Framework

Framework Well-Architected da AWS

Framework Well-Architected da AWS: Framework

Table of Contents

Framework Well-Architected da AWS

Data de publicação: 27 de junho de 2024 (Revisões do documento)

O AWS Well-Architected Framework ajuda a entender os prós e os contras das decisões que você toma ao criar sistemas na AWS. Ao usar o Framework, você terá acesso a práticas recomendadas de arquitetura para projetar e operar sistemas confiáveis, seguros, eficientes, econômicos e sustentáveis na nuvem.

Introdução

O AWS Well-Architected Framework ajuda a entender os prós e os contras das decisões que você toma ao criar sistemas na AWS. O uso do Framework ajuda você a aprender as práticas recomendadas de arquitetura para projetar e operar workloads seguras, confiáveis, eficientes, econômicas e sustentáveis na Nuvem AWS. Ele fornece uma maneira de avaliar de forma consistente suas arquiteturas em relação às práticas recomendadas e identificar áreas para melhorias. O processo para analisar uma arquitetura é uma conversa construtiva sobre decisões de arquitetura, e não um mecanismo de auditoria. Acreditamos que ter sistemas bem projetados aumenta significativamente a probabilidade de sucesso dos negócios.

Os arquitetos de soluções da AWS contam com vários anos de experiência em arquitetura de soluções em uma ampla variedade de segmentos de negócios verticais e casos de uso. Ajudamos a projetar e analisar as arquiteturas de milhares de clientes na AWS. Por meio dessa experiência, identificamos as práticas recomendadas e principais estratégias para a arquitetura de sistemas na nuvem.

O AWS Well-Architected Framework documenta um conjunto de perguntas fundamentais que ajudam a compreender se uma arquitetura específica se alinha bem às práticas recomendadas da nuvem. Ele fornece uma abordagem consistente para avaliar os sistemas em relação às qualidades que você espera dos sistemas modernos baseados em nuvem e a correção necessária para alcançar essas qualidades. À medida que a AWS continuar evoluindo e continuarmos a aprender mais com o trabalho com nossos clientes, aprimoraremos ainda mais a definição do Well-Architected.

Este Framework é destinado a pessoas que ocupam cargos de tecnologia, como diretores de tecnologia (CTOs), arquitetos, desenvolvedores e membros da equipe de operações. Ele descreve as práticas recomendadas e as estratégias da AWS a serem usadas ao projetar e operar uma workload na nuvem, além de fornecer links para detalhes de implementação e padrões de arquitetura

adicionais. Para obter mais informações, consulte a página inicial do AWS Well-Architected Framework.

A AWS também fornece um serviço gratuito para analisar suas workloads. O AWS Well-Architected Tool (AWS WA Tool) é um serviço na nuvem que fornece um processo consistente para analisar e medir a arquitetura usando o AWS Well-Architected Framework. O WA Tool da AWS fornece recomendações para tornar suas workloads mais confiáveis, seguras, eficientes e econômicas.

Para ajudar você a aplicar as práticas recomendadas, criamos os Laboratórios do AWS Well-Architected, os quais fornecem um repositório de código e documentação para ajudar a simplificar a implementação dessas práticas. Também nos juntamos a parceiros selecionados da Rede de Parceiros da AWS (APN) membros do Programa de Parceiros do AWS Well-Architected Esses parceiros da AWS têm profundo conhecimento sobre a AWS e podem ajudar você a analisar e melhorar suas workloads.

Definições

Todos os dias, os especialistas da AWS ajudam os clientes a projetar sistemas para aproveitar as práticas recomendadas na nuvem. Trabalhamos com você para oferecer vantagens e desvantagens arquitetônicas à medida que seus projetos evoluem. Conforme você implanta esses sistemas em ambientes dinâmicos, aprendemos como esses sistemas se desempenham e as consequências dessas vantagens e desvantagens.

Com base no que aprendemos, criamos o AWS Well-Architected Framework, que fornece um conjunto consistente de práticas recomendadas para os clientes e parceiros avaliarem arquiteturas e um conjunto de perguntas que você pode usar para avaliar o alinhamento de uma arquitetura com as práticas recomendadas da AWS.

O AWS Well-Architected Framework é baseado em seis pilares: excelência operacional, segurança, confiabilidade, eficiência de performance, otimização de custos e sustentabilidade.

Tabela 1. Os pilares do AWS Well-Architected Framework

Name (Nome)	Descrição
Excelência operacional	A capacidade de apoiar o desenvolvimento e executar workloads com eficácia, obter insights sobre as operações e melhorar continuamente

Name (Nome)	Descrição
	processos e procedimentos de suporte para oferecer valor empresarial.
Segurança	O pilar de segurança descreve como aproveitar as tecnologias de nuvem para proteger dados, sistemas e ativos de uma forma que possa melhorar sua postura de segurança.
Confiabilidade	O pilar Confiabilidade abrange a capacidade de uma workload de executar a função pretendida correta e consistentemente quando esperado. Isso inclui a capacidade de operar e testar a workload durante todo o ciclo de vida dela. Este documento fornece orientações detalhada s sobre as práticas recomendadas para a implementação de workloads confiáveis na AWS.
Eficiência de performance	A capacidade de usar recursos de computaçã o de forma eficiente para atender aos requisito s do sistema e manter essa eficiência à medida que a demanda muda e as tecnologias evoluem.
Otimização de custo	A capacidade de executar sistemas para agregar valor de negócio pelo menor preço.
Sustentabilidade	A possibilidade de melhorar continuamente os impactos sobre a sustentabilidade com a redução do consumo de energia e o aumento da eficiência de todos os componentes de uma workload por meio da maximização dos benefícios dos recursos provisionados e da minimização do total de recursos necessários.

No AWS Well-Architected Framework, usamos estes termos:

- Um componente é o código, configuração e recursos da AWS que, juntos, atendem a um requisito. Um componente geralmente é a unidade de propriedade técnica e é dissociada de outros componentes.
- O termo workload é usado para identificar um conjunto de componentes que entrega o valor empresarial. Uma workload é normalmente o nível de detalhes sobre o qual os líderes de negócios e tecnologia se comunicam.
- Pensamos na arquitetura como sendo a forma como os componentes trabalham juntos em uma workload. Como os componentes se comunicam e interagem é, com frequência, o foco dos diagramas de arquitetura.
- Os marcos assinalam as principais alterações em sua arquitetura à medida que evoluem ao longo do ciclo de vida do produto (design, implementação teste, ativação e produção).
- Dentro de uma organização, o portfólio de tecnologia é a coleção de workloads necessárias para o negócio operar.
- O nível de esforço refere-se à categorização da quantidade de tempo, esforço e complexidade que uma tarefa exige para implementação. Cada organização precisa considerar o tamanho e a especialização da equipe, além da complexidade da workload, a fim de ter contexto adicional para categorizar adequadamente o respectivo nível de esforço.
  - Alto: o trabalho pode demorar várias semanas ou vários meses. Ele poderia ser dividido em vários lançamentos, histórias e tarefas.
  - Médio: o trabalho pode demorar vários dias ou várias semanas. Ele poderia ser dividido em vários lançamentos e tarefas.
  - Baixo: o trabalho pode demorar várias horas ou vários dias. Ele poderia ser dividido em várias tarefas.

Ao arquitetar workloads, você obtém vantagens e desvantagens entre os pilares com base no contexto da sua empresa. Essas decisões comerciais podem determinar suas prioridades de engenharia. Você pode otimizar para melhorar o impacto sobre a sustentabilidade e reduzir os custos à custa da confiabilidade em ambientes de desenvolvimento ou, no caso de soluções essenciais à missão, otimizar a confiabilidade e aumentar os custos e o impacto sobre a sustentabilidade. Em soluções de comércio eletrônico, a performance pode afetar a receita e a propensão do cliente a comprar. Segurança e Excelência operacional geralmente não se envolvem em trocas com os outros pilares.

Sobre arquitetura

Em ambientes on-premises, os clientes geralmente têm uma equipe central de arquitetura de tecnologia que atua como uma sobreposição para outras equipes de produtos ou atributos para verificar se estão seguindo as práticas recomendadas. As equipes de arquitetura de tecnologia tipicamente incluem um conjunto de funções, como arquiteto técnico (infraestrutura), arquiteto de soluções (software), arquiteto de dados, arquiteto de redes e arquiteto de segurança. Muitas vezes, essas equipes usam o TOGAF ou o Zachman Framework como parte de um recurso de arquitetura corporativa.

Na AWS, preferimos distribuir os recursos entre equipes, em vez de termos uma equipe centralizada com esses recursos. Existem riscos na escolha de distribuir autoridade para tomada de decisões, por exemplo, verificar se as equipes atendem aos padrões internos. Atenuamos esses riscos de duas formas. Primeiro, adotamos práticas (processos, padrões, normas aceitas e formas de fazer as coisas) que têm como foco permitir que cada equipe tenha essa capacidade, e utilizamos especialistas que verificam se as equipes elevam o nível dos padrões que elas precisam cumprir. Em seguida, implementamos mecanismos que realizam verificações automatizadas para descobrir se os padrões estão sendo atendidos.

> "Boas intenções nunca funcionam, você precisa de bons mecanismos para fazer qualquer coisa acontecer", Jeff Bezos.

Isso significa substituir os melhores esforços humanos por mecanismos (muitas vezes automatizados) que examinam a conformidade com base em regras ou processos. Essa abordagem distribuída é apoiada pelos princípios de liderança da Amazon e estabelece uma cultura em todas as funções que é determinada pelo cliente do cliente. Trabalhar de trás para a frente é uma parte fundamental do nosso processo de inovação. Começamos com o cliente e o que ele quer, e deixamos isso definir e orientar nossos esforços. As equipes dedicadas ao cliente criam produtos em resposta a uma necessidade do cliente.

Para a arquitetura, isso significa que esperamos que todas as equipes tenham a capacidade de criar arquiteturas e seguir as práticas recomendadas. Para ajudar as novas equipes a obter essas capacidades ou as equipes existentes a elevar seus padrões, ativamos o acesso a uma comunidade virtual de engenheiros líderes que podem analisar os projetos e ajudá-las a entender quais são as práticas recomendadas da AWS. A comunidade de engenheiros líderes trabalha para que as práticas recomendadas sejam visíveis e acessíveis. Uma forma de fazer isso, por exemplo, é por meio de

palestras na hora do almoço focadas na aplicação das práticas recomendadas a exemplos reais. Essas conversas são gravadas e podem ser usadas como parte dos materiais de integração para novos membros da equipe.

As práticas recomendadas da AWS surgem de nossa experiência na execução de milhares de sistemas em escala da Internet. Preferimos usar dados para definir as práticas recomendadas, mas também usamos especialistas, como engenheiros líderes, para defini-las. À medida que os engenheiros líderes veem surgir novas práticas recomendadas, eles trabalham como uma comunidade para verificar se elas estão sendo seguidas pelas equipes. Com o tempo, essas práticas recomendadas são formalizadas em nossos processos internos de análise, bem como em mecanismos que reforçam a conformidade. O Well-Architected Framework é a implementação voltada para o cliente do nosso processo de análise interna, no qual codificamos nosso pensamento de engenharia principal nas funções de campo, como a arquitetura de soluções e equipes de engenharia internas. O Well-Architected Framework é um mecanismo escalável que permite que você aproveite esses aprendizados.

Seguindo a abordagem de uma comunidade de engenheiros líderes com propriedade distribuída de arquitetura, acreditamos que uma arquitetura corporativa do Well-Architected pode emergir, impulsionada pela necessidade do cliente. Líderes de tecnologia (como CTOs ou gerentes de desenvolvimento), realizando análises do Well-Architected em todas as workloads, permitirão uma melhor compreensão dos riscos no portfólio de tecnologia. Usando essa abordagem, você pode identificar temas entre as equipes que sua organização poderia abordar por mecanismos, treinamentos ou palestras na hora do almoço, em que seus engenheiros principais possam compartilhar seus pensamentos sobre áreas específicas com várias equipes.

Princípios gerais de projeto

O Well-Architected Framework identifica um conjunto de princípios gerais do projeto para facilitar um bom projeto na nuvem:

- Pare de tentar adivinhar suas necessidades de capacidade: se você tomar uma decisão ruim relacionada à capacidade ao implantar uma workload, poderá acabar com recursos ociosos caros ou lidando com as implicações da performance da capacidade limitada. Com a computação em nuvem, esses problemas terminaram. Você pode usar uma quantidade de capacidade qualquer de acordo com suas necessidades do momento e aumentar e diminuir a escala automaticamente.
- Teste seus sistemas em escala de produção: na nuvem, você pode criar um ambiente de teste em escala de produção sob demanda, concluir seus testes e desativar os recursos. Como você paga

somente pelo ambiente de teste quando está em execução, é possível simular seu ambiente ativo por uma fração do custo dos testes on-premises.

- Automatize com experimentação arquitetural em mente: a automação permite criar e replicar as workloads por custos baixos e evitar as despesas do trabalho manual. Você pode rastrear as alterações em sua automação, auditar o impacto e reverter para os parâmetros anteriores quando necessário.
- Considere arquiteturas evolucionárias: em um ambiente tradicional, as decisões de arquitetura são frequentemente implementadas como eventos estáticos e únicos, com algumas versões principais de um sistema durante sua vida útil. À medida que uma empresa e seu contexto continuam a evoluir, essas decisões iniciais podem prejudicar a capacidade do sistema de fornecer requisitos de negócios variáveis. Na nuvem, a capacidade de automatizar e testar sob demanda reduz o risco de impacto das alterações no projeto. Isso permite que os sistemas evoluam com o tempo, para que as empresas possam tirar proveito das inovações como prática padrão.
- Impulsione arquiteturas usando dados: na nuvem, você pode coletar dados sobre como suas escolhas de arquitetura afetam o comportamento da workload. Isso permite que você tome decisões baseadas em fatos sobre como melhorar sua workload. Sua infraestrutura de nuvem é código, portanto, você pode usar esses dados para informar suas escolhas e melhorias na arquitetura ao longo do tempo.
- Faça aprimoramentos com os game days: teste a performance e os processos de sua arquitetura, agendando regularmente dias de jogo para simular eventos em produção. Isso ajudará a compreender onde é possível aplicar melhorias e pode ajudar a desenvolver experiência organizacional ao lidar com eventos.

Os pilares do framework

Criar um sistema de software é como construir um edifício. Se a fundação não for sólida, problemas estruturais poderão prejudicar a integridade e a função do edifício. Ao arquitetar soluções de tecnologia, se você negligenciar os seis pilares (excelência operacional, segurança, confiabilidade, eficiência de performance, otimização de custos e sustentabilidade), poderá ser um desafio criar um sistema que atenda às suas expectativas e exigências. A incorporação desses pilares à sua arquitetura ajudará você a produzir sistemas estáveis e eficientes. Isso permitirá que você se concentre nos outros aspectos do projeto, como requisitos funcionais.

Pilares

- Excelência operacional
- Segurança
- Confiabilidade
- Eficiência de performance
- Otimização de custo
- Sustentabilidade

Excelência operacional

O pilar Excelência operacional inclui a capacidade de oferecer suporte ao desenvolvimento e de executar workloads com eficácia, obter insights sobre as operações e melhorar continuamente processos e procedimentos de suporte para oferecer valor empresarial.

O pilar Excelência operacional apresenta uma visão geral dos princípios de design, das práticas recomendadas e das perguntas. Recomendações sobre implementação estão disponíveis no whitepaper Pilar Excelência operacional.

Tópicos

- Princípios de design
- Definição
- Práticas recomendadas
- Recursos

Princípios de design

Os princípios de design para alcançar a excelência operacional na nuvem são:

- Organize as equipes em torno dos resultados comerciais: a capacidade de uma equipe alcançar resultados comerciais vem da visão de liderança, das operações eficazes e de um modelo operacional alinhado aos negócios. A liderança deve estar totalmente envolvida e comprometida com a transformação de CloudOps por meio de um modelo operacional de nuvem adequado que incentive as equipes a operar da maneira mais eficiente e atingir os resultados comerciais. O modelo operacional correto usa pessoas, processos e recursos tecnológicos para escalar, otimizar a produtividade e promover a diferenciação por meio de agilidade, capacidade de resposta e adaptação. A visão de longo prazo da organização é convertida em metas que são comunicadas em toda a empresa às partes interessadas e aos consumidores dos serviços de nuvem. As metas e os KPIs operacionais estão alinhados em todos os níveis. Essa prática sustenta o valor de longo prazo derivado da implementação dos princípios de design a seguir.
- Implemente observabilidade para insights acionáveis: obtenha uma compreensão abrangente do comportamento, da performance, da confiabilidade, do custo e da integridade da workload. Estabeleça indicadores-chave de performance (KPIs) e aproveite a telemetria de observabilidade para tomar decisões fundamentadas e agir imediatamente quando os resultados obtidos estiverem em risco. Melhore proativamente a performance, a confiabilidade e o custo com base em dados de observabilidade úteis.
- Automatize com segurança onde possível: na nuvem, você pode aplicar a mesma disciplina de engenharia usada para o código da aplicação em todo o ambiente. Você pode definir toda a workload e as respectivas operações (aplicações, infraestrutura, configuração e procedimentos) como código e atualizá-las. Em seguida, você pode automatizar as operações da workload iniciando-as em resposta a eventos. Na nuvem, você pode usar a segurança de automação configurando barreiras de proteção, incluindo controle de taxa, limites de erro e aprovações. Por meio de uma automação eficiente, você pode conseguir respostas consistentes a eventos, restringir erros humanos e reduzir o trabalho do operador.
- Faça alterações frequentes, pequenas e reversíveis: projete workloads escaláveis e com acoplamento fraco para permitir que os componentes sejam atualizados regularmente. Técnicas de implantação automatizadas, bem como mudanças menores e incrementais, reduzem o raio de expansão e permitem uma reversão mais rápida se ocorrerem falhas. Isso aumenta a confiança na entrega de mudanças benéficas à workload, mantendo a qualidade e possibilitando uma rápida adaptação às mudanças nas condições do mercado.

- Refine os procedimentos operacionais com frequência: à medida que você evolui suas workloads, desenvolva suas operações de forma adequada. À medida que usar procedimentos operacionais, procure oportunidades para melhorá-los. Organize revisões regularmente e valide se todos os procedimentos estão em vigor e se as equipes estão familiarizadas com eles. Ao identificar lacunas, atualize os procedimentos adequadamente. Comunique as atualizações dos procedimentos a todas as partes interessadas e equipes. Promova o aprendizado gamificado em suas operações para compartilhar as práticas recomendadas e instruir as equipes.
- Preveja a falha: maximize o sucesso operacional conduzindo cenários de falha para entender o perfil de risco da workload e seu impacto nos resultados comerciais. Teste a eficácia de seus procedimentos e a resposta de sua equipe em relação a essas falhas simuladas. Tome decisões embasadas para gerenciar riscos abertos identificados pelos testes.
- Aprenda com todos os eventos operacionais e métricas: promova melhorias com as lições aprendidas em todos os eventos e falhas operacionais. Compartilhe o que foi aprendido com as equipes e a organização inteira. Os aprendizados devem destacar dados e curiosidades sobre como as operações contribuem para os resultados comerciais.
- Use serviços gerenciados: reduza a carga operacional usando serviços gerenciados da AWS sempre que possível. Crie procedimentos operacionais em torno das interações com esses serviços.

Definição

Há quatro áreas de práticas recomendadas para excelência operacional na nuvem:

- Organização
- Preparar
- Operar
- Evoluir

A liderança da sua organização define os objetivos empresariais. Sua organização deve compreender requisitos e prioridades e usá-los para organizar e conduzir trabalhos para apoiar a obtenção de resultados empresariais. Sua workload deve emitir as informações necessárias para apoiá-la. A implementação de serviços para permitir a integração, a implantação e a entrega de sua workload criará um fluxo maior de alterações benéficas na produção por meio da automação de processos repetitivos.

Pode haver riscos inerentes à operação da workload. Compreenda esses riscos e tome uma decisão embasada para entrar em produção. Suas equipes devem ser capazes de oferecer suporte à sua workload. As métricas operacionais e de negócios derivadas dos resultados de negócios desejados permitirão que você compreenda a integridade da workload e das atividades operacionais enquanto você responde a incidentes. Suas prioridades mudarão à medida que suas necessidades de negócios e o ambiente de negócios mudarem. Use isso como um ciclo de comentários para promover continuamente melhorias para a sua organização e a operação da sua workload.

Práticas recomendadas

> **Note**
>
> Todas as perguntas de excelência operacional têm o prefixo OPS como abreviatura do pilar.

Tópicos

Organização

Suas equipes devem ter um entendimento comum de toda a sua workload, do papel que desempenham nela e dos objetivos de negócios compartilhados a fim de definir as prioridades que permitirão o êxito dos negócios. Prioridades bem definidas maximizarão os benefícios dos seus esforços. Avalie as necessidades de clientes internos e externos envolvendo as principais partes interessadas, incluindo equipes corporativas, de desenvolvimento e operacionais, a fim de determinar onde concentrar os esforços. A avaliação das necessidades do cliente verificará se você tem um entendimento completo do suporte necessário para obter resultados nos negócios. Esteja ciente das diretrizes ou obrigações definidas pela governança organizacional e de fatores externos, como requisitos de conformidade regulamentar e normas do setor, que podem exigir ou enfatizar um foco específico. Confirme se você tem os mecanismos para identificar alterações na governança interna e nos requisitos de conformidade externos. Se nenhum requisito for identificado, confirme se você aplicou a devida diligência para essa determinação. Analise suas prioridades regularmente para que elas possam ser atualizadas conforme as necessidades mudam.

Avalie ameaças à empresa (por exemplo, riscos e passivos empresariais e ameaças à segurança da informação) e mantenha essas informações em um registro de risco. Avalie o impacto dos riscos e as compensações entre interesses concorrentes ou abordagens alternativas. Por exemplo, a aceleração da velocidade de entrada no mercado de novos recursos pode ser enfatizada em relação à otimização de custos, ou você pode escolher um banco de dados relacional para dados não relacionais para simplificar o esforço de migração de um sistema. Gerencie benefícios e riscos para tomar decisões informadas ao determinar onde concentrar os esforços. Alguns riscos ou opções podem ser aceitáveis por um tempo. Talvez seja possível mitigar os riscos associados ou talvez seja inaceitável permitir que um risco permaneça; nesse caso, você tomará as devidas medidas para abordar o risco.

Suas equipes devem compreender o papel delas na obtenção de resultados empresariais. As equipes devem entender o papel delas no êxito de outras equipes e a função das outras equipes no êxito delas, além de ter objetivos comuns. Entender a responsabilidade, a propriedade, como as decisões são tomadas e quem tem autoridade para tomar decisões ajudará a concentrar os esforços e maximizar os benefícios das suas equipes. As necessidades de uma equipe são modeladas pelo cliente que ela auxilia, pela organização, pela formação da equipe e pelas características da workload. Não é sensato esperar que um modelo operacional único seja capaz de dar suporte a todas as equipes e suas respectivas workloads na sua organização.

Certifique-se de que haja proprietários identificados para cada componente de aplicação, workload, plataforma e infraestrutura, e que cada processo e procedimento tenha um proprietário identificado responsável pela definição e proprietários responsáveis pela performance.

Entender o valor empresarial de cada componente, processo e procedimento, da razão pela qual esses recursos estão em vigor ou de por que as atividades são executadas e por que essa propriedade existe informará as ações dos membros da equipe. Defina claramente as responsabilidades dos membros da equipe para que eles possam agir adequadamente e ter mecanismos para identificar responsabilidade e propriedade. Tenha mecanismos para solicitar adições, alterações e exceções para que você não restrinja a inovação. Defina contratos entre equipes que descrevem como elas trabalham juntas para apoiar umas às outras e seus resultados de negócios.

Forneça suporte aos membros da equipe para que eles possam ser mais eficazes na tomada de ações e no suporte aos resultados empresariais. A liderança sênior engajada deve definir expectativas e medir o sucesso. A liderança sênior deve ser a patrocinadora, a defensora e a motivadora da adoção das práticas recomendadas e da evolução da organização. Permita que os membros da equipe tomem medidas quando os resultados estiverem em risco, a fim de minimizar

o impacto, e os incentive a engajar os tomadores de decisão e as partes interessadas quando acharem que há algum risco, para resolvê-lo e evitar incidentes. Forneça comunicações oportunas, claras e acionáveis de riscos conhecidos e eventos planejados para que os membros da equipe possam tomar as medidas apropriadas e oportunas.

Incentive a experimentação para acelerar o aprendizado e manter os membros da equipe interessados e envolvidos. As equipes devem aumentar os conjuntos de habilidades para adotar novas tecnologias e apoiar mudanças na demanda e nas responsabilidades. Dê apoio e incentivo a isso, fornecendo um tempo estruturado e dedicado para o aprendizado. Garanta que os membros da equipe tenham os recursos (tanto ferramentas quanto pessoas) para serem bem-sucedidos e escalar para auxiliar os resultados empresariais. Aproveite a diversidade entre organizações para buscar várias perspectivas únicas. Use essa abordagem para aumentar a inovação, desafiar suas suposições e reduzir o risco de viés de confirmação. Aumente a inclusão, a diversidade e a acessibilidade em suas equipes para obter perspectivas benéficas.

Se houver requisitos externos de regulamentação ou conformidade aplicáveis à sua organização, use os recursos fornecidos pela Conformidade com a Nuvem AWS para ajudar a instruir suas equipes de modo que elas possam determinar o impacto em suas prioridades. O Well-Architected Framework enfatiza o aprendizado, a medição e a melhoria. Ele oferece uma abordagem consistente para avaliar arquiteturas e implementar designs que escalem ao longo do tempo. A AWS fornece o AWS Well-Architected Tool para ajudar você a analisar sua abordagem antes do desenvolvimento e o estado de suas workloads antes da produção e durante a produção. Você pode comparar as workloads com as práticas recomendadas de arquitetura da AWS mais recentes, monitorar seu status geral e obter insights sobre possíveis riscos. O AWS Trusted Advisor é uma ferramenta que fornece acesso a um conjunto essencial de verificações que recomendam otimizações capazes de ajudar a moldar suas prioridades. Os clientes Business e Enterprise Support recebem acesso a verificações adicionais com foco em segurança, confiabilidade, performance, otimização de custos e sustentabilidade que podem ajudar a moldar suas prioridades.

A AWS pode ajudar a instruir suas equipes sobre a AWS e os serviços que ela fornece para que compreendam melhor como as escolhas que elas fazem podem ter um impacto na workload. Use os recursos fornecidos pelo AWS Support (Centro de Conhecimentos da AWS, Fóruns de discussão da AWS e o AWS Support Center), bem como a documentação da AWS, para instruir suas equipes. Entre em contato com o AWS Support por meio do AWS Support Center para obter ajuda com relação a dúvidas sobre a AWS. A AWS também compartilha as práticas recomendadas e os padrões que aprendemos durante a operação da AWS na Amazon Builders' Library. Inúmeras outras informações úteis podem ser obtidas por meio do Blog da AWS e no podcast oficial da AWS. AWS A Training and Certification oferece treinamento por meio de cursos digitais autoguiados sobre os

fundamentos da AWS. Você também pode se inscrever em treinamento administrado por instrutor a fim de oferecer suporte adicional às suas equipes para o desenvolvimento de habilidades em serviços da AWS.

Usar ferramentas ou serviços que permitam controlar centralmente seus ambientes em todas as contas, como o AWS Organizations, para ajudar a gerenciar seus modelos operacionais. Serviços como o AWS Control Tower expandem esse recurso de gerenciamento, permitindo que você defina esquemas (compatíveis com modelos operacionais) para a configuração de contas, aplique governança contínua usando o AWS Organizations e automatize o provisionamento de novas contas. Provedores de serviços gerenciados como o AWS Managed Services e parceiros da AWS Managed Services ou os provedores de serviços gerenciados na Rede de Parceiros da AWS fornecem especialização na implementação de ambientes de nuvem e ajudam a atender os seus requisitos de segurança e conformidade e objetivos de negócios. A adição de serviços gerenciados ao seu modelo operacional pode economizar tempo e recursos, além de permitir que você mantenha as equipes internas reduzidas e focadas em resultados estratégicos que diferenciarão seus negócios, em vez de desenvolver novas habilidades e recursos.

As perguntas a seguir referem-se a essas considerações de excelência operacional. (Para obter uma lista de perguntas e práticas recomendadas de excelência operacional, consulte o Apêndice.)

OPS 1: Como determinar quais são suas prioridades?

Todas as pessoas devem compreender o papel delas na conquista do sucesso empresarial. Tenha objetivos compartilhados para definir as prioridades dos recursos. Isso maximizará os benefícios de seus esforços.

OPS 2 : Como estruturar sua organização para oferecer suporte aos seus resultados comerciais?

Suas equipes devem compreender o papel delas na obtenção de resultados empresariais. As equipes devem entender o papel delas no êxito de outras equipes e a função das outras equipes no êxito delas, além de ter objetivos comuns. Entender a responsabilidade, a propriedade, como as decisões são tomadas e quem tem autoridade para tomar decisões ajudará a concentrar os esforços e maximizar os benefícios das suas equipes.

OPS 3: Como a cultura organizacional oferece suporte aos resultados comerciais?

Forneça suporte aos membros da equipe para que eles possam ser mais eficazes na tomada de ações e no suporte aos resultados comerciais.

Em determinado momento, talvez você queira destacar um pequeno subconjunto de prioridades. Use uma abordagem equilibrada em longo prazo para garantir o desenvolvimento dos recursos necessários e o gerenciamento de riscos. Reveja as prioridades regularmente e atualize-as conforme as necessidades mudarem. Quando a responsabilidade e a propriedade não foram definidas ou não são conhecidas, você corre o risco de não realizar as ações necessárias em tempo hábil e de despender esforços redundantes e possivelmente conflitantes para atender a essas necessidades. A cultura organizacional tem impacto direto na satisfação com a tarefa e na retenção dos membros da equipe. Incentive o envolvimento e as habilidades dos membros da equipe para promover o sucesso da sua empresa. A experimentação é necessária para que a inovação ocorra e transforme ideias em resultados. Reconheça que um resultado indesejado é um experimento com êxito que identificou um caminho que não levará ao êxito.

Preparar

Para se preparar para a excelência operacional, é necessário entender suas workloads e os comportamentos esperados. Você poderá projetá-las para fornecer insights sobre seu status e criar os procedimentos para oferecer suporte a elas.

Projete sua workload para que as informações necessárias sejam fornecidas a fim de que você entenda seu estado interno (tais como métricas, logs, eventos e rastreamento) em todos os componentes, em apoio à observabilidade e à investigação de problemas. A observabilidade vai além do simples monitoramento, fornecendo uma compreensão abrangente do funcionamento interno de um sistema com base em suas saídas externas. Baseada em métricas, logs e rastreamentos, a observabilidade oferece insights profundos sobre o comportamento e a dinâmica do sistema. Com uma observabilidade eficaz, as equipes podem discernir padrões, anomalias e tendências, permitindo que abordem proativamente possíveis problemas e mantenham a integridade ideal do sistema. Identificar os indicadores-chave de performance (KPIs) é fundamental para garantir o alinhamento entre as atividades de monitoramento e os objetivos de negócios. Esse alinhamento garante que as equipes tomem decisões baseadas em dados usando métricas que realmente importam, otimizando a performance do sistema e os resultados comerciais. Além disso, a observabilidade capacita as empresas a serem proativas em vez de reativas. As equipes podem entender as relações de causa e efeito em seus sistemas, prevendo e prevenindo problemas em

vez de apenas reagir a eles. À medida que as workloads evoluem, é essencial revisitar e refinar a estratégia de observabilidade, garantindo que ela permaneça relevante e eficaz.

Adote abordagens que melhorem o fluxo de alterações na produção e permitam refatoração, feedbck rápido sobre a qualidade e correção de erros. Isso acelera as alterações benéficas que entram na produção, limita os problemas implantados e permite a rápida identificação e correção dos problemas introduzidos pelas atividades de implantação ou descobertos em seus ambientes.

Adote abordagens que forneçam feedback rápido sobre a qualidade e permitam recuperação rápida de alterações que não têm os resultados desejados. O uso dessas práticas reduz o impacto dos problemas introduzidos pela implantação de mudanças. Planeje alterações malsucedidas para que você possa responder mais rapidamente, se necessário, e testar e validar as alterações feitas. Mantenha-se a par das atividades planejadas em seus ambientes para que você possa gerenciar o risco de alterações que afetem as atividades planejadas. Enfatize alterações frequentes, pequenas e reversíveis para limitar o escopo das alterações. Isso resulta em solução de problemas e correção mais rápidas, com a opção de reverter uma alteração. Isso também significa que você pode conseguir o benefício de alterações valiosas com mais frequência.

Avalie a prontidão operacional de workload, processos, procedimentos e pessoal para compreender os riscos operacionais relacionados à workload. Use um processo consistente (incluindo listas de verificação manuais ou automatizadas) para saber quando você estiver pronto para trabalhar com sua workload ou fazer uma mudança. Isso também ajudará a encontrar as áreas que você deve abordar. Tenha runbooks que documentem suas atividades de rotina e playbooks que orientem seus processos para a resolução de problemas. Entenda os benefícios e os riscos para tomar decisões informadas e permitir que as alterações entrem na produção.

A AWS permite visualizar toda a workload (aplicações, infraestrutura, políticas, governança e operações) como código. Isso significa que você pode aplicar a mesma disciplina de engenharia usada para o código da aplicação a cada elemento da pilha e compartilhá-los entre equipes ou organizações para ampliar os benefícios dos esforços de desenvolvimento. Use operações como código na nuvem e a capacidade de experimentar com segurança para desenvolver sua workload, procedimentos de operações e praticar falhas. O uso do AWS CloudFormation permite que você tenha ambientes consistentes, com modelos, desenvolvimento de sandbox, teste e produção, com níveis crescentes de controle de operações.

As perguntas a seguir referem-se a essas considerações de excelência operacional.

OPS 4: Como implementar a observabilidade em sua workload?

Implemente a observabilidade na workload para poder entender seu estado e tomar decisões baseadas em dados com base nos requisitos de negócios.

OPS 5: Como reduzir defeitos, facilitar a correção e melhorar o fluxo na produção?

Adote abordagens que melhorem o fluxo de alterações na produção e permitam refatoração, feedback rápido sobre a qualidade e correção de erros. Isso acelera as alterações benéficas que entram na produção, limita os problemas implantados e alcança a rápida identificação e correção dos problemas introduzidos pelas atividades de implantação.

OPS 6: Como reduzir os riscos de implantação?

Adote abordagens que forneçam feedback rápido sobre a qualidade e alcancem recuperaç ão rápida de alterações que não têm os resultados desejados. O uso dessas práticas reduz o impacto dos problemas introduzidos pela implantação de mudanças.

OPS 7: Como saber se está tudo pronto para oferecer suporte a uma workload?

Avalie a prontidão operacional de sua workload, processos/procedimentos e pessoal para entender os riscos operacionais relacionados.

Invista na implementação de atividades operacionais como código para maximizar a produtividade do pessoal de operações, minimizar taxas de erro e permitir respostas automatizadas. Use estratégias "pre-mortem" para antecipar falhas e criar procedimentos, quando apropriado. Aplique metadados usando tags de recursos e AWS Resource Groups seguindo uma estratégia consistente de marcação com tags para identificar seus recursos. Identifique seus recursos de organização, contabilidade de custos e controles de acesso pensando na execução de atividades operacionais automatizadas. Adote práticas de implantação que aproveitem a elasticidade da nuvem para facilitar as atividades de desenvolvimento e a pré-implantação de sistemas para implementações mais rápidas. Ao fazer alterações nas listas de verificação usadas para avaliar suas workloads, planeje o que você fará com sistemas ativos que não estejam mais em conformidade.

Operar

A observabilidade permite que você se concentre em dados significativos e entenda as interações e os resultados da sua workload. Ao se concentrar em informações essenciais e eliminar dados desnecessários, você mantém uma abordagem direta para entender a performance da workload. É essencial não apenas coletar dados, mas também interpretá-los corretamente. Defina linhas de base claras e limites de alerta apropriados e monitore ativamente quaisquer desvios. Uma mudança em uma métrica-chave, especialmente quando correlacionada com outros dados, pode identificar áreas problemáticas específicas. Com a observabilidade, você está mais bem equipado para prever e enfrentar possíveis desafios, garantindo que sua workload opere sem problemas e atenda às necessidades de negócios.

A operação bem-sucedida de uma workload é medida pela obtenção de resultados de negócios e de clientes. Defina os resultados esperados, determine como o sucesso será medido e identifique as métricas que serão usadas nesses cálculos para determinar se a workload e as operações foram bem-sucedidas. A integridade operacional inclui a integridade da workload e a integridade e o sucesso de operações realizadas em apoio à workload (por exemplo, implantação e resposta a incidentes). Estabeleça linhas de base de métricas para melhoria, investigação e intervenção, colete e analise as métricas e valide seu entendimento sobre o sucesso das operações e como elas mudam ao longo do tempo. Use as métricas coletadas para determinar se você está satisfazendo as necessidades do cliente e da empresa e identifique áreas para melhoria.

É necessário um gerenciamento eficiente e eficaz dos eventos operacionais para alcançar a excelência operacional. Isso se aplica a eventos operacionais planejados e não planejados. Use runbooks estabelecidos para eventos bem compreendidos e use playbooks para ajudar na investigação e na resolução de problemas. Priorize respostas a eventos com base no impacto nos negócios e no cliente. Assegure que, caso um alerta seja gerado em resposta a um evento, exista um processo associado a ser executado com um proprietário especificamente identificado. Defina com antecedência o pessoal necessário para resolver um evento e inclua processos de encaminhamento para envolver pessoal adicional, conforme necessário, com base na urgência e no impacto. Identifique e envolva indivíduos com autoridade para tomar uma decisão sobre cursos de ação em que haverá um impacto nos negócios resultante de uma resposta de evento não abordada anteriormente.

Comunique o status operacional das workloads por meio de painéis e notificações adaptadas ao público-alvo (por exemplo, cliente, empresa, desenvolvedores, operações) para que eles possam tomar as ações adequadas, para que suas expectativas sejam gerenciadas e para que sejam informados quando as operações normais forem retomadas.

Na AWS, você pode gerar visualizações do painel sobre as métricas coletadas das workloads e nativamente na AWS. Você pode utilizar o CloudWatch ou aplicações de terceiros para agregar e apresentar visualizações das atividades operacionais em nível de negócios, workloads e operações. A AWS fornece insights das workloads por meio de recursos de registro em log como o AWS X-Ray, o CloudWatch, o CloudTrail e os Logs de fluxo da VPC para identificar problemas nas workloads a fim de ajudar na análise e correção da causa-raiz.

As perguntas a seguir referem-se a essas considerações de excelência operacional.

OPS 8: Como utilizar a observabilidade de workloads em sua organização?

Garanta a integridade ideal da workload usando a observabilidade. Utilize métricas, logs e rastreamentos relevantes para obter uma visão abrangente da performance da sua workload e resolver problemas com eficiência.

OPS 9: Como compreender a integridade das suas operações?

Defina, capture e analise as métricas de operações para obter visibilidade dos eventos de operações, para que você possa tomar as ações apropriadas.

OPS 10: Como gerenciar os eventos de workload e operações?

Prepare e valide procedimentos para responder a eventos, com o objetivo de minimizar a interrupção de sua workload.

Todas as métricas coletadas devem estar alinhadas a uma necessidade de negócios e aos resultados que elas auxiliam. Desenvolva respostas com script para eventos bem compreendidos e automatize a performance deles em resposta ao reconhecimento do evento.

Evoluir

Aprenda, compartilhe e melhore continuamente para manter a excelência operacional. Dedique ciclos de trabalho para fazer melhorias incrementais quase contínuas. Execute uma análise pós-incidente de todos os eventos que afetam o cliente. Identifique os fatores que contribuem e a ação preventiva para limitar ou evitar a recorrência. Comunique fatores contribuintes às comunidades

afetadas, conforme adequado. Avalie e priorize regularmente oportunidades de melhoria (por exemplo, solicitações de recursos, correção de problemas e requisitos de conformidade), incluindo a workload e os procedimentos operacionais.

Inclua ciclos de feedback nos procedimentos para identificar rapidamente áreas que podem ser melhoradas e aprender com a execução das operações.

Compartilhe as lições aprendidas entre as equipes para compartilhar os benefícios dessas lições. Analise as tendências nas lições aprendidas e execute análises retrospectivas entre as equipes de métricas de operações para identificar oportunidades e métodos de melhoria. Implemente alterações destinadas a trazer melhorias e avaliar os resultados para determinar o sucesso.

Na AWS, você pode exportar seus dados de log para o Amazon S3 ou enviar logs diretamente para o Amazon S3 para armazenamento de longo prazo. Usando o AWS Glue, você pode descobrir e preparar dados de log no Amazon S3 para estudo analítico, e armazenar metadados associados no AWS Glue Data Catalog. O Amazon Athena, por meio da integração nativa com o AWS Glue, pode ser usado para analisar dados de log, consultando-os com o SQL padrão. Uma ferramenta de inteligência de negócios como o Amazon QuickSight permite visualizar, explorar e analisar dados. Descoberta de tendências e eventos de interesse que podem promover melhorias.

A pergunta a seguir concentra-se nessas considerações de excelência operacional.

OPS 11: O que deve ser feito para que as operações evoluam?
Dedique tempo e recursos para a melhoria incremental praticamente contínua a fim de aumentar a eficácia e a eficiência das suas operações.

A evolução bem-sucedida das operações baseia-se em: pequenas melhorias frequentes; fornecer ambientes seguros e tempo para experimentar, desenvolver e testar melhorias; e ambientes em que o aprendizado com falhas é incentivado. O suporte de operações de ambientes de sandbox, desenvolvimento, teste e produção, com nível crescente de controles operacionais, facilita o desenvolvimento e aumenta a previsibilidade de resultados bem-sucedidos das alterações implementadas na produção.

Recursos

Consulte os recursos a seguir para saber mais sobre nossas práticas recomendadas para Excelência operacional.

Documentação

- DevOps e AWS

Whitepaper

- Pilar Excelência operacional

Vídeo

- DevOps na Amazon

Segurança

O pilar Segurança refere-se à capacidade de proteger dados, sistemas e ativos para utilizar as tecnologias de nuvem para melhorar sua segurança.

O pilar Segurança apresenta uma visão geral dos princípios de design, práticas recomendadas e perguntas. Recomendações sobre implementação estão disponíveis no whitepaper Pilar Segurança.

Tópicos

- Princípios de design
- Definição
- Práticas recomendadas
- Recursos

Princípios de design

Na nuvem, existem vários princípios que podem ajudá-lo a fortalecer a segurança da workload:

- Implementar uma base sólida de identidade: implemente o princípio do privilégio mínimo e separe as tarefas com autorização apropriada para cada interação com os recursos da AWS. Centralize o gerenciamento de identidades e procure eliminar a dependência de credenciais estáticas de longo prazo.

- Manter a rastreabilidade: monitore, alerte e examine ações e alterações em seu ambiente em tempo real. Integre a coleta de logs e métricas aos sistemas para investigar e executar ações automaticamente.
- Aplicar segurança em todas as camadas: aplique uma abordagem de defesa detalhada com vários controles de segurança. Aplique a todas as camadas (por exemplo, borda da rede, VPC, balanceamento de carga, cada instância e serviço de computação, sistema operacional, aplicação e código).
- Automatizar as práticas recomendadas de segurança: os mecanismos de segurança automatizados baseados em software aprimoram sua capacidade de dimensionar com segurança e de forma mais rápida e econômica. Crie arquiteturas seguras, incluindo a implementação de controles definidos e gerenciados como código em modelos controlados por versão.
- Proteger dados em trânsito e em repouso: classifique seus dados em níveis de confidencialidade e use mecanismos, como criptografia, tokenização e controle de acesso, quando apropriado.
- Manter as pessoas afastadas dos dados: crie mecanismos e ferramentas para reduzir ou eliminar a necessidade de acesso direto ou processamento manual de dados. Isso reduz o risco de erros de processamento ou modificação e erro humano ao manipular dados confidenciais.
- Preparar para eventos de segurança: prepare-se para um incidente com políticas e processos de gerenciamento e investigação de incidentes alinhados aos requisitos organizacionais. Execute simulações de resposta a incidentes e use ferramentas com automação para aumentar sua velocidade de identificação, investigação e recuperação.

Definição

Há sete áreas de práticas recomendadas para segurança na nuvem.

- Fundamentos de segurança
- Gerenciamento de identidade e acesso
- Detecção
- Proteção da infraestrutura
- Proteção de dados
- Resposta a incidentes
- Segurança da aplicação

Antes de projetar qualquer workload, coloque em vigor práticas que influenciem a segurança. Controle quem pode fazer o quê. Além disso, é útil conseguir identificar incidentes de segurança, proteger seus sistemas e serviços e manter a confidencialidade e a integridade dos dados por meio de proteção de dados. Você deve ter um processo bem definido e treinado para responder a incidentes de segurança. Essas ferramentas e técnicas são importantes porque ajudam a sustentar objetivos como evitar perdas financeiras ou cumprir obrigações regulatórias.

O modelo de responsabilidade compartilhada da AWS permite que as organizações que adotam a nuvem alcancem suas metas de segurança e conformidade. Como a AWS protege fisicamente a infraestrutura que suporta nossos serviços em nuvem, como cliente da AWS, você pode se concentrar no uso de serviços para atingir seus objetivos. A Nuvem AWS também oferece maior acesso aos dados de segurança e uma abordagem automatizada para responder a eventos de segurança.

Práticas recomendadas

Tópicos

Segurança

A pergunta a seguir concentra-se nessas considerações sobre segurança. (Para obter uma lista de perguntas e práticas recomendadas de segurança, consulte o Apêndice.)

SEC 1: Como operar com segurança sua workload?

Para operar seu workload com segurança, você deve aplicar as práticas recomendadas abrangentes em todas as áreas de segurança. Pegue os requisitos e processos que você definiu em excelência operacional em um nível organizacional e de workload e aplique-os a todas as áreas.

SEC 1: Como operar com segurança sua workload?

Manter-se atualizado com as recomendações da AWS, fontes do setor e a inteligência de ameaças ajuda você a desenvolver seu modelo de ameaças e objetivos de controle. A automação dos processos, testes e validação de segurança permite que você escale suas operações de segurança.

Na AWS, a segregação de workloads diferentes por conta, com base na respectiva função e nos requisitos de conformidade ou confidencialidade de dados, é uma abordagem recomendada.

Gerenciamento de identidade e acesso

O gerenciamento de identidades e acesso é parte essencial de um programa de segurança da informação, que garante que apenas usuários autorizados e autenticados possam acessar seus recursos e somente da forma que você pretender. Por exemplo, você deve definir entidades principais (ou seja, contas, usuários, funções e serviços que podem executar ações em sua conta), criar políticas alinhadas com essas entidades principais e implementar um gerenciamento forte de credenciais. Esses elementos de gerenciamento de privilégios formam o núcleo da autenticação e autorização.

Na AWS, o gerenciamento de privilégios é compatível principalmente com o serviço AWS Identity and Access Management (IAM), que permite controlar o acesso do usuário e programático a produtos e recursos da AWS. Você deve aplicar políticas granulares, que atribuem permissões a um usuário, grupo, função ou recurso. Você também pode exigir práticas de senha forte, como nível de complexidade, evitando reutilização e impondo autenticação multifator (MFA). Você pode usar federação com seu serviço de diretório atual. Para workloads que exigem que os sistemas tenham acesso à AWS, o IAM possibilita acesso seguro por meio de perfis, perfis de instância, federação de identidades e credenciais temporárias.

As perguntas a seguir referem-se a essas considerações sobre segurança.

SEC 2: Como gerenciar identidades para pessoas e máquinas?

Há dois tipos de identidades que você precisa gerenciar ao abordar a operação de workloads seguros da AWS. Entender o tipo de identidade de que você precisa para gerenciar e conceder acesso ajuda a garantir que as identidades corretas tenham acesso aos recursos certos nas condições certas.

SEC 2: Como gerenciar identidades para pessoas e máquinas?

Identidades humanas: seus administradores, desenvolvedores, operadores e usuários finais precisam de uma identidade para acessar seus ambientes e aplicações da AWS. Eles são membros da sua organização ou usuários externos com quem você colabora e que interagem com seus recursos da AWS por meio de um navegador da Web, aplicação do cliente ou ferramentas interativas de linha de comando.

Identidades de máquina: aplicações de serviço, ferramentas operacionais e workloads precisam de uma identidade para fazer solicitações a serviços da AWS, como para ler dados. Essas identidades incluem máquinas em execução em seu ambiente da AWS, como instâncias do Amazon EC2 ou funções do AWS Lambda. Você também pode gerenciar identidades de máquina para partes externas que precisam de acesso. Além disso, você pode ter máquinas fora da AWS que precisam de acesso aos seus ambientes da AWS.

SEC 3: Como gerenciar permissões para pessoas e máquinas?

Gerencie permissões para controlar o acesso a identidades de pessoas e máquinas que precisam de acesso à AWS e à sua workload. As permissões controlam quem pode acessar o quê e em quais condições.

As credenciais não devem ser compartilhadas entre usuários ou sistemas. O acesso do usuário deve ser concedido usando uma abordagem de privilégio mínimo, com práticas recomendadas que incluem requisitos de senha e imposição de MFA. O acesso programático, incluindo chamadas à API a serviços da AWS, deve ser realizado usando credenciais de privilégio limitado e temporárias como aquelas emitidas pelo AWS Security Token Service.

Os usuários precisam de acesso programático se quiserem interagir com a AWS de fora do AWS Management Console. A forma de conceder acesso programático depende do tipo de usuário que está acessando a AWS.

Para conceder acesso programático aos usuários, selecione uma das seguintes opções:

Qual usuário precisa de acesso programático?	Para	Por
Identificação da força de trabalho (Usuários gerenciados no Centro de Identidade do IAM)	Use credenciais temporárias para assinar solicitações programáticas para a AWS CLI, os SDKs da AWS ou as APIs da AWS.	Siga as instruções da interface que deseja utilizar. • Para a AWS CLI, consulte Configuração da AWS CLI para usar o AWS IAM Identity Center no Guia do usuário da AWS Command Line Interface. • Para os SDKs da AWS, ferramentas e APIs da AWS, consulte Autenticação do Centro de Identidade do IAM no Guia de referência de ferramentas e SDKs da AWS.
IAM	Use credenciais temporárias para assinar solicitações programáticas para a AWS CLI, os SDKs da AWS ou as APIs da AWS.	Siga as instruções em Como usar credenciais temporárias com recursos da AWS no Guia do usuário do IAM.
IAM	(Não recomendado) Use credenciais de longo prazo para assinar solicitações programáticas para a AWS CLI, os SDKs da AWS ou as APIs da AWS.	Siga as instruções da interface que deseja utilizar. • Para a AWS CLI, consulte Autenticação usando as credenciais de usuário do IAM no Guia do usuário da AWS Command Line Interface. • Para as ferramentas e SDKs da AWS, consulte

Qual usuário precisa de acesso programático?	Para	Por
		Autenticação usando as credenciais de longo prazo no Guia de referência de ferramentas e SDKs da AWS. • Para as APIs da AWS, consulte Gerenciamento de chaves de acesso de usuários do IAM no Guia do usuário do IAM.

A AWS fornece recursos que podem ajudá-lo no Identity and Access Management. Para ajudar a aprender as práticas recomendadas, explore nossos laboratórios práticos sobre gerenciamento de credenciais e autenticação, controle de acesso humano e controle de acesso programático.

Detecção

É possível usar controles de detecção para identificar uma potencial ameaça ou incidente de segurança. Eles são uma parte essencial das estruturas de governança e podem ser usados para apoiar um processo de qualidade, uma obrigação legal ou de conformidade e para os esforços de identificação e resposta a ameaças. Existem diferentes tipos de controles de detecção. Por exemplo, a realização de um inventário de ativos e seus atributos detalhados promove tomadas de decisão mais eficazes (e controles de ciclo de vida) para ajudar a estabelecer linhas de base operacionais. Você também pode usar a auditoria interna, um exame dos controles relacionados aos sistemas de informação, para garantir que as práticas atendam às políticas e aos requisitos e que você tenha definido as notificações de alerta automatizadas corretas com base nas condições definidas. Esses controles são fatores reativos importantes que podem ajudar sua organização a identificar e entender o escopo da atividade anômala.

Na AWS, você pode implementar controles de detecção processando logs, eventos e monitoramento que possibilitam auditoria, análise automatizada e alarmes. Os logs do CloudTrail, as chamadas à AWS API e o CloudWatch fornecem o monitoramento de métricas com alarmes, enquanto o AWS Config fornece o histórico de configuração. O Amazon GuardDuty é um serviço de detecção de ameaças gerenciado que monitora continuamente comportamentos mal-intencionados ou não

autorizados para ajudar a proteger contas e workloads da AWS. Logs em nível de serviço também estão disponíveis, por exemplo, você pode usar o Amazon Simple Storage Service (Amazon S3) para registrar em log as solicitações de acesso.

A pergunta a seguir concentra-se nessas considerações sobre segurança.

SEC 4: Como detectar e investigar eventos de segurança?

Capture e analise eventos de logs e métricas para obter visibilidade. Tomar medidas em relação aos eventos de segurança e possíveis ameaças para ajudar a proteger seu workload.

O gerenciamento de log é importante para uma workload do Well-Architected por motivos que vão de segurança ou análise forense a requisitos regulatórios ou legais. É fundamental analisar os logs e responder a eles para poder identificar possíveis incidentes de segurança. A AWS fornece funcionalidade que facilita a implementação do gerenciamento de logs, oferecendo a capacidade de definir um ciclo de vida de retenção de dados ou definir onde os dados serão preservados, arquivados ou eventualmente excluídos. Isso torna o processamento de dados previsível e confiável mais simples e econômico.

Proteção da infraestrutura

A proteção de infraestrutura abrange metodologias de controle, como defesa em profundidade, necessárias para atender às práticas recomendadas e obrigações organizacionais ou regulatórias. O uso dessas metodologias é fundamental para operações contínuas bem-sucedidas na nuvem ou on-premises.

Na AWS, é possível implementar inspeção de pacote stateful e stateless, seja usando tecnologias nativas da AWS ou produtos e serviços de parceiros disponíveis por meio do AWS Marketplace. Você deve usar o Amazon Virtual Private Cloud (Amazon VPC) para criar um ambiente privado, protegido e escalável em que seja possível definir sua topologia, incluindo gateways, tabelas de roteamento e sub-redes públicas e privadas.

As perguntas a seguir referem-se a essas considerações sobre segurança.

SEC 5: Como você protege seus recursos de rede?

Qualquer workload que tenha alguma forma de conectividade de rede, seja a Internet ou uma rede privada, exige várias camadas de defesa para ajudar a proteger contra ameaças externas e internas baseadas em rede.

SEC 6: Como você protege seus recursos de computação?

Os recursos computacionais em seu workload exigem várias camadas de defesa para ajudar na proteção contra ameaças externas e internas. Os recursos de computação incluem instâncias do EC2, contêineres, funções do AWS Lambda, serviços de banco de dados, dispositivos de IoT e muito mais.

É aconselhável usar várias camadas de defesa em qualquer tipo de ambiente. No caso de proteção de infraestrutura, muitos dos conceitos e métodos são válidos em modelos no on-premises e na nuvem. Impor proteção de limites, monitorar pontos de entrada e saída e registro em log, monitoramento e geração de alertas abrangentes são medidas essenciais para um plano eficaz de segurança da informação.

Os clientes da AWS são capazes de personalizar, ou reforçar, a configuração de um Amazon Elastic Compute Cloud (Amazon EC2), de um contêiner do Amazon Elastic Container Service (Amazon ECS) ou de uma instância do AWS Elastic Beanstalk, além de manter essa configuração em uma imagem de máquina da Amazon (AMI) imutável. Ao serem iniciados pelo Auto Scaling ou iniciados manualmente, todos os novos servidores virtuais (instâncias) iniciados com esse AMI recebem a configuração reforçada.

Proteção de dados

Antes de criar a arquitetura de qualquer sistema, práticas fundamentais que influenciam a segurança devem ser adotadas. Por exemplo, a classificação de dados fornece uma maneira de categorizar os dados organizacionais com base nos níveis de sensibilidade, e a criptografia protege os dados ao torná-los ininteligíveis ao acesso não autorizado. Essas ferramentas e técnicas são importantes porque ajudam a sustentar objetivos como evitar perdas financeiras ou cumprir obrigações regulatórias.

Na AWS, as seguintes práticas facilitam a proteção de dados:

- Como um cliente da AWS, você mantém controle total de seus dados.
- A AWS torna mais fácil criptografar e gerenciar chaves, incluindo a rotação regular de chaves, que pode ser facilmente automatizada pela AWS ou mantida por você.
- O registro em log detalhado com conteúdo importante, como acesso e alterações a arquivo, está disponível.
- A AWS desenvolveu sistemas de armazenamento para resiliência excepcional. Por exemplo, o Amazon S3 Standard, o S3 Standard – IA, o S3 One Zone-IA e o Amazon Glacier são todos projetados para oferecer 99,999999999% de durabilidade de objetos em um determinado ano. Esse nível de durabilidade corresponde a uma perda anual média esperada de 0,000000001% dos objetos.
- O versionamento, que pode fazer parte de um processo de gerenciamento de ciclo de vida de dados maior, é capaz de oferecer proteção contra substituições, exclusões e danos similares inadvertidos.
- A AWS nunca inicia a movimentação de dados entre regiões. O conteúdo colocado em uma região permanecerá nessa região, a menos que você explicitamente habilite um recurso ou utilize um serviço que desempenhe essa funcionalidade.

As perguntas a seguir referem-se a essas considerações sobre segurança.

SEC 7: Como você classifica seus dados?

A classificação fornece uma maneira de categorizar dados, com base na criticidade e sensibili dade, para ajudar você a determinar controles apropriados de proteção e retenção.

SEC 8: Como proteger seus dados em repouso?

Proteja seus dados em repouso implementando vários controles para reduzir o risco de acesso não autorizado ou manuseio incorreto.

SEC 9: Como proteger seus dados em trânsito?

Proteja seus dados em trânsito implementando vários controles para reduzir o risco de acesso ou perda não autorizados.

A AWS oferece vários meios de criptografar dados em repouso e em trânsito. Integramos recursos em nossos serviços que tornam mais fácil criptografar seus dados. Por exemplo, implementamos criptografia no lado do servidor (SSE) para o Amazon S3 a fim de facilitar para você armazenar seus dados em um formato criptografado. Você também pode providenciar que todo o processo de criptografia e descriptografia HTTPS (geralmente conhecido como terminação SSL) seja processado por Elastic Load Balancing (ELB).

Resposta a incidentes

Mesmo com controles preventivos e de detecção consolidados, sua organização ainda deve implementar processos para responder e mitigar o impacto potencial de incidentes de segurança. A arquitetura de sua workload afeta fortemente a capacidade de suas equipes de operar efetivamente durante um incidente, de isolar ou conter sistemas e de restaurar operações para um bom estado conhecido. Disponibilizar as ferramentas e o acesso antes de um incidente de segurança e praticar rotineiramente a resposta a incidentes durante os game days ajudará a garantir que sua arquitetura possa acomodar investigações e recuperação oportunas.

Na AWS, as seguintes práticas, facilitam a resposta eficaz a incidentes:

- Um log detalhado com conteúdo importante, como acesso e alterações em arquivos, está disponível.
- Os eventos podem ser processados automaticamente e acionar ferramentas que automatizam respostas usando as APIs da AWS.
- Você pode pré-provisionar ferramentas e uma “sala limpa” usando o AWS CloudFormation. Isso permite realizar análise forense em um ambiente seguro e isolado.

A pergunta a seguir concentra-se nessas considerações sobre segurança.

SEC 10: Como prever, responder a e se recuperar de incidentes?

A preparação é essencial para investigação, resposta e recuperação oportunas e eficazes de incidentes de segurança para ajudar a minimizar interrupções na sua organização.

Garanta acesso rápido de sua equipe de segurança e automatize o isolamento de instâncias, bem como a captura de dados e estado para análise forense.

Segurança de aplicações

A segurança de aplicações (AppSec) retrata o processo geral de como projetar, criar e testar as propriedades de segurança das workloads desenvolvidas por você. Você precisa treinar a equipe adequadamente em sua organização, entender as propriedades de segurança de sua infraestrutura de compilação e lançamento e utilizar a automação para identificar problemas de segurança.

Adotar testes de segurança de aplicações como parte regular do ciclo de vida de desenvolvimento de software (SDLC) e processos de pós-lançamento ajuda a garantir que você tenha um mecanismo estruturado para identificar, corrigir e impedir que problemas de segurança de aplicações entrem no ambiente de produção.

Sua metodologia de desenvolvimento de aplicações deve incluir controles de segurança à medida que você projeta, cria, implanta e opera suas workloads. Ao fazer isso, alinhe o processo para redução contínua de defeitos e redução da dívida técnica. Por exemplo, o uso de modelagem de ameaças na fase de design ajuda a detectar falhas de design precocemente, o que torna mais fácil e menos caro corrigi-las em contraposição a aguardar e mitigá-las posteriormente.

O custo e a complexidade para resolver defeitos geralmente serão menores quanto mais no princípio do SDLC você estiver. A forma mais fácil de resolver problemas é não os ter. Por isso, começar com um modelo de ameaças ajuda você a se concentrar nos resultados corretos da fase de design. À medida que seu programa de AppSec amadurece, é possível aumentar a quantidade de testes realizados usando automação, aumentar a fidelidade do feedback para os criadores e reduzir o tempo necessário para as avaliações de segurança. Todas essas ações melhoram a qualidade do software desenvolvido e aumentam a velocidade de entrega de recursos à produção.

Essas diretrizes de implementação focam quatro áreas: organização e cultura, segurança do pipeline, segurança no pipeline e gerenciamento de dependências. Cada área oferece um conjunto de princípios que você pode implementar, bem como uma visão completa de como projetar, desenvolver, criar, implantar e operar workloads.

Na AWS, há várias abordagens para lidar com seu programa de segurança de aplicações. Algumas dessas abordagens dependem de tecnologia, enquanto outras focam a equipe e aspectos organizacionais do programa de segurança de aplicações.

A pergunta a seguir concentra-se nessas considerações sobre segurança.

SEC 11: Como incorporar e validar as propriedades de segurança de aplicações durante o ciclo de vida de design, desenvolvimento e implantação?

Treinar a equipe, testar por meio da automação, entender as dependências e validar as proprieda des de segurança de ferramentas e aplicações ajuda a diminuir a probabilidade de problemas de segurança em workloads de produção.

Recursos

Consulte os recursos a seguir para saber mais sobre nossas práticas recomendadas para Segurança.

Documentação

- Segurança da nuvem da AWS
- Conformidade da AWS
- Blog de segurança da AWS
- Modelo de maturidade de segurança da AWS

Whitepaper

- Pilar de segurança
- Visão geral de segurança da AWS
- Risco e conformidade da AWS

Vídeo

- Estado de segurança da união da AWS
- Visão geral da responsabilidade compartilhada

Confiabilidade

O pilar Confiabilidade abrange a capacidade de uma workload de executar a função pretendida correta e consistentemente quando esperado. Isso inclui a capacidade de operar e testar a workload

durante todo o ciclo de vida dela. Este documento fornece orientações detalhadas sobre as práticas recomendadas para a implementação de workloads confiáveis na AWS.

O pilar Confiabilidade apresenta uma visão geral de princípios de design, práticas recomendadas e perguntas. Recomendações sobre implementação estão disponíveis no whitepaper Pilar Confiabilidade.

Tópicos

Princípios de design

Existem cinco princípios de design para confiabilidade na nuvem:

- Recuperar de falhas automaticamente: ao monitorar os indicadores-chave de performance (KPIs) de uma workload, você pode acionar a automação quando um limite é violado. Esses KPIs devem ser uma medida do valor comercial, e não dos aspectos técnicos da operação do serviço. Isso possibilita a notificação automática e o rastreamento de falhas, além de processos de recuperação automatizados que solucionam ou reparam a falha. Com uma automação mais sofisticada, é possível antecipar e corrigir falhas antes que elas ocorram.
- Teste os procedimentos de recuperação: em um ambiente on-premises, muitas vezes os testes são realizados para provar que a workload funciona em um cenário específico. Normalmente, o teste não é usado para validar estratégias de recuperação. Na nuvem, você pode testar o comportamento de falha da workload e validar os procedimentos de recuperação. É possível usar a automação para simular falhas diferentes ou para recriar cenários que levaram a falhas no passado. Essa abordagem expõe caminhos de falha que você pode testar e corrigir antes que um cenário de falha real ocorra, reduzindo assim o risco.
- Escale horizontalmente para aumentar a disponibilidade agregada da workload: substitua um recurso grande por vários recursos pequenos para reduzir o impacto de uma única falha na workload geral. Distribua as solicitações por vários recursos menores para garantir que elas não compartilhem um ponto de falha comum.
- Pare de tentar adivinhar a capacidade: uma causa comum de falha nas workloads on-premises é a saturação de recursos, quando as demandas impostas a uma workload excedem a respectiva

capacidade (esse muitas vezes é o objetivo dos ataques de negação de serviço). Na nuvem, você pode monitorar a demanda e a utilização da workload e automatizar a adição ou a remoção de recursos para manter o nível mais eficiente e atender à demanda, sem provisionamento excessivo ou subprovisionamento. Ainda há limites, mas algumas cotas podem ser controladas e outras podem ser gerenciadas. Consulte Gerenciar cotas de serviço e restrições.

- Gerencie alterações na automação: as alterações em sua infraestrutura devem ser feitas por meio de automação. Entre aquelas que devem ser gerenciadas estão as alterações na automação, que podem ser acompanhadas e analisadas.

Definição

Existem quatro áreas de práticas recomendadas para confiabilidade na nuvem:

- Fundamentos
- Arquitetura da workload
- Gerenciamento de alterações
- Gerenciamento de falhas

Para atingir a confiabilidade, você deve começar com as bases: um ambiente em que as cotas de serviço e a topologia de rede acomodam a workload. A arquitetura da workload do sistema distribuído deve ser projetada para evitar e mitigar falhas. A workload deve processar as alterações na demanda ou nos requisitos e ser projetada para detectar falhas e se reparar automaticamente.

Práticas recomendadas

Tópicos

Fundamentos

Os requisitos fundamentais são aqueles que têm um escopo que vai além de uma única workload ou projeto. Antes de criar a arquitetura de um sistema, é necessário instaurar os requisitos fundamentais

que influenciam a confiabilidade. Por exemplo, você deve ter largura de banda de rede suficiente no datacenter.

Com a AWS, a maioria desses requisitos fundamentais já está incorporada ou pode ser tratada conforme necessário. A nuvem foi projetada para ser praticamente ilimitada, portanto, é responsabilidade da AWS atender ao requisito de capacidade suficiente de rede e de computação, permitindo que você altere o tamanho e as alocações de recursos sob demanda.

As perguntas a seguir referem-se a essas considerações sobre confiabilidade. (Para obter uma lista de perguntas e práticas recomendadas de confiabilidade, consulte o Apêndice.)

REL 1: Como gerenciar as cotas e restrições de serviço?

Para arquiteturas de workload baseadas na nuvem, existem cotas de serviço (que também são chamadas de limites de serviço). Essas cotas existem para evitar o provisionamento acidental de mais recursos do que você precisa e para limitar as taxas de solicitação nas operações de API, a fim de proteger os serviços contra uso abusivo. Também há restrições de recursos, por exemplo, a taxa em que você pode propagar bits por um cabo de fibra óptica ou a quantidade de armazenamento em um disco físico.

REL 2: Como planejar a topologia de rede?

Geralmente, existem workloads em vários ambientes. Isso inclui vários ambientes de nuvem (acessíveis ao público e privados) e, possivelmente, a infraestrutura de datacenter existente . Os planos devem incluir considerações de rede, como conectividade intra e entre sistemas, gerenciamento de endereços IP públicos, gerenciamento de endereços IP privados e resolução de nomes de domínio.

Arquitetura da workload

Uma workload confiável começa com as decisões iniciais de projeto que envolvem tanto o software quanto a infraestrutura. Suas decisões de arquitetura afetarão o comportamento da workload em todos os pilares do Well-Architected. Para atingir a confiabilidade, há padrões específicos que devem ser seguidos.

Com a AWS, os desenvolvedores de workload podem usar suas linguagens e tecnologias preferidas. AWS Os SDKs eliminam a complexidade da codificação por meio de APIs específicas à linguagem

para os serviços da AWS. Esses SDKs e a possibilidade de escolher a linguagem permitem que os desenvolvedores implementem as práticas recomendadas de confiabilidade apresentadas neste documento. Os desenvolvedores também podem ler e descobrir como a Amazon cria e opera softwares na Amazon Builders' Library.

As perguntas a seguir referem-se a essas considerações sobre confiabilidade.

REL 3: Como projetar sua arquitetura de serviços de workload?

Use uma arquitetura orientada a serviços (SOA) ou uma arquitetura de microsserviços para criar workloads altamente escaláveis e confiáveis. A arquitetura orientada a serviços (SOA) é a prática de tornar componentes de software reutilizáveis por meio de interfaces de serviço. A arquitetura de microsserviços vai além para tornar os componentes menores e mais simples.

REL 4: Como projetar interações em um sistema distribuído para evitar falhas?

Os sistemas distribuídos dependem de redes de comunicação para interconectar componentes, como servidores ou serviços. A workload deve operar de forma confiável, apesar da perda de dados ou da latência nessas redes. Os componentes do sistema distribuído devem operar de uma maneira que não afete negativamente outros componentes ou a workload. Essas práticas recomendadas evitam falhas e melhoram o tempo médio entre falhas (MTBF).

REL 5: Como projetar interações em um sistema distribuído para mitigar ou resistir a falhas?

Os sistemas distribuídos dependem de redes de comunicação para interconectar componentes (como servidores ou serviços). Sua workload deve operar de forma confiável, apesar da perda de dados ou da latência nessas redes. Os componentes do sistema distribuído devem operar de uma maneira que não afete negativamente outros componentes ou a workload. Essas práticas recomendadas permitem que as workloads resistam a tensões ou falhas, recuperem-se mais rapidamente delas e reduzam o impacto de tais prejuízos. Como resultado, o tempo médio para recuperação (MTTR) é melhorado.

Gerenciamento de alterações

As alterações na workload ou no respectivo ambiente devem ser previstas e acomodadas para alcançar uma operação confiável da workload. As alterações incluem aquelas impostas à sua workload, como picos na demanda, bem como as internas, como implantações de recursos e patches de segurança.

Ao usar a AWS, você pode monitorar o comportamento de uma workload e automatizar a resposta aos KPIs. Por exemplo, a workload pode adicionar outros servidores à medida que recebe mais usuários. É possível controlar quem tem permissão para fazer alterações na workload e realizar auditorias no histórico dessas alterações.

As perguntas a seguir referem-se a essas considerações sobre confiabilidade.

REL 6: Como monitorar recursos de workload?

Logs e métricas são ferramentas avançadas para obter informações sobre a integridade da workload. Você pode configurar a workload para monitorar logs e métricas e enviar notificações quando os limites forem ultrapassados ou ocorrerem eventos significativos. O monitoramento permite que sua workload reconheça quando os limites de baixa performance são ultrapassados ou quando há falhas para que ela possa se recuperar automaticamente em resposta.

REL 7: Como projetar sua workload para se adaptar às mudanças na demanda?

Uma workload escalável fornece elasticidade para adicionar ou remover recursos automatic amente, de modo que eles correspondam perfeitamente à demanda atual em determinado momento.

REL 8: Como implementar uma alteração?

As alterações controladas são necessárias para implantar novas funcionalidades e garantir que as workloads e o ambiente operacional executem softwares conhecidos e possam ser corrigidos ou substituídos de maneira previsível. Se essas alterações não forem controladas, será difícil prever o efeito dessas alterações ou resolver os problemas que surgem por causa delas.

Quando você cria a arquitetura de uma workload para adicionar e remover recursos automaticamente em resposta às alterações na demanda, isso não apenas aumenta a confiabilidade, mas também garante que o sucesso nos negócios não se torne um fardo. Com o monitoramento implantado, sua equipe será automaticamente alertada quando os KPIs se desviarem das normas esperadas. O log automático de alterações em seu ambiente permite realizar auditorias e identificar rapidamente as ações que podem ter afetado a confiabilidade. Os controles de gerenciamento de alterações garantem que você possa impor as regras que oferecem a confiabilidade necessária.

Gerenciamento de falhas

Em qualquer sistema de complexidade razoável, espera-se que ocorram falhas. A confiabilidade exige que sua workload reconheça as falhas no momento em que elas ocorrem e tome medidas para evitar que elas prejudiquem a disponibilidade. As workloads devem ser capazes de resistir a falhas e reparar problemas automaticamente.

Com a AWS, você pode aproveitar a automação para reagir aos dados de monitoramento. Por exemplo, quando uma métrica específica ultrapassa um limite, você pode iniciar uma ação automatizada para solucionar o problema. Além disso, em vez de tentar diagnosticar e corrigir um recurso com falha que faz parte do seu ambiente de produção, você pode substituí-lo por um novo e executar a análise do recurso com falha fora de banda. Como a nuvem permite que você suporte versões temporárias de um sistema inteiro a baixo custo, é possível usar testes automatizados para verificar os processos de recuperação completos.

As perguntas a seguir referem-se a essas considerações sobre confiabilidade.

REL 9: Como fazer backup dos dados?

Faça backup de dados, aplicações e configurações para atender às suas necessidades de objetivos de tempo de recuperação (RTO) e objetivos de ponto de recuperação (RPO).

REL 10: Como usar o isolamento de falhas para proteger sua workload?

Os limites isolados de falhas limitam o efeito de uma falha em uma workload a um número limitado de componentes. Os componentes fora do limite não são afetados pela falha. Ao usar vários limites isolados de falhas, é possível limitar o impacto na workload.

REL 11: Como projetar a workload para resistir a falhas de componentes?

As workloads que exigem alta disponibilidade e baixo tempo médio até a recuperação (MTTR) devem ser projetadas visando a resiliência.

REL 12: Como testar a confiabilidade?

Depois de projetar a workload para resiliência à pressão da produção, o teste é a única maneira de garantir que ela opere conforme projetado e com a resiliência esperada.

REL 13: Como planejar a recuperação de desastres (DR)?

Implementar backups e componentes redundantes de workload é o ponto de partida da sua estratégia de DR. O RTO e o RPO são os objetivos para restaurar a workload. Defina-os de acordo com suas necessidades de negócios. Implemente uma estratégia para atender a esses objetivos, considerando os locais e a função dos recursos e dos dados da workload. A probabili dade de interrupção e o custo de recuperação também são fatores principais que ajudam a determinar o valor empresarial de fornecer a recuperação de desastres para uma workload.

Regularmente, faça backup dos dados e teste os arquivos de backup para garantir a capacidade de recuperação de erros tanto físicos quanto lógicos. Para gerenciar falhas, é essencial testar as workloads com frequência e de maneira automatizada por meio da indução de falhas e da observação do processo de recuperação. Faça isso periodicamente e também após alterações significativas na workload. Acompanhe ativamente os KPIs, como objetivo de tempo de recuperação (RTO) e objetivo de ponto de recuperação (RPO), para avaliar a resiliência de uma workload, principalmente em cenários de teste de falhas. O rastreamento dos KPIs ajudará você a identificar e mitigar os pontos únicos de falha. O objetivo é testar integralmente os processos de recuperação da workload para ter certeza de que é possível recuperar todos os seus dados e continuar a atender os clientes, mesmo diante de problemas contínuos. Seus processos de recuperação devem ser tão bem trabalhados quanto os processos de produção normais.

Recursos

Consulte os recursos a seguir para saber mais sobre nossas práticas recomendadas para Confiabilidade.

Documentação

- Documentação do AWS
- Infraestrutura global da AWS
- AWS Auto Scaling: como os planos de ajuste de escala funcionam
- O que é o AWS Backup?

Whitepaper

- Pilar Confiabilidade: AWS Well-Architected
- Implementar microsserviços na AWS

Eficiência de performance

O pilar Eficiência de performance inclui a capacidade de usar recursos de nuvem de maneira eficiente para atender aos requisitos de performance e manter essa eficiência à medida que a demanda muda e as tecnologias evoluem.

O pilar Eficiência de performance apresenta uma visão geral dos princípios de design, das práticas recomendadas e das perguntas. Recomendações sobre implementação estão disponíveis no whitepaper Pilar Eficiência de performance.

Tópicos

- Princípios de design
- Definição
- Práticas recomendadas
- Recursos

Princípios de design

Existem cinco princípios de design para eficiência de performance na nuvem:

- Democratize tecnologias avançadas: facilite a implementação de tecnologia avançada para a sua equipe delegando tarefas complexas ao seu fornecedor de nuvem. Em vez de solicitar que sua equipe de TI aprenda a hospedar e executar uma nova tecnologia, avalie a possibilidade de consumir a tecnologia como um serviço. Por exemplo, bancos de dados NoSQL, transcodificação de mídia e machine learning são tecnologias que exigem altos níveis de especialização. Na nuvem, essas tecnologias se tornam serviços que sua equipe pode consumir, permitindo que ela se concentre no desenvolvimento de produtos, em vez de provisionamento e gerenciamento de recursos.
- Tenha alcance global em poucos minutos: a implantação de sua workload em várias regiões da AWS em todo o mundo permite oferecer menor latência e uma melhor experiência para seus clientes a um custo mínimo.
- Use arquiteturas sem servidor: as arquiteturas sem servidor eliminam a necessidade de executar e manter servidores físicos para realizar atividades tradicionais de computação. Os serviços de armazenamento sem servidor, por exemplo, podem atuar como sites estáticos (eliminando a necessidade de servidores Web) e os serviços de eventos podem hospedar código. Isso elimina o fardo operacional do gerenciamento de servidores físicos e pode reduzir os custos transacionais, pois os serviços gerenciados operam em escala de nuvem.
- Experimente com mais frequência: com recursos virtuais e automatizáveis, você pode executar rapidamente testes comparativos usando diferentes tipos de instâncias, armazenamento ou configurações.
- Considere a solidariedade mecânica: entenda como os serviços de nuvem são consumidos e use sempre a abordagem tecnológica mais alinhada às suas metas de workload. Por exemplo, avalie padrões de acesso a dados ao selecionar abordagens de banco de dados ou armazenamento.

Definição

Existem cinco áreas de práticas recomendadas para eficiência de performance na nuvem:

- Seleção de arquitetura
- Computação e hardware
- Gerenciamento de dados
- Rede e entrega de conteúdo
- Processo e cultura

Adote uma abordagem impulsionada por dados para criar uma arquitetura de alta performance. Reúna dados sobre todos os aspectos da arquitetura, desde o design de alto nível até a seleção e a configuração dos tipos de recursos.

Analisar suas escolhas regularmente garante que você possa aproveitar a evolução contínua da Nuvem AWS. O monitoramento garante que você esteja ciente de qualquer desvio em relação à performance esperada. Faça concessões em sua arquitetura visando o aprimoramento da performance, como o uso de compactação ou armazenamento em cache, ou ainda a diminuição dos requisitos de consistência.

Práticas recomendadas

Tópicos

Seleção de arquitetura

A solução ideal para uma workload específica pode variar e, muitas vezes, as soluções combinam várias abordagens. As workloads do Well-Architected usam várias soluções e permitem diferentes recursos para aprimorar a performance.

Os recursos da AWS estão disponíveis em vários tipos e configurações, o que facilita encontrar uma abordagem que atenda melhor às suas necessidades. Também é possível encontrar opções que não são facilmente obtidas com infraestrutura on-premises. Um serviço gerenciado como o Amazon DynamoDB, por exemplo, fornece um banco de dados NoSQL totalmente gerenciado com latência de milissegundos de um dígito em qualquer escala.

As perguntas a seguir referem-se a essas considerações sobre a eficiência de performance. (Para obter uma lista de perguntas e práticas recomendadas sobre eficiência de performance, consulte o Apêndice.)

PERF 1: Como selecionar os recursos de nuvem e os padrões de arquitetura apropriados para sua workload?

Muitas vezes, é necessário empregar várias abordagens para obter a performance ideal em uma workload. Os sistemas com boa arquitetura usam várias soluções e recursos para aprimorar a performance.

Computação e hardware

A opção ideal de computação para uma workload específica pode variar de acordo com o design, os padrões de uso e as definições de configuração da aplicação. As arquiteturas podem usar diferentes opções de computação para vários componentes e permitir diferentes recursos para aprimorar a performance. A seleção da opção de computação incorreta para uma arquitetura pode levar a uma menor eficiência de performance.

Na AWS, a computação é disponibilizada em três formatos: instâncias, contêineres e funções:

- Instâncias são servidores virtualizados que permitem que você altere seus recursos com um botão ou uma chamada de API. Como as decisões de recursos na nuvem não são imutáveis, você pode testar diferentes tipos de servidores. Na AWS, essas instâncias de servidor virtual vêm em diferentes famílias e tamanhos e oferecem uma ampla variedade de capacidades, inclusive unidades de estado sólido (SSDs) e unidades de processamento gráfico (GPUs).
- Contêineres são um método de virtualização do sistema operacional que permite executar uma aplicação e suas dependências em processos isolados por recursos. O AWS Fargate é um serviço de computação sem servidor para contêineres, ou também é possível usar o Amazon EC2 se você precisar de controle sobre a instalação, a configuração e o gerenciamento do seu ambiente de computação. Você também pode escolher entre várias plataformas de orquestração de contêineres: Amazon Elastic Container Service (ECS) ou Amazon Elastic Kubernetes Service (EKS).
- As funções abstraem o ambiente de execução do código que você deseja aplicar. Por exemplo, o AWS Lambda permite executar código sem executar uma instância.

As perguntas a seguir referem-se a essas considerações sobre a eficiência de performance.

PERF 2: Como selecionar e usar recursos computacionais em sua workload?

A solução de computação mais eficiente para uma workload varia dependendo do design da aplicação, dos padrões de uso e das definições de configuração. As arquiteturas podem usar diferentes soluções de computação para vários componentes e podem ativar diferentes recursos para melhorar a performance. Selecionar a solução de computação incorreta para uma arquitetura pode levar a uma menor eficiência de performance.

Gerenciamento de dados

A solução de gerenciamento de dados ideal para um sistema específico varia conforme o tipo de dados (bloco, arquivo ou objeto), os padrões de acesso (aleatório ou sequencial), o throughput necessário, a frequência de acesso (online, offline, arquivamento), a frequência de atualização (WORM, dinâmica) e as restrições de disponibilidade e durabilidade. As workloads do Well-Architected usam datastores específicos que permitem que recursos diferentes melhorem a performance.

Na AWS, o armazenamento é disponibilizado em três formatos: objeto, bloco e arquivo:

- O armazenamento de objetos fornece uma plataforma escalável e durável para tornar os dados acessíveis a partir de qualquer local da Internet para conteúdo gerado pelo usuário, arquivamento ativo, computação sem servidor, armazenamento de big data ou backup e recuperação. O Amazon Simple Storage Service (Amazon S3) é um serviço de armazenamento de objetos que oferece escalabilidade líder do setor, disponibilidade de dados, segurança e performance. O Amazon S3 foi projetado para oferecer 99,999999999% (11 noves) de durabilidade e armazena dados para milhões de aplicações para empresas de todo o mundo.
- O armazenamento em bloco fornece armazenamento em blocos altamente disponível, consistente e de baixa latência para cada host virtual e é análogo ao armazenamento de conexão direta (DAS) ou a uma rede de área de armazenamento (SAN). O Amazon Elastic Block Store (Amazon EBS) foi projetado para workloads que exigem armazenamento persistente acessível por instâncias do EC2, o que ajuda você a ajustar aplicações com o custo, a performance e a capacidade de armazenamento corretos.
- O armazenamento de arquivos fornece acesso a um sistema de arquivos compartilhado entre vários sistemas. As soluções de armazenamento de arquivos, como o Amazon Elastic File System (Amazon EFS), são ideais para casos de uso como grandes repositórios de conteúdo, ambientes de desenvolvimento, armazenamentos de mídia ou diretórios iniciais de usuários. O Amazon

FSx torna fácil e eficiente iniciar e executar sistemas de arquivos populares para que você possa aproveitar os sofisticados conjuntos de recursos e a rápida performance de sistemas de arquivos de código aberto amplamente utilizados e licenciados comercialmente.

As perguntas a seguir referem-se a essas considerações sobre a eficiência de performance.

PERF 3: Como armazenar, gerenciar e acessar dados em sua workload?

A solução de armazenamento mais eficiente para um sistema varia em função de tipo de operação de acesso (bloco, arquivo ou objeto), padrões de acesso (aleatório ou sequencial), throughput necessário, frequência de acesso (online, offline, arquivamento), frequência de atualização (WORM, dinâmica) e disponibilidade e durabilidade. Os sistemas Well-Architected usam várias soluções de armazenamento e habilitam diferentes recursos para melhorar a performance e usar os recursos de modo eficiente.

Rede e entrega de conteúdo

A solução de rede ideal para uma workload varia com base em latência, requisitos de throughput, jitter e largura de banda. Restrições físicas, como recursos de usuário ou on-premises, determinam as opções de localização. Essas restrições podem ser compensadas com locais de borda ou posicionamento de recursos.

Na AWS, as redes são virtualizadas e estão disponíveis em vários tipos e configurações diferentes. Desse modo, é mais fácil atender às suas necessidades de rede. A AWS oferece recursos de produtos (por exemplo, redes avançada, instâncias otimizadas de rede do Amazon EC2, aceleração de transferências do Amazon S3 e Amazon CloudFront dinâmico) para otimizar o tráfego da rede. A AWS também oferece recursos de rede (por exemplo, roteamento de latência do Amazon Route 53, endpoints da Amazon VPC, AWS Direct Connect e AWS Global Accelerator) para reduzir a distância ou o jitter da rede.

As perguntas a seguir referem-se a essas considerações sobre a eficiência de performance.

PERF 4: Como selecionar e configurar os recursos de rede em sua workload?

Essa área de foco compartilha orientações e práticas recomendadas para projetar, configurar e operar soluções eficientes de rede e entrega de conteúdo na nuvem.

Processo e cultura

Ao arquitetar workloads, há princípios e práticas que você pode adotar para ajudar na melhor execução de workloads na nuvem eficientes e de alta performance. Para adotar uma cultura que promova a eficiência de performance das workloads na nuvem, considere estes princípios e práticas fundamentais:

Considere estes princípios fundamentais para construir essa cultura:

- Infraestrutura como código: defina sua infraestrutura como código usando abordagens como modelos do AWS CloudFormation CloudFormation. O uso de modelos permite colocar a infraestrutura no controle de origem junto com o código e as configurações de sua aplicação. Isso permite aplicar à sua infraestrutura as mesmas práticas usadas para desenvolver software, possibilitando uma iteração rápida.
- Pipeline de implantação: use um pipeline de integração e implantação contínuas (CI/CD) (por exemplo, repositório de código-fonte, sistemas de compilação, implantação e automação de teste) para implantar sua infraestrutura. Isso permite a você implantar de maneira repetível, consistente e econômica enquanto itera.
- Métricas bem-definidas: configure e monitore métricas para capturar os indicadores-chave de performance (KPIs). Recomendamos usar tanto de métricas técnicas quanto de negócios. Para aplicações móveis ou sites, métricas importantes são a captura do tempo até o primeiro byte ou renderização. Outras métricas geralmente aplicáveis incluem contagem de threads, taxa de coleta de resíduos e estados de espera. As métricas de negócios, como o custo cumulativo agregado por solicitação, podem alertar sobre maneiras de reduzir os custos. Considere com cuidado como você planeja interpretar as métricas. Por exemplo, você poderia escolher o máximo ou o 99º percentil, em vez da média.
- Teste a performance automaticamente:como parte do processo de implantação, inicie automaticamente os testes de performance após a aprovação bem-sucedida nos testes de execução mais rápida. A automação deve criar um novo ambiente, configurar as condições iniciais, como dados de teste, e então executar uma série de testes comparativos e de carga. Os resultados desses testes então devem ser vinculados de volta à compilação para que você possa rastrear as mudanças de performance ao longo do tempo. Para testes de execução longa, você pode tornar essa parte do pipeline assíncrona em relação ao restante da compilação. Como alternativa, é possível realizar testes de performance durante a noite usando instâncias spot do Amazon EC2.
- Geração de carga: você deve criar uma série de scripts de teste que repliquem jornadas sintéticas ou pré-gravadas do usuário. Esses scripts devem ser idempotentes e não acoplados, e talvez

você precise incluir scripts de pré-aquecimento para gerar resultados válidos. Seus scripts de teste devem replicar o máximo possível o comportamento do uso na produção. É possível usar soluções de software ou software como serviço (SaaS) para gerar a carga. Cogite o uso de soluções do AWS Marketplace e de instâncias spot: elas podem ser maneiras econômicas de gerar a carga.

- Visibilidade da performance: as métricas principais devem estar visíveis para a sua equipe, especialmente as métricas relacionadas a cada versão de compilação. Isso permite que você identifique qualquer tendência positiva ou negativa importante ao longo do tempo. Você também deve exibir métricas do número de erros ou exceções para garantir que esteja testando um sistema em funcionamento.
- Visualização: use técnicas de visualização que deixem claro onde os problemas de performance, hot spots, estados de espera ou baixa utilização estão ocorrendo. Sobreponha métricas de performance a diagramas de arquitetura: código ou gráficos de chamada podem ajudar a identificar problemas rapidamente.
- Revise os processos regularmente: arquiteturas com baixa performance geralmente são o resultado de um processo de análise de performance inexistente ou problemático. Se sua arquitetura está funcionando mal, a implementação de um processo de análise de performance permite promover melhorias iterativas.
- Otimização contínua: adote uma cultura para otimizar continuamente a eficiência de performance da da workload na nuvem.

As perguntas a seguir referem-se a essas considerações sobre a eficiência de performance.

PERF 5: Que processo você usa para oferecer maior eficiência de performance para sua workload?

Ao arquitetar workloads, há princípios e práticas que você pode adotar para ajudar na melhor execução de workloads na nuvem eficientes e de alta performance. Para adotar uma cultura que promova a eficiência de performance das workloads na nuvem, considere estes princípios e práticas fundamentais:

Recursos

Consulte os seguintes recursos para saber mais sobre nossas práticas recomendadas para eficiência de performance.

Documentação

- Otimização de performance do Amazon S3
- Amazon EBS Volume Performance (Performance de volumes do Amazon EBS)

Whitepaper

- Pilar Eficiência de performance

Vídeo

- AWS re:Invent 2019: Fundamentos do Amazon EC2 (CMP211-R2)
- AWS re:Invent 2019: Leadership session: considerações sobre o estado do armazenamento (STG201-L)
- AWS re:Invent 2019: Leadership session: bancos de dados com propósito específico da AWS (DAT209-L)
- AWS re:Invent 2019: Conectividade à AWS e arquiteturas de rede híbrida da AWS (NET317-R1)
- AWS re:Invent 2019: Capacitar o Amazon EC2 da próxima geração: mergulho profundo no Nitro System (CMP303-R2)
- AWS re:Invent 2019: Escalar para seus primeiros 10 milhões de usuários (ARC211-R)

Otimização de custo

O pilar Otimização de custos inclui a capacidade de executar sistemas para proporcionar valor comercial pelo menor preço.

O pilar Otimização de custos fornece uma visão geral dos princípios de design, práticas recomendadas e perguntas. Recomendações sobre implementação estão disponíveis no whitepaper Pilar Otimização de custos.

Tópicos

- Princípios de design
- Definição
- Práticas recomendadas

- Recursos

Princípios de design

Existem cinco princípios de design para a otimização de custos na nuvem:

- Implemente o gerenciamento financeiro na nuvem: para obter sucesso financeiro e acelerar a obtenção de valor empresarial na nuvem, invista em gerenciamento financeiro na nuvem e otimização de custos. Sua organização precisa dedicar tempo e recursos para criar aptidão nesse novo domínio de tecnologia e gerenciamento de uso. Semelhante à sua aptidão de Segurança ou Excelência operacional, é necessário desenvolver aptidão por meio da criação de conhecimento, programas, recursos e processos para se tornar uma organização econômica.
- Adote um modelo de consumo: pague somente pelos recursos computacionais necessários e aumente ou reduza o uso com base nos requisitos comerciais, não em previsões elaboradas. Por exemplo, ambientes de desenvolvimento e teste são geralmente usados apenas por oito horas ao dia durante a semana de trabalho. Você poderá parar esses recursos quando eles não estiverem em uso para obter uma economia potencial de 75% (40 horas versus 168 horas).
- Meça a eficiência geral: meça o resultado de negócios da workload e os custos associados à sua entrega. Use essa medida para identificar os ganhos obtidos com o aumento da capacidade e a redução de custos.
- Pare de gastar dinheiro em tarefas pesadas genéricas: a AWS faz o trabalho pesado das operações de datacenter, como o armazenamento em rack, o empilhamento e a alimentação de servidores. Ela também elimina a sobrecarga operacional do gerenciamento de sistemas operacionais e aplicações com serviços gerenciados. Isso permite que você mantenha o foco em seus clientes e projetos de negócios, e não na infraestrutura de TI.
- Analise e atribua despesas: a nuvem simplifica a identificação precisa do uso e do custo dos sistemas, o que permite a atribuição transparente de custos de TI aos proprietários de cada workload. Dessa forma, a medição do retorno sobre o investimento (ROI) é facilitada e os proprietários de workloads têm a oportunidade de otimizar recursos e reduzir custos.

Definição

Há cinco áreas de práticas recomendadas para otimização de custos na nuvem:

- Gerenciamento financeiro na nuvem
- Reconhecimento de despesas e usos

- Recursos economicamente eficientes
- Gerenciar recursos de demanda e fornecimento
- Otimização ao longo do tempo

Assim como ocorre com os outros pilares do Well-Architected Framework, é preciso escolher, por exemplo, entre otimizar para aumentar a velocidade de entrada no mercado ou para reduzir custos. Em alguns casos, é mais eficiente otimizar a velocidade para entrar no mercado rapidamente, enviar novos recursos ou cumprir um prazo, em vez de investir na otimização de custos inicial. Às vezes, as decisões de projeto são tomadas com base na pressa e não em dados, já que sempre existe a tentação de compensar "para garantir", em vez de dedicar tempo a realizar testes comparativos da implantação mais econômica. Isso pode levar a implantações com provisionamento excessivo e subotimizadas. No entanto, essa é uma escolha razoável quando você precisa mover sem alterações (lift-and-shift) recursos de seu ambiente on-premises para a nuvem rapidamente e então otimizar mais tarde. Investir na quantidade certa de esforço em uma estratégia de otimização de custos com antecedência permite aproveitar os benefícios econômicos da nuvem de modo mais rápido, conquistando assim uma adesão consistente às práticas recomendadas e evitando provisionamento excessivo desnecessário. As seções a seguir fornecem técnicas e práticas recomendadas para a implementação inicial e contínua do gerenciamento financeiro na nuvem e otimização de custos de suas workloads.

Práticas recomendadas

Tópicos

Gerenciamento financeiro na nuvem

Com a adoção da nuvem, as equipes de tecnologia inovam mais rapidamente devido à redução dos ciclos de implantação de aprovação, aquisição e infraestrutura. Uma nova abordagem ao gerenciamento financeiro na nuvem é necessária para obter valor empresarial e sucesso financeiro. Essa abordagem é o gerenciamento financeiro na nuvem, e ela cria recursos em toda a organização

por meio da implementação de criação, programas, recursos e processos de conhecimento em toda a organização.

Muitas organizações são compostas por várias unidades diferentes com prioridades diferentes. A capacidade de alinhar sua organização a um conjunto combinado de objetivos financeiros e fornecer a ela os mecanismos para alcançá-los criará uma organização mais eficiente. Uma organização capaz inovará e criará mais rapidamente, será mais ágil e se ajustará a todos os fatores internos ou externos.

Na AWS, você pode usar o Explorador de Custos e, opcionalmente, o Amazon Athena e o Amazon QuickSight com o Relatório de Custos e Uso (CUR) para fornecer conscientização de custos e uso em toda a organização. AWS O Budgets fornece notificações proativas para custo e uso. Os blogs da AWS oferecem informações sobre novos serviços e recursos para garantir que você esteja atualizado com os novos lançamentos de serviços.

As perguntas a seguir referem-se a essas considerações sobre otimização de custos. (Para obter uma lista de perguntas e práticas recomendadas de otimização de custos, consulte o Apêndice.).

COST 1: Como implementar o gerenciamento financeiro na nuvem?

A implementação do gerenciamento financeiro na nuvem ajuda as organizações a obterem valor empresarial e sucesso financeiro à medida que otimizam os custos e o uso e escalam na AWS.

Ao criar uma função de otimização de custos, use membros e complemente a equipe com especialistas em CFM e otimização de custos. Os membros existentes da equipe compreenderão como a organização funciona atualmente e como implementar melhorias com rapidez. Considere também incluir pessoas com conjuntos de habilidades complementares ou especializadas, como estudo analítico e gerenciamento de projetos.

Ao implementar a conscientização de custos na sua organização, melhore ou desenvolva programas e processos existentes. É muito mais rápido adicionar ao que já existe do que criar novos processos e programas novos. Isso resultará em resultados de maneira muito mais rápida.

Reconhecimento de despesas e usos

A flexibilidade e a agilidade proporcionadas pela nuvem incentivam a inovação, desenvolvimento e implantação em ritmo acelerado. Ela reduz os processos manuais e o tempo associado ao provisionamento da infraestrutura on-premises, incluindo a identificação de especificações de

hardware, negociação de cotações de preços, gerenciamento de pedidos de compra, programação de remessas e implantação dos recursos. No entanto, a facilidade de uso e a capacidade sob demanda praticamente ilimitada exigem uma nova forma de pensar sobre as despesas.

Muitas empresas são compostas por vários sistemas operados por várias equipes. A capacidade de atribuir custos de recursos à organização individual ou aos proprietários do produto gera um comportamento eficiente do uso e ajuda a reduzir o desperdício. A atribuição precisa de custos permite saber quais produtos são realmente rentáveis e permite tomar decisões mais informadas sobre alocação de orçamento.

Na AWS, você cria uma estrutura de contas com o AWS Organizations ou o AWS Control Tower, o que fornece separação e ajuda na alocação dos seus custos e do uso. Você também pode usar a marcação de recursos para aplicar informações empresariais e da organização ao seu uso e custo. Use o AWS Cost Explorer para obter visibilidade do custo e do uso ou crie estudos analíticos e painéis personalizados com o Amazon Athena e o Amazon QuickSight. O controle dos custos e do uso é feito com notificações por meio do AWS Budgets e de controles usando o AWS Identity and Access Management (IAM) e o Service Quotas.

As perguntas a seguir referem-se a essas considerações sobre otimização de custos.

COST 2: Como controlar o uso?

Estabeleça políticas e mecanismos para validar que os custos adequados são gerados enquanto os objetivos são alcançados. Ao empregar uma abordagem de verificação e equilíbrio, é possível inovar sem gastar demais.

COST 3: Como monitorar o uso e os custos?

Estabeleça políticas e procedimentos para monitorar e alocar adequadamente os custos. Isso permite medir e aprimorar a eficiência de custos dessa workload.

COST 4: Como desativar recursos?

Implemente o controle de alterações e o gerenciamento de recursos, desde o início do projeto até o fim da vida útil. Isso facilita o desligamento dos recursos não utilizados para reduzir o desperdíc io.

É possível usar tags de alocação de custos para categorizar e rastrear seus custos e uso da AWS. Quando você aplica tags aos recursos da AWS (como instâncias do EC2 ou buckets do S3), a AWS gera um relatório de custos e uso com seu uso e suas tags. Você pode aplicar tags que representam categorias da organização (como centros de custos, nomes de workload ou proprietários) para organizar os custos em vários serviços.

Use o nível correto de detalhes e granularidade no monitoramento e nos relatórios de custo e uso. Para obter insights e tendências de alto nível, use a granularidade diária com o AWS Cost Explorer. Para análise e inspeção mais aprofundadas, use a granularidade por hora no AWS Cost Explorer ou no Amazon Athena e no Amazon QuickSight com o Relatório de Custos e Uso (CUR) em uma granularidade por hora.

A combinação de recursos marcados com o rastreamento do ciclo de vida da entidade (funcionários, projetos) permite identificar recursos ou projetos órfãos que não estão mais gerando valor para a organização e devem ser desativados. Você pode configurar alertas de pagamento para ser notificado sobre gastos excessivos previstos.

Recursos economicamente eficientes

Usar as instâncias e os recursos adequados para sua workload é fundamental para reduzir os custos. Por exemplo, um processo de criação de relatórios pode demorar cinco horas para ser executado em um servidor pequeno, mas uma hora em um servidor grande que custa o dobro. Ambos os servidores fornecem o mesmo resultado, mas o servidor menor acarreta mais custos ao longo do tempo.

Uma workload bem projetada usa os recursos com o melhor custo-benefício, o que pode causar um impacto econômico positivo e considerável. Você também tem a oportunidade de usar serviços gerenciados para reduzir gastos. Por exemplo, em vez de manter servidores para entrega de e-mails, você pode usar um serviço que é pago individualmente por mensagem.

A AWS oferece diversas opções de definição de preço flexíveis e econômicas para você adquirir as instâncias do Amazon EC2 e de outros serviços que atendam às suas necessidades de forma mais eficiente. As Instâncias sob demanda permitem que você pague pela capacidade de computação por hora, sem nenhum compromisso mínimo necessário. Os Savings Plans e instâncias reservadas oferecem economias de até 75% em relação aos preços sob demanda. Com as instâncias spot, você pode aproveitar a capacidade não utilizada do Amazon EC2 e ter economias de até 90% em relação aos preços sob demanda. As instâncias spot são apropriadas para sistemas que aceitam o uso de uma frota de servidores em que os servidores individuais se movimentam dinamicamente, como servidores da Web sem estado, processamento de lotes ou ao usar HPC e big data.

A seleção do serviço adequado também pode reduzir o uso e os custos, como o CloudFront para minimizar a transferência de dados, ou reduzir os custos, e como ao usar o Amazon Aurora no Amazon RDS para remover gastos com licenças caras de banco de dados.

As perguntas a seguir referem-se a essas considerações sobre otimização de custos.

COST 5: Como avaliar o custo ao selecionar serviços?

O Amazon EC2, Amazon EBS e Amazon S3 são produtos fundamentais da AWS. Os produtos gerenciados, como Amazon RDS e Amazon DynamoDB, são serviços da AWS de nível superior ou de aplicação. Ao selecionar os produtos fundamentais e os serviços gerenciados adequados , é possível otimizar os custos dessa workload. Por exemplo, usando serviços gerenciados, é possível reduzir ou remover grande parte da sobrecarga administrativa e operacional, liberando você para trabalhar em aplicações e atividades relacionadas a negócios.

COST 6: Como atingir as metas de custo ao selecionar tamanho, número e tipo de recurso?

Escolha o tamanho e o número de recursos apropriados para a tarefa em mãos. Ao selecionar o tipo, tamanho e número mais econômicos, você minimiza o desperdício.

COST 7: Como usar modelos de preços para reduzir custos?

Use o modelo de preços mais adequado para seus recursos a fim de minimizar as despesas.

COST 8: Como planejar as cobranças de transferência de dados?

Planeje e monitore as cobranças de transferência de dados para tomar decisões de arquitetu ra que minimizam custos. Uma mudança arquitetônica pequena, porém eficaz, pode reduzir drasticamente os custos operacionais ao longo do tempo.

Ao considerar os gastos durante a escolha do serviço e usar ferramentas como o Explorador de Custos e o AWS Trusted Advisor para conferir regularmente seu uso da AWS, você pode monitorar ativamente a utilização e ajustar suas implantações de acordo com ela.

Gerenciar recursos de demanda e fornecimento

Ao migrar para a nuvem, você paga apenas pelo que precisa. Você pode fornecer recursos para atender à demanda da workload no momento em que eles são necessários, o que reduz a necessidade de um provisionamento em excesso que é caro e desperdiça recursos. Você também pode modificar a demanda usando um controle de utilização, buffer ou fila para suavizar a demanda e atendê-la com menos recursos, o que resulta em um custo menor, ou processá-la posteriormente com um serviço em lote.

Na AWS, você pode provisionar automaticamente os recursos para corresponderem à demanda da workload. O Auto Scaling que usa abordagens baseadas em demanda e tempo permitem que você adicione e remova recursos conforme necessário. Se você conseguir prever alterações na demanda, poderá economizar mais dinheiro e garantir que os recursos são compatíveis com as necessidades da sua workload. É possível usar o Amazon API Gateway para implementar o controle de utilização ou o Amazon SQS para implementar uma fila em sua workload. Ambos permitirão que você modifique a demanda nos componentes da workload.

As perguntas a seguir referem-se a essas considerações sobre otimização de custos.

COST 9: Como gerenciar a demanda e fornecer recursos?

Para uma workload com gasto e performance equilibrados, verifique se tudo o que você paga está sendo usado e evite instâncias significativamente subutilizadas. Uma métrica de utilizaçã o distorcida em ambas as direções tem um impacto adverso sobre a organização, tanto nos custos operacionais (redução na performance em decorrência de utilização excessiva) quanto em despesas desnecessárias na AWS (devido ao excesso de provisionamento).

Ao projetar para modificar a demanda e fornecer recursos, pense ativamente nos padrões de uso, no tempo necessário para provisionar novos recursos e na previsibilidade do padrão de demanda. Ao gerenciar a demanda, verifique se você tem uma fila ou um buffer corretamente dimensionado e se está respondendo à demanda da workload no período necessário.

Otimização ao longo do tempo

À medida que a AWS lança novos serviços e recursos, uma das práticas recomendadas é avaliar suas decisões sobre a arquitetura existente a fim de garantir que elas ofereçam o melhor custo-benefício. Conforme seus requisitos mudam, seja incisivo na desativação de recursos, de serviços inteiros e sistemas que não são mais necessários.

A implementação de novos recursos ou tipos de recursos pode otimizar sua workload de modo incremental, minimizando o esforço necessário para implementar a alteração. Isso proporciona melhorias contínuas na eficiência ao longo do tempo e garante que você permaneça na tecnologia mais atualizada para reduzir custos operacionais. Você também pode substituir ou adicionar novos componentes à workload por novos serviços. Isso pode fornecer aumentos significativos na eficiência. Portanto, é essencial revisar regularmente sua workload e implementar novos serviços e recursos.

As perguntas a seguir referem-se a essas considerações sobre otimização de custos.

COST 10: Como avaliar os novos serviços?

À medida que a AWS lança novos serviços e recursos, uma das práticas recomendadas é avaliar suas decisões sobre a arquitetura existente a fim de garantir que elas ofereçam o melhor custo-benefício.

Ao conferir regularmente suas implantações, analise como serviços mais novos podem ajudar você a economizar dinheiro. Por exemplo, o Amazon Aurora no Amazon RDS pode reduzir gastos com bancos de dados relacionados. Usar serviços sem servidor, como o Lambda, pode eliminar a necessidade de operar e gerenciar instâncias para executar código.

COST 11 Como avaliar o custo do esforço?

Avalie o custo do esforço para operações na nuvem, revise suas operações de nuvem demoradas e automatize-as para reduzir esforços humanos e custos adotando serviços da AWS relaciona dos, produtos de terceiros ou ferramentas personalizadas.

Recursos

Consulte os recursos a seguir para saber mais sobre nossas práticas recomendadas para Otimização de custos.

Documentação

- Documentação do AWS

Whitepaper

- Pilar da otimização de custos

Sustentabilidade

O pilar Sustentabilidade foca os impactos ambientais, especialmente a eficiência e o consumo de energia, que são fatores importantes para fundamentar ações diretas dos arquitetos visando reduzir o uso de recursos. Recomendações sobre implementação estão disponíveis no whitepaper Pilar Sustentabilidade.

Tópicos

- Princípios de design
- Definição
- Práticas recomendadas
- Recursos

Princípios de design

Há seis princípios de design para sustentabilidade na nuvem:

- Entenda seu impacto: meça o impacto da sua workload na nuvem e modele o impacto futuro que ela poderá causar. Inclua todas as fontes de impacto, inclusive aquelas resultantes do uso dos seus produtos pelo cliente e da desativação e descontinuação dos mesmos. Compare o resultado produtivo com o impacto total de suas workloads em nuvem analisando os recursos e as emissões exigidas por unidade de trabalho. Use esses dados para estabelecer indicadores-chave de performance (KPIs), avaliar maneiras de melhorar a produtividade enquanto reduz o impacto e estimar o impacto das mudanças propostas ao longo do tempo.
- Defina metas de sustentabilidade: para cada workload na nuvem, estabeleça metas de sustentabilidade de longo prazo, por exemplo, reduzir os recursos de computação e armazenamento exigidos por transação. Modele o retorno sobre o investimento para as melhorias de sustentabilidade das workloads e ofereça aos proprietários os recursos para os quais eles devem investir em metas de sustentabilidade. Planeje-se para o crescimento e projete suas workloads de forma que seu desenvolvimento resulte em uma intensidade de impacto menor com relação a uma unidade apropriada, como por usuário ou por transação. As metas ajudam você a

respaldar os objetivos de sustentabilidade mais amplos de sua empresa ou organização, identificar regressões e priorizar áreas para possível melhoria.

- Maximize a utilização: dimensione as workloads corretamente e implemente um design eficiente que garanta uma alta utilização e maximize a eficiência de energia do hardware subjacente. Dois hosts com 30% de utilização são menos eficientes do que um host com 60% devido ao consumo de energia de referência por host. Ao mesmo tempo, desligue ou minimize recursos, processamento e armazenamento ociosos para reduzir a energia total necessária para suprir a workload.
- Antecipe e adote ofertas de hardware e software novas e mais eficientes: apoie as melhorias preventivas que seus parceiros e fornecedores disponibilizam para ajudar você a reduzir o impacto das workloads na nuvem. Monitore e avalie continuamente ofertas de software e hardware novos e mais eficientes. Projete visando a flexibilidade para permitir a adoção rápida de novas tecnologias eficientes.
- Use serviços gerenciados: compartilhe serviços com uma ampla base de clientes ajuda a maximizar a utilização de recursos, o que reduz a quantidade de infraestrutura necessária para comportar as workloads na nuvem. Por exemplo, os clientes podem compartilhar o impacto de componentes comuns de um datacenter, como a energia e as redes, migrando workloads para a Nuvem AWS e adotando serviços gerenciados, como o AWS Fargate para contêineres sem servidor, onde a AWS trabalha em escala e é responsável por sua operação eficiente. Use serviços gerenciados que possam ajudar a minimizar seu impacto, como a migração automática de dados acessados com pouca frequência para o armazenamento com pouco acesso com as configurações do ciclo de vida do Amazon S3 ou o Amazon EC2 Auto Scaling para ajustar a capacidade de acordo com a demanda.
- Reduza o impacto downstream das suas workloads na nuvem: reduza a quantidade de energia ou recursos necessários para usar seus serviços. Reduza a necessidade de os clientes fazerem upgrade dos dispositivos para usar seus serviços. Teste o uso de parques de dispositivos para saber qual é o impacto esperado e teste com os clientes para entender o impacto atual do uso de seus serviços.

Definição

Há seis áreas de práticas recomendadas de segurança na nuvem:

- Seleção da região
- Alinhamento com a demanda

- Software e arquitetura
- Dados
- Hardware e serviços
- Processo e cultura

A sustentabilidade na nuvem é um esforço quase contínuo focado principalmente na redução e eficiência de energia em todos os componentes de uma workload, obtendo o máximo benefício dos recursos provisionados e minimizando o total de recursos necessários. Esse esforço pode abranger desde a seleção inicial de uma linguagem de programação eficiente, a adoção de algoritmos modernos, o uso de técnicas eficientes de armazenamento de dados, a implantação de uma infraestrutura computacional de tamanho correto e eficiente e a minimização dos requisitos de hardware de alta potência para o usuário final.

Práticas recomendadas

Tópicos

Seleção da região

A escolha da região para sua workload afeta significativamente seus KPIs, incluindo performance, custo e pegada de carbono. Para melhorar esses KPIs, escolha regiões para suas workloads com base em requisitos empresariais e metas de sustentabilidade.

A pergunta a seguir concentra-se nessas considerações sobre sustentabilidade. (Para uma lista de perguntas e práticas recomendadas sobre sustentabilidade consulte o Apêndice.)

SUS 1: Como selecionar regiões para sua workload?

A escolha da região para sua workload afeta significativamente seus KPIs, incluindo performance, custo e pegada de carbono. Para melhorar esses KPIs, escolha regiões para suas workloads com base em requisitos empresariais e metas de sustentabilidade.

Alinhamento com a demanda

A maneira como os usuários e as aplicações consomem suas workloads e outros recursos pode ajudar você a identificar melhorias para atingir metas de sustentabilidade. Escale a infraestrutura de forma que ela corresponda à demanda e use apenas os recursos mínimos necessários para oferecer suporte aos usuários. Alinhe os níveis de serviço às necessidades do cliente. Posicione os recursos a fim de limitar a rede necessária para que usuários e aplicações os consumam. Elimine ativos não utilizados. Forneça aos membros da sua equipe dispositivos compatíveis com suas necessidades e minimize o impacto na sustentabilidade.

A pergunta a seguir foca essas considerações sobre sustentabilidade.

SUS 2: Como alinhar recursos de nuvem à sua demanda?

A maneira como os usuários e as aplicações consomem suas workloads e outros recursos pode ajudar você a identificar melhorias para atingir metas de sustentabilidade. Escale a infraestrutura de forma que ela corresponda à demanda e use apenas os recursos mínimos necessários para oferecer suporte aos usuários. Alinhe os níveis de serviço às necessidades do cliente. Posicione os recursos a fim de limitar a rede necessária para que usuários e aplicações os consumam. Elimine ativos não utilizados. Forneça aos membros da sua equipe dispositivos compatíveis com suas necessidades e minimize o impacto na sustentabilidade.

Escale a infraestrutura com a carga do usuário: identifique períodos de utilização baixa ou sem utilização e escale os recursos para eliminar a capacidade em excesso e melhorar a eficiência.

Alinhe os SLAs às metas de sustentabilidade: defina e atualize as metas do Acordo de Nível de Serviço (SLA), como períodos de disponibilidade ou de retenção de dados de modo a minimizar o número de recursos exigidos para comportar sua workload e, ao mesmo tempo, continuar atendendo aos requisitos empresariais.

Reduza a criação e a manutenção de ativos não utilizados: analise os ativos de aplicações (como relatórios pré-compilados, conjuntos de dados e imagens estáticas) e os padrões de acesso aos ativos para identificar redundâncias, subutilização e possíveis alvos de desativação. Consolide ativos gerados com conteúdo redundante (por exemplo, relatórios mensais com saídas e conjuntos de dados que se sobreponham ou sejam comuns) para reduzir os recursos consumidos quando há duplicação de saídas. Desative ativos não utilizados (por exemplo, imagens de produtos que não são mais vendidos) para liberar os recursos consumidos e reduzir o número de recursos usados para comportar a workload.

Otimize o posicionamento geográfico das workloads: analise os padrões de acesso à rede para identificar de onde seus clientes estão se conectando geograficamente. Selecione regiões e serviços que reduzam a distância que o tráfego de rede deve percorrer para reduzir o total de recursos de rede necessários para comportar a workload.

Otimize os recursos dos membros da equipe para as atividades realizadas: otimize os recursos fornecidos aos membros da equipe para minimizar o impacto sobre a sustentabilidade e, ao mesmo tempo, atender às suas necessidades. Por exemplo, execute operações complexas, como renderização e compilação, em desktops de nuvem compartilhados e com muita utilização, em vez de sistemas de usuário único, de alta potência e subutilizados.

Software e arquitetura

Implemente padrões que suavizem os picos de carga e mantenham a alta utilização consistente de recursos implantados para minimizar os recursos consumidos. Os componentes podem ficar ociosos devido à falta de uso por conta das mudanças no comportamento do usuário ao longo do tempo. Revise os padrões e a arquitetura para consolidar os componentes subutilizados a fim de aumentar a utilização geral. Retire os componentes que não são mais necessários. Saiba qual é a performance dos componentes de sua workload e otimize os componentes que consomem a maioria dos recursos. Esteja ciente dos dispositivos que seus clientes usam para acessar seus serviços e implemente padrões a fim de minimizar a necessidade de upgrades de dispositivos.

A pergunta a seguir concentra-se nessas considerações sobre sustentabilidade.

SUS 3: Como aproveitar os padrões de software e arquitetura para apoiar as metas de sustentabilidade?

Implemente padrões que suavizem os picos de carga e mantenham a alta utilização consistente de recursos implantados para minimizar os recursos consumidos. Os componentes podem ficar ociosos devido à falta de uso por conta das mudanças no comportamento do usuário ao longo do

SUS 3: Como aproveitar os padrões de software e arquitetura para apoiar as metas de sustentabilidade?

tempo. Revise os padrões e a arquitetura para consolidar os componentes subutilizados a fim de aumentar a utilização geral. Retire os componentes que não são mais necessários. Saiba qual é a performance dos componentes de sua workload e otimize os componentes que consomem a maioria dos recursos. Esteja ciente dos dispositivos que seus clientes usam para acessar seus serviços e implemente padrões a fim de minimizar a necessidade de upgrades de dispositivos.

Otimize o software e a arquitetura para trabalhos assíncronos e agendados: use designs e arquiteturas eficientes de software para minimizar a média de recursos necessários por unidade de trabalho. Implemente mecanismos que resultem em uma utilização uniforme de componentes para reduzir os recursos ociosos entre as tarefas e minimizar o impacto de picos de carga.

Remova ou refatore componentes da workload com pouco ou nenhum uso: monitore a atividade da workload para identificar mudanças na utilização de componentes individuais ao longo do tempo. Remova os componentes que não são mais utilizados nem necessários e refatore os componentes pouco usados para reduzir o desperdício de recursos.

Otimize as áreas do código que consomem mais tempo ou recursos: monitore a atividade da workload para identificar os componentes da aplicação que consomem mais recursos. Otimize o código que é executado nesses componentes para minimizar o uso de recursos e, ao mesmo tempo, maximizar a performance.

Otimize o impacto nos dispositivos e equipamentos do cliente: conheça os dispositivos e o equipamento que os clientes usam para consumir seus serviços, o ciclo de vida esperado para eles e o impacto financeiro e na sustentabilidade pela substituição desses componentes. Implemente padrões e arquiteturas de software de modo a minimizar a necessidade de substituir dispositivos e fazer upgrade de equipamento. Por exemplo, implemente novos recursos usando código compatível com versões anteriores de sistemas operacionais e hardware mais antigos, ou gerencie o tamanho das cargas úteis para que elas não excedam a capacidade de armazenamento do dispositivo de destino.

Use padrões e arquiteturas de software que ofereçam suporte mais eficaz aos padrões de acesso e armazenamento de dados: entenda como os dados são usados em sua workload, consumidos por seus usuários, transferidos e armazenados. Escolha tecnologias com o mínimo de requisitos de armazenamento e processamento de dados.

Gerenciamento de dados

A pergunta a seguir concentra-se nessas considerações sobre sustentabilidade.

SUS 4 Como aproveitar as políticas e os padrões de gerenciamento de dados para apoiar as metas de sustentabilidade?

Implemente práticas de gerenciamento de dados para reduzir o armazenamento provisionado necessário para comportar a workload e os recursos exigidos para usá-la. Entenda seus dados e use tecnologias e configurações de armazenamento que promovam o valor empresarial dos dados de forma mais eficaz e a forma como eles são usados. Gerencie o ciclo de vida dos dados e opte por um armazenamento mais eficiente e com menor performance quando os requisitos diminuírem, excluindo os dados que não são mais necessários.

Implemente uma política de classificação de dados: classifique os dados para entender sua importância para os resultados comerciais. Use essas informações para determinar quando é possível migrar os dados para um armazenamento com uso mais eficiente de energia ou excluí-los de forma segura.

Use tecnologias que ofereçam suporte a padrões de acesso a dados e armazenamento: use armazenamento mais adequado à maneira como seus dados são acessados e armazenados a fim de reduzir os recursos provisionados e, ao mesmo tempo, comportar sua workload. Por exemplo, dispositivos de estado sólido (SSDs) usam mais energia do que unidades magnéticas e devem ser usados somente para casos de uso de dados ativos. Use armazenamento de classe de arquivamento com eficiência de energia para dados acessados com pouca frequência.

Use políticas de ciclo de vida para excluir dados desnecessários: gerencie o ciclo de vida de todos os seus dados e defina cronogramas de exclusão automática para minimizar os requisitos totais de armazenamento da workload.

Minimize o provisionamento excessivo no armazenamento em bloco: para reduzir o armazenamento total provisionado, crie um armazenamento em bloco com alocações por tamanho que sejam apropriadas para a workload. Use volumes elásticos para expandir o armazenamento à medida que os dados aumentam sem precisar redimensionar o armazenamento anexado aos recursos de computação. Analise regularmente volumes elásticos e reduza volumes com excesso de provisionamento para se ajustar ao tamanho de dados atual.

Remova dados desnecessários ou redundantes: duplique os dados somente quando necessário para reduzir o armazenamento total consumido. Use tecnologias de backup que eliminem dados duplicados em níveis de arquivo e bloco. Limite o uso de configurações de RAID, exceto quando necessário para atender aos SLAs.

Use sistemas de arquivos compartilhados ou armazenamento de objetos para acessar dados comuns: adote o armazenamento compartilhado e fontes únicas de verdade para evitar duplicação de dados e reduzir os requisitos totais de armazenamento da workload. Busque dados do armazenamento compartilhado somente conforme necessário. Desvincule volumes não usados para liberar recursos. Minimize a movimentação de dados entre redes: use o armazenamento compartilhado e acesse dados de datastores regionais para minimizar os recursos totais de rede exigidos para comportar a movimentação de dados da workload.

Faça backup dos dados somente quando for difícil recriá-los: para reduzir o consumo de armazenamento, faça backup somente de dados com valor empresarial ou que sejam necessários para atender aos requisitos de conformidade. Examine as políticas de backup e exclua armazenamentos temporários que não fornecem valor em um cenário de recuperação.

Hardware e serviços

Procure oportunidades para reduzir os impactos na sustentabilidade da workload fazendo mudanças nas suas práticas de gerenciamento de hardware. Minimize a quantidade de hardware necessária para provisionar e implantar e escolha o hardware e os serviços mais eficientes para sua workload específica.

A pergunta a seguir concentra-se nessas considerações sobre sustentabilidade.

SUS 5: Como selecionar e usar hardware e serviços em nuvem na arquitetura para apoiar os objetivos de sustentabilidade?

Procure oportunidades para reduzir os impactos na sustentabilidade da workload fazendo mudanças nas suas práticas de gerenciamento de hardware. Minimize a quantidade de hardware necessária para provisionar e implantar e escolha o hardware e os serviços mais eficientes para sua workload específica.

Use a quantidade mínima de hardware para atender às suas necessidades: ao usar os recursos da nuvem, é possível fazer alterações frequentes às implementações da workload. Atualize os componentes implantados conforme suas necessidades mudarem.

Use tipos de instâncias que causem o mínimo de impacto: monitore continuamente o lançamento de novos tipos de instância e aproveite as melhorias de eficiência de energia, incluindo os tipos de instância projetados para comportar workloads específicas, como treinamento e inferência de machine learning e transcodificação de vídeo.

Use serviços gerenciados: os serviços gerenciados transferem para a AWS a responsabilidade pela manutenção de uma média elevada de utilização e pela otimização da sustentabilidade do hardware implantado. Use serviços gerenciados para distribuir o impacto na sustentabilidade do serviço entre todos os locatários dele, reduzindo sua contribuição individual.

Otimize o uso de GPUs: unidades de processamento gráfico (GPUs) podem ser uma fonte de alto consumo de energia e várias workloads de GPU são altamente variáveis, como renderização, transcodificação e treinamento e modelagem de machine learning. Execute instâncias de GPUs somente pelo tempo necessário e desative-as com automação quando não precisar mais delas para reduzir o consumo de recursos.

Processo e cultura

Procure oportunidades para reduzir seu impacto na sustentabilidade fazendo mudanças nas suas práticas de desenvolvimento, teste e implantação.

A pergunta a seguir concentra-se nessas considerações sobre sustentabilidade.

SUS 6: Como os processos organizacionais contribuem para as metas de sustentabilidade?

Procure oportunidades para reduzir seu impacto na sustentabilidade fazendo mudanças nas suas práticas de desenvolvimento, teste e implantação.

Adote operações capazes de introduzir rapidamente melhorias de sustentabilidade: teste e valide as possíveis melhorias de sustentabilidade antes de implantá-las na produção. Considere o custo do teste ao calcular o benefício futuro potencial de uma melhoria. Desenvolva operações de teste de baixo custo para permitir pequenas melhorias.

Mantenha a workload atualizada: sistemas operacionais, bibliotecas e aplicações atualizados podem melhorar a eficiência da workload e facilitar a adoção de tecnologias mais eficientes. O software atualizado também pode incluir recursos para medir o impacto na sustentabilidade da workload com mais precisão, pois os fornecedores oferecem recursos para atender às suas próprias metas de sustentabilidade.

Aumente a utilização de ambientes de compilação: use a automação e a infraestrutura como código para ativar ambientes de pré-produção quando necessário e desativá-los quando não forem usados. Um padrão comum é programar períodos de disponibilidade que coincidam com as horas de trabalho dos membros da equipe de desenvolvimento. A hibernação é uma ferramenta útil para preservar o estado e colocar rapidamente as instâncias online apenas quando necessário. Use tipos de instância com capacidade de expansão, instâncias spot, serviços de banco de dados elásticos, contêineres e outras tecnologias para alinhar a capacidade de desenvolvimento e teste com o uso.

Use parques de dispositivos gerenciados para testar: os parques de dispositivos gerenciados distribuem o impacto na sustentabilidade da fabricação do hardware e do uso de recursos entre vários locatários. Os parques de dispositivos gerenciados oferecem diversos tipos de dispositivos para que você ofereça compatibilidade com componentes de hardware mais antigos e menos populares e evite o impacto sobre a sustentabilidade do cliente devido a atualizações desnecessárias de dispositivos.

Recursos

Consulte os recursos a seguir para saber mais sobre nossas práticas recomendadas para Sustentabilidade.

Whitepaper

- Pilar Sustentabilidade

Vídeo

- The Climate Pledge

O processo de revisão

A revisão das arquiteturas precisa ser feita de maneira consistente, com uma abordagem sem culpa que incentive o aprofundamento. Ela deve ser um processo leve (com duração de horas, e não dias) e semelhante a uma conversa, e não uma auditoria. O objetivo de revisar uma arquitetura é identificar quaisquer problemas críticos que possam precisar ser abordados ou áreas que possam ser melhoradas. O resultado da revisão é um conjunto de ações que devem melhorar a experiência de um cliente usando a workload.

Conforme discutido na seção "Sobre arquitetura", cada membro da equipe deve assumir a responsabilidade pela qualidade da própria arquitetura. Recomendamos que os membros da equipe que criam uma arquitetura usem o Well-Architected Framework para analisar continuamente sua arquitetura, em vez de realizar uma reunião formal de análise. Uma abordagem quase contínua permite que os membros da equipe atualizem as respostas à medida que a arquitetura evolui e melhorem a arquitetura à medida que você fornece recursos.

O AWS Well-Architected Framework está alinhado à forma como a AWS analisa sistemas e serviços internamente. Ele tem como premissa um conjunto de princípios do projeto que influenciam a abordagem arquitetônica e perguntas que verifica se as pessoas não negligenciam áreas que aparecem com frequência na análise de causa-raiz (RCA). Sempre que houver um problema significativo com um sistema interno, um serviço da AWS ou um cliente, examinaremos a RCA para ver se podemos melhorar os processos de análise que usamos.

As revisões devem ser aplicadas nos principais marcos do ciclo de vida do produto, logo no início da fase de projeto para evitar portas só de entrada que são difíceis de trocar e, em seguida, antes da data de entrada em operação. (Muitas decisões são reversíveis, de mão dupla. Essas decisões podem usar um processo leve. As portas só de entrada são difíceis ou impossíveis de reverter e exigem mais inspeção antes de serem fabricadas.) Depois que entrar em produção, sua workload continuará evoluindo à medida que você adicionar novos recursos e altera implementações de tecnologias. A arquitetura de uma workload muda com o tempo. É necessário seguir boas práticas de higiene para impedir as características arquitetônicas de se degradarem à medida que evoluírem. Ao fazer alterações significativas na arquitetura, você deve seguir um conjunto de processos de higiene, incluindo uma análise do Well-Architected.

Se você quiser usar a revisão como um snapshot único ou uma medida independente, precisará garantir que todas as pessoas certas participem da conversa. Muitas vezes, descobrimos que as revisões constituem a primeira vez em que a equipe realmente compreende o que implementou.

Uma abordagem que funciona bem ao analisar a workload de outra equipe é ter uma série de conversas informais sobre sua arquitetura, nas quais se pode ter as respostas para a maioria das perguntas. Em seguida, você pode continuar com uma ou duas reuniões para tirar dúvidas ou se aprofundar nas áreas de ambiguidade ou risco percebidas.

Aqui estão alguns itens sugeridos para facilitar suas reuniões:

- Uma sala de reuniões com quadros brancos
- Impressões diagramas ou notas de projeto
- Lista de ações de perguntas que exigem pesquisas fora de banda para responder (por exemplo, "ativamos ou não a criptografia?")

Após concluir uma revisão, você deverá ter uma lista de problemas que podem ser priorizados com base no contexto da sua empresa. Você também deverá considerar o impacto desses problemas no trabalho diário de sua equipe. Se você resolver esses problemas com antecedência, poderá disponibilizar mais tempo para trabalhar na criação de valor empresarial, em vez de resolver problemas recorrentes. Ao abordar os problemas, é possível atualizar a análise para ver como a arquitetura está melhorando.

Embora o valor de uma análise seja claro após sua realização, você pode descobrir que uma nova equipe pode ser resistente a princípio. Aqui estão algumas objeções que podem ser tratadas por meio da instrução da equipe sobre os benefícios de uma análise:

- "Estamos muito ocupados!" (Geralmente dito quando a equipe está se preparando para um lançamento importante.)
  - Se você está se preparando para um grande lançamento, provavelmente deseja que ele ocorra sem problemas. A revisão permitirá que você entenda os problemas que pode ter perdido.
  - Recomendamos fazer revisões no início do ciclo de vida do produto para descobrir riscos e desenvolver um plano de mitigação alinhado ao roteiro de entrega de recursos.
- "Não temos tempo para fazer nada com os resultados!" (Geralmente, quando há um evento que não pode ser adiado, como o Super Bowl, no qual estão focados.)
  - Esses eventos não podem ser adiados. Deseja realmente entrar nele sem conhecer os riscos em sua arquitetura? Mesmo se você não abordar todos esses problemas, ainda poderá ter playbooks para lidar com eles, caso ocorram.
- "Não queremos que outras pessoas saibam os segredos da implementação da nossa solução!"

- Se você apresentar as perguntas do Well-Architected Framework aos membros da equipe, eles verão que nenhuma das perguntas revela qualquer informação de propriedade comercial ou técnica.

Ao realizar várias análises com as equipes da sua organização, é possível identificar problemas temáticos. Por exemplo, você pode ver que um grupo de equipes tem grupos de problemas em um pilar ou tópico específico. Veja todas as análises de maneira holística e identifique quaisquer mecanismos, treinamento ou palestras de engenharia que possam ajudar a resolver esses problemas temáticos.

Conclusão

O AWS Well-Architected Framework fornece práticas recomendadas de arquitetura nos seis pilares para projetar e operar sistemas confiáveis, seguros, eficientes, econômicos e sustentáveis na nuvem. O Framework fornece um conjunto de perguntas que permite analisar uma arquitetura existente ou proposta. Ele também fornece um conjunto de práticas recomendadas da AWS para cada pilar. O uso do Framework em sua arquitetura o ajudará a produzir sistemas estáveis e eficientes, permitindo que você se concentre em seus requisitos funcionais.

Colaboradores

Os seguintes indivíduos e organizações contribuíram para este documento:

- Brian Carlson, líder de operações do Well-Architected, Amazon Web Services
- Ben Potter, líder de segurança do Well-Architected, Amazon Web Services
- Seth Eliot: líder de confiabilidade do Well-Architected, Amazon Web Services
- Eric Pullen, arquiteto sênior de soluções, Amazon Web Services
- Rodney Lester, arquiteto líder de soluções, Amazon Web Services
- Jon Steele, gerente técnico sênior de contas, Amazon Web Services
- Max Ramsay, arquiteto líder de soluções de segurança, Amazon Web Services
- Callum Hughes, arquiteto de soluções, Amazon Web Services
- Ben Mergen, arquiteto líder de soluções de custos, Amazon Web Services
- Chris Kozlowski, gerente técnico de contas especialista sênior, Enterprise Support, Amazon Web Services
- Alex Livingstone, arquiteto especialista em soluções líder, operações na nuvem, Amazon Web Services
- Paul Moran, tecnólogo líder, Enterprise Support, Amazon Web Services
- Peter Mullen, consultor, Professional Services, Amazon Web Services
- Chris Pates, gerente técnico de contas especialista sênior, Enterprise Support, Amazon Web Services
- Arvind Raghunathan, gerente técnico de contas especialista líder, Enterprise Support, Amazon Web Services
- Sam Mokhtari, arquiteto líder de soluções de eficiência, Amazon Web Services

Outras fontes de leitura

Centro de Arquitetura da AWS

AWS Compatibilidade da nuvem

Programa de parceiros do AWS Well-Architected

AWS Well-Architected Tool

Página inicial do AWS Well-Architected

Whitepaper Pilar Excelência operacional

Whitepaper Pilar Segurança

Whitepaper Pilar Confiabilidade

Whitepaper Pilar Eficiência de performance

Whitepaper Pilar Otimização de custos

Whitepaper Pilar Sustentabilidade

Amazon Builders' Library

Revisões do documento

Para ser notificado sobre atualizações desse whitepaper, inscreva-se no feed RSS.

Alteração	Descrição	Data
Orientação atualizada sobre práticas recomendadas	Atualizações em grande escala nas práticas recomendadas foram feitas em todo o pilar. Tanto a segurança quanto o custo receberam novas práticas recomendadas.	27 de junho de 2024
Atualização principal	Atualizações importantes nos pilares.	3 de outubro de 2023
Atualizações para o novo Framework	Atualizações nas práticas recomendadas com recomendações e adição de novas práticas recomendadas. Novas perguntas adicionad as aos pilares Segurança e Otimização de custos.	10 de abril de 2023
Atualização secundária	Adição da definição de nível de esforço e atualização das práticas recomendadas no apêndice.	20 de outubro de 2022
Whitepaper atualizado	Adição do pilar Sustentab ilidade e atualização dos links.	2 de dezembro de 2021
Atualização principal	Pilar de Sustentabilidade adicionado ao framework.	20 de novembro de 2021
Atualização secundária	Remoção de linguagem não inclusiva.	22 de abril de 2021

Atualização secundária	Correção de vários links.	10 de março de 2021
Atualização secundária	Pequenas alterações editoriais.	15 de julho de 2020
Atualizações para o novo Framework	Revisão e reescrita da maioria das perguntas e respostas.	8 de julho de 2020
Whitepaper atualizado	Adição do AWS Well-Architected Tool, de links para os laboratórios do AWS Well-Architected e parceiros do AWS Well-Architected, além de correções secundárias para possibilitar uma versão em várias linguagens do Framework.	1 de julho de 2019
Whitepaper atualizado	Revisão e reescrita da maioria das perguntas e respostas para garantir que as perguntas se concentrem em um tópico de cada vez. Isso fez com que algumas perguntas anteriores fossem divididas em várias perguntas. Adição de termos comuns às definições (workload, componente etc). Apresentação da pergunta no corpo principal alterada para incluir texto descritivo.	1 de novembro de 2018
Whitepaper atualizado	Atualizações para simplificar o texto de pergunta, padronizar respostas e melhorar a legibilidade.	1º de junho de 2018

Whitepaper atualizado	O trecho sobre excelência operacional foi movido para a frente dos pilares e reescrito para enquadrar outros pilares. Outros pilares foram atualizados para refletir a evolução da AWS.	1 de novembro de 2017
Whitepaper atualizado	Atualização do Framework para incluir o pilar Excelência operacional e revisão e atualização dos outros pilares para reduzir a duplicação e incorporar aprendizados da realização de análises com milhares de clientes.	1º de novembro de 2016
Atualizações menores	Atualização do apêndice com as informações atuais do Amazon CloudWatch Logs.	1º de novembro de 2015
Publicação inicial	Publicação do AWS Well-Architected Framework.	1 de outubro de 2015

> **Note**
>
> Para assinar as atualizações de RSS, você deve ter um plug-in RSS habilitado para o navegador que você está usando.

Versões do framework:

- 2023-10-03 (atual)
- 2023-04-10
- 2022-03-31

Apêndice: Perguntas e práticas recomendadas

Este apêndice resume todas as perguntas e práticas recomendadas do AWS Well-Architected Framework.

Pilares

- Excelência operacional
- Segurança
- Confiabilidade
- Eficiência de performance
- Otimização de custo
- Sustentabilidade

Excelência operacional

O pilar Excelência operacional inclui a capacidade de oferecer conformidade com o desenvolvimento e de executar workloads com eficácia, obter insights sobre as operações e melhorar continuamente processos e procedimentos de suporte para oferecer valor empresarial. Recomendações sobre implementação estão disponíveis no whitepaper Pilar Excelência operacional.

Áreas de práticas recomendadas

- Organização
- Preparar
- Operar
- Evoluir

Organização

Perguntas

- OPS 1. Como você determina quais são suas prioridades?
- OPS 2. Como você pode estruturar sua organização para oferecer suporte aos resultados comerciais?
- OPS 3. Como a cultura organizacional oferece suporte aos resultados comerciais?

OPS 1. Como você determina quais são suas prioridades?

Todos devem entender seu papel no sucesso dos negócios. Tenha objetivos compartilhados para definir as prioridades dos recursos. Isso maximizará os benefícios de seus esforços.

Práticas recomendadas

- OPS01-BP01 Avaliar as necessidades dos clientes
- OPS01-BP02 Avaliar as necessidades dos clientes internos
- OPS01-BP03 Avaliar os requisitos de governança
- OPS01-BP04 Avaliar os requisitos de conformidade
- OPS01-BP05 Avaliar o cenário de ameaças
- OPS01-BP06 Avaliar as compensações ao gerenciar benefícios e riscos

OPS01-BP01 Avaliar as necessidades dos clientes

Envolva as principais partes interessadas, incluindo equipes de negócios, de desenvolvimento e operacionais, a fim de determinar onde concentrar os esforços nas necessidades de clientes externos. Isso garantirá que você tenha um entendimento completo do suporte às operações, o que é necessário para obter os resultados desejados nos negócios.

Resultado desejado:

- Você trabalha de trás para frente partindo dos resultados do cliente.
- Você entende como as práticas operacionais apoiam os resultados e os objetivos comerciais.
- Você envolve todas as partes relevantes.
- Você tem mecanismos para registrar as necessidades do cliente.

Práticas comuns que devem ser evitadas:

- Você decidiu não oferecer suporte ao cliente fora do horário comercial principal, mas não analisou dados históricos de solicitação de suporte. Você não sabe se isso afetará seus clientes.
- Você está desenvolvendo um novo recurso, mas não envolveu seus clientes para descobrir se ele é desejado, em qual formato é desejado e sem experimentação para validar a necessidade e o método de entrega.

Benefícios de implementar esta prática recomendada: os clientes cujas necessidades são atendidas apresentam uma probabilidade muito maior de permanecerem como clientes. Avaliar e compreender as necessidades de clientes externos informará como você priorizará seus esforços para entregar valor empresarial.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Compreenda as necessidades empresariais: o sucesso nos negócios é possibilitado pelos objetivos e pelo entendimento compartilhados entre as partes interessadas, incluindo equipes de negócios, de desenvolvimento e operacionais.

Revise os objetivos, as necessidades e as prioridades de negócios dos clientes externos: envolva as principais partes interessadas, incluindo as equipes corporativas, de desenvolvimento e operacionais, para discutir as metas, as necessidades e as prioridades dos clientes externos. Isso garantirá que você tenha um entendimento completo do suporte às operações que é necessário para obter resultados nos negócios.

Estabeleça uma compreensão compartilhada: estabeleça um entendimento compartilhado das funções corporativas da workload, das funções de cada uma das equipes na operação da workload e de como esses fatores apoiam seus objetivos empresariais compartilhados entre os clientes internos e externos.

Recursos

Práticas recomendadas relacionadas:

- OPS11-BP03 Implementar loops de feedback

OPS01-BP02 Avaliar as necessidades dos clientes internos

Envolva as principais partes interessadas, incluindo equipes de negócios, de desenvolvimento e operacionais, ao determinar onde concentrar os esforços nas necessidades de clientes internos. Isso garantirá que você tenha um entendimento completo do suporte às operações necessário para obter resultados nos negócios.

Resultado desejado:

- Use as prioridades estabelecidas para concentrar os esforços de melhoria onde eles terão maior impacto (por exemplo, desenvolvendo habilidades de equipe, melhorando a performance da workload, reduzindo custos, automatizando runbooks ou aprimorando o monitoramento).
- Atualize suas prioridades conforme as necessidades mudam.

Práticas comuns que devem ser evitadas:

- Você decidiu alterar as alocações de endereços IP para as equipes de produtos, sem consultá-las, para facilitar o gerenciamento da rede. Você não sabe o impacto que isso terá em suas equipes de produtos.
- Você está implementando uma nova ferramenta de desenvolvimento, mas não envolveu seus clientes internos para descobrir se ela é necessária ou se é compatível com suas práticas existentes.
- Você está implementando um novo sistema de monitoramento, mas não entrou em contato com os clientes internos para descobrir se eles têm necessidades de monitoramento ou relatórios que devam ser consideradas.

Benefícios de implementar esta prática recomendada: avaliar e compreender as necessidades de clientes internos informará como você priorizará seus esforços para entregar valor comercial.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

- Compreenda as necessidades empresariais: o sucesso nos negócios é possibilitado pelos objetivos e pelo entendimento compartilhados entre as partes interessadas, incluindo equipes de negócios, de desenvolvimento e operacionais.
- Revise os objetivos, as necessidades e as prioridades de negócios dos clientes internos: envolva as principais partes interessadas, incluindo as equipes corporativas, de desenvolvimento e operacionais, para discutir as metas, as necessidades e as prioridades dos clientes internos. Isso garantirá que você tenha um entendimento completo do suporte às operações que é necessário para obter resultados nos negócios.
- Estabeleça uma compreensão compartilhada: estabeleça um entendimento compartilhado das funções corporativas da workload, das funções de cada uma das equipes na operação da workload e de como esses fatores apoiam seus objetivos empresariais compartilhados entre os clientes internos e externos.

Recursos

Práticas recomendadas relacionadas:

- OPS11-BP03 Implementar loops de feedback

OPS01-BP03 Avaliar os requisitos de governança

Governança refere-se a um conjunto de políticas, regras ou frameworks usados por uma empresa para atingir metas comerciais. Os requisitos de governança são gerados dentro da organização. Eles podem afetar os tipos de tecnologia que você escolhe ou influenciar a maneira como opera sua workload. Incorpore requisitos de governança organizacional em sua workload. Conformidade é a capacidade de demonstrar que você implementou os requisitos de governança.

Resultado desejado:

- Os requisitos de governança são incorporados ao design arquitetural e à operação da workload.
- Você pode fornecer prova de que seguiu os requisitos de governança.
- Os requisitos de governança são revistos e atualizados regularmente.

Práticas comuns que devem ser evitadas:

- Sua organização exige que a conta-raiz tenha autenticação multifator. Você não implementa esse requisito e a conta-raiz é comprometida.
- Durante o design da workload, você escolhe um tipo de instância que não é aprovado pelo departamento de TI. Você não consegue iniciar a workload e precisa começar a reprojetá-la.
- É obrigatório ter um plano de recuperação de desastres. Você não cria um, e a workload sofre uma interrupção prolongada.
- Sua equipe quer usar novas instâncias, mas seus requisitos de governança não foram atualizados para permiti-las.

Benefícios de implementar esta prática recomendada:

- A aderência aos requisitos de governança alinha sua workload às políticas da organização como um todo.
- Os requisitos de governança refletem os padrões e as práticas recomendas do setor para sua organização.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Identifique o requisito de governança trabalhando com as partes interessadas e as organizações de governança. Inclua os requisitos de governança em sua workload. Prepare-se para demonstrar prova de que você seguiu os requisitos de governança.

Exemplo de cliente

Na AnyCompany Retail, a equipe de operações em nuvem trabalha com as partes interessadas dentro da organização para desenvolver requisitos de governança. Por exemplo, eles proíbem acesso SSH a instâncias do Amazon EC2. Caso as equipes precisem de acesso ao sistema, elas deverão usar o AWS Systems Manager Session Manager. A equipe de operações em nuvem atualiza regularmente os requisitos de governança à medida que novos serviços são disponibilizados.

Etapas de implementação

1. Identifique as partes interessadas referentes à sua workload, incluindo quaisquer equipes centralizadas.
2. Trabalhe com as partes interessadas para identificar requisitos de governança.
3. Assim que gerar uma lista, priorize os itens de melhoria e comece a implementá-los na workload.
    a. Use serviços como o AWS Config para criar governança como código e validar se os requisitos de governança são seguidos.
    b. Se você usa i AWS Organizations, pode fazer uso das políticas de controle de serviços para implementar os requisitos de governança.
4. Forneça documentação que valide a implementação.

Nível de esforço do plano de implementação: Médio. A implementação de requisitos de governança pode exigir a reformulação da sua workload.

Recursos

Práticas recomendadas relacionadas:

- OPS01-BP04 Avaliar os requisitos de conformidade: A conformidade é como governança, mas acontece fora da organização.

Documentos relacionados:

- Gerenciamento e governança da AWS: guia do ambiente de nuvem
- Práticas recomendadas para Políticas de controle de serviços do AWS Organizations em um ambiente com várias contas
- Governança na Nuvem AWS: o equilíbrio certo entre agilidade e segurança
- O que é governança, risco e conformidade (GRC)?

Vídeos relacionados:

- Gestão e governança da AWS: configuração, conformidade e auditoria – AWS Online Tech Talks
- AWS re:Inforce 2019: Governança para a era da nuvem (DEM12-R1)
- AWS re:Invent 2020: Alcançar a conformidade como código usando o AWS Config
- AWS re:Invent 2020: Governança ágil na AWS GovCloud (US)

Exemplos relacionados:

- Exemplos de pacotes de conformidade da AWS Config

Serviços relacionados:

- AWS Config
- Políticas de controle de serviço do AWS Organizations

OPS01-BP04 Avaliar os requisitos de conformidade

Os requisitos de conformidade normativos, setoriais e internos são um importante motivador para definir as prioridades de sua organização. Seu framework de conformidade pode impedir que você use tecnologias ou localizações geográficas específicas. Realize a devida diligência se não for identificado nenhum framework de conformidade externo. Gere auditorias ou relatórios que validem a conformidade.

Se você anunciar que seu produto atende a padrões de conformidade específicos, deverá ter um processo interno para garantir a conformidade contínua. Os exemplos de padrões de conformidade incluem PCI DSS, FedRAMP e HIPAA. Os padrões de conformidade aplicáveis são determinados

por vários fatores, por exemplo, quais tipos de dados a solução armazena ou transmite e a quais regiões a solução oferece suporte.

Resultado desejado:

- Os requisitos de conformidade normativos, setoriais e internos são incorporados na seleção arquitetural.
- É possível validar a conformidade e gerar relatórios de auditoria.

Práticas comuns que devem ser evitadas:

- Partes da workload podem ser enquadradas no framework Payment Card Industry Data Security Standard (PCI-DSS), mas a workload armazena dados de cartões de crédito não criptografados.
- Seus desenvolvedores e arquitetos de software não estão cientes do framework de conformidade que sua organização deve adotar.
- A auditoria anual Systems and Organizations Control (SOC2) Tipo II será feita em breve e você não consegue verificar se esses controles estão em vigor.

Benefícios de implementar esta prática recomendada:

- Avaliar e compreender os requisitos de conformidade que se aplicam à sua workload informará como você prioriza seus esforços para entregar valor empresarial.
- Você escolhe as localizações e tecnologias corretas, que são congruentes com seu framework de conformidade.
- Quando a workload é projetada para ser auditável, é possível provar que você está seguindo seu framework de conformidade.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Implementar essa prática recomendada significa incorporar os requisitos de conformidade no processo de design da arquitetura. Os membros de sua equipe estão cientes do framework de conformidade necessário. Você valida a conformidade de acordo com o framework.

Exemplo de cliente

A AnyCompany Retail armazena informações de cartão de crédito dos clientes. Os desenvolvedores da equipe de armazenamento de cartões sabem que eles precisam respeitar o framework PCI-DSS. Eles adotaram medidas para verificar que as informações de cartão de crédito são armazenadas e acessadas com segurança de acordo com o framework PCI-DSS. Todo ano, eles trabalham com a equipe de segurança para validar a conformidade.

Etapas de implementação

1. Trabalhe com as equipes de segurança e governança para determinar quais frameworks de conformidade normativos, setoriais ou internos a workload deve seguir. Incorpore os frameworks de conformidade em sua workload.
   a. Valide a conformidade contínua dos recursos da AWS com serviços como AWS Compute Optimizer e AWS Security Hub.
2. Instrua os membros da equipe sobre os requisitos de conformidade para que possam operar e expandir a workload de acordo com eles. Os requisitos de conformidade devem ser incluídos nas escolhas de arquitetura e tecnologia.
3. Dependendo do framework de conformidade, talvez seja necessário gerar um relatório de auditoria ou conformidade. Trabalhe com sua organização para automatizar esse processo o máximo possível.
   a. Use serviços como AWS Audit Manager para gerar, validar a conformidade e gerar relatórios de auditoria.
   b. Você pode baixar documentos de segurança e conformidade da AWS com o AWS Artifact.

Nível de esforço do plano de implementação: Médio. A implementação de frameworks de conformidade pode ser um desafio. A geração de relatórios de auditoria e de documentos de conformidade aumenta ainda mais a complexidade.

Recursos

Práticas recomendadas relacionadas:

- SEC01-BP03 Identificar e validar objetivos do controle: os objetivos do controle de segurança são uma parte importante da conformidade geral.
- SEC01-BP06 Automatizar os testes e a validação de controles de segurança em pipelines: como parte de seus pipelines, valide os controles de segurança. Você também pode gerar documentação de conformidade para novas alterações.

- SEC07-BP02 Definir controles de proteção de dados: muitos frameworks de conformidade têm como base políticas de tratamento e armazenamento de dados.
- SEC10-BP03 Preparar recursos forenses: às vezes, os recursos forenses podem ser usados em auditorias de conformidade.

Documentos relacionados:

- Centro de Conformidade da AWS
- Recursos de conformidade do AWS
- Whitepaper Risco e conformidade da AWS
- Modelo de responsabilidade compartilhada da AWS
- Serviços da AWS em escopo por programas de conformidade

Vídeos relacionados:

- AWS re:Invent 2020: Alcançar a conformidade como código usando o AWS Compute Optimizer
- AWS re:Invent 2021: Conformidade, garantia e auditoria na nuvem
- AWS Summit ATL 2022: Implementar conformidade, garantia e auditoria na AWS (COP202)

Exemplos relacionados:

- PCI DSS e as Práticas Recomendadas de Segurança Básica da AWS na AWS

Serviços relacionados:

- AWS Artifact
- AWS Audit Manager
- AWS Compute Optimizer
- AWS Security Hub

OPS01-BP05 Avaliar o cenário de ameaças

Avalie as ameaças à empresa (por exemplo, concorrência, risco e passivos empresariais, riscos operacionais e ameaças à segurança da informação) e mantenha as informações atuais em um registro de risco. Inclua o impacto dos riscos ao determinar onde concentrar os esforços.

O Well-Architected Framework enfatiza o aprendizado, a medição e o aprimoramento. Ele oferece uma abordagem consistente para avaliar arquiteturas e implementar designs que escalem ao longo do tempo. A AWS fornece o AWS Well-Architected Tool para ajudar você a analisar sua abordagem antes do desenvolvimento e o estado de suas workloads antes e durante a produção. Você pode compará-las com as práticas recomendadas de arquitetura mais recentes da AWS, monitorar o status geral das workloads e receber insights sobre possíveis riscos.

Os clientes da AWS são elegíveis para uma revisão orientada do Well-Architected para suas workloads de missão crítica a fim de avaliar suas arquiteturas em relação às práticas recomendadas da AWS. Os clientes Enterprise Support são elegíveis para uma revisão~de operações que foi desenvolvida para ajudá-los a identificar lacunas em sua abordagem de operação na nuvem.

O envolvimento entre equipes dessas avaliações ajuda a estabelecer um entendimento comum de suas workloads e como as funções da equipe contribuem para o sucesso. As necessidades identificadas pela avaliação podem ajudar a moldar suas prioridades.

O AWS Trusted Advisor é uma ferramenta que fornece acesso a um conjunto principal de verificações que recomendam otimizações que podem ajudar a moldar suas prioridades. Os clientes Business e Enterprise Support recebem acesso a verificações adicionais com foco em segurança, confiabilidade, performance e otimização de custos que podem ajudar a moldar as prioridades.

Resultado desejado:

- Você revisa e age regularmente com base no Well-Architected e nos resultados do Trusted Advisor.
- Você está ciente do status do patch mais recente dos seus serviços.
- Você entende o risco e o impacto das ameaças conhecidas e toma medidas adequadas.
- Você implementa mitigações conforme necessário.
- Você fornece informações sobre as ações e o contexto.

Práticas comuns que devem ser evitadas:

- Você está usando uma versão antiga de uma biblioteca de software no seu produto. Você não está ciente das atualizações de segurança na biblioteca para problemas que podem ter um impacto indesejado na workload.
- Seu concorrente acabou de lançar uma versão do produto que lida com muitas das reclamações de seus clientes sobre seu produto. Você não priorizou a abordagem de nenhum desses problemas conhecidos.

- Os reguladores buscam empresas como a sua que não estejam em conformidade com os requisitos de conformidade normativa legais. Você não priorizou a abordagem de nenhum dos requisitos de conformidade pendentes.

Benefícios de implementar esta prática recomendada: identificar e compreender as ameaças à sua organização e à workload permite determinar quais ameaças devem ser resolvidas, a prioridade delas e os recursos necessários para isso.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

- Avalie o cenário de ameaças: avalie as ameaças aos negócios (como concorrência, riscos e responsabilidades comerciais, riscos operacionais e ameaças à segurança da informação), para que você possa incluir o impacto delas ao determinar onde concentrar os esforços.
  - Boletins de segurança mais recentes da AWS
  - AWS Trusted Advisor
- Mantenha um modelo de ameaças: estabeleça e mantenha um modelo de ameaças que identifique possíveis ameaças, mitigações planejadas e implementadas e a prioridade delas. Analise a probabilidade de as ameaças se manifestarem como incidentes, o custo de recuperação desses incidentes, o dano esperado causado e o custo para evitá-los. Revise as prioridades à medida que o conteúdo do modelo de ameaça muda.

Recursos

Práticas recomendadas relacionadas:

- SEC01-BP07 Identificar ameaças e priorizar mitigações usando um modelo de ameaça

Documentos relacionados:

- Conformidade da Nuvem AWS
- Boletins de segurança mais recentes da AWS
- AWS Trusted Advisor

Vídeos relacionados:

- AWS re:INFORCE 2023: Uma ferramenta para ajudar a melhorar sua modelagem de ameaças

OPS01-BP06 Avaliar as compensações ao gerenciar benefícios e riscos

Interesses conflitantes de várias partes podem dificultar a priorização de esforços, a criação de capacidades e a entrega de resultados alinhados às estratégias de negócios. Por exemplo, talvez seja solicitado que você acelere a comercialização de novos recursos em vez de otimizar os custos da infraestrutura de TI. Isso pode colocar duas partes interessadas em conflito. Nessas situações, é necessário encaminhar as decisões a uma autoridade superior a fim de resolver conflitos. Dados são necessários para remover o apego emocional do processo de tomada de decisões.

O mesmo desafio pode ocorrer em nível tático. Por exemplo, a escolha entre usar tecnologias de bancos de dados relacionais ou não relacionais pode ter um impacto significativo na operação de uma aplicação. É fundamental entender os resultados previsíveis de várias opções.

A AWS pode ajudar a instruir suas equipes sobre a AWS e os serviços que ela fornece para que compreendam melhor como as escolhas que elas fazem podem ter um impacto na workload. Use os recursos fornecidos pelo AWS Support (Centro de Conhecimentos da AWS, Fóruns de discussão da AWS e o AWS Support Center), bem como a documentação da AWS para instruir suas equipes. Em caso de dúvidas, entre em contato com o AWS Support.

A AWS também compartilha práticas recomendadas e padrões operacionais na Amazon Builders' Library. Inúmeras outras informações úteis podem ser obtidas por meio do Blog da AWS e no podcast oficial da AWS.

Resultado desejado: você tem uma estrutura de governança de tomada de decisão claramente definida para facilitar decisões importantes em todos os níveis da sua organização de fornecimento de nuvem. Esse framework inclui recursos como um registro de riscos, funções definidas que estão autorizadas a tomar decisões e modelos definidos para cada nível de decisão que pode ser tomada. O framework define com antecedência como os conflitos são resolvidos, quais dados precisam ser apresentados e como as opções são priorizadas, para que, uma vez tomadas as decisões, você possa se comprometer sem demora. O framework de tomada de decisões inclui uma abordagem padronizada para analisar e avaliar os benefícios e os riscos de cada decisão e entender as vantagens e as desvantagens. Isso pode incluir fatores externos, como a adesão aos requisitos de conformidade regulatória.

Práticas comuns que devem ser evitadas:

- Seus investidores solicitam que você demonstre conformidade com os Payment Card Industry Data Security Standards (PCI DSS). Você não pensa nas concessões entre atender a essa solicitação e continuar com seus esforços de desenvolvimento atuais. Em vez disso, você prossegue com os esforços de desenvolvimento sem demonstrar conformidade. Seus investidores deixam de apoiar sua empresa devido a preocupações com a segurança da plataforma e de seus investimentos.
- Você decidiu incluir uma biblioteca que um de seus desenvolvedores encontrou na Internet. Você não avaliou os riscos de adoção dessa biblioteca de origem desconhecida e não sabe se ela contém vulnerabilidades ou código mal-intencionado.
- A justificativa comercial original para sua migração foi baseada na modernização de 60% das workloads de aplicações. No entanto, devido a dificuldades técnicas, foi tomada a decisão de modernizar apenas 20%, ocasionando uma redução nos benefícios planejados de longo prazo, o aumento do trabalho do operador para que as equipes de infraestrutura apoiem manualmente os sistemas herdados e uma maior dependência do desenvolvimento de novos conjuntos de habilidades em suas equipes de infraestrutura que não estavam preparadas para essa mudança.

Benefícios de implementar esta prática recomendada: alinhar e apoiar totalmente as prioridades de negócios em nível de diretoria, compreender os riscos de alcançar o sucesso, tomar decisões informadas e agir adequadamente quando os riscos impedem as chances de sucesso. Compreender as implicações e as consequências de suas decisões ajuda você a priorizar suas opções e a possibilitar que os líderes cheguem a um acordo mais depressa, gerando melhores resultados comerciais. Identificar os benefícios gerados por suas escolhas e estar ciente dos riscos para a organização ajuda você a tomar decisões orientadas por dados, e não baseadas em histórias.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

O gerenciamento de benefícios e riscos deve ser definido por um órgão regulador que oriente os requisitos para a tomada de decisões importantes. Convém que as decisões sejam tomadas e priorizadas com base em como elas beneficiam a organização, com uma compreensão dos riscos envolvidos. Informações precisas são essenciais para a tomada de decisões organizacionais. Isso deve se basear em medições sólidas e ser definido por práticas comuns de análise de custo-benefício do setor. Para tomar esse tipo de decisão, encontre um equilíbrio entre autoridade centralizada e descentralizada. Sempre há concessões, e é importante entender como cada escolha afeta as estratégias definidas e os resultados comerciais desejados.

Etapas de implementação

1. Formalize as práticas de mensuração de benefícios dentro de um framework completo de governança de nuvem.
   a. Contrabalance o controle central da tomada de decisões e a autoridade descentralizada para algumas decisões.
   b. Entenda que processos fatigantes de tomada de decisões impostos a cada decisão podem diminuir seu ritmo.
   c. Incorpore fatores externos em seu processo de tomada de decisões (como requisitos de conformidade).
2. Estabeleça um framework de tomada de decisões acordado para vários níveis de decisão, incluindo quem deve desobstruir decisões sujeitas a interesses conflitantes.
   a. Centralize decisões unidirecionais que podem ser irreversíveis.
   b. Permita que as decisões bidirecionais sejam tomadas por líderes organizacionais de nível inferior.
3. Entenda e gerencie benefícios e riscos. Contrabalance os benefícios das decisões com os riscos envolvidos.
   a. Identifique os benefícios: identifique os benefícios com base nas metas, necessidades e prioridades da empresa. Os exemplos são os seguintes: impacto no caso de negócios, tempo até a comercialização, segurança, confiabilidade, performance e custo.
   b. Identifique os riscos: identifique os riscos com base nas metas, necessidades e prioridades da empresa. Os exemplos são os seguintes: tempo para comercialização, segurança, confiabilidade, performance e custo.
   c. Avalie os benefícios em comparação com os riscos e tome decisões informadas: determine o impacto dos benefícios e riscos com base nas metas, necessidades e prioridades de suas principais partes interessadas, incluindo negócios, desenvolvimento e operações. Avalie o valor do benefício em relação à probabilidade de realização do risco e o custo do seu impacto. Por exemplo, enfatizar a velocidade de entrada no mercado em vez da confiabilidade pode oferecer vantagem competitiva. No entanto, isso poderá causar tempo de atividade reduzido se houver problemas de confiabilidade.
4. Imponha de modo programático as principais decisões que automatizam sua adesão aos requisitos de conformidade.

5. Utilize frameworks e recursos conhecidos do setor, como análise do fluxo de valor e lean, para estabelecer uma referência comparativa para a performance do estado atual, bem como métricas de negócios, e defina iterações de progresso em direção a melhorias nessas métricas.

Nível de esforço do plano de implementação: Médio-Alto

Recursos

Práticas recomendadas relacionadas:

- OPS01-BP05 Avaliar o cenário de ameaças

Documentos relacionados:

- Elementos da cultura de Dia 1 da Amazon | Tomar decisões de alta qualidade em alta velocidade
- Governança na nuvem
- Gerenciamento e governança: ambiente de nuvem
- Governança na nuvem e na era digital: partes um e dois

Vídeos relacionados:

- Podcast | Jeff Bezos | Sobre como tomar decisões

Exemplos relacionados:

- Tomar decisões informadas usando dados (The DevOps Sagas)
- Usar o mapeamento do fluxo de valor do desenvolvimento para identificar restrições aos resultados de DevOps

OPS 2. Como você pode estruturar sua organização para oferecer suporte aos resultados comerciais?

Suas equipes devem compreender o papel delas na obtenção de resultados empresariais. As equipes devem entender o respectivo papel no êxito de outras equipes e o papel das demais equipes em seu próprio êxito e estabelecer objetivos compartilhados. Entender a responsabilidade, a

propriedade, como as decisões são tomadas e quem tem autoridade para tomar decisões ajudará a concentrar os esforços e maximizar os benefícios das suas equipes.

Práticas recomendadas

- OPS02-BP01 Recursos com proprietários identificados
- OPS02-BP02 Processos e procedimentos com proprietários identificados
- OPS02-BP03 Atividades de operações com proprietários identificados responsáveis pela performance
- OPS02-BP04 Mecanismos disponíveis para gerenciar responsabilidades e a relação de propriedade
- OPS02-BP05 Mecanismos para solicitar adições, alterações e exceções
- OPS02-BP06 As responsabilidades entre as equipes são predefinidas ou negociadas

OPS02-BP01 Recursos com proprietários identificados

Os recursos para sua workload devem ter proprietários identificados para fins de controle de alterações, resolução de problemas e outras funções. Proprietários são atribuídos para workloads, contas, infraestrutura, plataformas e aplicações. A propriedade é registrada usando ferramentas como um registro central ou metadados anexados aos recursos. O valor comercial dos componentes indica os processos e procedimentos aplicados a eles.

Resultado desejado:

- Os recursos têm proprietários identificados usando metadados ou um registro central.
- Os membros da equipe podem identificar quem é proprietários dos recursos.
- Quando possível, as contas têm um único proprietário.

Práticas comuns que devem ser evitadas:

- Os contatos alternativos para suas Contas da AWS não estão preenchidos.
- Os recursos não têm as tags que identificam as equipes às quais eles pertencem.
- Você tem uma fila ITSM sem mapeamento de e-mail.
- Duas equipes são proprietárias de uma mesma parte essencial da infraestrutura.

Benefícios de implementar esta prática recomendada:

- O controle de alterações para recursos é fácil com a atribuição de propriedade.
- Você pode envolver os proprietários corretos na resolução de problemas.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Defina o que significa propriedade para os casos de uso de recursos em seu ambiente. A propriedade pode se referir a quem supervisiona alterações no recurso e oferece suporte ao recurso durante a resolução de problemas ou a quem é responsável pela parte financeira. Especifique e registre proprietários para os recursos, incluindo nome, informações de contato, organização e equipe.

Exemplo de cliente

A AnyCompany Retail define propriedade como a equipe ou a pessoa proprietária das alterações e do suporte para os recursos. Ela utiliza o AWS Organizations para gerenciar as Contas da AWS. Os contatos de conta alternativos são configurados usando as caixas de entrada de grupo. Cada fila ITSM é mapeada em um alias de e-mail. As tags identificam quem é proprietário dos recursos da AWS. Para outras plataformas e infraestrutura, a empresa possui uma página de wiki que identifica informações sobre propriedade e contato.

Etapas de implementação

1. Para começar, identifique a propriedade para sua organização. A propriedade pode estar relacionada a quem é proprietário do risco referente ao recurso, quem é proprietário das alterações referentes ao recurso ou quem oferece suporte ao recurso na resolução de problemas. Propriedade também pode significar propriedade financeira ou administrativa do recurso.
2. Use o AWS Organizations para gerenciar contas. Você pode gerenciar contatos alternativos centralmente para as suas contas.
    a. O uso de endereços de e-mail ou de números de telefones pertencentes à empresa para informações de contato ajuda você a acessá-los mesmo quando os indivíduos aos quais eles pertencem não estiverem mais na organização. Por exemplo, crie listas de distribuição de e-mail separadas para faturamento, operações e segurança, e configure-as como contatos de Faturamento, Segurança e Operações em cada Conta da AWS ativa. Várias pessoas receberão notificações da AWS e poderão respondê-las, mesmo que alguém esteja de férias, mude de função ou saia da empresa.

   b. Se uma conta não for gerenciada pelo AWS Organizations, os contatos alternativos da conta ajudarão a AWS a entrar em contato com o pessoal apropriado, se necessário. Configure os contatos alternativos da conta para apontar para um grupo em vez de uma pessoa.
3. Use tags para identificar proprietários de recursos da AWS. Você pode especificar os proprietários e as respectivas informações de contato em tags separadas.
   a. Você pode usar regras do AWS Config para garantir que os recursos tenham as tags de propriedade necessárias.
   b. Para obter orientações detalhadas sobre como criar uma estratégia de marcação para sua organização, consulte o whitepaper sobre Práticas recomendadas de marcação com tags da AWS.
4. Use o Amazon Q Business, um assistente de conversação que usa IA generativa para melhorar a produtividade da força de trabalho, responder a perguntas e concluir tarefas com base nas informações dos seus sistemas corporativos.
   a. Conecte o Amazon Q Business à fonte de dados da sua empresa. O Amazon Q Business oferece conectores pré-construídos para mais de 40 fontes de dados compatíveis, incluindo Amazon Simple Storage Service (Amazon S3), Microsoft SharePoint, Salesforce e Atlassian Confluence. Para obter mais informações, consulte Conectores do Amazon Q Business.
5. Para outros recursos, plataformas e infraestrutura, crie uma documentação que identifique a propriedade. Ela deve ser acessível a todos os membros da equipe.

Nível de esforço do plano de implementação: Baixo. Utilize informações de contato da conta e tags para atribuir propriedade a recursos da AWS. Para outros recursos, você pode usar algo simples como uma tabela em uma wiki para registrar a propriedade e informações de contato ou usar uma ferramenta de ITSM para mapear a propriedade.

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP02 Processos e procedimentos com proprietários identificados
- OPS02-BP04 Mecanismos disponíveis para gerenciar responsabilidades e a relação de propriedade

Documentos relacionados:

- Gerenciamento de contas da AWS: atualizar as informações de contato

- AWS Organizations: atualizar contatos alternativos em sua organização
- Whitepaper Práticas recomendadas de marcação com tags da AWS
- Criar aplicações de IA generativa empresariais privadas e seguras com o Amazon Q Business e o AWS IAM Identity Center
- O Amazon Q Business, agora disponível ao público em geral, ajuda a aumentar a produtividade da força de trabalho com a IA generativa
- Blog de operações e migrações da Nuvem AWS: implementar controles de marcação com tags automatizados e centralizados com o AWS Config e o AWS Organizations
- Blog de segurança da AWS: estenda seus hooks de pré-confirmação com o AWS CloudFormation Guard
- Blog de DevOps da AWS: integrar o AWS CloudFormation Guard em pipelines de CI/CD

Workshops relacionados:

- Workshop da AWS: marcação com tags

Exemplos relacionados:

- Regras do AWS Config: Amazon EC2 com tags obrigatórias e valores válidos

Serviços relacionados:

- Regras do AWS Config: tags obrigatórias
- AWS Organizations

OPS02-BP02 Processos e procedimentos com proprietários identificados

Entenda quem tem a propriedade da definição de processos e procedimentos individuais, por que esses processos e procedimentos específicos são usados e por que essa propriedade existe. Entender os motivos pelos quais processos e procedimentos específicos são usados ajuda a identificar oportunidades de melhoria.

Resultado desejado: sua organização terá um conjunto bem definido e mantido de processos e procedimentos para tarefas operacionais. O processo e os procedimentos são armazenados em um local central e estarão disponíveis para os membros da equipe. Os processos e os procedimentos

são atualizados com frequência, por meio de uma propriedade claramente atribuída. Sempre que possível, scripts, modelos e documentos de automação são implementados como código.

Práticas comuns que devem ser evitadas:

- Os processos não são documentados. Scripts fragmentados podem existir em estações de trabalho de operadores isoladas.
- O conhecimento de como usar scripts é mantido por algumas pessoas ou informalmente como conhecimento da equipe.
- Um processo legado precisa ser atualizado, mas a propriedade da atualização não está clara e o autor original não faz mais parte da organização.
- Processos e scripts não podem ser descobertos, portanto, não estão prontamente disponíveis quando necessário (por exemplo, em resposta a um incidente).

Benefícios de implementar esta prática recomendada:

- Os processos e os procedimentos impulsionam seus esforços para operar as workloads.
- Novos membros da equipe se tornam efetivos mais rapidamente.
- Tempo reduzido para mitigar incidentes.
- Diferentes membros da equipe (e equipes) podem usar os mesmos processos e procedimentos de maneira consistente.
- As equipes podem escalar os processos com procedimentos repetíveis.
- Processos e procedimentos padronizados ajudam a mitigar o impacto da transferência de responsabilidades de workload entre equipes.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

- Os processos e procedimentos possuem proprietários identificados que são responsáveis por suas definições.
  - Identifique as atividades de operações realizadas para oferecer suporte às suas workloads. Documente essas atividades em um local que possa ser localizado.
  - Identifique de maneira única a pessoa ou equipe responsável pela especificação de uma atividade. Ela é responsável por verificar se a atividade pode ser executada com êxito por um membro da equipe devidamente qualificado com as permissões, as ferramentas e o acesso

corretos. Se houver problemas com a execução dessa atividade, os membros da equipe que a executam serão responsáveis por fornecer os comentários detalhados necessários para que a atividade seja melhorada.

- Registre a propriedade dos metadados do artefato da atividade por meio de serviços como o AWS Systems Manager, documentos e do AWS Lambda. Registre a propriedade de recursos usando tags ou grupos de recursos, especificando as informações de propriedade e de contato. Use o AWS Organizations para criar políticas de marcação com tags e capturar as informações de propriedade e de contato.

- Com o tempo, esses procedimentos devem ser evoluídos para ser executados como código, reduzindo a necessidade de intervenção humana.
  - Por exemplo, pense em funções do AWS Lambda, modelos do CloudFormation ou documentos de automação do AWS Systems Manager.
  - Execute o controle de versão nos repositórios apropriados.
  - Inclua uma marcação de recursos adequada para que os proprietários e a documentação possam ser facilmente identificados.

Exemplo de cliente

A AnyCompany Retail define propriedade como a equipe ou o indivíduo que é responsável pelos processos de uma aplicação ou grupos de aplicações (que compartilham práticas e tecnologias de arquitetura comuns). Inicialmente, os processos e os procedimentos são documentados como guias passo a passo no sistema de gerenciamento de documentos, que podem ser descobertos por meio de tags na Conta da AWS que hospeda a aplicação e em grupos específicos de recursos dentro da conta. Ela utiliza o AWS Organizations para gerenciar as Contas da AWS. Com o tempo, esses processos são convertidos em código e os recursos são definidos usando a infraestrutura como código (como modelos do CloudFormation ou do AWS Cloud Development Kit (AWS CDK)). Os processos operacionais se tornam documentos de automação nas funções do AWS Systems Manager ou AWS Lambda, os quais podem ser iniciados como tarefas agendadas, em resposta a eventos como os alarmes do AWS CloudWatch ou os eventos do AWS EventBridge ou iniciados por solicitações em uma plataforma de gerenciamento de serviços de TI (ITSM). Todo processo tem tags para identificar a propriedade. A documentação para a automação e o processo é mantida nas páginas wiki geradas pelo repositório de código do processo.

Etapas de implementação

1. Documente os processos e os procedimentos existentes.

a. Revise e mantenha-os atualizados.

b. Identifique um proprietário para cada processo ou procedimento.

c. Coloque-os sob controle de versão.

d. Sempre que possível, compartilhe processos e procedimentos entre workloads e ambientes que compartilham projetos de arquitetura.

2. Estabeleça mecanismos de feedback e melhoria.

a. Defina políticas sobre a frequência com que os processos devem ser revisados.

b. Defina processos para revisores e aprovadores.

c. Implemente problemas ou uma fila de emissão de tíquetes para que o feedback seja fornecido e rastreado.

d. Sempre que possível, os processos e os procedimentos devem ter pré-aprovação e classificação de risco de um conselho de aprovação de mudanças (CAB).

3. Verifique se os processos e os procedimentos estão acessíveis e detectáveis por aqueles que precisam executá-los.

a. Use tags para indicar onde os processos e os procedimentos podem ser acessados para a workload.

b. Use mensagens relevantes de erros e eventos para indicar os processos ou os procedimentos apropriados para resolver um problema.

c. Use wikis e gerenciamento de documentos e torne processos e procedimentos pesquisáveis de forma consistente em toda a organização.

4. Automatize quando apropriado.

a. As automações devem ser desenvolvidas quando os serviços e as tecnologias fornecem uma API.

b. Instrua adequadamente sobre os processos. Desenvolva as histórias e os requisitos dos usuários e para automatizar esses processos.

c. Avalie com êxito o uso de processos e procedimentos, com o rastreamento de problemas para apoiar a melhoria iterativa.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP01 Recursos com proprietários identificados
- OPS02-BP04 Mecanismos disponíveis para gerenciar responsabilidades e a relação de propriedade
- OPS11-BP04 Gerenciar o conhecimento

Documentos relacionados:

- Whitepaper da AWS: Introdução a DevOps na AWS
- Whitepaper da AWS: Práticas recomendadas para marcação de recursos da AWS com tags
- Whitepaper da AWS: Organizar seu ambiente da AWS usando várias contas
- Blog de operações e migrações da Nuvem AWS – Criar uma prática de automação na nuvem para excelência operacional: práticas recomendadas da AWS Managed Services
- Blog de operações e migrações da Nuvem AWS: implementar controles de marcação com tags automatizados e centralizados com o AWS Config e o AWS Organizations
- Blog de segurança da AWS: estenda seus hooks de pré-confirmação com o AWS CloudFormation Guard
- Blog de DevOps da AWS: integrar o AWS CloudFormation Guard em pipelines de CI/CD

Workshops relacionados:

- Workshop Excelência operacional no AWS Well-Architected
- Workshop da AWS: marcação com tags

Vídeos relacionados:

- Como automatizar operações de TI na AWS
- AWS re:Invent 2020: Automatize tudo com o AWS Systems Manager
- AWS re:Inforce 2022: Automatizar o gerenciamento e a conformidade de patches usando a AWS (NIS306)
- AWS Supports Você: Mergulho profundo no AWS Systems Manager

Serviços relacionados:

- AWS Systems Manager: automação

- AWS Service Management Connector

OPS02-BP03 Atividades de operações com proprietários identificados responsáveis pela performance

Entenda quem tem a responsabilidade de realizar atividades específicas em workloads definidas e por que essa responsabilidade existe. Entender quem tem a responsabilidade de realizar atividades informa quem realizará a atividade, valida o resultado e fornece feedback ao proprietário da atividade.

Resultado desejado:

Sua organização define claramente as responsabilidades de realizar atividades específicas em workloads definidas e reagir aos eventos gerados pela workload. A organização documenta a propriedade dos processos e o cumprimento e torna essas informações detectáveis. Você analisa e atualiza as responsabilidades quando ocorrem mudanças organizacionais, e as equipes monitoram e medem a performance das atividades de identificação de defeitos e ineficiência. Você implementa mecanismos de feedback para rastrear defeitos e aprimoramentos e apoiar a melhoria iterativa.

Práticas comuns que devem ser evitadas:

- Você não documenta responsabilidades.
- Scripts fragmentados estão presentes em estações de trabalho de operadores isoladas. Apenas algumas pessoas sabem como usá-las ou se referem informalmente a elas como conhecimento de equipe.
- Um processo herdado precisa ser atualizado, mas ninguém sabe quem é responsável pelo processo e o autor original não faz mais parte da organização.
- Processos e scripts não podem ser descobertos, portanto, não estão prontamente disponíveis quando necessário (por exemplo, em resposta a um incidente).

Benefícios de implementar esta prática recomendada:

- Você sabe quem é responsável por realizar uma atividade, a quem notificar quando uma ação é necessária e quem realiza a ação, valida o resultado e fornece feedback ao responsável pela atividade.
- Os processos e os procedimentos impulsionam seus esforços para operar as workloads.
- Novos membros da equipe se tornam efetivos mais rapidamente.

- Você reduz o tempo necessário para atenuar incidentes.
- Equipes diferentes usam os mesmos processos e procedimentos para realizar tarefas de maneira consistente.
- As equipes podem escalar os processos com procedimentos repetíveis.
- Processos e procedimentos padronizados ajudam a atenuar o impacto da transferência de responsabilidades de workload entre equipes.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Para começar a definir responsabilidades, comece com a documentação existente, como matrizes de responsabilidade, processos e procedimentos, perfis e responsabilidades, bem como ferramentas e automação. Revise e organize discussões sobre as responsabilidades pelos processos documentados. Revise com as equipes para identificar desalinhamentos entre as responsabilidades documentadas e os processos. Aborde os serviços oferecidos com os clientes internos dessa equipe para identificar as lacunas de expectativas entre as equipes.

Analise e resolva as discrepâncias. Identifique oportunidades de melhoria e procure atividades frequentemente solicitadas e que consomem muitos recursos, que normalmente são fortes candidatas a melhorias. Examine as práticas recomendadas, os padrões e as recomendações para simplificar e padronizar as melhorias. Registre oportunidades de melhoria e acompanhe as melhorias até a conclusão.

Com o tempo, esses procedimentos devem ser desenvolvidos para ser executados como código, reduzindo a necessidade de intervenção humana. Por exemplo, os procedimentos podem ser iniciados como funções do AWS Lambda, modelos do AWS CloudFormation ou documentos de automação do AWS Systems Manager. Verifique se esses procedimentos têm controle de versão nos repositórios apropriados e inclua a marcação de recursos adequada para que as equipes possam identificar prontamente os proprietários e a documentação. Documente a responsabilidade pela realização das atividades e, depois, monitore as automações para iniciação e operação bem-sucedidas, bem como a performance dos resultados desejados.

Exemplo de cliente

A AnyCompany Retail define propriedade como a equipe ou o indivíduo que é responsável pelos processos de uma aplicação ou grupos de aplicações que compartilham práticas e tecnologias de arquitetura comuns. Inicialmente, a empresa documenta os processos e os procedimentos como

guias passo a passo no sistema de gerenciamento de documentos. Ela torna os procedimentos detectáveis usando tags na Conta da AWS que hospeda a aplicação e em grupos específicos de recursos dentro da conta, usando o AWS Organizations para gerenciar as Contas da AWS. Com o tempo, a AnyCompany Retail converte esses processos em código e define recursos usando a infraestrutura como código (por meio de serviços como o CloudFormation ou de modelos do AWS Cloud Development Kit (AWS CDK)). Os processos operacionais se tornam documentos de automação no AWS Systems Manager ou nas funções do AWS Lambda, os quais podem ser iniciados como tarefas agendadas em resposta a eventos como os alarmes do Amazon CloudWatch ou os eventos do Amazon EventBridge ou iniciados por solicitações em uma plataforma de gerenciamento de serviços de TI (ITSM). Todos os processos têm tags para identificar quem é responsável por eles. As equipes gerenciam a documentação para a automação e o processo nas páginas wiki geradas pelo repositório de código do processo.

Etapas de implementação

1. Documente os processos e os procedimentos existentes.
    a. Revise e verifique se eles estão atualizados.
    b. Verifique se cada processo ou procedimento tem um proprietário.
    c. Submeta os procedimentos ao controle de versão.
    d. Sempre que possível, compartilhe processos e procedimentos entre workloads e ambientes que compartilham projetos de arquitetura.
2. Estabeleça mecanismos de feedback e melhoria.
    a. Defina políticas sobre a frequência com que os processos devem ser revisados.
    b. Defina processos para revisores e aprovadores.
    c. Implemente uma fila de problemas ou de tíquetes para fornecer e rastrear o feedback.
    d. Sempre que possível, forneça pré-aprovação e classificação de risco para processos e procedimentos de um conselho de aprovação de mudanças (CAB).
3. Torne os processos e os procedimentos acessíveis e detectáveis pelos usuários que precisam executá-los.
    a. Use tags para indicar onde os processos e os procedimentos podem ser acessados para a workload.
    b. Use mensagens relevantes de erros e eventos para indicar os processos ou os procedimentos apropriados para resolver o problema.
    c. Use wikis ou gerenciamento de documentos para tornar os processos e os procedimentos pesquisáveis de forma consistente em toda a organização.

4. Automatize quando for apropriado.
    a. Quando os serviços e as tecnologias fornecerem uma API, desenvolva automações.
    b. Verifique se os processos estão bem compreendidos e desenvolva as histórias e os requisitos dos usuários para automatizar esses processos.
    c. Avalie o uso bem-sucedido de processos e procedimentos, e faça rastreamento dos problemas para contribuir com a melhoria iterativa.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP01 Recursos com proprietários identificados
- OPS02-BP02 Processos e procedimentos com proprietários identificados
- OPS02-BP04 Mecanismos disponíveis para gerenciar responsabilidades e a relação de propriedade
- OPS02-BP05 Mecanismos disponíveis para identificar a responsabilidade e a relação de propriedade
- OPS11-BP04 Gerenciar o conhecimento

Documentos relacionados:

- Whitepaper da AWS | Introdução a DevOps na AWS
- Whitepaper da AWS | Práticas recomendadas para marcação de recursos da AWS com tags
- Whitepaper da AWS | Organizar seu ambiente da AWS usando várias contas
- Blog de operações e migrações da Nuvem AWS | Criar uma prática de automação na nuvem para excelência operacional: práticas recomendadas da AWS Managed Services
- Workshop da AWS: marcação com tags
- AWS Service Management Connector

Vídeos relacionados:

- Centro de Conhecimentos da AWS Live | Recursos de marcação com tags da AWS

- AWS re:Invent 2020 | Automatize tudo com o AWS Systems Manager
- AWS re:Inforce 2022 | Automatizar o gerenciamento e a conformidade de patches usando a AWS (NIS306)
- AWS Supports Você | Mergulho profundo no AWS Systems Manager

Exemplos relacionados:

- Workshop Excelência operacional no AWS Well-Architected

OPS02-BP04 Mecanismos disponíveis para gerenciar responsabilidades e a relação de propriedade

Entenda as responsabilidades do seu perfil e como você contribui para os resultados comerciais, pois esse entendimento fornece informações sobre como priorizar tarefas e por que sua função é importante. Isso ajuda os membros da equipe a reconhecer as necessidades e reagir adequadamente. Quando os membros da equipe conhecem seus perfis, eles podem estabelecer propriedade, identificar oportunidades de melhoria e entender como influenciar ou fazer as mudanças apropriadas.

Ocasionalmente, uma responsabilidade pode não ter um proprietário claro. Nessas situações, desenvolva um mecanismo para resolver essa lacuna. Crie um caminho de escalação bem definido para que alguém com autoridade atribua propriedade ou estabeleça um plano para atender à necessidade em questão.

Resultado desejado: as equipes de sua organização têm responsabilidades claramente definidas que incluem como elas estão relacionadas aos recursos, ações a serem executadas, processos e procedimentos. Essas responsabilidades alinham-se às responsabilidades e às metas da equipe, bem como às responsabilidades de outras equipes. Você documenta as rotas de escalação de forma consistente e detectável e insere essas decisões em artefatos de documentação, como matrizes de responsabilidade, definições de equipe ou páginas wiki.

Práticas comuns que devem ser evitadas:

- As responsabilidades da equipe são ambíguas ou mal definidas.
- A equipe não alinha perfis a responsabilidades.
- A equipe não alinha metas e objetivos às responsabilidades, o que torna difícil medir o sucesso.
- As responsabilidades dos membros da equipe não se alinham à equipe nem à organização em geral.

- Sua equipe não mantém as responsabilidades atualizadas, o que as torna inconsistentes com as tarefas realizadas por ela.
- Os caminhos de escalação para determinar responsabilidades não estão definidos ou não estão claros.
- Os caminhos de escalação não têm um proprietário de thread único para garantir uma resposta oportuna.
- Os perfis, as responsabilidades e os caminhos de escalação não são detectáveis e não estão prontamente disponíveis quando necessário (por exemplo, em resposta a um incidente).

Benefícios de implementar esta prática recomendada:

- Quando você entende quem tem responsabilidade ou propriedade, pode entrar em contato com a equipe ou o membro adequado para fazer uma solicitação ou fazer a transição de uma tarefa.
- Para reduzir o risco de inação e necessidades não atendidas, você identificou uma pessoa que tem autoridade para atribuir responsabilidade ou propriedade.
- Quando você define claramente o escopo de uma responsabilidade, os membros da equipe ganham autonomia e propriedade.
- Suas responsabilidades fundamentam as decisões que você toma, as ações que você realiza e suas atividades de entrega aos proprietários apropriados.
- É fácil identificar responsabilidades abandonadas porque você tem uma compreensão clara do que está fora da responsabilidade de sua equipe, o que ajuda você a encaminhar os assuntos para ter esclarecimentos.
- As equipes evitam confusões e tensões e podem gerenciar de forma mais adequada as workloads e os recursos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Identifique perfis e responsabilidades dos membros da equipe e verifique se eles entendem as expectativas do perfil que exercem. Torne essas informações detectáveis para que os membros da organização possam identificar com quem precisam entrar em contato (equipe ou indivíduo) em relação a necessidades específicas. À medida que as organizações buscam aproveitar as oportunidades de migrar e modernizar na AWS, os perfis e responsabilidades também podem mudar.

Mantenha suas equipes e os membros cientes das responsabilidades e treine-os adequadamente para realizar as tarefas durante essa mudança.

Determine o perfil ou a equipe que deve receber as escalações para identificar a responsabilidade e a propriedade. Essa equipe pode interagir com várias partes interessadas para tomar uma decisão. No entanto, ela deve assumir o gerenciamento da tomada de decisões.

Forneça mecanismos acessíveis para que os membros da organização descubram e identifiquem propriedade e responsabilidade. Esses mecanismos os ensinam com quem entrar em contato em relação a necessidades específicas.

Exemplo de cliente

A AnyCompany Retail concluiu recentemente uma migração de workloads de um ambiente on-premises para sua zona de pouso na AWS com uma abordagem de mover sem alterações (lift-and-shift). Ela realizou uma revisão das operações para refletir sobre como realiza tarefas operacionais comuns e verificou se a matriz de responsabilidades existente reflete as operações no novo ambiente. Ao migrar do ambiente on-premises para a AWS, ela reduziu as responsabilidades das equipes de infraestrutura relacionadas a hardware e à infraestrutura física. Essa mudança também revelou novas oportunidades de desenvolver o modelo operacional das workloads.

Embora tenha identificado, tratado e documentado a maioria das responsabilidades, ela também definiu rotas de escalação para todas as responsabilidades não detectadas ou que possam precisar mudar à medida que as práticas operacionais evoluem. Para examinar novas oportunidades de padronizar e melhorar a eficiência nas workloads, forneça acesso a ferramentas operacionais, como o AWS Systems Manager, e a ferramentas de segurança, como o AWS Security Hub e o Amazon GuardDuty. A AnyCompany Retail realiza uma análise das responsabilidades e da estratégia com base nas melhorias que ela deseja abordar primeiro. À medida que a empresa adota novas formas de trabalhar e padrões tecnológicos, ela atualiza a matriz de responsabilidades para adequá-la.

Etapas de implementação

1. Comece com a documentação existente. Alguns exemplos de documentos de origem típicos:
   a. Matrizes de responsabilidades ou responsáveis, aprovador, consultado e informado (RACI)
   b. Definições de equipe ou páginas wiki
   c. Definições e ofertas de serviços
   d. Descrições de perfis ou cargos
2. Revise e organize discussões sobre as responsabilidades documentadas:

   a. Revise com as equipes para identificar desalinhamentos entre as responsabilidades documentadas e as responsabilidades que a equipe normalmente executa.
   b. Aborde os possíveis serviços oferecidos pelos clientes internos para identificar lacunas nas expectativas entre as equipes.
3. Analise e resolva as discrepâncias.
4. Identifique oportunidades de melhoria.
   a. Identifique solicitações feitas com frequência e que consomem muitos recursos, as quais normalmente são fortes candidatas a melhorias.
   b. Procure práticas recomendadas, padrões e recomendações e simplifique e padronize as melhorias com essas orientações.
   c. Registre oportunidades de melhoria e acompanhe-as até a conclusão.
5. Se uma equipe ainda não tiver a responsabilidade de gerenciar e rastrear a atribuição de responsabilidades, identifique alguém na equipe para assumir essa responsabilidade.
6. Defina um processo para que as equipes solicitem esclarecimentos sobre responsabilidades.
   a. Analise o processo e verifique se ele está claro e é simples de usar.
   b. Certifique-se de que alguém seja responsável pelas escalações e faça o rastreamento até a conclusão.
   c. Estabeleça métricas operacionais para medir a eficácia.
   d. Crie mecanismos de feedback para verificar se as equipes podem destacar oportunidades de melhoria.
   e. Implemente um mecanismo para revisão periódica.
7. Documente em um local detectável e acessível.
   a. Wikis ou portais de documentação são escolhas comuns.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS01-BP06 Avaliar compensações
- OPS03-BP02 Os membros da equipe são capacitados para executar ações quando os resultados estão em risco
- OPS03-BP03 Incentivo à escalação

- OPS03-BP07 Fornecer recursos adequados às equipes
- OPS09-BP01 Medir metas operacionais e KPIs com métricas
- OPS09-BP03 Revisar as métricas operacionais e priorizar a melhoria
- OPS11-BP01 Adotar um processo para melhoria contínua

Documentos relacionados:

- Whitepaper da AWS: Introdução a DevOps na AWS
- Whitepaper da AWS: Framework de adoção da Nuvem AWS: perspectiva de operações
- Excelência operacional do AWS Well-Architected Framework: topologias do modelo operacional em nível de workload
- Recomendações da AWS: Criar seu modelo operacional de nuvem
- Recomendações da AWS: Criar uma matriz RACI ou RASCI para um modelo operacional na nuvem
- Blog de operações e migrações da Nuvem AWS: Agregar valor comercial com equipes da Cloud Platform
- Blog de operações e migrações da Nuvem AWS: Por que um modelo operacional na nuvem?
- Blog de DevOps da AWS: Como as organizações estão se modernizando para as operações na nuvem

Vídeos relacionados:

- AWS Summit Online: Modelos operacionais em nuvem para transformação acelerada
- AWS re:Invent 2023: Segurança na nuvem preparada para o futuro: um novo modelo operacional

OPS02-BP05 Mecanismos para solicitar adições, alterações e exceções

É possível fazer solicitações aos proprietários de processos, procedimentos e recursos. As solicitações incluem adições, alterações e exceções. Essas solicitações passam por um processo de gerenciamento de alterações. Tome decisões embasadas para aprovar solicitações quando elas forem viáveis e consideradas apropriadas após uma avaliação de benefícios e riscos.

Resultado desejado:

- É possível fazer solicitações para alterar processos, procedimentos e recursos com base na propriedade atribuída.
- As alterações são feitas de maneira deliberada, ponderando benefícios e riscos.

Práticas comuns que devem ser evitadas:

- Você precisa atualizar a maneira como implanta sua aplicação, mas não há como solicitar uma alteração no processo de implantação à equipe de operações.
- O plano de recuperação de desastres deve ser atualizado, mas não há nenhum proprietário identificado para solicitar alterações no plano.

Benefícios de implementar esta prática recomendada:

- Os processos, procedimentos e recursos podem evoluir à medida que os requisitos mudam.
- Os proprietários podem tomar decisões embasadas sobre quando realizar alterações.
- As alterações são feitas de maneira deliberada.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Para implementar esta prática recomendada, você precisará estar em uma posição em que possa solicitar alterações em processos, procedimentos e recursos. O processo de gerenciamento de alterações pode ser simples. Documente o processo de gerenciamento de alterações.

Exemplo de cliente

A AnyCompany Retail usa a matriz de atribuição de responsabilidades (RACI) para identificar quem é proprietário das alterações em processos, procedimentos e recursos. A empresa conta com um processo documentado de gerenciamento de alterações que é simples e fácil de seguir. Usando a matriz RACI e o processo, qualquer pessoa pode enviar solicitações de alteração.

Etapas de implementação

1. Identifique processos, procedimentos e recursos para sua workload e os proprietários de cada um. Documente-os no sistema de gerenciamento de conhecimento.
   a. Se você não implementou OPS02-BP01 Recursos com proprietários identificados, OPS02-BP02 Processos e procedimentos com proprietários identificados ou OPS02-BP03 Atividades

de operações com proprietários identificados responsáveis pela performance, comece com eles primeiro.

2. Trabalhe com as partes interessadas em sua organização para desenvolver um processo de gerenciamento de alterações. O processo deve abranger adições, alterações e exceções para recursos, processos e procedimentos.
   a. O Gerenciador de Alterações do AWS Systems Manager pode ser usado como uma plataforma de gerenciamento de alterações para recursos de workload.
3. Documente o processo de gerenciamento de alterações em seu sistema de gerenciamento de conhecimento.

Nível de esforço do plano de implementação: Médio. O desenvolvimento de um processo de gerenciamento de alterações deve estar alinhado a várias partes interessadas em sua organização.

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP01 Recursos com proprietários identificados: os recursos precisam de proprietários identificados antes de você criar um processo de gerenciamento de alterações.
- OPS02-BP02 Processos e procedimentos com proprietários identificados: os processos precisam de proprietários identificados antes de você criar um processo de gerenciamento de alterações.
- OPS02-BP03 Atividades de operações com proprietários identificados responsáveis pela performance: as atividades operacionais precisam de proprietários identificados antes de você criar um processo de gerenciamento de alterações.

Documentos relacionados:

- Recomendações da AWS: Manual básico para grandes migrações da AWS: criação de matrizes RACI
- Whitepaper Gerenciamento de alterações na nuvem

Serviços relacionados:

- Gerenciador de Alterações do AWS Systems Manager

OPS02-BP06 As responsabilidades entre as equipes são predefinidas ou negociadas

Tenha acordos definidos ou negociados entre as equipes que descrevam como elas trabalham e apoiam umas às outras (por exemplo, tempos de resposta, objetivos de nível de serviço ou acordos de serviço). Os canais de comunicação entre equipes são documentados. Ao entender o impacto do trabalho das equipes nos resultados de negócios e nos resultados de outras equipes e organizações, você conhece a priorização de tarefas delas e as ajuda a responder adequadamente.

Quando a responsabilidade e a propriedade não foram definidas ou são desconhecidas, você corre o risco de não abordar as atividades necessárias em tempo hábil e de despender esforços redundantes e possivelmente conflitantes para atender a essas necessidades.

Resultado desejado:

- Os acordos de trabalho ou apoio entre equipes são combinados e documentados.
- As equipes que apoiam ou trabalham umas com as outras definiram canais de comunicação e expectativas de resposta.

Práticas comuns que devem ser evitadas:

- Um problema ocorre na produção e duas equipes separadas iniciam a resolução de problemas de maneira independente. Esses esforços isolados estendem a interrupção.
- A equipe de operações necessita de assistência da equipe de desenvolvimento, mas nenhum tempo de resposta foi acordado. A solicitação está parada em uma lista de pendências.

Benefícios de implementar esta prática recomendada:

- As equipes sabem interagir e apoiar uma à outra.
- As expectativas quanto à capacidade de resposta são claras.
- Os canais de comunicação estão nitidamente definidos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

A implementação desta prática recomendada significa que não há ambiguidade quanto à forma como as equipes trabalham uma com a outra. Os acordos formais sistematizam de que maneira as

equipes trabalham juntas ou apoiam uma à outra. Os canais de comunicação entre as equipes são documentados.

Exemplo de cliente

A equipe de SRE da AnyCompany Retail tem um acordo de serviço com a equipe de desenvolvimento. Sempre que a equipe de desenvolvimento faz uma solicitação no sistema de tíquetes, ela pode esperar uma resposta em 15 minutos. Se não houver nenhuma interrupção no local, a equipe de SRE toma a dianteira na investigação e conta com o apoio da equipe de desenvolvimento.

Etapas de implementação

1. Trabalhando com as partes interessadas na organização, desenvolva acordos entre as equipes com base nos processos e procedimentos.
   a. Se um processo ou procedimento for compartilhado entre as duas equipes, desenvolva um runbook sobre como as equipes trabalharão juntas.
   b. Se houver dependências entre as equipes, estabeleça um SLA de resposta às solicitações.
2. Documente as responsabilidades no sistema de gerenciamento de conhecimento.

Nível de esforço do plano de implementação: Médio. Se não houver nenhum entendimento entre as equipes, talvez seja difícil chegar a um acordo com as partes interessadas na organização.

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP02 Processos e procedimentos com proprietários identificados: os proprietários do processo devem ser identificados antes de estabelecer acordos entre as equipes.
- OPS02-BP03 Atividades de operações com proprietários identificados responsáveis pela performance: os proprietários das atividades operacionais devem ser identificados antes de estabelecer acordos entre as equipes.

Documentos relacionados:

- AWS Executive Insights: Fortalecer a inovação com a "equipe de duas pizzas"
- Introdução a DevOps na AWS: equipes de duas pizzas

OPS 3. Como a cultura organizacional oferece suporte aos resultados comerciais?

Forneça suporte aos membros da equipe para que eles possam ser mais eficazes na tomada de ações e no suporte aos resultados comerciais.

Práticas recomendadas

- OPS03-BP01 Fornecer patrocínio executivo
- OPS03-BP02 Os membros da equipe estão capacitados para executar ações quando os resultados estão em risco
- OPS03-BP03 Incentivo à escalação
- OPS03-BP04 Comunicações rápidas, claras e acionáveis
- OPS03-BP05 Incentivo à experimentação
- OPS03-BP06 Os membros da equipe são incentivados a manter e a aumentar seus conjuntos de habilidades.
- OPS03-BP07 Fornecer recursos adequados às equipes

OPS03-BP01 Fornecer patrocínio executivo

No nível mais alto, a liderança sênior atua como patrocinadora executiva para definir claramente as expectativas e a direção dos resultados da organização, inclusive avaliando o sucesso. O patrocinador defende e promove a adoção das práticas recomendadas e a evolução da organização.

Resultado desejado: organizações que se esforçam para adotar, transformar e otimizar suas operações na nuvem estabelecem linhas claras de liderança e responsabilidade pelos resultados desejados. A organização compreende cada capacidade exigida pela organização para alcançar um novo resultado e atribui a propriedade às equipes funcionais para desenvolvimento. A liderança define ativamente essa direção, atribui propriedade, assume responsabilidade e define o trabalho. Como resultado, as pessoas em toda a organização podem se mobilizar, sentir-se inspiradas e trabalhar ativamente em direção aos objetivos desejados.

Práticas comuns que devem ser evitadas:

- Há uma obrigação de os proprietários de workloads migrá-las para a AWS sem um patrocinador e um plano claros para as operações na nuvem. Isso faz com que as equipes não colaborem conscientemente para melhorar e amadurecer a capacidade operacional. A falta de padrões de práticas recomendadas operacionais sobrecarrega as equipes (por exemplo, esforço do operador, plantões e dívidas técnicas), o que restringe a inovação.

- Um novo objetivo foi estabelecido em toda a organização de adotar uma tecnologia emergente sem fornecer liderança, patrocinador e estratégia. As equipes interpretam os objetivos de forma diferente, o que causa confusão sobre onde concentrar os esforços, por que eles são importantes e como medir o impacto. Consequentemente, a organização perde o ímpeto na adoção da tecnologia.

Benefícios de implementar esta prática recomendada: quando o patrocínio executivo comunica e compartilha claramente a visão, a direção e as metas, os membros da equipe sabem o que se espera deles. Indivíduos e equipes começam a concentrar intensamente os esforços na mesma direção para concretizar os objetivos definidos quando os líderes estão ativamente engajados. Como resultado, a organização maximiza a capacidade de sucesso. Ao avaliar o sucesso, você pode identificar melhor as barreiras ao sucesso para que elas possam ser resolvidas por meio da intervenção do patrocinador executivo.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

- Em cada fase da jornada para a nuvem (migração, adoção ou otimização), o sucesso exige envolvimento ativo no mais alto nível de liderança com um patrocinador executivo designado. O patrocinador executivo alinha a mentalidade, as habilidades e as formas de trabalhar da equipe com a estratégia definida.
  - Explique o porquê: traga clareza e explique o raciocínio por trás da visão e da estratégia.
  - Defina expectativas: defina e publique metas para suas organizações, incluindo como o progresso e o sucesso são medidos.
  - Monitore o cumprimento das metas: meça o alcance incremental das metas regularmente (não apenas a conclusão das tarefas). Compartilhe os resultados para que as ações apropriadas possam ser tomadas se os resultados estiverem em risco.
  - Forneça os recursos necessários para atingir suas metas: reúna pessoas e equipes para colaborar e criar as soluções certas que tragam os resultados definidos. Isso reduz ou elimina o atrito organizacional.
  - Defenda suas equipes: permaneça engajado com suas equipes para entender a performance de cada uma e se há fatores externos que as afetam. Identifique os obstáculos que estão impedindo o progresso das equipes. Aja em nome das suas equipes para ajudar a resolver obstáculos e eliminar obrigações desnecessárias. Quando suas equipes forem afetadas por fatores externos, reavalie os objetivos e ajuste as metas conforme apropriado.

- Impulsione a adoção de práticas recomendadas: reconheça as práticas recomendadas que oferecem benefícios quantificáveis e reconheça quem as cria e adota. Incentive ainda mais a adoção para ampliar os benefícios obtidos.
- Incentive a evolução de suas equipes: crie uma cultura de melhoria contínua e aprenda proativamente com o progresso feito e com as falhas. Incentive o crescimento e o desenvolvimento pessoal e organizacional. Use dados e histórias para desenvolver a visão e a estratégia.

Exemplo de cliente

A AnyCompany Retail está em processo de transformação dos negócios por meio da rápida reinvenção das experiências do cliente, do aumento da produtividade e da aceleração do crescimento via IA generativa.

Etapas de implementação

1. Estabeleça uma liderança unidirecional e designe um patrocinador executivo principal para liderar e promover a transformação.
2. Defina resultados comerciais claros de sua transformação e atribua propriedade e responsabilidade. Capacite o executivo principal com a autoridade para liderar e tomar decisões essenciais.
3. Verifique se sua estratégia de transformação está bem clara e amplamente comunicada pelo patrocinador executivo a todos os níveis da organização.
   a. Estabeleça objetivos comerciais claramente definidos para iniciativas de TI e nuvem.
   b. Documente as principais métricas de negócios para promover a transformação de TI e da nuvem.
   c. Comunique a visão de forma consistente a todas as equipes e indivíduos responsáveis por partes da estratégia.
4. Desenvolva matrizes de planejamento de comunicação que especifiquem qual mensagem precisa ser entregue a líderes, gerentes e colaboradores individuais específicos. Especifique a pessoa ou a equipe que deve entregar essa mensagem.
   a. Cumpra os planos de comunicação de forma consistente e confiável.
   b. Defina e gerencie as expectativas por meio de eventos presenciais regularmente.
   c. Aceite feedback sobre a eficácia das comunicações, ajuste-as e planeje adequadamente.

   d. Agende eventos de comunicação para entender proativamente os desafios das equipes e estabeleça um ciclo de feedback consistente que permita corrigir o curso quando necessário.
5. Mobilize ativamente cada iniciativa, do ponto de vista de liderança, para verificar se todas as equipes afetadas entendem os resultados que são responsáveis por alcançar.
6. Em cada reunião de status, os patrocinadores executivos devem procurar barreiras, inspecionar métricas estabelecidas, histórias ou feedback das equipes e medir o progresso em direção aos objetivos.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS03-BP04 Comunicações oportunas, claras e acionáveis
- OPS11-BP01 Adotar um processo para a melhoria contínua
- OPS11-BP07 Revisar as métricas de operações

Documentos relacionados:

- Desembaraçar o novelo organizacional: alinhamento elevado
- A transformação viva: abordagem pragmática às alterações
- Como se tornar uma empresa pronta para o futuro
- Sete obstáculos que devem ser evitados ao criar um CCoE
- Navegação na nuvem: indicadores-chave de performance para o sucesso

Vídeos relacionados:

- AWS re:Invent 2023: Um guia de IA generativa para líderes: usando a história para moldar o futuro (SEG204)

Exemplos relacionados:

- Prosci: Papel e importância do patrocinador principal

OPS03-BP02 Os membros da equipe estão capacitados para executar ações quando os resultados estão em risco

O comportamento cultural de apropriação encorajado pela liderança faz com que qualquer funcionário se sinta apto a agir em nome de toda a empresa, bem além do escopo definido para sua função e responsabilidade. Os funcionários podem agir para identificar proativamente os riscos à medida que eles surgem e tomar as medidas apropriadas. Essa cultura permite que os funcionários tomem decisões de alto valor com consciência situacional.

Por exemplo, a Amazon usa Princípios de liderança como diretrizes para impulsionar o comportamento desejado dos funcionários para avançar em situações, resolver problemas, lidar com conflitos e agir.

Resultado desejado: a equipe de liderança influenciou uma nova cultura que permite que indivíduos e equipes tomem decisões críticas, mesmo em níveis mais baixos da organização (desde que as decisões sejam definidas com permissões auditáveis e mecanismos de segurança). O fracasso não é desencorajado, e as equipes aprendem iterativamente a melhorar a tomada de decisão e suas respostas para enfrentar situações semelhantes daquele ponto em diante. Se as ações de alguém ocasionarem uma melhoria que possa beneficiar outras equipes, essa pessoa compartilha proativamente o conhecimento dessas ações. A liderança mede as melhorias operacionais e incentiva o indivíduo e a organização a adotar esses padrões.

Práticas comuns que devem ser evitadas:

- Não há diretrizes ou mecanismos claros em uma organização sobre o que fazer quando um risco é identificado. Por exemplo, quando um funcionário percebe um ataque de phishing, ele não se comunica com a equipe de segurança, fazendo com que grande parte da organização caia no ataque. Isso causa uma violação de dados.
- Seus clientes reclamam da indisponibilidade do serviço, que se deve principalmente a falhas nas implantações. Sua equipe de SRE é responsável pela ferramenta de implantação, e uma reversão automática das implantações está no roteiro de longo prazo. No lançamento recente de uma aplicação, um dos engenheiros criou uma solução para automatizar a reversão da aplicação para uma versão anterior. Embora a solução dele possa se tornar o padrão para equipes de SRE, outras equipes não a adotam, pois não há processo para rastrear essas melhorias. A organização continua sofrendo com falhas nas implantações, afetando os clientes e causando ainda mais sentimentos negativos.
- Para manter a conformidade, sua equipe de segurança da informação supervisiona um processo estabelecido há muito tempo para trocar as chaves SSH compartilhadas regularmente em

nome dos operadores que se conectam às instâncias do Linux do Amazon EC2. As equipes de segurança da informação demoram vários dias para concluir a troca das chaves, e você não consegue se conectar a essas instâncias. Ninguém dentro ou fora da equipe de segurança da informação sugere usar outras opções na AWS para ter o mesmo resultado.

Benefícios de implementar esta prática recomendada: ao descentralizar a autoridade para tomar decisões e capacitar suas equipes a tomar decisões importantes, você pode resolver os problemas mais rapidamente com o aumento das taxas de sucesso. Além disso, as equipes começam a perceber um senso de propriedade e que as falhas são aceitáveis. A experimentação torna-se um pilar cultural. Gerentes e diretores não se sentem microgerenciados em todos os aspectos do trabalho.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

1. Desenvolva uma cultura que reconheça a possibilidade de falhas.
2. Defina propriedade e responsabilidade claras para várias áreas funcionais dentro da organização.
3. Comunique a propriedade e a responsabilidade a todos para que as pessoas saibam quem pode ajudá-las a facilitar as decisões descentralizadas.
4. Defina suas decisões unidirecionais e bidirecionais para ajudar as pessoas a saber quando precisam escalar para níveis mais altos de liderança.
5. Crie a consciência organizacional de que todos os funcionários têm autonomia para agir em vários níveis quando os resultados correm risco. Forneça aos membros da equipe documentação sobre governança, níveis de permissão, ferramentas e oportunidades para praticar as habilidades necessárias para reagir de forma eficaz.
6. Dê aos membros da equipe a oportunidade de praticar as habilidades necessárias para reagir a várias decisões. Depois que os níveis de decisão forem definidos, realize game days para verificar se todos os colaboradores individuais entendem e podem demonstrar o processo.
    a. Forneça ambientes seguros alternativos em que processos e procedimentos possam ser testados e treinados.
    b. Reconheça e crie consciência de que os membros da equipe têm autoridade para agir quando o resultado tem um nível de risco predefinido.
    c. Defina a autoridade dos membros da equipe para realizar ações por meio da atribuição de permissões e acesso às workloads e aos componentes aos quais eles dão suporte.

7. Ofereça às equipes a capacidade de compartilhar seus aprendizados (sucessos e fracassos operacionais).
8. Capacite as equipes para desafiar o status quo e forneça mecanismos para rastrear e medir as melhorias, bem como seu impacto na organização.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS01-BP06 Avaliar as compensações ao gerenciar benefícios e riscos
- OPS02-BP05 Mecanismos disponíveis para identificar a responsabilidade e a relação de propriedade

Documentos relacionados:

- Publicação do blog da AWS | A empresa ágil
- Publicação do blog da AWS | Medição do sucesso: um paradoxo e um plano
- Publicação do blog da AWS | Deixar para trás: possibilitando a autonomia nas equipes
- Centralizar ou descentralizar?

Vídeos relacionados:

- re:Invent 2023 | Como não sabotar sua transformação (SEG201)
- re:Invent 2021 | Amazon Builders' Library: 25 anos de excelência operacional da Amazon
- Centralização versus descentralização

Exemplos relacionados:

- Usar registros de decisão de arquitetura para agilizar a tomada de decisões técnicas para um projeto de desenvolvimento de software

OPS03-BP03 Incentivo à escalação

Os membros da equipe são incentivados pela equipe de liderança a escalar questões e preocupações para partes interessadas e tomadores de decisão de alto nível se acreditarem que os resultados desejados estão em risco e os padrões esperados não estão sendo atendidos. Essa é uma característica da cultura da organização e é encorajada em todos os níveis. A escalação deve ser realizada de maneira antecipada e frequente para que os riscos possam ser identificados e evitar incidentes. A liderança não repreende as pessoas por escalarem um problema.

Resultado desejado: indivíduos de toda a organização sentem-se confortáveis em escalar os problemas para seus níveis imediatos e mais altos de liderança. A liderança estabeleceu deliberada e conscientemente expectativas de que suas equipes devem se sentir seguras para encaminhar qualquer problema. Existe um mecanismo para encaminhar problemas em cada nível da organização. Quando os funcionários encaminham problemas ao gerente, eles decidem em conjunto o nível de impacto e se o problema deve ser encaminhado. Para iniciar uma escalação, os funcionários devem incluir um plano de trabalho recomendado para resolver o problema. Se a gerência direta não agir em tempo hábil, os funcionários são incentivados a levar as questões ao mais alto nível de liderança se tiverem certeza de que os riscos para a organização justificam a escalação.

Práticas comuns que devem ser evitadas:

- Os líderes executivos não fazem perguntas investigativas suficientes durante a reunião de status do programa de transformação na nuvem para descobrir onde os problemas e as barreiras se encontram. Somente boas notícias são apresentadas como status. A CIO deixou claro que só gosta de ouvir boas notícias, pois qualquer desafio abordado faz com que o CEO pense que o programa está falhando.
- Você é engenheiro de operações na nuvem e percebe que o novo sistema de gerenciamento de conhecimento não está sendo amplamente adotado pelas equipes de aplicações. A empresa investiu um ano e vários milhões de dólares para implementar esse novo sistema de gerenciamento de conhecimento, mas as pessoas ainda estão criando seus runbooks localmente e compartilhando-os em uma nuvem organizacional compartilhada, o que torna difícil encontrar conhecimentos pertinentes às workloads aceitas. Você tenta chamar a atenção da liderança para isso porque o uso consistente desse sistema pode aumentar a eficiência operacional. Quando você leva isso para a diretora que lidera a implementação do sistema de gerenciamento de conhecimentos, ela o repreende porque isso contesta o investimento.
- A equipe de segurança da informação responsável por fortalecer os recursos de computação decidiu implementar um processo que exige a execução das verificações necessárias a fim

de garantir que as instâncias do EC2 estejam totalmente protegidas antes que a equipe de computação libere o recurso para uso. Isso criou um atraso de mais uma semana para que os recursos fossem implantados, o que viola o SLA dela. A equipe de computação tem medo de escalar a questão para o vice-presidente de nuvem, pois prejudica a imagem do vice-presidente de segurança da informação.

Benefícios de implementar esta prática recomendada:

Problemas complexos ou críticos são resolvidos antes que afetem os negócios. Menos tempo é desperdiçado. Os riscos são minimizados. As equipes tornam-se mais proativas e focadas nos resultados ao resolver problemas.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

A disposição e a capacidade de encaminhar problemas livremente em todos os níveis da organização são uma base organizacional e cultural que deve ser desenvolvida conscientemente por meio de treinamento enfatizado, comunicação de liderança, definição de expectativas e implantação de mecanismos em toda a organização em todos os níveis.

Etapas de implementação

1. Defina políticas, padrões e expectativas para sua organização.
    a. Garanta ampla adoção e compreensão de políticas, expectativas e padrões.
2. Incentive, treine e capacite os trabalhadores para escalação precoce e frequente quando os padrões não forem atendidos.
3. Reconheça de maneira organizacional que a escalação antecipada e frequente é a prática recomendada. Aceite que as escalações podem ser infundadas e que é melhor ter a chance de evitar um incidente do que perder essa oportunidade ao não encaminhar.
    a. Crie um mecanismo de escalação (como um sistema de cabos Andon).
    b. Mantenha procedimentos documentados que definam quando e como a escalação deve ocorrer.
    c. Defina o grupo de pessoas com autoridade crescente para tomar ou aprovar ações, bem como as informações de contato de cada parte interessada.
4. Quando a escalação ocorre, ela deve continuar até que o membro da equipe esteja convencido de que o risco foi mitigado por meio de ações orientadas pela liderança.

a. As escalações devem incluir:
    i. Descrição da situação e natureza do risco
    ii. Criticidade da situação
    iii. Quem ou o que é afetado
    iv. A dimensão do impacto
    v. A urgência em caso de impacto
    vi. Soluções sugeridas e planos de mitigação
b. Proteja os funcionários que escalam problemas. Adote uma política que proteja os membros da equipe contra retaliações se eles fizerem uma escalação a respeito de um responsável pela tomada de decisões ou uma parte interessada não responsiva. Tenha mecanismos implementados para identificar se isso está ocorrendo e que permita reagir da maneira adequada.

5. Incentive uma cultura de ciclos de feedback de melhoria contínua em tudo o que a organização produz. Os ciclos de feedback funcionam como pequenas escalações para os indivíduos responsáveis e identificam oportunidades de melhoria, mesmo quando a escalação não é necessária. As culturas de melhoria contínua instigam todos a serem mais proativos.
6. A liderança deve enfatizar periodicamente as políticas, os padrões, os mecanismos e o intuito de permitir escalação aberta e ciclos de feedback contínuos sem retribuição.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP05 Mecanismos para solicitar adições, alterações e exceções

Documentos relacionados:

- Como você promove uma cultura de melhoria e aprendizado contínios com o Andon e os sistemas de escalação?
- O Andon Cord (revolução da TI)
- Orientação de DevOps da AWS | Estabelecer caminhos claros de escalação e incentivar discordâncias construtivas

Vídeos relacionados:

- Jeff Bezos fala sobre como tomar decisões (e aumentar a velocidade)
- Sistema de produtos Toyota: interrupção da produção, um botão e um quadro elétrico Andon
- Andon Cord na manufatura LEAN

Exemplos relacionados:

- Como trabalhar com planos de escalação no Incident Manager

OPS03-BP04 Comunicações rápidas, claras e acionáveis

A liderança é responsável pela criação de comunicações fortes e eficazes, especialmente quando a organização adota novas estratégias, tecnologias ou formas de trabalhar. Os líderes precisam estabelecer expectativas para que todos os funcionários trabalhem de acordo com os objetivos da empresa. Desenvolva mecanismos de comunicação que criem e mantenham a conscientização entre as equipes responsáveis pela execução de planos financiados e patrocinados pela liderança. Faça uso da diversidade interorganizacional e ouça atentamente vários pontos de vista exclusivos. Use essa abordagem para aumentar a inovação, desafiar suas suposições e reduzir o risco de viés de confirmação. Promova a inclusão, a diversidade e a acessibilidade em suas equipes para ter pontos de vista benéficos.

Resultado desejado: sua organização cria estratégias de comunicação para lidar com o impacto da mudança na organização. As equipes permanecem informadas e motivadas para continuar trabalhando umas com as outras, e não umas contra as outras. Os indivíduos entendem a importância de seu papel para concretizar os objetivos declarados. O e-mail é apenas um mecanismo passivo de comunicação e é usado adequadamente. A gerência passa tempo com seus colaboradores individuais para ajudá-los a entender suas responsabilidades, as tarefas a serem concluídas e como o trabalho contribui para a missão geral. Quando necessário, os líderes engajam as pessoas diretamente em locais menores para transmitir mensagens e verificar se essas mensagens estão sendo entregues de forma eficaz. Como resultado de boas estratégias de comunicação, a organização tem uma performance igual ou superior às expectativas da liderança. A liderança incentiva e busca opiniões diversas dentro e entre as equipes.

Práticas comuns que devem ser evitadas:

- Sua organização tem um plano de cinco anos para migrar todas as workloads para a AWS. O caso comercial da nuvem inclui a modernização de 25% de todas as workloads para aproveitar

a tecnologia sem servidor. O diretor executivo de informação comunica essa estratégia aos subordinados diretos e espera que cada líder transmita essa apresentação em cascata para gerentes, diretores e colaboradores individuais sem nenhuma comunicação pessoal. O diretor executivo de informação recua e espera que a organização execute a nova estratégia.

- A liderança não fornece nem usa um mecanismo de feedback, aumentando a lacuna de expectativas, o que causa a paralisação dos projetos.
- Você deve fazer uma alteração em seus grupos de segurança, mas não recebe detalhes sobre qual alteração precisa ser feita, qual pode ser o impacto da mudança em todas as workloads e quando ela deve ocorrer. O gerente encaminha um e-mail do vice-presidente de segurança da informação e adiciona a mensagem "Faça isso acontecer".
- Foram feitas alterações em sua estratégia de migração que reduziram a taxa de modernização planejada de 25% para 10%. Isso tem efeitos subsequentes na organização de operações. A organização não foi informada dessa mudança estratégica e, portanto, não está preparada com capacidade qualificada suficiente para comportar um número maior de workloads movidas sem alterações (lift-and-shift) para a AWS.

Benefícios de implementar esta prática recomendada:

- Sua organização está bem informada sobre estratégias novas ou alteradas e age adequadamente, com forte motivação para ajudar umas às outras a alcançar os objetivos gerais e as métricas definidas pela liderança.
- Mecanismos existem e são usados para fornecer avisos oportunos aos membros da equipe sobre riscos conhecidos e eventos planejados.
- Novas formas de trabalhar (incluindo mudanças nas pessoas ou na organização, nos processos ou na tecnologia), além das habilidades necessárias, são adotadas de forma mais eficaz pela organização, que recebe os benefícios comerciais mais rapidamente.
- Os membros da equipe têm o contexto necessário para que as comunicações sejam recebidas e podem ser mais eficazes no trabalho.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Para implementar essa prática recomendada, você precisa trabalhar com as partes interessadas na organização para ajustar padrões de comunicação. Divulgue esses padrões para toda a organização. Para qualquer transição significativa de TI, uma equipe de planejamento estabelecida

pode gerenciar melhor o impacto da mudança no pessoal do que uma organização que ignora essa prática. Para organizações maiores, o gerenciamento de mudanças pode ser mais desafiador, pois é fundamental estabelecer uma forte adesão a uma nova estratégia junto a todos os colaboradores individuais. Na ausência dessa equipe de planejamento de transição, a liderança é totalmente responsável pela comunicação eficaz. Ao estabelecer uma equipe de planejamento de transição, designe membros da equipe para trabalhar com toda a liderança organizacional a fim de definir e gerenciar a comunicação eficaz em todos os níveis.

Exemplo de cliente

A AnyCompany Retail se inscreveu no Enterprise Support da AWS e depende de outros fornecedores terceirizados para suas operações na nuvem. A empresa usa chat e chatops como principal meio de comunicação para atividades operacionais. Alertas e outras informações são divulgados em canais específicos. Quando alguém precisa agir, essa pessoa expressa claramente o resultado desejado, e, em muitos casos, recebe um runbook ou playbook para uso nessas situações. As pessoas programam alterações importantes em sistemas de produção com um calendário de alterações.

Etapas de implementação

1. Estabeleça uma equipe central na organização que tenha a responsabilidade de criar e iniciar planos de comunicação para mudanças que ocorrem em vários níveis dentro da organização.
2. Institua a propriedade de um único segmento para obter supervisão. Dê às equipes individuais a capacidade de inovar de forma independente e contrabalance o uso de mecanismos consistentes, o que permite o nível certo de inspeção e visão direcional.
3. Trabalhe com as partes interessadas em toda a organização para chegar a um acordo acerca de padrões, práticas e planos de comunicação.
4. Verifique se a equipe principal de comunicação colabora com a liderança organizacional e do programa para criar mensagens para a equipe apropriada em nome dos líderes.
5. Crie mecanismos estratégicos de comunicação para gerenciar mudanças por meio de anúncios, calendários compartilhados, reuniões gerais e métodos presenciais ou individuais, para que os membros da equipe tenham expectativas adequadas sobre as ações que devem realizar.
6. Forneça o contexto, os detalhes e o tempo necessários (quando possível) para determinar se a ação é necessária. Quando uma ação for necessária, comunique-a junto com seu impacto.
7. Implemente ferramentas que facilitem a comunicação tática, como chat interno, e-mail e gerenciamento de conhecimentos.

8. Implemente mecanismos para medir e verificar se todas as comunicações geram os resultados desejados.
9. Estabeleça um ciclo de feedback que meça a eficácia de todas as comunicações, especialmente quando elas estão relacionadas à resistência às mudanças em toda a organização.
10. Para todas as Contas da AWS, estabeleça contatos alternativos para faturamento, segurança e operações. Idealmente, cada contato deve ser uma distribuição de e-mail em vez de um contato individual específico.
11. Estabeleça um plano de comunicação de escalação (inclusive escalação reversa) para interagir com as equipes internas e externas, incluindo suporte da AWS e outros fornecedores terceirizados.
12. Inicie e execute estratégias de comunicação de forma consistente ao longo da vida de cada programa de transformação.
13. Priorize ações que possam ser repetidas sempre que possível para automatizar com segurança em grande escala.
14. Quando a comunicação é necessária em cenários com ações automatizadas, o objetivo da comunicação deve ser informar as equipes para auditoria ou fazer parte do processo de gerenciamento de mudanças.
15. Analise as comunicações de seus sistemas de alerta em busca de falsos positivos ou alertas que são criados constantemente. Remova ou altere esses alertas para que eles sejam acionados quando há necessidade de intervenção humana. Se um alerta for acionado, forneça um runbook ou um playbook.
    a. Você pode usar os Documentos do AWS Systems Manager para criar playbooks e runbooks para alertas.
16. Mecanismos estão em vigor para fornecer notificações de riscos ou eventos planejados de maneira clara e acionável com aviso prévio em tempo suficiente para permitir respostas apropriadas. Use listas de e-mails ou canais por chat para enviar notificações antes dos eventos planejados.
    a. O AWS Chatbot pode ser usado para enviar alertas e responder a eventos na plataforma de mensagens da sua organização.
17. Forneça uma fonte de informações acessível em que eventos planejados possam ser descobertos. Forneça notificações de eventos planejados provenientes do mesmo sistema.
    a. O Calendário de Alterações do AWS Systems Manager pode ser usado para criar janelas de alteração quando mudanças podem ocorrer. Isso oferece aos membros da equipe um aviso prévio sobre quando eles podem fazer alterações com segurança.

18. Monitore notificações de vulnerabilidade e informações de patches para identificar vulnerabilidades nos riscos reais e potenciais associados aos componentes da workload. Forneça uma notificação aos membros da equipe para que eles possam agir.
    a. Você pode assinar os Boletins de Segurança da AWS para receber notificações de vulnerabilidades na AWS.
19. Busque opiniões e perspectivas diversas: incentive as contribuições de todos. Ofereça oportunidades de comunicação a grupos sub-representados. Alterne as funções e responsabilidades nas reuniões.
    a. Expanda as funções e responsabilidades: ofereça oportunidade para que os membros da equipe assumam funções que não poderiam assumir de outra forma. Eles poderão ganhar experiência e perspectiva com a função e com as interações com novos membros da equipe com os quais não interagiriam de outra forma. Eles levarão a experiência e o ponto de vista deles para a nova função e para os membros da equipe com os quais interagirem. À medida que a perspectiva aumenta, identifique oportunidades de negócios emergentes ou novas oportunidades de melhoria. Reveze tarefas comuns entre os membros de uma equipe que outras pessoas normalmente realizam para compreender as demandas e o impacto de realizá-las.
    b. Forneça um ambiente seguro e acolhedor: estabeleça políticas e controles que protejam a segurança física e mental dos membros da equipe em sua organização. Os membros da equipe devem poder interagir sem medo de sofrer represálias. Quando eles se sentem seguros e bem-vindos, as chances de se envolverem e serem produtivos também aumenta. Quanto mais diversificada sua organização, melhor será o entendimento das pessoas que você apoia, incluindo seus clientes. Quando os membros da equipe estiverem confortáveis, sentirem-se à vontade para falar e tiverem confiança de que serão ouvidos, será mais provável que compartilhem ideias valiosas (por exemplo, oportunidades de marketing, necessidades de acessibilidade, segmentos de mercado não atendidos, riscos não reconhecidos no seu ambiente).
    c. Estimule os membros da equipe a participar: forneça os recursos necessários para que seus funcionários participem totalmente de todas as atividades relacionadas ao trabalho. Os membros da equipe que enfrentam desafios diários desenvolvem habilidades para contorná-los. Essas habilidades desenvolvidas exclusivamente podem oferecer benefícios significativos para a sua organização. Apoie os membros da equipe com as acomodações necessárias para aumentar os benefícios que você poderá receber das contribuições de cada um.

Recursos

Práticas recomendadas relacionadas:

- OPS03-BP01 Fornecer patrocínio executivo
- OPS07-BP03 Usar runbooks para realizar procedimentos
- OPS07-BP04 Usar playbooks para investigar problemas

Documentos relacionados:

- Publicação no blog da AWS | Responsabilidade e capacitação são fundamentais para organizações ágeis de alta performance
- AWS Executive Insights | Aprenda a escalar a inovação, não a complexidade | Líderes de segmento único
- Boletins de segurança da AWS
- Open CVE
- Aplicação AWS Support no Slack para gerenciar casos de suporte
- Gerenciar recursos da AWS em seus canais do Slack com o AWS Chatbot

Exemplos relacionados:

- Laboratórios do Well-Architected: Gerenciamento de inventário e patches (Nível 100)

Serviços relacionados:

- AWS Chatbot
- Calendário de Alterações do AWS Systems Manager
- Documentos do AWS Systems Manager

OPS03-BP05 Incentivo à experimentação

A experimentação é um catalisador para transformar novas ideias em produtos e recursos. Ela acelera o aprendizado e mantém os membros da equipe interessados e envolvidos. Os membros da equipe são incentivados a experimentar com frequência para promover a inovação. Mesmo quando um resultado indesejado ocorre, é importante saber o que não se deve fazer. Os membros da equipe não são punidos por experimentos bem-sucedidos com resultados indesejados.

Resultado desejado:

- Sua organização incentiva a experimentação para promover a inovação.
- Os experimentos são usados como oportunidade de aprendizado.

Práticas comuns que devem ser evitadas:

- Você deseja executar um teste A/B, mas não há nenhum mecanismo para conduzir o experimento. Você implanta uma alteração de interface do usuário sem a possibilidade de testá-la. O resultado é uma experiência negativa para o cliente.
- Sua empresa tem apenas o ambiente de preparação e produção. Como não há ambiente de sandbox para experimentar novos recursos ou produtos, os experimentos devem ser realizados no ambiente de produção.

Benefícios de implementar esta prática recomendada:

- A experimentação promove a inovação.
- É possível reagir mais depressa ao feedback dos usuários por meio da experimentação.
- Sua organização desenvolve uma cultura de aprendizado.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Os experimentos devem ser conduzidos de maneira segura. Utilize vários ambientes para experimentar, sem colocar em risco os recursos da produção. Use testes A/B e sinalizadores de recursos para testar experimentos. Ofereça aos membros da equipe a possibilidade de conduzir experimentos em um ambiente de sandbox.

Exemplo de cliente

A AnyCompany Retail estimula a experimentação. Os membros da equipe podem usar 20% da semana de trabalho para experimentar ou aprender novas tecnologias. Eles têm um ambiente de sandbox no qual podem inovar. Testes A/B são usados para novos recursos com o objetivo de validá-los com um feedback de usuário real.

Etapas de implementação

1. Trabalhe com a liderança em toda a sua organização para favorecer a experimentação. Os membros da equipe devem ser incentivados a conduzir experimentos de maneira segura.
2. Ofereça aos membros da equipe um ambiente em que eles possa experimentar com segurança. Eles devem ter acesso a um ambiente semelhante ao de produção.
   a. Você pode usar uma Conta da AWS separada para criar um ambiente de sandbox para experimentação. O AWS Control Towerpode ser usado para provisionar essas contas.
3. Use sinalizadores de recursos e testes A/B para experimentar com segurança e coletar feedback dos usuários.
   a. O AWS AppConfig Feature Flags permite criar sinalizadores de recursos.
   b. Amazon CloudWatch Evidently pode ser usado para executar testes A/B em uma implantação limitada.
   c. Você pode usar versões do AWS Lambda para implantar uma nova versão de uma função para testes beta.

Nível de esforço do plano de implementação: Alto. A viabilização de um ambiente para experimentação de maneira segura para os membros da equipe conduzirem experimentos pode exigir um investimento significativo. Você também pode precisar modificar o código da aplicação para usar sinalizadores de recursos ou respaldar testes A/B.

Recursos

Práticas recomendadas relacionadas:

- OPS11-BP02 Executar análise pós-incidente: aprender com os incidentes é um fator importante para a inovação junto com a experimentação.
- OPS11-BP03 Implementar loops de feedback: os ciclos de feedback são uma parte importante da experimentação.

Documentos relacionados:

- Um olhar interno sobre a cultura da Amazon: experimentação, fracasso e obsessão pelo cliente
- Práticas recomendadas para criar e gerenciar contas de sandbox na AWS
- Criar uma cultura de experimentação viabilizada pela nuvem
- Possibilitar a experimentação e a inovação na nuvem na SulAmérica Seguros
- Experimentar mais, falhar menos

- Organizar seu ambiente da AWS usando várias contas: UO de sandbox
- Usar AWS AppConfig Feature Flags

Vídeos relacionados:

- Destaque da AWS On Air: Amazon CloudWatch Evidently | AWS Eventos
- Destaque da AWS On Air San Fran Summit 2022: Integração do AWS AppConfig Feature Flags ao Jira
- AWS re:Invent 2022: Uma implantação não é uma versão: controle seus lançamentos com sinalizadores de recursos (BOA305-R)
- Criar programaticamente uma Conta da AWS com o AWS Control Tower
- Configurar um ambiente da AWS com várias contas que use práticas recomendadas para o AWS Organizations

Exemplos relacionados:

- AWS Innovation Sandbox
- Princípio básico de personalização ponta a ponta para comércio eletrônico

Serviços relacionados:

- Amazon CloudWatch Evidently
- AWS AppConfig
- AWS Control Tower

OPS03-BP06 Os membros da equipe são incentivados a manter e a aumentar seus conjuntos de habilidades.

As equipes devem aumentar os conjuntos de habilidades para adotar novas tecnologias e apoiar mudanças na demanda e responsabilidades no apoio às suas workloads. O desenvolvimento das habilidades em novas tecnologias costuma ser uma fonte de satisfação dos membros da equipe e contribui para a inovação. Ofereça apoio aos membros da equipe na busca e atualização de certificações do setor que validem e reconheçam as suas habilidades crescentes. Treine profissionais em diferentes funções para promover a transferência de conhecimento e reduzir o risco

de impacto significativo quando você perde membros da equipe qualificados e experientes com conhecimento institucional. Reserve tempo estruturado e dedicado para o aprendizado.

A AWS fornece recursos, incluindo o Centro de recursos de conceitos básicos da AWS, Blogs da AWS, AWS Online Tech Talks, Eventos e webinars da AWS e os Laboratórios do AWS Well-Architected que oferecem orientação, exemplos e demonstrações detalhadas para ajudar a treinar suas equipes.

Recursos como o AWS Support, (AWS re:Post, AWS Support Center) e documentação da AWS ajudam a remover obstáculos técnicos e melhorar as operações. Entre em contato com o AWS Support por meio do AWS Support Center para obter ajuda para suas dúvidas.

A AWS também compartilha práticas recomendadas e padrões que aprendemos com a operação da AWS na Amazon Builders' Library e uma grande variedade de outros materiais educacionais úteis por meio do Blog da AWS e do Podcast oficial da AWS.

A Treinamento da AWS and Certification inclui treinamento gratuito por meio de cursos digitais individualizados, além de planos de aprendizado por função ou domínio. Você também pode se inscrever em treinamento administrado por instrutor a fim de oferecer suporte adicional às suas equipes para o desenvolvimento de habilidades em serviços da AWS.

Resultado desejado: sua organização avalia constantemente as lacunas de habilidades e as preenche com orçamento e investimento estruturados. As equipes incentivam os membros com atividades de aprimoramento, como a aquisição de certificações importantes do setor. As equipes aproveitam programas dedicados de compartilhamento cruzado de conhecimentos, como eventos de almoço, dias de imersão, hackathons e game days. Sua organização mantém os sistemas de conhecimento atualizados e relevantes para treinar os membros da equipe, incluindo treinamentos de integração para novos contratados.

Práticas comuns que devem ser evitadas:

- Na ausência de um programa de treinamento e orçamento estruturados, as equipes enfrentam incertezas ao tentar acompanhar a evolução da tecnologia, o que causa maior desgaste.
- Como parte da migração para a AWS, sua organização demonstra lacunas de habilidades e fluência variável na nuvem entre as equipes. Sem um esforço para aprimorar as habilidades, as equipes se veem sobrecarregadas com o gerenciamento herdado e ineficiente do ambiente de nuvem, o que causa maior esforço do operador. Esse esgotamento aumenta a insatisfação dos funcionários.

Benefícios de implementar esta prática recomendada: quando sua organização investe conscientemente no aprimoramento das habilidades de suas equipes, ela também ajuda a acelerar e escalar a adoção e a otimização da nuvem. Programas de aprendizado direcionados promovem a inovação e desenvolvem a capacidade operacional para que as equipes se preparem para lidar com eventos. As equipes investem conscientemente na implementação e na evolução das práticas recomendadas. O moral da equipe é alto e os membros valorizam a contribuição deles para os negócios.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Para adotar novas tecnologias, estimular a inovação e acompanhar as mudanças na demanda e nas responsabilidades para apoiar suas workloads, invista continuamente no crescimento profissional de suas equipes.

Etapas de implementação

1. Use programas estruturados de defesa da nuvem: o AWS Skills Guild oferece treinamento consultivo para aumentar a confiança nas habilidades de nuvem e estimular a cultura de aprendizado contínuo.
2. Forneça recursos didáticos: ofereça tempo estruturado dedicado, acesso a materiais de treinamento, recursos de laboratório e apoio à participação em conferências e organizações profissionais que forneçam oportunidades de aprendizado com educadores e colegas. Forneça aos membros da sua equipe júnior acesso aos membros seniores da equipe como mentores ou permita que os membros da equipe júnior acompanhem o trabalho dos seniores e sejam expostos a seus métodos e habilidades. Incentive o aprendizado sobre conteúdo não diretamente relacionado ao trabalho para ter uma perspectiva mais ampla.
3. Incentive o uso de recursos técnicos especializados: aproveite recursos como o AWS re:POST para ter acesso a conhecimento selecionado e a uma comunidade vibrante.
4. Crie e mantenha um repositório de conhecimento atualizado: use plataformas de compartilhamento de conhecimento, como wikis e runbooks. Crie sua própria fonte de conhecimento especializado reutilizável com o AWS re:Post Private para otimizar a colaboração, melhorar a produtividade e acelerar a integração de funcionários.
5. Aprendizado em equipe e engajamento entre equipes: planeje as necessidades de aprendizado contínuo dos membros da sua equipe. Ofereça oportunidades para que os membros da equipe se juntem a outras equipes (temporária ou permanentemente) para compartilhar habilidades e práticas recomendadas que beneficiam toda a organização.

6. Ofereça suporte à busca e à manutenção de certificações do setor: forneça apoio aos membros da equipe que conquistam e mantêm certificações do setor que validam o que aprenderam e reconheça as conquistas deles.

Nível de esforço do plano de implementação: Alto

Recursos

Práticas recomendadas relacionadas:

- OPS03-BP01 Fornecer patrocínio executivo
- OPS11-BP04 Gerenciar o conhecimento

Documentos relacionados:

- Whitepaper da AWS | Framework de adoção da nuvem: perspectiva das pessoas
- Investir em aprendizado contínuo para expandir o futuro da sua organização
- AWS Skills Guild
- Treinamento da AWS e certificação
- AWS Support
- AWS re:Post
- Centro de recursos de conceitos básicos da AWS
- Blogs da AWS
- Conformidade da Nuvem AWS
- Documentação do AWS
- Podcast oficial da AWS
- AWS Online Tech Talks
- Eventos e webinars da AWS
- Laboratórios do AWS Well-Architected
- Amazon Builders' Library

Vídeos relacionados:

- AWS re:Invent 2023 | Requalificar na velocidade da nuvem: transformar funcionários em empreendedores
- AWS re:Invent 2023 | Construir uma cultura de curiosidade por meio da gamificação

OPS03-BP07 Fornecer recursos adequados às equipes

Forneça a quantidade certa de membros proficientes da equipe, além de ferramentas e recursos para atender às necessidades da workload. Sobrecarregar os membros da equipe aumenta o risco de erro humano. Investimentos em ferramentas e recursos, como automação, podem aumentar a eficácia da equipe e ajudá-la a comportar um número maior de workloads sem exigir capacidade adicional.

Resultado desejado:

- Você contratou adequadamente sua equipe para ter as habilidades necessárias para operar workloads na AWS de acordo com o plano de migração. À medida que sua equipe aumentou ao longo do projeto de migração, ela adquiriu proficiência nas principais tecnologias da AWS que a empresa planeja usar ao migrar ou modernizar as aplicações.
- Você alinhou cuidadosamente seu plano de pessoal para fazer uso eficiente dos recursos, utilizando automação e fluxos de trabalho. Agora, uma equipe menor pode gerenciar uma maior infraestrutura em nome das equipes de desenvolvimento de aplicações.
- Com a mudança das prioridades operacionais, quaisquer restrições de pessoal são identificadas proativamente para proteger o sucesso das iniciativas de negócios.
- As métricas operacionais que relatam o esforço operacional (como fadiga de plantão ou chamadas em excesso) são revisadas para verificar se a equipe não está sobrecarregada.

Práticas comuns que devem ser evitadas:

- Sua equipe não aprimorou as habilidades da AWS ao concluir o plano plurianual de migração para a nuvem, o que arrisca o suporte das workloads e reduz o moral dos funcionários.
- Toda a sua organização de TI está adotando formas ágeis de trabalhar. A empresa está priorizando o portfólio de produtos e definindo métricas para quais recursos precisam ser desenvolvidos primeiro. Seu processo ágil não exige que as equipes atribuam “story points” aos planos de trabalho. Como resultado, é impossível saber o nível de capacidade necessário para a próxima quantidade de trabalho ou se você tem as habilidades certas atribuídas a ele.

- Um parceiro da AWS está migrando suas workloads e você não tem um plano de transição de suporte para suas equipes depois que o parceiro concluir o projeto de migração. Suas equipes têm dificuldade para oferecer suporte às workloads de forma eficiente.

Benefícios de implementar esta prática recomendada: você tem membros da equipe devidamente qualificados disponíveis em sua organização para acomodar as workloads. A alocação de recursos pode se adaptar às mudanças de prioridades sem afetar a performance. O resultado é que as equipes são proficientes em oferecer suporte às workloads e, ao mesmo tempo, maximizar o tempo de foco na inovação para os clientes, o que, por sua vez, aumenta a satisfação dos funcionários.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

O planejamento de recursos para sua migração para a nuvem deve ocorrer em um nível organizacional alinhado ao plano de migração, bem como ao modelo operacional desejado que está sendo implementado para comportar o novo ambiente de nuvem. Isso deve incluir a compreensão de quais tecnologias de nuvem são implantadas para as equipes de desenvolvimento de aplicações e negócios. A liderança em infraestrutura e operações deve planejar a análise de lacunas de habilidades, o treinamento e a definição de funções para engenheiros que lideram a adoção da nuvem.

Etapas de implementação

1. Defina critérios para o sucesso da equipe com métricas operacionais relevantes, como produtividade da equipe (por exemplo, custo para comportar uma workload ou horas gastas pelo operador durante incidentes).
2. Defina mecanismos de planejamento e inspeção da capacidade de recursos para verificar se o equilíbrio certo de capacidade qualificada está disponível quando necessário e pode ser ajustado ao longo do tempo.
3. Crie mecanismos (por exemplo, enviar uma pesquisa mensal às equipes) para entender os desafios relacionados ao trabalho que afetam as equipes (como aumento de responsabilidades, mudanças na tecnologia, perda de pessoal ou aumento de clientes atendidos).
4. Use esses mecanismos para interagir com equipes e identificar tendências que possam contribuir para os desafios de produtividade dos funcionários. Quando suas equipes forem afetadas por fatores externos, reavalie os objetivos e ajuste as metas conforme apropriado. Identifique os obstáculos que estão impedindo o progresso das equipes.

5. Analise regularmente se os recursos atualmente provisionados ainda são suficientes e se são necessários recursos adicionais e, em seguida, faça os ajustes apropriados nas equipes de suporte.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS03-BP06 Os membros da equipe são incentivados a manter e a aumentar seus conjuntos de habilidades
- OPS09-BP03 Revisar as métricas operacionais e priorizar a melhoria
- OPS10-BP01 Usar um processo para gerenciamento de eventos, incidentes e problemas
- OPS10-BP07 Automatizar respostas a eventos

Documentos relacionados:

- Nuvem AWS Adoption Framework: perspectiva das pessoas
- Como se tornar uma empresa pronta para o futuro
- Priorizar as capacidades dos funcionários para impulsionar o crescimento dos negócios
- Organização de alta performance: a equipe de duas pizzas da Amazon
- Como empresas maduras na nuvem são bem-sucedidas

Preparar

Perguntas

- OPS 4. Como implementar a observabilidade em sua workload?
- OPS 5. Como reduzir defeitos, facilitar a correção e melhorar o fluxo na produção?
- OPS 6. Como reduzir os riscos de implantação?
- OPS 7. Como saber se está tudo pronto para oferecer suporte a uma workload?

OPS 4. Como implementar a observabilidade em sua workload?

Implemente a observabilidade na workload para poder entender seu estado e tomar decisões baseadas em dados com base nos requisitos de negócios.

Práticas recomendadas

- OPS04-BP01 Identificar indicadores-chave de performance
- OPS04-BP02 Implementar a telemetria de aplicações
- OPS04-BP03 Implementar telemetria da experiência do usuário
- OPS04-BP04 Implementar a telemetria de dependências
- OPS04-BP05 Implementar rastreamento distribuído

OPS04-BP01 Identificar indicadores-chave de performance

A implementação da observabilidade em sua workload começa com a compreensão de seu estado e a tomada de decisões baseadas em dados de acordo com os requisitos de negócios. Uma das formas mais eficazes de garantir o alinhamento entre as atividades de monitoramento e os objetivos de negócios é definir e monitorar os indicadores-chave de performance (KPIs).

Resultado desejado: práticas de observabilidade eficientes que estão estreitamente alinhadas aos objetivos de negócios, garantindo que os esforços de monitoramento estejam sempre a serviço de resultados comerciais tangíveis.

Práticas comuns que devem ser evitadas:

- KPIs indefinidos: trabalhar sem KPIs claros pode levar ao monitoramento excessivo ou insuficiente, fazendo com que sinais vitais possam ser perdidos.
- KPIs estáticos: não revisitar ou refinar os KPIs à medida que a workload ou os objetivos de negócios evoluem.
- Desalinhamento: foco em métricas técnicas que não se correlacionam diretamente com os resultados comerciais ou são mais difíceis de correlacionar com problemas do mundo real.

Benefícios de implementar esta prática recomendada:

- Facilidade de identificação de problemas: os KPIs de negócios geralmente mostram os problemas com mais clareza do que as métricas técnicas. Uma queda em um KPI comercial pode identificar um problema com mais eficiência do que analisar várias métricas técnicas.

- Alinhamento comercial: garante que as atividades de monitoramento apoiem diretamente os objetivos de negócios.
- Eficiência: priorize os recursos de monitoramento e a atenção nas métricas que importam.
- Proatividade: reconheça e resolva os problemas antes que eles tenham implicações comerciais mais amplas.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Para definir com eficácia os KPIs da workload:

1. Comece com os resultados comerciais: antes de mergulhar nas métricas, entenda o resultado comercial desejado. É sobre aumento de vendas, maior engajamento do usuário ou tempos de resposta mais rápidos?
2. Correlacione métricas técnicas com objetivos de negócios: nem todas as métricas técnicas têm impacto direto nos resultados comerciais. Identifique aquelas que têm, mas muitas vezes é mais fácil identificar um problema usando um KPI comercial.
3. Use o Amazon CloudWatch: utilize o CloudWatch para definir e monitorar métricas que representam seus KPIs.
4. Revise e atualize regularmente os KPIs: à medida que sua workload e seus negócios evoluem, mantenha seus KPIs relevantes.
5. Envolva as partes interessadas: envolva as equipes técnicas e comerciais na definição e revisão dos KPIs.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- the section called “OPS04-BP02 Implementar a telemetria de aplicações”
- the section called “OPS04-BP03 Implementar telemetria da experiência do usuário”
- the section called “OPS04-BP04 Implementar a telemetria de dependências”
- the section called “OPS04-BP05 Implementar rastreamento distribuído”

Documentos relacionados:

- Práticas recomendadas de observabilidade da AWS
- Guia do usuário do CloudWatch
- Curso de desenvolvimento de habilidades de observabilidade da AWS

Vídeos relacionados:

- Desenvolver de uma estratégia de observabilidade

Exemplos relacionados:

- Workshop One Observability

OPS04-BP02 Implementar a telemetria de aplicações

A telemetria de aplicações serve como base para a observabilidade da workload. É fundamental emitir uma telemetria que ofereça informações práticas sobre o estado da sua aplicação e a obtenção de resultados técnicos e comerciais. Da solução de problemas à medição do impacto de um novo recurso ou à garantia do alinhamento com os indicadores-chave de performance (KPIs) de negócios, a telemetria de aplicações informa a maneira como você cria, opera e desenvolve sua workload.

Métricas, logs e rastreamentos formam os três pilares principais da observabilidade. Eles servem como ferramentas de diagnóstico que descrevem o estado de sua aplicação. Com o tempo, eles auxiliam na criação de linhas de base e na identificação de anomalias. No entanto, para garantir o alinhamento entre as atividades de monitoramento e os objetivos de negócios, é fundamental definir e monitorar os KPIs. Os KPIs de negócios geralmente facilitam a identificação de problemas em comparação com métricas técnicas isoladas.

Outros tipos de telemetria, como monitoramento de usuários reais (RUM) e transações sintéticas, complementam essas fontes de dados primárias. O RUM oferece informações sobre as interações do usuário em tempo real, enquanto as transações sintéticas simulam possíveis comportamentos do usuário, ajudando a detectar gargalos antes que usuários reais os encontrem.

Resultado desejado: obtenha insights acionáveis sobre a performance da sua workload. Esses insights permitem que você tome decisões proativas sobre otimização de performance, tenha maior estabilidade da workload, simplifique os processos de CI/CD e utilize recursos de forma eficaz.

Práticas comuns que devem ser evitadas:

- Observabilidade incompleta: negligência da incorporação da observabilidade em todas as camadas da workload, resultando em pontos cegos que podem obscurecer insights vitais sobre performance e comportamento do sistema.
- Visualização fragmentada dos dados: quando os dados estão espalhados por várias ferramentas e sistemas, torna-se difícil manter uma visão holística da integridade e da performance da sua workload.
- Problemas relatados pelo usuário: um sinal de que falta a detecção proativa de problemas por meio da telemetria e do monitoramento de KPI de negócios.

Benefícios de implementar esta prática recomendada:

- Tomada de decisão informada: com insights de telemetria e KPIs de negócios, você pode tomar decisões baseadas em dados.
- Eficiência operacional aprimorada: a utilização de recursos baseada em dados leva à redução de custos.
- Estabilidade aprimorada da workload: detecção e resolução mais rápidas de problemas, levando a um melhor tempo de atividade.
- Processos racionalizados de CI/CD: os insights dos dados de telemetria facilitam o refinamento dos processos e a entrega confiável de código.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Para implementar a telemetria de aplicações para sua workload, use serviços da AWS como o Amazon CloudWatch e o AWS X-Ray. O Amazon CloudWatch fornece um conjunto abrangente de ferramentas de monitoramento, permitindo que você observe seus recursos e aplicações em ambientes da AWS e on-premises. Ele coleta, rastreia e analisa métricas, consolida e monitora dados de log e reage às mudanças em seus recursos, aprimorando sua compreensão de como a workload opera. Em conjunto, o AWS X-Ray permite rastrear, analisar e depurar suas aplicações, oferecendo uma compreensão profunda do comportamento da workload. Com recursos como mapas de serviços, distribuições de latência e cronogramas de rastreamento, o AWS X-Ray fornece insights sobre a performance da workload e os gargalos que a afetam.

Etapas de implementação

1. Identifique quais dados coletar: determine as métricas essenciais, os logs e os rastreamentos essenciais que ofereceriam informações substanciais sobre a integridade, a performance e o comportamento da sua workload.
2. Implante o agente do CloudWatch: o agente do CloudWatch é fundamental na aquisição de métricas do sistema e da aplicação e de logs de sua workload e de sua infraestrutura subjacente. O agente do CloudWatch também pode ser usado para coletar OpenTelemetry ou rastreamentos do X-Ray e enviá-los ao X-Ray.
3. Implemente a detecção de anomalias para logs e métricas: use a detecção de anomalias do CloudWatch Logs e a detecção de anomalias do CloudWatch Metrics para identificar automaticamente atividades incomuns nas operações da aplicação. Essas ferramentas usam algoritmos de machine learning para detectar e alertar sobre anomalias, o que aprimora os recursos de monitoramento e acelera o tempo de resposta a possíveis interrupções ou ameaças à segurança. Configure esses recursos para gerenciar proativamente a integridade e a segurança das aplicações.
4. Proteja dados de log confidenciais: use a proteção de dados do Amazon CloudWatch Logs para mascarar informações confidenciais em seus logs. Esse recurso ajuda a manter a privacidade e a conformidade por meio da detecção automática e do mascaramento de dados confidenciais antes de serem acessados. Implemente o mascaramento de dados para tratar e proteger com segurança detalhes confidenciais, como informações de identificação pessoal (PII).
5. Defina e monitore os KPIs de negócios: estabeleça métricas personalizadas que se alinhem aos seus resultados de negócios.
6. Instrumente sua aplicação com o AWS X-Ray: além de implantar o agente CloudWatch, é fundamental instrumentar sua aplicação para emitir dados de rastreamento. Esse processo pode fornecer mais insights sobre o comportamento e a performance da workload.
7. Padronize a coleta de dados em toda a sua aplicação: padronize as práticas de coleta de dados em toda a aplicação. A uniformidade ajuda a correlacionar e analisar dados, fornecendo uma visão abrangente do comportamento da aplicação.
8. Implemente a observabilidade entre contas: aumente a eficiência do monitoramento entre várias Contas da AWS com a observabilidade entre contas do Amazon CloudWatch. Com esse recurso, é possível consolidar métricas, logs e alarmes de contas diferentes em uma única visualização, o que simplifica o gerenciamento e melhora os tempos de resposta para problemas identificados em todo o ambiente da AWS da organização.

9. Analise e aja com base em dados: quando a coleta e a normalização dos dados estiverem implementadas, use o Amazon CloudWatch para análise de métricas e logs e o AWS X-Ray para análise de rastreamento. Essa análise pode gerar informações cruciais sobre a integridade, a performance e o comportamento da workload, orientando o processo de tomada de decisão.

Nível de esforço do plano de implementação: Alto

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Definir KPIs da workload
- OPS04-BP03 Implementar a telemetria de atividades dos usuários
- OPS04-BP04 Implementar a telemetria de dependências
- OPS04-BP05 Implementar a rastreabilidade das transações

Documentos relacionados:

- Práticas recomendadas de observabilidade da AWS
- Guia do usuário do CloudWatch
- AWS X-Ray Guia do desenvolvedor
- Instrumentar sistemas distribuídos para visibilidade operacional
- Curso de desenvolvimento de habilidades de observabilidade da AWS
- Novidades do Amazon CloudWatch
- Novidades da AWS X-Ray

Vídeos relacionados:

- AWS re:Invent 2022: práticas recomendadas de observabilidade na Amazon
- AWS re:Invent 2022: desenvolver uma estratégia de observabilidade

Exemplos relacionados:

- Workshop One Observability
- Biblioteca de soluções da AWS: monitorar aplicações com o Amazon CloudWatch

OPS04-BP03 Implementar telemetria da experiência do usuário

É essencial obter insights profundos sobre as experiências dos clientes e as interações com sua aplicação. O monitoramento de usuários reais (RUM) e as transações sintéticas servem como ferramentas poderosas para essa finalidade. O RUM fornece dados sobre interações reais do usuário, oferecendo uma perspectiva não filtrada da satisfação do usuário, enquanto as transações sintéticas simulam as interações do usuário, ajudando a detectar possíveis problemas antes mesmo que eles afetem os usuários reais.

Resultado desejado: uma visão holística da experiência do cliente, detecção proativa de problemas e otimização das interações do usuário para oferecer experiências digitais perfeitas.

Práticas comuns que devem ser evitadas:

- Aplicações sem monitoramento de usuários reais (RUM):
  - Detecção atrasada de problemas: sem o RUM, talvez você não fique ciente dos gargalos ou problemas de performance até que os usuários reclamem. Essa abordagem reativa pode levar à insatisfação do cliente.
  - Falta de insights sobre a experiência do usuário: não usar o RUM significa perder dados cruciais que mostram como usuários reais interagem com sua aplicação, limitando sua capacidade de otimizar a experiência do usuário.
- Aplicações sem transações sintéticas:
  - Casos de borda perdidos: transações sintéticas ajudam você a testar caminhos e funções que podem não ser usados com frequência por usuários comuns, mas são essenciais para determinadas funções de negócios. Sem eles, esses caminhos podem ter problemas de funcionamento e passar despercebidos.
  - Verificação de problemas quando a aplicação não está sendo usada: testes sintéticos regulares podem simular momentos em que usuários reais não estão interagindo ativamente com sua aplicação, garantindo que o sistema sempre funcione corretamente.

Benefícios de implementar esta prática recomendada:

- Detecção proativa de problemas: identifique e resolva possíveis problemas antes que eles afetem usuários reais.
- Experiência otimizada do usuário: o feedback contínuo do RUM ajuda a refinar e aprimorar a experiência geral do usuário.

- Informações sobre a performance do dispositivo e do navegador: entenda a performance da sua aplicação em vários dispositivos e navegadores, permitindo uma maior otimização.
- Fluxos de trabalho de negócios validados: transações sintéticas regulares garantem que as principais funcionalidades e os caminhos críticos permaneçam operacionais e eficientes.
- Performance aprimorada da aplicação: utilize as informações coletadas de dados reais do usuário para melhorar a capacidade de resposta e a confiabilidade da aplicação.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Para aproveitar o RUM e as transações sintéticas na telemetria da atividade do usuário, a AWS oferece serviços como o Amazon CloudWatch RUM e o Amazon CloudWatch Synthetics. Métricas, logs e rastreamentos, juntamente com dados de atividades do usuário, fornecem uma visão abrangente do estado operacional da aplicação e da experiência do usuário.

Etapas de implementação

1. Implemente o Amazon CloudWatch RUM: integre sua aplicação ao CloudWatch RUM para coletar, analisar e apresentar dados reais do usuário.
   a. Use a biblioteca JavaScript RUM do CloudWatch para integrar o RUM à aplicação.
   b. Configure painéis para visualizar e monitorar dados reais do usuário.
2. Configure o CloudWatch Synthetics: crie canários ou rotinas com script para simular as interações do usuário com sua aplicação.
   a. Defina fluxos de trabalho e caminhos de aplicação críticos.
   b. Crie canários usando scripts do CloudWatch Synthetics para simular as interações do usuário nesses caminhos.
   c. Programe e monitore os canários para serem executados em intervalos específicos, garantindo verificações de performance consistentes.
3. Analise e aja sobre os dados: utilize dados de RUM e transações sintéticas para obter insights e tomar medidas corretivas quando anomalias forem detectadas. Use painéis e alarmes do CloudWatch para se manter informado.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Identificar indicadores-chave de performance
- OPS04-BP02 Implementar a telemetria de aplicações
- OPS04-BP04 Implementar a telemetria de dependências
- OPS04-BP05 Implementar rastreamento distribuído

Documentos relacionados:

- Guia do Amazon CloudWatch RUM
- Guia do Amazon CloudWatch Synthetics

Vídeos relacionados:

- Otimizar aplicações com base em insights do usuário final com o Amazon CloudWatch RUM
- Destaque da AWS On Air: Monitoramento de usuários reais para Amazon CloudWatch

Exemplos relacionados:

- Workshop One Observability
- Repositório do Git para cliente Web do Amazon CloudWatch RUM
- Usar o Amazon CloudWatch Synthetics para medir o tempo de carregamento da página

OPS04-BP04 Implementar a telemetria de dependências

A telemetria de dependências é essencial para monitorar a integridade e a performance dos serviços e componentes externos dos quais a workload depende. Ela fornece insights valiosos sobre acessibilidade, tempos limite e outros eventos críticos relacionados a dependências, como DNS, bancos de dados ou APIs de terceiros. Ao instrumentar sua aplicação para emitir métricas, logs e rastreamentos sobre essas dependências, você adquire uma compreensão mais clara dos possíveis gargalos, problemas de performance ou falhas que podem afetar a workload.

Resultado desejado: as dependências das quais a workload depende estão funcionando conforme o esperado, permitindo que você resolva problemas de forma proativa e garanta a performance ideal da workload.

Práticas comuns que devem ser evitadas:

- Negligenciar as dependências externas: focar apenas nas métricas internas da aplicação e negligenciar as métricas relacionadas às dependências externas.
- Ausência de monitoramento proativo: aguardar o surgimento de problemas em vez de monitorar continuamente a integridade e a performance da dependência.
- Monitoramento em silos: usar várias ferramentas de monitoramento diferentes, o que pode resultar em visualizações fragmentadas e inconsistentes da integridade da dependência.

Benefícios de implementar esta prática recomendada:

- Maior confiabilidade da workload: garantia de que as dependências externas estejam consistentemente disponíveis e tenham uma performance ideal.
- Detecção e resolução mais rápidas de problemas: identificação e resolução proativa de problemas com dependências antes que elas afetem a workload.
- Visão abrangente: obtenção de uma visão holística dos componentes internos e externos que influenciam a integridade da workload.
- Escalabilidade aprimorada da workload: compreensão dos limites de escalabilidade e das características de performance das dependências externas.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Implemente a telemetria de dependências começando com a identificação dos serviços, da infraestrutura e dos processos dos quais a workload depende. Quantifique quais são as boas condições quando essas dependências estão funcionando conforme o esperado e determine quais dados serão necessários para medi-las. Com essas informações, você pode criar painéis e alertas que forneçam insights para suas equipes de operações sobre o estado dessas dependências. Use ferramentas da AWS para descobrir e quantificar os impactos quando as dependências não tiverem a performance necessária. Revise continuamente sua estratégia para considerar as mudanças nas prioridades, metas e insights obtidos.

Etapas de implementação

Para implementar a telemetria de dependências de forma eficaz:

1. Identifique dependências externas: colabore com as partes interessadas para identificar as dependências externas das quais a workload depende. As dependências externas podem abranger serviços como bancos de dados externos, APIs de terceiros, rotas de conectividade de rede para outros ambientes e serviços de DNS. O primeiro passo para uma telemetria de dependências eficaz é entender de forma abrangente quais são essas dependências.
2. Desenvolver uma estratégia de monitoramento: depois de obter uma visão clara de suas dependências externas, elabore uma estratégia de monitoramento personalizada para elas. Isso envolve entender a importância de cada dependência, seu comportamento esperado e quaisquer contratos ou metas de nível de serviço associados (SLA ou SLTs). Configure alertas proativos para receber notificações sobre mudanças de status ou desvios de performance.
3. Use o monitoramento de rede: use o Internet Monitor e o Network Monitor para obter informações abrangentes sobre as condições globais da Internet e da rede. Essas ferramentas ajudam você a entender e reagir a interrupções ou degradações de performance que afetam as dependências externas.
4. Mantenha-se em dia com o AWS Health Dashboard: ele fornece alertas e orientações de correção quando a AWS está enfrentando eventos que podem impactar seus serviços.
    a. Monitore eventos do AWS Health com as regras do Amazon EventBridge ou integre-os programaticamente à AWS Health API para automatizar ações ao receber eventos do AWS Health. Podem ser ações gerais, como enviar todas as mensagens planejadas de eventos do ciclo de vida para uma interface de chat, ou ações específicas, como o início de um fluxo de trabalho em uma ferramenta de gerenciamento de serviços de TI.
    b. Se você usar o AWS Organizations, agregue eventos do AWS Health em todas as contas.
5. Instrumente sua aplicação com o AWS X-Ray: o AWS X-Ray fornece informações sobre a performance das aplicações e de suas respectivas dependências subjacentes. Ao rastrear as solicitações do início ao fim, você pode identificar gargalos ou falhas nos serviços ou componentes externos dos quais sua aplicação depende.
6. Use o Amazon DevOps Guru: esse serviço orientado por machine learning identifica problemas operacionais, prevê quando problemas críticos podem ocorrer e recomenda ações específicas a serem tomadas. Ele é inestimável para ter informações sobre dependências e determinar que elas não são a fonte dos problemas operacionais.
7. Monitore regularmente: monitore continuamente métricas e logs relacionados a dependências externas. Configure alertas para comportamento inesperado ou diminuição de performance.

8. Valide após as alterações: sempre que houver uma atualização ou alteração em qualquer uma das dependências externas, valide sua performance e verifique o alinhamento com os requisitos da sua aplicação.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Definir KPIs da workload
- OPS04-BP02 Implementar a telemetria de aplicações
- OPS04-BP03 Implementar a telemetria de atividades dos usuários
- OPS04-BP05 Implementar a rastreabilidade das transações
- OP08-BP04 Criar alertas acionáveis

Documentos relacionados:

- Guia do usuário do Amazon Personal AWS Health Dashboard
- Guia do usuário do AWS Internet Monitor
- AWS X-Ray Guia do desenvolvedor
- Guia do usuário do AWS DevOps Guru

Vídeos relacionados:

- Visibilidade sobre como as questões da Internet afetam a performance de aplicações
- Introdução ao Amazon DevOps Guru
- Gerenciar eventos do ciclo de vida dos recursos em grande escala com o AWS Health

Exemplos relacionados:

- Obter insights operacionais com AIOps usando o Amazon DevOps Guru
- AWS Health Aware
- Usar a filtragem baseada em tags para gerenciar o monitoramento e os alertas do AWS Health em grande escala

OPS04-BP05 Implementar rastreamento distribuído

O rastreamento distribuído oferece uma maneira de monitorar e visualizar solicitações à medida que elas percorrem vários componentes de um sistema distribuído. Ao capturar dados de rastreamento de várias fontes e analisá-los em uma visão unificada, as equipes podem entender melhor como as solicitações fluem, onde existem gargalos e onde os esforços de otimização devem se concentrar.

Resultado desejado: obtenha uma visão holística das solicitações que fluem pelo seu sistema distribuído, permitindo depuração precisa, performance otimizada e experiências de usuário aprimoradas.

Práticas comuns que devem ser evitadas:

- Instrumentação inconsistente: nem todos os serviços em um sistema distribuído são instrumentados para rastreamento.
- Ignorar a latência: foco apenas nos erros e sem considerar a latência ou as degradações graduais da performance.

Benefícios de implementar esta prática recomendada:

- Visão geral abrangente do sistema: visualização de todo o caminho das solicitações, da entrada à saída.
- Depuração aprimorada: identificação rápida de onde ocorrem falhas ou problemas de performance.
- Experiência de usuário aprimorada: monitoramento e otimização com base nos dados reais do usuário, garantindo que o sistema atenda às demandas do mundo real.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Comece identificando todos os elementos da workload que exigem instrumentação. Depois que todos os componentes forem contabilizados, utilize ferramentas como o AWS X-Ray e o OpenTelemetry para coletar dados de rastreamento para análise com ferramentas como o X-Ray e o Amazon CloudWatch ServiceLens Map. Faça avaliações regulares com desenvolvedores e complemente essas discussões com ferramentas como o Amazon DevOps Guru, o X-Ray Analytics e o X-Ray Insights para ajudar a fazer descobertas mais profundas. Estabeleça alertas a partir

de dados de rastreamento para notificar quando os resultados, conforme definido no plano de monitoramento da workload, estiverem em risco.

Etapas de implementação

Para implementar o rastreamento distribuído de forma eficaz:

1. Adote o AWS X-Ray: integre o X-Ray à sua aplicação para obter informações sobre seu comportamento, entender sua performance e identificar gargalos. Utilize o X-Ray Insights para análise automática de rastreamento.
2. Instrumente seus serviços: verifique se cada serviço, de uma função do AWS Lambda a uma instância do EC2, envia dados de rastreamento. Quanto mais serviços você instrumentar, mais clara será a visão completa.
3. Incorpore o monitoramento de usuários reais do CloudWatch e o monitoramento sintético: integre o monitoramento de usuários reais (RUM) e o monitoramento sintético com o X-Ray. Isso permite capturar experiências reais do usuário e simular as interações do usuário para identificar possíveis problemas.
4. Use o agente do CloudWatch: o agente pode enviar rastreamentos a partir do X-Ray ou do OpenTelemetry, aumentando a profundidade dos insights obtidos.
5. Use o Amazon DevOps Guru: o DevOps Guru usa dados do X-Ray, CloudWatch, AWS Config e AWS CloudTrail para fornecer recomendações práticas.
6. Analise os rastreamentos: revise regularmente os dados de rastreamento para discernir padrões, anomalias ou gargalos que possam afetar a performance da sua aplicação.
7. Configure alertas: configure alarmes no CloudWatch para padrões incomuns ou latências estendidas, permitindo o tratamento proativo de problemas.
8. Aprimoramento contínuo: revise sua estratégia de rastreamento à medida que os serviços são adicionados ou modificados para capturar todos os pontos de dados relevantes.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Identificar indicadores-chave de performance
- OPS04-BP02 Implementar a telemetria de aplicações
- OPS04-BP03 Implementar telemetria da experiência do usuário

- OPS04-BP04 Implementar a telemetria de dependências

Documentos relacionados:

- Guia do desenvolvedor do AWS X-Ray
- Guia do usuário do agente do Amazon CloudWatch
- Guia do usuário do Amazon DevOps Guru

Vídeos relacionados:

- Usar o AWS X-Ray Insights
- Destaque da AWS On Air: Observabilidade: Amazon CloudWatch e AWS X-Ray

Exemplos relacionados:

- Instrumentar sua aplicação para o AWS X-Ray

OPS 5. Como reduzir defeitos, facilitar a correção e melhorar o fluxo na produção?

Adote abordagens que melhoram o fluxo de alterações na produção, que acionem refatoração, feedback rápido sobre a qualidade e correção de erros. Isso acelera as alterações benéficas que entram na produção, limita os problemas implantados e alcança a rápida identificação e correção dos problemas introduzidos pelas atividades de implantação.

Práticas recomendadas

- OPS05-BP01 Usar controle de versão
- OPS05-BP02 Testar e validar alterações
- OPS05-BP03 Usar sistemas de gerenciamento de configuração
- OPS05-BP04 Usar sistemas de gerenciamento de compilação e de implantação
- OPS05-BP05 Executar o gerenciamento de patches
- OPS05-BP06 Compartilhar padrões de design
- OPS05-BP07 Implementar práticas para aprimorar a qualidade do código
- OPS05-BP08 Usar vários ambientes
- OPS05-BP09 Fazer alterações frequentes, pequenas e reversíveis

- OPS05-BP10 Automatizar totalmente a integração e a implantação

OPS05-BP01 Usar controle de versão

Use o controle de versão para ativar o rastreamento de alterações e liberações.

Muitos serviços da AWS oferecem recursos de controle de versão. Use um sistema controle de revisões ou código-fonte como o AWS CodeCommit para gerenciar código e outros artefatos, como modelos do AWS CloudFormation com controle de versão da sua infraestrutura.

Resultado desejado: suas equipes colaboram no código. Quando mesclado, o código é consistente e nenhuma alteração é perdida. Os erros são facilmente revertidos por meio do versionamento correto.

Práticas comuns que devem ser evitadas:

- Você está desenvolvendo e armazenando seu código na estação de trabalho. Você teve uma falha de armazenamento irrecuperável na estação de trabalho e seu código foi perdido.
- Depois de substituir o código existente pelas alterações, você reinicia a aplicação e ela deixa de ser operável. Não é possível reverter a alteração.
- Você tem um bloqueio de gravação em um arquivo de relatório que outra pessoa precisa editar. Ela entra em contato com você solicitando que você interrompa o trabalho para que ela possa concluir as tarefas.
- Sua equipe de pesquisa tem trabalhado em uma análise detalhada que moldará seu trabalho futuro. Alguém salvou acidentalmente a lista de compras sobre o relatório final. Não é possível reverter a alteração e você terá que recriar o relatório.

Benefícios de implementar esta prática recomendada: ao usar recursos de controle de versão, você pode reverter facilmente para estados e versões anteriores reconhecidamente bons e limitar o risco de perda de ativos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Mantenha ativos em repositórios controlados por versão. Fazer isso oferece suporte ao rastreamento de alterações, à implantação de novas versões, à detecção de alterações nas versões existentes e à reversão para versões anteriores (por exemplo, a reversão para um estado reconhecidamente bom no caso de uma falha). Integre os recursos de controle de versão dos sistemas de gerenciamento de configurações aos seus procedimentos.

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP04 Usar sistemas de gerenciamento de compilação e de implantação

Documentos relacionados:

- O que é AWS CodeCommit?

Vídeos relacionados:

- Introdução à AWS CodeCommit

OPS05-BP02 Testar e validar alterações

Cada alteração implantada deve ser testada para evitar erros na produção. A prática recomendada concentra-se em testar alterações do controle de versão na build de artefato. Além das alterações do código da aplicação, o teste deve incluir infraestrutura, configuração, controles de segurança e procedimentos de operações. O teste assume muitas formas, desde testes de unidade à análise dos componentes do software (SCA). Mova os testes mais para a esquerda na integração do software e o processo de entrega resultará em maior certeza da qualidade do artefato.

Sua organização deve desenvolver padrões de teste para todos os artefatos de software. Os testes automatizados reduzem o trabalho e evitam erros de testes manuais. Os testes manuais podem ser necessários em alguns casos. Os desenvolvedores precisam ter acesso aos resultados dos testes automatizados para criar loops de feedback que melhorem a qualidade do software.

Resultado desejado: as alterações do software são testadas antes de serem entregues. Os desenvolvedores têm acesso aos resultados e às validações dos testes. Sua organização tem um padrão de testes que se aplica a todas as alterações do software.

Práticas comuns que devem ser evitadas:

- Você implanta uma nova alteração do software sem nenhum teste. Ele não é executado na produção, o que ocasiona uma interrupção.
- Novos grupos de segurança são implantados com o AWS CloudFormation sem serem testados em um ambiente de pré-produção. Os grupos de segurança tornam sua aplicação inacessível para seus clientes.

- Um método é modificado, mas não há testes de unidade. O software falha quando é implantado em produção.

Benefícios de implementar esta prática recomendada: a taxa de falhas em alterações nas implantações de software é reduzida. A qualidade do software é aprimorada. Os desenvolvedores aumentaram a conscientização sobre a viabilidade do código deles. As políticas de segurança podem ser distribuídas com confiança para apoiar a conformidade da organização. Alterações da infraestrutura, como atualizações da política de ajuste de escala automático, são testadas com antecedência para atender às necessidades de tráfego.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Testes são realizados em todas as alterações, desde o código da aplicação à infraestrutura, como parte de sua prática de integração contínua. Os resultados dos testes são publicados para que os desenvolvedores tenham feedback rápido. Sua organização tem um padrão de testes de que todas as alterações devem ser aprovadas.

Use o poder da IA generativa com o Amazon Q Developer para melhorar a produtividade do desenvolvedor e a qualidade do código. O Amazon Q Developer inclui a geração de sugestões de código (com base em grandes modelos de linguagem), produção de testes unitários (incluindo condições de limite) e aprimoramentos de segurança de código por meio da detecção e correção de vulnerabilidades de segurança.

Exemplo de cliente

Como parte do pipeline de integração contínua, a AnyCompany Retail realiza alguns tipos de teste em todos os artefatos de software. Eles praticam desenvolvimento orientado a testes para que todo o software tenha testes de unidade. Depois que o artefato é criado, eles executam testes completos. Depois que a primeira etapa de testes é concluída, eles executam uma verificação de segurança da aplicação estática, que procura vulnerabilidades conhecidas. Os desenvolvedores recebem mensagens à medida que cada gate de testes é aprovado. Depois que todos os testes são concluídos, o artefato de software é armazenado em um repositório de artefatos.

Etapas de implementação

1. Trabalhe com partes interessadas em sua organização para desenvolver um padrão de testes para artefatos de software. Em quais testes padrão todos os artefatos devem ser aprovados? Há

requisitos de conformidade ou governança que devem ser incluídos na cobertura de testes? Você precisa realizar testes de qualidade de código? Quando os testes são concluídos, quem precisa saber?

1. A Arquitetura de referência do pipeline de implantação da AWS contém uma lista confiável de tipos de testes que podem ser conduzidos em artefatos de software como parte de um pipeline de integração.

2. Instrumente sua aplicação com os testes necessários com base em seu padrão de testes de software. Cada conjunto de testes deve ser concluído em menos de dez minutos. Os testes devem ser executados como parte de um pipeline de integração.
   a. Use o Amazon Q Developer, uma ferramenta generativa de IA que pode ajudar a criar casos de teste unitários (incluindo condições de limite), gerar funções usando código e comentários e implementar algoritmos conhecidos.
   b. Use o Amazon CodeGuru Reviewer para testar defeitos no código da sua aplicação.
   c. Você pode usar o AWS CodeBuild para realizar testes em artefatos de software.
   d. O AWS CodePipeline pode orquestrar seus testes de software em um pipeline.

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP01 Usar controle de versão
- OPS05-BP06 Compartilhar padrões de design
- OPS05-BP07 Implementar práticas para aprimorar a qualidade do código
- OPS05-BP10 Automatizar totalmente a integração e a implantação

Documentos relacionados:

- Adote uma abordagem de desenvolvimento orientado por testes
- Acelerar seu ciclo de vida de desenvolvimento de software com o Amazon Q
- O Amazon Q Developer, agora disponível ao público em geral, inclui visualizações de novos recursos para reimaginar a experiência do desenvolvedor
- A folha de dicas definitiva para usar o Amazon Q Developer em seu IDE
- workload Shift-Left: aproveitando a IA para a criação de testes
- Centro de desenvolvedores do Amazon Q

- Dez maneiras de criar aplicações mais rapidamente com o Amazon CodeWhisperer
- Olhar além da cobertura de código com o Amazon CodeWhisperer
- Práticas recomendadas para engenharia rápida com o Amazon CodeWhisperer
- Pipeline de teste do AWS CloudFormation automatizado com TaskCat e CodePipeline
- Criar um pipeline de CI/CD completo do AWS DevSecOps com ferramentas de código aberto SCA, SAST e DAST
- Conceitos básicos de testes de aplicações com tecnologia sem servidor)
- Meu pipeline de CI/CD é meu capitão de lançamentos
- Whitepaper Praticar a integração e entrega contínuas na AWS

Vídeos relacionados:

- Implementar uma API com o Amazon Q Developer Agent para desenvolvimento de software
- Instalar, configurar e usar o Amazon Q Developer com os IDEs da JetBrains (instruções)
- Dominar a arte do Amazon CodeWhisperer: playlist do YouTube
- AWS re:Invent 2020: Infraestrutura testável: teste de integração na AWS
- AWS Summit ANZ 2021: Conduzir uma estratégia de primeiro teste com o CDK e desenvolvimento orientado a testes
- Testar sua infraestrutura como código com o AWS CDK CDK

Recursos relacionados:

- Criar aplicações usando IA generativa com o Amazon CodeWhisperer
- Workshop do Amazon CodeWhisperer
- Arquitetura de referência do pipeline de implantação da AWS: aplicação
- Pipeline de DevSecOps de Kubernetes da AWS
- Workshop Política como código: desenvolvimento orientado a testes
- Executar testes de unidade para uma aplicação Node.js do GitHub usando o AWS CodeBuild
- Usar o Serverspec para o desenvolvimento orientado por testes de código de infraestrutura

Serviços relacionados:

- Amazon Q Developer
- Amazon CodeGuru Reviewer
- AWS CodeBuild
- AWS CodePipeline

OPS05-BP03 Usar sistemas de gerenciamento de configuração

Use os sistemas de gerenciamento de configuração para fazer e rastrear alterações nas configurações. Esses sistemas reduzem os erros causados pelos processos manuais e o nível de esforço para implantar as alterações.

O gerenciamento da configuração estática define valores ao inicializar um recurso que deve permanecer consistente durante todo o tempo de vida do recurso. O gerenciamento da configuração dinâmica define valores na inicialização que podem ou devem ser alterados durante o tempo de vida de um recurso. Por exemplo, é possível definir um recurso para ativar a funcionalidade em seu código por meio de uma alteração na configuração ou alterar o nível de detalhes do registro durante um incidente.

As configurações devem ser implantadas em um estado conhecido e consistente. Recomenda-se usar a inspeção automatizada para monitorar continuamente as configurações de recursos em todos os ambientes e regiões. Esses controles devem ser definidos como código e gerenciamento automatizados para garantir que as regras sejam aplicadas de forma consistente em todos os ambientes. As alterações nas configurações devem ser atualizadas por meio de procedimentos de controle de alterações acordados e aplicadas de forma consistente, respeitando o controle de versão. A configuração da aplicação deve ser gerenciada independentemente do código da aplicação e da infraestrutura. Isso permite uma implantação consistente em vários ambientes. As alterações na configuração não resultam na reconstrução ou reimplantação da aplicação.

Resultado desejado: você configura, valida e implanta como parte de seu pipeline de integração contínua e entrega contínua (CI/CD). Você monitora para validar se as configurações estão corretas. Isso minimiza qualquer impacto para usuários finais e clientes.

Práticas comuns que devem ser evitadas:

- Você atualiza manualmente a configuração do servidor Web em toda a frota e vários servidores não respondem devido a erros de atualização.
- Você atualiza manualmente a frota do servidor de aplicações ao longo de muitas horas. A inconsistência na configuração durante a alteração causa comportamentos inesperados.

- Alguém atualizou seus grupos de segurança e seus servidores Web não estão mais acessíveis. Sem saber o que foi alterado, você gasta muito tempo investigando o problema, ampliando o tempo de recuperação.
- Você coloca uma configuração de pré-produção em produção por meio de CI/CD sem validação. Você expõe usuários e clientes a dados e serviços incorretos.

Benefícios de implementar esta prática recomendada: a adoção de sistemas de gerenciamento de configurações reduz o nível de esforço para fazer e rastrear alterações, bem como a frequência de erros causados por procedimentos manuais. Os sistemas de gerenciamento de configuração fornecem garantias com relação aos requisitos regulatórios, de conformidade e de governança.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Os sistemas de gerenciamento de configuração são usados para rastrear e implementar alterações nas configurações de aplicações e ambientes. Os sistemas de gerenciamento de configuração também são usados para reduzir erros causados por processos manuais, tornar as alterações de configuração repetíveis e auditáveis e reduzir o nível de esforço.

Na AWS, é possível usar o AWS Config para monitorar continuamente suas configurações de recursos da AWS em todas as contas e regiões. Isso ajuda a rastrear o histórico da configuração, compreender como a alteração de uma configuração afeta outros recursos e auditá-la em relação a configurações esperadas ou desejadas, usando o Regras do AWS Config e o AWS Config Conformance Packs.

Para configurações dinâmicas em suas aplicações executadas em instâncias do Amazon EC2, AWS Lambda, contêineres, aplicações móveis ou dispositivos de IoT do Amazon EC2, você pode usar o AWS AppConfigpara configurá-los, valida-los, implantá-los e monitorá-los em seus ambientes.

Etapas de implementação

1. Identifique os proprietários da configuração.
    a. Informe os proprietários das configurações sobre quaisquer necessidades regulatórias, de conformidade ou de controle.
2. Identifique os itens de configuração e os resultados.
    a. Os itens de configuração são todas as configurações de aplicações e ambientes afetadas por uma implantação em seu pipeline de CI/CD.

   b. Os resultados incluem critérios de sucesso, validação e o que monitorar.
3. Selecione ferramentas para gerenciamento de configuração com base nos requisitos de seus negócios e no pipeline de entrega.
4. Considere implantações ponderadas, como implantações canário, para alterações significativas na configuração, a fim de minimizar o impacto de configurações incorretas.
5. Integre seu gerenciamento de configuração ao seu pipeline de CI/CD.
6. Valide todas as alterações enviadas.

Recursos

Práticas recomendadas relacionadas:

- OPS06-BP01 Preparar-se para alterações malsucedidas
- OPS06-BP02 Testar implantações
- OPS06-BP03 Utilizar estratégias de implantação seguras
- OPS06-BP04 Automatizar os testes e a reversão

Documentos relacionados:

- AWS Control Tower
- Acelerador de zona de pouso da AWS
- AWS Config
- O que é o AWS Config?
- AWS AppConfig
- O que é o AWS CloudFormation?
- Ferramentas de desenvolvedor AWS

Vídeos relacionados:

- AWS re:Invent 2022: governança e conformidade proativas para workloads da AWS
- AWS re:Invent 2020: Alcançar a conformidade como código usando o AWS Config
- Gerenciar e implantar configurações de aplicações com o AWS AppConfig

OPS05-BP04 Usar sistemas de gerenciamento de compilação e de implantação

Use sistemas de gerenciamento de compilação e implantação. Esses sistemas reduzem os erros causados pelos processos manuais e o nível de esforço para implantar as alterações.

Na AWS, é possível criar pipelines de integração contínua/implantação contínua (CI/CD) usando serviços como as Ferramentas de desenvolvedor da AWS (por exemplo, AWS CodeCommit, AWS CodeBuild, AWS CodePipeline, AWS CodeDeploy e AWS CodeStar).

Resultado desejado: seus sistemas de gerenciamento de compilação e implantação oferecem suporte ao sistema de integração contínua (CI/CD) de sua organização, que fornece recursos para automatizar implementações seguras com as configurações corretas.

Práticas comuns que devem ser evitadas:

- Depois de compilar o código no sistema de desenvolvimento e copiar o executável nos sistemas de produção, há uma falha na inicialização. Os arquivos de log locais indicam que a falha ocorreu devido à ausência de dependências.
- Você cria a aplicação com êxito com os novos recursos em seu ambiente de desenvolvimento e fornece o código à garantia de qualidade (QA). Ele falha no QA porque não há ativos estáticos.
- Na sexta-feira, após muito esforço, você consegue criar a aplicação manualmente em seu ambiente de desenvolvimento, incluindo os recursos recém-codificados. Na segunda-feira, você não consegue repetir as etapas que permitiram criar a aplicação com êxito.
- Você executa os testes que criou para a nova versão. Então você passa a próxima semana configurando um ambiente de teste e executando todos os testes de integração existentes, seguidos pelos testes de performance. O novo código tem um impacto inaceitável na performance e deve ser desenvolvido e testado novamente.

Benefícios de implementar esta prática recomendada: ao fornecer mecanismos para gerenciar atividades de criação e implantação, você reduz o nível de esforço para executar tarefas repetitivas, libera os membros da equipe para se concentrarem em tarefas criativas de alto valor e limita o surgimento de erros provenientes de procedimentos manuais.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Os sistemas de gerenciamento de compilação e implantação são usados para rastrear e implementar mudanças, reduzir erros causados por processos manuais e reduzir o nível de esforço necessário

para implantações seguras. Automatize totalmente o pipeline de integração e implantação desde o check-in do código até a compilação, o teste, a implantação e a validação. Isso reduz o tempo de espera, diminui os custos, incentiva o aumento da frequência de mudanças, reduz o nível de esforço e aumenta a colaboração.

Etapas de implementação

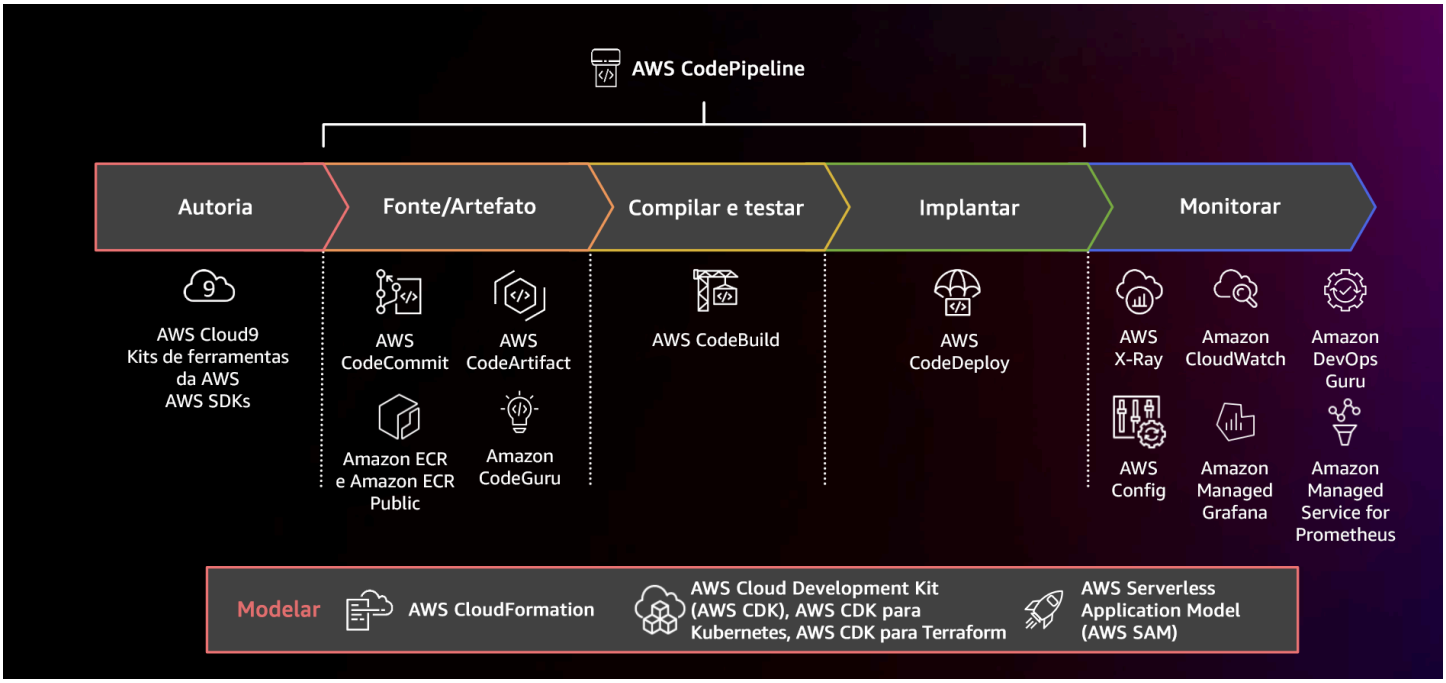


Diagrama que mostra um pipeline de CI/CD usando o AWS CodePipeline e serviços relacionados

1. Use o AWS CodeCommit para controlar versões, armazenar e gerenciar ativos (como documentos, código-fonte e arquivos binários).
2. Use o CodeBuild para compilar código-fonte, executar testes de unidade e produzir artefatos prontos para implantação.
3. Use o CodeDeploy como um serviço de implantação que automatiza implantações de aplicações em instâncias do Amazon EC2, instâncias on-premises, funções AWS Lambda com tecnologia sem servidor ou Amazon ECS.
4. Monitore suas implantações.

Recursos

Práticas recomendadas relacionadas:

- OPS06-BP04 Automatizar os testes e a reversão

Documentos relacionados:

- Ferramentas de desenvolvedor AWS
- O que é o AWS CodeCommit?
- O que é AWS CodeBuild?
- AWS CodeBuild
- O que é AWS CodeDeploy?

Vídeos relacionados:

- AWS re:Invent 2022: Práticas recomendadas do AWS Well-Architected para DevOps na AWS

OPS05-BP05 Executar o gerenciamento de patches

Execute o gerenciamento de patches para obter recursos, solucionar problemas e manter a conformidade com a governança. Automatize o gerenciamento de patches para reduzir erros causados por processos manuais, escalar e facilitar a realização de patches.

O gerenciamento de patches e vulnerabilidades faz parte de suas atividades de gerenciamento de benefícios e riscos. É preferível ter infraestruturas imutáveis e implantar workloads em bons estados verificados e conhecidos. Quando isso não é viável, a aplicação de patches é a opção restante.

O Amazon EC2 Image Builder fornece pipelines para atualizar imagens de máquinas. Como parte do gerenciamento de patches, considere utilizar imagens de máquina da Amazon (AMIs) com um pipeline de imagens de AMI ou imagens de contêiner com um pipeline de imagem Docker. Ao mesmo tempo, o AWS Lambda fornece padrões para runtimes personalizados e bibliotecas adicionais para remover vulnerabilidades.

Você deve gerenciar as atualizações das imagens de máquina da Amazon para Linux ou Windows Server usando o Amazon EC2 Image Builder. É possível usar o Amazon Elastic Container Registry (Amazon ECR) com seu pipeline existente para gerenciar imagens do Amazon ECS e gerenciar imagens do Amazon EKS. O Lambda inclui recursos de gerenciamento de versões.

A aplicação de patches não deve ser realizada em sistemas de produção sem antes testá-los em um ambiente seguro. Os patches só deverão ser aplicados se forem compatíveis com um resultado operacional ou comercial. Na AWS, é possível usar o AWS Systems Manager Patch Manager para automatizar o processo de aplicação de patches em sistemas gerenciados e programar a atividade usando as Janelas de manutenção do Systems Manager.

Resultado desejado: suas imagens de AMI e contêiner receberam os patches e estão atualizadas e prontas para o lançamento. É possível rastrear o status de todas as imagens implantadas e conhecer a conformidade do patch. Você também pode emitir relatórios do status atual e ter um processo para atender às suas necessidades de conformidade.

Práticas comuns que devem ser evitadas:

- Você recebe uma ordem para aplicar todos os novos patches de segurança em até duas horas, resultando em várias interrupções devido à incompatibilidade da aplicação com os patches.
- Uma biblioteca sem patches resulta em consequências indesejadas, pois partes desconhecidas usam vulnerabilidades dentro dela para acessar a workload.
- Você aplica patches nos ambientes do desenvolvedor automaticamente, sem notificar os desenvolvedores. Você recebe várias reclamações dos desenvolvedores afirmando que o ambiente deles não está funcionando conforme o esperado.
- Você não aplicou patches no software pronto para uso comercial em uma instância persistente. Quando você tiver um problema com o software e entrar em contato com o fornecedor, ele informará que a versão não é compatível e será necessário aplicar patches a um nível específico para receber assistência.
- Um patch lançado recentemente para o software de criptografia que você usou tem melhorias significativas de performance. Seu sistema sem patches tem problemas de performance que permanecem enquanto a aplicação de patches não é feita.
- Você é notificado sobre uma vulnerabilidade de dia zero que exige uma correção de emergência e precisa fazer isso em todos os seus ambientes manualmente.

Benefícios de implementar esta prática recomendada: ao estabelecer um processo de gerenciamento de patches, incluindo seus critérios de aplicação de patches e metodologia para distribuição em seus ambientes, você pode escalar e gerar relatórios sobre os níveis de patch. Isso fornece garantias sobre a aplicação de patches de segurança e garante uma visibilidade clara do status das correções conhecidas em vigor. Isso permite a adoção de recursos e capacidades desejados, a remoção rápida de problemas e a conformidade contínua com a governança. Implemente sistemas de gerenciamento de patches e automação para reduzir o nível de esforço na implantação de patches e limitar erros causados por processos manuais.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Aplique patches nos sistemas para corrigir problemas, obter os recursos ou capacidades desejados e permanecer em conformidade com a política de governança e os requisitos de suporte do fornecedor. Em sistemas imutáveis, implante com o conjunto de patches adequado para alcançar o resultado desejado. Automatize o mecanismo de gerenciamento de patches para reduzir o tempo decorrido na aplicação de patches, reduzir erros causados por processos manuais e reduzir o nível de esforço para corrigir.

Etapas de implementação

Para Amazon EC2 Image Builder:

1. Usando o Amazon EC2 Image Builder, especifique os detalhes do pipeline:
    a. Crie um pipeline de imagens e atribua um nome a ele
    b. Defina a programação e o fuso horário do pipeline
    c. Configure todas as dependências
2. Escolha uma fórmula:
    a. Selecione a fórmula existente ou crie uma nova.
    b. Selecione o tipo de imagem
    c. Nomeie e crie a versão da sua fórmula
    d. Selecione sua imagem base
    e. Adicione componentes de compilação e adicione ao registro de destino
3. Opcional: defina sua configuração de infraestrutura.
4. Opcional: defina as configurações.
5. Revise as configurações.
6. Mantenha a higiene da fórmula regularmente.

Para o Gerenciador de patches do Systems Manager:

1. Crie uma lista de referência de patches.
2. Selecione um método de operações de patch.
3. Habilite relatórios e verificações de conformidade.

Recursos

Práticas recomendadas relacionadas:

- OPS06-BP04 Automatizar os testes e a reversão

Documentos relacionados:

- O que é o Amazon EC2 Image Builder
- Criar um pipeline de imagens usando o Amazon EC2 Image Builder
- Criar um pipeline de imagens de contêiner
- Gerenciador de patches do AWS Systems Manager
- Trabalhar com o Patch Manager
- Trabalhar com relatórios de conformidade de patches
- Ferramentas de desenvolvedor da AWS

Vídeos relacionados:

- CI/CD para aplicações de tecnologia sem servidor na AWS
- Design com Ops em mente

Exemplos relacionados:

- Laboratórios do Well-Architected: Gerenciamento de inventário e patches
- Tutoriais do Gerenciador de patches do AWS Systems Manager

OPS05-BP06 Compartilhar padrões de design

Compartilhe práticas recomendadas entre equipes para aumentar a conscientização e maximizar os benefícios dos esforços de desenvolvimento. Documente-as e mantenha-as atualizadas à medida que sua arquitetura evolui. Se padrões compartilhados forem aplicados na sua organização, será fundamental que existam mecanismos para solicitar adições, alterações e exceções para os padrões. Sem essa opção, os padrões se tornam uma restrição à inovação.

Resultado desejado: os padrões de design são compartilhados entre as equipes nas organizações. Eles são documentados e mantidos atualizados de acordo com a evolução das práticas recomendadas.

Práticas comuns que devem ser evitadas:

- Cada uma das duas equipes de desenvolvimento criou um serviço de autenticação de usuários. Os usuários devem manter um conjunto separado de credenciais para cada parte do sistema que desejam acessar.
- Cada equipe gerencia sua própria infraestrutura. Um novo requisito de conformidade força uma alteração na infraestrutura e cada equipe o implementa de maneira diferente.

Benefícios de implementar esta prática recomendada: usar padrões compartilhados contribui para a adoção das práticas recomendadas e maximiza os benefícios dos esforços de desenvolvimento. A documentação e atualização dos padrões de design mantém a organização atualizada com relação às práticas recomendadas e aos requisitos de segurança e conformidade.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Compartilhe as práticas recomendadas, os padrões de design, as listas de verificação, os procedimentos operacionais, as orientações e os requisitos de governança entre equipes. Adote procedimentos para solicitar alterações, adições e exceções para padrões de design a fim de apoiar a melhoria e a inovação. As equipes devem estar cientes do conteúdo publicado. Adote um mecanismo para manter os padrões de design atualizados à medida que surgem novas práticas recomendadas.

Exemplo de cliente

A AnyCompany Retail tem uma equipe de arquitetura multifuncional que cria padrões de arquitetura de software. Essa equipe cria a arquitetura com conformidade e governança integradas. As equipes que adotam esses padrões compartilhados recebem os benefícios de ter a conformidade e governança integradas. Elas podem criar rapidamente com base no padrão de design. A equipe de arquitetura se reúne trimestralmente para avaliar os padrões de arquitetura e atualizá-los, se necessário.

Etapas de implementação

1. Identifique uma equipe multifuncional que seja responsável pelo desenvolvimento e pela atualização dos padrões de design. Essa equipe deverá trabalhar com as partes interessadas na organização para desenvolver os padrões de design, os procedimentos operacionais, as listas de

verificações, as orientações e os requisitos de governança. Documente os padrões de design e compartilhe-os na organização.

a. O AWS Service Catalog pode ser usado para criar portfólios representando os padrões de design usando infraestrutura como código. É possível compartilhar portfólios entre contas.

2. Tenha um mecanismo em vigor para manter os padrões de design atualizados à medida que novas práticas recomendadas são identificadas.
3. Se os padrões de design forem aplicados centralmente, tenha um processo para solicitar alterações, atualizações e isenções.

Nível de esforço do plano de implementação: Médio. O desenvolvimento de um processo para criar e compartilhar padrões de design pode exigir coordenação e cooperação com as partes interessadas na organização.

Recursos

Práticas recomendadas relacionadas:

- OPS01-BP03 Avaliar os requisitos de governança: os requisitos de governança influenciam os padrões de design.
- OPS01-BP04 Avaliar os requisitos de conformidade: a conformidade é um fator fundamental na criação dos padrões de design.
- OPS07-BP02 Garantir uma revisão consistente da prontidão operacional: as listas de verificação de prontidão operacional são um mecanismo para implementar os padrões de design ao projetar a workload.
- OPS11-BP01 Adotar um processo para melhoria contínua: a atualização dos padrões de design faz parte da melhoria contínua.
- OPS11-BP04 Gerenciar o conhecimento: como parte da sua prática de gerenciamento de conhecimento, documente e compartilhe os padrões de design.

Documentos relacionados:

- Automatizar AWS Backups com o AWS Service Catalog
- Conta do AWS Service Catalog aprimorada de fábrica
- Como o Expedia Group criou uma oferta de banco de dados como serviço (DBaaS) usando o AWS Service Catalog

- Manter a visibilidade sobre o uso dos padrões de arquitetura de nuvem
- Simplifique o compartilhamento de seus portfólios do AWS Service Catalog em uma configuração do AWS Organizations

Vídeos relacionados:

- Conceitos básicos do AWS Service Catalog
- AWS re:Invent 2020: gerenciar seus portfólios do AWS Service Catalog como um especialista

Exemplos relacionados:

- Arquitetura de referência do AWS Service Catalog
- Workshop do AWS Service Catalog

Serviços relacionados:

- AWS Service Catalog

OPS05-BP07 Implementar práticas para aprimorar a qualidade do código

Implemente práticas para aprimorar a qualidade do código e minimizar os defeitos. Alguns exemplos incluem desenvolvimento orientado por testes, análises de código, adoção de padrões e programação de pares. Incorpore essas práticas em seu processo de entrega e integração contínua.

Resultado desejado: sua organização usa práticas recomendadas como análises de código ou programação de pares para melhorar a qualidade do código. Os desenvolvedores e os operadores adotam práticas recomendadas de qualidade do código como parte do ciclo de vida de desenvolvimento de software.

Práticas comuns que devem ser evitadas:

- Você confirma o código para a ramificação principal da aplicação sem uma análise de código. A alteração é implantada automaticamente na produção e causa uma interrupção.
- Uma nova aplicação é desenvolvida sem nenhum teste de integração, completo ou de unidade. Não há como testar a aplicação antes da implantação.

- Sua equipe faz alterações manuais na produção para solucionar os defeitos. As alterações não passam por testes nem análises de código e não são capturadas nem registradas por processos contínuos de entrega e integração.

Benefícios de implementar esta prática recomendada: ao adotar práticas para melhorar a qualidade do código, é possível reduzir os problemas introduzidos na produção. A qualidade do código facilita o uso de práticas recomendadas, como programação de pares, análises de código e implementação de ferramentas de produtividade de IA.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Implemente práticas para melhorar a qualidade do código visando a minimizar os defeitos antes que eles sejam implantados. Use práticas como desenvolvimento orientado por testes, análises de código e programação de pares para aumentar a qualidade do desenvolvimento.

Use o poder da IA generativa com o Amazon Q Developer para melhorar a produtividade do desenvolvedor e a qualidade do código. O Amazon Q Developer inclui a geração de sugestões de código (com base em grandes modelos de linguagem), produção de testes unitários (incluindo condições de limite) e aprimoramentos de segurança de código por meio da detecção e correção de vulnerabilidades de segurança.

Exemplo de cliente

A AnyCompany Retail adota várias práticas para melhorar a qualidade do código. O desenvolvimento orientado por testes foi adotado com o padrão para escrever aplicações. Para alguns recursos novos, os desenvolvedores farão a programação de pares em conjunto durante um sprint. Cada pull request passa por uma análise de código feita por um desenvolvedor sênior antes de ser integrada e implantada.

Etapas de implementação

1. Adote práticas de qualidade de código como desenvolvimento orientado por testes, análises de código e programação de pares em seu processo de entrega e integração contínua. Use essas técnicas para melhorar a qualidade do software.
   a. Use o Amazon Q Developer, uma ferramenta de IA generativa que pode ajudar a criar casos de teste unitários (incluindo condições de limite), gerar funções usando código e comentários, implementar algoritmos conhecidos, detectar violações de políticas de segurança

e vulnerabilidades em seu código, detectar segredos, examinar infraestrutura como código (IaC) e código de documentos e aprender bibliotecas de código de terceiros mais rapidamente.

b. O Amazon CodeGuru Reviewer pode fornecer recomendações de programação para código Java e Python usando machine learning.

c. Você pode criar ambientes de desenvolvimento compartilhados com o AWS Cloud9, onde é possível colaborar no desenvolvimento de código.

Nível de esforço do plano de implementação: Médio. Há muitas maneiras de implementar essa prática recomendada, mas pode ser difícil garantir a adesão organizacional.

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP02 Testar e validar alterações
- OPS05-BP06 Compartilhar padrões de design

Documentos relacionados:

- Adote uma abordagem de desenvolvimento orientado por testes
- Acelerar seu ciclo de vida de desenvolvimento de software com o Amazon Q
- O Amazon Q Developer, agora disponível ao público em geral, inclui visualizações de novos recursos para reimaginar a experiência do desenvolvedor
- A folha de dicas definitiva para usar o Amazon Q Developer em seu IDE
- workload Shift-Left: aproveitando a IA para a criação de testes
- Centro de desenvolvedores do Amazon Q
- Dez maneiras de criar aplicações mais rapidamente com o Amazon CodeWhisperer
- Olhar além da cobertura de código com o Amazon CodeWhisperer
- Práticas recomendadas para engenharia rápida com o Amazon CodeWhisperer
- Guia do software Agile
- Meu pipeline de CI/CD é meu capitão de lançamentos
- Automatizar as revisões de código com o Amazon CodeGuru Reviewer
- Adote uma abordagem de desenvolvimento orientado por testes
- Como o DevFactory cria melhores aplicações com o Amazon CodeGuru

- Sobre a programação de pares
- RENGA Inc. automatiza as revisões de código com o Amazon CodeGuru)
- A arte do desenvolvimento ágil: desenvolvimento orientado por testes
- Por que as revisões de código são importantes (e economizam tempo!)

Vídeos relacionados:

- Implementar uma API com o Amazon Q Developer Agent para desenvolvimento de software
- Instalar, configurar e usar o Amazon Q Developer com os IDEs da JetBrains (instruções)
- Dominar a arte do Amazon CodeWhisperer: playlist do YouTube
- AWS re:Invent 2020: Melhoria contínua da qualidade do código com o Amazon CodeGuru
- AWS Summit ANZ 2021: Conduzir uma estratégia de primeiro teste com o CDK e desenvolvimento orientado a testes

Serviços relacionados:

- Amazon Q Developer
- Amazon CodeGuru Reviewer
- Amazon CodeGuru Profiler
- AWS Cloud9

OPS05-BP08 Usar vários ambientes

Use vários ambientes para experimentar, desenvolver e testar a workload. Use níveis crescentes de controles à medida que os ambientes se aproximam da produção para adquirir confiança de que sua workload operará conforme pretendido quando implantada.

Resultado desejado: você tem vários ambientes que refletem suas necessidades de conformidade e governança. Você testa e promove o código por meio de ambientes em seu caminho para a produção.

Práticas comuns que devem ser evitadas:

- Você está trabalhando em um desenvolvimento em um ambiente de desenvolvimento compartilhado e outro desenvolvedor substitui suas alterações de código.

- Os controles de segurança restritivos em seu ambiente de desenvolvimento compartilhado estão impedindo que você experimente novos serviços e recursos.
- Você realiza testes de carga em seus sistemas de produção e causa uma interrupção para seus usuários.
- Ocorreu um erro crítico na produção que resulta na perda de dados. No ambiente de produção, você tenta recriar as condições que levaram à perda de dados para identificar como isso aconteceu e impedir a recorrência. Para evitar mais perda de dados durante o teste, você é forçado a tornar indisponível a aplicação para seus usuários.
- Você está operando um serviço multilocatário e não consegue oferecer suporte a uma solicitação do cliente para um ambiente dedicado.
- Nem sempre você testa, mas, quando o faz, o teste acontece em seu ambiente de produção.
- Você acredita que a simplicidade de um único ambiente substitui o escopo do impacto das alterações dentro do ambiente.

Benefícios de implementar esta prática recomendada: é possível oferecer suporte a vários ambientes simultâneos de desenvolvimento, teste e produção sem criar conflitos entre desenvolvedores ou comunidades de usuários.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Use vários ambientes e forneça aos desenvolvedores ambientes de sandbox com controles minimizados para permitir a experimentação. Forneça ambientes de desenvolvimento individuais para ajudar o trabalho em paralelo, aumentando a agilidade do desenvolvimento. Implemente controles mais rigorosos nos ambientes ao se aproximar da produção para permitir que os desenvolvedores inovem. Use a infraestrutura como sistemas de gerenciamento de código e configuração para implantar ambientes que são configurados de maneira consistente com os controles presentes na produção para garantir que os sistemas operem conforme o esperado quando implantados. Quando os ambientes não estiverem em uso, desligue-os para evitar custos associados a recursos inativos (por exemplo, sistemas de desenvolvimento à noite e fins de semana). Implante ambientes equivalentes de produção ao carregar o teste para melhorar resultados válidos.

Recursos

Documentos relacionados:

- Agendador de instâncias na AWS
- O que é AWS CloudFormation?

OPS05-BP09 Fazer alterações frequentes, pequenas e reversíveis

Alterações frequentes, pequenas e reversíveis reduzem o escopo e o impacto de uma alteração. Quando usadas em conjunto com sistemas de gerenciamento de alterações, sistemas de gerenciamento de configuração e sistemas de compilação e entrega, alterações frequentes, pequenas e reversíveis reduzem o escopo e o impacto de uma mudança. Isso resulta em solução de problemas mais eficaz e correção mais rápida, com a opção de reverter alterações.

Práticas comuns que devem ser evitadas:

- Você implanta uma nova versão de sua aplicação trimestralmente com uma janela de alteração que significa que um serviço principal está desativado.
- Você frequentemente faz alterações no esquema do banco de dados sem rastrear as alterações nos sistemas de gerenciamento.
- Você realiza atualizações manuais no local, substituindo as instalações e configurações existentes e não tem um plano claro de reversão.

Benefícios de implantar esta prática recomendada: os esforços de desenvolvimento são mais rápidos com a implantação frequente de pequenas alterações. Quando as alterações são pequenas, é muito mais fácil identificar se elas têm consequências indesejadas e são mais fáceis de serem revertidas. Quando as alterações são reversíveis, há menos risco de implementar a alteração à medida que a recuperação é simplificada. O processo de mudança tem um risco reduzido e o impacto de uma alteração malsucedida é reduzido.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Use alterações frequentes, pequenas e reversíveis para reduzir o escopo e o impacto de uma mudança. Isso facilita a solução de problemas, ajuda a fazer uma correção mais rápida e oferece a opção de reverter uma alteração. Além disso, aumenta a taxa na qual você pode agregar valor aos negócios.

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP03 Usar sistemas de gerenciamento de configuração
- OPS05-BP04 Usar sistemas de gerenciamento de compilação e de implantação
- OPS06-BP04 Automatizar os testes e a reversão

Documentos relacionados:

- Implementar microsserviços na AWS
- Microservices: observabilidade

OPS05-BP10 Automatizar totalmente a integração e a implantação

Automatize a construção, a implantação e o teste da workload. Isso reduz os erros causados pelos processos manuais e reduz o esforço para implantar alterações.

Aplique metadados usando tags de recursos e AWS Resource Groups seguindo uma estratégia consistente de marcação com tags para identificar seus recursos. Marque com tags seus recursos de organização, contabilidade de custos, controles de acesso pensando na execução de atividades operacionais automatizadas.

Resultado desejado: os desenvolvedores usam ferramentas para entregar códigos e levá-los até a produção. Os desenvolvedores não precisam fazer login no AWS Management Console para fazer atualizações. Há uma trilha de auditoria completa de alterações e configurações, o que atende às necessidades de governança e conformidade. Os processos são repetíveis e padronizados entre as equipes. Os desenvolvedores podem se concentrar no desenvolvimento e na introdução de código, aumentando a produtividade.

Práticas comuns que devem ser evitadas:

- Na sexta-feira, você conclui a criação do novo código para a ramificação do recurso. Na segunda-feira, depois de executar os scripts de teste de qualidade de código e cada um dos scripts de teste de unidade, você registra seu código para a próxima versão agendada.
- Você tem a tarefa de codificar uma correção para um problema crítico que afeta um grande número de clientes em produção. Depois de testar a correção, você confirma o gerenciamento de alterações de e-mail e código para solicitar aprovação para implantá-lo na produção.
- Como desenvolvedor, você faz login no AWS Management Console para criar um novo ambiente de desenvolvimento usando métodos e sistemas que não são padrão.

Benefícios de implementar esta prática recomendada:: ao implementar sistemas automatizados de gerenciamento de criação e implantação, você reduz os erros causados por processos manuais e o esforço para implantar alterações, ajudando os membros da equipe a se concentrarem na entrega de valor para a empresa. Você aumenta a velocidade de entrega à medida que avança até a produção.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Você usa sistemas de gerenciamento de criação e implantação para rastrear e implementar alterações, reduzir erros causados por processos manuais e reduzir o nível de esforço. Automatize totalmente o pipeline de integração e implantação desde o check-in do código até a compilação, o teste, a implantação e a validação. Isso reduz o tempo de espera, aumenta a frequência de alterações, reduz o nível de esforço, aumenta a velocidade de entrada no mercado, resulta em maior produtividade e aumenta a segurança do seu código à medida que você o leva até a produção.

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP03 Usar sistemas de gerenciamento de configuração
- OPS05-BP04 Usar sistemas de gerenciamento de compilação e de implantação

Documentos relacionados:

- O que é AWS CodeBuild?
- O que é AWS CodeDeploy?

Vídeos relacionados:

- AWS re\:Invent 2022: Práticas recomendadas do AWS Well-Architected para DevOps na AWS

OPS 6. Como reduzir os riscos de implantação?

Adote abordagens que forneçam feedback rápido sobre a qualidade e alcancem recuperação rápida de alterações que não têm os resultados desejados. O uso dessas práticas reduz o impacto dos problemas introduzidos pela implantação de mudanças.

Práticas recomendadas

- OPS06-BP01 Preparar-se para alterações malsucedidas
- OPS06-BP02 Testar implantações
- OPS06-BP03 Utilizar estratégias de implantação seguras
- OPS06-BP04 Automatizar os testes e a reversão

OPS06-BP01 Preparar-se para alterações malsucedidas

Planeje reverter para um bom estado anterior ou realize reparos no ambiente de produção se a implantação causar um resultado indesejado. Ter uma política para estabelecer esse plano ajuda todas as equipes a desenvolver estratégias para se recuperar de alterações com falha. Alguns exemplos de estratégias são etapas de implantação e reversão, políticas de alteração, sinalizadores de atributos, isolamento de tráfego e mudança de tráfego. Uma única versão pode incluir várias alterações de componentes relacionadas. A estratégia deve fornecer a possibilidade de resistir ou se recuperar de uma falha de qualquer alteração de componente.

Resultado desejado: você preparou um plano de recuperação detalhado para a alteração, caso ela não tenha êxito. Além disso, você reduziu o tamanho da sua versão para minimizar o impacto potencial em outros componentes da workload. Como resultado, você reduziu o impacto nos negócios ao diminuir o possível tempo de inatividade decorrente de uma alteração malsucedida e aumentou a flexibilidade e a eficiência dos tempos de recuperação.

Práticas comuns que devem ser evitadas:

- Você executou uma implantação e sua aplicação se tornou instável, mas parece haver usuários ativos no sistema. Você precisa decidir se deseja reverter a alteração e afetar os usuários ativos ou esperar para reverter a alteração sabendo que, mesmo assim, os usuários podem ser afetados.
- Depois de fazer uma alteração de rotina, os novos ambientes ficam acessíveis, mas uma de suas sub-redes se tornou inacessível. Você precisa decidir se deseja reverter tudo ou tentar corrigir a sub-rede inacessível. Enquanto você estiver fazendo essa determinação, a sub-rede permanecerá inacessível.
- Seus sistemas não são arquitetados de uma forma que permita que sejam atualizados com versões menores. Como resultado, você tem dificuldade em reverter essas alterações em massa durante uma implantação com falha.
- Você não usa infraestrutura como código (IaC) e atualizações manuais foram feitas em sua infraestrutura que resultaram em uma configuração indesejada. Você não consegue rastrear e reverter com eficácia as alterações manuais.

- Como você não mediu o aumento da frequência das implantações, sua equipe não é incentivada a reduzir o tamanho das mudanças e melhorar seus planos de reversão para cada uma delas, gerando mais riscos e maiores taxas de falha.
- Você não mede a duração total de uma interrupção causada por alterações malsucedidas. A equipe não consegue priorizar e melhorar a eficácia do processo de implantação e do plano de recuperação.

Benefícios de implementar esta prática recomendada: ter um plano para se recuperar de mudanças malsucedidas minimiza o tempo médio de recuperação (MTTR) e reduz o impacto nos negócios.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

A adoção de uma política e prática documentadas e consistentes por parte das equipes de lançamento permitem que a organização planeje o que deve ocorrer se houver mudanças malsucedidas. A política deve permitir a correção em circunstâncias específicas. Seja qual for a situação, um plano de correção antecipada ou reversão deve ser bem documentado e testado antes da implantação na produção em tempo real, a fim de que o tempo necessário para reverter uma alteração seja minimizado.

Etapas de implementação

1. Documente as políticas que exigem que as equipes tenham planos efetivos para reverter as mudanças dentro de um período especificado.
   a. As políticas devem especificar quando uma situação de correção antecipada é permitida.
   b. Exija que um plano de reversão documentado seja acessível a todos os envolvidos.
   c. Especifique os requisitos de reversão (por exemplo, quando for constatado que foram implantadas alterações não autorizadas).
2. Analise o nível de impacto de todas as mudanças relacionadas a cada componente de uma workload.
   a. Permita que alterações repetíveis sejam padronizadas, modeladas e pré-autorizadas se seguirem um fluxo de trabalho consistente que imponha políticas de mudança.
   b. Reduza o impacto potencial de qualquer alteração diminuindo o tamanho dela para que a recuperação leve menos tempo e cause um impacto menor nos negócios.
   c. Garanta que os procedimentos de reversão revertam o código para um bom estado conhecido a fim de evitar incidentes sempre que possível.

3. Integre ferramentas e fluxos de trabalho para aplicar suas políticas de forma programática.
4. Torne os dados sobre as alterações visíveis para outros proprietários da workload a fim de melhorar a velocidade do diagnóstico de qualquer alteração malsucedida que não possa ser revertida.
    a. Avalie o sucesso dessa prática usando dados de mudança visíveis e identifique melhorias iterativas.
5. Use ferramentas de monitoramento para verificar o sucesso ou a falha de uma implantação a fim de acelerar a tomada de decisões sobre a reversão.
6. Meça a duração da interrupção durante uma alteração malsucedida para melhorar continuamente seus planos de recuperação.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS06-BP04 Automatizar os testes e a reversão

Documentos relacionados:

- AWS Builders Library | Garantir a segurança das reversões durante implantações
- Whitepaper da AWS | Gerenciamento de alterações na nuvem

Vídeos relacionados:

- re:Invent 2019 | A abordagem da Amazon a implantações de alta disponibilidade

OPS06-BP02 Testar implantações

Teste os procedimentos de lançamento na pré-produção usando a mesma configuração de implantação, controles de segurança, etapas e procedimentos da produção. Valide se todas as etapas implantadas foram concluídas conforme o esperado, como inspecionar arquivos, configurações e serviços. Teste mais detalhadamente todas as alterações com testes funcionais, de integração e de carga, além de qualquer monitoramento, como verificações de integridade. Ao fazer esses testes, você pode identificar problemas de implantação com antecedência, podendo planejá-los e mitigá-los antes da produção.

Você pode criar ambientes paralelos temporários para testar cada alteração. Automatize a implantação dos ambientes de teste usando a infraestrutura como código (IaC) para ajudar a reduzir a quantidade de trabalho envolvido e garantir estabilidade, consistência e entrega mais rápida de atributos.

Resultado desejado: a organização adota uma cultura de desenvolvimento orientada a testes que inclui testes de implantações. Isso garante que as equipes se concentrem em oferecer valor empresarial em vez de gerenciar lançamentos. As equipes são engajadas desde o início após a identificação dos riscos de implantação para determinar o curso apropriado da mitigação.

Práticas comuns que devem ser evitadas:

- Durante as versões de produção, implantações não testadas causam problemas frequentes que exigem soluções e encaminhamento.
- Sua versão contém infraestrutura como código (IaC) que atualiza os recursos existentes. Você não tem certeza se a IaC será executada com êxito ou causará impacto nos recursos.
- Você implanta um novo recurso na aplicação. Ele não funciona conforme o esperado e não há visibilidade até que o problema seja relatado pelos usuários afetados.
- Você atualiza seus certificados. Você instala acidentalmente os certificados nos componentes errados, o que não é detectado e afeta os visitantes do site porque não é possível estabelecer uma conexão segura.

Benefícios de implementar esta prática recomendada: testes extensivos na pré-produção dos procedimentos de implantação, considerando-se que as mudanças introduzidas por eles minimizam o impacto potencial na produção causado pelas etapas de implantação. Isso aumenta a confiança durante o lançamento para produção e minimiza o suporte operacional sem diminuir a velocidade das alterações que estão sendo entregues.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Testar seu processo de implantação é tão importante quanto testar as alterações resultantes da implantação. Isso poderá ser realizado testando-se suas etapas de implantação em um ambiente de pré-produção que se assemelhe o máximo possível à produção. Problemas comuns, como etapas de implantação incompletas ou incorretas, ou configurações incorretas, podem ser detectados como resultado antes da produção. Além disso, você pode testar suas etapas de recuperação.

Exemplo de cliente

Como parte do pipeline de integração e entrega contínuas (CI/CD), a AnyCompany Retail executa as etapas definidas necessárias para lançar atualizações de infraestrutura e software para seus clientes em um ambiente semelhante ao de produção. O pipeline é composto por pré-verificações para detectar desvios (detecção de alterações nos recursos executados fora da IaC) nos recursos antes da implantação, bem como validar as ações que a IaC realiza após seu início. Ele valida as etapas de implantação, como verificar se determinados arquivos e configurações estão em vigor e se os serviços estão em execução e respondendo corretamente às verificações de integridade no host local antes de serem registrados novamente no balanceador de carga. Além disso, todas as alterações sinalizam vários testes automatizados, como testes funcionais e de segurança, regressão, integração e carga.

Etapas de implementação

1. Execute verificações de pré-instalação para espelhar o ambiente de pré-produção na produção.
    a. Use a detecção de desvios para detectar quando os recursos foram alterados fora do AWS CloudFormation.
    b. Use conjuntos de alterações para validar se a intenção da atualização da pilha corresponde às ações que o AWS CloudFormation realiza quando o conjunto de alterações é iniciado.
2. Isso aciona uma etapa de aprovação manual ao AWS CodePipeline para autorizar a implantação no ambiente de pré-produção.
3. Use configurações de implantação, como arquivos AppSpec do AWS CodeDeploy, para definir as etapas de implantação e validação.
4. Quando aplicável, integre o AWS CodeDeploy a outros serviços da AWS ou integre o AWS CodeDeploy a produtos e serviços de parceiros.
5. Monitore implantações usando o Amazon CloudWatch, o AWS CloudTrail e as notificações de eventos do Amazon SNS.
6. Execute testes automatizados pós-implantação, incluindo testes funcionais, de segurança, regressão, integração e carga.
7. Solucione problemas de implantação.
8. A validação bem-sucedida das etapas anteriores deve iniciar um fluxo de trabalho de aprovação manual para autorizar a implantação na produção.

Nível de esforço do plano de implementação: Alto

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP02 Testar e validar alterações

Documentos relacionados:

- AWS Builders' Library | Automatizar implantações autônomas e seguras | Testar implantações
- Whitepaper da AWS | Praticar a integração e entrega contínuas na AWS
- A história da Apollo: o mecanismo de implantação da Amazon
- Como testar e depurar o AWS CodeDeploy localmente antes de enviar seu código
- Integrar testes de conectividade de rede com implantação da infraestrutura

Vídeos relacionados:

- re:Invent 2020 | Testar software e sistemas na Amazon

Exemplos relacionados:

- Tutorial | Implantar um serviço do Amazon ECS com um teste de validação

OPS06-BP03 Utilizar estratégias de implantação seguras

Implantações seguras de produção controlam o fluxo de mudanças benéficas com o objetivo de minimizar qualquer impacto percebido dessas alterações para os clientes. Os controles de segurança fornecem mecanismos de inspeção para validar os resultados desejados e limitar o escopo do impacto dos defeitos introduzidos pelas alterações ou das falhas de implantação. As implementações seguras podem incluir estratégias como sinalizadores de atributos e implantações one-box, contínuas (versões canário), imutáveis, de divisão de tráfego e azuis/verdes.

Resultado desejado: sua organização usa um sistema de integração e entrega contínuas (CI/CD) que fornece recursos para automatizar implementações seguras. As equipes devem usar estratégias apropriadas de implantação seguras.

Práticas comuns que devem ser evitadas:

- Você implanta uma alteração malsucedida em toda a produção de uma só vez. Como resultado, todos os clientes são afetados simultaneamente.
- Um defeito introduzido em uma implantação simultânea em todos os sistemas requer um lançamento de emergência. A correção para todos os clientes leva vários dias.
- O gerenciamento da versão de produção requer planejamento e participação de várias equipes. Isso restringe sua capacidade de atualizar atributos com frequência para seus clientes.
- Você executa uma implantação mutável modificando os sistemas existentes. Depois de descobrir que a alteração não foi bem-sucedida, você será forçado a modificar os sistemas novamente para restaurar a versão antiga, aumentando o seu tempo de recuperação.

Benefícios de implementar esta prática recomendada: as implantações automatizadas equilibram a velocidade das implementações com a entrega consistente de mudanças benéficas para os clientes. Limitar o impacto evita falhas de implantação dispendiosas e maximiza a capacidade das equipes de responder às falhas de forma eficiente.

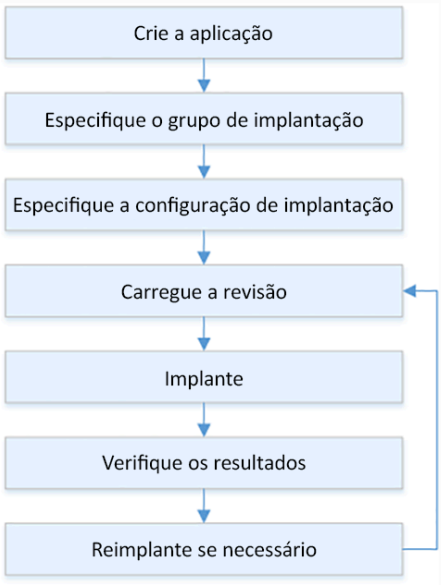
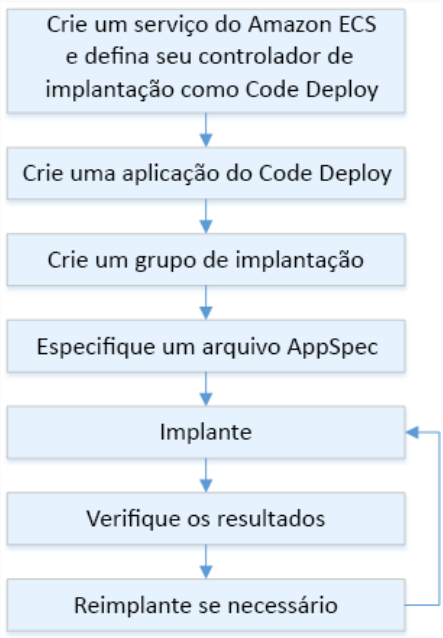
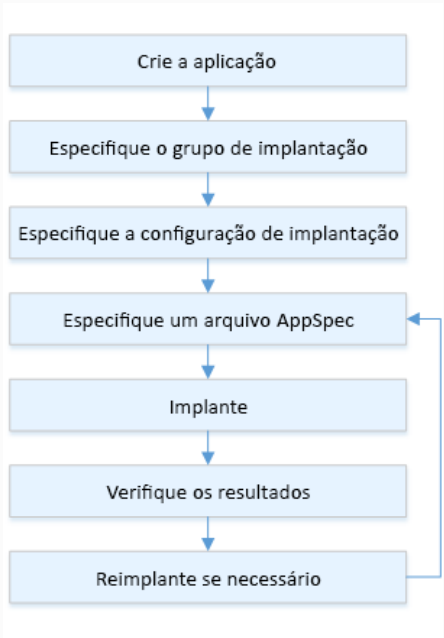
Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Falhas na entrega contínua podem levar à redução da disponibilidade do serviço e a uma experiência ruim para o cliente. Para maximizar a taxa de implantações bem-sucedidas, implemente controles de segurança no processo de lançamento de ponta a ponta para minimizar os erros de implantação e eliminar as falhas.

Exemplo de cliente

A AnyCompany Retail tem a missão de alcançar implantações com tempo de inatividade entre mínimo e zero, ou seja, sem impacto perceptível para seus usuários durante as implantações. Para fazer isso, a empresa estabeleceu padrões de implantação (consulte o diagrama de fluxo de trabalho a seguir), como implantações azuis/verdes e contínuas. Todas as equipes adotam um ou mais desses padrões no pipeline de CI/CD.

Fluxo de trabalho do CodeDeploy para Amazon EC2	Fluxo de trabalho do CodeDeploy para Amazon ECS	Fluxo de trabalho do CodeDeploy para Lambda
		

Etapas de implementação

1. Use um fluxo de trabalho de aprovação para iniciar a sequência das etapas de implantação na promoção para implantação.
2. Use um sistema de implantação automatizado, como o AWS CodeDeploy. As as opções de implantação do AWS CodeDeploy incluem implantações no local para EC2/on-premises e implantações azuis/verdes para EC2/on-premises, AWS Lambda e Amazon ECS (consulte o diagrama do fluxo de trabalho anterior).
   a. Quando aplicável, integre o AWS CodeDeploy a outros serviços da AWS ou integre o AWS CodeDeploy a produtos e serviços de parceiros.
3. Use implantações azuis/verdes para bancos de dados como Amazon Aurora e Amazon RDS.
4. Monitore implantações usando o Amazon CloudWatch, o AWS CloudTrail e as notificações de eventos do Amazon Simple Notification Service (Amazon SNS).
5. Realize testes automatizados pós-implantação, incluindo testes funcionais, de segurança, regressão, integração e testes de carga.
6. Solucione problemas de implantação.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP02 Testar e validar alterações
- OPS05-BP09 Fazer alterações frequentes, pequenas e reversíveis
- OPS05-BP10 Automatizar totalmente a integração e a implantação

Documentos relacionados:

- AWS Builders' Library | Automatizar implantações seguras e sem intervenção manual | Implantações de produção
- AWS Builders Library | Meu pipeline de CI/CD é meu capitão de lançamentos | Lançamentos de produção seguros e automáticos
- Whitepaper da AWS | Praticar integração e entrega contínuas na AWS | Métodos de implantação
- Guia do usuário do AWS CodeDeploy
- Trabalhar com configurações de implantação no AWS CodeDeploy
- Configurar uma implantação de versão canário do API Gateway
- Tipos de implantação do Amazon ECS
- Implantações azuis/verdes totalmente gerenciadas no Amazon Aurora e no Amazon RDS
- Implantações azuis/verdes com o AWS Elastic Beanstalk

Vídeos relacionados:

- re:Invent 2020 | Sem intervenção manual: como automatizar os pipelines de entrega contínua na Amazon
- re:Invent 2019 | A abordagem da Amazon a implantações de alta disponibilidade

Exemplos relacionados:

- Testar uma implantação azul/verde de exemplo no AWS CodeDeploy
- Workshop | Criar pipelines de CI/CD para implantações canário do Lambda usando o AWS CDK

- Workshop | Implantações canário e azuis/verdes para EKS e ECS
- Workshop | Criar um pipeline de CI/CD entre contas

OPS06-BP04 Automatizar os testes e a reversão

Para aumentar a velocidade, a confiabilidade e a confiança do seu processo de implantação, tenha uma estratégia para testes automatizados e recursos de reversão em ambientes de pré-produção e produção. Automatize os testes ao implantar na produção para simular interações entre humanos e sistemas que verifiquem as alterações que estão sendo implantadas. Automatize a reversão para voltar rapidamente a um estado anterior em boas condições. A reversão deve ser iniciada automaticamente em condições predefinidas, como quando o resultado desejado da alteração não é alcançado ou quando o teste automatizado falha. A automação dessas duas atividades melhora a taxa de sucesso das implantações, minimiza o tempo de recuperação e reduz o impacto potencial nos negócios.

Resultado desejado: os testes automatizados e as estratégias de reversão são integrados ao pipeline de integração e entrega contínuas (CI/CD). O monitoramento é capaz de validar seus critérios de sucesso e iniciar a reversão automática em caso de falha. Isso minimiza qualquer impacto para usuários finais e clientes. Por exemplo, quando todos os resultados do teste são satisfatórios, você promove seu código no ambiente de produção em que o teste de regressão automatizado é iniciado, utilizando os mesmos casos de teste. Se os resultados do teste de regressão não corresponderem às expectativas, a reversão automática será iniciada no fluxo de trabalho do pipeline.

Práticas comuns que devem ser evitadas:

- Seus sistemas não são arquitetados de uma forma que permita que sejam atualizados com versões menores. Como resultado, você tem dificuldade em reverter essas alterações em massa durante uma implantação com falha.
- O processo de implantação consiste em uma série de etapas manuais. Depois de implantar as alterações na workload, você inicia os testes pós-implantação. Após o teste, você percebe que a workload está inoperante e os clientes estão desconectados. Em seguida, você começa a reverter para a versão anterior. Todas essas etapas manuais atrasam a recuperação geral do sistema e causam um impacto prolongado para os clientes.
- Você dedicou tempo para desenvolver casos de teste automatizados para funcionalidades que não são usadas com frequência na aplicação, minimizando o retorno sobre o investimento no recurso de teste automatizado.

- Sua versão é composta de atualizações de aplicações, infraestrutura, patches e configuração que são independentes umas das outras. No entanto, você tem um único pipeline de CI/CD que fornece todas as alterações de uma só vez. Uma falha em um componente força você a reverter todas as alterações, tornando a reversão complexa e ineficiente.
- A equipe conclui o trabalho de codificação no primeiro sprint e inicia o trabalho no segundo sprint, mas seu plano não incluiu testes até o terceiro sprint. Como resultado, os testes automatizados revelaram defeitos do primeiro sprint que precisavam ser resolvidos antes que o teste dos resultados do segundo sprint pudesse ser iniciado, adiando todo o lançamento e desvalorizando seus testes automatizados.
- Seus casos de teste de regressão automatizados para a versão de produção estão completos, mas você não está monitorando a integridade da workload. Como você não tem visibilidade sobre se o serviço foi reiniciado, você não tem certeza se a reversão é necessária ou se ela já ocorreu.

Benefícios de implementar esta prática recomendada: o teste automatizado aumenta a transparência do processo de teste e a capacidade de abranger mais atributos em um período mais curto. Ao testar e validar as mudanças na produção, é possível identificar problemas imediatamente. A melhoria na consistência com ferramentas de teste automatizadas permite uma melhor detecção de defeitos. Ao reverter automaticamente para a versão anterior, o impacto sobre seus clientes é minimizado. A reversão automatizada acaba inspirando mais confiança em seus recursos de implantação ao reduzir o impacto nos negócios. No geral, esses recursos reduzem o tempo de entrega e, ao mesmo tempo, garantem a qualidade.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Automatize os testes dos ambientes implantados para confirmar os resultados desejados mais rapidamente. Automatize a reversão para um bom estado anterior conhecido quando os resultados predefinidos não forem alcançados, para minimizar o tempo de recuperação e reduzir os erros causados por processos manuais. Integre ferramentas de teste com seu fluxo de trabalho de pipeline para testar e minimizar as entradas manuais de forma consistente. Priorize a automação de casos de teste, como aqueles que mitigam os maiores riscos e precisam ser testados com frequência a cada alteração. Além disso, automatize a reversão com base em condições específicas predefinidas no plano de teste.

Etapas de implementação

1. Estabeleça um ciclo de vida de teste para o ciclo de vida de desenvolvimento que defina cada estágio do processo de teste, desde o planejamento dos requisitos até o desenvolvimento do caso de teste, a configuração da ferramenta, o teste automatizado e o encerramento do caso de teste.
   a. Crie uma abordagem de teste específica para workloads com base em sua estratégia geral de teste.
   b. Considere uma estratégia de teste contínuo, quando apropriado, durante o ciclo de vida do desenvolvimento.
2. Selecione ferramentas automatizadas para testes e reversões com base em seus requisitos de negócios e investimentos em pipeline.
3. Decida quais casos de teste você deseja automatizar e quais deverão ser executados manualmente. Eles podem ser definidos com base na prioridade do valor comercial do atributo que está sendo testado. Alinhe todos os membros da equipe a esse plano e verifique a responsabilidade pela realização de testes manuais.
   a. Aplique recursos de teste automatizados a casos de teste específicos que façam sentido para automação, como casos repetíveis ou executados com frequência, aqueles que exigem tarefas repetitivas ou aqueles que são necessários em várias configurações.
   b. Defina scripts de automação de testes, bem como os critérios de sucesso na ferramenta de automação, para que a automação contínua do fluxo de trabalho possa ser iniciada quando casos específicos falharem.
   c. Defina critérios de falha específicos para a reversão automatizada.
4. Priorize a automação de testes para gerar resultados consistentes com o desenvolvimento completo de casos de teste em que a complexidade e a interação humana têm um risco maior de falha.
5. Integre as ferramentas automatizadas de teste e reversão no pipeline de CI/CD.
   a. Desenvolva critérios claros de sucesso para as alterações.
   b. Monitore e observe para detectar esses critérios e reverter automaticamente as alterações quando critérios específicos de reversão forem atendidos.
6. Execute diferentes tipos de teste de produção automatizados, como:
   a. Teste A/B para mostrar resultados em comparação com a versão atual entre dois grupos de teste de usuários.
   b. Teste canário, que permite implantar a alteração em um subconjunto de usuários antes de lançá-la para todos.

   c. Teste de sinalização de atributos, que permite que a sinalização de um único atributo da nova versão seja ativada e desativada de fora da aplicação para que cada novo atributo possa ser validado individualmente.
   d. Teste de regressão para verificar novas funcionalidades com componentes inter-relacionados existentes.
7. Monitore os aspectos operacionais da aplicação, das transações e das interações com outras aplicações e componentes. Desenvolva relatórios para mostrar o sucesso das alterações por workload e identificar quais partes da automação e do fluxo de trabalho podem ser otimizadas ainda mais.
   a. Desenvolva relatórios de resultados de testes que ajudem você a tomar decisões rápidas sobre se os procedimentos de reversão devem ou não ser invocados.
   b. Implemente uma estratégia que permita a reversão automatizada com base em condições de falha predefinidas que resultam de um ou mais de seus métodos de teste.
8. Desenvolva seus casos de teste automatizados para permitir a reutilização em futuras alterações repetíveis.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS06-BP01 Preparar-se para alterações malsucedidas
- OPS06-BP02 Testar implantações

Documentos relacionados:

- AWS Builders Library | Garantir a segurança das reversões durante implantações
- Reimplantar e reverter uma implantação com a AWS CodeDeploy
- Oito práticas recomendadas para automatizar suas implantações com o AWS CloudFormation

Exemplos relacionados:

- Teste de interface do usuário sem servidor usando Selenium, AWS Lambda, AWS Fargate e as Ferramentas de desenvolvedor da AWS

Vídeos relacionados:

- re:Invent 2020 | Sem intervenção manual: como automatizar os pipelines de entrega contínua na Amazon
- re:Invent 2019 | A abordagem da Amazon a implantações de alta disponibilidade

OPS 7. Como saber se está tudo pronto para oferecer suporte a uma workload?

Avalie a prontidão operacional de sua workload, processos/procedimentos e pessoal para entender os riscos operacionais relacionados.

Práticas recomendadas

- OPS07-BP01 Garantir a capacidade da equipe
- OPS07-BP02 Garantir uma revisão consistente da prontidão operacional
- OPS07-BP03 Usar runbooks para realizar procedimentos
- OPS07-BP04 Usar playbooks para investigar problemas
- OPS07-BP05 Tomar decisões embasadas para implantar sistemas e alterações
- OPS07-BP06 Criar planos de suporte para workloads de produção

OPS07-BP01 Garantir a capacidade da equipe

Adote um mecanismo para validar que você tem o número adequado de funcionários treinados para fornecer suporte à workload. Eles devem ter treinamento para a plataforma e os serviços que compõem sua workload. Forneça a eles o conhecimento necessário para operar a workload. É necessário ter o número suficiente de funcionários treinados para oferecer suporte à operação da workload e solucionar os incidentes que ocorrerem. Tenha funcionários suficientes para que seja possível fazer uma rotação durante plantões e férias a fim de evitar a exaustão.

Resultado desejado:

- Há um número suficiente de funcionários treinados para oferecer suporte à workload quando ela estiver disponível.
- Você fornece treinamento para seus funcionários sobre software e serviços que compõem a workload.

Práticas comuns que devem ser evitadas:

- Implantar uma workload sem membros da equipe treinados para operar a plataforma e os serviços em uso.
- Não ter funcionários suficientes para oferecer suporte à rotaçõ~es de plantão ou folga de funcionários.

Benefícios de implementar esta prática recomendada:

- Ter membros da equipe qualificados possibilita o suporte eficaz da sua workload.
- Com um número suficiente de membros na equipe, é possível dar conta da workload e das rotações de plantão, reduzindo o risco de exaustão.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Valide se há um número suficiente de funcionários treinados para oferecer suporte à workload. Verifique se você tem membros da equipe suficientes para cobrir as atividades operacionais normais, incluindo rotações de plantão.

Exemplo de cliente

A AnyCompany Retail garante que as equipes que oferecem suporte à workload estejam completas e treinadas. Há engenheiros suficientes para oferecer suporte a uma rotação de plantão. Os funcionários têm treinamento referente ao software e à plataforma na qual a workload é criada e são incentivados a obter certificações. Há funcionários suficientes para que as pessoas possam tirar folgas enquanto mantêm o suporte à workload e à rotação de plantões.

Etapas de implementação

1. Atribua um número adequado de funcionários para operar e fornecer suporte à workload, incluindo tarefas de plantão.
2. Treine seus funcionários referente ao software e às plataformas que compõem a workload.
   a. A AWS Training and Certification oferece uma biblioteca de cursos sobre a AWS. Cursos pagos e gratuitos, online e presenciais, estão disponíveis.
   b. A AWS organiza eventos e webinars nos quais você aprende com especialistas da AWS.
3. Avalie regularmente o tamanho e as habilidades da equipe à medida que as condições operacionais e a workload mudam. Ajuste o tamanho e as habilidades da equipe para corresponderem aos requisitos operacionais.

Nível de esforço do plano de implementação: Alto. Contratar e treinar uma equipe para fornecer suporte a uma workload pode exigir um esforço significativo, mas traz benefícios substanciais de longo prazo.

Recursos

Práticas recomendadas relacionadas:

- OPS11-BP04 Gerenciar o conhecimento: os membros da equipe devem ter as informações necessárias para operar e fornecer suporte à workload. O gerenciamento de conhecimento é fundamental para isso.

Documentos relacionados:

- Eventos e webinars da AWS
- AWS Training and Certification

OPS07-BP02 Garantir uma revisão consistente da prontidão operacional

Use revisões de prontidão operacional (ORRs) para validar que você pode operar sua workload. A ORR é um mecanismo desenvolvido na Amazon para validar que as equipes podem operar as workloads com segurança. Uma ORR é um processo de análise e inspeção que usa uma lista de verificação de requisitos. Uma ORR é uma experiência de autoatendimento que as equipes usam para certificar suas workloads. As ORRs incluem práticas recomendadas de lições aprendidas de nossos anos de experiência na criação de software.

Uma lista de verificação de ORR é composta de recomendações de arquitetura, processo operacional, gerenciamento de evento e qualidade de lançamento. Nosso processo de Correção de erros (CoE) é um motivador principal desses itens. Sua própria análise pós-incidente deve impulsionar a evolução de sua própria ORR. Uma ORR não é apenas sobre seguir as práticas recomendadas, mas evitar a recorrência de eventos que você já viu. Por fim, os requisitos de segurança, governança e conformidade também podem ser incluídos em uma ORR.

Execute ORRs antes do lançamento de uma workload para disponibilidade geral e por todo o ciclo de vida de desenvolvimento do software. A execução da ORR antes do lançamento aumenta a capacidade de operar a workload com segurança. Execute a ORR periodicamente na workload para identificar qualquer desvio das práticas recomendadas. Você pode ter listas de verificação da ORR para o lançamento de outros serviços e ORRs para avaliações periódicas. Isso ajuda a manter

você em dia com as novas práticas recomendadas que surgem e incorporar as lições aprendidas da análise pós-incidente. À medida que seu uso da nuvem amadurece, é possível criar requisitos de ORR em sua arquitetura como padrões.

Resultado desejado: você tem uma lista de verificação da ORR com as práticas recomendadas para sua organização. As ORRs são realizadas antes do lançamento das workloads. As ORRs são executadas periodicamente ao longo do ciclo de vida da workload.

Práticas comuns que devem ser evitadas:

- Você lança uma workload sem saber se pode operá-la.
- Os requisitos de governança e segurança não estão incluídos na certificação de uma workload para o lançamento.
- As workloads não são reavaliadas periodicamente.
- As workloads são lançadas sem a aplicação dos procedimentos exigidos.
- Você vê a repetição das mesmas falhas da causa-raiz em várias workloads.

Benefícios de implementar esta prática recomendada:

- Suas workloads incluem práticas recomendadas de arquitetura, processo e gerenciamento.
- As lições aprendidas são incorporadas em seu processo de ORR.
- Os procedimentos exigidos estão em vigor no lançamento das workloads.
- As ORRs são executadas durante todo o ciclo de vida do software das workloads.

Nível de risco se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Uma ORR é composta por dois elementos: um processo e uma lista de verificação. O processo da ORR deve ser adotado pela organização e ter o apoio de um patrocinador executivo. No mínimo, as ORRs devem ser realizadas antes do lançamento da workload para disponibilidade geral. Execute a ORR ao longo de todo o ciclo de vida de desenvolvimento do software para mantê-la atualizada com as práticas recomendadas ou os novos requisitos. A lista de verificação da ORR deve incluir itens de configuração, requisitos de segurança e governança e práticas recomendadas de sua organização. Com o tempo, você pode usar serviços como AWS Config, AWS Security Hub e AWS Control Tower Guardrails para criar as práticas recomendadas do ORR em grades de proteção para a detecção automática de práticas recomendadas.

Exemplo de cliente

Depois de vários incidentes na produção, a AnyCompany Retail decidiu implementar um processo de ORR. Ela criou uma lista de verificação composta de práticas recomendadas, requisitos de governança e conformidade e lições aprendidas de interrupções. As novas workloads passam pelo processo de ORR antes do lançamento. Uma ORR é realizada anualmente para cada workload com um subconjunto de práticas recomendadas para incorporar novas práticas recomendadas e requisitos que são adicionados à lista de verificação da ORR. A AnyCompany Retail usava o AWS Config para detectar algumas das práticas recomendadas, acelerando o processo de ORR.

Etapas de implementação

Para saber mais sobre ORRs, leia o whitepaper Revisões de prontidão operacional (ORR). Ele fornece informações detalhadas sobre o histórico do processo de ORR, como criar sua própria prática de ORR e como desenvolver sua lista de verificação da ORR. As etapas a seguir são uma versão resumida desse documento. Para uma compreensão aprofundada do que são as ORRs e de como criar sua própria revisão, recomendamos a leitura desse whitepaper.

1. Reúna as principais partes interessadas, incluindo os representantes de segurança, operações e desenvolvimento.
2. Peça para cada parte interessada fornecer pelo menos um requisito. Para a primeira iteração, tente limitar o número de itens para trinta ou menos.
    - O Apêndice B: Perguntas de exemplo sobre ORR do whitepaper Revisões de prontidão operacional (ORR) contém exemplos de perguntas que você pode usar para começar.
3. Reúna seus requisitos em uma planilha.
    - Você pode usar lentes personalizadas no AWS Well-Architected Tool para desenvolver sua ORR e compartilhá-la entre suas contas e sua organização da AWS.
4. Identifique uma workload na qual realizar a ORR. O ideal seria em uma workload em pré-lançamento ou uma workload interna.
5. Execute a lista de verificação completa da ORR e anote as descobertas feitas. As descobertas poderão não ser corretas caso uma mitigação esteja ocorrendo. Para descobertas que não tenham uma mitigação, acrescente-as à sua lista de pendências e implemente-as antes do lançamento.
6. Continue a adicionar práticas recomendadas e requisitos à sua lista de verificação de ORR ao longo do tempo.

Os clientes do AWS Support com Enterprise Support podem solicitar o workshop Revisões de prontidão operacional a seus gerentes técnicos de conta. O workshop é uma sessão de trabalho retroativo interativa que permite desenvolver sua própria lista de verificação de ORR.

Nível de esforço do plano de implementação: Alto. Adotar uma prática de ORR em sua organização exige a adesão de um patrocinador executivo e das partes interessadas. Crie e atualize a lista de verificação com as opiniões de toda a sua organização.

Recursos

Práticas recomendadas relacionadas:

- OPS01-BP03 Avaliar os requisitos de governança: os requisitos de governança são uma opção natural para uma lista de verificação de ORR.
- OPS01-BP04 Avaliar os requisitos de conformidade: os requisitos de conformidade algumas vezes são incluídos em uma lista de verificação de ORR. Em outras, eles constituem um processo separado.
- OPS03-BP07 Fornecer recursos adequados às equipes: a capacidade da equipe é uma boa candidata para um requisito de ORR.
- OPS06-BP01 Preparar-se para alterações malsucedidas: um plano de reversão ou avanço deve ser estabelecido antes do lançamento da workload.
- OPS07-BP01 Garantir a capacidade da equipe: para acomodar uma workload, você deve ter o pessoal necessário.
- SEC01-BP03 Identificar e validar objetivos de controle: os objetivos de controle de segurança são excelentes requisitos de ORR.
- REL13-BP01 Definir objetivos de recuperação tempo de inatividade e perda de dados: planos de recuperação de desastres são um bom requisito de ORR.
- COST02-BP01 Desenvolver políticas com base nos requisitos da sua organização: políticas de gerenciamento de custos podem ser incluídas em sua lista de verificação de ORR.

Documentos relacionados:

- AWS Control Tower: barreiras de proteção no AWS Control Tower
- AWS Well-Architected Tool: perspectivas personalizadas
- Modelo de revisão de prontidão operacional, por Adrian Hornsby
- Whitepaper Revisões de prontidão operacional (ORR)

Vídeos relacionados:

- AWS Supports You | Criar uma Revisão de prontidão operacional (ORR) eficaz

Exemplos relacionados:

- Exemplo da perspectiva da Revisão de prontidão operacional (ORR)

Serviços relacionados:

- AWS Config
- AWS Control Tower
- AWS Security Hub
- AWS Well-Architected Tool

OPS07-BP03 Usar runbooks para realizar procedimentos

Um runbook é um processo documentado para alcançar um resultado específico. Os runbooks consistem em uma série de etapas seguidas por alguém para realizar alguma coisa. Os runbooks são usados em operações desde os primórdios da aviação. Nas operações na nuvem, usamos runbooks para reduzir riscos e alcançar os resultados desejados. Simplificando ao máximo, um runbook é uma lista de verificação para concluir uma tarefa.

Os runbooks são fundamentais para a operação de uma workload. Da integração de um novo membro da equipe à implantação de um lançamento importante, os runbooks são os processos codificados que fornecem resultados consistentes independentemente de quem os usa. Os runbooks devem ser publicados em um local central e ser atualizados à medida que o processo evolui, uma vez que a atualização dos runbooks é um aspecto fundamental de um processo de gerenciamento de mudanças. Eles também devem incluir orientação sobre tratamento de erros, ferramentas, permissões, exceções e encaminhamentos em caso de problema.

À medida que sua organização amadurece, comece a automatizar os runbooks. Comece com runbooks que sejam curtos e usados com frequência. Use linguagens de scripts para automatizar as etapas ou facilitar a realização delas. À medida que você automatiza os primeiros runbooks, você dedicará tempo à automação de runbooks mais complexos. Com o tempo, a maioria dos seus runbooks deverá ter algum nível de automação.

Resultado desejado: sua equipe tem um conjunto de guias detalhados para realizar tarefas de workload. Os runbooks contêm o resultado desejado, as ferramentas e as permissões necessárias e instruções para tratamento de erros. Eles são armazenados em um local central (sistema de controle de versão) e atualizados com frequência. Por exemplo, seus runbooks fornecem recursos para que suas equipes monitorem, se comuniquem e reajam a eventos do AWS Health para contas críticas durante alarmes de aplicações, problemas operacionais e eventos planejados do ciclo de vida.

Práticas comuns que devem ser evitadas:

- Depender da memória para concluir cada etapa de um processo.
- Implantar mudanças manualmente sem uma lista de verificação.
- Diferentes membros da equipe realizando o mesmo processo, mas com etapas ou resultados diferentes.
- Deixar que os runbooks fiquem desatualizados em relação às alterações no sistema e à automação.

Benefícios de implementar esta prática recomendada:

- Redução das taxas de erros em tarefas manuais.
- Operações realizadas de maneira consistente.
- Novos membros da equipe podem começar a realizar as tarefas mais cedo.
- Os runbooks podem ser automatizados para reduzir o esforço.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Os runbooks podem assumir diversos formatos dependendo do nível de maturidade da sua organização. No mínimo, devem consistir em um documento de texto detalhado. O resultado desejado deve estar claramente identificado. Documente claramente as permissões ou ferramentas especiais necessárias. Forneça orientação detalhada sobre tratamento de erros e encaminhamentos em caso de problema. Liste o proprietário do runbook e publique-o em um local central. Depois que o runbook estiver documentado, valide-o pedindo que outro membro da equipe o execute. À medida que os procedimentos evoluem, atualize os runbooks de acordo com seu processo de gerenciamento de mudanças.

Os runbooks em texto devem ser automatizados à medida que a organização amadurece. Ao usar serviços como o AWS Systems Manager Automations, é possível transformar texto plano em automações que podem ser executadas na workload. Essas automações podem ser executadas em resposta a eventos, reduzindo a sobrecarga operacional de manutenção da workload. AWS O Systems Manager Automation também fornece uma experiência de design visual com código simples para criar runbooks de automação com mais facilidade.

Exemplo de cliente

A AnyCompany Retail precisa realizar atualizações no esquema de banco de dados durante as implantações de software. A equipe de operações na nuvem trabalhou com a equipe de administração do banco de dados para criar um runbook para implantação manual dessas alterações. O runbook lista cada etapa do processo em um formato de lista de verificação. Ele inclui uma seção sobre tratamento de erros em caso de problema. A equipe de operações na nuvem publicou o runbook na wiki interna junto com outros runbooks. Ela planeja automatizar o runbook em um sprint futuro.

Etapas de implementação

Se você não tem um repositório de documentos, um repositório de controle de versão é um ótimo lugar para começar a criar a biblioteca de runbooks. Você pode criar runbooks usando Markdown. Disponibilizamos um modelo de runbook que pode ser usado para começar a criar runbooks.

```
# Runbook Title
## Runbook Info
| Runbook ID | Description | Tools Used | Special Permissions | Runbook Author | Last
 Updated | Escalation POC |
|-------|-------|-------|-------|-------|-------|-------|
| RUN001 | What is this runbook for? What is the desired outcome? | Tools | Permissions
 | Your Name | 2022-09-21 | Escalation Name |
## Steps
1. Step one
2. Step two
```

1. Se você não tiver um repositório de documentação ou uma wiki, crie um repositório de controle de versão no sistema de controle de versão.
2. Identifique um processo que não tenha um runbook. Um processo ideal é um que seja realizado quase regularmente, que tenha poucas etapas e que tenha falhas de baixo impacto.
3. No repositório de documentos, crie um rascunho de documento em Markdown usando o modelo. Preencha o Título do runbook e os campos obrigatórios em Informações do runbook.

4. Começando com a primeira etapa, preencha a parte Etapas do runbook.
5. Dê o runbook para um membro da equipe. Peça que ele use o runbook para validar as etapas. Se algo estiver faltando ou não estiver claro, atualize o runbook.
6. Disponibilize o runbook em seu armazenamento interno de documentos. Depois, informe sua equipe e outras partes interessadas.
7. Com o passar do tempo, você terá uma biblioteca de runbooks. À medida que essa biblioteca cresce, comece a trabalhar na automatização dos runbooks.

Nível de esforço do plano de implementação: Baixo. O padrão mínimo para um runbook é um guia de texto detalhado. A automatização dos runbooks pode aumentar o esforço de implementação.

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP02 Processos e procedimentos com proprietários identificados
- OPS07-BP04 Usar playbooks para investigar problemas
- OPS10-BP01 Usar um processo para gerenciamento de eventos, incidentes e problemas
- OPS10-BP02 Adotar um processo por alerta
- OPS11-BP04 Gerenciar o conhecimento

Documentos relacionados:

- AWS Well-Architected Framework: conceitos: desenvolvimento de runbooks
- Como alcançar excelência operacional usando playbooks e runbooks automatizados
- AWS Systems Manager: trabalhar com runbooks
- Playbook de migração para grandes migrações da AWS – Tarefa 4: como melhorar runbooks de migração
- Como usar runbooks do AWS Systems Manager Automation para resolver tarefas operacionais

Vídeos relacionados:

- AWS re:Invent 2019: Guia de faça você mesmo para runbooks, relatórios de incidentes e resposta a incidentes

- Como automatizar as operações de TI na AWS | Amazon Web Services
- Integrar scripts ao AWS Systems Manager

Exemplos relacionados:

- Laboratórios do Well-Architected: Automatização de operações com playbooks e runbooks
- Publicação no blog da AWS: Criar uma prática de automação de nuvem para excelência operacional: práticas recomendadas do AWS Managed Services
- AWS Systems Manager: orientações sobre automação
- AWS Systems Manager: runbook para restauração de um volume raiz usando o snapshot mais recente
- Criar um runbook de resposta a incidentes da AWS usando cadernos Jupyter e o CloudTrail Lake
- Gitlab: runbooks
- Rubix: uma biblioteca de Python para criação de runbooks em cadernos Jupyter
- Usar o Document Builder para criar um runbook personalizado

Serviços relacionados:

- AWS Systems Manager Automation

OPS07-BP04 Usar playbooks para investigar problemas

Os playbooks são guias detalhados usados para investigar incidentes. Quando incidentes ocorrem, os playbooks são usados para investigar, definir o escopo do impacto e identificar a causa-raiz. Os playbooks são usados em diversos cenários, desde falhas em implantações até incidentes de segurança. Em muitos casos, os playbooks identificam a causa-raiz que um runbook costuma mitigar. Os playbooks são essenciais aos planos de resposta a incidentes de sua organização.

Um bom playbook abrange vários aspectos importantes. Ele guia o usuário, detalhadamente, ao longo do processo de descoberta. Considerando várias perspectivas, quais etapas devem ser seguidas para diagnosticar um incidente? Defina claramente no playbook se ferramentas especiais ou permissões elevadas são necessárias. Ter um plano de comunicação para atualizar as partes interessadas sobre o status da investigação é essencial. Em situações em que a causa-raiz ainda não foi identificada, o playbook deve ter um plano de escalação. Se a causa-raiz tiver sido identificada, o playbook deverá indicar um runbook que descreva como resolvê-la. Os playbooks

devem ser armazenados em um local central e atualizados com frequência. Caso os playbooks sejam usados para alertas específicos, forneça às equipes indicadores para o playbook no alerta.

À medida que sua organização continuar amadurecendo, automatize seus playbooks. Comece com playbooks para abordar incidentes de baixo risco. Use scripts para automatizar as etapas de descoberta. Tenha runbooks complementares para mitigar as causas-raiz comuns.

Resultado desejado: sua organização tem playbooks para incidentes comuns. Os playbooks são armazenados em um local central e estão disponíveis para os membros da equipe. Os playbooks são atualizados com frequência. Runbooks complementares são criados para todas as causas-raiz conhecidas.

Práticas comuns que devem ser evitadas:

- Não há uma maneira padrão de investigar um incidente.
- Os membros da equipe precisam confiar na própria memória ou no conhecimento institucional para solucionar uma falha na implantação.
- Os novos membros da equipe aprendem a investigar os problemas por meio de tentativa e erro.
- As práticas recomendadas para a investigação dos problemas não são compartilhadas entre as equipes.

Benefícios de implementar esta prática recomendada:

- Os playbooks impulsionam seus esforços para mitigar os incidentes.
- Diferentes membros da equipe podem usar o mesmo playbook para identificar uma causa-raiz de maneira consistente.
- As causas-raiz conhecidas podem ter runbooks desenvolvidos para elas, diminuindo o tempo de recuperação.
- Os playbooks permitem que os membros da equipe comecem a contribuir o quanto antes.
- As equipes podem escalar seus processos com playbooks repetíveis.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A maneira que você cria e usa os playbooks depende da maturidade da sua organização. Se você é iniciante na nuvem, crie playbooks no formato de texto em um repositório central de documentos.

À medida que sua organização amadurecer, os playbooks poderão passar a ser semiautomatizados com linguagens de script, como Python. Esses scripts podem ser executados em um caderno Jupyter para acelerar a descoberta. As organizações avançadas têm playbooks totalmente automatizados para problemas comuns que são corrigidos automaticamente com runbooks.

Comece a criar seus playbooks listando incidentes comuns que ocorrem com sua workload. Para começar, escolha playbooks para incidentes com baixo risco e nos quais a causa-raiz tenha sido restrita a poucos problemas. Quando já tiver playbooks para os cenários mais simples, passe para cenários de alto risco ou cenários em que a causa-raiz não é bem conhecida.

Seus playbooks em texto deverão ser automatizados à medida que sua organização amadurecer. Usando serviços como o AWS Systems Manager Automations, o texto simples pode ser transformado em automações. Essas automações podem ser executadas na workload para acelerar as investigações. Elas podem ser ativadas em resposta a eventos, o que reduz o tempo necessário para descobrir e resolver incidentes.

Os clientes podem usar o AWS Systems Manager Incident Manager para responder a incidentes. Esse serviço fornece uma interface única para fazer a triagem de incidentes, informar as partes interessadas durante a descoberta e a mitigação e permitir a colaboração durante todo o incidente. Ele usa o AWS Automations para acelerar a detecção e a recuperação.

Exemplo de cliente

Um incidente na produção afetou a AnyCompany Retail. O engenheiro de plantão usou um playbook para investigar o problema. À medida que foi avançando pelas etapas, ele manteve as principais partes interessadas informadas, as quais estão identificadas no playbook. O engenheiro identificou a causa-raiz como uma condição de corrida em um serviço de backend. Usando um runbook, o engenheiro reiniciou o serviço, colocando a AnyCompany Retail online novamente.

Etapas de implementação

Se você não tem um repositório de documentos, sugerimos criar um repositório de controle de versão para a biblioteca de playbooks. É possível criar os playbooks usando o Markdown, que é compatível com a maioria dos sistemas de automação de playbooks. Se você estiver iniciando do zero, use o modelo de exemplo de playbook a seguir.

```
# Playbook Title
## Playbook Info
| Playbook ID | Description | Tools Used | Special Permissions | Playbook Author | Last
 Updated | Escalation POC | Stakeholders | Communication Plan |
```

```
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| RUN001 | What is this playbook for? What incident is it used for? | Tools |
 Permissions | Your Name | 2022-09-21 | Escalation Name | Stakeholder Name | How will
 updates be communicated during the investigation? |
## Steps
1. Step one
2. Step two
```

1. Se você não tiver um repositório de documentos ou uma wiki, crie um repositório de controle de versão para seus playbooks no sistema de controle de versão.
2. Identifique um problema comum que requer investigação. Ele deve ser um cenário em que a causa-raiz está limitada a poucos problemas e a resolução é de baixo risco.
3. Usando o modelo Markdown, preencha a seção Nome do playbook e os campos em Informações do playbook.
4. Preencha as etapas de resolução de problemas. Seja o mais claro possível sobre quais ações devem ser executadas ou quais áreas devem ser investigadas.
5. Dê o playbook a um membro da equipe e peça para essa pessoa analisá-lo a fim de validá-lo. Caso algo esteja faltando ou não esteja claro, atualize o playbook.
6. Publique o playbook no repositório de documentos e informe sua equipe e as partes interessadas.
7. Essa biblioteca de playbooks crescerá à medida que você adicionar outros playbooks. Quando tiver vários playbooks, comece a automatizá-los usando ferramentas como o AWS Systems Manager Automations a fim de manter a automação e os playbooks sincronizados.

Nível de esforço do plano de implementação: Baixo. Os playbooks devem ser documentos de texto armazenados em um local central. Organizações mais consolidadas passarão a automatizar os respectivos playbooks.

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP02 Processos e procedimentos com proprietários identificados
- OPS07-BP03 Usar runbooks para realizar procedimentos
- OPS10-BP01 Usar um processo para gerenciamento de eventos, incidentes e problemas
- OPS10-BP02 Adotar um processo por alerta
- OPS11-BP04 Gerenciar o conhecimento

Documentos relacionados:

- AWS Well-Architected Framework: conceitos: desenvolvimento de playbooks
- Como alcançar excelência operacional usando playbooks e runbooks automatizados
- AWS Systems Manager: trabalhar com runbooks
- Como usar runbooks do AWS Systems Manager Automation para resolver tarefas operacionais

Vídeos relacionados:

- AWS re:Invent 2019: Guia de faça você mesmo para runbooks, relatórios de incidentes e resposta a incidentes (SEC318-R1)
- AWS Systems Manager Incident Manager: workshops virtuais da AWS
- Integrar scripts ao AWS Systems Manager

Exemplos relacionados:

- Framework do playbook do cliente da AWS
- AWS Systems Manager: orientações sobre automação
- Criar um runbook de resposta a incidentes da AWS usando cadernos Jupyter e o CloudTrail Lake
- Rubix: uma biblioteca Python para criação de runbooks em cadernos Jupyter
- Usar o Document Builder para criar um runbook personalizado
- Laboratórios do Well-Architected: Automatização de operações com playbooks e runbooks
- Laboratórios do Well-Architected: Playbook de resposta a incidentes com o Jupyter

Serviços relacionados:

- AWS Systems Manager Automation
- AWS Systems Manager Incident Manager

OPS07-BP05 Tomar decisões embasadas para implantar sistemas e alterações

Adote processos para lidar com as alterações com e sem êxito feitas na workload. Uma estratégia pre-mortem é um exercício em que uma equipe simula uma falha para desenvolver estratégias de mitigação. Use as estratégias pre-mortem para antecipar falhas e criar procedimentos, quando

apropriado. Avalie os benefícios e os riscos de implantar alterações na workload. Verifique se todas as alterações estão em conformidade com a governança.

Resultado desejado:

- Você toma decisões embasadas ao implantar alterações na workload.
- As alterações estão em conformidade com a governança.

Práticas comuns que devem ser evitadas:

- Implantar uma alteração em nossa workload sem um processo para lidar com uma implantação com falha.
- Fazer alterações no ambiente de produção que estão fora da conformidade com os requisitos de governança.
- Implantar uma nova versão da workload sem estabelecer uma referência para a utilização de recursos.

Benefícios de implementar esta prática recomendada:

- Você está preparado para alterações sem êxito na workload.
- As alterações na workload estão em conformidade com as políticas de governança.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Use estratégias pre-mortem para desenvolver processos para lidar com alterações sem êxito. Documente os processos de alterações sem êxito. Garanta que todas as alterações estejam em conformidade com a governança. Avalie os benefícios e os riscos de implantar alterações na workload.

Exemplo de cliente

A AnyCompany Retail realiza estratégias pre-mortem regularmente para validar seus processos de alterações sem êxito. Os processos são documentados em uma Wiki compartilhada e atualizados regularmente. Todas as alterações estão em conformidade com os requisitos de governança.

Etapas de implementação

1. Tome decisões embasadas ao implantar alterações na workload. Estabeleça e revise os critérios de uma implantação bem-sucedida. Desenvolva cenários ou critérios que acionariam a reversão de uma alteração. Pondere os benefícios de implantar alterações considerando os riscos de uma alteração sem êxito.
2. Verifique se todas as alterações estão em conformidade com as políticas de governança.
3. Use estratégias pre-mortem para alterações sem êxito e documente as estratégias de migração. Realize um exercício de simulação para modelar uma alteração sem êxito e validar os procedimentos de reversão.

Nível de esforço do plano de implementação: Moderado. Implementar uma prática de estratégias pre-mortem requer coordenação e esforço das partes interessadas na organização

Recursos

Práticas recomendadas relacionadas:

- OPS01-BP03 Avaliar os requisitos de governança: os requisitos de governança são um fator fundamental para determinar se uma alteração deve ser implementada.
- OPS06-BP01 Preparar-se para alterações malsucedidas: estabeleça planos para mitigar uma implantação sem êxito e use estratégias pre-mortem para validá-los.
- OPS06-BP02 Testar implantações: toda alteração de software deve ser testada adequadamente antes da implantação para reduzir os defeitos na produção.
- OPS07-BP01 Garantir a capacidade da equipe: ter um número suficiente de funcionários treinados para fornecer suporte à workload é essencial para tomar uma decisão embasada quanto à implantação de uma alteração no sistema.

Documentos relacionados:

- Amazon Web Services: risco e conformidade
- Modelo de responsabilidade compartilhada da AWS
- Governança na Nuvem AWS: o equilíbrio certo entre agilidade e segurança

OPS07-BP06 Criar planos de suporte para workloads de produção

Habilite o suporte para qualquer software e quaisquer serviços dos quais sua workload de produção dependa. Selecione um nível de suporte apropriado para atender às necessidades de nível de

serviço da produção. Planos de suporte para essas dependências são necessários no caso de interrupção de um serviço ou de um problema de software. Documente os planos de suporte e como solicitar suporte para todos os fornecedores de serviços e software. Implemente mecanismos que verifiquem se os pontos de contato do suporte são mantidos atualizados.

Resultado desejado:

- Implemente planos de suporte para software e serviços dos quais as workloads de produção dependem.
- Escolha um plano de suporte apropriado com base nas necessidades de nível de serviço.
- Documente os planos de suporte, os níveis de suporte e como solicitar suporte.

Práticas comuns que devem ser evitadas:

- Você não tem nenhum plano de suporte junto a um fornecedor de software essencial. Sua workload é afetada por isso e você não pode fazer nada para agilizar a correção ou obter atualizações em tempo hábil do fornecedor.
- Um desenvolvedor que era o principal ponto de contato com um fornecedor de software deixou a empresa. Você não consegue entrar em contato com o suporte do fornecedor diretamente. Você precisa despender tempo pesquisando e navegando por sistemas de contato genéricos, aumentando o tempo requerido para responder quando necessário.
- Uma interrupção ocorre na produção relacionada a um fornecedor de software. Não há nenhuma documentação sobre como abrir um caso de suporte.

Benefícios de implementar esta prática recomendada:

- Com o nível de suporte apropriado, você é capaz de obter uma resposta no espaço de tempo requerido para atender às necessidades de nível de serviço.
- Como um cliente com suporte, você pode encaminhar a questão se houver problemas na produção.
- Os fornecedores de software e serviços podem ajudar na resolução de problemas durante um incidente.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Habilite planos de suporte para qualquer software e quaisquer serviços dos quais sua workload de produção dependa. Estabeleça planos de suporte apropriados para atender a necessidades de nível de serviço. Para clientes da AWS, isso significa habilitar o AWS Business Support ou superior em quaisquer contas em que você tenha workloads de produção. Entre em contato regularmente com os fornecedores de suporte para obter atualizações sobre ofertas, processos e contatos de suporte. Documente como solicitar suporte de fornecedores de software e serviços e sobre como encaminhar problemas se houver uma interrupção. Implemente mecanismos para manter os contatos de suporte atualizados.

Exemplo de cliente

Na AnyCompany Retail, todas dependências de software e serviços comerciais contam com planos de suporte. Por exemplo, eles têm o AWS Enterprise Support ativado em todas as contas com workloads de produção. Qualquer desenvolvedor pode abrir um caso de suporte quando há um problema. Há uma página de wiki com informações sobre como solicitar suporte, a quem notificar e as práticas recomendadas para agilizar um caso.

Etapas de implementação

1. Trabalhe com as partes interessadas em sua organização para identificar fornecedores de software e serviços dos quais sua workload dependa. Documente essas dependências.
2. Determine as necessidades de nível de serviço para sua workload. Selecione um plano de suporte alinhado a elas.
3. Para software e serviços comerciais, estabeleça um plano de suporte com os fornecedores.
    a. A assinatura do AWS Business Support ou superior para todas as contas de produção fornece um tempo de resposta rápido do AWS Support e é altamente recomendada. Se você não tiver suporte premium, precisará de um plano de ação para lidar com os problemas, o que requer a ajuda do AWS Support. O AWS Support oferece um conjunto de ferramentas e tecnologia, pessoas e programas projetados para ajudar você de forma proativa a otimizar a performance, reduzir custos e inovar com maior rapidez. AWS O Business Support oferece benefícios adicionais, incluindo acesso ao AWS Trusted Advisor e ao AWS Personal Health Dashboard e tempos de resposta mais rápidos.
4. Documente o plano de suporte em sua ferramenta de gerenciamentos de conhecimentos. Inclua como solicitar suporte, a quem notificar se for aberto um caso de suporte e como encaminhar o problema durante um incidente. Uma wiki é um bom mecanismo para possibilitar que todos façam

as atualizações necessárias na documentação quando forem informados sobre alterações em processos ou contatos de suporte.

Nível de esforço do plano de implementação: Baixo. A maioria dos fornecedores de software e serviços oferece planos de suporte que requerem adesão. Documentar e compartilhar práticas recomendadas no sistema de gerenciamento de conhecimentos garante que sua equipe saiba o que fazer quando houver um problema na produção.

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP02 Processos e procedimentos com proprietários identificados

Documentos relacionados:

- Planos do AWS Support

Serviços relacionados:

- AWS Business Support
- AWS Enterprise Support

Operar

Perguntas

- OPS 8. Como utilizar a observabilidade da workload em sua organização?
- OPS 9. Como compreender a integridade das suas operações?
- OPS 10. Como gerenciar os eventos de workload e operações?

OPS 8. Como utilizar a observabilidade da workload em sua organização?

Garanta a integridade ideal da workload usando a observabilidade. Utilize métricas, logs e rastreamentos relevantes para obter uma visão abrangente da performance da sua workload e resolver problemas com eficiência.

Práticas recomendadas

- OPS08-BP01 Analisar métricas da workload
- OPS08-BP02 Analisar logs de workloads
- OPS08-BP03 Analisar rastreamentos de workload
- OPS08-BP04 Criar alertas acionáveis
- OPS08-BP05 Criar painéis

OPS08-BP01 Analisar métricas da workload

Depois de implementar a telemetria de aplicações, analise regularmente as métricas coletadas. Embora a latência, as solicitações, os erros e a capacidade (ou cotas) forneçam informações sobre a performance do sistema, é fundamental priorizar a análise das métricas de resultados comerciais. Isso garante que você esteja tomando decisões orientadas por dados alinhadas aos seus objetivos de negócios.

Resultado desejado: insights precisos sobre a performance da workload que impulsionam decisões baseadas em dados, garantindo o alinhamento com os objetivos de negócios.

Práticas comuns que devem ser evitadas:

- Análise das métricas isoladamente, sem considerar seu impacto nos resultados comerciais.
- Confiança excessiva em métricas técnicas e, ao mesmo tempo, marginalização das métricas de negócios.
- Revisão pouco frequente das métricas, perdendo oportunidades de tomada de decisão em tempo real.

Benefícios de implementar esta prática recomendada:

- Compreensão aprimorada da correlação entre performance técnica e resultados comerciais.
- Processo de tomada de decisão aprimorado baseado em dados em tempo real.
- Identificação proativa e mitigação de problemas antes que eles afetem os resultados comerciais.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Utilize ferramentas como o Amazon CloudWatch para realizar análises métricas. Serviços da AWS como a deteção de anomalias do CloudWatch e o Amazon DevOps Guru podem ser usados para

detectar anomalias, especialmente quando os limites estáticos são desconhecidos ou quando os padrões de comportamento são mais adequados para a detecção de anomalias.

Etapas de implementação

1. Analise e revise: revise e interprete regularmente suas métricas de workload.
   a. Priorize as métricas de resultados comerciais em vez das métricas puramente técnicas.
   b. Entenda a importância de picos, quedas ou padrões em seus dados.
2. Utilize o Amazon CloudWatch: use o Amazon CloudWatch para uma visão centralizada e uma análise aprofundada.
   a. Configure painéis do CloudWatch para visualizar suas métricas e compará-las ao longo do tempo.
   b. Use percentis no CloudWatch para obter uma visão clara da distribuição métrica, o que pode ajudar na definição de SLAs e na compreensão de valores discrepantes.
   c. Configure a detecção de anomalias do CloudWatch para identificar padrões incomuns sem depender de limites estáticos.
   d. Implemente a observabilidade entre contas do CloudWatch para monitorar e solucionar problemas de aplicações que abrangem várias contas em uma região.
   e. Use o CloudWatch Metric Insights para consultar e analisar dados métricos em contas e regiões, identificando tendências e anomalias.
   f. Aplique o CloudWatch Metric Math para transformar, agregar ou realizar cálculos em suas métricas para obter insights mais profundos.
3. Use o Amazon DevOps Guru: incorpore o Amazon DevOps Guru por sua detecção de anomalias aprimorada por machine learning para identificar sinais precoces de problemas operacionais em suas aplicações sem servidor e corrigi-los antes que afetem seus clientes.
4. Otimize com base em insights: tome decisões informadas baseadas na análise das métricas para ajustar e melhorar as workloads.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Identificar indicadores-chave de performance
- OPS04-BP02 Implementar a telemetria de aplicações

Documentos relacionados:

- The Wheel Blog: como enfatizar a importância de revisar continuamente as métricas
- O percentil é importante
- Usar o AWS Cost Anomaly Detection
- Observabilidade entre contas do CloudWatch
- Consultar métricas com o CloudWatch Metrics Insights

Vídeos relacionados:

- Habilitar a observabilidade entre contas no Amazon CloudWatch
- Introdução ao Amazon DevOps Guru
- Analisar continuamente as métricas usando o AWS Cost Anomaly Detection

Exemplos relacionados:

- Workshop One Observability
- Obter insights operacionais com AIOps usando o Amazon DevOps Guru

OPS08-BP02 Analisar logs de workloads

Analisar regularmente os logs da workload é essencial para obter uma compreensão mais profunda dos aspectos operacionais da sua aplicação. Ao filtrar, visualizar e interpretar com eficiência os dados de log, é possível otimizar continuamente a performance e a segurança das aplicações.

Resultado desejado: informações ricas sobre o comportamento e as operações da aplicação derivadas de uma análise completa de log, garantindo a detecção e mitigação proativas de problemas.

Práticas comuns que devem ser evitadas:

- Negligenciar a análise dos logs até um problema crítico surgir.
- Não usar o conjunto completo de ferramentas disponíveis para análise de logs, deixando para trás insights essenciais.
- Confiar exclusivamente na revisão manual dos logs, sem utilizar os recursos de automação e consulta.

Benefícios de implementar esta prática recomendada:

- Identificação proativa de gargalos operacionais, ameaças à segurança e outros possíveis problemas.
- Utilização eficiente dos dados de log para otimização contínua da aplicação.
- Compreensão aprimorada do comportamento da aplicação, auxiliando na depuração e solução de problemas.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

O Amazon CloudWatch Logs é uma ferramenta poderosa para análise de logs. Recursos integrados, como o CloudWatch Logs Insights e o Contributor Insights, tornam intuitivo e eficiente o processo de derivação de informações significativas dos logs.

Etapas de implementação

1. Configure o CloudWatch Logs: configure aplicações e serviços para enviar logs ao CloudWatch Logs.
2. Use a detecção de anomalias de log: utilize a detecção de anomalias do Amazon CloudWatch Logs para identificar e alertar automaticamente sobre padrões de log incomuns. Essa ferramenta ajuda você a gerenciar proativamente anomalias nos logs e detectar possíveis problemas com antecedência.
3. Configure o CloudWatch Logs Insights: use o CloudWatch Logs Insights para pesquisar e analisar dados de log de forma interativo.
   a. Crie consultas para extrair padrões, visualizar dados de log e obter insights acionáveis.
   b. Use a análise de padrões do CloudWatch Logs Insights para analisar e visualizar padrões de log frequentes. Esse recurso ajuda você a entender tendências operacionais comuns e possíveis discrepâncias em seus dados de logs.
   c. Use a comparação (diff) do CloudWatch Logs para realizar análises diferenciais entre diferentes períodos de tempo ou grupos de logs. Use esse recurso para identificar mudanças e avaliar seus impactos na performance ou no comportamento do sistema.
4. Monitore registros em tempo real com o Live Tail: use o Amazon CloudWatch Logs Live Tail para visualizar dados de log em tempo real. Você pode monitorar ativamente as atividades operacionais da aplicação à medida que elas ocorrem, o que oferece visibilidade imediata da performance do sistema e dos possíveis problemas.

5. Aproveite o Contributor Insights: use o CloudWatch Contributor Insights para identificar os principais oradores em dimensões de alta cardinalidade, como endereços IP ou agentes de usuário.
6. Implemente filtros métricos do CloudWatch Logs: configure os filtros métricos do CloudWatch Logs para converter dados de log em métricas acionáveis. Isso permite que você defina alarmes ou analise melhor os padrões.
7. Implemente a observabilidade entre contas do Amazon CloudWatch: monitore e solucione problemas de aplicações que abrangem várias contas em uma região.
8. Revisão e aprimoramento periódicos: revise periodicamente suas estratégias de análise de log para capturar todas as informações relevantes e otimizar continuamente a performance da aplicação.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Identificar indicadores-chave de performance
- OPS04-BP02 Implementar a telemetria de aplicações
- OPS08-BP01 Analisar métricas da workload

Documentos relacionados:

- Analisar logs de dados com o CloudWatch Logs Insights
- Usar o CloudWatch Contributor Insights
- Criar e gerenciar de filtros de métrica do CloudWatch Logs

Vídeos relacionados:

- Analisar logs de dados com o CloudWatch Logs Insights
- Usar o CloudWatch Contributor Insights para analisar dados de alta cardinalidade

Exemplos relacionados:

- Exemplos de consultas do CloudWatch Logs

- Workshop One Observability

OPS08-BP03 Analisar rastreamentos de workload

Analisar dados de rastreamento é crucial para obter uma visão abrangente da jornada operacional de uma aplicação. Ao visualizar e compreender as interações entre vários componentes, a performance pode ser ajustada, os gargalos identificados e as experiências do usuário aprimoradas.

Resultado desejado: obtenha uma visibilidade clara das operações distribuídas da sua aplicação, permitindo uma resolução mais rápida de problemas e uma experiência de usuário aprimorada.

Práticas comuns que devem ser evitadas:

- Ignorar dados de rastreamento, confiando apenas em logs e métricas.
- Não correlacionar dados de rastreamento com logs associados.
- Ignorar as métricas derivadas de rastreamentos, como latência e taxas de falhas.

Benefícios de implementar esta prática recomendada:

- Aprimoramento da solução de problemas e redução do tempo médio de resolução (MTTR).
- Obtenção de insights sobre dependências e seu impacto.
- Identificação e correção rápidas de problemas de performance.
- Uso de métricas derivadas de rastreamento para uma tomada de decisão informada.
- Experiências de usuário aprimoradas por meio de interações otimizadas de componentes.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

O AWS X-Ray oferece um pacote abrangente para análise de dados de rastreamento, fornecendo uma visão holística das interações de serviços, monitorando as atividades do usuário e detectando problemas de performance. Recursos como ServiceLens, X-Ray Insights, X-Ray Analytics Analytics e Amazon DevOps Guru aprimoram o detalhamento dos insights acionáveis derivados de dados de rastreamento.

Etapas de implementação

As seguintes etapas oferecem uma abordagem estruturada para implementar com eficácia a análise de dados de rastreamento usando serviços da AWS:

1. Integre o AWS X-Ray: certifique-se de que o X-Ray esteja integrado às suas aplicações para capturar dados de rastreamento.
2. Analise as métricas do X-Ray: mergulhe nas métricas derivadas dos rastreamentos do X-Ray, como latência, taxas de solicitação, taxas de falhas e distribuições de tempo de resposta, usando o mapa de serviços para monitorar a integridade da aplicação.
3. Use o ServiceLens: aproveite o mapa do ServiceLens para melhorar a observabilidade de seus serviços e aplicações. Isso permite a visualização integrada de rastreamentos, métricas, logs, alarmes e outras informações de integridade.
4. Habilite o X-Ray Insights:
    a. Ative o X-Ray Insights para detecção automática de anomalias em rastreamentos.
    b. Examine os insights para identificar padrões e determinar as causas-raiz, como maiores taxas de falhas ou latências.
    c. Consulte o cronograma de insights para realizar uma análise cronológica dos problemas detectados.
5. Use o X-Ray Analytics: o X-Ray Analytics permite que você explore minuciosamente os dados de rastreamento, identifique padrões e extraia insights.
6. Use grupos no X-Ray: crie grupos no X-Ray para filtrar rastreamentos com base em critérios como alta latência, permitindo uma análise mais direcionada.
7. Incorpore o Amazon DevOps Guru: envolva o Amazon DevOps Guru para se beneficiar dos modelos de machine learning que identificam anomalias operacionais nos rastreamentos.
8. Use o CloudWatch Synthetics: use o CloudWatch Synthetics para criar canários para monitorar continuamente seus endpoints e fluxos de trabalho. Esses canários podem se integrar ao X-Ray para fornecer dados de rastreamento para uma análise detalhada das aplicações que estão sendo testadas.
9. Use o monitoramento de usuários reais (RUM): Com o AWS X-Ray e o CloudWatch RUM, é possível analisar e depurar o caminho da solicitação a partir dos usuários finais da aplicação até os serviços subsequentes gerenciados pela AWS. Isso ajuda a identificar tendências e erros de latência que afetam os usuários finais.
10. Correlacione com logs: correlacione dados de rastreamento com registros relacionados na visualização de rastreamento do X-Ray para obter uma perspectiva granular do comportamento

da aplicação. Isso permite que você visualize eventos de log diretamente associados às transações rastreadas.

11. Implemente a observabilidade entre contas do Amazon CloudWatch: monitore e solucione problemas de aplicações que abrangem várias contas em uma região.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS08-BP01 Analisar métricas da workload
- OPS08-BP02 Analisar logs de workloads

Documentos relacionados:

- Usar o ServiceLens para monitorar a integridade da aplicação
- Explorar dados de rastreamento com o X-Ray Analytics
- Detectar anomalias em rastreamentos com o X-Ray Insights
- Monitorar continuamente com o CloudWatch Synthetics

Vídeos relacionados:

- Analisar e depurar aplicações usando Amazon CloudWatch Synthetics e o AWS X-Ray
- Usar o AWS X-Ray Insights

Exemplos relacionados:

- Workshop One Observability
- Implementar o X-Ray com AWS Lambda
- Modelos canário do CloudWatch Synthetics

OPS08-BP04 Criar alertas acionáveis

Detectar e responder prontamente aos desvios no comportamento da sua aplicação é crucial. É essencial reconhecer quando os resultados baseados em indicadores-chave de performance (KPIs)

estão em risco ou quando surgem anomalias inesperadas. Basear alertas em KPIs garante que os sinais que você recebe estejam diretamente vinculados ao impacto comercial ou operacional. Essa abordagem de alertas acionáveis promove respostas proativas e ajuda a manter a performance e a confiabilidade do sistema.

Resultado desejado: receba alertas imediatos, relevantes e acionáveis para rápida identificação e mitigação de possíveis problemas, especialmente quando os resultados dos KPI estão em risco.

Práticas comuns que devem ser evitadas:

- A configuração de muitos alertas não críticos gera fadiga de alertas.
- A não priorização de alertas com base em KPIs dificulta a compreensão do impacto comercial dos problemas.
- A não abordagem das causas-raiz ocasiona alertas repetitivos para o mesmo problema.

Benefícios de implementar esta prática recomendada:

- Redução da fadiga de alertas ao se concentrar em alertas acionáveis e relevantes.
- Maior disponibilidade e confiabilidade do sistema por meio da detecção e mitigação proativas de problemas.
- Colaboração em equipe aprimorada e resolução mais rápida de problemas por meio da integração com ferramentas conhecidas de alerta e comunicação.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Para criar um mecanismo de alerta eficaz, é fundamental usar métricas, logs e dados de rastreamento que sinalizem quando os resultados com base nos KPIs estão em risco ou quando anomalias são detectadas.

Etapas de implementação

1. Determine indicadores-chave de performance (KPIs): identifique KPIs da sua aplicação. Os alertas devem estar vinculados a esses KPIs para refletir com precisão o impacto nos negócios.
2. Implemente a detecção de anomalias:

- Use a detecção de anomalias do Amazon CloudWatch: configure a detecção de anomalias do Amazon CloudWatch para detectar automaticamente padrões incomuns, o que ajuda você a gerar alertas somente para anomalias genuínas.
- Use o AWS X-Ray Insights:
  a. Configure o X-Ray Insights para detectar anomalias nos dados de rastreamento.
  b. Configure notificações para que o X-Ray Insights seja alertado sobre problemas detectados.
- Integre-se ao Amazon DevOps Guru:
  a. Utilize o Amazon DevOps Guru devido a seus recursos de machine learning na detecção de anomalias operacionais com dados existentes.
  b. Navegue até as configurações de notificação no DevOps Guru para configurar alertas de anomalias.

3. Implemente alertas acionáveis: crie alertas que forneçam informações adequadas para ação imediata.
   1. Monitore eventos do AWS Health com as regras do Amazon EventBridge ou integre-se programaticamente à AWS Health API para automatizar ações ao receber eventos do AWS Health. Podem ser ações gerais, como enviar todas as mensagens planejadas de eventos do ciclo de vida para uma interface de chat, ou ações específicas, como o início de um fluxo de trabalho em uma ferramenta de gerenciamento de serviços de TI.
4. Reduza a fadiga dos alertas: minimize os alertas não críticos. Quando as equipes se tornam sobrecarregadas com vários alertas insignificantes, elas podem não perceber problemas críticos, o que diminui a eficácia geral do mecanismo de alerta.
5. Configure alarmes compostos: use os alarmes compostos do Amazon CloudWatch para consolidar vários alarmes.
6. Integre com ferramentas de alerta: incorpore ferramentas como Ops Genie e PagerDuty.
7. Engage o AWS Chatbot: integre o AWS Chatbot para retransmitir alertas para o Amazon Chime, o Microsoft Teams e o Slack.
8. Alerta com base em logs: use filtros de métrica de log no CloudWatch para criar alarmes com base em eventos de log específicos.
9. Revise e repita: revisite e refine regularmente as configurações de alerta.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Identificar indicadores-chave de performance
- OPS04-BP02 Implementar a telemetria de aplicações
- OPS04-BP03 Implementar telemetria da experiência do usuário
- OPS04-BP04 Implementar a telemetria de dependências
- OPS04-BP05 Implementar rastreamento distribuído
- OPS08-BP01 Analisar métricas da workload
- OPS08-BP02 Analisar logs de workloads
- OPS08-BP03 Analisar rastreamentos de workload

Documentos relacionados:

- Usar alarmes do Amazon CloudWatch
- Criar um alarme composto
- Criar um alarme do CloudWatch baseado na detecção de anomalias
- Notificações do DevOps Guru
- Notificações de insights do X-Ray
- Monitorar, operar e solucionar problemas de seus recursos da AWS com ChatOps interativos
- Guia de integração do Amazon CloudWatch | PagerDuty
- Integrar o Opsgenie com o Amazon CloudWatch

Vídeos relacionados:

- Criar alarmes compostos no Amazon CloudWatch
- Visão geral do AWS Chatbot
- Destaque da AWS On Air: comandos mutantes no AWS Chatbot

Exemplos relacionados:

- Alarmes, gerenciamento de incidentes e remediação na nuvem com o Amazon CloudWatch

- Tutorial: criar uma regra do Amazon EventBridge que envia notificações para o AWS Chatbot
- Workshop One Observability

OPS08-BP05 Criar painéis

Os painéis são a visão voltada para o ser humano dos dados de telemetria das workloads. Embora forneçam uma interface visual vital, eles não devem substituir os mecanismos de alerta, mas sim complementá-los. Quando elaborados com cuidado, eles não apenas oferecem insights rápidos sobre a integridade e a performance do sistema, como também podem apresentar às partes interessadas informações em tempo real sobre os resultados empresariais e o impacto dos problemas.

Resultado desejado:

Insights claros e acionáveis sobre a integridade do sistema e dos negócios usando representações visuais.

Práticas comuns que devem ser evitadas:

- Painéis complicados demais e com muitas métricas.
- Confiar em painéis sem alertas para detecção de anomalias.
- Não atualizar os painéis à medida que as workloads evoluem.

Benefícios de implementar esta prática recomendada:

- Visibilidade imediata de métricas e KPIs críticos do sistema.
- Comunicação e compreensão aprimoradas com as partes interessadas.
- Visão rápida do impacto dos problemas operacionais.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Painéis centrados nos negócios

Os painéis personalizados para os KPIs de negócios envolvem uma gama maior de partes interessadas. Embora essas pessoas possam não estar interessadas nas métricas do sistema, elas

estão interessadas em entender as implicações comerciais desses números. Um painel centrado nos negócios garante que todas as métricas técnicas e operacionais monitoradas e analisadas estejam sincronizadas com as metas empresariais abrangentes. Esse alinhamento fornece clareza, garantindo que todos estejam em sintonia sobre o que é essencial e o que não é. Além disso, painéis que destacam os KPIs de negócios tendem a ser mais acionáveis. As partes interessadas podem entender rapidamente a integridade das operações, as áreas que precisam de atenção e o impacto potencial nos resultados empresariais.

Com isso em mente, ao criar seus painéis, garanta que haja um equilíbrio entre métricas técnicas e KPIs comerciais. Ambos são vitais, mas atendem a públicos diferentes. O ideal é que você tenha painéis que forneçam uma visão holística da integridade e da performance do sistema e, ao mesmo tempo, enfatizem os principais resultados comerciais e suas implicações.

Os painéis do Amazon CloudWatch são páginas iniciais personalizáveis no console do CloudWatch que você pode usar para monitorar seus recursos em uma única visualização, até mesmo os recursos distribuídos em diferentes contas e Regiões da AWS.

Etapas de implementação

1. Crie um painel básico: crie um novo painel no CloudWatch e dê a ele um nome descritivo.
2. Use widgets do Markdown: antes de mergulhar nas métricas, use os widgets do Markdown para adicionar contexto textual na parte superior do seu painel. Isso deve explicar o que o painel abrange, a importância das métricas representadas e também pode conter links para outros painéis e ferramentas de solução de problemas.
3. Crie variáveis do painel: incorpore variáveis do painel onde apropriado para permitir visualizações dinâmicas e flexíveis do painel.
4. Crie widgets de métricas: adicione widgets de métricas para visualizar várias métricas que sua aplicação emite, adaptando esses widgets para representar com eficácia a integridade do sistema e os resultados empresariais.
5. Consultas do Log Insights: use o CloudWatch Logs Insights para obter métricas acionáveis de seus logs e exibir esses insights no painel.
6. Configure alarmes: integre os alarmes do CloudWatch ao seu painel para ter uma visão rápida de qualquer métrica que esteja ultrapassando seus limites.
7. Use o Contributor Insights: incorpore o CloudWatch Contributor Insights para analisar campos de alta cardinalidade e obter uma compreensão mais clara dos principais colaboradores do seu recurso.

8. Crie widgets personalizados: para necessidades específicas não atendidas pelos widgets padrão, considere criar widgets personalizados. Eles podem ser extraídos de várias fontes de dados ou representar dados de maneiras exclusivas.
9. Use o AWS Health Dashboard: use o AWS Health Dashboard para obter informações mais detalhadas sobre a integridade da sua conta, eventos e mudanças futuras que possam afetar seus serviços e recursos. Você também pode obter uma visão centralizada dos eventos de integridade no AWS Organizations ou criar seus próprios painéis personalizados (para obter mais detalhes, consulte Exemplos relacionados).
10. Itere e refine: à medida que sua aplicação evolui, revise regularmente o painel para garantir sua relevância.

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Identificar indicadores-chave de performance
- OPS08-BP01 Analisar métricas da workload
- OPS08-BP02 Analisar logs de workloads
- OPS08-BP03 Analisar rastreamentos de workload
- OPS08-BP04 Criar alertas acionáveis

Documentos relacionados:

- Como criar painéis para visibilidade operacional
- Usar painéis do Amazon CloudWatch

Vídeos relacionados:

- Criar painéis do CloudWatch entre contas e entre regiões
- AWS re:Invent 2021: como obter visibilidade corporativa com painéis de operação do Nuvem AWS

Exemplos relacionados:

- Workshop One Observability
- Monitoramento de aplicações com o Amazon CloudWatch

- Painéis e insights de inteligência de eventos do AWS Health
- Visualizar eventos do AWS Health usando o Amazon Managed Grafana

OPS 9. Como compreender a integridade das suas operações?

Defina, capture e analise as métricas de operações para obter visibilidade dos eventos de operações, para que você possa tomar as ações apropriadas.

Práticas recomendadas

- OPS09-BP01 Medir metas operacionais e KPIs com métricas
- OPS09-BP02 Comunicar o status e as tendências para garantir a visibilidade da operação
- OPS09-BP03 Revisar as métricas operacionais e priorizar melhorias

OPS09-BP01 Medir metas operacionais e KPIs com métricas

Obtenha metas e KPIs que definam o sucesso das operações de sua organização e determine se as métricas os refletem. Defina linhas de base como ponto de referência e reavalie regularmente. Desenvolva mecanismos para coletar essas métricas das equipes para avaliação.

Resultado desejado:

- As metas e os KPIs das equipes de operações da organização foram publicados e compartilhados.
- Métricas que refletem esses KPIs são estabelecidas. Os exemplos podem incluir:
  - Profundidade da fila de tíquetes ou idade média do tíquete
  - Contagem de tíquetes agrupada por tipo de problema
  - Tempo gasto trabalhando em problemas com ou sem um procedimento operacional padronizado (SOP)
  - Tempo gasto na recuperação de uma falha no envio de código
  - Volume de chamadas

Práticas comuns que devem ser evitadas:

- Os prazos de implantação são perdidos porque os desenvolvedores são contratados para realizar tarefas de solução de problemas. As equipes de desenvolvimento demandam mais pessoal, mas não conseguem quantificar quantos precisam porque o tempo perdido não pode ser medido.

- Um atendimento de Nível 1 foi configurado para lidar com chamadas de usuários. Com o tempo, mais workloads foram adicionadas, mas nenhum número de funcionários foi alocado para o atendimento de Nível 1. A satisfação do cliente sofre à medida que os tempos de atendimento aumentam e os problemas ficam mais tempo sem resolução, mas a gerência não vê indicadores disso, impedindo qualquer ação.
- Uma workload problemática foi transferida para uma equipe de operações separada para manutenção. Diferentemente de outras workloads, a nova não foi fornecida com documentação e runbooks adequados. Dessa forma, as equipes passam mais tempo solucionando problemas e lidando com falhas. No entanto, não há métricas que documentem isso, o que dificulta a prestação de contas.

Benefícios de implementar esta prática recomendada: onde o monitoramento da workload mostra o estado de nossas aplicações e serviços, as equipes de operações de monitoramento fornecem aos proprietários uma visão das mudanças entre os consumidores dessas workloads, como as mudanças nas necessidades de negócios. Meça a eficácia dessas equipes e avalie-as em relação às metas de negócios, criando métricas que possam refletir o estado das operações. As métricas podem destacar problemas de suporte ou identificar quando há desvios de uma meta de nível de serviço.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Agende um horário com líderes de negócios e partes interessadas para determinar as metas gerais do serviço. Determine quais devem ser as tarefas de várias equipes de operações e quais desafios elas podem enfrentar. Com isso, pense em indicadores-chave de performance (KPIs) que possam refletir essas metas operacionais. Pode ser a satisfação do cliente, o tempo desde a concepção do recurso até a implantação, o tempo médio de resolução de problemas e outros.

Trabalhando a partir de KPIs, identifique as métricas e as fontes de dados que podem refletir melhor essas metas. A satisfação do cliente pode ser uma combinação de várias métricas, como tempos de espera ou resposta de chamadas, índices de satisfação e tipos de problemas levantados. Os tempos de implantação podem ser a soma do tempo necessário para testes e implantação e quaisquer correções pós-implantação que precisem ser adicionadas. As estatísticas que mostram o tempo gasto em diferentes tipos de problemas (ou a contagem desses problemas) podem fornecer uma visão de onde é necessário um esforço direcionado.

Recursos

Documentos relacionados:

- Amazon QuickSight: usar KPIs
- Amazon CloudWatch: usar métricas
- Criar painéis
- Como rastrear KPIs de otimização de custos com o painel de KPI

OPS09-BP02 Comunicar o status e as tendências para garantir a visibilidade da operação

É necessário conhecer o estado de suas operações e a direção das tendências para identificar quando os resultados podem estar em risco, se trabalho adicional pode ou não receber apoio ou os efeitos que as mudanças causaram em suas equipes. Durante eventos operacionais, ter páginas de status que os usuários e as equipes operacionais possam consultar para obter informações pode reduzir a pressão nos canais de comunicação e disseminar informações de forma proativa.

Resultado desejado:

- Os líderes de operações têm uma visão rápida para ver em que tipo de volume de chamadas suas equipes estão operando e quais esforços podem estar em andamento, como implantações.
- Os alertas são disseminados para as partes interessadas e comunidades de usuários quando ocorrem impactos nas operações normais.
- A liderança da organização e as partes interessadas podem verificar uma página de status em resposta a um alerta ou impacto e obter informações sobre um evento operacional, como pontos de contato, informações sobre tíquetes e tempos estimados de recuperação.
- Os relatórios são disponibilizados para a liderança e outras partes interessadas para mostrar estatísticas operacionais, como volumes de chamadas durante um período de tempo, índices de satisfação do usuário, números de tíquetes pendentes e suas idades.

Práticas comuns que devem ser evitadas:

- Uma workload diminui, deixando um serviço indisponível. O volume de chamadas aumenta à medida que os usuários solicitam saber o que está acontecendo. Os gerentes aumentam o volume de solicitações para saber quem está resolvendo um problema. Várias equipes de operações duplicam esforços na tentativa de investigar.

- O desejo por uma nova capacidade faz com que vários funcionários sejam transferidos para um esforço de engenharia. Nenhum preenchimento é fornecido e os tempos de resolução de problemas aumentam. Essas informações não são capturadas e a liderança toma conhecimento do problema somente após várias semanas de comentários de insatisfação do usuário.

Benefícios de implementar esta prática recomendada: durante eventos operacionais em que a empresa é afetada, muito tempo e energia podem ser desperdiçados com a consulta de informações por várias equipes em uma tentativa de entender a situação. Ao estabelecer páginas de status e painéis amplamente divulgados, as partes interessadas podem obter rapidamente informações, como se um problema foi detectado ou não, quem liderou o problema ou quando é esperado um retorno às operações normais. Isso permite que os membros da equipe dediquem mais tempo à resolução de problemas e passem menos tempo comunicando o status a outras pessoas.

Além disso, painéis e relatórios podem fornecer informações aos tomadores de decisão e às partes interessadas para ver como as equipes de operações são capazes de responder às necessidades de negócios e como seus recursos estão sendo alocados. Isso é crucial para determinar se os recursos adequados estão disponíveis para apoiar os negócios.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Crie painéis que mostrem as principais métricas atuais para suas equipes de operações e as torne facilmente acessíveis, tanto para os líderes de operações quanto para a gerência.

Crie páginas de status que possam ser atualizadas rapidamente para mostrar quando um incidente ou evento está ocorrendo, quem é o proprietário e quem está coordenando a resposta. Compartilhe todas as etapas ou soluções alternativas que os usuários devem considerar nesta página e divulgue amplamente a localização. Incentive os usuários a verificar esse local primeiro quando confrontados com um problema desconhecido.

Colete e forneça relatórios que mostrem a integridade das operações ao longo do tempo e distribua-os aos líderes e tomadores de decisão para ilustrar o trabalho das operações junto com os desafios e as necessidades.

Compartilhe entre as equipes essas métricas e relatórios que melhor refletem as metas e os KPIs e onde eles foram influentes na promoção da mudança. Dedique tempo a essas atividades para aumentar a importância das operações dentro das equipes e entre elas.

Recursos

Documentos relacionados:

- Avaliar o progresso
- Criar painéis para visibilidade da operação

Soluções relacionadas:

- Operações de dados

OPS09-BP03 Revisar as métricas operacionais e priorizar melhorias

Reservar tempo e dedicar recursos para analisar o estado das operações garante que atender à linha de negócios do dia a dia continue sendo uma prioridade. Reúna líderes de operações e partes interessadas para revisar regularmente as métricas, reafirmar ou modificar metas e objetivos e priorizar melhorias.

Resultado desejado:

- Os líderes de operações e a equipe se reúnem regularmente para revisar as métricas durante um determinado período de relatório. Os desafios são comunicados, as vitórias são celebradas e as lições aprendidas são compartilhadas.
- As partes interessadas e os líderes de negócios são regularmente informados sobre o estado das operações e solicitados a fornecer informações sobre metas, KPIs e iniciativas futuras. As compensações entre prestação de serviços, operações e manutenção são discutidas e contextualizadas.

Práticas comuns que devem ser evitadas:

- Um novo produto é lançado, mas as equipes operacionais de nível 1 e nível 2 não são adequadamente treinadas para prestar suporte nem recebem pessoal adicional. Métricas que mostram a diminuição nos tempos de resolução de tíquetes e o aumento nos volumes de incidentes não são vistas pelos líderes. Uma ação é tomada semanas depois, quando os números de assinaturas começam a cair à medida que usuários insatisfeitos saem da plataforma.
- Um processo manual para realizar a manutenção de uma workload está em vigor há muito tempo. Embora o desejo de automatizar estivesse presente, essa era uma prioridade baixa, considerando a baixa importância do sistema. No entanto, com o tempo, o sistema cresceu em importância e

agora esses processos manuais consomem a maior parte do tempo das operações. Nenhum recurso está agendado para fornecer mais ferramentas às operações, causando o esgotamento da equipe à medida que as workloads aumentam. A liderança percebe o que está acontecendo quando é relatado que funcionários estão indo trabalhar para outros concorrentes.

Benefícios de implementar esta prática recomendada: em algumas organizações, pode ser um desafio alocar o mesmo tempo e atenção dedicados à prestação de serviços e a novos produtos ou ofertas. Quando isso ocorre, a linha de negócios pode sofrer enquanto o nível de serviço esperado se deteriora lentamente. Isso ocorre porque as operações não mudam e evoluem com o crescimento dos negócios e logo podem ser deixadas para trás. Sem uma análise regular dos insights que as operações coletam, o risco para a empresa pode se tornar visível somente quando for tarde demais. Ao alocar tempo para revisar métricas e procedimentos tanto entre a equipe de operações quanto com a liderança, o papel crucial que as operações desempenham permanece visível e os riscos podem ser identificados muito antes de atingirem níveis críticos. As equipes de operações obtêm uma visão melhor das mudanças e iniciativas comerciais iminentes, permitindo que esforços proativos sejam realizados. A visibilidade da liderança nas métricas operacionais mostra o papel que essas equipes desempenham na satisfação do cliente, tanto interno quanto externo, e permite que ela avalie melhor as opções de prioridades ou garanta que as operações tenham tempo e recursos para mudar e evoluir com novas iniciativas de negócios e workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Dedique tempo para analisar as métricas operacionais entre as partes interessadas e as equipes operacionais e analisar os dados do relatório. Coloque esses relatórios nos contextos das metas e objetivos da organização para determinar se eles estão sendo cumpridos. Identifique fontes de ambiguidade onde as metas não são claras ou onde pode haver conflitos entre o que é pedido e o que é fornecido.

Identifique onde o tempo, as pessoas e as ferramentas podem ajudar nos resultados das operações. Determine quais KPIs isso afetaria e quais deveriam ser as metas de sucesso. Revise regularmente para garantir que as operações tenham recursos suficientes para apoiar a linha de negócios.

Recursos

Documentos relacionados:

- Amazon Athena

- Referência de métricas e dimensões do Amazon CloudWatch
- Amazon QuickSight
- AWS Glue
- AWS Glue Data Catalog
- Coletar métricas e logs de instâncias do Amazon EC2 e servidores on-premises com o agente do Amazon CloudWatch
- Usar métricas do Amazon CloudWatch

OPS 10. Como gerenciar os eventos de workload e operações?

Prepare e valide procedimentos para responder a eventos, com o objetivo de minimizar a interrupção de sua workload.

Práticas recomendadas

- OPS10-BP01 Usar um processo para gerenciamento de eventos, incidentes e problemas
- OPS10-BP02 Ter um processo por alerta
- OPS10-BP03 Priorizar eventos operacionais com base no impacto nos negócios
- OPS10-BP04 Definir caminhos de escalação
- OPS10-BP05 Definir um plano de comunicação com o cliente para interrupções
- OPS10-BP06 Comunicar o status por meio de painéis
- OPS10-BP07 Automatizar respostas a eventos

OPS10-BP01 Usar um processo para gerenciamento de eventos, incidentes e problemas

O gerenciamento eficiente de eventos, incidentes e problemas é fundamental para manter a integridade e a performance da workload. É crucial reconhecer e compreender as diferenças entre esses elementos para desenvolver uma estratégia eficaz de resposta e resolução. Estabelecer e seguir um processo bem definido para cada aspecto ajuda sua equipe a lidar de forma rápida e eficaz com qualquer desafio operacional que surgir.

Resultado desejado: sua organização gerencia com eficiência eventos, incidentes e problemas operacionais por meio de processos bem documentados e armazenados de maneira centralizada. Esses processos são atualizados de forma consistente para refletir as mudanças, simplificando o manuseio e mantendo a alta confiabilidade do serviço e a performance da workload.

Práticas comuns que devem ser evitadas:

- Você responde de forma reativa, em vez de proativa, aos eventos.
- Abordagens inconsistentes são adotadas para diferentes tipos de eventos ou incidentes.
- Sua organização não analisa e nem aprende com os incidentes para evitar futuras ocorrências.

Benefícios de implementar esta prática recomendada:

- Processos de resposta simplificados e padronizados.
- Impacto reduzido dos incidentes nos serviços e nos clientes.
- Resolução rápida de problemas.
- Melhoria contínua nos processos operacionais.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Implementar essa prática recomendada significa que você está monitorando os eventos da workload. Você tem processos para lidar com incidentes e problemas. Os processos são documentados, compartilhados e atualizados com frequência. Os problemas são identificados, priorizados e corrigidos.

Compreender eventos, incidentes e problemas

- Eventos: um evento pode ser uma observação de uma ação, ocorrência ou alteração de estado. Os eventos podem ser planejados ou não e podem ter origens internas ou externas à workload.
- Incidentes: os incidentes são eventos que exigem uma resposta, como interrupções não planejadas ou degradações da qualidade do serviço. Eles representam interrupções que precisam de atenção imediata para restaurar a operação normal da workload.
- Problemas: problemas são as causas subjacentes de um ou mais incidentes. Identificar e resolver problemas envolve aprofundar-se nos incidentes para evitar futuras ocorrências.

Etapas de implementação

Eventos

1. Monitorar eventos:

- Implemente a observabilidade e utilize a observabilidade da workload.
- As ações de monitoramento tomadas por um usuário, função ou serviço da AWS são registradas como eventos no AWS CloudTrail.
- Responda às mudanças operacionais em suas aplicações em tempo real com o Amazon EventBridge.
- Avalie, monitore e registre constantemente as alterações na configuração de recursos com o AWS Config.

2. Criar processos:
   - Desenvolva um processo para avaliar quais eventos são importantes e exigem monitoramento. Isso envolve definir limites e parâmetros para atividades normais e anormais.
   - Determine os critérios que transformam um evento em um incidente. Isso pode ser baseado na gravidade, no impacto nos usuários ou no desvio do comportamento esperado.
   - Analise regularmente os processos de monitoramento e resposta a eventos. Isso inclui analisar incidentes anteriores, ajustar limites e refinar os mecanismos de alerta.

Incidentes

1. Responder a incidentes:
   - Use insights das ferramentas de observabilidade para identificar e responder rapidamente a incidentes.
   - Implemente o Ops Center do AWS Systems Manager para agregar, organizar e priorizar itens e incidentes operacionais.
   - Use serviços como o Amazon CloudWatch e o AWS X-Ray para análises e soluções de problemas mais aprofundadas.
   - Considere o AWS Managed Services (AMS) para melhorar o gerenciamento de incidentes, aproveitando suas capacidades proativas, preventivas e de detecção. O AMS amplia o suporte operacional com serviços como monitoramento, detecção e resposta a incidentes e gerenciamento de segurança.
   - Os clientes Enterprise Support podem usar a Detecção e Resposta a Incidentes da AWS, que fornece monitoramento proativo e gerenciamento de incidentes contínuos para workloads de produção.
2. Criar um processo de gerenciamento de incidentes:

- Estabeleça um processo estruturado de gerenciamento de incidentes, incluindo funções claras, protocolos de comunicação e etapas para resolução.
- Integre o gerenciamento de incidentes com ferramentas como o AWS Chatbot para usufruir de respostas e coordenação eficientes.
- Categorize os incidentes por gravidade, com planos de resposta a incidentes predefinidos para cada categoria.

3. Aprender e melhorar:
   - Conduza análises pós-incidentes para entender as causas-raiz e a eficácia da resolução.
   - Atualize e melhore constantemente os planos de resposta com base em análises e práticas em evolução.
   - Documente e compartilhe as lições aprendidas entre as equipes para melhorar a resiliência operacional.
   - Os clientes Enterprise Support podem solicitar o workshop Gerenciamento de incidentes ao respectivo gerente técnico da conta. Esse workshop guiado testa seu plano de resposta a incidentes e ajuda você a identificar áreas para melhoria.

Problemas

1. Identificar problemas:
   - Use dados de incidentes anteriores para identificar padrões recorrentes que possam indicar problemas sistêmicos mais profundos.
   - Utilize ferramentas como o AWS CloudTrail e o Amazon CloudWatch para analisar tendências e descobrir problemas subjacentes.
   - Envolva equipes multifuncionais, incluindo operações, desenvolvimento e unidades de negócios, para obter perspectivas diversas sobre as causas principais dos problemas.
2. Criar um processo de gerenciamento de problemas:
   - Desenvolva um processo estruturado para gerenciamento de problemas com foco em soluções de longo prazo em vez de soluções rápidas.
   - Incorpore técnicas de análise das causas-raiz (RCA) para investigar e compreender as causas subjacentes dos incidentes.
   - Atualize políticas, procedimentos e infraestrutura operacionais com base nas descobertas para evitar recorrência.
3. Continuar melhorando:

- Promova uma cultura de aprendizado e aprimoramento constantes, incentivando as equipes a identificar e resolver possíveis problemas de forma proativa.
- Analise e revise regularmente os processos e ferramentas de gerenciamento de problemas para se alinhar aos cenários de negócios e tecnologia em evolução.
- Compartilhe insights e práticas recomendadas em toda a organização para criar um ambiente operacional mais resiliente e eficiente.

4. Envolver o AWS Support:
    - Use os recursos de suporte da AWS, como o AWS Trusted Advisor, para receber orientação proativa e recomendações de otimização.
    - Os clientes Enterprise Support podem acessar programas especializados, como o AWS Countdown, para obter suporte durante eventos críticos.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Identificar indicadores-chave de performance
- OPS04-BP02 Implementar a telemetria de aplicações
- OPS07-BP03 Usar runbooks para realizar procedimentos
- OPS07-BP04 Usar playbooks para investigar problemas
- OPS08-BP01 Analisar métricas da workload
- OPS11-BP02 Executar análise pós-incidente

Documentos relacionados:

- Guia de resposta a incidentes de segurança da AWS
- Detecção e resposta a incidentes da AWS
- Framework de adoção da Nuvem AWS: perspectiva de operações: gerenciamento de incidentes e problemas
- Gerenciamento de incidentes na era de DevOps e SRE
- PagerDuty: o que é gerenciamento de incidentes?

Vídeos relacionados:

- As principais dicas de resposta a incidentes da AWS
- AWS re:Invent 2022: Amazon Builders' Library: 25 anos de excelência operacional da Amazon
- AWS re:Invent 2022: Detecção e resposta a incidentes na AWS (SUP201)
- Introdução ao AWS Systems Manager Incident Manager

Exemplos relacionados:

- Serviços proativos da AWS: workshop de gerenciamento de incidentes
- Como automatizar a resposta a incidentes com o PagerDuty e o AWS Systems Manager Incident Manager
- Engajar os respondedores de incicentes com escalas de plantão na AWS Systems Manager Incident Manager
- Melhorar a visibilidade e a colaboração durante o tratamento de incidentes na AWS Systems Manager Incident Manager
- Relatórios de incidentes e solicitações de serviço no AMS

Serviços relacionados:

- Amazon EventBridge

OPS10-BP02 Ter um processo por alerta

Estabelecer um processo claro e definido para cada alerta em seu sistema é essencial para um gerenciamento eficaz e eficiente de incidentes. Essa prática garante que cada alerta leve a uma resposta específica e acionável, melhorando a confiabilidade e a capacidade de resposta de suas operações.

Resultado desejado: cada alerta inicia um plano de resposta específico e bem definido. Sempre que possível, as respostas são automatizadas, com propriedade clara e um caminho de escalação definido. Os alertas estão vinculados a uma base de conhecimento atualizada para que qualquer operador possa responder de forma consistente e eficaz. As respostas são rápidas e uniformes em todos os setores, aumentando a eficiência e a confiabilidade operacionais.

Práticas comuns que devem ser evitadas:

- Os alertas não têm um processo de resposta predefinido, o que leva a resoluções improvisadas e atrasadas.
- A sobrecarga de alertas faz com que alertas importantes sejam ignorados.
- Os alertas são tratados de forma inconsistente devido à falta de propriedade e responsabilidade claras.

Benefícios de implementar esta prática recomendada:

- Redução da fadiga dos alertas ao gerar apenas alertas acionáveis.
- Diminuição do tempo médio de resolução (MTTR) para problemas operacionais.
- Diminuição do tempo médio de investigação (MTTI), o que ajuda a reduzir o MTTR.
- Capacidade aprimorada para escalar respostas operacionais.
- Consistência e confiabilidade aprimoradas no tratamento de eventos operacionais.

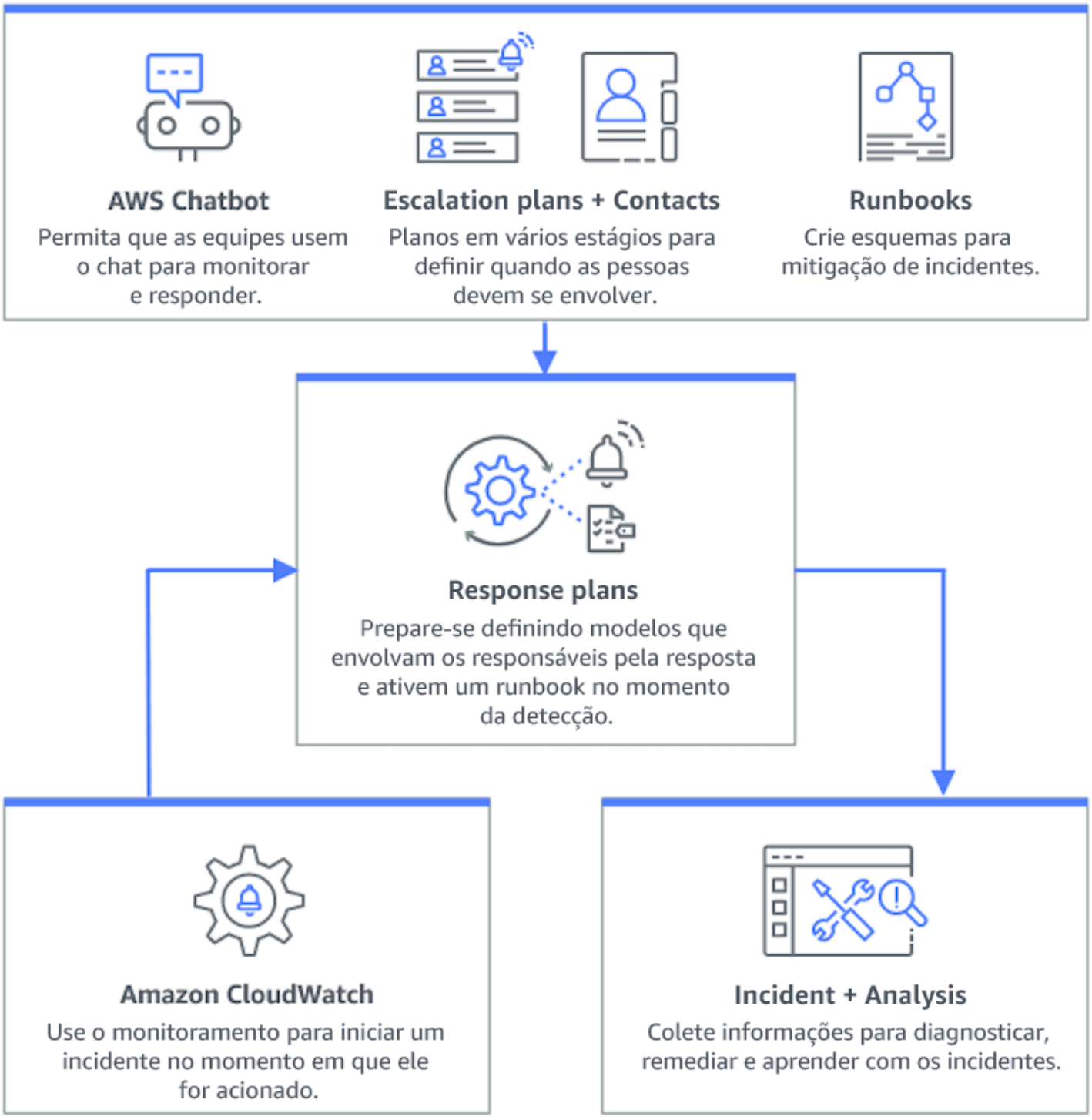
Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Ter um processo por alerta envolve estabelecer um plano de resposta claro para cada alerta, automatizar as respostas sempre que possível e refinar constantemente esses processos com base no feedback operacional e nos requisitos em evolução.

Etapas de implementação

O seguinte diagrama ilustra o fluxo de trabalho de gerenciamento de incidentes dentro do AWS Systems Manager Incident Manager. Ele foi projetado para responder rapidamente a problemas operacionais, criando automaticamente incidentes em resposta a eventos específicos do Amazon CloudWatch ou Amazon EventBridge. Quando um incidente é criado, automática ou manualmente, o Incident Manager centraliza o gerenciamento do incidente, organiza as informações relevantes dos recursos da AWS e inicia planos de resposta predefinidos. Isso inclui executar runbooks de automação do Systems Manager Automation para ação imediata, bem como criar um item de trabalho operacional principal no OpsCenter para rastrear tarefas e análises relacionadas. Esse processo simplificado acelera e coordena a resposta a incidentes em todo o seu ambiente da AWS.



1. Use alarmes compostos: crie alarmes compostos no CloudWatch para agrupar alarmes relacionados, reduzindo o ruído e permitindo respostas mais significativas.
2. Integre os alarmes do Amazon CloudWatch ao Incident Manager: configure os alarmes do CloudWatch para criar automaticamente incidentes no AWS Systems Manager Incident Manager.

3. Integre o Amazon EventBridge ao Incident Manager: crie regras do EventBridge para reagir a eventos e criar incidentes usando planos de resposta definidos.
4. Prepare-se para incidentes no Incident Manager:
    - Estabeleça planos de resposta detalhados no Incident Manager para cada tipo de alerta.
    - Estabeleça canais de chat via AWS Chatbot conectados aos planos de resposta no Incident Manager, facilitando a comunicação em tempo real durante incidentes em plataformas como Slack, Microsoft Teams e Amazon Chime.
    - Incorpore os runbooks do Systems Manager Automation no Incident Manager para gerar respostas automatizadas aos incidentes.

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Identificar indicadores-chave de performance
- OPS08-BP04 Criar alertas acionáveis

Documentos relacionados:

- Framework de adoção da Nuvem AWS: perspectiva de operações: gerenciamento de incidentes e problemas
- Usar alarmes do Amazon CloudWatch
- Configurar o AWS Systems Manager Incident Manager
- Como se preparar para incidentes no Incident Manager

Vídeos relacionados:

- As principais dicas de resposta a incidentes da AWS

Exemplos relacionados:

- Workshops da AWS: AWS Systems Manager Incident Manager – Automatizar a resposta a incidentes em eventos de segurança

OPS10-BP03 Priorizar eventos operacionais com base no impacto nos negócios

Responder prontamente aos eventos operacionais é fundamental, mas nem todos os eventos são iguais. Ao priorizar com base no impacto nos negócios, você também prioriza o tratamento de eventos com o potencial de graves consequências, como segurança, perdas financeiras, violações regulatórias ou danos à reputação.

Resultado desejado: as respostas aos eventos operacionais são priorizadas com base no possível impacto nas operações e nos objetivos de negócios. Isso torna as respostas eficientes e eficazes.

Práticas comuns que devem ser evitadas:

- Cada evento é tratado com o mesmo nível de urgência, causando confusão e atrasos na resolução de problemas críticos.
- Você não consegue distinguir entre eventos de alto e baixo impacto, o que leva à má alocação de recursos.
- Sua organização carece de uma estrutura de priorização clara, o que acarreta em respostas inconsistentes aos eventos operacionais.
- Os eventos são priorizados com base na ordem em que são relatados, e não em seu impacto nos resultados de negócios.

Benefícios de implementar esta prática recomendada:

- Garante que as funções críticas da empresa recebam atenção em primeiro lugar, minimizando possíveis danos.
- Melhora a alocação de recursos durante vários eventos simultâneos.
- Melhora a capacidade da organização de manter a confiança e atender aos requisitos regulatórios.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Em se tratando de vários eventos operacionais, uma abordagem estruturada de priorização com base no impacto e na urgência é essencial. Essa abordagem ajuda você a tomar decisões embasadas, direcionar esforços para onde eles são mais necessários e reduzir o risco à continuidade dos negócios.

Etapas de implementação

1. Avalie o impacto: desenvolva um sistema de classificação para avaliar a gravidade dos eventos em termos de possível impacto nas operações e nos objetivos de negócios. O seguinte exemplo mostra as categorias de impacto:

Nível de impacto	Descrição
Alta	Afeta muitos funcionários ou clientes, tem alto impacto financeiro, alto dano à reputação ou ferimentos.
Médio	Afeta grupos de funcionários ou clientes, tem impacto financeiro moderado ou dano moderado à reputação.
Baixo	Afeta funcionários ou clientes individualmente, tem baixo impacto financeiro ou baixo dano à reputação.

2. Avalie a urgência: defina os níveis de urgência da rapidez com que um evento precisa de uma resposta, considerando fatores como segurança, implicações financeiras e acordos de serviço (SLAs). O seguinte exemplo demonstra as categorias de urgência:

Nível de urgência	Descrição
Alta	Aumento exponencial dos danos, impacto no trabalho urgente, escalação iminente ou usuários ou grupos VIP afetados.
Médio	Os danos aumentam com o tempo ou um único usuário ou grupo VIP é afetado.
Baixo	Os danos marginais aumentam com o tempo ou trabalho não urgente é afetado.

3. Crie uma matriz de priorização:
   - Use uma matriz para fazer a referência cruzada das informações sobre impacto e urgência, atribuindo níveis de prioridade a diferentes combinações.

- Torne a matriz acessível e capaz de ser compreendida por todos os membros da equipe responsáveis pelas respostas aos eventos operacionais.
- O seguinte exemplo de matriz exibe a gravidade do incidente de acordo com a urgência e o impacto:

Urgência e impacto	Alta	Médio	Baixo
Alta	Crítico	Urgente	Alta
Médio	Urgente	Alta	Normal
Baixo	Alta	Normal	Baixo

4. Treine e comunique: treine as equipes de resposta sobre a matriz de priorização e a importância de segui-la durante um evento. Comunique o processo de priorização a todas as partes interessadas para definir expectativas claras.
5. Integre à resposta a incidentes:
   - Incorpore a matriz de priorização em seus planos e ferramentas de resposta a incidentes.
   - Automatize a classificação e a priorização de eventos sempre que possível para acelerar os tempos de resposta.
   - Os clientes Enterprise Support podem usar a Detecção e Resposta a Incidentes da AWS, que fornece monitoramento proativo e gerenciamento de incidentes contínuos para workloads de produção.
6. Revise e adapte: analise regularmente a eficácia do processo de priorização e faça ajustes com base no feedback e nas mudanças no ambiente de negócios.

Recursos

Práticas recomendadas relacionadas:

- OPS03-BP03 Incentivo à escalação
- OPS08-BP04 Criar alertas acionáveis
- OPS09-BP01 Medir metas operacionais e KPIs com métricas

Documentos relacionados:

- Atlassian: como entender os níveis de severidade dos incidentes
- Mapa de processos de TI: prioridade de incidentes na lista de verificação

OPS10-BP04 Definir caminhos de escalação

Estabeleça caminhos claros de escalação em seus protocolos de resposta a incidentes para facilitar ações rápidas e eficazes. Isso inclui especificar solicitações de escalação, detalhar o processo de escalação e pré-aprovar ações para agilizar a tomada de decisões e reduzir o tempo médio de resolução (MTTR).

Resultado desejado: um processo estruturado e eficiente que encaminha os incidentes para a equipe apropriada, minimizando os tempos de resposta e o impacto.

Práticas comuns que devem ser evitadas:

- A falta de clareza sobre os procedimentos de recuperação leva a respostas improvisadas durante incidentes críticos.
- A ausência de permissões e propriedade definidas ocasiona atrasos quando uma ação urgente é necessária.
- As partes interessadas e os clientes não são informados de acordo com as expectativas.
- Decisões importantes estão atrasadas.

Benefícios de implementar esta prática recomendada:

- Resposta simplificada a incidentes por meio de procedimentos de escalação predefinidos.
- Tempo de inatividade reduzido com ações pré-aprovadas e propriedade clara.
- Melhor alocação de recursos e ajustes no nível de suporte de acordo com a gravidade do incidente.
- Comunicação aprimorada com as partes interessadas e os clientes.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Caminhos de escalação definidos adequadamente são cruciais para uma resposta rápida a incidentes. O AWS Systems Manager Incident Manager oferece suporte à configuração de planos de

escalação estruturada e agendamentos de plantão, que alertam a equipe certa para que ela esteja pronta para agir quando ocorrerem incidentes.

Etapas de implementação

1. Configure solicitações de escalação: configure os alarmes do CloudWatch para criar um incidente no AWS Systems Manager Incident Manager.
2. Configure escalas de plantão: crie escalas de plantão no Incident Manager que se alinhem aos seus caminhos de escalação. Equipe o pessoal de plantão com as permissões e ferramentas necessárias para agir rapidamente.
3. Detalhe os procedimentos detalhados de escalação:
   - Determine as condições específicas sob as quais um incidente deve ser escalado.
   - Crie planos de escalação no Incident Manager.
   - Os canais de escalação devem consistir em um contato ou em uma escala de plantão.
   - Defina as funções e responsabilidades da equipe em cada nível de escalação.
4. Aprove previamente as ações de mitigação: colabore com os tomadores de decisão para pré-aprovar ações para cenários previstos. Use runbooks do Systems Manager Automation integrados ao Incident Manager para acelerar a resolução de incidentes.
5. Especifique a propriedade: identifique claramente os proprietários internos de cada etapa do caminho de escalação.
6. Detalhe as escalações de terceiros:
   - Documente os acordos de serviço (SLAs) de terceiros e alinhe-os às metas internas.
   - Defina protocolos claros para a comunicação com o fornecedor durante incidentes.
   - Integre os contatos do fornecedor às ferramentas de gerenciamento de incidentes para acesso direto.
   - Realize exercícios regulares que incluam cenários de resposta de terceiros.
   - Mantenha as informações de escalação de fornecedores bem documentadas e facilmente acessíveis.
7. Treine e ensaie os planos de escalação: treine sua equipe no processo de escalação e realize exercícios regulares de resposta a incidentes ou encenações. Os clientes Enterprise Support podem solicitar um workshop sobre gerenciamento de incidentes.
8. Continue a aprimorar: analise com frequência a eficácia de seus caminhos de escalação. Atualize seus processos com base nas lições aprendidas com os post-mortems de incidentes e com o feedback contínuo.

Nível de esforço do plano de implementação: Moderado

Recursos

Práticas recomendadas relacionadas:

- OPS08-BP04 Criar alertas acionáveis
- OPS10-BP02 Ter um processo por alerta
- OPS11-BP02 Executar análise pós-incidente

Documentos relacionados:

- Planos de escalação da AWS Systems Manager Incident Manager
- Como trabalhar com escalas de plantão no Incident Manager
- Criar e gerenciar runbooks
- Gerenciamento de acesso elevado temporário com o AWS IAM Identity Center
- Atlassian: políticas de escalação para o gerenciamento efetivo de incidentes

OPS10-BP05 Definir um plano de comunicação com o cliente para interrupções

A comunicação eficaz durante interrupções é fundamental para manter a confiança e a transparência com os clientes. Um plano de comunicação bem definido ajuda sua organização a compartilhar informações de forma rápida e clara, interna e externamente, durante incidentes.

Resultado desejado:

- Um plano de comunicação robusto que informa de maneira eficaz os clientes e as partes interessadas sobre interrupções.
- Transparência na comunicação para criar confiança e reduzir a ansiedade do cliente.
- Minimiza o impacto das interrupções na experiência do cliente e nas operações comerciais.

Práticas comuns que devem ser evitadas:

- A comunicação inadequada ou atrasada leva à confusão e insatisfação do cliente.
- Mensagens excessivamente técnicas ou vagas não transmitem o impacto real sobre os usuários.
- Não há uma estratégia de comunicação predefinida, resultando em mensagens inconsistentes e reativas.

Benefícios de implementar esta prática recomendada:

- Maior confiança e satisfação do cliente por meio de uma comunicação proativa e clara.
- Redução da carga depositada sobre as equipes de suporte ao abordar de maneira preventiva as preocupações dos clientes.
- Gerenciamento e recuperação mais eficazes depois de incidentes.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A criação de um plano de comunicação abrangente para interrupções envolve vários aspectos, desde a escolha dos canais certos até a elaboração da mensagem e do tom. O plano deve tanto ser adaptável e escalável quanto atender a diferentes cenários de interrupção.

Etapas de implementação

1. Defina perfis e responsabilidades:
    - Atribua um gerente de incidentes principal para supervisionar as atividades de resposta a incidentes.
    - Atribua um gerente de comunicações responsável por coordenar todas as comunicações externas e internas.
    - Inclua o gerente de suporte para fornecer comunicação consistente por meio de tíquetes de suporte.
2. Identifique os canais de comunicação: selecione canais como chat interno, e-mail, SMS, redes sociais, notificações na aplicação e páginas de status. Esses canais devem ser resilientes e funcionar de forma independente durante interrupções.
3. Comunique-se com os clientes de forma rápida, clara e frequente:
    - Desenvolva modelos para vários cenários de comprometimento do serviço, enfatizando a simplicidade e os detalhes essenciais. Inclua informações sobre a deficiência do serviço, o tempo esperado de resolução e o impacto.
    - Use o Amazon Pinpoint para alertar os clientes usando notificações push, notificações na aplicação, e-mails, mensagens de texto, mensagens de voz e mensagens em canais personalizados.

- Use o Amazon Simple Notification Service (Amazon SNS) para alertar os assinantes programaticamente ou por e-mail, notificações push em telefones celulares e mensagens de texto.
- Comunique o status por meio de um painel público do Amazon CloudWatch.
- Incentive o engajamento nas redes sociais:
  - Monitore ativamente as redes sociais para entender a percepção do cliente.
  - Publique em plataformas de rede social para fazer atualizações públicas e engajar a comunidade.
  - Prepare modelos para uma comunicação consistente e clara nas redes sociais.

4. Coordene a comunicação interna: implemente protocolos internos usando ferramentas como o AWS Chatbot para coordenação e comunicação de equipes. Use os painéis do CloudWatch para comunicar o status.
5. Organize a comunicação com ferramentas e serviços dedicados:
   - Use o AWS Systems Manager Incident Manager com o AWS Chatbot para configurar canais de chat dedicados para comunicação e coordenação internas em tempo real durante incidentes.
   - Use os runbooks do AWS Systems Manager Incident Manager para automatizar as notificações enviadas aos clientes por meio do Amazon Pinpoint, do Amazon SNS ou de ferramentas de terceiros, como plataformas de rede social, durante incidentes.
   - Incorpore fluxos de trabalho de aprovação nos runbooks para, opcionalmente, revisar e autorizar todas as comunicações externas antes do envio.
6. Pratique e melhore:
   - Realize treinamentos sobre o uso de ferramentas e estratégias de comunicação. Capacite as equipes a tomar decisões rápidas durante incidentes.
   - Teste o plano de comunicação por meio de exercícios ou game days frequentes. Use esses testes para refinar as mensagens e avaliar a eficácia dos canais.
   - Implemente mecanismos de feedback para avaliar a eficácia da comunicação durante incidentes. Continue desenvolvendo o plano de comunicação com base no feedback e nas mudanças necessárias.

Nível de esforço do plano de implementação: Alto

Recursos

Práticas recomendadas relacionadas:

- OPS07-BP03 Usar runbooks para realizar procedimentos
- OPS10-BP06 Comunicar o status por meio de painéis
- OPS11-BP02 Executar análise pós-incidente

Documentos relacionados:

- Atlassian: práticas recomendadas de comunicação de incidentes
- Atlassian: como redigir uma boa atualização de status
- PagerDuty: um guia para a comunicação de incidentes

Vídeos relacionados:

- Atlassian: criar seu próprio plano de comunicação de incidentes: modelos de incidentes

Exemplos relacionados:

- Painel do AWS Health
- Exemplos de atualizações de status da AWS

OPS10-BP06 Comunicar o status por meio de painéis

Use painéis como uma ferramenta estratégica para transmitir o status operacional em tempo real e as principais métricas para diferentes públicos, incluindo equipes técnicas internas, liderança e clientes. Esses painéis oferecem uma representação visual centralizada da integridade do sistema e da performance dos negócios, aumentando a transparência e a eficiência na tomada de decisões.

Resultado desejado:

- Os painéis fornecem uma visão abrangente do sistema e das métricas comerciais relevantes para diferentes partes interessadas.
- As partes interessadas podem acessar as informações operacionais de forma proativa, reduzindo a necessidade de solicitações frequentes de status.
- A tomada de decisões em tempo real é aprimorada durante operações e incidentes normais.

Práticas comuns que devem ser evitadas:

- Os engenheiros que participam de uma chamada de gerenciamento de incidentes precisam de atualizações de status para se atualizarem.
- Confiar em relatórios manuais para gerenciamento, o que leva a atrasos e possíveis imprecisões.
- As equipes de operações são frequentemente interrompidas para atualizações de status durante incidentes.

Benefícios de implementar esta prática recomendada:

- Capacita as partes interessadas com acesso imediato a informações críticas, promovendo a tomada de decisões embasada.
- Reduz as ineficiências operacionais minimizando os relatórios manuais e as frequentes consultas de status.
- Aumenta a transparência e a confiança por meio da visibilidade em tempo real da performance do sistema e das métricas de negócios.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Os painéis comunicam com eficácia o status dos sistemas e métricas de negócios e podem ser adaptados às necessidades de diferentes grupos de público. Ferramentas como os painéis do Amazon CloudWatch e o Amazon QuickSight ajudam você a criar painéis interativos em tempo real para monitoramento de sistemas e business intelligence.

Etapas de implementação

1. Identifique as necessidades das partes interessadas: determine as necessidades específicas de informações de diferentes grupos de público, como equipes técnicas, liderança e clientes.
2. Escolha as ferramentas certas: selecione as ferramentas apropriadas, como os painéis do Amazon CloudWatch para monitoramento do sistema e o Amazon QuickSight para inteligência comercial interativa.
3. Crie painéis eficazes:
    - Crie painéis para apresentar claramente métricas e KPIs relevantes, garantindo que sejam compreensíveis e acionáveis.
    - Incorpore visualizações em nível de sistema e de negócios, conforme necessário.

- Inclua painéis de alto nível (para visões gerais amplas) e de baixo nível (para análises detalhadas).
- Integre alarmes automatizados em painéis para destacar problemas críticos.
- Anote painéis com métricas, limites e metas importantes para visibilidade imediata.

4. Integre fontes de dados:
   - Use o Amazon CloudWatch para agregar e exibir métricas de vários serviços da AWS e consultar métricas de outras fontes de dados, criando uma visão unificada das métricas comerciais e de integridade do seu sistema.
   - Use recursos como o CloudWatch Logs Insights para consultar e visualizar dados de log de diferentes aplicações e serviços.
5. Forneça acesso por autoatendimento:
   - Compartilhe os painéis do CloudWatch com partes interessadas relevantes para acessar informações por autoatendimento usando recursos de compartilhamento de painéis.
   - Garanta que os painéis sejam facilmente acessíveis e forneçam informações atualizadas e em tempo real.
6. Atualize e refine com frequência:
   - Atualize e refine constantemente os painéis para se alinharem às necessidades comerciais em evolução e ao feedback das partes interessadas.
   - Analise com frequência os painéis para mantê-los relevantes e eficazes a fim de transmitir as informações necessárias.

Recursos

Práticas recomendadas relacionadas:

- OPS08-BP05 Criar painéis

Documentos relacionados:

- Criar painéis para visibilidade operacional
- Usar painéis do Amazon CloudWatch
- Criar painéis flexíveis com variáveis de painel
- Compartilhar painéis do CloudWatch
- Métricas de consulta de outras fontes de dados

- Adicionar um widget personalizado a um painel do CloudWatch

Exemplos relacionados:

- Workshop One Observability: painéis

OPS10-BP07 Automatizar respostas a eventos

Automatizar as respostas a eventos é essencial para operações rápidas, consistentes e sem erros. Crie processos simplificados e use ferramentas para gerenciar e responder automaticamente aos eventos, minimizando as intervenções manuais e aprimorando a eficácia operacional.

Resultado desejado:

- Redução de erros humanos e tempos de resolução mais rápidos por meio de automação.
- Tratamento de eventos operacionais consistente e confiável.
- Eficiência operacional e confiabilidade do sistema aprimoradas.

Práticas comuns que devem ser evitadas:

- O tratamento manual de eventos leva a atrasos e erros.
- A automação é negligenciada em tarefas críticas e repetitivas.
- Tarefas manuais repetitivas levam à fadiga de alertas e à negligência de problemas críticos.

Benefícios de implementar esta prática recomendada:

- Aceleração das respostas aos eventos, reduzindo o tempo de inatividade do sistema.
- Operações confiáveis com tratamento automatizado e consistente de eventos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Incorpore a automação para criar fluxos de trabalho operacionais eficientes e minimizar as intervenções manuais.

Etapas de implementação

1. Identifique oportunidades de automação: determine tarefas repetitivas a serem automatizadas, como correção de problemas, ampliação de tíquetes, gerenciamento de capacidade, ajuste de escala, implantações e testes.
2. Identifique prompts de automação:
   - Avalie e defina condições ou métricas específicas que iniciam respostas automatizadas usando ações de alarme do Amazon CloudWatch.
   - Use o Amazon EventBridge para responder a eventos em serviços da AWS, workloads personalizadas e aplicações SaaS.
   - Considere eventos de iniciação, como entradas de log específicas, limites de métricas de performance ou mudanças de estado em recursos da AWS.
3. Implemente a automação orientada por eventos:
   - Use os runbooks de automação do AWS Systems Manager para simplificar as tarefas de manutenção, implantação e correção.
   - A criação de incidentes no Incident Manager reúne e adiciona automaticamente detalhes sobre os recursos da AWS envolvidos no incidente.
   - Monitore proativamente as cotas usando o Quota Monitor para AWS.
   - Ajuste automaticamente a capacidade do AWS Auto Scaling para manter a disponibilidade e a performance.
   - Automatize os pipelines de desenvolvimento com o Amazon CodeCatalyst.
   - Faça um teste preliminar ou monitore continuamente endpoints e APIs usando monitoramento sintético.
4. Faça a mitigação de riscos por meio de automação:
   - Implemente respostas de segurança automatizadas para lidar rapidamente com os riscos.
   - Use o AWS Systems Manager State Manager para reduzir desvios de configuração.
   - Corrija os recursos não compatíveis automaticamente com o Regras do AWS Config

Nível de esforço do plano de implementação: Alto

Recursos

Práticas recomendadas relacionadas:

- OPS08-BP04 Criar alertas acionáveis

- OPS10-BP02 Ter um processo por alerta

Documentos relacionados:

- Usar runbooks do Systems Manager Automation com o Incident Manager
- Criar incidentes no Incident Manager
- Cotas de serviço da AWS
- Monitorar o uso de recursos e enviar notificações ao se aproximar das cotas
- AWS Auto Scaling
- O que é o Amazon CodeCatalyst?
- Usar alarmes do Amazon CloudWatch
- Usar ações de alarmes no Amazon CloudWatch
- Corrigir recursos fora de conformidade com o Regras do AWS Config
- Criar métricas de eventos de log usando filtros
- AWS Systems Manager State Manager

Vídeos relacionados:

- Criar runbooks de automação com o AWS Systems Manager
- Como automatizar operações de TI na AWS
- Regras de automação do AWS Security Hub
- Como começar seu projeto rapidamente com esquemas do Amazon CodeCatalyst

Exemplos relacionados:

- Tutorial do Amazon CodeCatalyst: Criar um projeto com o esquema de aplicação Web de três níveis moderna
- Workshop One Observability
- Responder a incidentes usando o Incident Manager

Evoluir

Pergunta

OPS 11. Como evoluir as operações?

Dedique tempo e recursos para a melhoria incremental praticamente contínua a fim de aumentar a eficácia e a eficiência das suas operações.

Práticas recomendadas

OPS11-BP01 Adotar um processo para melhoria contínua

Avalie a workload em relação às práticas recomendadas de arquitetura interna e externa. Realize análises frequentes e intencionais da workload. Priorize as oportunidades de melhoria na cadência de desenvolvimento de software.

Resultado desejado:

- Analise a workload em relação às práticas recomendadas de arquitetura com frequência.
- Atribua às oportunidades de melhoria a mesma prioridade que os recursos do processo de desenvolvimento de software.

Práticas comuns que devem ser evitadas:

- Não realizar uma análise de arquitetura na workload desde que ela foi implantada há vários anos.
- Atribuir uma prioridade menor às oportunidades de melhoria. Em comparação com os novos recursos, essas oportunidades permanecem pendentes.

- Não há um padrão para implementar modificações nas práticas recomendadas da organização.

Benefícios de implementar esta prática recomendada:

- A workload é mantida atualizada em relação às práticas recomendadas de arquitetura.
- Você desenvolveu a workload de forma intencional.
- Você pode utilizar as práticas recomendadas da organização para melhorar todas as workloads.
- Você tem ganhos marginais que têm um impacto cumulativo, o que gera maior eficiência.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Realize frequentemente uma revisão arquitetônica da workload. Usando práticas recomendadas internas e externas, avalie a workload e identifique oportunidades de melhoria. Priorize as oportunidades de melhoria na cadência de desenvolvimento de software.

Etapas de implementação

1. Realize revisões de arquitetura periódicas da workload de produção com uma frequência previamente combinada. Use um padrão de arquitetura documentado que inclua práticas recomendadas específicas da AWS.
   a. Use os padrões definidos internamente para essas avaliações. Se não houver um padrão interno, use o AWS Well-Architected Framework.
   b. Use o AWS Well-Architected Tool para criar uma perspectiva personalizada das práticas recomendadas internas e realizar a análise da arquitetura.
   c. Entre em contato com o arquiteto de soluções ou o gerente técnico de contas da AWS para realizar uma análise guiada do Well-Architected Framework para sua workload.
2. Priorize as oportunidades de melhoria identificadas durante a análise em seu processo de desenvolvimento de software.

Nível de esforço do plano de implementação: Baixo. É possível usar o AWS Well-Architected Framework para realizar sua análise de arquitetura anual.

Recursos

Práticas recomendadas relacionadas:

- OPS11-BP02 Executar análise pós-incidente
- OPS11-BP08 Documentar e compartilhar as lições aprendidas
- OPS04 Como implementar a observabilidade

Documentos relacionados:

- AWS Well-Architected Tool: perspectivas personalizadas
- Whitepaper do AWS Well-Architected: O processo de revisão
- Personalizar avaliações do Well-Architected usando Custom Lenses e o AWS Well-Architected Tool
- Implementar o ciclo de vida do AWS Well-Architected Custom Lenses em sua organização

Vídeos relacionados:

- Laboratórios do Well-Architected: Nível 100: Lentes Personalizadas no AWS Well-Architected Tool
- AWS re:Invent 2023: Como escalar as práticas recomendadas da AWS Well-Architected em toda a sua organização

Exemplos relacionados:

- AWS Well-Architected Tool

OPS11-BP02 Executar análise pós-incidente

Revise os eventos que afetam o cliente e identifique os fatores contribuintes e as ações preventivas. Use essas informações para desenvolver mitigações e limitar ou evitar recorrência. Desenvolva procedimentos para respostas rápidas e eficazes. Comunique os fatores contribuintes e as ações corretivas conforme apropriado, de acordo com o público-alvo.

Resultado desejado:

- Você estabelece processos de gerenciamento de incidentes que incluem análise pós-incidente.
- Você tem planos de observabilidade para coletar dados sobre eventos.
- Com esses dados, você entende e coleta métricas que apoiam seu processo de análise pós-incidente.
- Você aprende com os incidentes para melhorar os resultados futuros.

Práticas comuns que devem ser evitadas:

- Você administra um servidor de aplicações. Todas as sessões ativas são encerradas aproximadamente a cada 23 horas e 55 minutos. Você tentou identificar o que está errado no servidor de aplicações. Você suspeita que possa ser um problema de rede, mas não consegue obter colaboração da equipe da rede, pois ela está muito ocupada para ajudar. Você não tem um processo predefinido a seguir para obter suporte e coletar as informações necessárias para determinar o que está acontecendo.
- Houve de dados em sua workload. Esta é a primeira vez que isso acontece e a causa não é óbvia. Você decide que não é importante porque pode recriar os dados. A perda de dados começa a ocorrer com maior frequência, afetando seus clientes. Isso também cria uma sobrecarga operacional adicional à medida que você restaura os dados ausentes.

Benefícios de implementar esta prática recomendada:

- Você tem um processo predefinido para determinar componentes, condições, ações e eventos que contribuíram para um incidente, ajudando a identificar oportunidades de melhoria.
- Você usa dados da análise pós-incidente para fazer melhorias.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Use um processo para determinar fatores contribuintes. Revise todos os incidentes de impacto do cliente. Tenha um processo para identificar e documentar as causas de um incidente para que você possa desenvolver atenuações para limitar ou impedir a recorrência e para desenvolver procedimentos para respostas rápidas e eficazes. Comunique as causas principais do incidente conforme apropriado e adapte a comunicação ao seu público-alvo. Compartilhe os aprendizados abertamente em sua organização.

Etapas de implementação

1. Colete métricas como mudança na implantação, mudança de configuração, hora de início do incidente, hora do alarme, hora do engajamento, hora de início da mitigação e hora de resolução do incidente.
2. Descreva os principais pontos do cronograma para entender os eventos do incidente.
3. Faça as seguintes perguntas:

a. Você pode melhorar o tempo de detecção?
b. Há atualizações nas métricas e alarmes que detectariam o incidente mais cedo?
c. Você pode melhorar o tempo até o diagnóstico?
d. Há atualizações em seus planos de resposta ou planos de escalação que envolveriam os respondentes corretos mais cedo?
e. Você pode melhorar o tempo de mitigação?
f. Existe alguma etapa do runbook ou playbook que você pode adicionar ou melhorar?
g. Você pode evitar que futuros incidentes ocorram?

4. Crie listas de verificação e ações. Acompanhe e realize todas as ações.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS11-BP01 Adotar um processo para melhoria contínua
- OPS 4 Como implementar a observabilidade

Documentos relacionados:

- Como realizar uma análise pós-incidente no Incident Manager
- Revisão da prontidão operacional

OPS11-BP03 Implementar loops de feedback

Os loops de feedback fornecem insights que levam a ações concretas e orientam a tomada de decisões. Crie loops de feedback em seus procedimentos e workloads. Isso ajuda a identificar problemas e áreas que precisam de melhorias. Eles também validam os investimentos feitos em melhorias. Esses loops de feedback são a base para o aprimoramento contínuo da sua workload.

Os ciclos de feedback se dividem em duas categorias: feedback imediato e análise retrospectiva. O feedback imediato é coletado por meio da avaliação da performance e dos resultados das atividades de operações. Esse feedback é proveniente de membros da equipe, de clientes ou do resultado automático da atividade. O feedback imediato é recebido de elementos como testes A/B e do envio de novos recursos e é essencial para antecipar-se à falha.

A análise retrospectiva é realizada regularmente para obter feedback da avaliação de resultados e métricas operacionais ao longo do tempo. Essa retrospectiva ocorre ao final de um sprint, com certa frequência ou após grandes lançamentos ou eventos. Esse tipo de loop de feedback valida investimentos em operações ou na workload. Ele ajuda a medir o sucesso e valida sua estratégia.

Resultado desejado: o feedback imediato e a análise retrospectiva são usados para promover melhorias. Há um mecanismo para obter o feedback de usuários e membros da equipe. A análise retrospectiva é usada para identificar tendências que promovem melhorias.

Práticas comuns que devem ser evitadas:

- Você lança um novo recurso, mas não há uma maneira de receber feedback de clientes sobre ele.
- Depois de investir em melhorias de operações, você não realiza uma retrospectiva para validá-las.
- Você coleta feedback dos clientes, mas não os avalia regularmente.
- Os loops de feedback levam a itens de ação propostos, mas não estão incluídos no processo de desenvolvimento de software.
- Os clientes não recebem feedback sobre as melhorias que propuseram.

Benefícios de implementar esta prática recomendada:

- É possível trabalhar partindo do feedback do cliente para criar novos recursos.
- A cultura da sua organização pode reagir às mudanças mais rapidamente.
- As tendências são usadas para identificar oportunidades de melhoria.
- As retrospectivas validam os investimentos feitos na workload e nas operações.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

A implementação dessa prática recomendada significa que você usa tanto o feedback imediato como a análise de retrospectiva. Esses loops de feedback geram melhorias. Há muitos mecanismos para o feedback imediato, incluindo pesquisas, enquetes com clientes ou formulários de feedback. Sua organização também pode usar as retrospectivas para identificar oportunidades de melhoria e validar iniciativas.

Exemplo de cliente

A AnyCompany Retail criou um formulário online pelo qual os clientes podem fornecer feedback ou relatar problemas. Durante as reuniões semanais, o feedback dos usuários é avaliado pela equipe de desenvolvimento de software. O feedback é usado regularmente para conduzir a evolução da plataforma. É feita uma retrospectiva ao final de cada sprint para identificar itens que eles desejam melhorar.

Etapas de implementação

1. Feedback imediato
   - Você precisa de um mecanismo para receber feedback de clientes e membros da equipe. Suas atividades de operações também podem ser configuradas para oferecer feedback automático.
   - Sua organização precisa de um processo para avaliar esse feedback, determinar o que precisa ser melhorado e programar a melhoria.
   - O feedback deve ser adicionado ao seu processo de desenvolvimento de software.
   - À medida que você faz melhorias, faça o rastreamento com quem enviou o feedback.
     - É possível usar o OpsCenter do AWS Systems Manager para criar e monitorar essas melhorias como OpSitems.
2. Análise retrospectiva
   - Faça retrospectivas ao final de um ciclo de desenvolvimento, com certa frequência ou após um grande lançamento.
   - Faça uma reunião de retrospectiva com as partes interessadas envolvidas na workload.
   - Crie três colunas em um quadro branco ou uma planilha: "Parar", "Iniciar" e "Manter".
     - Parar aplica-se ao que você deseja que a equipe pare de fazer.
     - Iniciar é para ideias em que você deseja começar a trabalhar.
     - Manter refere-se aos itens que você deseja continuar fazendo.
   - Caminhe pela sala e colete o feedback das partes interessadas.
   - Priorize o feedback. Atribua ações e partes interessadas aos itens "Iniciar" e "Manter".
   - Adicione as ações ao processo de desenvolvimento de software e comunique as atualizações de status às partes interessadas à medida que as melhorias são implementadas.

Nível de esforço do plano de implementação: Médio. Para implementar essa prática recomendada, você precisa de uma maneira para receber feedback imediato e analisá-lo. Além disso, é necessário estabelecer um processo de análise retrospectiva.

Recursos

Práticas recomendadas relacionadas:

- OPS01-BP01 Avaliar as necessidades dos clientes: loops de feedback são um mecanismo para coletar as necessidades de clientes externos.
- OPS01-BP02 Avaliar as necessidades dos clientes internos: as partes interessadas internas podem usar loops de feedback para comunicar necessidades e requisitos.
- OPS11-BP02 Executar análise pós-incidente: a análise pós-incidente é uma forma importante de análise retrospectiva conduzida após os incidentes.
- OPS11-BP07 Fazer revisões das métricas de operações: as avaliações das métricas de operações identificam tendências e áreas para melhorias.

Documentos relacionados:

- Sete obstáculos que devem ser evitados ao criar um CCoE
- Playbook da equipe Atlassian: retrospectivas
- Definições de e-mail: loops de feedback
- Como estabelecer loops de feedback com base na avaliação do AWS Well-Architected Framework
- Metodologia IBM Garage: fazer uma retrospectiva
- Investopedia: o ciclo de PDCS
- Como maximizar a eficácia do desenvolvedor, por Tim Cochran
- Whitepaper Revisões de prontidão operacional (ORR): iteração
- ITIL CSI: melhoria contínua nos serviços
- Quando a Toyota conheceu o comércio eletrônico: confiança na Amazon

Vídeos relacionados:

- Como criar loops de feedback de clientes eficazes

Exemplos relacionados:

- Astuto: ferramenta de código aberto para feedback de clientes
- Soluções da AWS: QnABot na AWS

- Fider: uma plataforma para organizar feedback de clientes

Serviços relacionados:

- AWS Systems Manager OpsCenter

OPS11-BP04 Gerenciar o conhecimento

O gerenciamento de conhecimento ajuda os membros da equipe a encontrar as informações necessárias para realizar suas tarefas. Nas organizações de aprendizagem, as informações são compartilhadas livremente, o que promove a capacitação das pessoas. As informações podem ser descobertas ou pesquisadas. As informações são precisas e atualizadas. Mecanismos existem para criar informações, atualizar informações existentes e arquivar informações desatualizadas. O exemplo mais comum de uma plataforma de gerenciamento de conhecimento é um sistema de gerenciamento de conteúdo como uma wiki.

Resultado desejado:

- Os membros da equipe têm acesso a informações precisas e atualizadas.
- As informações podem ser pesquisadas.
- Existem mecanismos para adicionar, atualizar e arquivar informações.

Práticas comuns que devem ser evitadas:

- Não há um armazenamento de conhecimento centralizado. Os membros da equipe gerenciam suas próprias notas em suas máquinas locais.
- Você tem uma wiki hospedada pela própria empresa, mas nenhum mecanismo para gerenciar informações, o que resulta em informações desatualizadas.
- Alguém identifica a ausência de informações, mas não há nenhum processo para solicitar sua adição à wiki da equipe. Essa pessoa adiciona as informações por conta própria, mas deixa de realizar uma etapa, o que resulta em uma interrupção.

Benefícios de implementar esta prática recomendada:

- Os membros da equipe são capacitados, pois as informações são compartilhadas livremente.

- Os novos membros da equipe passam pelo processo de integração mais rapidamente, pois a documentação está atualizada e pode ser pesquisada.
- As informações são precisas, levam a ações concretas e são enviadas em tempo hábil.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

O gerenciamento de conhecimento é uma faceta importante das organizações de aprendizagem. Para começar, é necessário ter um repositório central para armazenar seu conhecimento (como um exemplo comum, uma wiki hospedada pela própria empresa). É necessário desenvolver processos para adicionar, atualizar e arquivar conhecimento. Desenvolva padrões para o que deve ser documentado e permita que todos contribuam.

Exemplo de cliente

A AnyCompany Retail hospeda uma wiki interna em que todo o conhecimento é armazenado. Os membros da equipe são incentivados a adicionar informações na base de conhecimento à medida que realizam suas tarefas diárias. Trimestralmente, uma equipe multifuncional avalia quais páginas estão mais desatualizadas e determina se elas devem ser arquivadas ou atualizadas.

Etapas de implementação

1. Comece identificando o sistema de gerenciamento de conteúdo em que o conhecimento será armazenado. Obtenha o consentimento das partes interessadas em sua organização.
    a. Se você não tiver um sistema de gerenciamento de conteúdo, considere desenvolver uma wiki hospedada pela própria empresa ou usar um repositório de controle de versão como ponto de partida.
2. Desenvolva runbooks para adicionar, atualizar e arquivar informações. Instrua a equipe sobre esses processos.
3. Identifique quais conhecimentos devem ser armazenados no sistema de gerenciamento de conteúdo. Comece com as atividades diárias (runbooks e playbooks) realizadas pelos membros da equipe. Trabalhe com as partes interessadas para priorizar qual conhecimento deve ser adicionado.
4. Periodicamente, trabalhe com as partes interessadas para identificar informações desatualizadas e arquive-as ou atualize-as.

Nível de esforço do plano de implementação: Médio. Se você não tiver um sistema de gerenciamento de conteúdo, defina uma wiki hospedada pela própria empresa ou um repositório de documentos com controle de versão.

Recursos

Práticas recomendadas relacionadas:

- OPS11-BP08 Documentar e compartilhar as lições aprendidas: o gerenciamento de conhecimento facilita o compartilhamento de informações sobre as lições aprendidas.

Documentos relacionados:

- Atlassian: Gerenciamento do conhecimento

Exemplos relacionados:

- DokuWiki
- Gollum
- MediaWiki
- Wiki.js

OPS11-BP05 Definir fatores de melhoria

Identifique os fatores de melhoria para ajudar a avaliar e priorizar oportunidades com base em dados e ciclos de feedback. Explore oportunidades de melhoria nos sistemas e nos processos e automatize sempre que apropriado.

Resultado desejado:

- Você rastreia dados de todo o ambiente.
- Você correlaciona eventos e atividades aos resultados comerciais.
- Você pode comparar e contrastar entre ambientes e sistemas.
- Você mantém um histórico detalhado de atividades das implantações e dos resultados.
- Você coleta dados para apoiar o procedimento de segurança.

Práticas comuns que devem ser evitadas:

- Coletar dados de todo o ambiente, mas não correlacionar eventos e atividades.
- Coletar dados detalhados de toda a propriedade, gerando atividade e custos elevados do Amazon CloudWatch e do AWS CloudTrail. No entanto, você não usa esses dados de forma significativa.
- Não levar em conta os resultados comerciais ao definir os fatores de melhoria.
- Não medir os efeitos dos novos recursos.

Benefícios de implementar esta prática recomendada:

- O impacto das motivações baseadas em eventos ou investimentos emocionais ao determinar os critérios de melhoria é minimizado.
- Você reage a eventos de negócios, não apenas a eventos técnicos.
- Você mede o ambiente para identificar áreas de melhoria.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

- Compreenda as motivações para melhoria: só faça alterações em um sistema quando o resultado desejado for compatível.
  - Capacidades desejadas: avalie as capacidades e os recursos desejados ao avaliar oportunidades de melhoria.
    - Novidades da AWS
  - Problemas inaceitáveis: avalie problemas, erros e vulnerabilidades inaceitáveis ao avaliar oportunidades de melhoria. Acompanhe as opções dimensionamento correto e busque oportunidades de otimização.
    - Boletins de segurança mais recentes da AWS
    - AWS Trusted Advisor
    - Cloud Intelligence Dashboards
  - Requisitos de conformidade: avalie as atualizações e as alterações necessárias para manter a conformidade com a regulamentação e com a política, ou para permanecer sob o suporte de terceiros ao analisar as oportunidades de melhoria.
    - Conformidade da AWS
    - Programas de conformidade da AWS
    - Últimas notícias sobre conformidade com a AWS

Recursos

Práticas recomendadas relacionadas:

- OPS01 Prioridades da organização
- OPS02 Relacionamentos e propriedades
- OPS04-BP01 Identificar indicadores-chave de performance
- OPS08 Utilizar a observabilidade da workload
- OPS09 Como compreender a integridade operacional
- OPS11-BP03 Implementar loops de feedback

Documentos relacionados:

- Amazon Athena
- Amazon QuickSight
- Conformidade da AWS
- Últimas notícias sobre conformidade com a AWS
- Programas de conformidade da AWS
- AWS Glue
- Boletins de segurança mais recentes da AWS
- AWS Trusted Advisor
- Exportar seus dados de log para o Amazon S3
- Novidades da AWS
- Os imperativos da inovação centrada no cliente
- Transformação digital: modismo ou necessidade estratégica?

Vídeos relacionados

- AWS re:Invent 2023: Melhorar a eficiência operacional e a resiliência com o AWS Support (SUP310)

OPS11-BP06 Validar insights

Revise os resultados e as respostas da análise com equipes multifuncionais e proprietários de negócios. Use essas revisões para estabelecer um entendimento comum, identificar impactos adicionais e determinar cursos de ação. Ajuste as respostas conforme apropriado.

Resultado desejado:

- Você revisa os insights regularmente com proprietários de empresas. Os empresários fornecem contexto adicional aos insights recém-adquiridos.
- Você analisa os insights e solicita feedback de pares técnicos e compartilha seu aprendizado entre as equipes.
- Você publica dados e insights para que outras equipes técnicas e comerciais analisem. Você pensa no aprendizado de novas práticas de outros departamentos.
- Você resume e analisa novos insights com os líderes seniores. Os líderes seniores usam novos insights para definir a estratégia.

Práticas comuns que devem ser evitadas:

- Você lança um novo recurso. Esse recurso muda alguns comportamentos dos clientes. Sua observabilidade não leva em conta essas mudanças. Você não quantifica os benefícios dessas mudanças.
- Você envia uma nova atualização e deixa de atualizar sua CDN. O cache da CDN não é mais compatível com a versão mais recente. Você mede a porcentagem de solicitações com erros. Todos os seus usuários relatam erros de HTTP 400 ao se comunicarem com servidores de backend. Você investiga os erros do cliente e descobre que, por ter medido a dimensão errada, seu tempo foi desperdiçado.
- Seu contrato de nível de serviço estipula 99,9% de tempo de atividade e seu objetivo de ponto de recuperação é de quatro horas. O proprietário do serviço afirma que o sistema tem zero tempo de inatividade. Você implementa uma solução de replicação cara e complexa que desperdiça tempo e dinheiro.

Benefícios de implementar esta prática recomendada:

- Ao validar insights com proprietários de empresas e especialistas, você estabelece um entendimento comum e orienta as melhorias de maneira mais eficaz.
- Você descobre problemas ocultos e os leva em conta em decisões futuras.

- Seu foco passa dos resultados técnicos para os resultados comerciais.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

- Valide insights: envolva-se com proprietários de empresas e especialistas para garantir que haja entendimento e concordância comuns sobre o significado dos dados coletados. Identifique preocupações adicionais, possíveis impactos e determine as ações.

Recursos

Práticas recomendadas relacionadas:

- OPS01-BP06 Avaliar as compensações ao gerenciar benefícios e riscos
- OPS02-BP06 As responsabilidades entre as equipes são predefinidas ou negociadas
- OPS11-BP03 Implementar loops de feedback

Documentos relacionados:

- Como projetar um Centro de Excelência da Nuvem (CCoE)

Vídeos relacionados:

- Criar observabilidade para aumentar a resiliência

OPS11-BP07 Fazer revisões das métricas de operações

Realize regularmente revisões das métricas de operações com participantes de equipes diferentes de várias áreas do negócio. Use essas revisões para identificar oportunidades de melhorias e possíveis ações e compartilhar as lições aprendidas. Procure oportunidades para melhorar em todos os seus ambientes (por exemplo, desenvolvimento, teste e produção).

Resultado desejado:

- Você analisa frequentemente métricas que afetam os negócios.
- Você detecta e analisa anomalias por meio de suas capacidades de observabilidade.
- Você usa dados para apoiar os resultados e as metas de negócios.

Práticas comuns que devem ser evitadas:

- Sua janela de manutenção interrompe uma promoção significativa no varejo. A empresa continua sem saber que existe uma janela de manutenção padrão que poderá ser atrasada se houver outros eventos que afetam os negócios.
- Você sofreu uma paralisação prolongada porque costuma usar uma biblioteca desatualizada na organização. Desde então, você migrou para uma biblioteca compatível. As outras equipes da organização não sabem que estão em risco.
- Você não analisa regularmente o cumprimento dos SLAs do cliente. Você está tendendo a não cumprir os SLAs dos clientes. Há penalidades financeiras relacionadas ao não cumprimento de SLAs dos clientes.

Benefícios de implementar esta prática recomendada:

- Ao se reunir regularmente para analisar métricas de operações, eventos e incidentes, você mantém um entendimento comum entre as equipes.
- Sua equipe se reúne rotineiramente para analisar métricas e incidentes, o que permite tomar medidas sobre os riscos e reconhecer os SLAs dos clientes.
- Você compartilha as lições aprendidas, as quais fornecem dados para priorização e melhorias direcionadas para os resultados comerciais.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

- Realize regularmente revisões das métricas de operações com participantes de equipes diferentes de várias áreas do negócio.
- Envolva as partes interessadas, incluindo as equipes de negócios, desenvolvimento e operações, para validar suas descobertas de feedback imediato e análise retrospectiva e para compartilhar as lições aprendidas.
- Use suas ideias para identificar oportunidades de melhoria e possíveis cursos de ação.

Recursos

Práticas recomendadas relacionadas:

- OPS08-BP05 Criar painéis

- OPS09-BP03 Revisar as métricas operacionais e priorizar a melhoria
- OPS10-BP01 Usar um processo para gerenciamento de eventos, incidentes e problemas

Documentos relacionados:

- Amazon CloudWatch
- Referência de métricas e dimensões do Amazon CloudWatch
- Publicar métricas personalizadas
- Usar métricas do Amazon CloudWatch
- Painéis e visualizações com o CloudWatch

OPS11-BP08 Documentar e compartilhar as lições aprendidas

Documente e compartilhe as lições aprendidas das atividades operacionais, para que possa usá-las internamente e entre equipes. Você deve compartilhar o que suas equipes aprendem para aumentar os benefícios em toda a organização. Compartilhe informações e recursos para evitar erros previsíveis e facilitar os esforços de desenvolvimento, e concentre-se na entrega dos recursos desejados.

Use o AWS Identity and Access Management (IAM) para definir permissões que permitem acesso controlado aos recursos que você deseja compartilhar dentro e entre contas.

Resultado desejado:

- Você usa os repositórios com controle de versão para compartilhar bibliotecas de aplicações, procedimentos com script, documentações de procedimentos e outras documentações do sistema.
- Você compartilha seus padrões de infraestrutura como modelos com controle de versão do AWS CloudFormation.
- Você revisa as lições aprendidas entre as equipes.

Práticas comuns que devem ser evitadas:

- Você sofreu uma paralisação prolongada porque a organização geralmente usa bibliotecas com erros. Desde então, você migrou para uma biblioteca confiável. As outras equipes na organização não sabem que estão em risco. Ninguém documenta e compartilha a experiência com essa biblioteca e não está ciente do risco.

- Você identificou um caso de borda em um microsserviço compartilhado internamente que causa a queda das sessões. Atualizou suas chamadas para o serviço para evitar esse caso de borda. As outras equipes da organização não sabem que estão em risco.
- Você encontrou uma maneira de reduzir significativamente os requisitos de utilização da CPU para um dos microsserviços. Você não sabe se alguma outra equipe poderia aproveitar essa técnica.

Benefícios de implementar esta prática recomendada: compartilhe as lições aprendidas para apoiar a melhoria e maximizar os benefícios da experiência.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

- Documente e compartilhe as lições aprendidas: tenha procedimentos para documentar as lições aprendidas com a execução de atividades operacionais e análises retrospectivas, para que possam ser usadas por outras equipes.
- Compartilhe os aprendizados: tenha procedimentos para compartilhar lições aprendidas e produções associadas entre as equipes. Por exemplo, compartilhe procedimentos atualizados, orientações, governança e práticas recomendadas por meio de uma wiki acessível. Compartilhe scripts, códigos e bibliotecas por meio de um repositório comum.
  - Delegar acesso ao seu ambiente da AWS
  - Compartilhar um repositório do AWS CodeCommit

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP06 As responsabilidades entre as equipes são predefinidas ou negociadas
- OPS05-BP01 Usar controle de versão
- OPS05-BP06 Compartilhar padrões de design
- OPS11-BP03 Implementar loops de feedback
- OPS11-BP07 Revisar as métricas de operações

Documentos relacionados:

- Reduzir atrasos em projetos com uma solução de documentos como código

Vídeos relacionados:

- Delegar acesso ao seu ambiente da AWS
- AWS Supports You | Explorar a simulação teórica de gerenciamento de incidentes

OPS11-BP09 Alocar tempo para fazer melhorias

Dedique tempo e recursos em seus processos para possibilitar melhorias incrementais contínuas.

Resultado desejado:

- Você cria duplicações temporárias de ambientes, o que reduz o risco, o esforço e o custo de testes e experimentações.
- Esses ambientes duplicados podem ser usados para testar as conclusões de sua análise, experimentar e desenvolver e testar as melhorias planejadas.
- Você realiza game days e usa o Fault Injection Service (FIS) para fornecer os controles e as barreiras de proteção de que as equipes precisam para realizar experimentos em um ambiente semelhante ao de produção.

Práticas comuns que devem ser evitadas:

- Há um problema de performance conhecido no servidor de aplicações. Ele é adicionado ao backlog por trás de cada implementação de recurso planejada. Se a taxa de adição de recursos planejados permanecer constante, o problema de performance nunca será resolvido.
- Para apoiar a melhoria contínua, você aprova administradores e desenvolvedores usando todo o tempo extra para selecionar e implementar melhorias. As melhorias nunca são concluídas.
- A aceitação operacional está completa e você não testa as práticas operacionais novamente.

Benefícios de implementar esta prática recomendada: ao dedicar tempo e recursos em seus processos, você possibilita melhorias incrementais contínuas.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

- Aloque tempo para fazer melhorias: dedique tempo e recursos em seus processos para realizar melhorias incrementais contínuas.

- Implemente alterações para melhorar e avaliar os resultados para determinar o sucesso.
- Se os resultados não satisfizerem as metas e a melhoria ainda for uma prioridade, procure ações alternativas.
- Simule workloads de produção durante os game days e use o que aprendeu com essas simulações para melhorar.

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP08 Usar vários ambientes

Vídeos relacionados:

- AWS re:Invent 2023: Melhorar a resiliência de aplicações com o AWS Fault Injection Service

Segurança

O pilar Segurança refere-se à capacidade de proteger dados, sistemas e ativos para utilizar as tecnologias de nuvem para melhorar sua segurança. Recomendações sobre implementação estão disponíveis no whitepaper Pilar Segurança.

Áreas de práticas recomendadas

- Fundamentos de segurança
- Gerenciamento de identidade e acesso
- Detecção
- Proteção da infraestrutura
- Proteção de dados
- Resposta a incidentes
- Segurança de aplicações

Fundamentos de segurança

Pergunta

SEC 1. Como você opera seu workload com segurança?

Para operar seu workload com segurança, você deve aplicar as práticas recomendadas abrangentes em todas as áreas de segurança. Pegue os requisitos e processos que você definiu em excelência operacional em um nível organizacional e de workload e aplique-os a todas as áreas. Manter-se atualizado com as recomendações do setor e da AWS e a inteligência contra ameaças ajuda você a desenvolver seu modelo de ameaças e seus objetivos de controle. A automação de processos, testes e validação de segurança permite escalar suas operações de segurança.

Práticas recomendadas

- SEC01-BP01 Separar as workloads usando contas
- SEC01-BP02 Proteger as propriedades e o usuário-raiz das contas
- SEC01-BP03 Identificar e validar objetivos de controle
- SEC01-BP04 Manter-se em dia com ameaças e recomendações de segurança
- SEC01-BP05 Reduzir o escopo do gerenciamento de segurança
- SEC01-BP06 Automatizar a implantação de controles de segurança padrão
- SEC01-BP07 Identificar ameaças e priorizar mitigações usando um modelo de ameaça
- SEC01-BP08 Avaliar e implementar regularmente novos serviços e recursos de segurança

SEC01-BP01 Separar as workloads usando contas

Estabeleça barreiras de proteção e isolamento entre workloads e ambientes (como de produção, desenvolvimento e teste) por meio de uma estratégia de várias contas. A separação em nível de conta é altamente recomendável, pois ela oferece um limite de isolamento robusto para segurança, faturamento e acesso.

Resultado desejado: uma estrutura de contas que isola operações em nuvem, workloads não relacionadas e ambientes em contas separadas, aumentando a segurança em toda a infraestrutura de nuvem.

Práticas comuns que devem ser evitadas:

- Colocação de várias workloads não relacionadas com diferentes níveis de confidencialidade na mesma conta.

- Estrutura de unidade organizacional (UO) definida de forma inadequada.

Benefícios de implementar esta prática recomendada:

- Redução do escopo de impacto se uma workload for acessada acidentalmente.
- Governança central de acesso a serviços, recursos e regiões da AWS.
- Manutenção da segurança da infraestrutura de nuvem com políticas e administração centralizada de serviços de segurança.
- Criação de contas automatizada e processo de manutenção.
- Auditoria centralizada da infraestrutura de conformidade e requisitos regulatórios.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

As Contas da AWS oferecem um limite de isolamento de segurança entre workloads ou recursos que operam em diferentes níveis de confidencialidade. Para utilizar esse limite de isolamento, a AWS oferece ferramentas para gerenciar em grande escala suas workloads de nuvem por meio de uma estratégia de várias contas. Para obter orientação sobre os conceitos, padrões e implementação de uma estratégia de várias contas na AWS, consulte Organizar seu ambiente da AWS usando várias contas.

Quando você tem várias Contas da AWS no gerenciamento central, elas devem ser organizadas em uma hierarquia definida por camadas de unidades organizacionais (UOs). Desse modo, os controles de segurança podem ser organizados e aplicados às UOs e às contas-membro, estabelecendo controles preventivos consistentes nas contas-membro da organização. Os controles de segurança são herdados, permitindo que você filtre as permissões disponíveis para as contas-membro localizadas em níveis inferiores de uma hierarquia de UOs. Um bom design aproveita essa herança para reduzir o número e a complexidade das políticas de segurança necessárias para obter os controles de segurança desejados para cada conta-membro.

AWS Organizations e AWS Control Tower são dois serviços que podem ser usados para implementar e gerenciar essa estrutura de várias contas em seu ambiente da AWS. O AWS Organizations permite que você organize contas em uma hierarquia definida por uma ou mais camadas de UOs, onde cada UO contém várias contas-membro. As políticas de controle de serviços (SCPs) permitem que o administrador da organização estabeleça controles preventivos granulares nas contas-membro, e o AWS Config pode ser usado para estabelecer controles proativos e de

detetive nas contas-membro. Muitos serviços da AWS se integram ao AWS Organizations para fornecer controles administrativos delegados e realizar tarefas específicas do serviço em todas as contas-membro da organização.

Em cima do AWS Organizations, o AWS Control Tower fornece uma configuração de práticas recomendadas com um clique para um ambiente da AWS de várias contas com uma zona de pouso. A zona de pouso é o ponto de entrada para o ambiente de várias contas estabelecido pelo Control Tower. O Control Tower oferece vários benefícios em relação ao AWS Organizations. Três benefícios que oferecem governança aprimorada de contas são:

- Controles de segurança obrigatórios e integrados que são aplicados automaticamente às contas admitidas na organização.
- Controles opcionais que podem ser ativados ou desativados em determinado conjunto de UOs.
- O AWS Control Tower Account Factory fornece implantação automatizada de contas contendo linhas de base e opções de configuração pré-aprovadas em sua organização.

Etapas de implementação

1. Projete uma estrutura de unidade organizacional: uma estrutura de unidade organizacional projetada adequadamente reduz a carga de gerenciamento necessária para criar e manter políticas de controle de serviços e outros controles de segurança. A estrutura da unidade organizacional deve estar alinhada às necessidades da empresa, à sensibilidade dos dados e à estrutura da workload.
2. Crie uma zona de pouso para seu ambiente de várias contas: uma zona de pouso fornece uma base consistente de segurança e infraestrutura a partir da qual sua organização pode desenvolver, lançar e implantar workloads rapidamente. Você pode usar uma zona de pouso personalizada ou o AWS Control Tower para orquestrar seu ambiente.
3. Estabeleça barreiras de proteção: implemente proteções de segurança consistentes para seu ambiente em sua zona de pouso. O AWS Control Tower fornece uma lista de controles obrigatórios e opcionais que podem ser implantados. Os controles obrigatórios são implantados automaticamente na implementação do Control Tower. Leia a lista de controles opcionais e altamente recomendados e implemente controles adequados às suas necessidades.
4. Restrinja o acesso a regiões recém-adicionadas: para novas Regiões da AWS, recursos do IAM como usuários e perfis são propagados somente para as regiões que você especificar. Essa ação pode ser executada por meio do console ao usar o Control Tower ou ajustando as políticas de permissão do IAM no AWS Organizations.

5. Considere o AWS CloudFormation StackSets: o StackSets ajuda a implantar recursos, incluindo políticas, perfis e grupos do IAM, em diferentes contas e regiões da Contas da AWS por meio de um modelo aprovado.

Recursos

Práticas recomendadas relacionadas:

- SEC02-BP04 Confiar em um provedor de identidades centralizado

Documentos relacionados:

- AWS Control Tower
- Diretrizes de auditoria de segurança da AWS
- Práticas recomendadas do IAM
- Usar o CloudFormation StackSets para provisionar recursos em várias regiões e Contas da AWS
- Perguntas frequentes sobre o Organizations
- Terminologia e conceitos do AWS Organizations
- Práticas recomendadas para Políticas de controle de serviços do AWS Organizations em um ambiente com várias contas
- Guia de referência de gerenciamento de contas da AWS
- Organizar seu ambiente da AWS usando várias contas

Vídeos relacionados:

- Permitir a adoção da AWS em escala por meio de automação e governança
- Práticas recomendadas de segurança à maneira do Well-Architected
- Criar e gerenciar várias contas usando o AWS Control Tower
- Habilitar o Control Tower para organizações existentes

Workshops relacionados:

- Dia de imersão no Control Tower

SEC01-BP02 Proteger as propriedades e o usuário-raiz das contas

O usuário-raiz é o mais privilegiado de uma Conta da AWS, com acesso administrativo integral a todos os recursos da conta, e em alguns casos não pode ser restringido por políticas de segurança. Desabilitar o acesso programático ao usuário-raiz, estabelecer controles apropriados para ele e evitar o uso rotineiro desse usuário ajuda a reduzir o risco de exposição acidental das credenciais raiz e o subsequente comprometimento do ambiente de nuvem.

Resultado desejado: proteger o usuário-raiz ajuda a reduzir a chance de que danos acidentais ou intencionais ocorram devido ao uso indevido das credenciais do usuário-raiz. Estabelecer controles de detecção também pode alertar o pessoal apropriado ações são postas em prática com o usuário-raiz.

Práticas comuns que devem ser evitadas:

- Utilizar o usuário-raiz para outras tarefas que não sejam aquelas que exigem credenciais do usuário-raiz.
- Negligenciar os testes dos planos de contingência regularmente a fim de verificar a funcionalidade da infraestrutura, dos processos e dos funcionários essenciais durante uma emergência.
- Considerar apenas o fluxo típico de login de contas e não considerar nem testar métodos de recuperação de contas alternativos.
- Não lidar com DNS, servidores de e-mail e operadoras de telefonia como parte do perímetro de segurança essencial, pois eles são usados no fluxo de recuperação de contas.

Benefícios de implementar esta prática recomendada: proteger o acesso ao usuário-raiz aumenta a confiança de que as ações em sua conta são controladas e auditadas.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

A AWS oferece muitas ferramentas para ajudar a proteger sua conta. No entanto, como algumas dessas medidas não estão habilitadas por padrão, é necessário implementá-las diretamente. Leve em consideração essas recomendações como etapas fundamentais para proteger sua Conta da AWS. Ao implementar essas etapas, é importante criar um processo para avaliar e monitorar os controles de segurança de forma contínua.

Ao criar uma Conta da AWS pela primeira vez, você começa com uma identidade que tem acesso completo a todos os recursos e serviços da AWS na conta. Essa identidade é chamada de usuário-

raiz da Conta da AWS. Você pode fazer login como usuário-raiz usando o endereço de e-mail e a senha que usou para criar a conta. Devido ao acesso elevado concedido ao usuário-raiz da AWS, limite o uso do usuário-raiz da AWS à realização de tarefas que o necessitem especificamente dele. As credenciais de login do usuário-raiz devem ser bem protegidas, e a autenticação multifator (MFA) sempre deve ser usada para o usuário-raiz da Conta da AWS.

Além do fluxo de autenticação normal para fazer login com seu usuário-raiz usando um nome de usuário, senha e o dispositivo de autenticação multifator (MFA), há fluxos de recuperação de contas para fazer login com seu usuário-raiz da Conta da AWS com o endereço de e-mail e o número de telefone associados à sua conta. Dessa forma, é igualmente importante proteger a conta de e-mail do usuário-raiz para a qual o e-mail de recuperação é enviado e o número de telefone associado à conta. Além disso, considere possíveis dependências circulares em que o endereço de e-mail associado ao usuário-raiz é hospedado em servidores de e-mail ou recursos de serviço de nome de domínio (DNS) da mesma Conta da AWS.

Quando o AWS Organizations é usado, há várias Contas da AWS, e cada uma tem um usuário-raiz. Uma conta é designada como a conta de gerenciamento e várias camadas de contas-membro podem ser adicionadas à conta de gerenciamento. Priorize a proteção do usuário-raiz de sua conta de gerenciamento e, depois, os usuários-raiz das contas-membro. A estratégia para proteger o usuário-raiz de sua conta de gerenciamento pode diferir da utilizada nos usuários raiz de suas contas-membro, e é possível implementar controles de segurança preventivos nos usuários-raiz dessas contas.

Etapas de implementação

As etapas de implementação a seguir são recomendadas para estabelecer controles para o usuário-raiz. Onde aplicável, as recomendações são cruzadas com o CIS AWS Foundations Benchmark versão 1.4.0. Além dessas etapas, consulte as diretrizes de práticas recomendadas do AWS para proteger sua Conta da AWS e seus recursos.

Controles preventivos

1. Configure informações de contato precisas para a conta.
    a. Essas informações são usadas para o fluxo de recuperação de senha perdida, o fluxo de recuperação de conta de dispositivo MFA perdida e para comunicações com sua equipe sobre segurança crítica.
    b. Utilize um endereço de e-mail hospedado por seu domínio corporativo, preferencialmente uma lista de distribuição, como o endereço de e-mail do usuário-raiz. O uso de uma lista de

distribuição em vez da conta de e-mail de um indivíduo oferece redundância e continuidade adicionais para o acesso à conta raiz por longos períodos.

c. O número de telefone listado nas informações de contato deve ser um telefone dedicado e seguro para esse fim. O número de telefone não deve ser listado nem compartilhado com ninguém.

2. Não crie chaves de acesso para o usuário-raiz. Se houver chaves de acesso, remova-as (CIS 1.4).
   a. Elimine todas as credenciais programáticas de longa duração (chaves de acesso e secretas) para o usuário-raiz.
   b. Se as chaves de acesso do usuário-raiz já existirem, você deverá fazer a transição dos processos usando essas chaves para usar chaves de acesso temporárias de um pefil do AWS Identity and Access Management (IAM) e, em seguida, excluir as chaves de acesso do usuário-raiz.
3. Determine se você precisa armazenar credenciais para o usuário-raiz.
   a. Ao usar o AWS Organizations para criar contas-membro, a senha inicial do usuário-raiz em novas contas-membro é definida como um valor aleatório que não é exposto a você. Considere usar o fluxo de redefinição de senha da sua conta de gerenciamento do AWS Organizations para obter acesso à conta-membro, se necessário.
   b. Para Contas da AWS autônomas ou a conta de gerenciamento do AWS Organizations, considere criar e armazenar de forma segura as credenciais do usuário-raiz. Use MFA para o usuário-raiz
4. Ative os controles preventivos para os usuários-raiz das contas-membro em ambientes de várias contas da AWS.
   a. Considere usar a barreira de proteção Não permitir a criação de chaves de acesso raiz para o usuário-raiz para contas-membro.
   b. Considere usar a barreira de proteção Não permitir ações como o usuário-raiz para contas-membro.
5. Se você precisar de credenciais para o usuário-raiz:
   a. Use uma senha complexa.
   b. Ative a autenticação multifator (MFA) para o usuário-raiz, especialmente para contas (pagantes) de gerenciamento do AWS Organizations (CIS 1.5).
   c. Considere o uso de dispositivos de MFA de hardware para ter resiliência e segurança, pois os dispositivos de uso único reduzem as chances de os dispositivos que contêm seus códigos

de MFA serem reutilizados para outros fins. Garanta que os dispositivos de MFA de hardware alimentados por bateria sejam substituídos regularmente. (CIS 1.6)

- Para configurar a MFA para o usuário-raiz, siga as instruções para criar uma MFA virtual ou um dispositivo com MFA de hardware.

d. Considere inscrever vários dispositivos de MFA para backup. Até 8 dispositivos de MFA são permitidos por conta.

- Observe que a inscrição de mais de um dispositivo de MFA para o usuário-raiz desativa automaticamente o fluxo para recuperar sua conta se o dispositivo de MFA for perdido.

e. Armazene a senha com segurança e considere as dependências circulares se for armazenar a senha eletronicamente. Não armazene a senha de uma forma que exija o acesso à mesma Conta da AWS para obtê-la.

6. Opcional: considere estabelecer um cronograma de rotação de senha periódica para o usuário-raiz.

- As práticas recomendadas de gerenciamento de credenciais dependem de seus requisitos regulatórios e de política. Os usuários-raiz protegidos por MFA não dependem da senha como um único fator de autenticação.
- Alterar a senha do usuário-raiz periodicamente reduz o risco de que uma senha exposta inadvertidamente possa ser usada indevidamente.

Controles de detecção

- Crie alarmes para detectar o uso das credenciais de usuário-raiz (CIS 1.7). O Amazon GuardDuty pode monitorar e alertar sobre o uso da credencial da API do usuário-raiz por meio da descoberta RootCredentialUsage.
- Avalie e implemente os controles de detetive incluídos no pacote de conformidade do pilar Segurança do AWS Well-Architected para AWS Config, ou se estiver usando o AWS Control Tower, os controles altamente recomendados disponíveis no Control Tower.

Orientação operacional

- Determine quem na organização deve ter acesso às credenciais do usuário-raiz.
  - Use uma regra de duas pessoas de forma que um indivíduo tenha acesso a todas as credenciais necessárias e MFA para obter acesso de usuário-raiz.

  - Verifique se é a organização, e não um único indivíduo, que mantém controle sobre o número de telefone e alias de e-mail associados à conta (que são utilizados para redefinição de senha e fluxo de redefinição de MFA).
- Utilize o usuário-raiz apenas como uma exceção (CIS 1.7).
  - O usuário-raiz da AWS não deve ser usado para tarefas diárias, mesmo que sejam tarefas administrativas. Faça login somente como usuário-raiz para realizar tarefas da AWS que exijam o usuário-raiz. Todas as outras ações devem ser realizadas por outros usuários com perfis apropriados.
- Confira periodicamente se o acesso ao usuário-raiz está funcionando de forma que os procedimentos sejam testados antes de uma situação de emergência que exija o uso das credenciais do usuário-raiz.
- Verifique periodicamente se o endereço de e-mail associado à conta e os listados em Contatos alternativos funcionam. Monitore as caixas de entrada de e-mail em busca de notificações de segurança que poderia receber de <abuse@amazon.com>. Além disso, garanta que todos os números de telefone associados à conta estejam funcionando.
- Prepare um procedimento de resposta a incidentes para responder ao mau uso da conta de usuário-raiz. Consulte o Guia de resposta a incidentes de segurança da AWS e as práticas recomendadas na seção Resposta a Incidentes do whitepaper Pilar Segurança para obter mais informações sobre como criar uma estratégia de resposta a incidentes para sua Conta da AWS.

Recursos

Práticas recomendadas relacionadas:

- SEC01-BP01 Separar as workloads usando contas
- SEC02-BP01 Usar mecanismos de início de sessão fortes
- SEC03-BP02 Conceder acesso de privilégio mínimo
- SEC03-BP03 Estabelecer processo de acesso de emergência
- SEC10-BP05 Provisionar acesso previamente

Documentos relacionados:

- AWS Control Tower
- Diretrizes de auditoria de segurança da AWS

- Práticas recomendadas do IAM
- Amazon GuardDuty — alerta de uso da credencial de usuário-raiz
- Orientação passo a passo sobre o monitoramento do uso da credencial de usuário-raiz via CloudTrail
- Tokens MFA aprovados para uso com o AWS
- Implementar o acesso de emergência na AWS
- Os 10 principais itens de segurança para melhorar em sua Conta da AWS
- O que devo fazer se perceber uma atividade não autorizada em minha Conta da AWS?

Vídeos relacionados:

- Permitir a adoção da AWS em escala por meio de automação e governança
- Práticas recomendadas de segurança à maneira do Well-Architected
- Limitar o uso de credenciais de usuário-raiz da AWS: AWS re:inforce 2022: Práticas recomendadas de segurança com o AWS IAM

Exemplos e laboratórios relacionados:

- Laboratório: configuração da Conta da AWS e do usuário-raiz

SEC01-BP03 Identificar e validar objetivos de controle

Com base em seus requisitos de conformidade e riscos identificados no modelo de ameaça, derive e valide os objetivos de controle e os controles que você precisa aplicar à workload. A validação contínua de objetivos de controle e controles ajuda a medir a eficácia da mitigação de riscos.

Resultado desejado: os objetivos de controle de segurança da sua empresa estão bem definidos e alinhados aos seus requisitos de conformidade. Os controles são implementados e aplicados por meio de automação e políticas, bem como continuamente avaliados quanto à respectiva eficácia para alcançar seus objetivos. As evidências de eficácia em determinado momento e durante um período de tempo podem ser facilmente relatadas aos auditores.

Práticas comuns que devem ser evitadas:

- Os requisitos regulatórios, as expectativas de mercado e os padrões do setor de garantia de segurança não são claros para sua empresa.

- Seus frameworks de segurança cibernética e seus objetivos de controle estão desalinhados em relação aos requisitos de sua empresa.
- A implementação de controles não se alinha de maneira consistente e mensurável aos seus objetivos de controle.
- Você não usa a automação para relatar a eficácia de seus controles.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Há muitos frameworks comuns de segurança cibernética que podem servir de base para seus objetivos de controle de segurança. Considere os requisitos regulatórios, as expectativas de mercado e os padrões do setor aplicáveis à sua empresa a fim de determinar quais frameworks atendem melhor às suas necessidades. Os exemplos incluem AICPA SOC 2, HITRUST, PCI-DSS, ISO 27.001 e NIST SP 800-53.

Com relação aos objetivos de controle identificados, entenda como os serviços da AWS que você consome ajudam a atingi-los. Use o AWS Artifact para encontrar documentação e relatórios alinhados às suas estruturas de destino que descrevam o escopo de responsabilidade coberto pela AWS e orientações para o escopo restante que é de sua responsabilidade. Para obter mais orientações específicas do serviço, conforme elas se alinham a várias declarações de controle da estrutura, consulte os Guias de conformidade do cliente da AWS.

Ao definir os controles que viabilizam seus objetivos, codifique a imposição usando controles preventivos e automatize a mitigação usando controles de detecção. Ajude a evitar configurações e ações de recursos fora de conformidade em todo o seu AWS Organizations por meio do uso políticas de controle de serviços (SCP). Implemente regras no AWS Config para monitorar e relatar recursos fora de conformidade e, em seguida, mude as regras para um modelo de fiscalização quando estiver confiante em seu comportamento. Para implantar conjuntos de regras predefinidas e gerenciadas que se alinham às suas estruturas de segurança cibernética, avalie o uso de padrões do AWS Security Hub como sua primeira opção. O padrão de Práticas de Segurança Básica da AWS (FSBP) e o CIS AWS Foundations Benchmark são bons pontos de partida e têm controles que se alinham a muitos objetivos que são compartilhados em vários frameworks padrão. Onde o Security Hub não tem intrinsecamente as detecções de controle desejadas, ele pode ser complementado usando pacotes de conformidade do AWS Config.

Use pacotes de parceiros da APN recomendados pela equipe AWS Global Security and Compliance Acceleration (GSCA) para obter assistência de consultores de segurança, agências de consultoria,

sistemas de coleta de evidências e relatórios, auditores e outros serviços complementares quando necessário.

Etapas de implementação

1. Avalie frameworks comuns de segurança cibernética e alinhe seus objetivos de controle aos escolhidos.
2. Obtenha documentação relevante sobre orientações e responsabilidades pelo uso de seu framework usando o AWS Artifact. Entenda quais partes da conformidade enquadram-se no modelo de responsabilidade compartilhada da AWS e quais partes são de sua responsabilidade.
3. Use SCPs, políticas de recursos, políticas de confiança de perfil e outras barreiras de proteção para evitar configurações e ações de recursos fora de conformidade.
4. Avalie a implantação de padrões do Security Hub e de pacotes de conformidade do AWS Config que se alinhem aos seus objetivos de controle.

Recursos

Práticas recomendadas relacionadas:

- SEC03-BP01 Definir requisitos de acesso
- SEC04-BP01 Configurar o registro em log de serviços e aplicações
- SEC07-BP01 Compreender seu esquema de classificação de dados
- OPS01-BP03 Avaliar os requisitos de governança
- OPS01-BP04 Avaliar os requisitos de conformidade
- PERF01-BP05 Usar políticas e arquiteturas de referência
- COST02-BP01 Desenvolver políticas com base nos requisitos da sua organização

Documentos relacionados:

- Guias de conformidade do cliente da AWS

Ferramentas relacionadas:

- AWS Artifact

SEC01-BP04 Manter-se em dia com ameaças e recomendações de segurança

Mantenha-se em dia com as ameaças e mitigações mais recentes monitorando as publicações de inteligência contra ameaças do setor e os feeds de dados para atualizações. Avalie as ofertas de serviços gerenciados que são atualizadas automaticamente com base nos dados de ameaças mais recentes.

Resultado desejado: você se mantém informado à medida que as publicações do setor são atualizadas com as ameaças e recomendações mais recentes. Você usa a automação para detectar possíveis vulnerabilidades e exposições à medida que identifica novas ameaças. Você toma medidas de mitigação contra essas ameaças. Você adota serviços da AWS que são atualizados automaticamente com a inteligência de ameaças mais recente.

Práticas comuns que devem ser evitadas:

- Não ter um mecanismo confiável e repetível para se manter em dia com as últimas informações sobre ameaças.
- Manter um inventário manual do portfólio de tecnologia, das workloads e das dependências que exigem análise humana de possíveis vulnerabilidades e exposições.
- Não ter mecanismos para atualizar workloads e dependências para as versões mais recentes disponíveis que ofereçam mitigações de ameaças conhecidas.

Benefícios de implementar esta prática recomendada: usar fontes de inteligência de ameaças para se manter atualizado reduz o risco de perder mudanças importantes no cenário de ameaças que podem afetar seus negócios. Ter a automação implementada para verificar, detectar e corrigir possíveis vulnerabilidades ou exposições em workloads e dependências pode ajudar você a reduzir os riscos de forma rápida e previsível em comparação com as alternativas manuais. Isso ajuda a controlar o tempo e os custos relacionados à mitigação de vulnerabilidades.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Analise publicações confiáveis de inteligência contra ameaças para ficar por dentro do cenário de ameaças. Consulte a base de conhecimento MITRE ATT&CK para obter documentação sobre táticas, técnicas e procedimentos (TTPs) conhecidos de adversários. Analise a lista de Vulnerabilidades e exposições comuns (CVE) da MITRE para se manter informado sobre vulnerabilidades conhecidas em produtos nos quais você confia. Entenda os riscos críticos das

aplicações Web com o popular projeto OWASP Top 10 do Open Worldwide Application Security Project (OWASP).

Mantenha-se em dia com os eventos de segurança da AWS e as etapas de correção recomendadas com os boletins de segurança da AWS para CVEs.

Para reduzir o esforço geral e as despesas indiretas para se manter em dia, considere usar serviços da AWS que incorporam automaticamente novas informações sobre ameaças ao longo do tempo. Por exemplo, o Amazon GuardDuty se mantém atualizado com a inteligência de ameaças do setor para detectar comportamentos anômalos e assinaturas de ameaças em suas contas. O Amazon Inspector mantém automaticamente um banco de dados dos CVEs que usa para seus recursos de verificação contínua atualizados. O AWS WAF e o AWS Shield Advanced fornecem grupos de regras gerenciados que são atualizados automaticamente à medida que novas ameaças surgem.

Revise o pilar de excelência operacional do Well-Architected para gerenciamento e correção automatizados de frotas.

Etapas de implementação

- Assine atualizações de publicações de inteligência contra ameaças que sejam relevantes para sua empresa e setor. Assine os Boletins de segurança da AWS.
- Considere a adoção de serviços que incorporem automaticamente novas informações sobre ameaças, como o Amazon GuardDuty e o Amazon Inspector.
- Implemente uma estratégia de gerenciamento e correção de frotas que se alinhe às práticas recomendadas do pilar Excelência operacional do Well-Architected.

Recursos

Práticas recomendadas relacionadas:

- SEC01-BP07 Identificar ameaças e priorizar mitigações usando um modelo de ameaça
- OPS01-BP05 Avaliar o cenário de ameaças
- OPS11-BP01 Adotar um processo para melhoria contínua

SEC01-BP05 Reduzir o escopo do gerenciamento de segurança

Determine se você pode reduzir seu escopo de segurança usando serviços da AWS que transferem o gerenciamento de determinados controles para a AWS (serviços gerenciados). Esses serviços

podem ajudar a reduzir suas tarefas de manutenção de segurança, como provisionamento de infraestrutura, configuração de software, aplicação de patches ou backups.

Resultado desejado: você considera o escopo do seu gerenciamento de segurança ao selecionar AWS serviços para sua workload. O custo referente a despesas gerais de gerenciamento e a tarefas de manutenção (o custo total de propriedade ou TCO) é ponderado em relação ao custo dos serviços que você seleciona, além de outras considerações do Well-Architected. Você incorpora a documentação de controle e conformidade da AWS em seus procedimentos de avaliação e verificação de controle.

Práticas comuns que devem ser evitadas:

- Implantar workloads sem entender completamente o modelo de responsabilidade compartilhada referente aos serviços que você seleciona.
- Hospedar bancos de dados e outras tecnologias em máquinas virtuais sem ter avaliado um serviço gerenciado equivalente.
- Não incluir tarefas de gerenciamento de segurança no custo total de propriedade de tecnologias de hospedagem em máquinas virtuais em comparação com as opções de serviços gerenciados.

Benefícios de implementar esta prática recomendada: o uso de serviços gerenciados pode reduzir sua carga geral de gerenciar controles de segurança operacional, o que pode reduzir seus riscos de segurança e o custo total de propriedade. O tempo que de outra forma seria gasto em determinadas tarefas de segurança pode ser reinvestido em tarefas que agregam maior valor aos negócios. Os serviços gerenciados também podem reduzir o escopo dos requisitos de conformidade ao transferir alguns requisitos de controle para a AWS.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Há várias maneiras de integrar os componentes da workload na AWS. A instalação e a execução de tecnologias em instâncias do Amazon EC2 geralmente exigem que você assuma a maior parte da responsabilidade geral pela segurança. Para ajudar a diminuir a sobrecarga de operar determinados controles, identifique serviços gerenciados da AWS que reduzam o escopo da sua parte do modelo de responsabilidade compartilhada e entenda como é possível usá-los em sua arquitetura atual. Os exemplos incluem o uso do Amazon Relational Database Service (Amazon RDS) para implantação de bancos de dados, do Amazon Elastic Kubernetes Service (Amazon EKS) ou do Amazon Elastic Container Service (Amazon ECS) para orquestrar contêineres ou usar opções com tecnologia sem

servidor. Ao criar novas aplicações, pense em quais serviços podem ajudar a reduzir o tempo e o custo referentes à implementação e ao gerenciamento de controles de segurança.

Os requisitos de conformidade também podem ser um fator na seleção de serviços. Os serviços gerenciados podem mudar a conformidade de alguns requisitos relacionados à AWS. Converse com sua equipe de conformidade sobre quanto ela se sente confortável para auditar os aspectos dos serviços que você opera e gerencia e aceitar declarações de controle em relatórios de auditoria relevantes da AWS. Você pode fornecer os artefatos de auditoria encontrados no AWS Artifact para seus auditores ou reguladores como evidência dos controles de segurança da AWS. Você também pode usar a orientação de responsabilidade fornecida por alguns dos artefatos de auditoria da AWS para projetar sua arquitetura, junto com os Guias de conformidade do cliente da AWS. Essas orientações ajudam a determinar os controles de segurança adicionais que você deve implementar para atender aos casos de uso específicos do seu sistema.

Ao usar serviços gerenciados, familiarize-se com o processo de atualização de recursos para versões mais recentes (por exemplo, atualizar a versão de um banco de dados gerenciado pelo Amazon RDS ou o runtime de uma linguagem de programação para um perfil do AWS Lambda). Embora o serviço gerenciado possa realizar essa operação para você, configurar o momento da atualização e entender o impacto em suas operações continua sendo sua responsabilidade. Ferramentas como essas AWS Health podem ajudar você a rastrear e gerenciar essas atualizações em todos os seus ambientes.

Etapas de implementação

1. Avalie os componentes da workload que podem ser substituídos por um serviço gerenciado.
    a. Se você estiver migrando uma workload para a AWS, considere a redução do gerenciamento (tempo e despesas) e a diminuição do risco ao avaliar se deve redefinir a hospedagem, refatorar, redefinir a plataforma, reformular, recompilar ou substituir a workload. Às vezes, investimentos adicionais no início de uma migração podem gerar economias significativas no longo prazo.
2. Pense em implementar serviços gerenciados (por exemplo, o Amazon RDS) em vez de instalar e gerenciar suas próprias implantações de tecnologia.
3. Use as orientações sobre responsabilidade no AWS Artifact para ajudar a determinar os controles de segurança que você deve implementar para a workload.
4. Mantenha um inventário dos recursos em uso e esteja sempre a par de novos serviços e abordagens a fim de identificar novas oportunidades para reduzir o escopo.

Recursos

Práticas recomendadas relacionadas:

- PERF02-BP01 Selecionar as melhores opções de computação para as workloads
- PERF03-BP01 Usar um datastore com propósitos específicos que melhor atenda aos requisitos de acesso e armazenamento de dados
- SUS05-BP03 Usar serviços gerenciados

Documentos relacionados:

- Eventos de ciclo de vida planejados para o AWS Health

Ferramentas relacionadas:

- AWS Health
- AWS Artifact
- Guias de conformidade do cliente da AWS

Vídeos relacionados:

- Como faço para migrar para uma instância de banco de dados Amazon RDS ou Aurora MySQL usando o AWS DMS?
- AWS re:Invent 2023: Gerenciar eventos do ciclo de vida dos recursos em grande escala com o AWS Health

SEC01-BP06 Automatizar a implantação de controles de segurança padrão

Aplique práticas modernas de DevOps ao desenvolver e implantar controles de segurança que são padrão em seus ambientes da AWS. Defina controles e configurações de segurança padrão usando modelos de infraestrutura como código (IaC), capture alterações em um sistema de controle de versão, teste as alterações como parte de um pipeline de CI/CD e automatize a implantação de mudanças em seus ambientes da AWS.

Resultado desejado: os modelos de IaC capturam controles de segurança padronizados e os comprometem com um sistema de controle de versão. Os pipelines de CI/CD estão em locais que

detectam mudanças e automatizam os testes e a implantação de seus ambientes da AWS. Barreiras de proteção estão em vigor para detectar e emitir alertas sobre configurações incorretas nos modelos antes de prosseguir com a implantação. As workloads são implantadas em ambientes em que há controles padrão em vigor. As equipes têm acesso para implantar configurações de serviço aprovadas por meio de um mecanismo de autoatendimento. Existem estratégias seguras de backup e recuperação para controlar configurações, scripts e dados relacionados.

Práticas comuns que devem ser evitadas:

- Fazer alterações manuais nos controles de segurança padrão por meio de um console da web ou uma interface de linha de comandos.
- Contar com equipes de workload individuais para implementar manualmente os controles definidos por uma equipe central.
- Contar com uma equipe central de segurança para implantar controles em nível de workload a pedido de uma equipe de workload.
- Permitir que as mesmas pessoas ou equipes desenvolvam, testem e implantem scripts de automação de controle de segurança sem a separação adequada de deveres ou freios e contrapesos.

Benefícios de implementar esta prática recomendada: o uso de modelos para definir seus controles de segurança padrão permite rastrear e comparar as alterações ao longo do tempo usando um sistema de controle de versão. Usar a automação para testar e implantar alterações gera padronização e previsibilidade, aumentando as chances de uma implantação bem-sucedida e reduzindo as tarefas manuais repetitivas. Fornecer um mecanismo de autoatendimento para as equipes de workload implantarem serviços e configurações aprovados reduz o risco de configuração incorreta e uso indevido. Isso também ajuda as equipes a incorporar controles logo no início no processo de desenvolvimento.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Ao seguir as práticas descritas em SEC01-BP01 Separar workloads por meio de contas, você acabará com várias Contas da AWS para diferentes ambientes que você gerencia usando o AWS Organizations. Embora cada um desses ambientes e workloads possam precisar de controles de segurança distintos, você pode padronizar alguns deles em toda a sua organização. Isso inclui integração de provedores de identidades centralizados, definição de redes e firewalls e

configuração de locais padrão para armazenar e analisar logs. Da mesma forma que você pode usar infraestrutura como código (IaC) para aplicar o mesmo rigor do desenvolvimento do código da aplicação ao provisionamento da infraestrutura, você também pode usar o IaC para definir e implantar seus controles de segurança padrão.

Sempre que possível, defina seus controles de segurança de forma declarativa, como em AWS CloudFormation, e armazene-os em um sistema de controle de origem. Use práticas de DevOps para automatizar a implantação de seus controles para obter lançamentos mais previsíveis, realizar testes automatizados usando ferramentas como o AWS CloudFormation Guard e detectar desvios entre os controles implantados e a configuração desejada. É possível usar serviços como AWS CodePipeline, AWS CodeBuild e AWS CodeDeploy para construir um pipeline de CI/CD. Considere a orientação em Organizar seu ambiente da AWS usando várias contas para configurar esses serviços em suas próprias contas, separadas de outros pipelines de implantação.

Você também pode definir modelos para padronizar a definição e a implantação de serviços, configurações e Contas da AWS. Essa técnica permite que uma equipe central de segurança gerencie essas definições e as forneça às equipes de workload por meio de uma abordagem de autoatendimento. Uma maneira de conseguir isso é usar o Service Catalog, onde você pode publicar modelos como produtos que as equipes de workload podem incorporar em suas próprias implantações de pipeline. Se você estiver usando o AWS Control Tower, alguns modelos e controles estarão disponíveis como ponto de partida. O Control Tower também fornece o recurso Account Factory, permitindo que as equipes de workload criem novas Contas da AWS usando os padrões definidos por você. Com esse recurso, não é preciso depender de uma equipe central para aprovar e criar contas quando elas são identificadas como necessárias pelas equipes de workload. Talvez essas contas precisem isolar diferentes componentes da workload com base em motivos como a função que eles desempenham, a confidencialidade dos dados que estão sendo processados ou o comportamento desses componentes.

Etapas de implementação

1. Determine como você armazenará e manterá os modelos em um sistema de controle de versão.
2. Crie pipelines de CI/CD para testar e implantar modelos. Defina testes para verificar se há configurações incorretas e se os modelos estão de acordo com os padrões da sua empresa.
3. Crie um catálogo de modelos padronizados para as equipes de workload implantarem serviços e Contas da AWS de acordo com suas necessidades.
4. Implemente estratégias seguras de backup e recuperação para suas configurações de controle, scripts e dados relacionados.

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP01 Usar controle de versão
- OPS05-BP04 Usar sistemas de gerenciamento de compilação e implantação
- REL08-BP05 Implantar alterações com automação
- SUS06-BP01 Adotar métodos que podem introduzir rapidamente melhorias na sustentabilidade

Documentos relacionados:

- Organizar seu ambiente da AWS usando várias contas

Exemplos relacionados:

- Automatize a criação de contas e o provisionamento de recursos usando Service Catalog, o AWS Organizations e o AWS Lambda
- Fortaleça o pipeline de DevOps e proteja os dados com o AWS Secrets Manager, o AWS KMS e o AWS Certificate Manager

Ferramentas relacionadas:

- AWS CloudFormation Guard
- Acelerador de zona de pouso na AWS

SEC01-BP07 Identificar ameaças e priorizar mitigações usando um modelo de ameaça

Realize a modelagem de ameaças para identificar e manter um registro atualizado de possíveis ameaças e mitigações associadas para sua workload. Priorize suas ameaças e adapte as mitigações de controles de segurança para prevenir, detectar e responder. Revise e mantenha isso no contexto de sua workload e no cenário de segurança em evolução.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

O que é modelagem de ameaças?

"A modelagem de ameaças funciona para identificar, comunicar e entender ameaças e mitigações no contexto da proteção de algo de valor." – Modelagem de ameaças a aplicações do Open Web Application Security Project (OWASP)

Por que você deveria usar um modelo de ameaça?

Os sistemas são complexos e se tornam cada vez mais intrincados e qualificados com o passar do tempo, oferecendo maior valor empresarial e maior satisfação e engajamento do cliente. Isso significa que as decisões de design de TI precisam considerar um número cada vez maior de casos de uso. Essa complexidade e o número de permutações de caso de uso geralmente tornam as abordagens não estruturadas ineficazes para encontrar e mitigar ameaças. Em vez disso, você precisa de uma abordagem sistemática para enumerar as possíveis ameaças ao sistema, elaborar mitigações e priorizá-las a fim de garantir que os recursos limitados de sua organização tenham impacto máximo na melhoria do procedimento geral de segurança do sistema.

A modelagem de ameaças foi projetada para oferecer essa abordagem sistemática com o objetivo de encontrar e resolver problemas na fase inicial do processo de design, quando as mitigações têm custo e esforço relativamente baixos em comparação com a fase posterior do ciclo de vida. Essa abordagem se alinha ao princípio de segurança shift-left do setor. Por fim, a modelagem de ameaças é integrada ao processo de gerenciamento de riscos de uma organização e ajuda a impulsionar as decisões sobre quais controles implementar usando uma abordagem orientada a ameaças.

Quando a modelagem de ameaças deve ser realizada?

Inicie a modelagem de ameaças o quanto antes no ciclo de vida de sua workload. Isso oferece a você maior flexibilidade sobre o que fazer com as ameaças identificadas. Muito semelhante aos bugs de software, quanto mais cedo você identificar as ameaças, mais econômico será resolvê-las. Um modelo de ameaças é um documento ativo e deve continuar a evoluir à medida que suas workloads mudam. Revise seus modelos de ameaça no decorrer do tempo, inclusive quando há uma alteração importante ou uma alteração no cenário de ameaças ou ao adotar um novo recurso ou serviço.

Etapas de implementação

Como podemos realizar a modelagem de ameaças?

Há muitas formas diferentes de realizar a modelagem de ameaças. Muito semelhante às linguagens de programação, há vantagens e desvantagens em cada uma, e é necessário escolher a forma mais adequada para você. Uma abordagem é começar com o Quadro de quatro perguntas para modelagem de ameaças do Shostack, que apresenta perguntas abertas para fornecer estrutura ao seu exercício de modelagem de ameaças:

1. Em que estão trabalhando?

   A finalidade dessa pergunta é ajudar você a entender e chegar a um acordo sobre o sistema que você está construindo e os detalhes sobre ele que são relevantes para a segurança. Criar um modelo ou diagrama é a forma mais popular de responder a essa pergunta, pois ajuda a visualizar o que você está criando, por exemplo, usando um diagrama de fluxo de dados. Escrever as suposições e os detalhes importantes sobre seu sistema também ajuda a definir o que está no escopo. Isso permite que todos que estão contribuindo para o modelo de ameaças se concentrem na mesma coisa e evitem desvios demorados para tópicos fora do escopo (inclusive versões desatualizadas do sistema). Por exemplo, se você está criando uma aplicação Web, provavelmente não vale a pena criar uma modelagem de ameaças da sequência de inicialização confiável do sistema operacional para clientes de navegador, pois não há nenhuma possibilidade de seu design ter influência nisso.

2. O que pode acontecer de errado?

   É nessa fase que você identifica ameaças ao seu sistema. Ameaças são ações ou eventos acidentais ou intencionais que causam impactos indesejados que podem afetar a segurança de seu sistema. Sem um claro entendimento do que pode dar errado, não há o que fazer sobre isso.

   Não há uma lista canônica do que pode dar errado. A criação dessa lista exige um brainstorming e a colaboração entre todos os indivíduos de sua equipe e pessoas relevantes envolvidas no exercício de modelagem de ameaças. Você pode ajudar seu brainstorming usando um modelo para identificar ameaças, como o STRIDE, que sugere diferentes categorias para avaliação: falsificação, adulteração, repúdio, divulgação de informações, negação de serviço e elevação de privilégios. Além disso, talvez você queira ajudar no brainstorming revisando as listas e pesquisas existentes em busca de inspiração, incluindo o OWASP Top 10, o HiTrust Threat Catalog e o catálogo de ameaças da sua própria organização.

3. O que vamos fazer sobre isso?

   Como no caso da primeira pergunta, não há uma lista canônica de todas as mitigações possíveis. A entradas nessa etapa são as ameaças identificadas, as pessoas e as áreas de melhoria da etapa anterior.

   Segurança e conformidade são uma responsabilidade compartilhada entre você e a AWS. É importante entender que ao perguntar "O que vamos fazer a respeito?" você também está perguntando "Quem é responsável por fazer algo a respeito?". Entender o equilíbrio entre suas responsabilidades e as da AWS ajuda a definir o escopo de seu exercício de modelagem de

ameaças para as mitigações que estão sob seu controle, que, geralmente, são uma combinação de opções de configuração de serviços da AWS e suas mitigações específicas do sistema.

Para a parte da AWS da responsabilidade compartilhada, você descobrirá que os serviços da AWS estão no escopo de muitos programas de conformidade. Esses programas ajudam você a entender os controles sólidos implementados na AWS para manter a segurança e a conformidade da nuvem. Os relatórios de auditoria desses programas estão disponíveis para download para clientes da AWS no AWS Artifact.

Seja quais forem os serviços da AWS que você está utilizando, sempre há um elemento de responsabilidade do cliente, e as mitigações alinhadas a essas responsabilidades devem ser incluídas em seu modelo de ameaças. Para mitigações de controle de segurança dos próprios serviços da AWS, convém considerar a implementação de controles de segurança em todos os domínios; por exemplo, domínios como gerenciamento de identidade e acesso (autenticação e autorização), proteção de dados (em repouso e em trânsito), segurança de infraestrutura, registro em log e monitoramento. A documentação de cada serviço da AWS conta com um capítulo de segurança dedicado que fornece orientação sobre os controles de segurança a serem considerados como mitigações. É importante considerar o código que você está escrevendo e suas dependências e pensar nos controles que você poderia implementar para resolver essas ameaças. Esses controles podem ser coisas como validação de entrada, tratamento de sessão e tratamento de limites. Com frequência, a maioria das vulnerabilidades é introduzida em código personalizado. Por isso, concentre-se nessa área.

4. Fizemos um bom trabalho?

   O objetivo é a sua equipe e a organização aprimorarem a qualidade dos modelos de ameaças e a velocidade na qual você está realizando a modelagem de ameaças no decorrer do tempo. Essas melhorias vêm de uma combinação entre prática, aprendizado, instrução e revisão. Para se aprofundar e colocar a mão na massa, é recomendável que você e sua equipe concluam o workshop ou curso de treinamento Modelagem de ameaças da maneira certa para construtores. Além disso, se você estiver procurando orientação sobre como integrar a modelagem de ameaças ao ciclo de vida de desenvolvimento de aplicações da sua organização, consulte Como abordar a modelagem de ameaças no blog de segurança da AWS.

Compositor de ameaças

Para obter ajuda e orientação na execução da modelagem de ameaças, considere usar a ferramenta Threat Composer, que visa reduzir o tempo de obtenção de valor na modelagem de ameaças. Essa ferramenta ajuda você a:

- Escrever declarações de ameaças úteis alinhadas à gramática de ameaças que funcionem em um fluxo de trabalho natural não linear
- Gerar um modelo de ameaça legível por humanos.
- Gerar um modelo de ameaça legível por máquina para permitir tratar os modelos de ameaças como código.
- Identificar rapidamente as áreas de melhoria de qualidade e de cobertura usando o painel do Insights.

Para obter mais referências, visite o Threat Composer e alterne para o Exemplo de espaço de trabalho definido pelo sistema.

Recursos

Práticas recomendadas relacionadas:

- SEC01-BP03 Identificar e validar objetivos de controle
- SEC01-BP04 Manter-se em dia com ameaças e recomendações de segurança
- SEC01-BP05 Reduzir o escopo do gerenciamento de segurança
- SEC01-BP08 Avaliar e implementar regularmente novos serviços e recursos de segurança

Documentos relacionados:

- Como abordar a modelagem de ameaças (Blog de segurança da AWS)
- NIST: Guia para modelagem de ameaças a sistemas centrada em dados

Vídeos relacionados:

- AWS Summit ANZ 2021: Como abordar a modelagem de ameaças
- AWS Summit ANZ 2022: Escalar a segurança: otimizar para entrega rápida e segura

Treinamento relacionado:

- Modelagem de ameaças da maneira certa para criadores: treinamento individualizado virtual do AWS Skill Builder
- Modelagem de ameaças da maneira certa para construtores: workshop da AWS

Ferramentas relacionadas:

- Compositor de ameaças

SEC01-BP08 Avaliar e implementar regularmente novos serviços e recursos de segurança

Avalie e implemente serviços e recursos de segurança da AWS e de parceiros da AWS que ajudem você a desenvolver o procedimento de segurança de suas workloads.

Resultado desejado: você tem uma prática padrão em vigor que informa sobre novos recursos e serviços lançados pela AWS e os parceiros da AWS. Você avalia como esses novos recursos influenciam o design dos controles novos e atuais de seus ambientes e workloads.

Práticas comuns que devem ser evitadas:

- Não assinar blogs e feeds RSS da AWS para tomar conhecimento rapidamente de novos recursos e serviços relevantes.
- Recorrer a notícias e atualizações sobre serviços e recursos de segurança de fontes secundárias.
- Não incentivar os usuários da AWS da sua organização a se manterem informados sobre as atualizações mais recentes

Benefícios de implementar esta prática recomendada: ao se manter atualizado sobre os novos serviços e recursos de segurança, você pode tomar decisões informadas sobre a implementação de controles em seus ambientes e workloads na nuvem. Essas fontes ajudam a aumentar a conscientização sobre o cenário de segurança em constante evolução e como os serviços da AWS podem ser usados para impedir ameaças novas e emergentes.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

A AWS informa os clientes sobre novos serviços e recursos de segurança por meio de vários canais:

- Novidades da AWS

- Notícias do blog da AWS
- Blog de segurança da AWS
- Boletins de segurança da AWS
- Visão geral da documentação da AWS

Você pode assinar um tópico do AWS Daily Feature Updates usando o Amazon Simple Notification Service (Amazon SNS) para obter um resumo diário abrangente das atualizações. Alguns serviços de segurança, como o Amazon GuardDuty e o AWS Security Hub, fornecem seus próprios tópicos de SNS para manter você em dia com os novos padrões, descobertas e outras atualizações desses serviços específicos.

Novos serviços e recursos também são anunciados e descritos em detalhes durante conferências, eventos e webinars realizados em todo o mundo a cada ano. Em destaque está a conferência anual de segurança AWS re:Inforce e a conferência mais geral AWS re:Invent. Os canais de notícias da AWS mencionados anteriormente compartilham esses anúncios de conferências sobre segurança e outros serviços, e você pode assistir a sessões educacionais aprofundadas on-line no canal AWS Events no YouTube.

Você também pode perguntar à equipe da sua Conta da AWS sobre as atualizações e recomendações mais recentes do serviço de segurança. Para entrar em contato com sua equipe, use o formulário de Suporte de Vendas se não tiver as informações de contato direto. Da mesma forma, se você se inscreveu no AWS Enterprise Support, receberá atualizações semanais do seu gerente técnico de contas (TAM) e poderá agendar uma reunião regular de revisão com ele.

Etapas de implementação

1. Assine vários blogs e boletins com o leitor de RSS de sua preferência ou o tópico de atualizações diárias de recursos do SNS.
2. Avalie de quais eventos da AWS você deve participar para saber em primeira mão sobre novos recursos e serviços.
3. Agende reuniões com a equipe da sua Conta da AWS para esclarecer dúvidas sobre a atualização de serviços e recursos de segurança.
4. Considere a possibilidade de assinar o Enterprise Support para consultar regularmente um gerente técnico de contas (TAM).

Recursos

Práticas recomendadas relacionadas:

- PERF01-BP01 Conhecer e compreender os serviços e recursos de nuvem disponíveis
- COST01-BP07 Manter-se em dia com os novos lançamentos de serviços

Gerenciamento de identidade e acesso

Perguntas

- SEC 2. Como gerenciar a autenticação de pessoas e máquinas?
- SEC 3. Como você gerencia as permissões para pessoas e máquinas?

SEC 2. Como gerenciar a autenticação de pessoas e máquinas?

Há dois tipos de identidade que você deverá gerenciar para operar workloads seguras da AWS. Entender o tipo de identidade de que você precisa para gerenciar e conceder acesso ajuda a garantir que as identidades corretas tenham acesso aos recursos certos nas condições certas.

Identidades humanas: seus administradores, desenvolvedores, operadores e usuários finais precisam de uma identidade para acessar seus ambientes e aplicações da AWS. Eles são membros da sua organização ou usuários externos com quem você colabora e que interagem com seus recursos da AWS por meio de um navegador da Web, aplicação do cliente ou ferramentas interativas de linha de comando.

Identidades de máquina: aplicações de serviço, ferramentas operacionais e workloads precisam de uma identidade para fazer solicitações a serviços da AWS, como para ler dados. Essas identidades incluem máquinas em execução em seu ambiente da AWS, como instâncias do Amazon EC2 ou funções do AWS Lambda. Você também pode gerenciar identidades de máquina para partes externas que precisam de acesso. Além disso, você pode ter máquinas fora da AWS que precisam de acesso aos seus ambientes da AWS.

Práticas recomendadas

- SEC02-BP01 Usar mecanismos de início de sessão fortes
- SEC02-BP02 Usar credenciais temporárias
- SEC02-BP03 Armazenar e usar segredos com segurança

- SEC02-BP04 Confiar em um provedor de identidades centralizado
- SEC02-BP05 Auditar e fazer a rotação das credenciais periodicamente
- SEC02-BP06 Utilizar grupos de usuários e atributos

SEC02-BP01 Usar mecanismos de início de sessão fortes

Os inícios de sessão (autenticação com credenciais de login) podem apresentar riscos quando não são usados mecanismos, como autenticação multifator (MFA), especialmente em situações em que as credenciais de login foram divulgadas acidentalmente ou podem ser deduzidas com facilidade. Utilize mecanismos de início de sessão fortes para reduzir essas riscos exigindo MFA e políticas de senhas fortes.

Resultado desejado: reduza os riscos de acesso não intencional às credenciais na AWS usando mecanismos de início de sessão robustos para usuários do AWS Identity and Access Management (IAM), o usuário-raiz da Conta da AWS, o AWS IAM Identity Center (sucessor do AWS Single Sign-On) e provedores de identidade terceirizados. Isso significa exigir MFA, impor políticas de senhas fortes e detectar comportamento de login anômalo.

Práticas comuns que devem ser evitadas:

- Não impor uma política de senhas fortes para suas identidades incluindo senhas complexas e MFA.
- Compartilhar as mesmas credenciais entre usuários diferentes.
- Não utilizar controles de detecção para logins suspeitos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Há muitas formas de identidades humanas fazerem iniciarem sessão na AWS. É prática recomendada da AWS confiar em um provedor de identidades centralizado utilizando federação (federação direta ou via AWS IAM Identity Center) ao realizar a autenticação na AWS. Nesse caso, você deve estabelecer um processo de início de sessão seguro com seu provedor de identidades ou o Microsoft Active Directory.

Ao abrir pela primeira vez uma Conta da AWS, você começa com um usuário-raiz da Conta da AWS. Você só deve usar o usuário-raiz da conta para configurar o acesso para seus usuários (e para tarefas que exijam o usuário-raiz). É importante ativar o MFA para o usuário-raiz da conta

imediatamente após abrir sua conta Conta da AWS e proteger o usuário-raiz usando o guia de práticas recomendadas da AWS.

Se você criar usuários no AWS IAM Identity Center, proteja o processo de início de sessão nesse serviço. Para identidades de consumidores, é possível usar grupos de usuários do Amazon Cognito e proteger o processo de início de sesão nesse serviço ou usar um dos provedores de identidades aos quais os grupos de usuários do Amazon Cognito oferecem suporte.

Se você estiver usando usuários do AWS Identity and Access Management (IAM), poderá proteger processo de início de sessão usando o IAM.

Seja qual for o método de início de sessão, é essencial impor uma política de login forte.

Etapas de implementação

Veja a seguir as recomendações gerais de início de sessão forte. As configurações reais que você configurar deverão ser definidas pela política da sua empresa ou usar um padrão como o NIST 800-63.

- Solicite a MFA. É prática recomendada do IAM exigir MFA para identidades humanas e workloads. A ativação da MFA oferece uma camada adicional de segurança que exige que os usuários forneçam credenciais de início de sessão e uma senha de uso único (OTP) ou uma string gerada e verificada criptograficamente por um dispositivo de hardware.
- Imponha um comprimento mínimo de senha, que é um fator essencial da força da senha.
- Imponha complexidade para tornar as senhas mais difíceis de deduzir.
- Permitir que os usuários troquem suas próprias senhas.
- Crie identidades individuais em vez de credenciais compartilhadas. Com a criação de identidades individuais, é possível fornecer a cada usuário um conjunto exclusivo de credenciais de segurança. Os usuários individuais oferecem a capacidade de auditar a atividade de cada usuário.

Recomendações do Centro de Identidade do IAM:

- O Centro de Identidade do IAM fornece uma política de senha predefinida ao usar o diretório padrão que estabelece o tamanho, a complexidade e os requisitos de reutilização da senha.
- Ative o MFA e defina a configuração contextual ou sempre ativa do MFA quando a fonte de identidade for o diretório padrão, o AWS Managed Microsoft AD ou o AD Connector.
- Permita que os usuários registrem seus próprios dispositivos de MFA.

Recomendações do diretório de grupos de usuários do Amazon Cognito:

- Configure as opções de força da senha.
- Exija MFA para os usuários.
- Use as configurações de segurança avançadas dos grupos de usuários do Amazon Cognito para recursos como a autenticação adaptativa, a qual pode bloquear logins suspeitos.

Recomendações para usuários do IAM:

- Em teoria, você está utilizando Centro de Identidade do IAM ou federação direta. No entanto, talvez você precise de usuários do IAM. Nesse caso, defina uma política de senhas para usuários do IAM. A política de senhas pode ser usada para definir requisitos como extensão mínima ou a obrigatoriedade de uso de caracteres não alfabéticos.
- Crie uma política do IAM para impor o início de sessão com MFA para que os usuários possam gerenciar suas próprias senhas e dispositivos de MFA.

Recursos

Práticas recomendadas relacionadas:

- SEC02-BP03 Armazenar e usar segredos com segurança
- SEC02-BP04 Confiar em um provedor de identidades centralizado
- SEC03-BP08 Compartilhar recursos com segurança em sua organização

Documentos relacionados:

- Política de senhas do AWS IAM Identity Center
- Política de senhas de usuários do IAM
- Definir da senha do usuário-raiz da Conta da AWS
- Política de senhas do Amazon Cognito
- Credenciais da AWS
- Práticas recomendadas de segurança do IAM

Vídeos relacionados:

- Gerenciar permissões de usuário em escala com o AWS IAM Identity Center
- Como dominar a identidade em cada camada do bolo

SEC02-BP02 Usar credenciais temporárias

Ao realizar qualquer tipo de autenticação, é melhor utilizar credenciais temporárias em vez de credenciais de longo prazo a fim de reduzir ou eliminar riscos como credenciais que são divulgadas acidentalmente, compartilhadas ou roubadas.

Resultado desejado: para reduzir o risco de credenciais de longo prazo, use credenciais temporárias sempre que possível para identidades humanas e de máquinas. Credenciais de longo prazo criam muitos riscos. Por exemplo, é possível fazer upload delas em código para repositórios públicos do GitHub. Ao utilizar credenciais temporárias, você reduz significativamente as chances de comprometimento das credenciais.

Práticas comuns que devem ser evitadas:

- Desenvolvedores que usam chaves de acesso de longo prazo de usuários do IAM em vez de obter credenciais temporárias da CLI usando federação.
- Desenvolvedores que incorporam chaves de acesso de longo prazo no código e fazem upload desse código para repositórios públicos do Git.
- Desenvolvedores que incorporam chaves de acesso de longo prazo em aplicações móveis que, depois, são disponibilizadas em lojas de aplicações.
- Usuários que compartilham chaves de acesso de longo prazo com outros usuários ou funcionários que deixam a empresa e não devolvem as chaves de acesso de longo prazo.
- Utilizar chaves de acesso de longo prazo para identidades de máquina quando é possível usar credenciais temporárias.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Utilize credenciais de segurança temporárias em vez de credenciais de longo prazo para todas as solicitações à AWS API e CLI. As solicitações de API e CLI para serviços da AWS devem, em quase todos os casos, ser assinadas usando chaves de acesso da AWS. Essas solicitações podem ser assinadas com credenciais temporárias ou de longo prazo. A única vez em que você deve usar credenciais de longo prazo, também conhecidas como chaves de acesso de longo prazo, é se

estiver usando um usuário do IAM ou o usuário-raiz da Conta da AWS. Quando você se federa à AWS ou assume um perfil do IAM por meio de outros métodos, credenciais temporárias são geradas. Mesmo quando você acessa o AWS Management Console utilizando credenciais de login, credenciais temporárias são geradas para você fazer chamadas para serviços da AWS. Há poucas situações nas quais você precisa de credenciais de longo prazo, e é possível realizar quase todas as tarefas usando credenciais temporárias.

Evitar o uso de credenciais de longo prazo em favor de credenciais temporárias deve andar lado a lado com uma estratégia de reduzir o uso de usuários do IAM em favor da federação e de perfis do IAM. Embora usuários do IAM tenham sido usados para identidades humanas e de máquina no passado, agora recomendamos não utilizá-los para evitar os riscos de utilizar chaves de acesso de longo prazo.

Etapas de implementação

Para identidades humanas, como funcionários, administradores, desenvolvedores, operadores e clientes:

- Recomenda-se confiar em um provedor de identidade centralizado e exigir que os usuários humanos usem federação com um provedor de identidades para acessar a AWS usando credenciais temporárias. A federação para seus usuários pode ser feita com federação direta para cada Conta da AWS ou usando o AWS IAM Identity Center e um provedor de identidades escolhido por você. A federação oferece uma série de vantagens em comparação com a utilização de usuários do IAM que vão além de eliminar credenciais de longo prazo. Seus usuários também podem solicitar credenciais temporárias na linha de comando para federação direta ou usando o Centro de Identidade do IAM. Isso significa que há poucos casos de uso que exigem usuários do IAM ou credenciais de longo prazo para seus usuários.
- Ao conceder a terceiros, como provedores de software como serviço (SaaS), acesso aos recursos em sua Conta da AWS, você pode usar funções entre contas e políticas baseadas em recursos.
- Se precisar conceder às aplicações para consumidores ou clientes acesso aos seus recursos da AWS, você pode usar bancos de identidades do Amazon Cognito ou grupos de usuários do Amazon Cognito para fornecer credenciais temporárias. As permissões para as credenciais são configuradas por meio de perfis do IAM. Você também pode definir um perfil do IAM separado com permissões limitadas para usuários convidados que não são autenticados.

Para identidades de máquina, talvez seja necessário utilizar credenciais de longo prazo. Nesses casos, exija que as workloads usem credenciais temporárias com perfis do IAM para acessar a AWS.

- Para o Amazon Elastic Compute Cloud (Amazon EC2), é possível usar perfis para o Amazon EC2.

- O AWS Lambda permite configurar um perfil de execução do Lambda para conceder ao serviço permissões para realizar ações AWS usando credenciais temporárias. Há muitos outros modelos semelhantes para os serviços da AWS concederem credenciais temporárias utilizando perfis do IAM.
- Para dispositivos de IoT, você pode usar o provedor de credenciais do AWS IoT Core para solicitar credenciais temporárias.
- Para sistemas on-premises ou sistemas que funcionam fora da AWS e que precisam de acesso a recursos da AWS, é possível usar o IAM Roles Anywhere.

Há cenários em que credenciais temporárias não são uma opção e talvez seja necessário usar credenciais de longo prazo. Nessas situações, audite e alterne as credenciais periodicamente e alterne as chaves de acesso regularmente para casos de uso que exigem credenciais de longo prazo. Alguns exemplos que podem exigir credenciais de longo prazo incluem plug-ins do WordPress e clientes da AWS de terceiros. Em situações em que você precisa usar credenciais de longo prazo, ou para credenciais que não sejam chaves de acesso da AWS, como logins de bancos de dados, é possível usar um serviço projetado para lidar com o gerenciamento de segredos, como o AWS Secrets Manager. O Secrets Manager facilita o gerenciamento, a rotação e o armazenamento seguro de segredos criptografados usando os serviços compatíveis. Para obter mais informações sobre a mudança de credenciais de longo prazo, consulte a mudança de chaves de acesso

Recursos

Práticas recomendadas relacionadas:

- SEC02-BP03 Armazenar e usar segredos com segurança
- SEC02-BP04 Confiar em um provedor de identidades centralizado
- SEC03-BP08 Compartilhar recursos com segurança em sua organização

Documentos relacionados:

- Credenciais de segurança temporárias
- Credenciais da AWS
- Práticas recomendadas de segurança do IAM
- Perfis do IAM

- Centro de Identidade do IAM
- Provedores de identidade e federação
- Fazer a rotação das chave de acesso
- Soluções para parceiros de segurança: acesso e controle
- O usuário-raiz da conta da AWS

Vídeos relacionados:

- Gerenciar permissões de usuário em escala com o AWS IAM Identity Center
- Como dominar a identidade em cada camada do bolo

SEC02-BP03 Armazenar e usar segredos com segurança

Uma workload exige um recurso automatizado para comprovar a identidade dela em bancos de dados, recursos e serviços de terceiros. Isso é feito com o uso de credenciais de acesso secretas, como chaves de acesso de API, senhas e tokens do OAuth. Utilizar um serviço com propósito específico para armazenar, gerenciar e fazer a rotação de credenciais ajuda a reduzir a probabilidade de comprometimento dessas credenciais.

Resultado desejado: implementar um mecanismo para gerenciar com segurança credenciais da aplicação que atinja as seguintes metas:

- Identificar quais segredos são necessários para a workload.
- Reduzir o número de credenciais de longo prazo necessárias substituindo-as por credenciais de curto prazo quando possível.
- Estabelecer um armazenamento seguro e uma rotação automatizada das credenciais de longo prazo restantes.
- Auditar o acesso aos segredos existentes na workload.
- Monitorar continuamente para confirmar que nenhum segredo seja incorporado ao código-fonte durante o processo de desenvolvimento.
- Reduzir a probabilidade de divulgação acidental de credenciais.

Práticas comuns que devem ser evitadas:

- Ausência de rotação de credenciais.

- Armazenar credenciais de longo prazo em código-fonte ou arquivos de configuração.
- Armazenar credenciais em repouso não criptografadas.

Benefícios de implementar esta prática recomendada:

- Os segredos são armazenados com criptografia em repouso e em trânsito.
- O acesso às credenciais é controlado por meio de uma API (pense nisso como uma máquina de venda automática de credenciais).
- O acesso a uma credencial (de leitura e gravação) é auditado e registrado.
- Separação de preocupações: a rotação de credenciais é realizada por um componente separado que pode ser segregado do restante da arquitetura.
- Os segredos são automaticamente distribuídos sob demanda em componentes de software e a rotação ocorre em um local central.
- O acesso às credenciais pode ser controlado de forma detalhada.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

No passado, as credenciais usadas para realizar a autenticação em bancos de dados, APIs de terceiros, tokens e outros segredos podiam ser incorporadas em código-fonte ou em arquivos do ambiente. A AWS oferece vários mecanismos para armazenar essas credenciais com segurança, alterná-las automaticamente e auditar o uso delas.

A melhor forma de abordar o gerenciamento de segredos é seguir as orientações de remover, substituir e fazer a rotação. A credencial mais segura é a que você não precisa armazenar, gerenciar nem processar. Pode haver credenciais que não sejam mais necessárias ao funcionamento da workload que podem ser removidas com segurança.

Para credenciais que ainda são necessárias ao funcionamento adequado da workload, pode haver uma oportunidade de substituir uma credencial de longo prazo por uma credencial temporária ou de curto prazo. Por exemplo, em vez de codificar uma chave de acesso secreta da AWS, considere substituir essa credencial de longo prazo por uma temporária utilizando perfis do IAM.

Em algumas situações, talvez alguns segredos de longa duração não possam removidos ou substituídos. Esses segredos podem ser armazenados em um serviço, como o AWS Secrets

Manager, onde podem ser armazenados, gerenciados e ter a rotação feita centralmente de tempos em tempos.

Uma auditoria do código-fonte da workload e os arquivos de configuração podem revelar muitos tipos de credencial. A seguinte tabela resume as estratégias para lidar com tipos comuns de credenciais:

Tipo de credencial	Descrição	Estratégia sugerida
Chaves de acesso ao IAM	Chaves secretas e de acesso do AWS IAM usadas para assumir perfis do IAM dentro de uma workload	Substitua: em vez disso, use perfis do IAM atribuídos às instâncias de computação (como Amazon EC2 ou AWS Lambda). Para interoperabilidade com terceiros que precisam de acesso a recursos em sua Conta da AWS, pergunte se eles oferecem suporte ao acesso entre contas da AWS. Para aplicações móveis, considere usar credenciais temporárias nos bancos de identidades do Amazon Cognito (identidades federadas). Para workloads executadas fora da AWS, considere o usar o IAM Roles Anywhere ou o AWS Systems Manager Hybrid Activations.
Chaves SSH	Chaves privadas do Secure Shell usadas para fazer login em instâncias do EC2 Linux, manualmente ou como parte de um processo automatizado	Substitua: use o AWS Systems Manager ou o EC2 Instance Connect para fornecer acesso programático e humano às instâncias do EC2 usando perfis do IAM.
Credenciais de aplicações e bancos de dados	Senhas: sequência de texto simples	Rotação: armazene as credenciais no AWS Secrets

Tipo de credencial	Descrição	Estratégia sugerida
		Manager e estabeleça a rotação automática, se possível.
Credenciais do Amazon RDS e do Aurora Admin Database	Senhas: sequência de texto simples	Substitua: use a integração do Secrets Manager com o Amazon RDS ou o Amazon Aurora. Além disso, alguns tipos de banco de dados do RDS podem usar perfis do IAM em vez de senhas para alguns casos de uso (para obter mais detalhes, consulte Autenticação de banco de dados do IAM).
Tokens OAuth	Tokens secretos: sequência de texto simples	Rotação: armazene tokens no AWS Secrets Manager e configure a rotação automátic a.
Chaves e tokens de API	Tokens secretos: sequência de texto simples	Rotação: armazene no AWS Secrets Manager e estabeleç a a rotação automática, se possível.

Um prática não recomendada comum é incorporar chaves de acesso do IAM a código-fonte, arquivos de configuração ou aplicações móveis. Quando uma chave de acesso do IAM for necessária para se comunicar com um serviço da AWS, use credenciais de segurança temporárias (de curto prazo). Essas credenciais de curto prazo podem ser fornecidas por meio de perfis do IAM para instâncias do EC2, funções de execução para funções Lambda, perfis do IAM do Cognito para acesso de usuários móveis e políticas do IoT Core para dispositivos de IoT. Ao interagir com terceiros, prefira delegar acesso a um perfil do IAM com o acesso necessário aos recursos da sua conta em vez de configurar um usuário do IAM e enviar para terceiros a chave de acesso secreta desse usuário.

Há muitos casos em que a workload exige o armazenamento dos segredos necessários para interoperar com outros serviços e recursos. O AWS Secrets Manager foi criado especificamente para gerenciar com segurança essas credenciais, bem como o armazenamento, o uso e a rotação de tokens de API, senhas e outras credenciais.

O AWS Secrets Manager oferece cinco recursos principais para garantir o armazenamento e o manuseio seguros de credenciais confidenciais: criptografia em repouso, criptografia em trânsito, auditoria abrangente, controle de acesso refinado e rotação extensível de credenciais. Outros serviços de gerenciamento de segredos de parceiros da AWS ou soluções desenvolvidas localmente que oferecem recursos e garantias semelhantes também são aceitáveis.

Etapas de implementação

1. Identifique caminhos de código contendo credenciais codificadas usando ferramentas automatizadas, como o Amazon CodeGuru.
   a. Utilize o Amazon CodeGuru para verificar seus repositórios de código. Quando a revisão estiver concluída, filtre `Type=Secrets` no CodeGuru para encontrar linhas de código problemáticas.
2. Identifique credenciais que possam ser removidas ou substituídas.
   a. Identifique credenciais não mais necessárias e marque-as para remoção.
   b. Para chaves secretas da AWS incorporadas ao código-fonte, substitua-as por perfis do IAM associados aos recursos necessários. Se parte da sua workload for externa à AWS, mas exigir credenciais do IAM para acessar os recursos da AWS, considere usar o IAM Roles Anywhere ou o AWS Systems Manager Hybrid Activations.
3. Para outros segredos duradouros de terceiros que exijam o uso da estratégia de rotação, integre o Secrets Manager ao seu código para recuperar segredos de terceiros em tempo de execução.
   a. O console do CodeGuru pode criar automaticamente um segredo no Secrets Manager usando as credenciais descobertas.
   b. Integre a recuperação de segredos do Secrets Manager ao código da sua aplicação.
      i. As funções do Lambda sem servidor podem usar uma extensão do Lambda independente de linguagem.
      ii. Para instâncias ou contêineres do EC2, a AWS fornece exemplos de código do lado do cliente para recuperar segredos do Secrets Manager em várias linguagens de programação populares.
4. Revise periodicamente sua base de código e verifique novamente para confirmar se não há novos segredos adicionados ao código.

   a. Considere usar uma ferramenta como git-secrets para evitar a confirmação de novos segredos em seu repositório de código-fonte.
5. Monitore a atividade do Secrets Manager em busca de indicações de uso inesperado, acesso inadequado a segredos ou tentativas de exclusão de segredos.
6. Reduza a exposição humana às credenciais. Restrinja o acesso a credenciais de leitura, gravação e modificação a um perfil do IAM dedicado a esse fim, e apenas forneça acesso para assumir o perfil a um pequeno subconjunto de usuários operacionais.

Recursos

Práticas recomendadas relacionadas:

- SEC02-BP02 Usar credenciais temporárias
- SEC02-BP05 Auditar e fazer a rotação das credenciais periodicamente

Documentos relacionados:

- Conceitos básicos do AWS Secrets Manager
- Provedores de identidade e federação
- Amazon CodeGuru apresenta o detector de segredos
- Como o AWS Secrets Manager usa o AWS Key Management Service.
- Criptografia e descriptografia de segredos no Secrets Manager
- Entradas de blog do Secrets Manager
- Amazon RDS anuncia integração com AWS Secrets Manager

Vídeos relacionados:

- Práticas recomendadas para gerenciar, recuperar e fazer a rotação de segredos em escala
- Encontrar segredos codificados com o detector de segredos do Amazon CodeGuru
- Proteger segredos para workloads híbridas usando o AWS Secrets Manager

Workshops relacionados:

- Armazenar, recuperar e gerenciar credenciais confidenciais no AWS Secrets Manager

- Ativações híbridas do AWS Systems Manager

SEC02-BP04 Confiar em um provedor de identidades centralizado

Para identidades da força de trabalho (funcionários e prestadores de serviços), confie em um provedor de identidade que permita gerenciar identidades em um local centralizado. Isso facilita o gerenciamento do acesso em várias aplicações e sistemas, pois você está criando, atribuindo, gerenciando, revogando e auditando o acesso de um único local.

Resultado desejado: você tem um provedor de identidade centralizado no qual gerencia centralmente os usuários da força de trabalho, as políticas de autenticação (como a exigência de autenticação multifator (MFA)) e a autorização para sistemas e aplicações (como atribuir acesso com base na associação ou nos atributos do grupo de um usuário). Os usuários da sua força de trabalho fazem login no provedor de identidade central e se federam (autenticação única) a aplicações internas e externas, eliminando a necessidade de os usuários se lembrarem de várias credenciais. Seu provedor de identidade é integrado aos seus sistemas de recursos humanos (RH) para que as mudanças de pessoal sejam automaticamente sincronizadas com seu provedor de identidade. Por exemplo, se alguém deixar sua organização, você poderá revogar automaticamente o acesso a aplicações e sistemas federados (inclusive a AWS). Você habilitou o registro em log detalhado de auditoria em seu provedor de identidade e está monitorando esses logs em busca de comportamentos incomuns do usuário.

Práticas comuns que devem ser evitadas:

- Você não usa federação e autenticação única. Os usuários da sua força de trabalho criam contas de usuário e credenciais separadas em várias aplicações e sistemas.
- Você não automatizou o ciclo de vida das identidades dos usuários da força de trabalho, por exemplo, integrando seu provedor de identidade aos seus sistemas de RH. Quando um usuário deixa sua organização ou muda de função, você segue um processo manual para excluir ou atualizar seus registros em várias aplicações e sistemas.

Benefícios de implementar esta prática recomendada: ao usar um provedor de identidades centralizado, você tem um único local para gerenciar as identidades e políticas dos usuários da força de trabalho, a capacidade de atribuir acesso às aplicações a usuários e grupos e a capacidade de monitorar a atividade de login do usuário. Ao se integrar aos seus sistemas de recursos humanos (RH), quando um usuário muda de função, essas alterações são sincronizadas com o provedor de identidade e atualizam automaticamente as aplicações e permissões atribuídas. Quando um usuário

sai da sua organização, sua identidade é automaticamente desativada no provedor de identidade, revogando seu acesso a aplicações e sistemas federados.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Orientação para usuários da força de trabalho que acessam a AWS

Usuários da força de trabalho em sua organização, como funcionários e prestadores de serviços, podem precisar acessar a AWS usando o AWS Management Console ou a AWS Command Line Interface (AWS CLI) para realizar suas funções de trabalho. Você pode conceder acesso à AWS aos usuários da sua força de trabalho federando a partir de seu provedor de identidade centralizado para a AWS em dois níveis: federação direta para cada Conta da AWS ou federação para várias contas em sua organização da AWS.

- Para federar os usuários da sua força de trabalho diretamente com cada Conta da AWS, é possível usar um provedor de identidade centralizado para federar o AWS Identity and Access Management na conta em questão. A flexibilidade do IAM permite que você habilite um SAML 2.0 ou um provedor de identidade Open ID Connect (OIDC) separado para cada Conta da AWS e atributos de usuário federados para controle de acesso. Os usuários da sua força de trabalho usarão o navegador da Web para fazer login no provedor de identidade fornecendo suas respectivas credenciais (como senhas e códigos de token MFA). O provedor de identidade emite uma declaração SAML para o navegador, que é enviada ao URL de login do AWS Management Console para permitir que o usuário faça autenticação única no AWS Management Console assumindo um perfil do IAM. Seus usuários também podem obter credenciais da API da AWS temporárias para uso no AWS CLI ou AWS SDKs do AWS STS assumindo o perfil do IAM usando uma declaração SAML do provedor de identidade.
- Para federar os usuários da sua força de trabalho com várias contas em sua organização da AWS, é possível usar o AWS IAM Identity Center para gerenciar centralmente o acesso dos usuários a Contas da AWS e aplicações. Você ativa o Identity Center para sua organização e configura sua fonte de identidade. O IAM Identity Center fornece um diretório de origem de identidade padrão que você pode usar para gerenciar seus usuários e grupos. Como alternativa, você pode escolher uma fonte de identidade externa conectando-se ao seu provedor de identidade externo usando SAML 2.0 e provisionando automaticamente usuários e grupos usando o SCIM ou conectando-se ao seu Microsoft AD Directory usando o AWS Directory Service. Depois que uma fonte de identidade é configurada, você pode atribuir acesso a usuários e grupos a Contas da AWS definindo políticas de privilégios mínimos em seus conjuntos de permissões. Os usuários da sua

força de trabalho podem se autenticar por meio de seu provedor de identidade central para entrar no portal de acesso da AWS e fazer login único nas aplicações em nuvem atribuídas a Contas da AWS eles. Este tópico descreve como configurar a AWS CLI v2 para autenticar o usuário com o Centro de Identidade e obter credenciais para executar comandos da AWS CLI. O Centro de Identidade também permite acesso com login único a aplicações da AWS como os portais do Amazon SageMaker Studio e os portais do AWS IoT Sitewise Monitor.

Depois de seguir as orientações anteriores, os usuários da sua força de trabalho não precisarão mais utilizar usuários e grupos do IAM para operações normais ao gerenciar workloads na AWS. Em vez disso, seus usuários e grupos serão gerenciados fora da AWS e os usuários poderão acessar recursos da AWS como identidade federada. As identidades federadas usam os grupos definidos pelo seu provedor de identidade centralizado. Você deve identificar e remover grupos do IAM, usuários do IAM e credenciais de usuário de longa duração (senhas e chaves de acesso) que não são mais necessárias nas suas Contas da AWS. Você pode encontrar credenciais não utilizadas usando relatórios de credenciais do IAM, excluir os usuários do IAM correspondentes e excluir grupos do IAM. Você pode aplicar uma Política de controle de serviços (SCP) à sua organização que ajuda a impedir a criação de novos usuários e grupos do IAM, impondo esse acesso à AWS por meio de identidades federadas.

Orientação para usuários das suas aplicações

Você pode gerenciar as identidades dos usuários das suas aplicações, como uma aplicação móvel, usando o Amazon Cognito como seu provedor de identidade centralizado. O Amazon Cognito habilita a autenticação, autorização e gerenciamento de usuários para suas aplicações Web e móveis. O Amazon Cognito fornece um armazenamento de identidades que pode ser escalado para milhões de usuários, oferece suporte à federação de identidades sociais e corporativas e oferece recursos avançados de segurança para ajudar a proteger seus usuários e negócios. É possível integrar sua aplicação Web ou móvel personalizada ao Amazon Cognito para adicionar autenticação de usuário e controle de acesso a suas aplicações em minutos. Desenvolvido com base em padrões de identidade abertos, como SAML e Open ID Connect (OIDC), o Amazon Cognito oferece suporte a vários regulamentos de conformidade e se integra aos recursos de desenvolvimento de frontend e backend.

Etapas de implementação

Etapas para usuários da força de trabalho acessarem a AWS

- Federe os usuários da sua força de trabalho à AWS usando um provedor de identidade centralizado de acordo com uma das seguintes abordagens:
  - Use o Centro de Identidade do IAM para habilitar a autenticação única para várias Contas da AWS em sua organização da AWS via federação com seu provedor de identidade.
  - Use o IAM para conectar seu provedor de identidade diretamente a cada Conta da AWS, permitindo acesso federado refinado.
- Identifique e remova usuários e grupos do IAM que são substituídos por identidades federadas.

Etapas para usuários das suas aplicações

- Use o Amazon Cognito como um provedor de identidades centralizado para suas aplicações.
- Integre suas aplicações personalizadas com o Amazon Cognito usando o OpenID Connect e o OAuth. Você pode desenvolver suas aplicações personalizadas usando as bibliotecas do Amplify que fornecem interfaces simples para integração com uma variedade de serviços da AWS, como o Amazon Cognito para autenticação.

Recursos

Práticas recomendadas do Well-Architected relacionadas:

- SEC02-BP06 Utilizar grupos de usuários e atributos
- SEC03-BP02 Conceder acesso de privilégio mínimo
- SEC03-BP06 Gerenciar o acesso com base no ciclo de vida

Documentos relacionados:

- Federação de identidades na AWS
- Práticas recomendadas de segurança no IAM
- Práticas recomendadas do AWS Identity and Access Management
- Conceitos básicos da administração delegada no Centro de Identidade do IAM
- Como usar políticas gerenciadas pelo cliente no Centro de Identidade do IAM para casos de uso avançados
- AWS CLI v2: fornecedor de credenciais do Centro de Identidade do IAM

Vídeos relacionados:

- AWS re:Inforce 2022: Mergulho profundo no AWS Identity and Access Management (IAM)
- AWS re:Invent 2022: Simplificar o acesso da sua força de trabalho com o Centro de Identidade do IAM
- AWS re:Invent 2018: Dominar a identidade em todos os aspectos

Exemplos relacionados:

- Workshop: Usar o AWS IAM Identity Center para conseguir um gerenciamento de identidade forte
- Workshop: Identidade sem servidor

Ferramentas relacionadas:

- Parceiros de competência Segurança da AWS: gerenciamento de identidade e acesso
- AWS IAM Identity Center

SEC02-BP05 Auditar e fazer a rotação das credenciais periodicamente

Audite e faça a rotação das credenciais periodicamente para limitar o período durante o qual as credenciais podem ser usadas para acessar seus recursos. Credenciais de longo prazo criam muitos riscos, e estes podem ser reduzidos por meio da rotação periódica das credenciais de longo prazo.

Resultado desejado: implemente a rotação de credenciais para ajudar a reduzir os riscos associados ao uso de credenciais a longo prazo. Audite e corrija regularmente a não conformidade com políticas de rotação de credenciais.

Práticas comuns que devem ser evitadas:

- Não auditar o uso de credenciais.
- Utilizar credenciais de longo prazo desnecessariamente.
- Utilizar credenciais de longo prazo e não fazer sua rotação regularmente.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Quando você não puder contar com credenciais temporárias e exigir credenciais de longo prazo, faça uma auditoria das credenciais para garantir que os controles definidos, por exemplo, autenticação multifator (MFA), sejam aplicados, sofram rotação periódica e tenham o nível de acesso apropriado.

A validação periódica, preferencialmente por meio de uma ferramenta automatizada, é necessária para verificar se os controles corretos são aplicados. Para identidades humanas, exija que os usuários alterem suas senhas periodicamente e substituam chaves de acesso por credenciais temporárias. Ao migrar de usuários do AWS Identity and Access Management (IAM) para identidades centralizadas, você pode gerar um relatório de credenciais para auditar seus usuários.

Também recomendamos implementar e monitorar a MFA no provedor de identidades. É possível configurar o Regras do AWS Config ou usar padrões de segurança do AWS Security Hub para monitorar se os usuários configuraram a MFA. Considere utilizar o IAM Roles Anywhere para fornecer credenciais temporárias para identidades de máquina. Em situações em que o uso de perfis do IAM e credenciais temporárias não é possível, é necessário realizar auditoria frequente e fazer a rotação das chaves de acesso.

Etapas de implementação

- Audite as credenciais periodicamente: a auditoria das identidades configuradas em seu provedor de identidades e no IAM ajuda a garantir que somente identidades autorizadas tenham acesso à sua workload. Essas identidades podem incluir, entre outros, usuários do IAM, do AWS IAM Identity Center, do Active Directory ou usuários em um provedor de identidades upstream diferente. Por exemplo, remova as pessoas que saem da organização os perfis entre contas que não são mais necessários. Estabeleça um processo para auditar periodicamente as permissões para os serviços acessados por uma entidade do IAM. Isso ajuda a identificar as políticas que você precisa modificar a fim de remover todas as permissões não utilizadas. Use relatórios de credenciais e o AWS Identity and Access Management Access Analyzer para auditar credenciais e permissões do IAM. Você pode usar o Amazon CloudWatch para configurar alarmes para chamadas de API específicas chamadas dentro do seu ambiente. AWS O Amazon GuardDuty também pode alertar você sobre atividades inesperadas, o que pode indicar acesso excessivamente permissivo ou acesso não intencional às credenciais do IAM.
- Faça a rotação das credenciais regularmente: quando não conseguir usar credenciais temporárias, faça a rotação das chaves de acesso do IAM de longo prazo regularmente (máximo a cada 90 dias). Se uma chave de acesso for divulgada acidentalmente sem seu conhecimento, isso limitará

o período de uso das credenciais para acessar seus recursos. Para obter mais informações sobre a rotação de chaves de acesso para usuários do IAM, consulte Fazer a rotação das chave de acesso.

- Revise suas permissões do IAM: para melhorar a segurança da sua conta da Conta da AWS, você deve revisar e monitorar regularmente cada uma de suas políticas do IAM. Verifique se as políticas seguem o princípio de privilégio mínimo.
- Considere automatizar a criação e as atualizações de recursos do IAM: o IAM Identity Center automatiza muitas tarefas do IAM, como gerenciamento de perfis e políticas. Como alternativa, o AWS CloudFormation pode ser usado para automatizar a implantação de recursos do IAM, como perfis e políticas, para reduzir a chance de erros humanos, pois os modelos podem ser verificados e ter controle de versão.
- Use o IAM Roles Anywhere para substituir usuários do IAM por identidades de máquina: o IAM Roles Anywhere permite que você use perfis em áreas que você tradicionalmente não poderia, como servidores locais. O IAM Roles Anywhere utiliza um certificado X.509 confiável para realizar a autenticação na AWS e receber credenciais temporárias. O uso do IAM Roles Anywhere evita a necessidade de fazer a rotação dessas credenciais, pois credenciais de longo prazo não são mais armazenadas em seu ambiente on-premises. Você precisará monitorar e fazer a rotação do certificado X.509 à medida que ele se aproxima da validade.

Recursos

Práticas recomendadas relacionadas:

- SEC02-BP02 Usar credenciais temporárias
- SEC02-BP03 Armazenar e usar segredos com segurança

Documentos relacionados:

- Conceitos básicos do AWS Secrets Manager
- Práticas recomendadas do IAM
- Provedores de identidade e federação
- Soluções para parceiros de segurança: acesso e controle
- Credenciais de segurança temporárias
- Obter relatórios de credenciais da sua Conta da AWS

Vídeos relacionados:

- Práticas recomendadas para gerenciar, recuperar e fazer a rotação de segredos em escala
- Gerenciar permissões de usuário em escala com o AWS IAM Identity Center
- Como dominar a identidade em cada camada do bolo

Exemplos relacionados:

- Laboratório do Well-Architected: Limpeza automatizada de usuários do IAM
- Laboratório do Well-Architected: Implantação automatizada de grupos e perfis do IAM

SEC02-BP06 Utilizar grupos de usuários e atributos

A definição de permissões de acordo com grupos de usuários e atributos ajuda a reduzir o número e a complexidade das políticas, simplificando o cumprimento do princípio do privilégio mínimo. Você pode usar grupos de usuários para gerenciar permissões para várias pessoas em um só lugar com base na função que elas desempenham em sua organização. Os atributos, como departamento ou localização, podem ampliar o escopo de permissão quando as pessoas realizam uma função que, embora semelhante, envolve diferentes subconjuntos de recursos.

Resultado desejado: é possível aplicar alterações nas permissões com base na função a todos os usuários que executam essa função. A associação a grupos e os atributos governam as permissões de usuário, reduzindo a necessidade de gerenciar permissões para cada usuário. Os grupos e atributos que você define em seu provedor de identidades (IdP) são propagados automaticamente para seus ambientes da AWS.

Práticas comuns que devem ser evitadas:

- Gerenciar permissões para usuários individuais e duplicá-las entre vários usuários.
- Definir grupos em um nível muito alto, concedendo permissões excessivamente amplas.
- Definir grupos em um nível muito detalhado, criando duplicação e confusão em termos de associação.
- Usar grupos com permissões duplicadas em subconjuntos de recursos quando, em vez disso, é possível usar atributos.
- Não gerenciar grupos, atributos e associações por meio de um provedor de identidades padronizado integrado aos seus ambientes da AWS.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

As permissões da AWS são definidas em documentos chamados de políticas, os quais são associados a uma entidade principal, como usuário, grupo, perfil ou recurso. Para sua força de trabalho, isso permite definir grupos com base na função desempenhada pelos usuários dentro da organização, e não nos recursos que estão sendo acessados. Por exemplo, um grupo `WebAppDeveloper` pode ter uma política anexada para configurar um serviço como o Amazon CloudFront em uma conta de desenvolvimento. Um grupo `AutomationDeveloper` pode ter algumas permissões do CloudFront em comum com o grupo `WebAppDeveloper`. Essas permissões podem ser capturadas em uma política separada e associadas aos dois grupos, em vez de fazer com que os usuários de ambas as funções pertençam a um grupo `CloudFrontAccess`.

Além dos grupos, você pode usar atributos para controlar melhor o escopo do acesso. Por exemplo, você pode ter um atributo `Project` para os usuários do seu grupo `WebAppDeveloper` para definir o escopo do acesso a recursos específicos do projeto. O uso dessa técnica elimina a necessidade de ter grupos diferentes para desenvolvedores de aplicações que estão trabalhando em diferentes projetos se, em outras circunstâncias, as permissões deles forem as mesmas. A forma como você se refere aos atributos nas políticas de permissão baseia-se na respectiva origem, sejam eles definidos como parte do seu protocolo de federação (por ex., SAML, OIDC ou SCIM) ou como declarações SAML personalizadas, ou definidos dentro do Centro de Identidade do IAM.

Etapas de implementação

1. Estabeleça onde você definirá grupos e atributos.
   a. Seguindo as orientações em SEC02-BP04 Confiar em um provedor de identidades centralizado, é possível determinar se há necessidade de definir grupos e atributos no seu provedor de identidade, no Centro de Identidade do IAM, ou usar grupos de usuários do IAM em uma conta específica.
2. Defina grupos.
   a. Determine seus grupos com base na função e no escopo de acesso necessário.
   b. Se estiver definindo no Centro de Identidade do IAM, crie grupos e associe o nível de acesso desejado usando conjuntos de permissões.
   c. Se estiver definindo em um provedor de identidades externo, determine se o provedor atende ao protocolo SCIM e considere habilitar o provisionamento automático no Centro de Identidade do IAM. Esse recurso sincroniza a criação, associação e exclusão de grupos entre seu provedor e o Centro de Identidade do IAM.

3. Defina atributos.
    a. Se estiver usando um provedor de identidades externo, os protocolos SCIM e SAML 2.0 fornecem determinados atributos por padrão. Atributos adicionais podem ser definidos e transmitidos por meio de declarações SAML usando o nome do atributo `https://aws.amazon.com/SAML/Attributes/PrincipalTag`.
    b. Se estiver definindo no Centro de Identidade do IAM, habilite o recurso de controle de acesso por atributo (ABAC) e defina os atributos conforme desejado.
4. Defina o escopo das permissões com base em grupos e atributos.
    a. Considere incluir condições em suas políticas de permissão que comparem os atributos da entidade principal aos atributos dos recursos que estão sendo acessados. Por exemplo, é possível definir uma condição para permitir o acesso a um recurso somente se o valor de uma chave de condição `PrincipalTag` corresponder ao valor de uma chave `ResourceTag` com o mesmo nome.

Recursos

Práticas recomendadas relacionadas:

- SEC02-BP04 Confiar em um provedor de identidades centralizado
- SEC03-BP02 Conceder acesso de privilégio mínimo
- COST02-BP04 Implementar grupos e perfis

Documentos relacionados:

- Práticas recomendadas do IAM
- Gerenciar identidades no Centro de Identidade do IAM
- O que é ABAC para AWS?
- ABAC no Centro de Identidade do IAM

Vídeos relacionados:

- Gerenciar permissões de usuário em grande escala com o Centro de Identidade do AWS IAM
- Como dominar a identidade em cada camada do bolo

SEC 3. Como você gerencia as permissões para pessoas e máquinas?

Gerencie permissões para controlar o acesso a identidades de pessoas e máquinas que precisam de acesso à AWS e à sua workload. As permissões controlam quem pode acessar o quê e em quais condições.

Práticas recomendadas

- SEC03-BP01 Definir requisitos de acesso
- SEC03-BP02 Conceder acesso de privilégio mínimo
- SEC03-BP03 Estabelecer processo de acesso de emergência
- SEC03-BP04 Reduzir as permissões continuamente
- SEC03-BP05 Definir barreiras de proteção de permissões para sua organização
- SEC03-BP06 Gerenciar o acesso com base no ciclo de vida
- SEC03-BP07 Analisar o acesso público e entre contas
- SEC03-BP08 Compartilhar recursos com segurança em sua organização
- SEC03-BP09 Compartilhar recursos com terceiros de forma segura

SEC03-BP01 Definir requisitos de acesso

Cada componente ou recurso de seu workload precisa ser acessado por administradores, usuários finais ou outros componentes. Tenha uma definição clara de quem ou o que deve ter acesso a cada componente, escolha o tipo de identidade e o método de autenticação e autorização apropriados.

Práticas comuns que devem ser evitadas:

- Codificação rígida ou armazenamento de segredos em sua aplicação.
- Concessão de permissões personalizadas a cada usuário.
- Uso de credenciais de longa duração.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Cada componente ou recurso de seu workload precisa ser acessado por administradores, usuários finais ou outros componentes. Tenha uma definição clara de quem ou o que deve ter acesso a cada componente, escolha o tipo de identidade e o método de autenticação e autorização apropriados.

O acesso regular a Contas da AWS dentro da organização deve ser fornecido usando acesso federado ou um provedor de identidade centralizado. Você também deve centralizar o gerenciamento de identidade e garantir que haja uma prática estabelecida para integrar o acesso à AWS ao ciclo de vida de acesso dos funcionários. Por exemplo, quando um funcionário muda para um cargo com um nível de acesso diferente, sua associação ao grupo também deve mudar para refletir os novos requisitos de acesso.

Ao definir os requisitos de acesso para identidades não humanas, determine quais aplicações e componentes precisam de acesso e como as permissões são concedidas. O uso de perfis do IAM criados com o modelo de acesso de privilégio mínimo é uma abordagem recomendada. AWS As políticas gerenciadas fornecem políticas do IAM predefinidas que abordam a maioria dos casos de uso comuns.

Serviços da AWS, como AWS Secrets Manager e o AWS Systems Manager Parameter Store, podem ajudar a desacoplar os segredos da aplicação ou workload com segurança em casos em que não é viável usar perfis do IAM. No Secrets Manager, é possível estabelecer uma rotação automática das suas credenciais. É possível usar o Systems Manager para referenciar parâmetros em seus scripts, comandos, documentos do SSM, configurações e fluxos de trabalho de automação usando o nome exclusivo que você especificou ao criar o parâmetro.

É possível usar o AWS Identity and Access Management Roles Anywhere para obter credenciais de segurança temporárias no IAM para workloads executadas fora da AWS. Suas workloads podem usar as mesmas políticas IAM e os mesmos perfis do IAM que você usa com aplicações da AWS para acessar recursos da AWS.

Quando possível, prefira credenciais temporárias de curta duração em vez de credenciais estáticas de longa duração. Para cenários em que você precisa de usuários do com acesso programático e credenciais de longo prazo, use as informações de última utilização da chave de acesso para fazer a rotação e remover chaves de acesso.

Os usuários precisam de acesso programático se quiserem interagir com a AWS de fora do AWS Management Console. A forma de conceder acesso programático depende do tipo de usuário que está acessando a AWS.

Para conceder acesso programático aos usuários, selecione uma das seguintes opções:

Qual usuário precisa de acesso programático?	Para	Por
Identificação da força de trabalho (Usuários gerenciados no Centro de Identidade do IAM)	Use credenciais temporárias para assinar solicitações programáticas para a AWS CLI, os SDKs da AWS ou as APIs da AWS.	Siga as instruções da interface que deseja utilizar. • Para a AWS CLI, consulte Configuração da AWS CLI para usar o AWS IAM Identity Center no Guia do usuário da AWS Command Line Interface. • Para os SDKs da AWS, ferramentas e APIs da AWS, consulte Autenticação do Centro de Identidade do IAM no Guia de referência de ferramentas e SDKs da AWS.
IAM	Use credenciais temporárias para assinar solicitações programáticas para a AWS CLI, os SDKs da AWS ou as APIs da AWS.	Siga as instruções em Como usar credenciais temporárias com recursos da AWS no Guia do usuário do IAM.
IAM	(Não recomendado) Use credenciais de longo prazo para assinar solicitações programáticas para a AWS CLI, os SDKs da AWS ou as APIs da AWS.	Siga as instruções da interface que deseja utilizar. • Para a AWS CLI, consulte Autenticação usando as credenciais de usuário do IAM no Guia do usuário da AWS Command Line Interface. • Para as ferramentas e SDKs da AWS, consulte

Qual usuário precisa de acesso programático?	Para	Por
		Autenticação usando as credenciais de longo prazo no Guia de referência de ferramentas e SDKs da AWS. • Para as APIs da AWS, consulte Gerenciamento de chaves de acesso de usuários do IAM no Guia do usuário do IAM.

Recursos

Documentos relacionados:

- Controle de acesso por atributo (ABAC)
- AWS IAM Identity Center
- IAM Roles Anywhere
- Políticas gerenciadas pela AWS para o IAM Identity Center
- Condições de política do AWS IAM
- Casos de uso do IAM
- Remover credenciais desnecessárias
- Trabalhar com políticas do
- Como controlar o acesso a recursos da AWS com base em Conta da AWS, UO ou organização
- Identificar, organizar e gerenciar segredos facilmente usando a pesquisa avançada no AWS Secrets Manager

Vídeos relacionados:

- Torne-se um mestre e políticas do IAM em no máximo 60 minutos
- Separação de deveres, privilégio mínimo, delegação e CI/CD

- Simplificação do gerenciamento de identidade e acesso para inovação

SEC03-BP02 Conceder acesso de privilégio mínimo

É prática recomendada conceder somente o acesso de que as identidades precisam para realizar ações em recursos específicos e sob condições específicas. Use grupos e atributos de identidade para definir permissões dinamicamente em escala, em vez de definir permissões para usuários individuais. Por exemplo, é possível permitir o acesso de um grupo de desenvolvedores para gerenciar apenas recursos de seu próprio projeto. Dessa forma, se um desenvolvedor sair do projeto, seu acesso será automaticamente revogado sem que seja necessário alterar as políticas de acesso adjacentes.

Resultado desejado: os usuários devem ter apenas as permissões necessárias para realizar suas tarefas. Os usuários devem ter acesso apenas a ambientes de produção para realizar uma tarefa específica dentro de um período limitado e o acesso deve ser revogado quando a tarefa é concluída. As permissões devem ser revogadas quando não forem mais necessárias, incluindo quando um usuário for atribuído a um projeto diferente ou mudar de cargo. Os privilégios de administrador devem ser concedidos apenas a um grupo pequeno de administradores confiáveis. As permissões devem ser revistas regularmente para evitar desvios de permissão. Contas de máquina ou sistema devem ter apenas o mínimo de permissões necessárias para concluir suas tarefas.

Práticas comuns que devem ser evitadas:

- Usar como padrão a concessão de permissões de administrador aos usuários.
- Usar o usuário-raiz para atividades diárias.
- Criar políticas permissivas demais, mas sem privilégios completos de administrador.
- Não revisar as permissões para entender se elas permitem o acesso de privilégio mínimo.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

O princípio do privilégio mínimo afirma que as identidades só devem ter permissão para realizar o menor conjunto de ações necessárias para cumprir uma tarefa específica. Isso equilibra a usabilidade, eficiência e segurança. Operar sobre esse princípio ajuda a limitar acesso não intencional e a rastrear quem tem acesso a quais recursos. Usuários e perfis do IAM não têm permissões por padrão. O usuário-raiz tem acesso total por padrão e deve ser rigorosamente controlado, monitorado e usado somente para tarefas que exijam acesso de usuário-raiz.

Políticas do IAM são usadas para conceder explicitamente permissões aos perfis do IAM ou recursos específicos. Por exemplo, políticas com base em identidade podem ser anexadas a grupos do IAM, enquanto buckets do S3 podem ser controlados por políticas baseadas em recursos.

Ao criar e associar uma política do IAM, você pode especificar as ações de serviço, os recursos e as condições que devem ser verdadeiras para que a AWS permita ou negue o acesso. A AWS oferece suporte a uma variedade de condições para ajudar você a reduzir o acesso. Por exemplo, usando a chave de condição `PrincipalOrgID`, você poderá negar ações se o solicitante não fizer parte da sua organização da AWS.

Você também pode controlar as solicitações feitas pelos serviços da AWS em seu nome, como o AWS CloudFormation criando uma função do AWS Lambda usando a chave de condição `CalledVia`. Aplique camadas de tipos diferentes de políticas para estabelecer a defesa em profundidade e limitar as permissões gerais de seus usuários. É possível restringir as permissões que podem ser concedidas e sob quais condições. Por exemplo, você pode permitir que suas equipes de aplicações criem suas próprias políticas do IAM para os sistemas que elas criam, mas também deve aplicar um limite de permissão para limitar o máximo de permissões que o sistema pode receber.

Etapas de implementação

- Implementar políticas de privilégio mínimo: atribua políticas de acesso com privilégio mínimo a grupos e perfis do IAM para refletir a função do usuário ou a função que você definiu.
  - Baseie as políticas no uso da API: uma forma de determinar as permissões necessárias é revisar os logs da AWS CloudTrail. Essa revisão permite que você crie permissões personalizadas para as ações do usuário dentro da AWS. O IAM Access Analyzer pode gerar automaticamente uma política do IAM baseada na atividade. É possível usar o IAM Access Advisor em nível de organização ou conta para rastrear as últimas informações acessadas para uma política específica.
- Considere o uso de políticas gerenciadas pela AWS para funções de trabalho. Pode ser difícil saber por onde começar ao criar políticas de permissões mais estritas. A AWS gerencia políticas para cargos comuns, como faturamento, administradores de banco de dados e cientistas de dados. Essas políticas podem ajudar a diminuir o acesso dos usuários ao determinar como implementar as políticas de privilégio mínimo.
- Remova permissões desnecessárias: remova as permissões que não são necessárias e reduza as políticas excessivamente permissivas. A geração de políticas do IAM Access Analyzer pode ajudar a ajustar as políticas de permissões.

- Garanta que os usuários tenham acesso limitado aos ambientes de produção: os usuários só devem ter acesso aos ambientes de produção com um caso de uso válido. Depois de o usuário realizar as tarefas específicas para as quais o acesso à produção foi necessário, o acesso deve ser revogado. Limitar o acesso a ambientes de produção ajuda a prevenir eventos não intencionais e que causam impacto à produção, além de diminuir o escopo do impacto do acesso não intencional.
- Considere usar limites de permissões: um limite de permissões é um recurso avançado para usar uma política gerenciada que define o número máximo de permissões que uma política baseada em identidade pode conceder a uma entidade do IAM. O limite de permissões de uma entidade permite que a entidade execute somente as ações permitidas por ambas as políticas baseadas em identidade e seus limites de permissões.
- Considere usar tags de recursos para obter permissões: um modelo de controle de acesso baseado em atributos que usa tags de recursos permite conceder acesso com base na finalidade, no proprietário, no ambiente ou em outros critérios do recurso. Por exemplo, você pode usar tags de recurso para diferenciar entre ambientes de desenvolvimento e produção. Ao usar essas tags, é possível restringir os desenvolvedores ao ambiente de desenvolvimento. Ao combinar tags e políticas de permissões, é possível alcançar um acesso restrito ao recurso sem precisar definir políticas complicadas e personalizadas para cada cargo.
- Use políticas de controle de serviços para o AWS Organizations. As políticas de controle de serviço controlam centralmente o máximo de permissões disponíveis para contas-membro em sua organização. É importante notar que as políticas de controle de serviço permitem restringir as permissões do usuário-raiz nas contas-membro. Considere também usar o AWS Control Tower, que fornece controles gerenciados prescritivos que enriquecem o AWS Organizations. Também é possível definir os seus próprios controles no Control Tower.
- Estabeleça uma política de ciclo de vida do usuário para sua organização: as políticas de ciclo de vida do usuário definem tarefas a serem executadas quando os usuários são integrados à AWS, mudam de função ou escopo de trabalho ou não precisam mais de acesso à AWS. As análises de permissões devem ser feitas durante todas as etapas do ciclo de vida do usuário para verificar se as permissões estão adequadamente restritas e para evitar desvios nas permissões.
- Estabeleça um cronograma regular para revisar as permissões e remover todas as permissões desnecessárias: revise regularmente o acesso dos usuários para verificar se os usuários não têm acesso excessivamente permissivo. O AWS Config e o IAM Access Analyzer podem ajudar na auditoria das permissões do usuário.
- Estabeleça uma matriz de funções de trabalho: uma matriz de funções de trabalho visualiza as várias funções e níveis de acesso necessários em sua presença da AWS. Com uma matriz de

cargos, você pode definir e separar as permissões com base nas responsabilidades do usuário dentro da sua organização. Use grupos em vez de aplicar permissões diretamente a usuários ou funções individuais.

Recursos

Documentos relacionados:

- Conceder privilégio mínimo
- Limites de permissões para entidades do IAM
- Técnicas para criar políticas do IAM de privilégio mínimo
- O IAM Access Analyzer facilita a implementação de permissões de privilégio mínimo ao gerar políticas do IAM com base na atividade de acesso
- Delegar o gerenciamento de permissões aos desenvolvedores usando os limites de permissões do IAM
- Refinar permissões usando as informações de último acesso
- Tipos de política do IAM e quando usá-las
- Testar as políticas do IAM com o simulador de políticas do IAM
- Barreiras de proteção no AWS Control Tower
- Arquiteturas de confiança zero: uma perspectiva da AWS
- Como implementar o princípio de privilégio mínimo com o CloudFormation StackSets
- Controle de acesso por atributo (ABAC)
- Reduzir o escopo da política pela visualização das atividades do usuário
- Visualizar acesso do perfil
- Usar a marcação com tags para organizar seu ambiente e impulsionar a responsabilidade
- Estratégias de marcação com tags da AWS
- Marcando recursos do AWS

Vídeos relacionados:

- Gerenciamento de permissões de última geração
- Confiança zero: uma perspectiva da AWS

Exemplos relacionados:

- Laboratório: limites de permissões do IAM que delegam a criação de perfis
- Laboratório: Controle de acesso por tags do IAM para EC2

SEC03-BP03 Estabelecer processo de acesso de emergência

Crie um processo que permita acesso emergencial às suas workloads no caso improvável de um problema com seu provedor de identidades centralizado.

Crie processos para diferentes modos de falha que poderiam resultar em um evento de emergência. Por exemplo, em circunstâncias normais, os usuários da sua força de trabalho são federados na nuvem usando um provedor de identidades centralizado (SEC02-BP04) para gerenciar suas workloads. No entanto, se o provedor de identidades centralizado falhar ou a configuração da federação na nuvem for modificada, talvez os usuários de sua força de trabalho não consigam se federar na nuvem. Um processo de acesso de emergência permite que administradores autorizados acessem seus recursos de nuvem por meios alternativos (como uma forma alternativa de federação ou acesso direto do usuário) para corrigir problemas com sua configuração de federação ou workloads. O processo de acesso de emergência é usado até o mecanismo normal de federação ser restaurado.

Resultado desejado:

- Você definiu e documentou os modos de falha que são considerados uma emergência: considere suas circunstâncias normais e os sistemas dos quais seus usuários dependem para gerenciar suas workloads. Pense em como cada uma dessas dependências pode falhar e causar uma situação de emergência. Talvez você considere as perguntas e as práticas recomendadas do pilar Confiabilidade úteis para identificar modos de falha e arquitetar sistemas mais resilientes para minimizar a probabilidade de falhas.
- Você documentou as etapas que devem ser seguidas para confirmar uma falha como emergência. Por exemplo, é possível exigir que os administradores de identidade confiram o status de seus provedores de identidade primário e de reserva e, se nenhum dos dois estiver disponível, declarar um evento de emergência por falha do provedor de identidades.
- Você definiu um processo de acesso de emergência específico de cada tipo de modo de emergência ou falha. Ser específico pode reduzir a tentação de seus usuários de abusar de um processo geral para todos os tipos de emergência. Seus processos de acesso de emergência descrevem as circunstâncias em que cada processo deve ser usado e, inversamente, as situações

em que o processo não deve ser usado e apontam para processos alternativos que podem ser aplicados.

- Seus processos são bem documentados com instruções detalhadas e playbooks que podem ser seguidos com rapidez e eficiência. Lembre-se de que um evento de emergência pode ser um momento estressante para os usuários e eles podem estar sob extrema pressão de tempo. Portanto, desenvolva o processo para ser o mais simples possível.

Práticas comuns que devem ser evitadas:

- Você não tem processos de acesso de emergência bem documentados e bem testados. Os usuários não estão preparados para uma emergência e seguem processos improvisados quando um evento de emergência ocorre.
- Seus processos de acesso de emergência dependem dos mesmos sistemas (como um provedor de identidades centralizado) que seus mecanismos de acesso normais. Isso significa que uma falha desse sistema pode afetar os mecanismos de acesso normal e de emergência e prejudicar sua capacidade de se recuperar da falha.
- Seus processos de acesso de emergência são usados em situações não emergenciais. Por exemplo, os usuários muitas vezes utilizam de forma indevida os processos de acesso de emergência, pois acham mais fácil fazer alterações diretamente do que enviá-las por meio de um pipeline.
- Seus processos de acesso de emergência não geram logs suficientes para auditar os processos, ou os logs não são monitorados para alertar sobre o possível uso indevido dos processos.

Benefícios de implementar esta prática recomendada:

- Com processos de acesso de emergência bem documentados e testados, é possível reduzir o tempo gasto pelos usuários para responder e resolver um evento de emergência. Isso pode resultar em menos tempo de inatividade e maior disponibilidade dos serviços fornecidos aos seus clientes.
- É possível rastrear cada solicitação de acesso de emergência e detectar e alertar sobre tentativas não autorizadas de uso indevido do processo para eventos não emergenciais.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Esta seção fornece orientação para criar processos de acesso de emergência para vários modos de falha relacionados às workloads implantadas na AWS, começando com uma orientação comum que se aplica a todos os modos de falha e seguida por uma orientação específica com base no tipo de modo de falha.

Orientação comum para todos os modos de falha

Pense no seguinte ao projetar um processo de acesso de emergência para um modo de falha:

- Documente as pré-condições e as suposições do processo: quando o processo deve ou não ser usado. Isso ajuda a detalhar o modo de falha e documentar suposições, como o estado de outros sistemas relacionados. Por exemplo, o processo do Modo de falha 2 pressupõe que o provedor de identidades está disponível, mas a configuração na AWS foi modificada ou expirou.
- Pré-crie os recursos necessários para o processo de acesso de emergência (SEC10-BP05). Por exemplo, crie previamente a Conta da AWS de acesso de emergência com usuários e perfis do IAM e os perfis do IAM entre contas em todas as contas da workload. Isso verifica se esses recursos estão prontos e disponíveis quando um evento de emergência ocorre. Ao pré-criar recursos, você não depende das APIs do ambiente de gerenciamento da AWS (usadas para criar e modificar recursos da AWS) que podem estar indisponíveis em caso de emergência. Além disso, ao pré-criar recursos do IAM, não é necessário considerar possíveis atrasos devido a consistência eventual.
- Inclua processos de acesso de emergência como parte dos planos de gerenciamento de incidentes (SEC1-BP02). Documente como os eventos de emergência são acompanhados e comunicados a outras pessoas na organização, como equipes de colegas, sua liderança e, quando aplicável, externamente a seus clientes e parceiros de negócios.
- Defina o processo de solicitação de acesso de emergência no sistema de fluxo de trabalho de solicitação de serviço existente, caso haja um. Normalmente, esses sistemas de fluxo de trabalho permitem criar formulários de admissão para coletar informações sobre a solicitação, rastrear a solicitação em cada estágio do fluxo de trabalho e adicionar etapas de aprovação automatizadas e manuais. Relacione cada solicitação a um evento de emergência correspondente acompanhado no sistema de gerenciamento de incidentes. Ter um sistema uniforme para acessos de emergência permite que você acompanhe essas solicitações em um único sistema, analise as tendências de uso e melhore os processos.
- Verifique se os processos de acesso de emergência só podem ser iniciados por usuários autorizados e exigem aprovações dos colegas ou da gerência do usuário, conforme apropriado.

O processo de aprovação deve operar de forma eficaz dentro e fora do horário comercial. Defina como as solicitações de aprovação permitirão aprovadores secundários se os aprovadores primários não estiverem disponíveis e forem encaminhadas para a cadeia de gerenciamento até serem aprovadas.

- Verifique se o processo gera logs e eventos de auditoria detalhados para tentativas bem-sucedidas e fracassadas de obter acesso de emergência. Monitore o processo de solicitação e o mecanismo de acesso de emergência para detectar uso indevido ou acessos não autorizados. Correlacione a atividade com eventos de emergência contínuos do sistema de gerenciamento de incidentes e alerte quando as ações ocorrerem fora dos períodos esperados. Por exemplo, você deve monitorar e alertar sobre atividades na Conta da AWS de acesso de emergência, pois ela nunca deve ser usada em operações normais.
- Teste os processos de acesso de emergência periodicamente para verificar se as etapas estão claras e garantir o nível correto de acesso com rapidez e eficiência. Seus processos de acesso de emergência devem ser testados como parte das simulações de resposta a incidentes (SEC10-BP07) e dos testes de recuperação de desastres (REL13-BP03).

Modo de falha 1: o provedor de identidades usado para federação na AWS não está disponível

Conforme descrito em SEC02-BP-04 Confiar em um provedor de identidade federado, recomendamos confiar em um provedor de identidades centralizado para federar os usuários de sua força de trabalho e conceder acesso a Contas da AWS. Você pode federar em várias Contas da AWS na organização da AWS usando o Centro de Identidade do IAM ou federar em Contas da AWS individuais usando o IAM. Nos dois casos, os usuários da força de trabalho se autenticam com seu provedor de identidades centralizado antes de serem redirecionados a um endpoint de login da AWS para SSO.

No caso improvável do provedor de identidades centralizado não estar disponível, os usuários da sua força de trabalho não poderão se federar nas Contas da AWS nem gerenciar as workloads. Nesse evento de emergência, é possível fornecer um processo de acesso de emergência para um pequeno grupo de administradores acessar as Contas da AWS a fim de realizar tarefas essenciais que não podem esperar até que seus provedores de identidades centralizados estejam online novamente. Por exemplo, seu provedor de identidades fica indisponível por quatro horas e, durante esse período, você precisa modificar os limites superiores de um grupo do Amazon EC2 Auto Scaling em uma conta de produção para lidar com um aumento inesperado no tráfego de clientes. Seus administradores de emergência devem seguir o processo de acesso de emergência a fim de obter acesso à Conta da AWS de produção específica e fazer as alterações necessárias.

O processo de acesso de emergência depende de uma Conta da AWS de acesso de emergência pré-criada usada exclusivamente para acesso de emergência e tem recursos da AWS (como perfis e usuários do IAM) para apoiar o processo de acesso de emergência. Durante as operações normais, ninguém deve acessar a conta de acesso de emergência, e você deve monitorar e alertar sobre o uso indevido dessa conta (para receber mais detalhes, consulte a seção Orientação comum anterior).

A conta de acesso de emergência tem perfis do IAM de acesso de emergência com permissões para assumir perfis entre contas nas Contas da AWS que exigem acesso de emergência. Esses perfis do IAM são pré-criados e configurados com políticas de confiança que confiam nos perfis do IAM da conta de emergência.

O processo de acesso de emergência pode usar uma das seguintes abordagens:

- É possível pré-criar um conjunto de usuários do IAM para seus administradores de emergência na conta de acesso de emergência com senhas fortes e tokens de MFA associados. Esses usuários do IAM têm permissões para assumir os perfis do IAM que permitem o acesso entre contas à Conta da AWS onde o acesso de emergência é necessário. Recomendamos criar o menor número possível de usuários e atribuir cada um a um único administrador de emergência. Durante uma emergência, um usuário administrador de emergência entra na conta de acesso de emergência usando sua senha e código de token MFA, muda para o perfil do IAM de acesso de emergência na conta de emergência e, por fim, para o perfil do IAM de acesso de emergência na conta da workload para realizar a ação de alteração de emergência. A vantagem dessa abordagem é que cada usuário do IAM é atribuído a um administrador de emergência, e é possível saber qual usuário fez login analisando os eventos do CloudTrail. A desvantagem é que você precisa manter vários usuários do IAM com as respectivas senhas de longa duração e tokens de MFA associados.
- É possível usar o usuário-raiz da Conta da AWS de emergência para entrar na conta de acesso de emergência, assumir o perfil do IAM para acesso de emergência e assumir o perfil entre contas na conta da workload. Recomendamos definir uma senha forte e vários tokens de MFA para o usuário-raiz. Também recomendamos armazenar a senha e os tokens de MFA em um cofre de credenciais corporativo seguro que imponha autenticação e autorização fortes. Você deve proteger a senha e os fatores de redefinição de tokens de MFA: defina o endereço de e-mail da conta como uma lista de distribuição de e-mail monitorada pelos administradores de segurança na nuvem e o número de telefone da conta como um número de telefone compartilhado que também seja monitorado pelos administradores de segurança. A vantagem dessa abordagem é que há um conjunto de credenciais de usuário-raiz para gerenciar. A desvantagem é que, como se trata de um usuário compartilhado, vários administradores podem fazer login como usuário-raiz. Você

deve fazer auditoria dos eventos de log do cofre corporativo para identificar qual administrador fez check-out da senha do usuário-raiz.

Modo de falha 2: a configuração do provedor de identidades na AWS foi modificada ou expirou

Para permitir que os usuários de sua força de trabalho sejam federados nas Contas da AWS, é possível configurar o Centro de Identidade do IAM com um provedor de identidades externo ou criar um provedor de identidades do IAM (SEC02-BP04). Normalmente, você os configura importando um documento XML de metadados SAML fornecido pelo provedor de identidades. O documento XML de metadados inclui um certificado X.509 correspondente a uma chave privada que o provedor de identidades usa para assinar as declarações SAML.

Essas configurações no lado da AWS podem ser modificadas ou excluídas por engano por um administrador. Em outro cenário, o certificado X.509 importado para a AWS pode expirar, e um novo XML de metadados com um novo certificado ainda não foi importado para a AWS. Os dois cenários podem interromper a federação na AWS para os usuários de sua força de trabalho, ocasionando uma emergência.

Nesse evento de emergência, você pode fornecer aos seus administradores de identidade acesso à AWS para resolver os problemas de federação. Por exemplo, seu administrador de identidade usa o processo de acesso de emergência para fazer login na Conta da AWS de acesso de emergência, muda para um perfil na conta de administrador do Centro de Identidade e atualiza a configuração do provedor de identidades externo importando o documento XML de metadados SAML mais recente do provedor de identidades para reativar a federação. Após a federação ser corrigida, os usuários da sua força de trabalho continuarão usando o processo operacional normal para federar em suas contas da workload.

Você pode seguir as abordagens detalhadas no Modo de falha 1 anterior para criar um processo de acesso de emergência. É possível conceder permissões de privilégio mínimo aos seus administradores de identidade a fim de acessar somente a conta de administrador do Centro de Identidade e realizar ações no Centro de Identidade nessa conta.

Modo de falha 3: interrupção do Centro de Identidade

No caso improvável de uma interrupção do Centro de Identidade do IAM ou da Região da AWS, recomendamos definir uma configuração que possa ser usada para conceder acesso temporário ao AWS Management Console.

O processo de acesso de emergência usa a federação direta do provedor de identidades no IAM em uma conta de emergência. Para obter detalhes sobre o processo e as considerações de design, consulte Configurar o acesso de emergência ao AWS Management Console.

Etapas de implementação

Etapas comuns para todos os modos de falha

- Crie uma Conta da AWS dedicada aos processos de acesso de emergência. Crie previamente os recursos do IAM necessários na conta, como perfis ou usuários do IAM e, opcionalmente, provedores de identidades do IAM. Além disso, crie previamente perfis do IAM entre contas nas Contas da AWS da workload com relacionamentos de confiança com os perfis do IAM correspondentes na conta de acesso de emergência. É possível usar o AWS CloudFormation StackSets com AWS Organizations para criar esses recursos nas contas-membro da sua organização.

- Crie políticas de controle de serviços (SCP) do AWS Organizations para negar a exclusão e a modificação dos perfis do IAM entre contas nas Contas da AWS-membro.
- Ative o CloudTrail para a Conta da AWS de acesso de emergência e envie os eventos da trilha a um bucket central do S3 em sua Conta da AWS de coleção de logs. Se você estiver usando o AWS Control Tower para configurar e controlar seu ambiente de várias contas da AWS, todas as contas que você criar usando o AWS Control Tower ou inscrever no AWS Control Tower terão o CloudTrail ativado por padrão e serão enviadas a um bucket do S3 em uma Conta da AWS de arquivo de log dedicado.
- Monitore a atividade da conta de acesso de emergência criando regras do EventBridge que correspondam ao login do console e à atividade da API pelos perfis do IAM de emergência. Envie notificações ao seu centro de operações de segurança quando ocorrerem atividades fora de um evento de emergência contínuo acompanhado no sistema de gerenciamento de incidentes.

Etapas adicionais para o Modo de falha 1: o provedor de identidades usado para federar na AWS não está disponível; Modo de falha 2: a configuração do provedor de identidades na AWS foi modificada ou expirou

- Crie previamente recursos de acordo com o mecanismo escolhido para acesso de emergência:
  - Utilizar usuários do IAM: crie previamente os usuários do IAM com senhas fortes e dispositivos MFA associados.

- Utilizar o usuário-raiz da conta de emergência: configure o usuário-raiz com uma senha forte e armazene a senha no seu cofre de credenciais corporativo. Associe vários dispositivos físicos de MFA ao usuário-raiz e armazene os dispositivos em locais que possam ser acessados rapidamente pelos membros de sua equipe de administradores de emergência.

Etapas adicionais para o Modo de falha 3: interrupção do Centro de Identidade

- Conforme detalhado em Configurar o acesso de emergência ao AWS Management Console, na Conta da AWS de acesso de emergência, crie um provedor de identidades do IAM para ativar a federação direta de SAML a partir do provedor de identidades.
- Crie grupos de operações de emergência no IdP sem membros.
- Crie perfis do IAM correspondentes aos grupos de operações de emergência na conta de acesso de emergência.

Recursos

Práticas recomendadas do Well-Architected relacionadas:

- SEC02-BP04 Confiar em um provedor de identidades centralizado
- SEC03-BP02 Conceder acesso de privilégio mínimo
- SEC10-BP02 Desenvolver planos de gerenciamento de incidentes
- SEC10-BP07 Promover game days

Documentos relacionados:

- Configurar o acesso de emergência ao AWS Management Console
- Habilitar o acesso de usuários federados SAML 2.0 ao AWS Management Console
- Acesso de emergência

Vídeos relacionados:

- AWS re:Invent 2022: Simplificar o acesso da sua força de trabalho com o Centro de Identidade do IAM
- AWS re:Inforce 2022: Mergulho profundo no AWS Identity and Access Management (IAM)

Exemplos relacionados:

- Perfil de acesso de emergência da AWS
- Framework do playbook do cliente da AWS
- Exemplos de playbook de resposta a incidentes da AWS

SEC03-BP04 Reduzir as permissões continuamente

À medida que suas equipes determinarem o acesso de que precisam, remova as permissões desnecessárias e estabeleça processos de análise para obter permissões de privilégio mínimo. Monitore e remova continuamente identidades e permissões não utilizadas para acesso humano e de máquina.

Resultado desejado: as políticas de permissão devem seguir o princípio de privilégio mínimo. À medida que os cargos e os perfis se tornem mais bem definidos, suas políticas de permissões precisam ser analisadas para remover permissões desnecessárias. Essa abordagem reduz o escopo do impacto caso as credenciais sejam expostas de forma acidental ou sejam acessadas sem autorização.

Práticas comuns que devem ser evitadas:

- Usar como padrão a concessão de permissões de administrador aos usuários.
- Criar políticas permissivas demais, mas sem privilégios completos de administrador.
- Manter as políticas de permissão quando não são mais necessárias.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Enquanto as equipes e os projetos estiverem apenas começando, políticas de permissão permissivas podem ser usadas para inspirar inovação e agilidade. Por exemplo, em um ambiente de desenvolvimento ou teste, os desenvolvedores podem receber acesso a uma ampla gama de serviços da AWS. Recomendamos avaliar o acesso de forma contínua e restringir o acesso somente àqueles serviços e ações de serviço necessários para concluir o trabalho atual. Recomendamos essa avaliação para identidades humanas e de máquina. Identidades de máquina, às vezes, denominadas contas de sistema ou serviço, são identidades que fornecem acesso da AWS a aplicações ou servidores. Esse acesso é especialmente importante em um ambiente de produção,

em que as permissões excessivamente permissivas podem causar um grande impacto e expor dados dos clientes.

A AWS oferece vários métodos para ajudar a identificar usuários, perfis, permissões e credenciais não utilizados. A AWS também pode ajudar a analisar a atividade de acesso dos usuários e dos perfis do IAM, como chaves de acesso associadas, e o acesso aos recursos da AWS, como objetos em buckets do Amazon S3. A geração de políticas do AWS Identity and Access Management Access Analyzer pode auxiliar você a criar políticas de permissão restritivas com base nos serviços e nas ações reais com os quais uma entidade principal interage. O controle de acesso por atributo (ABAC) pode ajudar a simplificar o gerenciamento de permissões, pois você pode fornecer permissões aos usuários usando seus atributos em vez de anexar políticas de permissões diretamente a cada usuário.

Etapas de implementação

- Use o AWS Identity and Access Management Access Analyzer: o IAM Access Analyzer ajuda a identificar os recursos em sua organização e suas contas, como buckets do Amazon Simple Storage Service (Amazon S3) ou perfis do IAM que são compartilhados com uma entidade externa.
- Use a geração de políticas do IAM Access Analyzer: a geração de políticas do IAM Access Analyzer ajuda você a criar políticas de permissão refinadas com base na atividade de acesso de um usuário ou perfil do IAM.
- Determine um prazo e uma política de uso aceitáveis para usuários e funções do IAM: use o carimbo de data/hora do último acesso para identificar usuários e perfis não utilizados e removê-los. Revise as informações de serviço e ação acessadas mais recentemente para identificar e definir o escopo das permissões para usuários e perfis específicos. Por exemplo, você pode usar as informações acessadas mais recentemente para identificar as ações específicas do Amazon S3 exigidas pelo perfil da aplicação e restringir o acesso do perfil apenas a essas ações. Recursos de informações acessadas mais recentemente estão disponíveis no AWS Management Console e de maneira programática para permitir que você As incorpore aos fluxos de trabalho de infraestrutura e ferramentas automatizadas.
- Considere registrar em log eventos de dados no AWS CloudTrail: por padrão, o CloudTrail não registra eventos de dados em log, como atividades em nível de objeto do Amazon S3 (por exemplo, `GetObject` e `DeleteObject`) ou atividades de tabela do Amazon DynamoDB (por exemplo, `PutItem` e `DeleteItem`). Considere ativar o registro em log desses eventos para determinar quais usuários e perfis precisam acessar objetos do Amazon S3 ou itens de tabelas do DynamoDB específicos.

Recursos

Documentos relacionados:

- Conceder privilégio mínimo
- Remover credenciais desnecessárias
- O que é o AWS CloudTrail?
- Trabalhar com políticas do
- Registrar em log e monitorar no DynamoDB
- Usar o registro em log de eventos do CloudTrail para buckets e objetos do Amazon S3
- Obter relatórios de credenciais da sua Conta da AWS

Vídeos relacionados:

- Torne-se um mestre e políticas do IAM em no máximo 60 minutos
- Separação de deveres, privilégio mínimo, delegação e CI/CD
- AWS re:Inforce 2022: Mergulho profundo no AWS Identity and Access Management (IAM)

SEC03-BP05 Definir barreiras de proteção de permissões para sua organização

Use barreiras de proteção de permissões para reduzir o escopo das permissões disponíveis que podem ser concedidas a entidades principais. A cadeia de avaliação da política de permissões inclui suas grades de proteção para determinar as permissões efetivas de uma entidade principal ao tomar decisões de autorização. Você pode definir barreiras de proteção usando uma abordagem baseada em camadas. Aplique algumas barreiras de proteção de maneira abrangente em toda a organização e outras de forma detalhada às sessões de acesso temporário.

Resultado desejado: você obtém um isolamento claro dos ambientes usando Contas da AWS sparapdpas. As políticas de controle de serviços (SCPs) são usadas para definir barreiras de proteção de permissões em toda a organização. As barreiras de proteção mais amplas são definidas nos níveis hierárquicos mais próximos da raiz da sua organização e as barreiras de proteção mais rígidas são definidas mais perto do nível das contas individuais. Quando aceitas, as políticas de recursos definem as condições que uma entidade principal deve satisfazer para obter acesso a um recurso. As políticas de recursos também abrangem o conjunto de ações permitidas, quando apropriado. Os limites de permissão são impostos às entidades principais que gerenciam as

permissões da workload, delegando o gerenciamento de permissões aos proprietários de workload individuais.

Práticas comuns que devem ser evitadas:

- Criar Contas da AWS-membro dentro de uma organização da AWS, mas não usar SCPs para restringir o uso e as permissões disponíveis para suas credenciais de raiz.
- Atribuir permissões com base em privilégio mínimo, mas não colocar barreiras de proteção no conjunto máximo de permissões que podem ser concedidas.
- Confiar na base de negação implícita do AWS IAM para restringir as permissões, confiando que as políticas não concederão uma permissão explícita indesejada.
- Executar vários ambientes de workload na mesma Conta da AWS e confiar em mecanismos como VPCs, tags ou políticas de recursos para impor limites de permissão.

Benefícios de implementar esta prática recomendada: as barreiras de proteção de permissões ajudam a criar confiança de que permissões indesejadas não podem ser concedidas, mesmo quando uma política de permissão tenta fazer isso. Isso pode simplificar a definição e o gerenciamento de permissões, reduzindo o escopo máximo das permissões que precisam ser consideradas.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Recomendamos usar uma abordagem baseada em camadas para definir barreiras de proteção de permissões para sua organização. Essa abordagem reduz sistematicamente o conjunto máximo de permissões possíveis à medida que camadas adicionais são aplicadas. Isso ajuda você a conceder acesso com base no princípio de privilégio mínimo, reduzindo o risco de acesso indesejado devido à configuração incorreta de alguma política.

A primeira etapa para estabelecer barreiras de proteção de permissões é isolar workloads e ambientes em Contas da AWS separadas. As entidades principais de uma conta não podem acessar recursos em outra conta sem permissão explícita para fazer isso, mesmo quando as duas contas estão na mesma organização da AWS ou na mesma unidade organizacional (UO). Você pode usar UOs para agrupar contas que deseja administrar como uma única unidade.

A próxima etapa é reduzir o conjunto máximo de permissões que você pode conceder às entidades principais nas contas-membro da sua organização. Você pode usar políticas de controle de serviço (SCPs) para essa finalidade, as quais podem ser aplicadas a uma UO ou a uma conta. As SCPs

podem impor controles de acesso comuns, como restringir o acesso a determinadas Regiões da AWS, ajudar a impedir a exclusão de recursos ou desabilitar ações de serviço possivelmente arriscadas. As SCPs que você aplica à raiz da sua organização afetam apenas as contas-membro, mas não a conta de gerenciamento. As SCPs regem apenas as entidades principais da sua organização. As SCPs não regem entidades principais externas à sua organização que estão acessando seus recursos.

Outra etapa é usar políticas de recursos do IAM para definir o escopo das ações disponíveis que você pode realizar nos recursos por elas governados, juntamente com quaisquer condições que o diretor interino deva atender. Isso pode ser tão amplo quanto permitir todas as ações, desde que a entidade principal faça parte da sua organização (usando a chave de condição PrincipalOrgID), ou tão granular quanto permitir apenas ações específicas de um perfil do IAM específico. É possível adotar uma abordagem semelhante com as condições das políticas de confiança de perfis do IAM. Se uma política de confiança de recursos ou perfis nomear explicitamente uma entidade principal na mesma conta que o perfil ou o recurso por ela governado, essa entidade principal não precisará de uma política do IAM anexada que conceda as mesmas permissões. Se a entidade principal estiver em uma conta diferente da conta do recurso, ela precisará de uma política do IAM anexada que conceda essas permissões.

Muitas vezes, uma equipe de workload quer gerenciar as permissões exigidas pela workload em questão. Isso, por sua vez, pode exigir que a equipe crie políticas de permissão e perfis do IAM. É possível capturar o escopo máximo de permissões que a equipe pode conceder em um limite de permissão do IAM e associar esse documento a um perfil do IAM que a equipe pode usar para gerenciar seus perfis do IAM e permissões. Essa abordagem pode proporcionar à equipe a capacidade concluir seu trabalho e, ao mesmo tempo, reduzir os riscos de acesso administrativo ao IAM.

Uma etapa mais granular é implementar técnicas de gerenciamento de acesso privilegiado (PAM) e gerenciamento de acesso elevado temporário (TEAM). Um exemplo de PAM é exigir que as entidades principais realizem a autenticação multifator antes de executar ações privilegiadas. Para obter mais informações, consulte Configuração de acesso à API protegido por MFA. O TEAM exige uma solução que gerencie a aprovação e o período durante o qual uma entidade principal pode ter acesso elevado. Uma abordagem é adicionar temporariamente a entidade principal à política de confiança de perfis referente a um perfil do IAM que tenha acesso elevado. Outra abordagem é, em operação normal, reduzir o escopo das permissões concedidas a uma entidade principal por um perfil do IAM usando uma política de sessão e, em seguida, suspender temporariamente essa restrição durante o período aprovado. Para saber mais sobre as soluções validadas pela AWS e por parceiros selecionados, consulte Acesso elevado temporário.

Etapas de implementação

1. Isole workloads e ambientes em Contas da AWS separadas.
2. Use SCPs para reduzir o conjunto máximo de permissões que podem ser concedidas às entidades principais nas contas dos membros da sua organização.
   a. Recomendamos adotar uma abordagem de lista de permissões para redigir as SCPs que negam todas as ações, exceto aquelas permitidas, bem como descrever as condições sob as quais elas são permitidas. Defina primeiro os recursos que você deseja controlar e, em seguida, o efeito de negar. Use o elemento NotAction para negar todas as ações, exceto aquelas que você especificar. Associe isso a uma condição NotLike para definir quando essas ações são permitidas, se aplicável, como StringNotLike e ArnNotLike.
   b. Consulte Exemplos de políticas de controle de serviço.
3. Use políticas de recursos do IAM para reduzir o escopo e especificar as condições das ações permitidas nos recursos. Use condições nas políticas de confiança de perfis do IAM para criar restrições aos perfis assumidos.
4. Atribua limites de permissão do IAM a perfis do IAM que as equipes de workload podem usar para gerenciar perfis e permissões do IAM de suas próprias workloads.
5. Avalie as soluções de PAM e TEAM com base em suas necessidades.

Recursos

Documentos relacionados:

- Perímetro de dados na AWS
- Estabelecer barreiras de proteção para permissões usando perímetros de dados
- Lógica da avaliação de política

Exemplos relacionados:

- Exemplos de políticas de controle de serviço

Ferramentas relacionadas:

- Solução da AWS: gerenciamento de acesso elevado temporário
- Soluções validadas de parceiros de segurança para TEAM

SEC03-BP06 Gerenciar o acesso com base no ciclo de vida

Monitore e ajuste as permissões concedidas às entidades principais (usuários, funções e grupos) durante todo o ciclo de vida em sua organização. Ajuste as associações de grupo à medida que os usuários mudarem de função e remova o acesso quando um usuário sair da organização.

Resultado desejado: você monitora e ajusta as permissões em todo o ciclo de vida dos diretores da organização, reduzindo o risco de privilégios desnecessários. Você concede acesso apropriado ao criar um usuário. Você modifica o acesso à medida que as responsabilidades do usuário mudam e remove o acesso quando o usuário não está mais ativo ou sai da organização. Você gerencia centralmente as alterações em seus usuários, perfis e grupos. Você usa a automação para propagar alterações em seus ambientes da AWS.

Práticas comuns que devem ser evitadas:

- Você concede antecipadamente privilégios de acesso excessivos ou amplos às identidades que se estendem além daqueles exigidos inicialmente.
- Você não revisa nem ajusta os privilégios de acesso à medida que as funções e responsabilidades das identidades mudam ao longo do tempo.
- Você mantém identidades inativas ou encerradas com privilégios de acesso ativos. Isso aumenta o risco de acesso não autorizado.
- Você não automatiza o gerenciamento dos ciclos de vida das identidades.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Gerencie e ajuste cuidadosamente os privilégios de acesso concedidos às identidades (como usuários, funções, grupos) durante todo o ciclo de vida. Esse ciclo de vida inclui a fase inicial de integração, mudanças contínuas em perfis e responsabilidades e eventual desligamento ou rescisão. Gerencie proativamente o acesso com base no estágio do ciclo de vida para manter o nível de acesso adequado. Siga o princípio de privilégio mínimo para reduzir o risco de privilégios de acesso excessivos ou desnecessários.

Você pode gerenciar o ciclo de vida dos usuários do IAM diretamente na Conta da AWS ou por meio de federação no provedor de identidades de seu quadro de funcionários para o Centro de Identidade do AWS IAM. Para usuários do IAM, é possível criar, modificar e excluir usuários e as respectivas permissões associadas na Conta da AWS. No caso de usuários federados, você pode usar o Centro

de Identidade do IAM para gerenciar o respectivo ciclo de vida sincronizando as informações de usuários e grupos do provedor de identidades da sua organização por meio do protocolo System for Cross-Domain Identity Management (SCIM).

O SCIM é um protocolo de padrão aberto para provisionamento e desprovisionamento automatizados de identidades de usuários em diferentes sistemas. Ao integrar seu provedor de identidades ao Centro de Identidade do IAM usando o SCIM, você pode sincronizar automaticamente as informações do usuário e do grupo para ajudar a validar que os privilégios de acesso sejam concedidos, modificados ou revogados com base nas alterações na fonte de identidade autorizada da sua organização.

À medida que as funções e responsabilidades dos funcionários mudam em sua organização, ajuste os respectivos privilégios de acesso de maneira correspondente. Você pode usar os conjuntos de permissões do Centro de Identidade do IAM para definir diferentes funções ou responsabilidades de trabalho e associá-las a políticas e permissões apropriadas do IAM. Quando a função de um funcionário muda, você pode atualizar o conjunto de permissões atribuído para refletir as novas responsabilidades. Verifique se ele tem o acesso necessário e segue o princípio de privilégio mínimo.

Etapas de implementação

1. Defina e documente um processo de ciclo de vida do gerenciamento de acesso, incluindo procedimentos para concessão de acesso inicial, revisões periódicas e desligamento.
2. Implemente perfis, grupos e limites de permissões do IAM para gerenciar o acesso coletivamente e impor os níveis máximos de acesso permitidos.
3. Integre-se a um provedor de identidades federado (como Microsoft Active Directory, Okta, Ping Identity) como fonte confiável para uso de informações de usuários e grupos usando o Centro de Identidade do IAM.
4. Use o protocolo SCIM para sincronizar informações de usuários e grupos do provedor de identidades com o repositório de identidades do Centro de Identidade do IAM.
5. Crie conjuntos de permissões no Centro de Identidade do IAM que representem diferentes cargos ou responsabilidades em sua organização. Defina as políticas e permissões apropriadas do IAM para cada conjunto de permissões.
6. Implemente análises regulares de acesso, revogação imediata do acesso e melhoria contínua do processo do ciclo de vida do gerenciamento de acesso.
7. Ofereça treinamento e conhecimento aos funcionários sobre as práticas recomendadas de gerenciamento de acesso.

Recursos

Práticas recomendadas relacionadas:

- SEC02-BP04 Confiar em um provedor de identidades centralizado

Documentos relacionados:

- Gerenciar sua fonte de identidade
- Gerenciar identidades no Centro de Identidade do IAM
- Como usar o AWS Identity and Access Management Access Analyzer
- Geração de políticas do IAM Access Analyzer

Vídeos relacionados:

- AWS re:Inforce 2023: Gerenciar o acesso elevado temporário com o AWS IAM Identity Center
- AWS re:Invent 2022: Simplificar o acesso da sua força de trabalho com o Centro de Identidade do IAM
- AWS re:Invent 2022: Aproveitar o poder das políticas do IAM e controlar permissões com o Access Analyzer

SEC03-BP07 Analisar o acesso público e entre contas

Monitore continuamente as descobertas que destacam o acesso público e entre contas. Limite o acesso público e o acesso entre contas somente aos recursos específicos que exigem esse tipo de acesso.

Resultado desejado: saiba quais de seus recursos da AWS são compartilhados e com quem. Monitore e audite continuamente seus recursos compartilhados para verificar se eles são compartilhados apenas com entidades principais autorizadas.

Práticas comuns que devem ser evitadas:

- Não manter um inventário dos recursos compartilhados.
- Não seguir um processo de aprovação do acesso público ou entre contas aos recursos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Se a sua conta estiver no AWS Organizations, você poderá conceder acesso aos recursos à toda a organização, a unidades organizacionais específicas ou a contas individuais. Se sua conta não for membro de uma organização, você poderá compartilhar recursos com contas individuais. Você pode conceder acesso direto entre contas usando políticas baseadas em recursos — por exemplo, políticas de bucket do Amazon Simple Storage Service (Amazon S3) — ou permitindo que um principal em outra conta assuma uma função do IAM em sua conta. Ao utilizar políticas de recursos, verifique se o acesso é concedido apenas a entidades principais autorizadas. Defina um processo para aprovar todos os recursos que devem ser acessíveis publicamente.

O AWS Identity and Access Management Access Analyzer segurança comprovada para identificar todos os caminhos de acesso a um recurso de fora de sua conta. Ele revisa as políticas de recursos continuamente e relata descobertas de acesso público e entre contas para facilitar a análise de acesso potencialmente amplo. Considere a configuração do IAM Access Analyzer com o AWS Organizations para verificar se você tem visibilidade em todas as suas contas. O IAM Access Analyzer também permite que você visualize as descobertas antes de implantar permissões de recursos. Isso permite validar que as alterações de política concedam apenas o acesso público e entre contas pretendido aos seus recursos. Ao designar para acesso a várias contas, você pode usar políticas de confiança para controlar em quais casos um perfil pode ser assumido. Por exemplo, você pode usar a chave de condição `PrincipalOrgId` para negar uma tentativa de assumir uma função fora da sua AWS Organizations.

O AWS Config pode relatar recursos que estão configurados incorretamente e, por meio de verificações de políticas do AWS Config, pode detectar recursos com acesso público configurado. Serviços como AWS Control Tower e AWS Security Hub simplificam as barreiras de proteção e as verificações de implantação em um AWS Organizations para identificar e corrigir recursos publicamente expostos. Por exemplo, AWS Control Tower tem uma barreira de proteção gerenciada que pode detectar se algum snapshot do Amazon EBS pode ser restaurado por Contas da AWS.

Etapas de implementação

- Considere usar o AWS Config para AWS Organizations: o AWS Config permite agregar descobertas de várias contas em um AWS Organizations na conta de um administrador delegado. Isso fornece uma visão abrangente e permite que você implante Regras do AWS Config em várias contas para detectar recursos acessíveis ao público.

- Configure o AWS Identity and Access Management Access Analyzer. O IAM Access Analyzer ajuda você a identificar os recursos em sua organização e suas contas, como buckets do Amazon S3 ou perfis do IAM, que são compartilhados com uma entidade externa.
- Use a remediação automática no AWS Config para responder a mudanças na configuração de acesso público dos buckets do Amazon S3: você pode ativar automaticamente as configurações de bloqueio de acesso público para buckets do Amazon S3.
- Implemente monitoramento e alertas para identificar se os buckets do Amazon S3 se tornaram públicos: você deve ter monitoramento e alertas em vigor para identificar quando o Bloqueio de Acesso Público do Amazon S3 está desativado e se os buckets do Amazon S3 se tornam públicos. Além disso, se você estiver usando o AWS Organizations, poderá criar uma política de controle de serviços que impeça alterações nas políticas de acesso público do Amazon S3. O AWS Trusted Advisor procura buckets do Amazon S3 que têm permissões de acesso livre. As permissões de bucket que concedem acesso de upload ou exclusão a todos criam possíveis problemas de segurança, pois permitem que qualquer pessoa adicione, modifique ou remova itens em um bucket. A verificação do Trusted Advisor examina as permissões de bucket explícitas e as políticas de bucket associadas que podem substituir as permissões de bucket. Você também pode utilizar o AWS Config para monitorar seus buckets do Amazon S3 para acesso público. Para obter mais informações, consulte Como usar o AWS Config para monitorar e responder a buckets do Amazon S3 que permitem acesso público. Ao revisar o acesso, é importante considerar quais tipos de dados estão contidos nos buckets do Amazon S3. O Amazon Macie ajuda a descobrir e proteger dados confidenciais, como PII, PHI e credenciais, como chaves privadas ou da AWS.

Recursos

Documentos relacionados:

- Como usar o AWS Identity and Access Management Access Analyzer
- Biblioteca de controles do AWS Control Tower
- Padrão de práticas recomendas de segurança básica da AWS
- Regras gerenciadas do AWS Config
- Referência de verificação do AWS Trusted Advisor
- Monitorar resultados da verificação do AWS Trusted Advisor com o Amazon EventBridge
- Habilitar regras do AWS Config em todas as contas na sua organização
- AWS Config e AWS Organizations
- Disponibilizar publicamente sua AMI para uso no Amazon EC2

Vídeos relacionados:

- Práticas recomendadas para proteger seu ambiente de várias contas
- Mergulho profundo no IAM Access Analyzer

SEC03-BP08 Compartilhar recursos com segurança em sua organização

À medida que o número de workloads aumenta, talvez você precise compartilhar o acesso aos recursos nessas workloads ou fornecer os recursos várias vezes nas contas. É possível usar constructos para compartimentalizar seu ambiente, por exemplo, para ter ambientes de desenvolvimento, teste e produção. No entanto, ter constructos de separação não impede que você compartilhe com segurança. Ao compartilhar componentes que se sobrepõem, você pode reduzir a sobrecarga operacional e possibilitar uma experiência consistente sem precisar adivinhar o que ignorou ao criar o mesmo recurso várias vezes.

Resultado desejado: minimize o acesso não intencional usando métodos seguros para compartilhar recursos em sua organização e ajudar na sua iniciativa de prevenção de perda de dados. Reduza sua sobrecarga operacional em comparação com o gerenciamento de componentes individuais, reduza os erros gerados pela criação manual do mesmo componente várias vezes e aumente a escalabilidade das suas workloads. É possível se beneficiar da redução de tempo para a resolução em cenários de falhas em vários pontos e aumentar sua confiança na determinação de quando um componente não é mais necessário. Para obter recomendações sobre a análise de recursos compartilhados externamente, consulte SEC03-BP07 Analisar o acesso público e entre contas.

Práticas comuns que devem ser evitadas:

- Ausência de um processo para monitorar de forma contínua e alertar automaticamente sobre o compartilhamento externo inesperado.
- Ausência de referência sobre o que deve ou não ser compartilhado.
- Adotar como padrão uma política amplamente aberta em vez de compartilhar explicitamente quando necessário.
- Criar manualmente recursos básicos que se sobrepõem quando necessário.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Projete seus controles e padrões de acesso para reger o consumo de recursos compartilhados com segurança e somente com entidades confiáveis. Monitore recursos compartilhados e revise o acesso a eles de forma contínua e seja alertado sobre o compartilhamento inadequado ou inesperado. Revise Analisar o acesso público e entre contas para saber como estabelecer uma governança para limitar o acesso externo somente aos recursos que o exijam e estabelecer um processo para monitorar continuamente e enviar alertas de forma automática.

O compartilhamento entre contas no AWS Organizations é aceito por vários serviços da AWS, como AWS Security Hub, Amazon GuardDuty e AWS Backup. Esses serviços possibilitam compartilhar os dados em uma conta central, acessá-los ou gerenciar recursos e dados dessa conta. Por exemplo, o AWS Security Hub pode transferir as descobertas de contas individuais para uma conta central onde é possível visualizar todas elas. O AWS Backup pode fazer um backup de um recurso e compartilhá-lo entre contas. É possível usar o AWS Resource Access Manager (AWS RAM) para compartilhar outros recursos comuns, como sub-redes de VPC e anexos do gateway de trânsito, AWS Network Firewall ou Amazon SageMaker Pipelines.

Para restringir sua conta para compartilhar apenas recursos dentro de sua organização, use políticas de controle de serviços (SCPs) para impedir o acesso a entidades externas. Ao compartilhar recursos, combine controles baseados em identidade e controles de rede para criar um perímetro de dados para sua organização e ajudar a se proteger contra acesso não intencional. Um perímetro de dados é um conjunto de barreiras de proteção preventivas que ajudam a garantir que apenas suas identidades confiáveis acessem recursos confiáveis das redes esperadas. Esses controles impõem limites apropriados sobre quais recursos podem ser compartilhados e impedir o compartilhamento ou a exposição de recursos que não devem ser permitidos. Por exemplo, como parte do seu perímetro de dados, você pode usar as políticas de endpoint da VPC e a condição `AWS:PrincipalOrgId` para garantir que as identidades que acessam seus buckets do Amazon S3 pertençam à sua organização. É importante observar que as SCPs não se aplicam a perfis vinculados a serviços ou entidades principais de serviços da AWS.

Ao usar o Amazon S3, desative as ACLs do seu bucket do Amazon S3 e use as políticas do IAM para definir o controle de acesso. Para restringir o acesso a uma origem do Amazon S3 do Amazon CloudFront, migre da identidade do acesso de origem (OAI) para um controle de acesso de origem (OAC) compatível com recursos adicionais, incluindo criptografia do lado do servidor com o AWS Key Management Service.

Em alguns casos, convém permitir o compartilhamento de recursos fora de sua organização ou conceder a terceiros acesso aos seus recursos. Para obter recomendações sobre o gerenciamento de permissões para compartilhar recursos externamente, consulte Gerenciamento de permissões.

Etapas de implementação

1. Use AWS Organizations.

   O AWS Organizations é um serviço de gerenciamento de contas que permite consolidar várias Contas da AWS em uma organização criada e gerencia centralmente por você. É possível agrupar suas contas em unidades organizacionais (UOs) e anexar políticas diferentes a cada UO a fim de ajudar a atender às suas necessidades orçamentárias, de segurança e conformidade. Também é possível controlar como serviços de inteligência artificial (IA) e machine learning (ML) da AWS podem coletar e armazenar dados e usar o gerenciamento de várias contas dos serviços da AWS integrados ao Organizations.

2. Integre o AWS Organizations aos serviços da AWS.

   Quando você habilita um serviço da AWS para executar tarefas em seu nome nas contas-membro da organização, o AWS Organizations cria um perfil vinculado ao serviço do IAM para esse serviço em cada conta-membro. Você deve gerenciar o acesso confiável usando o AWS Management Console, as APIs da AWS ou a AWS CLI. Para obter recomendações sobre como ativar o acesso confiável, consulte Usar o AWS Organizations com outros serviços da AWS e Serviços da AWS que você pode usar com o Organizations.

3. Estabeleça um perímetro de dados.

   O perímetro da AWS geralmente é representado como uma organização gerenciada pelo AWS Organizations. Junto com redes e sistemas on-premises, o acesso a recursos da AWS é o que muitos consideram o perímetro de My AWS. O objetivo do perímetro é garantir que o acesso seja permitido se a identidade e o recurso forem confiáveis e a rede for esperada.

   a. Defina e implante os perímetros.

      Siga as etapas descritas em Implementação do perímetro no whitepaper "Construir um perímetro na AWS" para cada condição de autorização. Para obter recomendações sobre a proteção da camada de rede, consulte Proteger redes.

   b. Monitore e alerte de forma contínua.

      O AWS Identity and Access Management Access Analyzer ajuda a identificar os recursos da organização e as contas que são compartilhados com entidades externas. É possível integrar o

IAM Access Analyzer ao AWS Security Hub para enviar e agregar descobertas de um recurso do IAM Access Analyzer ao Security Hub com o objetivo de ajudar a analisar a postura de segurança do seu ambiente. Para integrar, habilite tanto o IAM Access Hub quanto o Security Hub em cada região em cada conta. Também é possível usar o Regras do AWS Config para auditar a configuração e alertar a parte apropriada usando o AWS Chatbot com AWS Security Hub. Em seguida, você pode usar documentos de automação do AWS Systems Manager para corrigir recursos fora de conformidade.

c. Para obter recomendações sobre monitoramento e alertas contínuos sobre recursos compartilhados externamente, consulte Analisar o acesso público e entre contas.

4. Use o compartilhamento de recursos nos serviços da AWS e restrinja adequadamente.

   Muitos serviços da AWS permitem compartilhar recursos com outra conta ou direcionar um recurso em outra conta, como imagens de máquina da Amazon (AMIs) e o AWS Resource Access Manager (AWS RAM). Restrinja a API `ModifyImageAttribute` para especificar as contas confiáveis com as quais a AMI será compartilhada. Especifique a condição `ram:RequestedAllowsExternalPrincipals` ao usar o AWS RAM para restringir o compartilhamento somente à sua organização, ajudando assim a impedir o acesso de identidades não confiáveis. Para orientações e recomendações, consulte Compartilhamento de recursos e destinos externos.

5. Use AWS RAM para compartilhar com segurança em uma conta ou com outras Contas da AWS.

   O AWS RAM ajuda você a compartilhar com segurança os recursos criados com perfis e usuários em sua conta e em outras Contas da AWS. Em um ambiente de várias contas, o AWS RAM permite criar um recurso uma vez e compartilhá-lo com outras contas. Essa abordagem ajuda a reduzir sua sobrecarga operacional enquanto oferece consistência, visibilidade e capacidade de auditoria por meio de integrações com o Amazon CloudWatch e o AWS CloudTrail, o que você não recebe ao utilizar o acesso entre contas.

   Se você tiver recursos que compartilhou anteriormente usando uma política baseada em recursos, poderá usar a API `PromoteResourceShareCreatedFromPolicy` ou equivalente para promover o compartilhamento de recursos em um compartilhamento de recursos do AWS RAM completo.

   Em alguns casos, convém realizar etapas adicionais para compartilhar recursos. Por exemplo, para compartilhar um snapshot criptografado, é necessário compartilhar uma chave do AWS KMS.

Recursos

Práticas recomendadas relacionadas:

- SEC03-BP07 Analisar o acesso público e entre contas
- SEC03-BP09 Compartilhar recursos com terceiros de forma segura
- SEC05-BP01 Criar camadas de rede

Documentos relacionados:

- O proprietário do bucket concede permissão entre contas para objetos que não possui
- Como usar políticas de confiança com o IAM
- Como criar um perímetro de dados na AWS
- Como usar um ID externo ao conceder acesso aos seus recursos da AWS para terceiros
- Serviços da AWS que você pode usar com o AWS Organizations
- Estabelecer um perímetro de dados na AWS: permitir que somente identidades confiáveis acessem os dados da empresa

Vídeos relacionados:

- Acesso granular com o AWS Resource Access Manager
- Como proteger seu perímetro de dados com endpoints da VPC
- Estabelecer um perímetro de dados na AWS

Ferramentas relacionadas:

- Exemplos de políticas de perímetro de dados

SEC03-BP09 Compartilhar recursos com terceiros de forma segura

A segurança de seu ambiente de nuvem não para na sua organização. Sua organização pode contar com terceiros para gerenciar uma parte de seus dados. O gerenciamento de permissões para o sistema gerenciado por terceiros deve seguir a prática de acesso just-in-time utilizando o princípio de privilégio mínimo com credenciais temporárias. Ao trabalhar em parceria com terceiros, é possível reduzir o escopo do impacto e o risco de acesso acidental.

Resultado desejado: credenciais do AWS Identity and Access Management (IAM) de longo prazo, chaves de acesso do IAM e chaves secretas associadas a um usuário podem ser usadas por qualquer pessoa, desde que sejam válidas e ativas. O uso de um perfil do IAM e de credenciais temporárias ajuda você a melhorar seu procedimento de segurança geral reduzindo o esforço para manter credenciais de longo prazo, inclusive o gerenciamento e a sobrecarga operacional dessas informações sigilosas. Ao utilizar um identificador universalmente exclusivo (UUID) para o ID externo na política de confiança do IAM e manter as políticas do IAM anexadas ao perfil do IAM sob seu controle, é possível fazer auditoria e garantir que o acesso concedido a terceiros não seja permissivo demais. Para obter recomendações sobre a análise de recursos compartilhados externamente, consulte SEC03-BP07 Analisar o acesso público e entre contas.

Práticas comuns que devem ser evitadas:

- Utilizar a política de confiança padrão do IAM sem nenhuma condição.
- Utilizar credenciais e chaves de acesso de longo prazo do IAM.
- Reutilizar IDs externos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Talvez você deseje permitir o compartilhamento de recursos fora do AWS Organizations ou conceder a terceiros acesso à sua conta. Por exemplo, um parceiro (terceiro) pode oferecer uma solução de monitoramento que precise acessar recursos em sua conta. Nesses casos, crie um perfil entre contas do IAM somente com os privilégios necessários para o parceiro. Além disso, defina uma política de confiança usando a condição de ID externo. Ao utilizar um ID externo, você ou o parceiro pode gerar um ID exclusivo para cada cliente, terceiro ou locação. O ID exclusivo não deve ser controlado por ninguém, exceto por você, depois de criado. O parceiro deve implementar um processo para relacionar o ID externo ao cliente de forma segura, auditável e reproduzível.

Também é possível usar o IAM Roles Anywhere para gerenciar perfis do IAM para aplicações fora da AWS que usam APIs da AWS.

Se o parceiro não precisar mais de acesso ao seu ambiente, remova o perfil. Evite fornecer credenciais de longo prazo para terceiros. Esteja ciente de outros serviços da AWS compatíveis com compartilhamento. Por exemplo, o AWS Well-Architected Tool permite compartilhar workloads com outras Contas da AWS e o AWS Resource Access Managerajuda você a compartilhar com segurança um recurso da AWS pertencente a você com outras contas.

Etapas de implementação

1. Use perfis entre contas para fornecer acesso a contas externas.

   Os perfis entre contas reduzem a quantidade de informações confidenciais armazenadas por contas externas e por terceiros para atender os clientes. Os perfis entre contas possibilitam a você conceder acesso a recursos da AWS em sua conta de forma segura a terceiros, como AWS Partners ou outras contas em sua organização e, ao mesmo tempo, manter a capacidade de gerenciar e auditar esse acesso.

   O parceiro pode oferecer serviço a você a partir de uma infraestrutura híbrida ou, como alternativa, extrair dados de um local externo. O IAM Roles Anywhere ajuda você a permitir que workloads de terceiros interajam com segurança com suas workloads da AWS e a reduzir ainda mais a necessidade de credenciais de longo prazo.

   Você não deve usar credenciais ou chaves de acesso de longo prazo associadas a usuários para conceder acesso a contas externas. Em vez disso, utilize perfis entre contas para conceder acesso entre contas.
2. Use um ID externo com terceiros.

   O uso de um ID externo permite que você determine quem pode assumir um perfil em uma política de confiança do IAM. A política de confiança pode exigir que o usuário que assume o perfil imponha a condição e o destino no qual ele está operando. Também oferece uma forma para o proprietário da conta permitir que a função seja assumida apenas em determinadas circunstâncias. A função principal do ID externo é abordar e impedir o problema de substituto confuso.

   Utilize um ID externo se você for proprietário de uma Conta da AWS e tiver configurado um perfil para terceiros que acesse outras Contas da AWS além da sua, ou quando você pode assumir perfis em nome de clientes diferentes. Trabalhe com terceiros ou a AWS Partner para estabelecer uma condição de ID externo a ser incluída na política de confiança do IAM.
3. Use IDs externos universalmente exclusivos.

   Implemente um processo que gere um valor exclusivo aleatório para um ID externo, como um identificador universalmente exclusivo (UUID). Um parceiro que reutilize IDs externos entre diferentes clientes não resolve o problema de substituto confuso porque o cliente A pode ser capaz de visualizar dados do cliente B utilizando o ARN do perfil do cliente B junto com o ID externo duplicado. Em um ambiente de vários locatários em que um parceiro atende a vários

clientes com diferentes Contas da AWS, o parceiro deve usar um ID exclusivo diferente como o ID externo de cada Conta da AWS. O parceiro é responsável por detectar IDs externos duplicados e mapear de forma segura cada cliente ao seu respectivo ID externo. O parceiro deve testar para verificar se ele pode assumir o perfil somente ao especificar o ID externo. O parceiro deve evitar armazenar o ARN do perfil do cliente e o ID externo até que este seja necessário.

O ID externo não é tratado como segredo, mas ele não pode ser um valor facilmente dedutível, como um número de telefone, um nome ou o ID da conta. Torne o ID externo um campo somente leitura de forma que o ID externo não possa ser alterado com o fim de representar a configuração.

Você ou o parceiro podem gerar o ID externo. Defina um processo para determinar quem é responsável pela geração do ID. Seja qual for a entidade que crie o ID externo, o parceiro impõe a exclusividade e os formatos de forma consistente entre os clientes.

4. Deprecie as credenciais de longo prazo fornecidas pelo cliente.

   Deprecie o uso de credenciais de longo prazo e use perfis entre clientes ou o IAM Roles Anywhere. Se você precisar utilizar credenciais de longo prazo, estabeleça um plano para migrar para um acesso baseado em perfil. Para obter detalhes sobre o gerenciamento de chaves, consulte Gerenciamento de identidades. Trabalhe também com a equipe da sua Conta da AWS e o parceiro para estabelecer um runbook de mitigação de riscos. Para obter recomendações sobre como responder e mitigar o impacto potencial de incidentes de segurança, consulte Resposta a incidentes.

5. Verifique se a configuração possui recomendações ou é automatizada.

   A política criada para acesso entre contas em suas contas deve seguir o princípio de privilégio mínimo. O terceiro deve fornecer um documento de política de perfil ou um mecanismo de configuração automatizada que utilize um modelo do AWS CloudFormation ou um equivalente para você. Isso reduz a chance de erros associados à criação manual de políticas e oferece uma trilha auditável. Para obter mais informações sobre como usar um modelo do AWS CloudFormation para criar funções entre contas, consulte Funções entre contas.

   O terceiro deve fornecer um mecanismo de configuração automatizado e auditável. No entanto, ao utilizar o documento de política de perfis que descreve o acesso necessário, você deve automatizar a configuração do perfil. Com um modelo do AWS CloudFormation ou equivalente, monitore alterações com detecção de desvios como parte da prática de auditoria.

6. Considere as alterações.

Sua estrutura de contas, sua necessidade de terceiros ou a oferta de serviço pode sofrer alterações. Você deve antecipar alterações e falhas e planejar adequadamente com as pessoas, o processo e a tecnologia corretos. Audite o nível de acesso que você concede periodicamente e implemente métodos de detecção para ser alertado sobre alterações inesperadas. Monitore e audite o uso do perfil e o datastore dos IDs externos. Você deve estar preparado para revogar o acesso de terceiros, seja de forma temporária ou permanente, como resultado de alterações ou padrões de acesso inesperados. Além disso, meça o impacto de sua operação de revogação, inclusive o tempo para realizá-la, as pessoas envolvidas, o custo e o impacto de outros recursos.

Para obter recomendações sobre métodos de detecção, consulte as práticas recomendadas de detecção.

Recursos

Práticas recomendadas relacionadas:

- SEC02-BP02 Usar credenciais temporárias
- SEC03-BP05 Definir barreiras de proteção de permissões para sua organização
- SEC03-BP06 Gerenciar o acesso com base no ciclo de vida
- SEC03-BP07 Analisar o acesso público e entre contas
- SEC04 Detecção

Documentos relacionados:

- O proprietário do bucket concede permissão entre contas para objetos que não possui
- Como usar políticas de confiança com perfis do IAM
- Delegar acesso entre Contas da AWS usando perfis do IAM
- Como faço para acessar recursos em outra Conta da AWS usando o IAM?
- Práticas recomendadas de segurança no IAM
- Lógica de avaliação de política entre contas
- Como usar um ID externo ao conceder acesso aos seus recursos da AWS para terceiros
- Coletar informações de recursos da AWS CloudFormation criados em contas externas com recursos personalizados
- Usar IDs externos com segurança para acessar contas da AWS pertencentes a terceiros

- Estender os perfis do IAM para workloads fora do IAM com o IAM Roles Anywhere

Vídeos relacionados:

- Como faço para permitir que usuários ou perfis em uma Conta da AWS separada tenham acesso à minha Conta da AWS?
- AWS re:Invent 2018: Torne-se um mestre e políticas do IAM em no máximo 60 minutos
- Centro de Conhecimentos da AWS Live: Práticas recomendadas e decisões de design do IAM

Exemplos relacionados:

- Laboratório do Well-Architected: Suposição de perfil do IAM entre contas do Lambda (Nível 300)
- Configurar o acesso entre contas ao Amazon DynamoDB
- Ferramenta de consulta de rede do AWS STS

Detecção

Pergunta

- SEC 4. Como detectar e investigar eventos de segurança?

SEC 4. Como detectar e investigar eventos de segurança?

Capture e analise eventos de logs e métricas para obter visibilidade. Tomar medidas em relação aos eventos de segurança e possíveis ameaças para ajudar a proteger seu workload.

Práticas recomendadas

- SEC04-BP01 Configurar o registro em log de serviços e aplicações
- SEC04-BP02 Capturar logs, descobertas e métricas em locais padronizados
- SEC04-BP03 Correlacionar e enriquecer alertas de segurança
- SEC04-BP04 Iniciar a correção de recursos fora de conformidade

SEC04-BP01 Configurar o registro em log de serviços e aplicações

Retenha logs de eventos de segurança de serviços e aplicações. Esse é um princípio fundamental de segurança para auditoria, investigações e casos de uso operacionais e um requisito de segurança

comum orientado por padrões, políticas e procedimentos de governança, risco e conformidade (GRC).

Resultado desejado: uma organização deve ser capaz de recuperar de forma confiável e consistente logs de eventos de segurança de serviços e aplicações da AWS em tempo hábil quando necessário para cumprir um processo ou obrigação interna, como uma resposta a incidente de segurança. Considere centralizar os logs para obter os melhores resultados operacionais.

Práticas comuns que devem ser evitadas:

- Os logs são armazenados de forma perpétua ou excluídos muito precocemente.
- Todos podem acessar os logs.
- Contar inteiramente com processos manuais para uso e governança de logs.
- Armazenar todo e qualquer tipo de log para uma eventual necessidade.
- Conferir a integridade dos logs apenas quando necessário.

Benefícios de implementar esta prática recomendada: implemente um mecanismo de análise de causa-raiz (RCA) para incidentes de segurança e uma fonte de evidência para suas obrigações de governança, risco e conformidade.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Durante uma investigação de segurança ou outros casos de uso com base em seus requisitos, você precisa ser capaz de analisar os logs relevantes a fim de registrar e entender o escopo total e a linha do tempo do incidente. Os logs também são necessários para geração de alertas indicando que ocorreram determinadas ações de interesse. É essencial selecionar, ativar, armazenar e configurar mecanismos de consulta, recuperação e alertas.

Etapas de implementação

- Selecione e use fontes de log. Antes de uma investigação de segurança, você precisa capturar logs relevantes para reconstruir de forma retroativa a atividade em uma Conta da AWS. Selecione fontes de logs relevantes para suas workloads.

  Os critérios de seleção de fonte de logs devem se basear nos casos de uso necessários à sua empresa. Estabeleça uma trilha para cada Conta da AWS utilizando o AWS CloudTrail ou uma trilha do AWS Organizations e configure um bucket do Amazon S3 para ela.

O AWS CloudTrail é um serviço de registro em log que rastreia chamadas de API feitas em uma Conta da AWS capturando a atividade do serviço da AWS. Ele é ativado por padrão com uma retenção de 90 dias de eventos de gerenciamento que podem ser recuperados por meio do histórico de eventos do CloudTrail via AWS Management Console, AWS CLI ou AWS SDK. Para maior retenção e visibilidade dos eventos de dados, crie uma trilha do CloudTrail e associe-a a um bucket do Amazon S3 e, opcionalmente, a um grupo de log do Amazon CloudWatch. Como alternativa, você pode criar um CloudTrail Lake, o qual retém os logs do CloudTrail por até sete anos e fornece um recurso de consulta baseado em SQL

A AWS recomenda que os clientes que usam VPC ativem os logs de tráfego de rede e DNS usando Logs de fluxo da VPC e Logs de consulta do Amazon Route 53 Resolver, respectivamente, e os transmitam para um bucket do Amazon S3 ou um grupo de logs do CloudWatch. É possível criar um log de fluxo de VPC, uma sub-rede ou uma interface de rede. Para logs de fluxo de VPC, é possível ser seletivo em relação a como e onde usar os logs de fluxo para reduzir o custo.

Logs do AWS CloudTrail, logs de fluxo de VPC e logs de consulta do Route 53 Resolver são as fontes básicas de registro em log para oferecer compatibilidade com investigações de segurança na AWS. Também é possível usar o Amazon Security Lake para coletar, normalizar e armazenar esses dados de log no formato Apache Parquet e no Open Cybersecurity Schema Framework (OCSF), que está pronto para consulta. O Security Lake também é compatível com outros logs da AWS e logs de fontes de terceiros.

Os serviços da AWS podem gerar logs não capturados pelas fontes de log básicas, como logs do Elastic Load Balancing, logs do AWS WAF, logs de gravador do AWS Config, descobertas do Amazon GuardDuty, logs de auditoria do Amazon Elastic Kubernetes Service (Amazon EKS) e logs de sistema operacional e aplicações e de instâncias do Amazon EC2. Para obter uma lista completa das opções de registro e monitoramento, consulte o Apêndice A: Definições de capacidade de nuvem – Log e eventos do Guia de Resposta a Incidentes de Segurança da AWS.

- Pesquise recursos de log para cada serviço e aplicação da AWS: cada serviço e aplicação da AWS oferece opções de armazenamento de log, cada uma com seus próprios recursos de retenção e ciclo de vida. Os dois serviços de armazenamento de logs mais comuns são o Amazon Simple Storage Service (Amazon S3) e o Amazon CloudWatch. Para períodos de retenção longos, é recomendável utilizar o Amazon S3 para seus recursos de economia e ciclo de vida flexíveis. Se a opção de registro em log principal for o Amazon CloudWatch Logs, como opção, você deve considerar o arquivamento de logs menos acessados no Amazon S3.

- Selecione o armazenamento de logs: a escolha do armazenamento de logs geralmente está relacionada à ferramenta de consulta que você usa, aos recursos de retenção, à familiaridade e ao custo. As principais opções para armazenamento de logs são um bucket do Amazon S3 ou um grupo de logs do CloudWatch.

  Um bucket do Amazon S3 oferece armazenamento econômico e durável com uma política de ciclo de vida opcional. Os logs armazenados em buckets do Amazon S3 podem ser consultados com serviços como o Amazon Athena.

  Um grupo de logs do CloudWatch oferece armazenamento durável e um recurso de consultas incorporado por meio do CloudWatch Logs Insights.

- Identifique a retenção apropriada de logs: ao usar um bucket do Amazon S3 ou um grupo de logs do CloudWatch para armazenar logs, você deve estabelecer ciclos de vida adequados para cada fonte de log para otimizar os custos de armazenamento e recuperação. Os clientes geralmente têm entre três meses a um ano de logs prontamente disponíveis para consultas, com retenção de até sete anos. A escolha de disponibilidade e retenção deve se alinhar aos seus requisitos de segurança e um composto de atribuições regulatórias, estatutárias e de negócios.

- Use o registro em log para cada serviço e aplicação da AWS com políticas adequadas de retenção e ciclo de vida: para cada serviço ou aplicação da AWS em sua organização, procure a orientação específica de configuração do log:
  - Configurar trilha do AWS CloudTrail
  - Configurar Logs de fluxo da VPC
  - Configurar a exportação de descobertas do Amazon GuardDuty
  - Configurar a gravação do AWS Config
  - Configurar o tráfego de ACL da Web do AWS WAF
  - Configurar logs de tráfego de rede do AWS Network Firewall
  - Configurar logs de acesso do Elastic Load Balancing
  - Configurar logs de consulta do Amazon Route 53 Resolver
  - Configurar logs do Amazon RDS
  - Configurar logs do ambiente de gerenciamento do Amazon EKS
  - Configurar o agente do Amazon CloudWatch para instâncias do Amazon EC2 e servidores on-premises

- Selecione e implemente mecanismos de consulta para logs: para consultas em logs, é possível usar o CloudWatch Logs Insights para dados armazenados em grupos de log do CloudWatch, e

o Amazon Athena e o Amazon OpenSearch Service para dados armazenados no Amazon S3. Também é possível usar ferramentas de consulta de terceiros, como um serviço de gerenciamento de eventos e informações de segurança (SIEM).

O processo para selecionar uma ferramenta de consulta de log deve considerar as pessoas, o processo e os aspectos de tecnologia de suas operações de segurança. Selecione uma ferramenta que atenda aos requisitos operacionais, de negócios e segurança, esteja acessível e possa receber manutenção no longo prazo. Lembre-se de que as ferramentas de consulta de logs funcionam da forma ideal quando o número de logs a serem verificados é mantido dentro dos limites da ferramenta. Não é incomum ter várias ferramentas de consulta devido a restrições financeiras ou técnicas.

Por exemplo, você pode usar uma ferramenta de gerenciamento de eventos e informações de segurança (SIEM) de terceiros para realizar consultas para os últimos 90 dias de dados, mas usar o Athena para realizar consultas além de 90 dias devido ao custo de ingestão de logs de um SIEM. Seja qual for a implementação, garanta que sua abordagem minimize o número de ferramentas necessárias para maximizar a eficiência operacional, especialmente durante a investigação de um evento de segurança.

- Use logs para alertas: a AWS fornece alertas por meio de vários serviços de segurança.
  - O AWS Config monitora e registra as configurações de recursos da AWS e permite automatizar as tarefas de avaliação e correção em relação às configurações desejadas.
  - O Amazon GuardDuty é um serviço de detecção de ameaças que monitora continuamente atividades mal-intencionadas e comportamentos não autorizados para proteger suas Contas da AWS e workloads. O GuardDuty ingere, agrega e analisa informações de fontes, como eventos de gerenciamento e dados do AWS CloudTrail, logs de DNS, logs de fluxo de VPC e logs de auditoria do Amazon EKS. O GuardDuty extrai fluxos de dados independentes diretamente do CloudTrail, dos logs de fluxo de VPC, dos logs de consulta ao DNS e do Amazon EKS. Não é necessário gerenciar políticas de bucket do Amazon S3 nem modificar a forma de coletar e armazenar logs. Ainda é recomendável reter esses logs para sua própria investigação e fins de conformidade.
  - O AWS Security Hub fornece um único local que agrega, organiza e prioriza alertas de segurança ou descobertas de vários serviços da AWS e produtos opcionais de terceiros para oferecer uma visão abrangente dos alertas de segurança e do status de conformidade.

  Você também pode utilizar mecanismos de geração de alertas personalizados para alertas de segurança não cobertos por esses serviços ou para alertas específicos relevantes para o seu

ambiente. Para obter informações sobre como criar esses alertas e detecções, consulte Detecção no Guia de resposta a incidentes de segurança da AWS.

Recursos

Práticas recomendadas relacionadas:

- SEC04-BP02 Capturar logs, descobertas e métricas em locais padronizados
- SEC07-BP04 Definir o gerenciamento escalável do ciclo de vida dos dados
- SEC10-BP06 Implantar ferramentas previamente

Documentos relacionados:

- Guia de resposta a incidentes de segurança da AWS
- Conceitos básicos do Amazon Security Lake
- Conceitos básicos do Amazon CloudWatch Logs
- Soluções de segurança de parceiros: log e monitoramento

Vídeos relacionados:

- AWS re:Invent 2022: Lançamento do Amazon Security Lake

Exemplos relacionados:

- Ativador de log assistido para AWS
- Exportação do histórico de de descobertas do AWS Security Hub

Ferramentas relacionadas:

- Snowflake para segurança cibernética

SEC04-BP02 Capturar logs, descobertas e métricas em locais padronizados

As equipes de segurança confiam em logs e descobertas para analisar eventos que podem indicar atividades não autorizadas ou alterações não intencionais. Para agilizar essa análise, capture logs

e descobertas de segurança em locais padronizados. Fazer isso disponibiliza pontos de dados de interesse para correlação e pode simplificar a integração de ferramentas.

Resultado desejado: você tem uma abordagem padronizada para coletar, analisar e visualizar dados de log, descobertas e métricas. As equipes de segurança podem correlacionar, analisar e visualizar com eficiência os dados de segurança em sistemas diferentes para descobrir possíveis eventos de segurança e identificar anomalias. Os sistemas de gerenciamento de eventos e informações de segurança (SIEM) ou outros mecanismos são integrados para consultar e analisar dados de log e oferecer respostas oportunas, rastreamento e encaminhamento de eventos de segurança.

Práticas comuns que devem ser evitadas:

- As equipes são proprietárias e gerenciam de forma independente o registro em log e a coleta de métricas que são inconsistentes com a estratégia de registro em log da organização.
- As equipes não têm controles de acesso adequados para restringir a visibilidade e a alteração dos dados coletados.
- As equipes não controlam logs, descobertas e métricas de segurança como parte da política de classificação de dados.
- As equipes negligenciam os requisitos de soberania e localização dos dados ao configurar as coletas de dados.

Benefícios de implementar esta prática recomendada: uma solução de registro em log padronizada para coletar e consultar dados e eventos de registro melhora os insights derivados das informações neles contidas. Configurar um ciclo de vida automatizado para os dados de log coletados pode reduzir os custos incorridos pelo armazenamento de logs. É possível criar um controle de acesso refinado para as informações de log coletadas de acordo com a confidencialidade dos dados e os padrões de acesso necessários para as equipes. Você pode integrar ferramentas para correlacionar, visualizar e obter insights dos dados.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

O aumento do uso da AWS em uma organização resulta em um número crescente de workloads e ambientes distribuídos. Como cada um desses ambientes e workloads gera dados sobre as atividades dentro deles, capturar e armazenar esses dados localmente representa um desafio para as operações de segurança. As equipes de segurança usam ferramentas, como sistemas de gerenciamento de eventos e informações de segurança (SIEM), para coletar dados de fontes

distribuídas e submetê-los a fluxos de trabalho de correlação, análise e resposta. Isso requer o gerenciamento de um conjunto complexo de permissões para acessar as várias fontes de dados e despesas operacionais indiretas adicionais na operação dos processos de extração, transformação e carregamento (ETL).

Para superar esses desafios, considere agregar todas as fontes relevantes de dados de log de segurança em uma conta de arquivamento de logs, conforme descrito em Como organizar seu ambiente da AWS usando várias contas. Isso inclui todos os dados relacionados à segurança da sua workload e dos logs gerados pelos serviços da AWS, como AWS CloudTrail, AWS WAF, Elastic Load Balancing e Amazon Route 53. Há vários benefícios resultantes da captura desses dados em locais padronizados em uma Conta da AWS separada com as devidas permissões entre contas. Essa prática ajuda a evitar a violação de logs em workloads e ambientes comprometidos, fornece um único ponto de integração para ferramentas adicionais e oferece um modelo mais simplificado para configurar a retenção de dados e o ciclo de vida. Avalie os impactos da soberania de dados, dos escopos de conformidade e de outras regulamentações para determinar se vários locais de armazenamento de dados de segurança e períodos de retenção são necessários.

Para facilitar a captura e padronização de logs e descobertas, avalie o Amazon Security Lake em sua conta de arquivamento de logs. É possível configurar o Security Lake para ingerir automaticamente dados de fontes comuns, como CloudTrail, Route 53, Amazon EKS e logs de fluxo de VPC. Também é possível configurar AWS Security Hub como fonte de dados no Security Lake, permitindo que você correlacione descobertas de outros serviços da AWS, como Amazon GuardDuty e Amazon Inspector, com seus dados de log. Você também pode usar integrações de fontes de dados de terceiros ou configurar fontes de dados personalizadas. Todas as integrações padronizam seus dados no formato Open Cybersecurity Schema Framework (OCSF) e são armazenadas em buckets do Amazon S3 como arquivos Parquet, eliminando a necessidade de processamento de ETL.

O armazenamento de dados de segurança em locais padronizados fornece recursos avançados de análise. A AWS recomenda implantar ferramentas para análise de segurança que operem em um ambiente da AWS em uma conta do Security Tooling separada da sua conta de arquivamento de logs. Essa abordagem permite implantar controles detalhados para proteger a integridade e a disponibilidade dos logs e do processo de gerenciamento de logs, diferentemente das ferramentas que os acessam. Considere usar serviços, como o Amazon Athena, para executar consultas sob demanda que correlacionam várias fontes de dados. Também é possível integrar ferramentas de visualização, como o Amazon QuickSight. As soluções baseadas em IA estão se tornando cada vez mais disponíveis e podem desempenhar funções como traduzir descobertas em resumos legíveis por humanos e interação em linguagem natural. Essas soluções geralmente são mais fáceis de integrar por terem um local de armazenamento de dados padronizado para consulta.

Etapas de implementação

1. Crie as contas do Log Archive e do Security Tooling
    a. Usando o AWS Organizations, crie as contas Log Archive e Security Tooling em uma unidade organizacional de segurança. Se estiver usando o AWS Control Tower para gerenciar sua organização, as contas de arquivo de logs e de ferramentas de segurança serão criadas automaticamente para você. Configure perfis e permissões para acessar e administrar essas contas conforme necessário.
2. Configure seus locais de dados de segurança padronizados
    a. Determine sua estratégia para criar locais de dados de segurança padronizados. É possível conseguir isso por meio de opções como abordagens comuns de arquitetura de data lake, produtos de dados de terceiros ou Amazon Security Lake. A AWS recomenda capturar dados de segurança de Regiões da AWS opted-in para suas contas, mesmo quando não estiverem em uso ativo.
3. Configurar a publicação da fonte de dados em seus locais padronizados
    a. Identifique as fontes de seus dados de segurança e configure-as para publicação em seus locais padronizados. Avalie as opções para exportar dados automaticamente no formato desejado, em vez daquelas em que é necessário desenvolver processos de ETL. Com o Amazon Security Lake, é possível coletar dados de fontes da AWS compatíveis e sistemas integrados de terceiros.
4. Configurar ferramentas para acessar seus locais padronizados
    a. Configure ferramentas (como o Amazon Athena e o Amazon QuickSight) ou soluções de terceiros para ter o acesso necessário aos locais padronizados. Configure essas ferramentas para operar fora da conta de ferramentas de segurança com acesso de leitura entre contas à conta de arquivo de logs quando aplicável. Crie assinantes no Amazon Security Lake para fornecer a essas ferramentas acesso aos seus dados.

Recursos

Práticas recomendadas relacionadas:

- SEC01-BP01 Separar workloads usando contas
- SEC07-BP04 Definir o gerenciamento do ciclo de vida dos dados
- SEC08-BP04 Aplicar controle de acesso
- OPS08-BP02 Analisar logs de workloads

Documentos relacionados:

- Whitepapers da AWS: Organizar seu ambiente da AWS usando várias contas
- Recomendações da AWS: Arquitetura de referência de segurança da AWS (AWS SRA)
- Recomendações da AWS: Guia de log e monitoramento para proprietários de aplicações

Exemplos relacionados:

- Agregar, pesquisar e visualizar dados de log de fontes distribuídas com o Amazon Athena e o Amazon QuickSight
- Como visualizar descobertas do Amazon Security Lake com o Amazon QuickSight
- Gerar insights baseados em IA para o Amazon Security Lake usando o Amazon SageMaker Studio e o Amazon Bedrock
- Identificar anomalias de segurança cibernética em dados do Amazon Security Lake usando o Amazon SageMaker
- Ingerir, transformar e entregar eventos publicados pelo Amazon Security Lake para o Amazon OpenSearch Service
- Como usar o AWS Security Hub e o Amazon OpenSearch Service para SIEM

Ferramentas relacionadas:

- Amazon Security Lake
- Integrações de parceiros do Amazon Security Lake
- Open Cybersecurity Schema Framework (OCSF)
- Amazon Athena
- Amazon QuickSight
- Amazon Bedrock

SEC04-BP03 Correlacionar e enriquecer alertas de segurança

Atividades inesperadas podem gerar vários alertas de segurança de diferentes fontes, exigindo mais correlação e enriquecimento para entender o contexto completo. Implemente a correlação automatizada e o enriquecimento de alertas de segurança para ajudar a obter identificações e respostas mais precisas a incidentes.

Resultado desejado: à medida que a atividade gera alertas diferentes em seus ambientes e workloads, mecanismos automatizados correlacionam dados e enriquecem esses dados com informações adicionais. Esse pré-processamento apresenta uma compreensão mais detalhada do evento, o que ajuda os investigadores a determinar a importância do evento e se ele constitui um incidente que requer uma resposta formal. Esse processo reduz a carga sobre suas equipes de monitoramento e investigação.

Práticas comuns que devem ser evitadas:

- Diferentes grupos de pessoas investigam descobertas e alertas gerados por sistemas diferentes, a menos que seja exigido de outra forma pelos requisitos de separação de deveres.
- Sua organização canaliza todos os dados de detecção e alerta de segurança para locais padrão, mas exige que os investigadores realizem a correlação e o enriquecimento manualmente.
- Você depende exclusivamente da inteligência dos sistemas de detecção de ameaças para relatar descobertas e determinar a gravidade.

Benefícios de implementar esta prática recomendada: a correlação e o enriquecimento automatizados de alertas ajudam a reduzir a carga cognitiva geral e a preparação manual de dados exigidas de seus investigadores. Essa prática pode reduzir o tempo necessário para determinar se o evento representa um incidente e iniciar uma resposta formal. O contexto adicional também ajuda a avaliar com precisão a verdadeira gravidade de um evento, pois ela pode ser maior ou menor do que o sugerido por qualquer alerta.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Os alertas de segurança podem vir de várias fontes diferentes na AWS, incluindo:

- Serviços como Amazon GuardDuty, AWS Security Hub, Amazon Macie, Amazon Inspector, AWS Config, AWS Identity and Access Management Access Analyzer e Analisador de Acesso à Rede
- Alertas de análises automatizadas de logs de serviços, infraestrutura e aplicações da AWS, como do Security Analytics for Amazon OpenSearch Service.
- Alarmes em resposta a alterações em sua atividade de faturamento de fontes como Amazon CloudWatch, Amazon EventBridge ou AWS Budgets.
- Fontes de terceiros, como feeds de inteligência de ameaças e soluções de parceiros de segurança da AWS Partner Network

- Contato via AWS Trust & Safety ou por outras fontes, como clientes ou funcionários internos.

Em sua forma mais fundamental, os alertas contêm informações sobre quem (a entidade principal ou identidade) está fazendo o quê (a ação tomada) a quê (os recursos afetados). Em cada uma dessas fontes, identifique se há maneiras de criar associações nos identificadores referentes a essas identidades, ações e recursos como base para realizar a correlação. Isso poderia ser integrar fontes de alerta a uma ferramenta de gerenciamento de eventos e informações de segurança (SIEM) para realizar a correlação automatizada para você, criar seus próprios pipelines e processamento de dados ou uma combinação de ambos.

Um exemplo de serviço que pode realizar a correlação para você é o Amazon Detective. O Detective realiza a ingestão contínua de alertas de várias fontes da AWS e de terceiros e usa diferentes formas de inteligência com o objetivo de montar um grafo visual das respectivas relações para auxiliar nas investigações.

Embora a gravidade inicial de um alerta ajude na priorização, o contexto em que o alerta aconteceu determina sua verdadeira gravidade. Como exemplo, o Amazon GuardDuty pode alertar que uma instância do Amazon EC2 em sua workload está consultando um nome de domínio inesperado. O GuardDuty pode atribuir por conta própria uma baixa criticidade a esse alerta. Entretanto, a correlação automatizada com outras atividades em torno do momento do alerta pode revelar que várias centenas de instâncias do EC2 foram implantadas pela mesma identidade, o que aumenta os custos operacionais gerais. Nesse caso, o GuardDuty pode publicar esse contexto de evento correlacionado como um novo alerta de segurança e definir a gravidade como alta, o que agilizaria ações futuras.

Etapas de implementação

1. Identifique fontes de informações sobre alertas de segurança. Entenda como os alertas desses sistemas representam identidade, ação e recursos para determinar onde a correlação é possível.
2. Estabeleça um mecanismo para capturar alertas de diferentes fontes. Considere serviços como Security Hub, EventBridge e CloudWatch para essa finalidade.
3. Identifique fontes para correlação e enriquecimento de dados. Exemplos de fontes incluem CloudTrail, logs de fluxo de VPC, Amazon Security Lake e logs de infraestrutura e aplicações.
4. Integre os alertas às fontes de correlação e enriquecimento de dados para criar contextos de eventos de segurança mais detalhados e determinar a gravidade.
   a. O Amazon Detective, ferramentas de SIEM ou outras soluções de terceiros podem realizar determinado nível de ingestão, correlação e enriquecimento automaticamente.

b. Você também pode usar serviços da AWS para criar seus próprios alertas. Por exemplo, você pode invocar uma função do AWS Lambda para executar uma consulta do Amazon Athena no AWS CloudTrail ou no Amazon Security Lake e publicar os resultados no EventBridge.

Recursos

Práticas recomendadas relacionadas:

- SEC10-BP03 Preparar recursos forenses
- OPS08-BP04 Criar alertas acionáveis
- REL06-BP03 Enviar notificações (processamento e emissão de alarmes em tempo real)

Documentos relacionados:

- Guia de resposta a incidentes de segurança da AWS

Exemplos relacionados:

- Como enriquecer as descobertas do AWS Security Hub com metadados da conta
- Como usar o AWS Security Hub e o Amazon OpenSearch Service para SIEM

Ferramentas relacionadas:

- Amazon Detective
- Amazon EventBridge
- AWS Lambda
- Amazon Athena

SEC04-BP04 Iniciar a correção de recursos fora de conformidade

Seus controles de detecção podem emitir alertas sobre recursos que não estão em conformidade com seus requisitos de configuração. É possível iniciar correções definidas de maneira programática, tanto manual quanto automaticamente, para corrigir esses recursos e ajudar a minimizar possíveis impactos. Definir correções programaticamente permite tomar medidas rápidas e consistentes.

Embora a automação possa aprimorar as operações de segurança, você deve implementá-la e gerenciá-la com cuidado. Estabeleça mecanismos apropriados de supervisão e controle para verificar se as respostas automatizadas são eficazes e precisas e estão alinhadas com as políticas organizacionais e a propensão ao risco.

Resultado desejado: você define os padrões de configuração de recursos junto com as etapas de correção quando os recursos são detectados como fora de conformidade. Sempre que possível, você definiu as correções programaticamente para que elas possam ser iniciadas de modo manual ou por meio de automação. Existem sistemas de detecção para identificar recursos fora de conformidade e publicar alertas em ferramentas centralizadas que são monitoradas por suas equipes de segurança. Essas ferramentas comportam a execução das correções programáticas, tanto manual quanto automaticamente. As correções automáticas têm mecanismos apropriados de supervisão e controle para governar o respectivo uso.

Práticas comuns que devem ser evitadas:

- Você implementa a automação, mas não consegue testar e validar minuciosamente as ações de correção. Isso pode resultar em consequências indesejadas, como interrupção de operações comerciais legítimas ou instabilidade no sistema.
- Você melhora os tempos de resposta e os procedimentos por meio da automação, mas sem monitoramento e mecanismos adequados que permitam a intervenção e avaliação humanas quando necessário.
- Você depende exclusivamente de correções, em vez de tê-las como parte de um programa mais amplo de resposta e recuperação de incidentes.

Benefícios de implementar esta prática recomendada: as correções automáticas podem responder a configurações incorretas mais rapidamente do que os processos manuais, o que ajuda a minimizar possíveis impactos nos negócios e reduzir a janela de oportunidade para usos não intencionais. Quando você define as remediações de forma programática, elas são aplicadas de forma consistente, o que reduz o risco de erro humano. A automação também pode lidar com um volume maior de alertas de forma simultânea, o que é particularmente importante em ambientes que operam em grande escala.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Conforme descrito em SEC01-BP03 Identificar e validar objetivos de controle, serviços como o AWS Config podem ajudar a monitorar a configuração de recursos em suas contas para atender às suas necessidades. Quando recursos fora de conformidade são detectados, recomendamos configurar o envio de alertas para uma solução de gerenciamento de postura de segurança (CSPM) na nuvem, por exemplo, AWS Security Hub, para ajudar na correção do problema. Essas soluções fornecem um local central para os investigadores de segurança monitorarem os problemas e adotarem medidas corretivas.

Embora algumas situações em que há recursos fora de conformidade sejam únicas e exijam avaliação humana para ser corrigidas, outras têm uma resposta padrão que você pode definir programaticamente. Por exemplo, uma resposta padrão a um grupo de segurança da VPC configurado incorretamente pode ser remover as regras não permitidas e notificar o responsável. As respostas podem ser definidas em funções do AWS Lambda, documentos do AWS Systems Manager Automation ou por meio de outros ambientes de código de sua preferência. O ambiente deve estar apto a se autenticar na AWS usando um perfil do IAM com as permissões mínimas necessárias para tomar medidas corretivas.

Depois de definir a remediação desejada, você poderá determinar seus meios preferidos para iniciá-la. O AWS Config pode iniciar remediações para você. Se você estiver usando o Security Hub, poderá fazer isso por meio de ações personalizadas que publicam as informações de descoberta no Amazon EventBridge. Uma regra do EventBridge pode então iniciar a correção. Você pode configurar a ação personalizada no Security Hub para ser executada de forma automática ou manual.

Para remediação programática, recomendamos que manter logs e auditorias abrangentes das ações tomadas, bem como de seus resultados. Revise e analise esses logs para avaliar a eficácia dos processos automatizados e identificar áreas de melhoria. Capture registros no Amazon CloudWatch Logs e resultados de correções como notas de descoberta no Security Hub.

Como ponto de partida, considere a Resposta de Segurança Automatizada na AWS, que oferece correções criadas previamente para resolver configurações incorretas de segurança comuns.

Etapas de implementação

1. Analise e priorize os alertas.
    a. Consolide alertas de segurança de vários serviços da AWS no Security Hub para ter visibilidade, priorização e correção centralizadas.
2. Desenvolva correções.

   a. Use serviços como o Systems Manager e o AWS Lambda para executar correções programáticas.

3. Configure como as correções são iniciadas.
   a. Usando o Systems Manager, defina ações personalizadas para publicar descobertas no EventBridge. Configure essas ações para serem iniciadas manual ou automaticamente.
   b. Também é possível usar o Amazon Simple Notification Service (SNS) para enviar notificações e alertas às partes interessadas relevantes (como equipes de segurança ou equipes de resposta a incidentes) para intervenção manual ou escalação, se necessário.
4. Revise e analise os logs de correção em prol da eficácia e melhoria.
   a. Envie a saída do log ao CloudWatch Logs. Capture resultados como notas de descoberta no Security Hub.

Recursos

Práticas recomendadas relacionadas:

- SEC06-BP03 Reduzir o gerenciamento manual e o acesso interativo

Documentos relacionados:

- Guia de resposta a incidentes de segurança da AWS: Detecção

Exemplos relacionados:

- Resposta de segurança automatizada na AWS
- Monitorar pares de chaves de instância do EC2 com o AWS Config
- Crie regras personalizadas do AWS Config usando políticas do AWS CloudFormation Guard
- Corrija automaticamente instâncias e clusters de banco de dados não criptografados do Amazon RDS

Ferramentas relacionadas:

- AWS Systems Manager Automation
- Resposta de segurança automatizada na AWS

Proteção da infraestrutura

Perguntas

- SEC 5. Como você protege seus recursos de rede?
- SEC 6. Como você protege seus recursos computacionais?

SEC 5. Como você protege seus recursos de rede?

Qualquer workload que tenha alguma forma de conectividade de rede, seja a Internet ou uma rede privada, exige várias camadas de defesa para ajudar a proteger contra ameaças externas e internas baseadas em rede.

Práticas recomendadas

- SEC05-BP01 Criar camadas de rede
- SEC05-BP02 Controlar o fluxo de tráfego dentro das camadas de rede
- SEC05-BP03 Implementar proteção baseada em inspeção
- SEC05-BP04 Automatizar a proteção da rede

SEC05-BP01 Criar camadas de rede

Segmente a topologia de rede em diferentes camadas com base nos agrupamentos lógicos dos componentes da workload e de acordo com a confidencialidade dos dados e os requisitos de acesso. Diferencie os componentes que exigem acesso de entrada pela internet, como endpoints públicos da web e aqueles que precisam apenas de acesso interno, como bancos de dados.

Resultado desejado: as camadas de sua rede fazem parte de uma abordagem integral de defesa aprofundada à segurança que complementa a estratégia de autenticação e autorização de identidade de suas workloads. As camadas estão posicionadas de acordo com a confidencialidade dos dados e os requisitos de acesso, com mecanismos apropriados de fluxo e controle de tráfego.

Práticas comuns que devem ser evitadas:

- Você cria todos os recursos em uma única VPC ou sub-rede.
- Você constrói as camadas de rede sem considerar os requisitos de confidencialidade dos dados, o comportamento dos componentes ou a funcionalidade.
- Você usa VPCs e sub-redes como padrão para todas as considerações de camada de rede e não considera como os serviços gerenciados da AWS influenciam sua topologia.

Benefícios de implementar esta prática recomendada: estabelecer camadas de rede é a primeira etapa para restringir caminhos desnecessários na rede, especialmente aqueles que levam a sistemas e dados críticos. Desse modo, o acesso de agentes não autorizados à sua rede e a outros recursos dentro dela torna-se mais difícil. Camadas de rede discretas reduzem de forma favorável o escopo da análise para sistemas de inspeção, por exemplo, para detecção de intrusões ou prevenção de malware. Isso reduz a possibilidade de falsos positivos e a sobrecarga de processamento desnecessária.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Quando se projeta uma arquitetura de workload, é comum separar os componentes em diferentes camadas com base nas respectivas responsabilidades. Por exemplo, uma aplicação web pode ter uma camada de apresentação, uma camada de aplicação e uma camada de dados. É possível adotar uma abordagem semelhante ao projetar sua topologia de rede. Os controles de rede subjacentes podem ajudar a aplicar os requisitos de acesso aos dados da workload. Por exemplo, em uma arquitetura de aplicação web de três camadas, você pode armazenar seus arquivos de camada de apresentação estática no Amazon S3 e servi-los desde uma rede de entrega de conteúdo (CDN), como o Amazon CloudFront. A camada de aplicação pode ter endpoints públicos que um Application Load Balancer (ALB) serve em uma sub-rede pública da Amazon VPC (semelhante a uma zona desmilitarizada, ou DMZ), com serviços de backend implantados em sub-redes privadas. A camada de dados, que hospeda recursos como bancos de dados e sistemas de arquivos compartilhados, pode residir em diferentes sub-redes privadas dos recursos da camada de aplicação. Em cada um desses limites de camada (CDN, sub-rede pública, sub-rede privada), é possível implantar controles que permitam somente a entrada do tráfego autorizado.

De modo semelhante à modelagem de camadas de rede com base na finalidade funcional dos componentes da workload, considere também a confidencialidade dos dados que estão sendo processados. Usando o exemplo de aplicação web, embora todos os serviços de workload possam residir na camada de aplicação, serviços diferentes podem processar dados com diferentes níveis de confidencialidade. Nesse caso, dividir a camada de aplicação usando várias sub-redes privadas, diferentes VPCs na mesma Conta da AWS ou até mesmo diferentes VPCs em diferentes Contas da AWS para cada nível de confidencialidade de dados pode ser apropriado de acordo com seus requisitos de controle.

Uma consideração adicional sobre as camadas de rede é a consistência do comportamento dos componentes da workload. Continuando com o exemplo, na camada de aplicação, você pode ter serviços que aceitem entradas de usuários finais ou integrações de sistemas externos que são

inerentemente mais arriscadas do que as entradas de outros serviços. Os exemplos incluem uploads de arquivos, scripts de código para execução, verificação de e-mails e assim por diante. Colocar esses serviços em uma camada de rede própria ajuda a criar um limite de isolamento mais forte em torno deles e pode impedir que o comportamento exclusivo de cada um deles crie alertas falsos positivos nos sistemas de inspeção.

Como parte do seu design, considere como o uso de serviços gerenciados da AWS influencia sua topologia de rede. Explore como serviços como o Amazon VPC Lattice podem ajudar a facilitar a interoperabilidade dos componentes da workload nas camadas da rede. Ao usar o AWS Lambda, implante nas sub-redes da VPC, a menos que haja motivos específicos para não fazê-lo. Determine onde a VPC termina e o AWS PrivateLink pode simplificar a adesão às políticas de segurança que limitam o acesso aos gateways da Internet.

Etapas de implementação

1. Revise a arquitetura da sua workload. Agrupe logicamente os componentes e serviços com base nas funções às quais eles atendem, na confidencialidade dos dados que estão sendo processados e no respectivo comportamento.
2. Com relação a componentes que respondem a solicitações da internet, considere usar balanceadores de carga ou outros proxies para fornecer endpoints públicos. Explore mudanças nos controles de segurança usando serviços gerenciados, como CloudFront, Amazon API Gateway, Elastic Load Balancing e AWS Amplify para hospedar endpoints públicos.
3. Para componentes executados em ambientes computacionais, como instâncias do Amazon EC2, contêineres do AWS Fargate ou funções do Lambda, implante-os em sub-redes privadas com base em seus grupos desde a primeira etapa.
4. Para serviços da AWS totalmente gerenciados, como Amazon DynamoDB, Amazon Kinesis ou Amazon SQS, considere usar endpoints da VPC como padrão para acesso por endereços IP privados.

Recursos

Práticas recomendadas relacionadas:

- REL02 Planejar a topologia da rede
- PERF04-BP01 Compreender como a rede afeta a performance

Vídeos relacionados:

- AWS re:Invent 2023: Fundamentos de rede na AWS

Exemplos relacionados:

- Exemplos de VPC
- Acessar aplicações de contêiner de forma privada no Amazon ECS usando o AWS Fargate, o AWS PrivateLink e um Network Load Balancer
- Servir conteúdo estático em um bucket do Amazon S3 por meio de uma VPC usando o Amazon CloudFront

SEC05-BP02 Controlar o fluxo de tráfego dentro das camadas de rede

Dentro das camadas da sua rede, use uma segmentação adicional para restringir o tráfego somente aos fluxos necessários para cada workload. Primeiro, concentre-se em controlar o tráfego entre a Internet ou outros sistemas externos para uma workload e seu ambiente (tráfego norte-sul). Depois, observe os fluxos entre diferentes componentes e sistemas (tráfego leste-oeste).

Resultado desejado: você permite somente os fluxos de rede necessários para que os componentes de suas workloads se comuniquem uns com os outros e com seus clientes e com quaisquer outros serviços dos quais eles dependam. Seu design considera questões como comparação entre entradas e saídas públicas e privadas, classificação de dados, regulamentações regionais e requisitos de protocolo. Sempre que possível, você favorece fluxos ponto a ponto em vez de emparelhamento de rede como parte de um princípio de design de privilégio mínimo.

Práticas comuns que devem ser evitadas:

- Você adota uma abordagem de segurança de rede baseada em perímetro e controla apenas o fluxo de tráfego no limite das camadas de sua rede.
- Você presume que todo o tráfego dentro de uma camada de rede está autenticado e autorizado.
- Você aplica controles para o tráfego de entrada ou de saída, mas não para ambos.
- Você depende exclusivamente dos componentes da workload e dos controles de rede para autenticar e autorizar o tráfego.

Benefícios de implementar esta prática recomendada: essa prática ajuda a reduzir o risco de movimentação não autorizada em sua rede e adiciona uma camada extra de autorização às suas workloads. Ao realizar o controle do fluxo de tráfego, você pode restringir o escopo do impacto de um incidente de segurança e acelerar a detecção e a resposta.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Embora as camadas de rede ajudem a estabelecer os limites em torno de componentes da workload que atendem a funções, níveis de confidencialidade de dados e comportamentos semelhantes, você pode criar um nível de controle de tráfego bem mais refinado usando técnicas para segmentar ainda mais os componentes dentro dessas camadas, seguindo o princípio de privilégio mínimo. Na AWS, as camadas de rede são definidas principalmente via sub-redes de acordo com faixas de endereços IP em uma Amazon VPC. As camadas também podem ser definidas usando diferentes VPCs, como para agrupar ambientes de microsserviços por domínio de negócios. Ao usar várias VPCs, medie o roteamento usando um AWS Transit Gateway. Embora isso forneça controle de tráfego em um nível de camada 4 (endereços IP e intervalos de portas) usando grupos de segurança e tabelas de rotas, você pode obter mais controle usando serviços adicionais como AWS PrivateLink, Firewall de DNS do Amazon Route 53 Resolver, AWS Network Firewall e AWS WAF.

Entenda e faça um inventário do fluxo de dados e dos requisitos de comunicação de workloads em termos de partes que iniciam a conexão, portas, protocolos e camadas de rede. Avalie os protocolos disponíveis para estabelecer conexões e transmitir dados para selecionar aqueles que atendam aos seus requisitos de proteção (por exemplo, HTTPS em vez de HTTP). Capture esses requisitos nos limites de suas redes e dentro de cada camada. Depois que esses requisitos forem identificados, explore as opções para permitir que apenas o tráfego necessário flua em cada ponto de conexão. Um bom ponto de partida é usar grupos de segurança em sua VPC, pois eles podem ser anexados a recursos que usam uma interface de rede elástica (ENI), como instâncias do Amazon EC2, tarefas do Amazon ECS, pods do Amazon EKS ou bancos de dados do Amazon RDS. Ao contrário de um firewall de camada 4, um grupo de segurança pode ter uma regra que permite o tráfego de outro grupo de segurança por meio do respectivo identificador, minimizando as atualizações à medida que os recursos dentro do grupo mudam ao longo do tempo. Você também pode filtrar o tráfego por meio de regras de entrada e saída usando grupos de segurança.

Quando o tráfego se move entre as VPCs, é comum usar o emparelhamento de VPCs para roteamento simples ou o AWS Transit Gateway para roteamento complexo. Com essas abordagens, você facilita os fluxos de tráfego entre o intervalo de endereços IP das redes de origem e de destino. No entanto, se sua workload exigir apenas fluxos de tráfego entre componentes específicos em VPCs diferentes, considere usar uma conexão ponto a ponto usando o AWS PrivateLink. Para isso, identifique qual serviço deve atuar como produtor e qual deve atuar como consumidor. Implante um balanceador de carga compatível para o produtor, ative o PrivateLink adequadamente e, em seguida, aceite uma solicitação de conexão do consumidor. Em seguida, o serviço do produtor recebe um

endereço IP privado da VPC do consumidor que o consumidor pode usar para fazer solicitações subsequentes. Essa abordagem reduz a necessidade de emparelhar as redes. Inclua os custos de processamento de dados e balanceamento de carga como parte da avaliação do PrivateLink.

Embora os grupos de segurança e o PrivateLink ajudem a controlar o fluxo entre os componentes de suas workloads, outra consideração importante é como controlar quais domínios DNS seus recursos podem acessar (se houver). Dependendo da configuração DHCP das suas VPCs, é possível considerar dois serviços da AWS diferentes para essa finalidade. A maioria dos clientes usa o serviço DNS padrão do Route 53 Resolver (também chamado de servidor Amazon DNS ou AmazonProvideDDNS) disponível para VPCs no endereço +2 de seu intervalo CIDR. Com essa abordagem, é possível criar regras de firewall de DNS e associá-las à VPC para determinar quais ações devem ser realizadas para as listas de domínios que você fornece.

Se você não estiver usando o Route 53 Resolver, ou se quiser complementar o Resolver com recursos mais detalhados de inspeção e controle de fluxo, além da filtragem de domínio, considere implantar um AWS Network Firewall. Esse serviço inspeciona pacotes individuais usando regras sem estado ou com estado para determinar se deve negar ou permitir o tráfego. Você pode adotar uma abordagem semelhante para filtrar o tráfego de entrada da web para seus endpoints públicos usando o AWS WAF. Para obter mais orientações sobre esses serviços, consulte SEC05-BP03 Implementar proteção baseada em inspeção.

Etapas de implementação

1. Identifique os fluxos de dados necessários entre os componentes das workloads.
2. Aplique vários controles com uma abordagem de defesa profunda para tráfego de entrada e saída, incluindo o uso de grupos de segurança e tabelas de rotas.
3. Use firewalls para definir um controle refinado sobre o tráfego de rede que entra, sai e atravessa suas VPCs, como o Firewall de DNS do Route 53 Resolver, o AWS Network Firewall e o AWS WAF. Considere usar o AWS Firewall Manager para configurar e gerenciar centralmente suas regras de firewall em toda a organização.

Recursos

Práticas recomendadas relacionadas:

- REL03-BP01 Escolher como segmentar a workload
- SEC09-BP02 Impor a criptografia em trânsito

Documentos relacionados:

- Práticas recomendadas de segurança para a VPC
- Dicas de otimização de rede da AWS
- Orientação para segurança de rede na AWS
- Proteger o tráfego de rede de saída da sua VPC na Nuvem AWS

Ferramentas relacionadas:

- AWS Firewall Manager

Vídeos relacionados:

- Arquiteturas de referência do AWS Transit Gateway para muitas VPCs
- Aceleração e proteção de aplicações com Amazon CloudFront, AWS WAF e AWS Shield
- AWS re:Inforce 2023: Firewalls e onde colocá-los

Exemplos relacionados:

- Laboratório: CloudFront para aplicações Web

SEC05-BP03 Implementar proteção baseada em inspeção

Configure pontos de inspeção de tráfego entre as camadas de rede para garantir que os dados em trânsito correspondam aos padrões e categorias esperados. Analise padrões, metadados e fluxos de tráfego para ajudar a identificar, detectar e responder a eventos com maior eficiência.

Resultado desejado: o tráfego que passa entre suas camadas de rede é inspecionado e autorizado. As decisões de permissão e negação baseiam-se em regras explícitas, inteligência contra ameaças e desvios dos comportamentos de referência. As proteções tornam-se mais rígidas à medida que o tráfego aproxima-se dos dados confidenciais.

Práticas comuns que devem ser evitadas:

- Confiar somente em regras de firewall baseadas em portas e protocolos. Não aproveitar os sistemas inteligentes.

- Criar regras de firewall com base em padrões específicos de ameaças atuais que estão sujeitos a alterações.
- Inspecionar somente o tráfego que transita de sub-redes privadas para públicas ou de sub-redes públicas para a internet.
- Não ter uma visão de referência do tráfego da rede para comparar com anomalias de comportamento.

Benefícios de implementar esta prática recomendada: os sistemas de inspeção permitem que você crie regras inteligentes, como permitir ou negar tráfego somente quando houver determinadas condições nos dados de tráfego. Beneficie-se de conjuntos de regras gerenciados pela AWS e por parceiros, com base na inteligência contra ameaças mais recente, à medida que o cenário de ameaças muda ao longo do tempo. Isso reduz as despesas indiretas de manter regras e pesquisar indicadores de comprometimento, reduzindo o potencial de falsos positivos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Mantenha um controle refinado sobre seu tráfego de rede com e sem estado usando o AWS Network Firewall ou outros firewalls e sistemas de prevenção de intrusões (IPS) no AWS Marketplace você pode implantar por trás de um Gateway Load Balancer (GWLB). O AWS Network Firewall oferece suporte a especificações IPS de código aberto compatíveis com Suricata para ajudar a proteger sua workload.

Tanto o AWS Network Firewall quanto as soluções de fornecedor que usam um GWLB comportam diferentes modelos de implantação de inspeção em linha. Por exemplo, você pode realizar a inspeção por VPC, centralizar em uma VPC de inspeção ou implantar em um modelo híbrido em que o tráfego leste-oeste flui por meio de uma VPC de inspeção e a entrada da internet é inspecionada por VPC. Outra consideração é se a solução comporta o desempacotamento do Transport Layer Security (TLS), permitindo a inspeção detalhada de pacotes para fluxos de tráfego iniciados em qualquer direção. Para obter mais informações e detalhes sobre essas configurações, consulte o Guia de práticas recomendadas do AWS Network Firewall.

Se você usa soluções que realizam inspeções fora de banda, como análise pcap de dados de pacotes de interfaces de rede operando em modo promíscuo, é possível configurar o espelhamento de tráfego de VPC. O tráfego espelhado é computado na largura de banda disponível de suas interfaces e está sujeito às mesmas cobranças de transferência de dados que o tráfego não

espelhado. É possível ver se as versões virtuais desses appliances estão disponíveis no AWS Marketplace, o que pode oferecer suporte à implantação em linha por trás de um GWLB.

Para componentes que operam por meio de protocolos baseados em HTTP, proteja sua aplicação contra ameaças comuns com um firewall de aplicações Web (WAF). O AWS WAF é um firewall de aplicações Web que permite monitorar e bloquear solicitações HTTP(S) que correspondem a suas regras configuráveis antes de enviar para o Amazon API Gateway, o Amazon CloudFront, o AWS AppSync ou um Application Load Balancer. Considere a inspeção detalhada de pacotes ao avaliar a implantação do firewall de aplicações Web, pois alguns exigem que você encerre o TLS antes da inspeção de tráfego. Para começar a usar o AWS WAF, é possível usar AWS Managed Rules em combinação com as suas próprias regras ou usar as integrações de parceiros existentes.

Você pode gerenciar centralmente o AWS WAF, o AWS Shield Advanced, o AWS Network Firewall e os grupos de segurança da Amazon VPC em toda a sua organização da AWS com o AWS Firewall Manager.

Etapas de implementação

1. Determine se você pode definir um escopo amplo das regras de inspeção, como por meio de uma VPC de inspeção, ou se precisa de uma abordagem mais detalhada por VPC.
2. Para soluções de inspeção em linha:
    a. Se estiver usando o AWS Network Firewall, crie regras, políticas de firewall e o próprio firewall. Após a configuração, você poderá rotear o tráfego para o endpoint do firewall para permitir a inspeção.
    b. Se estiver usando um appliance de terceiros com um Gateway Load Balancer (GWLB), implante e configure seu appliance em uma ou mais zonas de disponibilidade. Em seguida, crie o GWLB, o serviço de endpoint e o endpoint e configure o roteamento para o tráfego.
3. Para soluções de inspeção fora de banda:
    1. Ative o espelhamento de tráfego da VPC em interfaces nas quais o tráfego de entrada e saída deve ser espelhado. É possível usar regras do Amazon EventBridge para invocar uma função do AWS Lambda que ative o espelhamento de tráfego em interfaces quando são criados recursos. Aponte as sessões de espelhamento de tráfego para o Network Load Balancer na frente do appliance que processa o tráfego.
4. Para soluções de tráfego da Web de entrada:
    a. Para configurar o AWS WAF, primeiro configure uma lista de controle de acesso à web (ACL da web). A ACL da web é um conjunto de regras com uma ação padrão processada em série

(permitir ou negar) que define como o WAF lida com o tráfego. Você pode criar seus próprios grupos e regras ou usar grupos de regras gerenciadas da AWS em sua ACL da Web.

b. Assim que a ACL da Web for configurada, ela poderá ser associada a um recurso da AWS (como um Application Load Balancer, uma API REST do API Gateway ou uma distribuição do CloudFront) para começar a proteger o tráfego da Web.

Recursos

Documentos relacionados:

- O que é espelhamento de tráfego?
- Implementar a inspeção de tráfego em linha usando appliances de segurança de terceiros
- Exemplos de arquiteturas do AWS Network Firewall com roteamento
- Arquitetura de inspeção centralizada com o AWS Gateway Balancer e o AWS Transit Gateway

Exemplos relacionados:

- Práticas recomendadas para implantar o balanceador de carga do gateway
- Configuração de inspeção TLS para tráfego de saída criptografado e AWS Network Firewall

Ferramentas relacionadas:

- AWS Marketplace IDS/IPS

SEC05-BP04 Automatizar a proteção da rede

Automatize a implantação de suas proteções de rede usando práticas de DevOps, como infraestrutura como código (IaC) e pipelines de CI/CD. Essas práticas podem ajudar você a monitorar alterações nas proteções da rede por meio de um sistema de controle de versão, reduzir o tempo necessário para implantar alterações e detectar se as proteções de rede se desviam da configuração desejada.

Resultado desejado: você define as proteções de rede com modelos e as compromete em um sistema de controle de versão. Quando novas alterações são feitas, os pipelines automatizados são iniciados para orquestrar os respectivos testes e a implantação. Verificações de políticas e outros testes estáticos estão em vigor para validar as alterações antes da implantação. Você implanta as

alterações em um ambiente de preparação para validar se os controles estão operando conforme o esperado. A implantação nos ambientes de produção também é executada automaticamente quando os controles são aprovados.

Práticas comuns que devem ser evitadas:

- Contar com equipes de workload individuais para que definam sua pilha de rede completa, proteções e automações. Não publicar os aspectos padrão da pilha de rede e das proteções de maneira centralizada para as equipes de workload consumirem.
- Contar com uma equipe de rede central para definir todos os aspectos da rede, proteções e automações. Não delegar aspectos específicos da workload da pilha de rede e das proteções à equipe da workload em questão.
- Conseguir o equilíbrio certo de centralização e delegação entre a equipe de rede e as equipes de workload, mas não aplicar padrões consistentes de teste e implantação aos modelos de IaC e pipelines de CI/CD. Não capturar as configurações necessárias em ferramentas que verificam a aderência aos modelos.

Benefícios de implementar esta prática recomendada: o uso de modelos para definir suas proteções de rede permite rastrear e comparar as alterações ao longo do tempo com um sistema de controle de versão. Usar automação para testar e implantar alterações gera padronização e previsibilidade, aumentando as chances de uma implantação bem-sucedida e reduzindo configurações manuais repetitivas.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Vários controles de proteção de rede descritos em SEC05-BP02 Controlar fluxos de tráfego em suas camadas de rede e SEC05-BP03 Implementar a proteção baseada em inspeção são fornecidos com sistemas de regras gerenciados que podem ser atualizados automaticamente com base nas informações mais recentes sobre ameaças. Exemplos de proteção de seus endpoints da Web incluem regras gerenciadas pelo AWS WAF e mitigação de DDoS na camada de aplicações automática do AWS Shield Advanced. Use grupos de regras gerenciados pelo AWS Network Firewall para se manter atualizado com listas de domínios de baixa reputação e assinaturas de ameaças.

Além das regras gerenciadas, recomendamos usar práticas de DevOps para automatizar a implantação dos recursos de rede, das proteções e das regras que você especificar. Você pode capturar essas definições no AWS CloudFormation ou em outra ferramenta de infraestrutura como

código (IaC) de sua escolha, confirmá-las em um sistema de controle de versão e implantá-las usando pipelines de CI/CD. Use essa abordagem para obter os benefícios tradicionais de DevOps para gerenciar seus controles de rede, como lançamentos mais previsíveis, testes automatizados usando ferramentas como o AWS CloudFormation Guarde detecção de desvios entre o ambiente implantado e a configuração desejada.

Com base nas decisões tomadas como parte de SEC05-BP01 Criar camadas de rede, você pode adotar uma abordagem de gerenciamento central para criar VPCs dedicadas aos fluxos de entrada, saída e inspeção. Conforme descrito na Arquitetura de referência de segurança da AWS (AWS SRA), é possível definir essas VPCs em uma conta de infraestrutura de rede dedicada. Você pode usar técnicas semelhantes para definir centralmente as VPCs usadas pelas workloads em outras contas, os respectivos grupos de segurança, as implantações do AWS Network Firewall, as regras do Route 53 Resolver, as configurações do firewall de DNS e outros recursos de rede. Você pode compartilhar esses recursos com suas outras contas com o AWS Resource Access Manager. Com essa abordagem, é possível simplificar os testes automatizados e a implantação dos controles de rede na conta de rede, o que resulta em apenas um destino para gerenciar. É possível fazer isso em um modelo híbrido no qual você implanta e compartilha determinados controles centralmente e delega outros controles às equipes individuais de workload e respectivas contas.

Etapas de implementação

1. Estabeleça a propriedade para definir quais aspectos da rede e das proteções são definidos centralmente e quais as equipes de workload podem manter.
2. Crie ambientes para testar e implantar alterações na rede e nas respectivas proteções. Por exemplo, use uma conta de teste de rede e uma conta de produção de rede.
3. Determine como você armazenará e manterá os modelos em um sistema de controle de versão. Os modelos centrais podem ser armazenados em um repositório diferente dos repositórios de workload, enquanto os modelos de workload podem ser armazenados em repositórios específicos para essa workload.
4. Crie pipelines de CI/CD para testar e implantar modelos. Defina testes para verificar se há configurações incorretas e se os modelos estão de acordo com os padrões da sua empresa.

Recursos

Práticas recomendadas relacionadas:

- SEC01-BP06 Automatizar a implantação de controles de segurança padrão

Documentos relacionados:

- Arquitetura de referência de segurança da AWS: Conta de rede

Exemplos relacionados:

- Arquitetura de referência do pipeline de implantação da AWS
- NetDevSecOps para modernizar as implantações de rede da AWS
- Integrar testes de segurança do AWS CloudFormation com relatórios do AWS Security Hub e do AWS CodeBuild

Ferramentas relacionadas:

- AWS CloudFormation
- AWS CloudFormation Guard
- cfn_nag

SEC 6. Como você protege seus recursos computacionais?

Os recursos computacionais em seu workload exigem várias camadas de defesa para ajudar na proteção contra ameaças externas e internas. Os recursos de computação incluem instâncias do EC2, contêineres, funções do AWS Lambda, serviços de banco de dados, dispositivos de IoT e muito mais.

Práticas recomendadas

- SEC06-BP01 Realizar o gerenciamento de vulnerabilidades
- SEC06-BP02 Provisionar computação com base em imagens reforçadas
- SEC06-BP03 Reduzir o gerenciamento manual e o acesso interativo
- SEC06-BP04 Validar a integridade do software
- SEC06-BP05 Automatizar a proteção da computação

SEC06-BP01 Realizar o gerenciamento de vulnerabilidades

Verifique e corrija com frequência vulnerabilidades no código, nas dependências e na infraestrutura para se proteger contra novas ameaças.

Resultado desejado: crie e mantenha um programa de gerenciamento de vulnerabilidades. Examine e corrija regularmente recursos como instâncias do Amazon EC2, contêineres do Amazon Elastic Container Service (Amazon ECS) e workloads do Amazon Elastic Kubernetes Service (Amazon EKS). Configure janelas de manutenção para recursos gerenciados pela AWS, como bancos de dados do Amazon Relational Database Service (Amazon RDS). Use a verificação de código estático para inspecionar a existência de problemas comuns no código-fonte da aplicação. Considere testes de penetração de aplicações da Web se sua organização tiver as habilidades obrigatórias ou puder contratar assistência externa.

Práticas comuns que devem ser evitadas:

- Não ter um programa de gerenciamento de vulnerabilidades.
- Realizar a aplicação de patches do sistema sem considerar a gravidade ou formas de evitar riscos.
- Utilizar software que ultrapassou a data de fim de vida útil (EOL) indicada pelo fornecedor.
- Implantar código em produção antes de analisar a existência de problemas de segurança.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Um programa de gerenciamento de vulnerabilidades inclui avaliação de segurança, identificação de problemas, priorização e realização de operações de patch como parte da solução dos problemas. A automação é a chave para verificar de forma contínua as workloads quanto a problemas e exposição acidental à rede e realização de remediação. Automatizar a criação e atualizar os recursos economiza tempo e reduz o risco de erros de configuração que criam mais problemas. Um programa de gerenciamento de vulnerabilidades bem projetado também deve considerar testes de vulnerabilidades durante o desenvolvimento e os estágios de implantação do ciclo de vida do software. Implementar o gerenciamento de vulnerabilidades durante o desenvolvimento e a implantação ajuda a reduzir a chance de uma vulnerabilidade atingir seu ambiente de produção.

A implementação de um programa de gerenciamento de vulnerabilidades exige uma boa compreensão do modelo de responsabilidade compartilhada da AWS e de como ele se relaciona com suas workloads específicas. Segundo o modelo de responsabilidade compartilhada, a AWS é responsável por proteger a infraestrutura da Nuvem AWS. Essa infraestrutura é composta por hardware, software, redes e instalações que executam os serviços da Nuvem AWS. Você é responsável pela segurança na nuvem, por exemplo, os dados reais, a configuração de segurança e as tarefas de gerenciamento de instâncias do Amazon EC2 e por garantir que seus objetos do

Amazon S3 sejam classificados e configurados corretamente. Sua abordagem ao gerenciamento de vulnerabilidades também pode variar dependendo dos serviços consumidos. Por exemplo, a AWS gerencia a aplicação de patches para nosso serviço de banco de dados relacional gerenciado, o Amazon RDS, mas você seria responsável pela colocação de patches em bancos de dados auto-hospedados.

A AWS oferece diversos serviços para ajudar em seu programa de gerenciamento de vulnerabilidades. O Amazon Inspector verifica continuamente as workloads da AWS em busca de problemas de software e acesso não intencional à rede. AWS Systems Manager O Patch Manager ajuda a gerenciar a aplicação de patches em suas instâncias do Amazon EC2. O Amazon Inspector e o Systems Manager podem ser visualizados no AWS Security Hub, um serviço de gerenciamento de postura de segurança na nuvem que ajuda a automatizar as verificações de segurança da AWS e centralizar os alertas de segurança.

O Amazon CodeGuru pode ajudar a identificar possíveis problemas em aplicações Java e Python usando análise estática de código.

Etapas de implementação

- Configure o Amazon Inspector: o Amazon Inspector detecta automaticamente instâncias do Amazon EC2 recém-lançadas, funções do Lambda e imagens de contêineres elegíveis enviadas ao Amazon ECR e as examina imediatamente em busca de problemas de software, defeitos potenciais e exposição não intencional na rede.
- Examine o código-fonte: verifique bibliotecas e dependências em busca de problemas e defeitos. O Amazon CodeGuru pode examinar e fornecer recomendações para corrigir problemas comuns de segurança em aplicações Java e Python. A OWASP Foundation publica uma lista de ferramentas de análise de código-fonte (também conhecidas como ferramentas SAST).
- Implemente um mecanismo para verificar e corrigir seu ambiente existente, bem como fazer a verificação como parte de um processo de criação de pipeline de CI/CD: implemente um mecanismo para verificar e corrigir problemas em suas dependências e sistemas operacionais para ajudar a se proteger contra novas ameaças. Execute esse mecanismo regularmente. O gerenciamento de vulnerabilidades de software é essencial para entender onde é necessário aplicar patches ou resolver problemas de software. Priorize a correção de possíveis problemas de segurança incorporando avaliações de vulnerabilidade no início de seu pipeline de integração/entrega contínua (CI/CD). Sua abordagem pode variar com base nos serviços da AWS que você está consumindo. Para verificar possíveis problemas no software executado em instâncias do Amazon EC2, adicione o Amazon Inspector ao seu pipeline para ser alertado e para interromper o processo de compilação se problemas ou defeitos potenciais forem detectados. O Amazon

Inspector monitora continuamente os recursos. Você também pode utilizar produtos de código aberto, como OWASP Dependency-Check, Snyk, OpenVAS, gerenciadores de pacotes e ferramentas de AWS Partner para gerenciamento de vulnerabilidades.

- Use o AWS Systems Manager: você é responsável pelo gerenciamento de patches para seus recursos da AWS, incluindo instâncias do Amazon Elastic Compute Cloud (Amazon EC2), imagens de máquina da Amazon (AMIs) e outros recursos de computação. AWS Systems Manager O Patch Manager automatiza o processo de aplicação de patches em instâncias gerenciadas com atualizações relacionadas à segurança e outros tipos de atualizações. O Patch Manager pode ser utilizado para aplicar patches em instâncias do Amazon EC2 para sistemas operacionais e aplicações, como aplicações da Microsoft, pacotes de serviços Windows e atualizações de versão secundária para instâncias baseadas em Linux. Além do Amazon EC2, o Patch Manager também pode ser utilizado para aplicar patches em servidores on-premises.

  Para obter uma lista de sistemas operacionais compatíveis, consulte Sistemas operacionais compatíveis no Guia do usuário do Systems Manager. Você pode verificar instâncias para visualizar somente um relatório de patches ausentes ou verificar e instalar automaticamente todos os patches ausentes.
- Use o AWS Security Hub: o Security Hub fornece uma visualização abrangente de seu estado de segurança na AWS. Ele coleta dados de segurança em vários serviços da AWS e fornece essas descobertas em um formato padronizado, permitindo que você priorize as descobertas de segurança em todos os serviços da AWS.
- Use o AWS CloudFormation: o AWS CloudFormation é um serviço de infraestrutura como código (IaC) que pode ajudar no gerenciamento de vulnerabilidades automatizando a implantação de recursos e padronizando a arquitetura de recursos em várias contas e ambientes.

Recursos

Documentos relacionados:

- AWS Systems Manager
- Visão geral da segurança do AWS Lambda
- Amazon CodeGuru
- Gerenciamento de vulnerabilidades aprimorado e automatizado para workloads na nuvem com um novo Amazon Inspector
- Automatize o gerenciamento e a correção de vulnerabilidades na AWS usando o Amazon Inspector e o AWS Systems Manager: parte 1

Vídeos relacionados:

- Proteger serviços com tecnologia sem servidor e em contêineres
- Práticas recomendadas de segurança para o serviço de metadados da instância do Amazon EC2

SEC06-BP02 Provisionar computação com base em imagens reforçadas

Ofereça menos oportunidades de acesso indesejado aos ambientes de runtime implantando-os com base em imagens reforçadas. Adquira somente dependências de runtime, como imagens de contêiner e bibliotecas de aplicações, de registros confiáveis e verifique as respectivas assinaturas. Crie seus próprios registros privados para armazenar imagens e bibliotecas confiáveis para uso nos processos de criação e implantação.

Resultado desejado: seus recursos computacionais são provisionados a partir de imagens de referência reforçadas. Você recupera dependências externas, como imagens de contêiner e bibliotecas de aplicações, somente de registros confiáveis e verifica as respectivas assinaturas. Elas são armazenadas em registros privados para que seus processos de compilação e implantação as consultem. Você verifica e atualiza imagens e dependências regularmente para ajudar a oferecer proteção contra qualquer vulnerabilidade recém-descoberta.

Práticas comuns que devem ser evitadas:

- Adquirir imagens e bibliotecas de registros confiáveis, mas não verificar a respectiva assinatura nem realizar verificações de vulnerabilidades antes de colocá-las em uso.
- Reforçar as imagens, mas não testá-las regularmente em busca de novas vulnerabilidades ou atualizá-las para a versão mais recente.
- Instalar ou não remover pacotes de software que não são necessários durante o ciclo de vida previsto da imagem.
- Confiar apenas na aplicação de patches para manter os recursos de computação de produção atualizados. Ao utilizar apenas a aplicação de patches, os recursos de computação podem se desviar do padrão reforçado com o passar do tempo. A aplicação de patches também pode não conseguir remover malware instalado por um agente de ameaças durante um evento de segurança.

Benefícios de implementar esta prática recomendada: o reforço de imagens ajuda a reduzir o número de caminhos disponíveis em seu ambiente de runtime que podem permitir acesso não intencional

a usuários ou serviços não autorizados. Ele também pode reduzir o escopo do impacto caso ocorra algum acesso indesejado.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Para reforçar seus sistemas, comece com as versões mais recentes de sistemas operacionais, imagens de contêiner e bibliotecas de aplicações. Aplique patches aos problemas conhecidos. Minimize o sistema removendo quaisquer aplicações, serviços, drivers de dispositivo, usuários padrão e outras credenciais desnecessários. Execute qualquer outra ação necessária, como desabilitar portas para criar um ambiente que tenha somente os recursos e capacidades essenciais para as workloads. Com base nesse parâmetro, você pode instalar software, agentes ou outros processos necessários para finalidades como monitoramento da workload ou gerenciamento de vulnerabilidades.

É possível reduzir a carga de reforçar os sistemas usando orientações fornecidas por fontes confiáveis, como o Center for Internet Security (CIS) e os Guias de implementação técnica de segurança (STIGs) da Defense Information Systems Agency (DISA). Recomendamos começar com uma imagem de máquina da Amazon (AMI) publicada pela AWS ou um parceiro da APN e use o AWS EC2 Image Builder para automatizar a configuração de acordo com uma combinação apropriada de controles CIS e STIG.

Embora existam imagens reforçadas e fórmulas do EC2 Image Builder disponíveis que aplicam as recomendações do CIS ou do STIG da DISA, talvez você veja que sua configuração impede que seu software seja executado com êxito. Nessa situação, você pode começar com uma imagem base não reforçada, instalar o software e, em seguida, aplicar incrementalmente os controles do CIS para testar o respectivo impacto. Com relação a qualquer controle do CIS que impeça a execução do software, teste se é possível implementar as recomendações de fortalecimento mais refinadas em um STIG da DISA. Acompanhe os diferentes controles do CIS e as configurações do STIG da DISA que você pode aplicar com sucesso. Use-os para definir adequadamente suas fórmulas de reforço de imagem no EC2 Image Builder.

Para workloads em contêineres, imagens reforçadas do Docker estão disponíveis no repositório público do Amazon Elastic Container Registry (ECR). Você pode usar o EC2 Image Builder para reforçar imagens de contêiner, bem como AMIs.

Semelhante aos sistemas operacionais e às imagens de contêiner, você pode obter pacotes de código (ou bibliotecas) de repositórios públicos por meio de ferramentas como pip, npm, Maven e

NuGet. Recomendamos gerenciar pacotes de código integrando repositórios privados, como os do AWS CodeArtifact, a repositórios públicos confiáveis. Com essa integração, você não precisa se preocupar em lidar com a recuperação, o armazenamento e a manutenção de pacotes atualizados. Seus processos de criação de aplicações podem então obter e testar a versão mais recente desses pacotes, bem como a aplicação, usando técnicas como análise de composição de software (SCA), testes estáticos de segurança de aplicações (SAST) e testes dinâmicos de segurança de aplicações (DAST).

Para workloads sem servidor que usam o AWS Lambda, simplifique o gerenciamento de dependências de pacotes usando camadas do Lambda. Use camadas do Lambda para configurar um conjunto de dependências padrão que são compartilhadas em diferentes funções em um arquivo independente. Você pode criar e manter camadas por meio de seu próprio processo de criação, fornecendo um meio centralizado para manter as funções atualizadas.

Etapas de implementação

- Reforce os sistemas operacionais. Use imagens básicas de fontes confiáveis como base para criar AMIs reforçadas. Use o EC2 Image Builder para ajudar a personalizar o software instalado em suas imagens.
- Reforce os recursos em contêineres. Configure recursos em contêineres para atender a práticas recomendadas de segurança. Ao usar contêineres, implemente a varredura de imagens do ECR no pipeline de compilação e regularmente no repositório de imagens para procurar CVEs nos contêineres.
- Ao usar a implementação sem servidor com o AWS Lambda, use camadas do Lambda para separar o código da função da aplicação e as bibliotecas dependentes compartilhadas. Configure a assinatura de código para Lambda para garantir que apenas código confiável seja executado em suas funções do Lambda.

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP05 Realizar o gerenciamento de patches

Vídeos relacionados:

- Mergulho profundo na segurança do AWS Lambda

Exemplos relacionados:

- Criar rapidamente uma AMI compatível com STIG usando o EC2 Image Builder
- Criar imagens de contêiner melhores
- Usar camadas do Lambda para simplificar seu processo de desenvolvimento
- Desenvolver e implantar camadas do AWS Lambda usando um framework sem servidor
- Criar um pipeline de CI/CD completo do AWS DevSecOps com ferramentas de código aberto SCA, SAST e DAST

SEC06-BP03 Reduzir o gerenciamento manual e o acesso interativo

Use a automação para realizar tarefas de implantação, configuração, manutenção e investigação sempre que possível. Considere usar o acesso manual aos recursos de computação em casos de procedimentos de emergência ou em ambientes seguros (sandbox) quando a automação não estiver disponível.

Resultado desejado: scripts programáticos e documentos de automação (runbooks) capturam ações autorizadas em seus recursos computacionais. Os runbooks são iniciados automaticamente por meio de sistemas de detecção de alterações ou manualmente quando a avaliação humana é necessária. O acesso direto aos recursos de computação só é disponibilizado em situações de emergência quando a automação não está disponível. Todas as atividades manuais são registradas em log e incorporadas a um processo de análise para aprimorar continuamente os recursos de automação.

Práticas comuns que devem ser evitadas:

- Usar o acesso interativo a instâncias do Amazon EC2 com protocolos como SSH ou RDP.
- Manter logins de usuários individuais, como `/etc/passwd` ou usuários locais do Windows.
- Compartilhar uma senha ou chave privada para acessar uma instância entre vários usuários.
- Instalar software e criar ou atualizar manualmente arquivos de configuração.
- Atualizar ou aplicar patches manualmente no software.
- Fazer login em uma instância para solucionar problemas.

Benefícios de implementar esta prática recomendada: a execução de ações com automação ajuda a reduzir o risco operacional de alterações não intencionais e configurações incorretas. Eliminar o uso do Secure Shell (SSH) e do Remote Desktop Protocol (RDP) para acesso interativo reduz o escopo do acesso aos seus recursos de computação. Fazer isso elimina um caminho comum para

ações não autorizadas. Capturar suas tarefas de gerenciamento de recursos de computação em documentos de automação e scripts programáticos oferece um mecanismo para definir e auditar todo o escopo das atividades autorizadas em um nível de detalhes refinado.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Fazer login em uma instância é uma abordagem clássica à administração de sistemas. Após a instalação do sistema operacional do servidor, os usuários normalmente fazem login manualmente para configurar o sistema e instalar o software desejado. Durante o ciclo de vida do servidor, os usuários podem fazer login para realizar atualizações de software, aplicar patches, alterar configurações e solucionar problemas.

No entanto, o acesso manual apresenta vários riscos. Ele exige um servidor que escuta solicitações, como um serviço SSH ou RDP, o que pode fornecer um possível caminho para acessos não autorizados. Ele também aumenta o risco de erro humano associado à execução de etapas manuais. Isso pode resultar em incidentes de workload, corrompimento ou destruição de dados ou outros problemas de segurança. O acesso humano também exige proteções contra o compartilhamento de credenciais, o que cria uma sobrecarga adicional de gerenciamento.

Para mitigar esses riscos, você pode implementar uma solução de acesso remoto baseada em agente, como o AWS Systems Manager. AWS Systems Manager O agente (SSM Agent) inicia um canal criptografado e, portanto, não depende da escuta de solicitações iniciadas externamente. Considere configurar o SSM Agent para estabelecer esse canal em um endpoint da VPC.

O Systems Manager oferece controle refinado sobre como você pode interagir com suas instâncias gerenciadas. Você define as automações a serem executadas, quem pode executá-las e quando elas podem ser executadas. O Systems Manager pode aplicar patches, instalar software e fazer alterações na configuração sem ter acesso interativo à instância. O Systems Manager também pode fornecer acesso a um shell remoto e registrar cada comando invocado e sua saída durante a sessão nos logs e no Amazon S3. O AWS CloudTrail registra as invocações das APIs do Systems Manager para inspeção.

Etapas de implementação

1. Instale o AWS Systems Manager Agent (SSM Agent) nas suas instâncias do Amazon EC2. Verifique se o SSM Agent está incluído e foi iniciado automaticamente como parte da configuração básica da AMI.

2. Verifique se as funções do IAM associadas aos seus perfis de instância do EC2 incluem a política `AmazonSSMManagedInstanceCore` gerenciada pelo IAM.
3. Desabilite o SSH, o RDP e outros serviços de acesso remoto em execução nas instâncias. Você pode fazer isso executando scripts configurados na seção de dados do usuário dos seus modelos de lançamento ou criando AMIs personalizadas com ferramentas como o EC2 Image Builder.
4. Verifique se as regras de entrada do grupo de segurança aplicáveis às instâncias do EC2 não permitem acesso na porta 22/tcp (SSH) ou na porta 3389/tcp (RDP). Implemente a detecção e o alerta de grupos de segurança configurados incorretamente usando serviços como o AWS Config.
5. Defina automações, runbooks e comandos de execução apropriados no Systems Manager. Use políticas do IAM para definir quem pode realizar essas ações e as condições sob as quais elas são permitidas. Teste essas automações minuciosamente em um ambiente de não produção. Invoque essas automações quando necessário, em vez de acessar a instância de forma interativa.
6. Use o AWS Systems Manager Session Manager para fornecer acesso interativo às instâncias quando necessário. Ative o log de atividades da sessão para manter uma trilha de auditoria no Amazon CloudWatch Logs ou no Amazon S3.

Recursos

Práticas recomendadas relacionadas:

- REL08-BP04 Implantar usando infraestrutura imutável

Exemplos relacionados:

- Substituir o acesso SSH para reduzir a sobrecarga de gerenciamento e segurança com o AWS Systems Manager

Ferramentas relacionadas:

- AWS Systems Manager

Vídeos relacionados:

- Controlar o acesso da sessão do usuário à instâncias no Gerenciador de Sessões do AWS Systems Manager

SEC06-BP04 Validar a integridade do software

Use a verificação criptográfica para validar a integridade dos artefatos de software (incluindo imagens) que a workload usa. Assine criptograficamente seu software como uma proteção contra alterações não autorizadas executadas em seus ambientes de computação.

Resultado desejado: todos os artefatos são obtidos de fontes confiáveis. Os certificados do site do fornecedor são validados. Os artefatos baixados são verificados criptograficamente com base na respectiva assinatura. Seu próprio software é assinado e verificado criptograficamente por seus ambientes de computação.

Práticas comuns que devem ser evitadas:

- Confiar em sites de fornecedores de boa reputação para obter artefatos de software, mas ignorar os avisos de expiração de certificado. Prosseguir com os downloads sem confirmar se os certificados são válidos.
- Validar certificados de sites de fornecedores, mas não verificar criptograficamente os artefatos baixados desses sites.
- Confiar apenas em resumos ou hashes para validar a integridade do software. Os hashes estabelecem que os artefatos não foram modificados da versão original, mas não validam a respectiva fonte.
- Não assinar seu próprio software, código ou biblioteca, mesmo quando usados apenas em suas próprias implantações.

Benefícios de implementar esta prática recomendada: validar a integridade dos artefatos dos quais sua workload depende ajuda a impedir a entrada de malware em seus ambientes computacionais. Assinar seu software ajuda a impedir a execução não autorizada em seus ambientes de computação. Proteja sua cadeia de suprimentos de software assinando e verificando o código.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

As imagens do sistema operacional, as imagens de contêiner e os artefatos de código geralmente são distribuídos com verificações de integridade disponíveis, como por meio de um resumo ou hash. Isso permite que os clientes verifiquem a integridade calculando o hash da carga útil e validando se ele é o mesmo que o publicado. Embora essas verificações ajudem a verificar se a carga não foi adulterada, elas não validam que a carga veio da fonte original (sua procedência). A verificação

de procedência exige um certificado emitido por uma autoridade confiável para assinar o artefato digitalmente.

Se você estiver usando um software ou artefatos baixados na workload, verifique se o provedor fornece uma chave pública para verificação de assinatura digital. Veja alguns exemplos de como a AWS fornece uma chave pública e instruções de verificação para o software que publicamos:

- EC2 Image Builder: verificar a assinatura do download da instalação do AWS TOE
- AWS Systems Manager: verificar a assinatura do SSM Agent
- Amazon CloudWatch: verificar a assinatura do pacote do agente do CloudWatch

Incorpore a verificação de assinatura digital aos processos que você usa para obter e reforçar imagens, conforme discutido em SEC06-BP02 Provisionar a computação com base em imagens reforçadas.

Você pode usar o AWS Signer para ajudar a gerenciar a verificação de assinaturas, bem como seu próprio ciclo de vida de assinatura de código para seu próprio software e artefatos. Tanto o AWS Lambda quanto o Amazon Elastic Container Registry fornecem integrações com o Signer para verificar as assinaturas do seu código e imagens. Usando os exemplos na seção “Recursos”, você pode incorporar o Signer aos pipelines de integração e entrega contínuas (CI/CD) para automatizar a verificação de assinaturas e a assinatura de código e imagens.

Recursos

Documentos relacionados:

- Assinatura criptográfica para contêineres
- Práticas recomendadas para ajudar a proteger sua imagem de contêiner: crie um pipeline usando AWS Signer
- Assinatura de imagens de contêineres com o AWS Signer e o Amazon EKS
- Configurar a assinatura de código para o AWS Lambda
- Práticas recomendadas e padrões avançados para assinatura de código do Lambda
- Assinatura de código usando CA privada do AWS Certificate Manager e chaves assimétricas do AWS Key Management Service

Exemplos relacionados:

- Automatizar a assinatura de código do Lambda com o Amazon CodeCatalyst e o AWS Signer
- Assinar e validar artefatos OCI com o AWS Signer

Ferramentas relacionadas:

- AWS Lambda
- AWS Signer
- AWS Certificate Manager
- AWS Key Management Service
- AWS CodeArtifact

SEC06-BP05 Automatizar a proteção da computação

Automatize as operações de proteção da computação para reduzir a necessidade de intervenção humana. Use a verificação automatizada para detectar possíveis problemas em seus recursos de computação e corrigir com respostas programáticas automatizadas ou operações de gerenciamento de frota. Incorpore a automação em seus processos de CI/CD para implantar workloads confiáveis com dependências atualizadas.

Resultado desejado: sistemas automatizados realizam todas as verificações e correções dos recursos computacionais. Você usa a verificação automatizada para determinar se as imagens e dependências do software são provenientes de fontes confiáveis e não foram adulteradas. As workloads são verificadas automaticamente em busca de dependências atualizadas e assinadas para estabelecer a confiabilidade em ambientes computacionais da AWS. As correções automatizadas são iniciadas quando recursos fora de conformidade são detectados.

Práticas comuns que devem ser evitadas:

- Seguir a prática de infraestrutura imutável, mas sem ter uma solução para correção emergencial ou substituição de sistemas de produção.
- Usar a automação para corrigir recursos configurados incorretamente, mas sem ter um mecanismo de substituição manual instalado. Podem surgir situações em que você precise ajustar os requisitos e suspender as automações até fazer essas alterações.

Benefícios de implementar esta prática recomendada: a automação pode reduzir o risco de acesso e uso não autorizados de seus recursos computacionais. Isso ajuda a evitar que configurações

incorretas entrem nos ambientes de produção e a detectar e corrigir configurações incorretas caso elas ocorram. A automação também ajuda a detectar acesso e uso não autorizados de recursos de computação para reduzir o tempo de resposta. Isso, por sua vez, pode reduzir o escopo geral do impacto do problema.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

É possível aplicar as automações descritas nas práticas do pilar de segurança para proteger seus recursos de computação. SEC06-BP01 Realizar o gerenciamento de vulnerabilidades descreve como você pode usar o Amazon Inspector em seus pipelines de CI/CD e para verificar continuamente seus ambientes de runtime em busca de vulnerabilidades e exposições comuns (CVEs) conhecidas. Você pode usar o AWS Systems Manager para aplicar patches ou reimplantar com base em novas imagens por meio de runbooks automatizados para manter sua frota computacional atualizada com o software e as bibliotecas mais recentes. Use essas técnicas para reduzir a necessidade de processos manuais e acesso interativo aos seus recursos de computação. Consulte SEC06-BP03 Reduzir o gerenciamento manual e o acesso interativo para saber mais.

A automação também desempenha um papel na implantação de workloads confiáveis, descritas em SEC06-BP02 Provisionar computação com base em imagens reforçadas e SEC06-BP04 Validar a integridade do software. É possível usar serviços como EC2 Image Builder, AWS Signer, AWS CodeArtifact e Amazon Elastic Container Registry (ECR) para baixar, verificar, construir e armazenar imagens reforçadas e aprovadas e dependências de código. Com o Inspector, cada um desses serviços pode desempenhar um papel no processo de CI/CD, de forma que a workload chegue à produção somente quando for confirmado que suas dependências estão atualizadas e provêm de fontes confiáveis. Sua workload também é assinada para que ambientes computacionais da AWS, como AWS Lambda e Amazon Elastic Kubernetes Service (EKS), possam verificar se ela não foi adulterada antes de permitir sua execução.

Além desses controles preventivos, você também pode usar a automação nos controles de detecção para seus recursos de computação. Como exemplo, o AWS Security Huboferece o padrão NIST 800-53 Rev. 5, que inclui verificações como [EC2.8] As instâncias do EC2 devem usar o Instance Metadata Service Version 2 (IMDSv2). O IMDSv2 usa as técnicas de autenticação de sessão, bloqueando solicitações que contêm um cabeçalho HTTP X-Forwarded-For e um TTL de rede de 1 para interromper o tráfego proveniente de fontes externas e recuperar informações sobre a instância do EC2. Essa verificação no Security Hub pode detectar quando as instâncias do EC2 usam o IMDSv1 e iniciar a autocorreção. Saiba mais sobre detecção e remediações automatizadas em SEC04-BP04 Iniciar a correção para recursos fora de conformidade.

Etapas de implementação

1. Automatize a criação de AMIs seguras, em conformidade e reforçadas com o EC2 Image Builder. Você pode produzir imagens que incorporem controles dos padrões de referência do Center for Internet Security (CIS) ou do Security Technical Implementation Guide (STIG) com base em imagens básicas da AWS e de parceiros da APN.
2. Automatize o gerenciamento de configuração. Aplique e valide configurações seguras automaticamente em seus recursos de computação usando um serviço ou uma ferramenta de gerenciamento de configuração.
   a. Gerenciamento automatizado de configurações usando o AWS Config
   b. Gerenciamento automatizado da postura de segurança e conformidade usando o AWS Security Hub
3. Automatize a aplicação de patches ou a substituição de instâncias do Amazon Elastic Compute Cloud (Amazon EC2). AWS O Gerenciador de Patches do Systems Manager automatiza o processo de aplicação de patches em instâncias gerenciadas com atualizações relacionadas à segurança e com outros tipos de atualizações. Você pode usar o Patch Manager para aplicar patches de sistemas operacionais e aplicações.
   a. Gerenciador de patches do AWS Systems Manager
4. Automatize a verificação de recursos de computação em busca de vulnerabilidades e exposições comuns (CVEs) e incorpore soluções de verificação de segurança em seu pipeline de criação.
   a. Amazon Inspector
   b. Verificação de imagens do ECR
5. Considere o Amazon GuardDuty para detecção automática de malware e ameaças para proteger os recursos computacionais. O GuardDuty também pode identificar possíveis problemas quando uma função do AWS Lambda é invocada em seu ambiente da AWS.
   a. Amazon GuardDuty
6. Considere as soluções dos parceiros da AWS. AWS Os parceiros oferecem produtos líderes do setor que são equivalentes, idênticos ou se integram aos controles existentes nos seus ambientes on-premises. Esses produtos complementam os serviços existentes da AWS para que você possa implantar uma arquitetura de segurança abrangente e obter uma experiência mais uniforme em seus ambientes na nuvem e on-premises.
   a. Segurança da infraestrutura

Recursos

Práticas recomendadas relacionadas:

- SEC01-BP06 Automatizar a implantação de controles de segurança padrão

Documentos relacionados:

- Obtenha todos os benefícios do IMDSv2 e desative o IMDSv1 em sua infraestrutura da AWS

Vídeos relacionados:

- Práticas recomendadas de segurança para o serviço de metadados da instância do Amazon EC2

Proteção de dados

Perguntas

- SEC 7. Como classificar meus dados?
- SEC 8. Como você protege seus dados em repouso?
- SEC 9. Como você protege seus dados em trânsito?

SEC 7. Como classificar meus dados?

A classificação fornece uma maneira de categorizar dados, com base na criticidade e sensibilidade, para ajudar você a determinar controles apropriados de proteção e retenção.

Práticas recomendadas

- SEC07-BP01 Compreender seu esquema de classificação de dados
- SEC07-BP02 Aplicar controles de proteção de dados com base na confidencialidade dos dados
- SEC07-BP03 Automatizar a identificação e a classificação
- SEC07-BP04 Definir o gerenciamento escalável do ciclo de vida dos dados

SEC07-BP01 Compreender seu esquema de classificação de dados

Compreenda a classificação dos dados que a workload está processando, os requisitos de tratamento, os processos de negócios associados, onde os dados são armazenados e quem é

o proprietário dos dados. Seu esquema de classificação e tratamento de dados deve considerar os requisitos legais e de conformidade aplicáveis à workload e quais controles de dados são necessários. Compreender os dados é a primeira etapa da jornada de classificação de dados.

Resultado desejado: os tipos de dados presentes em sua workload são bem compreendidos e documentados. Controles apropriados estão em vigor para proteger os dados confidenciais com base na respectiva classificação. Esses controles regem fatores como: quem tem permissão para acessar os dados e com que finalidade, onde eles são armazenados, a respectiva política de criptografia e como as chaves de criptografia são gerenciadas, o ciclo de vida dos dados e os requisitos de retenção, os processos de destruição apropriados, quais processos de backup e recuperação estão em vigor e a auditoria do acesso.

Práticas comuns que devem ser evitadas:

- Não ter uma política formal de classificação de dados em vigor para definir os níveis de confidencialidade dos dados e os requisitos de tratamento
- Não ter uma boa compreensão dos níveis de confidencialidade dos dados na workload e não capturar essas informações na documentação de arquitetura e operações
- Não aplicar os controles apropriados sobre os dados com base na confidencialidade e nos respectivos requisitos, conforme descrito em sua política de classificação e tratamento de dados
- Não fornecer feedback sobre os requisitos de classificação e tratamento de dados aos proprietários das políticas.

Benefícios de implementar esta prática recomendada: essa prática elimina a ambiguidade em relação ao tratamento adequado dos dados em sua workload. A aplicação de uma política formal que defina os níveis de confidencialidade dos dados em sua organização e as proteções necessárias pode ajudar você a cumprir as regulamentações legais e outros atestados e certificações de segurança cibernética. Os proprietários das workloads podem ter certeza sobre onde os dados confidenciais estão armazenados e quais controles de proteção estão em vigor. Capturá-los na documentação ajuda os novos membros da equipe a compreendê-los melhor e a manter os controles no início de sua gestão. Essas práticas também podem ajudar a reduzir os custos ao dimensionar corretamente os controles para cada tipo de dados.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Ao projetar uma workload, você pode considerar opções para proteger os dados confidenciais de forma intuitiva. Por exemplo, em uma aplicação multilocatário, é intuitivo considerar os dados de cada locatário como confidenciais e implementar proteções para impedir que um locatário acesse os dados de outro locatário. Da mesma forma, você pode projetar controles de acesso intuitivamente para que apenas os administradores modifiquem os dados e os outros usuários tenham acesso somente leitura ou não tenham nenhum acesso.

Com esses níveis de confidencialidade de dados definidos e capturados na política, bem como os respectivos requisitos de proteção de dados, você pode identificar formalmente quais dados residem na workload. Em seguida, é possível determinar se os controles corretos estão em vigor, se os controles podem ser auditados e quais respostas são apropriadas se os dados forem tratados incorretamente.

Para ajudar a categorizar onde os dados confidenciais estão presentes em sua workload, considere usar tags de recursos quando disponíveis. Por exemplo, você pode aplicar uma tag que tenha uma chave de tag de `Classification` e um valor de tag de `PHI` para informações de saúde protegidas (PHI) e outra tag que tenha uma chave de tag de `Sensitivity` e um valor de tag de `High`. Serviços como o AWS Config podem então ser usados para monitorar esses recursos em busca de alterações e alertar se eles forem modificados de uma forma que os tire da conformidade com seus requisitos de proteção (como alterar as configurações de criptografia). Você pode capturar a definição padrão de suas chaves de tag e valores aceitáveis usando políticas de tag, um recurso do AWS Organizations. Não é recomendável que a chave ou o valor da tag contenha dados privados ou confidenciais.

Etapas de implementação

1. Entenda o esquema de classificação de dados e os requisitos de proteção da sua organização.
2. Identifique os tipos de dados confidenciais processados pelas workloads.
3. Verifique se os dados confidenciais estão sendo armazenados e protegidos na workload de acordo com sua política. Use técnicas, como testes automatizados, para auditar a eficácia de seus controles.
4. Considere usar a marcação em nível de recursos e dados, quando disponível, para marcar o nível de confidencialidade dos dados e outros metadados operacionais que possam ajudar no monitoramento e resposta a incidentes.
   a. As políticas de tag do AWS Organizations podem ser usadas para impor padrões de marcação.

Recursos

Práticas recomendadas relacionadas:

- SUS04-BP01 Implementar uma política de classificação de dados

Documentos relacionados:

- Whitepaper Classificação de dados
- Práticas recomendadas para marcação de recursos da AWS com tags

Exemplos relacionados:

- Sintaxe e exemplos de políticas de tags AWS Organizations

Ferramentas relacionadas

- AWS Tag Editor

SEC07-BP02 Aplicar controles de proteção de dados com base na confidencialidade dos dados

Aplique controles de proteção de dados que ofereçam um nível apropriado de controle para cada classe de dados definida em sua política de classificação. Essa prática pode permitir que você proteja dados confidenciais contra acesso e uso não autorizados, preservando a disponibilidade e o uso dos dados.

Resultado desejado: você tem uma política de classificação que define os diferentes níveis de sensibilidade dos dados em sua organização. Para cada um desses níveis de confidencialidade, você tem diretrizes claras publicadas para serviços e locais de armazenamento e manuseio aprovados e as respectivas configurações necessárias. Você implementa os controles para cada nível conforme o nível de proteção necessário e os custos correspondentes. Você dispõe de monitoramento e alertas para detectar se há dados em locais não autorizados, se eles são processados em ambientes não autorizados, se eles são acessados por agentes não autorizados ou se a configuração dos serviços relacionados está fora de conformidade.

Práticas comuns que devem ser evitadas:

- Aplicar o mesmo nível de controles de proteção em todos os dados. Isso pode levar ao superprovisionamento de controles de segurança para dados com baixo nível de confidencialidade ou à proteção insuficiente dos dados altamente confidenciais.
- Não envolver as partes interessadas relevantes das equipes de segurança, conformidade e negócios ao definir os controles de proteção de dados.
- Ignorar as despesas operacionais indiretas e os custos associados à implementação e manutenção dos controles de proteção de dados.
- Não realizar revisões periódicas de controle de proteção de dados para manter o alinhamento com as políticas de classificação.

Benefícios de implementar esta prática recomendada: ao alinhar seus controles ao nível de classificação de seus dados, sua organização pode investir em níveis mais altos de controle quando necessário. Isso pode incluir a ampliação dos recursos de proteção, monitoramento, medição, correção e geração de relatórios. Nas circunstâncias em que é apropriado usar menos controles, você pode melhorar a acessibilidade e a completude dos dados para seu quadro de funcionários, clientes ou membros. Essa abordagem possibilita que sua organização use os dados da forma mais flexível possível e, ao mesmo tempo, cumpra os requisitos de proteção de dados.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

A implementação de controles de proteção de dados com base nos níveis de confidencialidade dos dados envolve várias etapas importantes. Primeiro, identifique os diferentes níveis de confidencialidade de dados em sua arquitetura de workload (como público, interno, confidencial e restrito) e avalie onde você armazena e processa esses dados. Em seguida, defina limites de isolamento em torno dos dados com base no respectivo nível de confidencialidade. Recomendamos separar os dados em diferentes Contas da AWS usando políticas de controle de serviços (SCPs) para restringir serviços e ações permitidos para cada nível de confidencialidade de dados. Dessa forma, é possível criar limites de isolamento robustos e aplicar o princípio de privilégio mínimo.

Após a definição dos limites de isolamento, implemente controles de proteção apropriados com base nos níveis de confidencialidade dos dados. Consulte as práticas recomendadas para Proteger dados em repouso e Proteger dados em trânsito para implementar controles relevantes, como criptografia, controles de acesso e auditoria. Considere técnicas como tokenização ou anonimização para reduzir o nível de confidencialidade dos dados. Simplifique a aplicação de políticas de dados consistentes em sua empresa com um sistema centralizado para tokenização e destokenização.

Monitore e teste continuamente a eficácia dos controles implementados. Analise e atualize regularmente o esquema de classificação de dados, as avaliações de risco e os controles de proteção à medida que as ameaças e o cenário de dados de sua organização evoluírem. Alinhe os controles de proteção de dados implementados com as regulamentações, os padrões e os requisitos legais relevantes do setor. Além disso, ofereça conscientização e treinamento sobre segurança para ajudar os funcionários a entender o esquema de classificação de dados e suas responsabilidades no tratamento e proteção de dados confidenciais.

Etapas de implementação

1. Identifique os níveis de classificação e confidencialidade dos dados em sua workload.
2. Defina limites de isolamento para cada nível e determine uma estratégia de imposição.
3. Avalie os controles definidos por você que regem acesso, criptografia, auditoria, retenção e outras questões exigidas por sua política de classificação de dados.
4. Avalie as opções para reduzir o nível de confidencialidade dos dados quando apropriado, como usar tokenização ou anonimização.
5. Verifique os controles usando testes e monitoramento automatizados dos recursos configurados.

Recursos

Práticas recomendadas relacionadas:

- PERF03-BP01 Usar um datastore com propósitos específicos que melhor atenda aos requisitos de acesso e armazenamento de dados
- COST04-BP05 Impor políticas de retenção de dados

Documentos relacionados:

- Whitepaper Classificação de dados
- Práticas recomendadas de segurança, identidade e conformidade
- Práticas recomendadas do AWS KMS.
- Práticas recomendadas e recursos de criptografia para serviços da AWS

Exemplos relacionados:

- Criar uma solução de tokenização sem servidor para mascarar dados confidenciais

- Como usar a tokenização para melhorar a segurança dos dados e reduzir o escopo da auditoria

Ferramentas relacionadas:

- AWS Key Management Service (AWS KMS)
- AWS CloudHSM
- AWS Organizations

SEC07-BP03 Automatizar a identificação e a classificação

Automatizar a identificação e a classificação de dados pode ajudar a implementar os controles corretos. O uso da automação para ampliar a determinação manual reduz o risco de erro humano e a exposição.

Resultado desejado: você pode verificar se os controles adequados estão em vigor com base em sua política de classificação e manuseio. Ferramentas e serviços automatizados ajudam a identificar e classificar o nível de confidencialidade dos dados. A automação também ajuda a monitorar continuamente os ambientes para detectar e alertar se os dados estão sendo armazenados ou manipulados de forma não autorizada para que medidas corretivas possam ser tomadas rapidamente.

Práticas comuns que devem ser evitadas:

- Confiar apenas em processos manuais para identificação e classificação de dados, os quais podem ser propensos a erros e demorados. Isso pode resultar em uma classificação de dados ineficiente e inconsistente, especialmente à medida que os volumes de dados aumentam.
- Não ter mecanismos para monitorar e gerenciar ativos de dados em toda a organização.
- Ignorar a necessidade de monitoramento e classificação contínuos dos dados conforme eles se movimentam e evoluem dentro da organização.

Benefícios de implementar esta prática recomendada: automatizar a identificação e a classificação de dados pode levar a uma aplicação mais consistente e precisa dos controles de proteção de dados, reduzindo o risco de erro humano. A automação também pode fornecer visibilidade sobre o acesso e a movimentação de dados confidenciais, ajudando a detectar o tratamento não autorizado e a adotar medidas corretivas.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Embora a avaliação humana seja frequentemente usada para classificar dados durante as fases iniciais de projeto de uma workload, considere a possibilidade de ter sistemas que automatizem a identificação e a classificação dos dados de teste como um controle preventivo. Por exemplo, os desenvolvedores podem receber uma ferramenta ou serviço para verificar dados representativos e determinar o nível de confidencialidade. Na AWS, é possível carregar conjuntos de dados no Amazon S3 e examiná-los usando o Amazon Macie, o Amazon Comprehend ou o Amazon Comprehend Medical. Da mesma forma, considere a possibilidade de verificar os dados como parte dos testes de unidade e integração para detectar onde não deve haver dados confidenciais. Utilizar alertas sobre dados confidenciais nesse estágio pode destacar brechas nas proteções antes da implantação na produção. Outros recursos do AWS Glue, como detecção de dados confidenciais no Amazon SNS e no Amazon CloudWatch, também podem ser usados para detectar PII e aplicar as medidas mitigadoras necessárias. Com relação a quaisquer ferramentas ou serviços automatizados, entenda como eles definem dados confidenciais e complemente-os com outras soluções humanas ou automatizadas para resolver qualquer brecha conforme necessário.

Como controle de detecção, use o monitoramento contínuo dos ambientes para detectar se os dados confidenciais estão sendo armazenados fora de conformidade. Isso pode ajudar a detectar determinadas situações, como publicação ou cópia de dados confidenciais em arquivos de log ou para um ambiente de análise de dados sem a devida desidentificação ou edição. Com relação aos dados armazenados no Amazon S3, os dados confidenciais podem ser monitorados continuamente usando o Amazon Macie.

Etapas de implementação

1. Execute uma verificação inicial dos ambientes para identificação e classificação automatizadas.
    a. Uma verificação inicial completa dos dados pode ajudar a gerar uma compreensão abrangente do local em que os dados confidenciais residem nos ambientes. Quando uma verificação completa não for necessária inicialmente ou não puder ser concluída antecipadamente devido ao custo, avalie se as técnicas de amostragem de dados são adequadas para alcançar seus resultados. Por exemplo, o Amazon Macie pode ser configurado para realizar uma ampla operação automatizada de descoberta de dados confidenciais nos buckets do S3. Esse recurso usa técnicas de amostragem para realizar de forma econômica uma análise preliminar do local em que os dados confidenciais residem. Uma análise mais detalhada dos buckets do S3 pode então ser realizada usando um trabalho de descoberta de dados confidenciais. Outros datastores também podem ser exportados para o S3 para serem verificados pelo Macie.
2. Configure verificações contínuas dos ambientes.

   a. O recurso automatizado de descoberta de dados confidenciais do Macie pode ser usado para realizar verificações contínuas dos ambientes. Os buckets do S3 conhecidos e autorizados a armazenar dados confidenciais podem ser excluídos usando uma lista de permissões em no Macie.

3. Incorpore identificação e classificação nos processos de compilação e teste.

   a. Identifique as ferramentas que os desenvolvedores podem usar para verificar a confidencialidade dos dados enquanto as workloads estão sendo desenvolvidas. Use essas ferramentas como parte dos testes de integração para emitir alertas quando houver dados confidenciais inesperados e evitar implantações adicionais.

4. Implemente um sistema ou um runbook para tomar medidas quando dados confidenciais forem encontrados em locais não autorizados.

Recursos

Documentos relacionados:

- AWS Glue: Detectar e processar dados confidenciais
- Usar identificadores de dados gerenciados no Amazon SNS
- Amazon CloudWatch Logs: ajude a proteger dados de log confidenciais com mascaramento

Exemplos relacionados:

- Habilitar a classificação de dados para o banco de dados do Amazon RDS com o Macie
- Detectar dados confidenciais no DynamoDB com o Macie

Ferramentas relacionadas:

- Amazon Macie
- Amazon Comprehend
- Amazon Comprehend Medical
- AWS Glue

SEC07-BP04 Definir o gerenciamento escalável do ciclo de vida dos dados

Entenda os requisitos do ciclo de vida dos dados relacionados aos seus diferentes níveis de classificação e tratamento de dados. Isso pode incluir como os dados são tratados quando entram pela primeira vez em seu ambiente, como os dados são transformados e as regras para sua destruição. Considere fatores como períodos de retenção, acesso, auditoria e rastreamento da procedência.

Resultado desejado: você classifica os dados o mais próximo possível do ponto e da hora da ingestão. Quando a classificação de dados exige mascaramento, tokenização ou outros processos que reduzam o nível de confidencialidade, você executa essas ações o mais próximo possível do ponto e hora de ingestão.

Você exclui os dados de acordo com sua política quando não é mais apropriado mantê-los e com base na respectiva classificação.

Práticas comuns que devem ser evitadas:

- Implementar uma abordagem única de gerenciamento do ciclo de vida dos dados sem considerar os diferentes níveis de confidencialidade e requisitos de acesso.
- Considerar o gerenciamento do ciclo de vida somente do ponto de vista dos dados utilizáveis ou dos dados submetidos a backup, mas não de ambos.
- Supor que os dados que entraram na workload são válidos, sem estabelecer o respectivo valor ou procedência.
- Confiar na durabilidade dos dados como substituto dos backups e da proteção de dados.
- Reter os dados depois que eles já perderam a utilidade e após o período de retenção exigido.

Benefícios de implementar esta prática recomendada: uma estratégia de gerenciamento do ciclo de vida de dados bem definida e escalável ajuda a manter a conformidade regulatória, melhora a segurança dos dados, otimiza os custos de armazenamento e permite o acesso e o compartilhamento eficientes dos dados ao mesmo tempo que os controles adequados são mantidos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Os dados em uma workload geralmente são dinâmicos. A forma que eles assumem ao entrar no ambiente da workload pode ser diferente de quando são armazenados ou usados em lógica de

negócios, relatórios, análises ou machine learning. Além disso, a importância dos dados pode mudar com o tempo. Alguns dados são de natureza temporal e perdem o valor à medida que se tornam obsoletos. Considere como essas mudanças nos dados afetam a avaliação em seu esquema de classificação de dados e controles associados. Sempre que possível, use um mecanismo de ciclo de vida automatizado, como as políticas de ciclo de vida do Amazon S3 e o Amazon Data Lifecycle Manager, para configurar seus processos de retenção, arquivamento e expiração de dados.

Diferencie os dados que estão disponíveis para uso e os dados armazenados como backup. Considere usar o AWS Backup para automatizar o backup de dados em todos os serviços da AWS. Os snapshots do Amazon EBS oferecem uma forma de copiar um volume do EBS e armazená-lo usando recursos do S3, incluindo ciclo de vida, proteção de dados e acesso a mecanismos de proteção. Dois desses mecanismos são o Bloqueio de Objetos do S3 e o AWS Backup Vault Lock, que podem fornecer segurança e controle adicionais sobre seus backups. Gerencie a separação clara de deveres e acesso para backups. Isole os backups no nível da conta para manter a separação do ambiente afetado durante um evento.

Outro aspecto do gerenciamento do ciclo de vida é registrar o histórico dos dados à medida que eles progridem em sua workload, o que é chamado de rastreamento da procedência dos dados. Desse modo, você pode ter certeza de que sabe de onde os dados vieram, quais transformações foram realizadas, qual proprietário ou processo fez essas alterações e quando. Ter esse histórico ajuda a solucionar problemas e investigações durante possíveis eventos de segurança. Por exemplo, você pode registrar metadados sobre transformações em uma tabela do Amazon DynamoDB. Em um data lake, você pode manter cópias dos dados transformados em diferentes buckets do S3 para cada estágio do pipeline de dados. Armazene as informações do esquema e do carimbo de data/hora em um AWS Glue Data Catalog. Independentemente da sua solução, considere os requisitos dos usuários finais para determinar as ferramentas apropriadas e necessárias para oferecer um relatório sobre a procedência dos dados. Isso ajudará você a determinar a melhor forma de rastrear a procedência.

Etapas de implementação

1. Analise os tipos de dados, os níveis de confidencialidade e os requisitos de acesso da workload para classificar os dados e definir estratégias apropriadas de gerenciamento do ciclo de vida.
2. Projete e implemente políticas de retenção de dados e processos automatizados de destruição que se alinhem aos requisitos legais, regulatórios e organizacionais.
3. Estabeleça processos e automação para monitoramento, auditoria e ajuste contínuos de estratégias, controles e políticas de gerenciamento do ciclo de vida dos dados à medida que os requisitos e as regulamentações da workload evoluem.

Recursos

Práticas recomendadas relacionadas:

- COST04-BP05 Impor políticas de retenção de dados
- SUS04-BP03 Usar políticas para gerenciar o ciclo de vida de conjuntos de dados

Documentos relacionados:

- Whitepaper Classificação de dados
- Esquema da AWS para defesa contra ransomware
- Orientação de DevOps: melhorar a rastreabilidade com o rastreamento da procedência de dados

Exemplos relacionados:

- Como proteger dados confidenciais durante todo o seu ciclo de vida na AWS
- Como construir linhagem de dados para data lakes usando o AWS Glue, o Amazon Neptune e o Spline

Ferramentas relacionadas:

- AWS Backup
- Amazon Data Lifecycle Manager
- AWS Identity and Access Management Access Analyzer

SEC 8. Como você protege seus dados em repouso?

Proteja seus dados em repouso implementando vários controles para reduzir o risco de acesso não autorizado ou manuseio incorreto.

Práticas recomendadas

- SEC08-BP01 Implementar o gerenciamento seguro de chaves
- SEC08-BP02 Aplicar criptografia em repouso
- SEC08-BP03 Automatizar a proteção de dados em repouso
- SEC08-BP04 Aplicar controle de acesso

SEC08-BP01 Implementar o gerenciamento seguro de chaves

O gerenciamento seguro de chaves inclui o armazenamento, a rotação, o controle de acesso e o monitoramento do material essencial necessário para proteger os dados em repouso para sua workload.

Resultado desejado: um mecanismo de gerenciamento de chaves escalável, repetível e automatizado. O mecanismo deve fornecer a capacidade de impor o acesso de privilégio mínimo ao material essencial e fornecer o equilíbrio correto entre disponibilidade, confidencialidade e integridade das chaves. O acesso às chaves deve ser monitorado e o material da chave deve ser rotacionado por meio de um processo automatizado. O material da chave nunca deve estar acessível para identidades humanas.

Práticas comuns que devem ser evitadas:

- Acesso humano a material de chave não criptografado.
- Criação de algoritmos criptográficos personalizados.
- Permissões excessivamente amplas para acessar materiais importantes.

Benefícios de implementar esta prática recomendada: ao estabelecer um mecanismo seguro de gerenciamento de chaves para sua workload, você pode ajudar a proteger seu conteúdo contra acesso não autorizado. Além disso, você pode estar sujeito a requisitos regulatórios para criptografar seus dados. Uma solução eficaz de gerenciamento de chaves pode fornecer mecanismos técnicos alinhados a essas regulamentações para proteger o material das chaves.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Muitos requisitos regulatórios e práticas recomendadas incluem a criptografia de dados em repouso como um controle de segurança fundamental. Para cumprir esse controle, sua workload precisa de um mecanismo para armazenar e gerenciar com segurança o material de chave usado para criptografar seus dados em repouso.

A AWS oferece o AWS Key Management Service (AWS KMS) para fornecer armazenamento durável, seguro e redundante para chaves do AWS KMS. Muitos serviços da AWS se integram ao AWS KMS para oferecer suporte à criptografia de seus dados. O AWS KMS usa módulos de segurança de hardware validados pelo FIPS 140-2 Nível 3 para proteger suas chaves. Não há mecanismo para exportar chaves do AWS KMS em texto simples.

Ao implantar workloads usando uma estratégia de várias contas, é considerado prática recomendada manter as chaves do AWS KMS na mesma conta da workload que as utiliza. Nesse modelo distribuído, a responsabilidade pelo gerenciamento das chaves do AWS KMS é da equipe de aplicações. Em outros casos de uso, as organizações podem optar por armazenar as chaves do AWS KMS em uma conta centralizada. Essa estrutura centralizada requer políticas adicionais para permitir o acesso entre contas necessário para que a conta da workload acesse as chaves armazenadas na conta centralizada, mas pode ser mais aplicável em casos de uso em que uma única chave é compartilhada entre várias Contas da AWS.

Independentemente de onde o material da chave esteja armazenado, o acesso à chave deve ser rigorosamente controlado por meio do uso de políticas de chave e políticas do IAM. Políticas de chave são a principal forma de controlar o acesso a uma chave do AWS KMS. Além disso, concessões à chave do AWS KMS podem fornecer acesso a serviços da AWS para criptografar e descriptografar dados em seu nome. Reserve tempo para analisar as práticas recomendadas para controle de acesso às suas chaves do AWS KMS.

É prática recomendada monitorar o uso de chaves de criptografia para detectar padrões de acesso incomuns. As operações realizadas usando chaves gerenciadas pela AWS e chaves gerenciadas pelo cliente armazenadas no AWS KMS podem ser registradas no AWS CloudTrail e devem ser revisadas periodicamente. Atenção especial deve ser dada ao monitoramento dos principais eventos de destruição. Para mitigar a destruição acidental ou maliciosa de material de chave, os eventos de destruição da chave não excluem o material da chave imediatamente. Tentativas de excluir chaves no AWS KMS estão sujeitas a um período de espera cujo padrão é de 30 dias, o que dá aos administradores tempo para revisar essas ações e reverter a solicitação, se necessário.

A maioria dos serviços da AWS usam o AWS KMS de forma transparente para você. Seu único requisito é decidir se quer usar uma chave gerenciada pela AWS ou gerenciada pelo cliente. Se sua workload exigir o uso direto do AWS KMS para criptografar ou descriptografar dados, a prática recomendada é usar criptografia envelopada para proteger seus dados. O SDK de criptografia da AWS pode fornecer primitivas de criptografia do lado do cliente às suas aplicações para implementar a criptografia envelopada e integrar com o AWS KMS.

Etapas de implementação

1. Determine as opções apropriadas de gerenciamento de chaves (gerenciadas pela AWS ou gerenciadas pelo cliente) para a chave.

- Para facilitar o uso, a AWS oferece, para a maioria dos serviços, chaves pertencentes à AWS e gerenciadas pela AWS que fornecem capacidade de criptografia em repouso sem a necessidade de gerenciar materiais ou políticas de chaves.
- Ao usar chaves gerenciadas pelo cliente, considere o armazenamento de chaves padrão para fornecer o melhor equilíbrio entre agilidade, segurança, soberania de dados e disponibilidade. Outros casos de uso podem exigir o uso de armazenamentos de chaves personalizadas com o AWS CloudHSM ou o repositório de chaves externo.

2. Analise a lista de serviços que você está usando para sua workload para entender como o AWS KMS se integra ao serviço. Por exemplo, as instâncias do EC2 podem usar volumes criptografados do EBS, verificando se os snapshots do Amazon EBS criados com base nesses volumes também são criptografados usando uma chave gerenciada pelo cliente e mitigando a divulgação acidental de dados de snapshots não criptografados.
   - Como os serviços da AWS usam o AWS KMS
   - Para obter informações detalhadas sobre as opções de criptografia oferecidas por um serviço da AWS, consulte o tópico Criptografia em repouso no manual do usuário ou no Guia do desenvolvedor do serviço.
3. Implemente o AWS KMS: o AWS KMS simplifica a criação e o gerenciamento de chaves e o controle do uso da criptografia em uma ampla variedade de serviços da AWS e em suas aplicações.
   - Conceitos básicos: AWS Key Management Service (AWS KMS)
   - Reserve tempo para analisar as práticas recomendadas para controle de acesso às suas chaves do AWS KMS.
4. Considere o AWS Encryption SDK: use a integração do AWS Encryption SDK com o AWS KMS quando sua aplicação precisar criptografar dados no lado do cliente.
   - AWS Encryption SDK
5. Habilite o IAM Access Analyzer para revisar e notificar automaticamente se houver políticas de chaves do AWS KMS excessivamente amplas.
6. Habilite o Security Hub para receber notificações se houver políticas de chaves configuradas incorretamente, chaves agendadas para exclusão ou chaves sem a rotação automática ativada.
7. Determine o nível de registro em log apropriado para suas chaves do AWS KMS. Como as chamadas para o AWS KMS, incluindo eventos somente para leitura, são registradas em log, os logs do CloudTrail associados ao AWS KMS podem se tornar volumosos.

- Algumas organizações preferem registrar a atividade de log do AWS KMS em uma trilha separada. Para obter mais detalhes, consulte a seção Registrar em log as chamadas de API do AWS KMS com o CloudTrail do Guia do desenvolvedor do AWS KMS.

Recursos

Documentos relacionados:

- AWS Key Management Service
- Ferramentas e serviços criptográficos da AWS
- Proteção de dados do Amazon S3 usando criptografia
- Criptografia de envelope
- Promessa de soberania digital
- Desmistificação das operações de chave do AWS KMS, traga sua própria chave, armazenamento de chaves personalizado e portabilidade de texto cifrado
- Detalhes criptográficos do AWS Key Management Service

Vídeos relacionados:

- Como funciona a criptografia na AWS
- Como proteger seu armazenamento em bloco na AWS
- Proteção de dados na AWS: usar bloqueios, chaves, assinaturas e certificados

Exemplos relacionados:

- Implemente mecanismos avançados de controle de acesso usando o AWS KMS

SEC08-BP02 Aplicar criptografia em repouso

O uso de criptografia para dados em repouso é indispensável. A criptografia mantém a confidencialidade dos dados sigilosos em caso de acesso não autorizado ou divulgação acidental.

Resultado desejado: os dados privados devem ser criptografados por padrão quando em repouso. A criptografia ajuda a manter a confidencialidade dos dados e oferece uma camada adicional de proteção contra a divulgação intencional ou acidental ou exfiltração de dados. Os dados

criptografados não podem ser lidos nem acessados sem ser descriptografados primeiro. Todos os dados armazenados não criptografados devem ser inventariados e controlados.

Práticas comuns que devem ser evitadas:

- Não utilizar configurações de criptografia por padrão.
- Conceder acesso excessivamente permissivo para chaves de descriptografia.
- Não monitorar o uso de chaves de criptografia e descriptografia.
- Armazenar dados não criptografados.
- Utilizar a mesma chave de criptografia para todos os dados, seja qual for o uso, os tipos e a classificação de dados.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Mapeie as chaves de criptografia às classificações de dados em suas workloads. Essa abordagem ajuda a proteger contra o acesso excessivamente permissivo ao usar uma única chave de criptografia ou um número muito pequeno de chaves de criptografia para seus dados (consulte SEC07-BP01 Compreender seu esquema de classificação de dados).

O AWS Key Management Service (AWS KMS) integra-se a muitos serviços da AWS para facilitar a criptografia de seus dados em repouso. Por exemplo, no Amazon Simple Storage Service (Amazon S3), é possível definir a criptografia padrão em um bucket para que todos os novos objetos sejam criptografados automaticamente. Ao utilizar o AWS KMS, considere o nível de restrição necessário para os dados. Chaves do AWS KMS controladas por serviço e padrão são gerenciadas e utilizadas em seu nome pelo AWS. Para dados sigilosos que exijam acesso refinado à chave de criptografia subjacente, considere chaves gerenciadas pelo cliente (CMKs). Você tem total controle sobre as CMKs, como gerenciamento de rotação e acesso pelo uso de políticas de chave.

Além disso, o Amazon Elastic Compute Cloud (Amazon EC2) e o Amazon S3 oferecem suporte à aplicação da criptografia por meio da definição da criptografia padrão. Você pode usar o Regras do AWS Config para verificar automaticamente se está usando criptografia, por exemplo, para volumes do Amazon Elastic Block Store (Amazon EBS), instâncias do Amazon Relational Database Service (Amazon RDS) e buckets do Amazon S3.

A AWS também oferece operações de criptografia do lado do cliente, possibilitando que você criptografe os dados antes de fazer seu upload para a nuvem. O AWS Encryption SDK fornece

uma maneira de criptografar seus dados usando criptografia envelopada. Você fornece a chave de encerramento e o AWS Encryption SDK gera uma chave de dados exclusiva para cada objeto de dados que ele criptografa. Considere utilizar o AWS CloudHSM se precisar de um módulo de segurança de hardware de um locatário (HSM) gerenciado. O AWS CloudHSM possibilita gerar, importar e gerenciar chaves criptográficas em um HSM validado de nível 3 FIPS 140-2. Alguns casos de uso do AWS CloudHSM incluem proteger chaves privadas para emitir uma autoridade de certificado (CA) e ativar a criptografia de dados transparente (TDE) para bancos de dados Oracle. O AWS CloudHSM Client SDK oferece software que possibilita criptografar dados do lado do cliente com chaves armazenadas no AWS CloudHSM antes de fazer upload de seus dados para AWS. O Amazon DynamoDB Encryption Client também possibilita criptografar e assinar itens antes de fazer upload para uma tabela do DynamoDB.

Etapas de implementação

- Aplique criptografia em repouso para o Amazon S3: implemente a criptografia padrão do bucket do Amazon S3.

  Configure a criptografia padrão para novos volumes do Amazon EBS: especifique que você deseja que todos os volumes do Amazon EBS recém-criados sejam criados em formato criptografado, com a opção de usar a chave padrão fornecida pela AWS ou uma chave que você criar.

  Configure imagens de máquina da Amazon (AMIs) criptografadas: copiar uma AMI existente com a criptografia configurada criptografará automaticamente os volumes raiz e snapshots.

  Configure a criptografia do Amazon RDS: configure a criptografia para seus clusters de banco de dados do Amazon RDS e snapshots em repouso usando a opção de criptografia.

  Crie e configure chaves do AWS KMS com políticas que limitam o acesso das entidades principais apropriadas para cada classificação de dados: por exemplo, crie uma chave do AWS KMS para criptografar dados de produção e uma chave diferente para criptografar dados de desenvolvimento ou teste. Você também pode conceder acesso de chave a outras Contas da AWS. Considere ter contas diferentes para seus ambientes de desenvolvimento e produção. Se seu ambiente de produção precisar descriptografar artefatos na conta de desenvolvimento, você poderá editar a política de CMK utilizada para criptografar os artefatos de desenvolvimento a fim de conferir à conta de produção a capacidade de descriptografar esses artefatos. O ambiente de produção pode, então, ingerir os dados descriptografados para uso na produção.

Configure a criptografia em serviços da AWS adicionais: para outros serviços da AWS que você usa, revise a documentação de segurança desse serviço para determinar as opções de criptografia do serviço.

Recursos

Documentos relacionados:

- AWS Crypto Tools
- AWS Encryption SDK
- Whitepaper Detalhes criptográficos do AWS KMS
- AWS Key Management Service
- Ferramentas e serviços criptográficos da AWS
- Criptografia do Amazon EBS
- Criptografia padrão para volumes do Amazon EBS
- Como criptografar recursos do Amazon RDS
- Como faço para habilitar a criptografia padrão em um bucket do Amazon S3?
- Proteção de dados do Amazon S3 usando criptografia

Vídeos relacionados:

- Como funciona a criptografia na AWS
- Como proteger seu armazenamento em bloco na AWS

SEC08-BP03 Automatizar a proteção de dados em repouso

Use a automação para validar e aplicar controles de dados em repouso. Use a verificação automatizada para detectar configurações incorretas de soluções de armazenamento de dados e realize correções por meio de resposta programática automatizada sempre que possível. Incorpore a automação nos processos de CI/CD para detectar configurações incorretas de armazenamento de dados antes que elas sejam implantadas na produção.

Resultado desejado: sistemas automatizados examinam e monitoram os locais de armazenamento de dados em busca de configurações incorretas de controles, acesso não autorizado e uso inesperado. A detecção de locais de armazenamento configurados incorretamente inicia correções

automatizadas. Processos automatizados criam backups de dados e armazenam cópias imutáveis fora do ambiente original.

Práticas comuns que devem ser evitadas:

- Não considerar as opções para habilitar as configurações de criptografia por padrão, onde compatíveis.
- Não considerar eventos de segurança, além dos eventos operacionais, ao formular uma estratégia automatizada de backup e recuperação.
- Não impor configurações de acesso público para serviços de armazenamento.
- Não monitorar e auditar os controles para proteger os dados em repouso.

Benefícios de implementar esta prática recomendada: a automação ajuda a evitar o risco de configuração incorreta dos locais de armazenamento de dados. Isso ajuda a evitar que configurações incorretas entrem nos ambientes de produção. Essa prática recomendada também ajuda a detectar e corrigir configurações incorretas, caso elas ocorram.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A automação é um tema em todas as práticas para proteger os dados em repouso. SEC01-BP06 Automatizar a implantação de controles de segurança padrão descreve como é possível capturar a configuração de seus recursos usando modelos de infraestrutura como código (IaC), como o AWS CloudFormation. Esses modelos estão comprometidos com um sistema de controle de versão e são usados para implantar recursos da AWS por meio de um pipeline de CI/CD. Essas técnicas também se aplicam à automação da configuração de soluções de armazenamento de dados, como configurações de criptografia em buckets do Amazon S3.

Você pode verificar as configurações definidas nos modelos de IaC para verificar se há erros de configuração nos pipelines de CI/CD usando regras no AWS CloudFormation Guard. Você pode monitorar configurações que ainda não estão disponíveis no CloudFormation ou em outras ferramentas de IaC em busca de configurações incorretas com AWS Config. Os alertas que o Config gera para configurações incorretas podem ser corrigidos automaticamente, conforme descrito em SEC04-BP04 Iniciar a correção de recursos fora de conformidade.

Usar a automação como parte da estratégia de gerenciamento de permissões também é um componente essencial das proteções de dados automatizadas. SEC03-BP02 Conceder acesso de

privilégio mínimo e SEC03-BP04 Reduzir permissões continuamente descrevem a configuração de políticas de acesso de privilégio mínimo que são continuamente monitoradas pelo AWS Identity and Access Management Access Analyzer para gerar descobertas quando a permissão pode ser reduzida. Além da automação para monitoramento de permissões, é possível configurar o Amazon GuardDuty para observar comportamentos anômalos de acesso aos dados em seus volumes do EBS (por meio de uma instância do EC2), buckets do S3 e bancos de dados do Amazon Relational Database Service compatíveis.

A automação também desempenha um papel para detectar o armazenamento de dados confidenciais em locais não autorizados. SEC07-BP03 Automatizar a identificação e a classificação descreve como o Amazon Macie pode monitorar seus buckets do S3 em busca de dados confidenciais inesperados e gerar alertas que podem iniciar uma resposta automática.

Siga as práticas de REL09 Backup de dados para desenvolver uma estratégia automatizada de backup e recuperação de dados. O backup e a recuperação de dados são importantes para a recuperação tanto de eventos de segurança quanto de eventos operacionais.

Etapas de implementação

1. Capture a configuração de armazenamento de dados em modelos de IaC. Use verificações automatizadas nos pipelines de CI/CD para detectar configurações incorretas.
   a. É possível usar para seus modelos de IaC e o CloudFormation Guard para verificar se há erros de configuração nos modelos.
   b. Use o AWS Config para executar regras em um modo de avaliação proativa. Use essa configuração como uma etapa em seu pipeline de CI/CD para verificar a conformidade de um recurso antes de criá-lo.
2. Monitore os recursos em busca de configurações incorretas de armazenamento de dados.
   a. Configure o AWS Config para monitorar os recursos de armazenamento de dados em busca de alterações nas configurações de controle e gerar alertas para invocar ações de remediação ao detectar uma configuração incorreta.
   b. Consulte SEC04-BP04 Iniciar a correção para recursos fora de conformidade para obter mais orientações sobre correções automatizadas.
3. Monitore e reduza continuamente as permissões de acesso aos dados por meio da automação.
   a. O IAM Access Analyzer pode ser executado continuamente para gerar alertas quando as permissões podem ser potencialmente reduzidas.
4. Monitore e emita alertas sobre comportamentos anômalos de acesso aos dados.

   a. O GuardDuty observa tanto as assinaturas de ameaças conhecidas quanto os desvios dos comportamentos de acesso básicos para recursos de armazenamento de dados, como volumes do EBS, buckets do S3 e bancos de dados do RDS.

5. Monitore e emita alertas sobre dados confidenciais armazenados em locais inesperados.
   a. Use o Amazon Macie para examinar continuamente seus buckets do S3 em busca de dados confidenciais.

6. Automatize backups seguros e criptografados dos dados.
   a. O AWS Backupé um serviço gerenciado que cria backups criptografados e seguros de várias fontes de dados na AWS. O Elastic Disaster Recovery permite copiar workloads completas do servidor e manter a proteção contínua dos dados com um objetivo de ponto de recuperação (RPO) medido em segundos. Você pode configurar os dois serviços para que funcionem juntos e automatizem a criação de backups de dados e os copiem para locais de failover. Isso pode ajudar a manter os dados disponíveis quando eles forem afetados por eventos operacionais ou de segurança.

Recursos

Práticas recomendadas relacionadas:

- SEC01-BP06 Automatizar a implantação de controles de segurança padrão
- SEC03-BP02 Conceder acesso de privilégio mínimo
- SEC03-BP04 Reduzir as permissões continuamente
- SEC04-BP04 Iniciar a correção de recursos fora de conformidade
- SEC07-BP03 Automatizar a identificação e a classificação
- REL09-BP02 Proteger e criptografar backups
- REL09-BP03 Fazer backup de dados automaticamente

Documentos relacionados:

- Recomendação da AWS: Criptografar automaticamente volumes novos e existentes do Amazon EBS
- Gerenciamento de riscos de ransomware na AWS usando o CSF (Cyber Security Framework) do NIST

Exemplos relacionados:

- Como usar regras proativas do AWS Config e hooks do AWS CloudFormation proativos para evitar a criação de recursos de nuvem fora de conformidade
- Automatizar e gerenciar centralmente a proteção de dados para o Amazon S3 com o AWS Backup
- AWS re:Invent 2023: Implementar proteção proativa de dados usando snapshots do Amazon EBS
- AWS re:Invent 2022: Criar e automatizar para alcançar resiliência com proteção de dados moderna

Ferramentas relacionadas:

- AWS CloudFormation Guard
- Registro de regras do AWS CloudFormation Guard
- IAM Access Analyzer
- Amazon Macie
- AWS Backup
- Elastic Disaster Recovery

SEC08-BP04 Aplicar controle de acesso

Para ajudar a proteger seus dados em repouso, implemente o controle de acesso utilizando mecanismos como isolamento e versionamento e aplique o princípio de privilégio mínimo. Evite conceder acesso público aos seus dados.

Resultado desejado: verifique se somente usuários autorizados podem acessar os dados com base em necessidade estrita de conhecimento. Proteja seus dados com backups regulares e versionamento a fim de impedir a modificação ou a exclusão de dados intencionais ou acidentais. Isole dados críticos de outros dados a fim de proteger a confidencialidade e a integridade desses dados.

Práticas comuns que devem ser evitadas:

- Armazenar dados com requisitos de confidencialidade ou classificações diferentes juntos.
- Utilizar permissões excessivamente tolerantes em chaves de descriptografia.
- Classificar dados de modo inadequado.
- Não reter backups detalhados de dados importantes.

- Conceder acesso persistente a dados de produção.
- Não auditar o acesso aos dados nem rever as permissões regularmente.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Vários controles podem ajudar a proteger seus dados em repouso, por exemplo, acesso (utilizando privilégio mínimo), isolamento e versionamento. Deve ser feita a auditoria de acesso aos seus dados com os mecanismos de detecção, como AWS CloudTrail e os logs de nível de serviço, como os logs de acesso do Amazon Simple Storage Service (Amazon S3). Você deve inventariar quais dados são acessíveis publicamente e criar um plano para reduzir a quantidade de dados disponíveis ao longo do tempo.

O Amazon S3 Glacier Vault Lock e o Amazon S3 Object Lock fornecem controle de acesso obrigatório para os objetos no Amazon S3. Assim que uma política de cofre é bloqueada com a opção de conformidade, nem mesmo o usuário-raiz pode alterá-la até que o bloqueio expire.

Etapas de implementação

- Aplique controle de acesso: aplique o controle de acesso com privilégio mínimo, incluindo acesso a chaves de criptografia.
- Separe os dados com base em diferentes níveis de classificação: use Contas da AWS diferentes para níveis de classificação de dados e gerencie essas contas usando o AWS Organizations.
- Revise as políticas do AWS Key Management Service (AWS KMS): revise o nível de acesso concedido nas políticas do AWS KMS.
- Revise as permissões de bucket e objeto do Amazon S3: revise regularmente o nível de acesso concedido em políticas de bucket do S3. Uma das práticas recomendadas é evitar buckets que possam ser lidos ou gravados publicamente. Considere o uso do AWS Config para detectar buckets que estão disponíveis publicamente e do Amazon CloudFront para fornecer conteúdo do Amazon S3. Garanta que os buckets que não devem permitir acesso público sejam configurados adequadamente para evitar o acesso público. Por padrão, todos os buckets do S3 são privados e só ser acessados por usuários que receberam explicitamente esse acesso.
- Use o AWS IAM Access Analyzer: o IAM Access Analyzer analisa buckets do Amazon S3 e gera uma descoberta quando uma política do S3 concede acesso a uma entidade externa.
- Use o versionamento do Amazon S3 e o bloqueio de objetos quando apropriado.

- Use o Inventário Amazon S3: o Inventário Amazon S3 é uma das ferramentas que podem ser usadas para auditar e gerar relatórios sobre o status de replicação e criptografia de seus objetos do S3.
- Revise as permissões do Amazon EBS e de compartilhamento de AMIs: as permissões de compartilhamento podem permitir que imagens e volumes sejam compartilhados com Contas da AWS externas à sua workload.
- Revise os compartilhamentos do AWS Resource Access Manager periodicamente para determinar se os recursos devem continuar sendo compartilhados. O Resource Access Manager possibilita compartilhar recursos, como políticas do AWS Network Firewall, regras do Amazon Route 53 Resolver e sub-redes em suas Amazon VPCs. Faça auditoria em recursos compartilhados regularmente e interrompa o compartilhamento dos que não precisam mais ser compartilhados.

Recursos

Práticas recomendadas relacionadas:

- SEC03-BP01 Definir requisitos de acesso
- SEC03-BP02 Conceder acesso de privilégio mínimo

Documentos relacionados:

- Whitepaper Detalhes criptográficos do AWS KMS
- Introdução ao gerenciamento de permissões de acesso aos recursos do Amazon S3
- Visão geral do gerenciamento de acesso a recursos do AWS KMS
- Regras do AWS Config
- Amazon S3 + Amazon CloudFront: uma combinação feita na nuvem
- Usar versionamento
- Bloquear objetos usando o bloqueio de objetos do Amazon S3
- Compartilhar um snapshot do Amazon EBS
- AMIs compartilhadas
- Hospedar uma aplicação de página única no Amazon S3

Vídeos relacionados:

- Como proteger seu armazenamento em bloco na AWS

SEC 9. Como você protege seus dados em trânsito?

Proteja seus dados em trânsito implementando vários controles para reduzir o risco de acesso ou perda não autorizados.

Práticas recomendadas

- SEC09-BP01 Implementar o gerenciamento seguro de chaves e certificados
- SEC09-BP02 Impor a criptografia em trânsito
- SEC09-BP03 Autenticar as comunicações de rede

SEC09-BP01 Implementar o gerenciamento seguro de chaves e certificados

Os certificados Transport Layer Security (TLS) são usados para proteger as comunicações de rede e estabelecer a identidade de sites, recursos e workloads na Internet, bem como em redes privadas.

Resultado desejado: um sistema seguro de gerenciamento de certificados que pode provisionar, implantar, armazenar e renovar certificados em uma infraestrutura de chave pública (PKI). Um mecanismo seguro de gerenciamento de chaves e certificados evita que o material da chave privada do certificado seja divulgado e renova automaticamente o certificado periodicamente. Ele também se integra a outros serviços para fornecer comunicações de rede seguras e identidade para os recursos da máquina na workload. O material da chave nunca deve estar acessível para identidades humanas.

Práticas comuns que devem ser evitadas:

- Executar etapas manuais durante os processos de implantação ou renovação de certificados.
- Não prestar a devida atenção à hierarquia da autoridade de certificação (CA) ao criar uma CA privada.
- Usar certificados autoassinados para recursos públicos.

Benefícios de implementar esta prática recomendada:

- Simplificar o gerenciamento de certificados por meio de implantação e renovação automatizadas.
- Incentivar a criptografia de dados em trânsito usando certificados TLS.

- Aumentar a segurança e a auditabilidade das ações de certificação realizadas pela autoridade de certificação.
- Organizar as tarefas de gerenciamento em diferentes camadas da hierarquia da CA.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

As workloads modernas fazem uso extensivo de comunicações de rede criptografadas usando protocolos de PKI, como TLS. O gerenciamento de certificados PKI pode ser complexo, mas o provisionamento, a implantação e a renovação automatizados de certificados podem reduzir o atrito associado ao gerenciamento deles.

A AWS oferece dois serviços para gerenciar certificados de PKI de uso geral: AWS Certificate Manager e AWS Private Certificate Authority (AWS Private CA). O ACM é o principal serviço usado pelos clientes para provisionar, gerenciar e implantar certificados para uso em workloads públicas e privadas da AWS. O ACM emite certificados usando AWS Private CA e se integra a muitos outros serviços gerenciados da AWS para fornecer certificados TLS seguros para workloads.

A AWS Private CA permite estabelecer a própria autoridade de certificação raiz ou subordinada e emitir certificados TLS por meio de uma API. É possível usar esses tipos de certificado em cenários em que você controla e gerencia a cadeia de confiança do lado do cliente da conexão TLS. Além dos casos de uso do TLS, a AWS Private CA pode ser usada para emitir certificados para pods do Kubernetes, atestados de produtos de dispositivos Matter, assinatura de código e outros casos de uso com um modelo personalizado. Também é possível usar o IAM Roles Anywhere para fornecer credenciais do IAM temporárias para workloads on-premises que receberam certificados X.509 assinados pela CA privada.

Além do ACM e AWS Private CA, o AWS IoT Core fornece suporte especializado para provisionamento, gerenciamento e implantação de certificados PKI em dispositivos de IoT. O AWS IoT Core fornece mecanismos especializados para integrar dispositivos de IoT em sua infraestrutura de chave pública em escala.

Considerações para estabelecer uma hierarquia de CA privada

Quando é necessário estabelecer uma CA privada, é importante tomar cuidado especial para projetar adequadamente a hierarquia da CA com antecedência. É prática recomendada implantar cada nível de sua hierarquia de CA em Contas da AWS separadas ao criar uma hierarquia de CA privada. Essa etapa intencional reduz a área de superfície de cada nível na hierarquia da CA, simplificando

a descoberta de anomalias nos dados de log do CloudTrail e reduzindo o escopo de acesso ou impacto se houver acesso não autorizado a uma das contas. A CA raiz deve residir em uma própria conta separada e deve ser usada somente para emitir um ou mais certificados de CA intermediários.

Depois, crie uma ou mais CAs intermediárias em contas separadas da conta da CA raiz para emitir certificados para usuários finais, dispositivos ou outras workloads. Por fim, emita certificados da CA raiz para as CAs intermediárias, que, por sua vez, emitirão certificados para os usuários finais ou dispositivos. Para obter mais informações sobre como planejar a implantação de CA e projetar a hierarquia de CA, incluindo planejamento de resiliência, replicação entre regiões, compartilhamento de CAs na organização e muito mais, consulte Planejar sua implantação da AWS Private CA.

Etapas de implementação

1. Determine os serviços da AWS relevantes e necessários para seu caso de uso:
    - Muitos casos de uso podem aproveitar a infraestrutura de chave pública da AWS existente usando o AWS Certificate Manager. O ACM pode ser usado para implantar certificados TLS para servidores Web, balanceadores de carga ou outros usos para certificados publicamente confiáveis.
    - Considere o AWS Private CA quando precisar estabelecer a própria hierarquia de autoridade de certificação privada ou precisar acessar certificados exportáveis. O ACM pode então ser usado para emitir vários tipos de certificados de entidade final utilizando a AWS Private CA.
    - Para casos de uso em que os certificados devem ser provisionados em grande escala para dispositivos incorporados de Internet das Coisas (IoT), considere usar o AWS IoT Core.
2. Implemente a renovação automática do certificado sempre que possível:
    - Use a renovação gerenciada pelo ACM para certificados emitidos pelo ACM junto com serviços gerenciados da AWS integrados.
3. Estabeleça trilhas de auditoria e registro em log:
    - Habilite os Logs do CloudTrail para monitorar o acesso às contas que detêm autoridades de certificação. Considere configurar a validação da integridade do arquivo de log no CloudTrail para verificar a autenticidade dos dados de log.
    - Gere e revise periodicamente relatórios de auditoria que listam os certificados que sua CA privada emitiu e revogou. Esses relatórios podem ser exportados para um bucket do S3.
    - Ao implantar uma CA privada, você também precisará estabelecer um bucket do S3 para armazenar a lista de revogação de certificados (CRL). Para obter orientação sobre como configurar esse bucket do S3 com base nos requisitos da workload, consulte Planejar uma lista de revogação de certificados (CRL).

Recursos

Práticas recomendadas relacionadas:

- SEC02-BP02 Usar credenciais temporárias
- SEC08-BP01 Implementar o gerenciamento seguro de chaves
- SEC09-BP03 Autenticar as comunicações de rede

Documentos relacionados:

- Como hospedar e gerenciar toda uma infraestrutura de certificados privados na AWS
- Como proteger uma hierarquia de CA privada do ACM em escala empresarial para o setor automotivo e de manufatura
- Práticas recomendadas de CA privada
- Como usar o AWS RAM para compartilhar sua CA privada do ACM entre contas

Vídeos relacionados:

- Como ativar a CA privada do AWS Certificate Manager (workshop)

Exemplos relacionados:

- Workshop de CA privada
- Workshop de gerenciamento de dispositivos IoT (incluindo provisionamento de dispositivos)

Ferramentas relacionadas:

- Plug-in para o gerenciador de certificados do Kubernetes para uso da AWS Private CA

SEC09-BP02 Impor a criptografia em trânsito

Aplique os requisitos de criptografia definidos com base em políticas, obrigações regulatórias e padrões da organização para cumprir os requisitos organizacionais, legais e de conformidade. Utilize somente protocolos com criptografia ao transmitir dados sigilosos para fora da sua nuvem privada virtual (VPC). A criptografia ajuda a manter a confidencialidade dos dados mesmo quando os dados passam por redes não confiáveis.

Resultado desejado: todos os dados devem ser criptografados em trânsito usando protocolos TLS seguros e pacotes de criptografia. O tráfego de rede entre seus recursos e a Internet deve ser criptografado para reduzir o acesso não autorizado aos dados. O tráfego de rede exclusivamente em seu ambiente interno da AWS deve ser criptografado com TLS sempre que possível. A rede interna da AWS é criptografada por padrão e o tráfego de rede em uma VPC não pode ser adulterado nem interceptado a menos que uma parte não autorizada tenha obtido acesso ao recurso que esteja gerando o tráfego (como instâncias do Amazon EC2 e contêineres do Amazon ECS). Considere proteger o tráfego de rede para rede com uma rede privada virtual (VPN) IPsec.

Práticas comuns que devem ser evitadas:

- Utilizar versões obsoletas de SSL, TLS e componentes do pacote de criptografia (por exemplo, SSL v3.0, chaves RSA de 1024 bits e criptografia RC4).
- Permitir tráfego não criptografado (HTTP) para ou de recursos voltados para o público.
- Não monitorar e substituir certificados X.509 antes da validade.
- Utilizar certificados X.509 autoassinados para TLS.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Os serviços da AWS fornecem endpoints HTTPS usando TLS para comunicação, fornecendo criptografia em trânsito quando se comunicam com as APIs da AWS. Protocolos não seguros, como HTTP, podem ser auditados e bloqueados em uma VPC por meio do uso de grupos de segurança. As solicitações HTTP também podem ser redirecionadas automaticamente para HTTPS no Amazon CloudFront ou em um Application Load Balancer. Você tem controle total sobre seus recursos de computação para implementar a criptografia em trânsito em seus serviços. Além disso, você pode usar a conectividade de VPN em sua VPC a partir de uma rede externa ou do AWS Direct Connect para facilitar a criptografia do tráfego. Verifique se seus clientes estão fazendo chamadas para APIs da AWS usando pelo menos o TLS 1.2, já que a AWS descontinuou o uso de versões anteriores do TLS em junho de 2023. A AWS recomenda utilizar o TLS 1.3. Soluções de terceiros estão disponíveis no AWS Marketplace, caso você tenha requisitos especiais.

Etapas de implementação

- Aplique a criptografia em trânsito: os requisitos de criptografia definidos devem se basear nos mais recentes padrões e práticas recomendadas e permitir apenas protocolos seguros. Por exemplo,

configure um grupo de segurança para permitir o protocolo HTTPS apenas para a um Application Load Balancer ou instância do Amazon EC2.

- Configure protocolos seguros em serviços de borda: configure o HTTPS com o Amazon CloudFront e use um perfil de segurança apropriado para sua postura de segurança e caso de uso.
- Use uma VPN para conectividade externa: considere usar uma VPN IPsec para proteger conexões ponto a ponto ou rede a rede para ajudar a garantir a privacidade e a integridade dos dados.
- Configure protocolos seguros em balanceadores de carga: selecione uma política de segurança que forneça os pacotes de criptografia mais fortes suportados pelos clientes que se conectarão ao receptor. Crie um receptor HTTPS para seu Application Load Balancer.
- Configure protocolos seguros no Amazon Redshift: configure seu cluster para exigir uma conexão Secure Socket Layer (SSL) ou Transport Layer Security (TLS).
- Configure protocolos seguros: revise a documentação do serviço da AWS para determinar os recursos de criptografia em trânsito.
- Configure o acesso seguro ao fazer o upload para buckets do Amazon S3: use os controles de política do bucket do Amazon S3 para impor o acesso seguro aos dados.
- Considere usar o AWS Certificate Manager: o ACM permite que você provisione, gerencie e implante certificados TLS públicos para uso com serviços da AWS.
- Considere usar o AWS Private Certificate Authoritypara necessidades de PKI privado: a AWS Private CA permite criar hierarquias de autoridade de certificação (CA) privada para emitir certificados X.509 de entidade final que podem ser usados para criar canais TLS criptografados.

Recursos

Documentos relacionados:

- Usar HTTPS com o CloudFront
- Conectar sua VPC a redes remotas usando a AWS Virtual Private Network
- Criar um receptor HTTPS para seu Application Load Balancer
- Tutorial: configurar o SSL/TLS no Amazon Linux 2
- Usar SSL/TLS para criptografar uma conexão com uma instância de um banco de dados
- Configurar as opções de segurança para conexões

SEC09-BP03 Autenticar as comunicações de rede

Verifique a identidade das comunicações usando protocolos que oferecem suporte à autenticação, como Transport Layer Security (TLS) ou IPsec.

Projete a workload para usar protocolos de rede seguros e autenticados sempre que uma comunicação entre serviços, aplicações ou usuários for feita. O uso de protocolos de rede compatíveis com a autenticação e a autorização fornece maior controle sobre os fluxos de rede e reduz o impacto causado por acessos não autorizados.

Resultado desejado: uma workload com fluxos de tráfego de plano de dados e ambiente de gerenciamento bem definidos entre os serviços. Os fluxos de tráfego usam protocolos de rede autenticados e criptografados quando tecnicamente viável.

Práticas comuns que devem ser evitadas:

- Fluxos de tráfego não criptografados ou não autenticados na workload.
- Reutilizar credenciais de autenticação em vários usuários ou entidades.
- Confiar apenas nos controles de rede como um mecanismo de controle de acesso.
- Criar um mecanismo de autenticação personalizado em vez de depender de mecanismos de autenticação padrão do setor.
- Fluxos de tráfego excessivamente permissivos entre componentes de serviço ou outros recursos na VPC.

Benefícios de implementar esta prática recomendada:

- Limita o escopo do impacto do acesso não autorizado a uma parte da workload.
- Fornece um nível mais alto de garantia de que as ações são executadas somente por entidades autenticadas.
- Melhora o desacoplamento de serviços definindo e aplicando claramente as interfaces de transferência de dados pretendidas.
- Melhora o monitoramento, o log e a resposta a incidentes por meio da atribuição de solicitações e interfaces de comunicação bem definidas.
- Oferece defesa profunda para as workloads combinando controles de rede com controles de autenticação e de autorização.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Os padrões de tráfego de rede da workload podem ser caracterizados em duas categorias:

- O tráfego leste-oeste representa fluxos de tráfego entre serviços que compõem uma workload.
- O tráfego norte-sul representa fluxos de tráfego entre a workload e os consumidores.

Embora seja uma prática comum criptografar o tráfego norte-sul, é menos comum proteger o tráfego leste-oeste usando protocolos autenticados. As práticas modernas de segurança recomendam que o design da rede por si só não conceda um relacionamento confiável entre duas entidades. Quando dois serviços podem residir dentro de um limite de rede comum, criptografar, autenticar e autorizar as comunicações ainda são práticas recomendadas entre esses serviços.

Como exemplo, as APIs de serviço da AWS usam o protocolo de assinatura AWS Signature Version 4 (SigV4) para autenticar o chamador, independentemente da rede de origem da solicitação. Essa autenticação garante que as APIs da AWS possam verificar a identidade que solicitou a ação e que essa identidade possa ser combinada com políticas para tomar uma decisão de autorização a fim de determinar se a ação deve ser permitida ou não.

Serviços como o Amazon VPC Lattice e o Amazon API Gateway permitem que você use o mesmo protocolo de assinatura SigV4 para adicionar autenticação e autorização ao tráfego leste-oeste em suas próprias workloads. Se recursos fora do seu ambiente da AWS precisarem se comunicar com serviços que exigem autenticação e autorização baseadas em SIGV4, é possível usar o AWS Identity and Access Management (IAM) Roles Anywhere no recurso externo à AWS para adquirir credenciais temporárias da AWS. Essas credenciais podem ser usadas para assinar solicitações para serviços que usam o SigV4 para autorizar o acesso.

Outro mecanismo comum para autenticar o tráfego leste-oeste é a autenticação mútua TLS (mTLS). Muitas aplicações da Internet das Coisas (IoT), aplicações business to business e microsserviços usam o mTLS para validar a identidade de ambos os lados de uma comunicação TLS por meio do uso de certificados X.509 do lado do cliente e do servidor. Esses certificados podem ser emitidos pela AWS Private Certificate Authority (AWS Private CA). Você pode usar serviços como o Amazon API Gateway e o AWS App Mesh para fornecer autenticação mTLS para comunicação entre workloads ou dentro de uma mesma workload. Embora o mTLS forneça informações de autenticação aos dois lados de uma comunicação TLS, ele não fornece um mecanismo de autorização.

Por fim, o OAuth 2.0 e o OpenID Connect (OIDC) são dois protocolos normalmente usados para controlar o acesso dos usuários aos serviços, mas agora também estão se tornando populares

para o tráfego entre serviços. O API Gateway fornece um autorizador JSON Web Token (JWT), permitindo que as workloads restrinjam o acesso às rotas de API usando JWTs emitidos por provedores de identidade OIDC ou OAuth 2.0. Os escopos do OAuth2 podem ser usados como uma fonte para decisões básicas de autorização, mas as verificações de autorização ainda precisam ser implementadas na camada da aplicação, e os escopos do OAuth2 em si não atendem a necessidades de autorização mais complexas.

Etapas de implementação

- Defina e documente seus fluxos de rede de workload: a primeira etapa na implementação de uma estratégia de defesa aprofundada é definir os fluxos de tráfego da sua workload.
  - Crie um diagrama de fluxo de dados que defina claramente como os dados são transmitidos entre os diferentes serviços que compõem a workload. Esse diagrama é a primeira etapa para aplicar esses fluxos por meio de canais de rede autenticados.
  - Instrumente a workload nas fases de desenvolvimento e testes para validar se o diagrama de fluxo de dados reflete com precisão o comportamento da workload em tempo de execução.
  - Um diagrama de fluxo de dados também pode ser útil ao realizar uma simulação de modelagem de ameaças, conforme descrito em SEC01-BP07 Identificar ameaças e priorizar mitigações usando um modelo de ameaça.
- Estabeleça controles de rede: considere os recursos da AWS para estabelecer controles de rede alinhados aos seus fluxos de dados. Embora os limites da rede não devam ser o único controle de segurança, eles fornecem uma camada na estratégia de defesa profunda para proteger a workload.
  - Use grupos de segurança para estabelecer, definir e restringir fluxos de dados entre recursos.
  - Considere usar o AWS PrivateLink para se comunicar com a AWS e com serviços de terceiros compatíveis com o AWS PrivateLink. Os dados enviados por meio de um endpoint da interface do AWS PrivateLink permanecem na estrutura da rede da AWS e não atravessam a Internet pública.
- Implemente autenticação e autorização em todos os serviços em sua workload: escolha o conjunto de serviços da AWS mais adequado para fornecer fluxos de tráfego autenticados e criptografados em sua workload.
  - Considere o Amazon VPC Lattice para proteger a comunicação entre serviços. O VPC Lattice pode usar a autenticação SigV4 combinada com políticas de autenticação para controlar o acesso entre serviços.

  - Para comunicação entre serviços usando mTLS, considere o API Gateway ou o App Mesh. A AWS Private CA pode ser usada para estabelecer uma hierarquia de CA privada capaz de emitir certificados para uso com mTLS.
  - Ao fazer a integração com serviços usando OAuth 2.0 ou OIDC, considere o API Gateway usando o autorizador JWT.
  - Para comunicação entre sua workload e dispositivos de IoT, considere usar o AWS IoT Core, que fornece várias opções para criptografia e autenticação de tráfego de rede.
- Monitore o acesso não autorizado: monitore continuamente canais de comunicação não intencionais, entidades principais não autorizadas tentando acessar recursos protegidos e outros padrões de acesso impróprios.
  - Se estiver usando o VPC Lattice para gerenciar o acesso aos seus serviços, considere habilitar e monitorar os logs de acesso do VPC Lattice. Esses logs de acesso incluem informações sobre a entidade solicitante, informações de rede que incluem a VPC de origem e de destino e os metadados da solicitação.
  - Considere habilitar os Logs de fluxo da VPC para capturar metadados em fluxos de rede e analisar periodicamente se há anomalias.
  - Consulte o Guia de resposta a incidentes de segurança da AWS e a seção Resposta a Incidentes do pilar Segurança do AWS Well-Architected Framework para obter mais orientações sobre planejamento, simulação e resposta a incidentes de segurança.

Recursos

Práticas recomendadas relacionadas:

- SEC03-BP07 Analisar o acesso público e entre contas
- SEC02-BP02 Usar credenciais temporárias
- SEC01-BP07 Identificar ameaças e priorizar mitigações usando um modelo de ameaça

Documentos relacionados:

- Avaliar métodos de controle de acesso para proteger as APIs do Amazon API Gateway
- Configurar a autenticação TLS mútua para uma API REST
- Como proteger endpoints HTTP do API Gateway com o autorizador JWT
- Autorizar chamadas diretas para serviços da AWS usando o provedor de credenciais do AWS IoT Core

- Guia de resposta a incidentes de segurança da AWS

Vídeos relacionados:

- AWS re:invent 2022: Introdução ao VPC Lattice
- AWS re:Invent 2020: Autenticação da API sem servidor para APIs HTTP na AWS

Exemplos relacionados:

- Workshop sobre Amazon VPC Lattice
- Zero-Trust Episódio 1: workshop sobre o Phantom Service Perimeter

Resposta a incidentes

Pergunta

- SEC 10. Como prever, responder e se recuperar de incidentes?

SEC 10. Como prever, responder e se recuperar de incidentes?

Mesmo com controles preventivos e de detecção consolidados, sua organização deve implementar mecanismos para responder e atenuar o impacto potencial de incidentes de segurança. Sua preparação afeta muito a capacidade das equipes operarem efetivamente durante um incidente, isolarem, conterem e analisarem problemas e restaurarem as operações para um estado adequado conhecido. Implementar as ferramentas e o acesso antes de um incidente de segurança e praticar rotineiramente game days para validar a resposta a incidentes ajudam a garantir que você possa se recuperar enquanto minimiza interrupções empresariais.

Práticas recomendadas

- SEC10-BP01 Identificar equipes e recursos externos fundamentais
- SEC10-BP02 Desenvolver planos de gerenciamento de incidentes
- SEC10-BP03 Preparar recursos forenses
- SEC10-BP04 Desenvolver e testar playbooks de resposta a incidentes de segurança
- SEC10-BP05 Provisionar acesso previamente
- SEC10-BP06 Implantar ferramentas previamente

- SEC10-BP07 Executar simulações
- SEC10-BP08 Estabelecer um framework para aprender com os incidentes

SEC10-BP01 Identificar equipes e recursos externos fundamentais

Identifique as equipes, as obrigações legais e os recursos internos e externos que ajudam sua organização a responder a um incidente.

Resultado desejado: você tem uma lista dos principais funcionários, suas informações de contato e as funções que eles desempenham ao responder a um evento de segurança. Você revisa essas informações regularmente e as atualiza para refletir mudanças de equipes do ponto de vista das ferramentas internas e externas. Você considera todos os provedores de serviços e fornecedores terceirizados ao documentar essas informações, incluindo parceiros de segurança, provedores de nuvem e aplicações de software como serviço (SaaS). Durante um evento de segurança, as equipes estão preparadas com o nível apropriado de responsabilidade, contexto e acesso para resposta e recuperação.

Práticas comuns que devem ser evitadas:

- Não manter uma lista atualizada das principais equipes com informações de contato, funções e responsabilidades para responder a eventos de segurança.
- Supor que todos saibam quais são as pessoas, as dependências, a infraestrutura e as soluções necessárias para resposta a um evento e recuperação.
- Não ter um documento ou repositório de conhecimentos que represente a infraestrutura principal ou o design da aplicação.
- Não ter processos de integração adequados para que novos funcionários contribuam eficazmente para uma resposta a eventos de segurança, como a realização de simulações de eventos.
- Não ter um caminho de encaminhamento estabelecido quando as equipes principais estão temporariamente indisponíveis ou não respondem durante eventos de segurança.

Benefícios de implementar esta prática recomendada: essa prática reduz a triagem e o tempo de resposta gastos na identificação do pessoal certo e de seus perfis durante um evento. Minimize o tempo perdido durante um evento mantendo uma lista atualizada das principais equipes e de suas funções para que você possa convocar as pessoas certas para a triagem e se recuperar de um evento.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Identifique o pessoal-chave em sua organização: mantenha uma lista de contatos do pessoal de sua organização que você precisa envolver. Revise e atualize regularmente essas informações em caso de movimentação de equipes, como mudanças organizacionais, promoções e mudanças de equipe. Isso é especialmente importante para funções importantes, como gerentes de incidentes, respondedores a incidentes e líderes de comunicação.

- Gerente de incidentes: os gerentes de incidentes têm autoridade geral durante a resposta ao evento.
- Respondedores a incidentes: os respondedores a incidentes são responsáveis pelas atividades de investigação e correção. Essas pessoas podem diferir com base no tipo de evento, mas normalmente são desenvolvedores e equipes operacionais responsáveis pela aplicação afetada.
- Líder de comunicação: o líder de comunicação é responsável pelas comunicações internas e externas, especialmente com órgãos públicos e reguladores e clientes.
- Especialistas no assunto (SME): no caso de equipes distribuídas e autônomas, recomendamos que você identifique um SME para workloads de missão crítica. Eles oferecem insights sobre a operação e a classificação de dados das workloads críticas envolvidas no evento.

Considere usar o recurso AWS Systems Manager Incident Manager para capturar os principais contatos, definir um plano de resposta, automatizar cronogramas de plantão e criar planos de escalação. Automatize e faça a rotação de toda a equipe por meio de um cronograma de plantão para que a responsabilidade pela workload seja compartilhada entre os respectivos responsáveis. Isso favorece boas práticas, como emitir métricas e logs relevantes e definir limites de alarme importantes para a workload.

Identifique parceiros externos: as empresas usam ferramentas criadas por provedores de software independentes (ISVs), parceiros e subcontratados para criar soluções diferenciadas para seus clientes. Envolva as principais equipes dessas partes que possam ajudar na resposta e recuperação de um incidente. Recomendamos se inscrever no nível apropriado do AWS Support para obter acesso imediato a especialistas da AWS por meio de um caso de suporte. Considere acordos semelhantes com todos os provedores de soluções essenciais para as workloads. Alguns eventos de segurança exigem que as empresas de capital aberto notifiquem os órgãos públicos e reguladores relevantes sobre o evento e os impactos. Mantenha e atualize as informações de contato dos departamentos relevantes e das pessoas responsáveis.

Etapas de implementação

1. Configure uma solução de gerenciamento de incidentes.
   a. Considere implantar o Incident Manager em sua conta de ferramentas de segurança.
2. Defina contatos em sua solução de gerenciamento de incidentes.
   a. Defina pelo menos dois tipos de canal de contato para cada contato (como SMS, telefone ou e-mail) para garantir a acessibilidade durante um incidente.
3. Defina um plano de resposta.
   a. Identifique os contatos mais adequados a serem mobilizados durante um incidente. Defina planos de encaminhamento alinhados às funções das equipes a serem mobilizadas, em vez de contatos individuais. Considere incluir contatos que possam ter a responsabilidade de informar entidades externas, mesmo que eles não sejam diretamente mobilizados para resolver o incidente.

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP03 Atividades de operações com proprietários identificados responsáveis pela performance

Documentos relacionados:

- Guia de resposta a incidentes de segurança da AWS

Exemplos relacionados:

- Framework do playbook do cliente da AWS
- Como se preparar e responder a incidentes de segurança no ambiente da AWS

Ferramentas relacionadas:

- AWS Systems Manager Incident Manager

Vídeos relacionados:

- A abordagem da Amazon à segurança durante o desenvolvimento

SEC10-BP02 Desenvolver planos de gerenciamento de incidentes

O primeiro documento a ser desenvolvido para resposta a incidentes é o plano de resposta a incidentes. O plano de resposta a incidentes foi projetado para ser a base de seu programa e estratégia de resposta a incidentes.

Benefícios de implementar esta prática recomendada: o desenvolvimento de processos de resposta a incidentes completos e claramente definidos é fundamental para um programa de resposta a incidentes bem-sucedido e escalável. Quando um evento de segurança ocorre, etapas e fluxos de trabalho claros poderão ajudar você a responder em tempo hábil. Talvez você já tenha processos de resposta a incidentes existentes. Independentemente do seu estado atual, é importante atualizar, repetir e testar seus processos de resposta a incidentes regularmente.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Um plano de gerenciamento de incidentes é fundamental para responder, mitigar e se recuperar de possíveis impactos de incidentes de segurança. Um plano de gerenciamento de incidentes é um processo estruturado de identificação, correção e resposta em tempo hábil a incidentes de segurança.

A nuvem tem muitos dos mesmos requisitos e perfis operacionais encontrados em um ambiente on-premises. Ao criar um plano de gerenciamento de incidentes, é importante definir estratégias de resposta e recuperação que se alinhem melhor aos seus resultados empresariais e requisitos de conformidade. Por exemplo, se você opera workloads na AWS em conformidade com o FedRAMP dos Estados Unidos, é útil aderir ao NIST SP 800-61 Guia de tratamento de segurança de computadores. Da mesma forma, ao operar workloads com informações de identificação pessoal (PII) da Europa, considere cenários como a maneira como você deve se proteger e responder a incidentes relacionados à residência de dados, conforme exigido pelo Regulamento Geral de Proteção de Dados (RGPD) da UE.

Ao criar um plano de gerenciamento de incidentes para suas workloads na AWS, comece com o Modelo de responsabilidade compartilhada da AWS para criar uma abordagem de defesa aprofundada para a resposta a incidentes. Nesse modelo, a AWS gerencia a segurança da nuvem, e você é responsável pela segurança na nuvem. Isso significa que você mantém o controle e é responsável pelos controles de segurança que escolhe implementar. O Guia de resposta a

incidentes de segurança da AWS detalha os conceitos e as orientações básicas para criar um plano de gerenciamento de incidentes centrado na nuvem.

Um plano de gerenciamento de incidentes eficaz deve ser continuamente trabalhado e permanecer atualizado com relação às suas metas de operações na nuvem. Considere o uso dos planos de implementação detalhados abaixo à medida que cria e evolui seu plano de gerenciamento de incidentes.

Etapas de implementação

Definir funções e responsabilidades

Lidar com eventos de segurança exige disciplina interorganizacional e uma inclinação para a ação. Em sua estrutura organizacional, deve haver muitas pessoas responsáveis, atribuídas, consultadas ou mantidas informadas durante um incidente, como representantes de recursos humanos (RH), da equipe executiva e do setor jurídico. Considere essas funções e responsabilidades e se algum terceiro deve estar envolvido. Observe que muitas regiões têm leis locais que determinam o que deve e o que não deve ser feito. Embora possa parecer burocrático criar um grafo de pessoas responsáveis, atribuídas, consultadas e informadas (RACI) para seus planos de resposta de segurança, isso facilita a comunicação rápida e direta e descreve claramente a liderança em diferentes estágios do evento.

Durante um incidente, incluir os proprietários e os desenvolvedores de aplicações e recursos afetados é fundamental porque eles são especialistas no assunto (PMEs) que podem fornecer informações e contexto para ajudar a medir o impacto. Pratique e construa relacionamentos com os desenvolvedores e os proprietários de aplicações antes de confiar na experiência deles para responder a incidentes. Proprietários de aplicações ou PMEs, como administradores ou engenheiros de nuvem, podem precisar agir em situações em que o ambiente não seja familiar ou tenha complexidade, ou em que os respondentes não tenham acesso.

Por fim, parceiros confiáveis podem estar envolvidos na investigação ou na resposta, pois podem oferecer experiência adicional e um controle valioso. Se você não tiver essas habilidades em sua própria equipe, contrate uma parte externa para obter assistência.

Entender as equipes de resposta e o suporte da AWS

- AWS Support
  - O AWS Support oferece uma variedade de planos que permitem conceder acesso a ferramentas e conhecimentos que oferecem suporte ao sucesso e à integridade operacional das soluções da

AWS. Se precisar de suporte técnico e mais recursos para ajudar a planejar, implantar e otimizar seu ambiente da AWS, selecione um plano de suporte mais adequado ao seu caso de uso da AWS.

  - Considere o Support Center no AWS Management Console (é necessário iniciar sessão) como ponto central de contato para obter suporte para problemas que afetam seus recursos da AWS. O acesso ao AWS Support é controlado pelo AWS Identity and Access Management. Para mais informações sobre como obter acesso aos recursos da AWS Support, consulte Conceitos básicos do AWS Support.
- Equipe de Resposta a Incidentes de Clientes (CIRT) da AWS
  - A Equipe de Resposta a Incidentes de Clientes (CIRT) da AWS é uma equipe global da AWS especializada que está disponível 24 horas por dia, 7 dias por semana, para prestar assistência aos clientes durante eventos de segurança ativos no lado do cliente do Modelo de responsabilidade compartilhada da AWS.
  - Ao apoiar você, a CIRT da AWS presta assistência na triagem e na recuperação de um evento de segurança ativo na AWS. A equipe pode ajudar na análise da causa-raiz por meio do uso de logs de serviço da AWS e fornecer recomendações para recuperação. Ela também podem fornecer recomendações de segurança e práticas recomendadas para ajudar você a evitar eventos de segurança no futuro.
  - Os clientes da AWS podem solicita a ajuda da CIRT da AWS por meio de um caso do AWS Support
- Suporte de resposta a DDoS
  - A AWS oferece o AWS Shield, que fornece um serviço gerenciado de proteção contra negação de serviço distribuída (DDoS) para proteger aplicações Web executadas na AWS. O Shield fornece detecção permanente e mitigações automáticas em linha que podem minimizar o tempo de inatividade e a latência das aplicações para que não seja necessário envolver o AWS Support para usufruir da proteção contra DDoS. O Shield possui dois níveis: AWS Shield Standard e AWS Shield Advanced. Para saber mais sobre as diferenças entre esses dois níveis, consulte a Documentação de recursos do Shield.
- AWS Managed Services (AMS)
  - O AWS Managed Services (AMS) oferece gerenciamento contínuo de sua infraestrutura da AWS para que você possa se concentrar em suas aplicações. Ao implementar práticas recomendadas para manter sua infraestrutura, o AMS ajuda a reduzir a sobrecarga e os riscos operacionais. O AMS automatiza atividades comuns, como solicitações de alteração, monitoramento, gerenciamento de patches, segurança e serviços de backup, além de disponibilizar serviços de ciclo de vida total para provisionar, executar e apoiar a sua infraestrutura.

- O AMS assume a responsabilidade de implantar um pacote de controles de detecção de segurança e fornece uma primeira linha de resposta aos alertas 24 horas por dia, 7 dias por semana. Quando um alerta é iniciado, o AMS segue um conjunto padrão de guias e playbooks automatizados para verificar uma resposta consistente. Esses playbooks são compartilhados com os clientes do AMS durante a integração para que eles possam desenvolver e coordenar uma resposta com o AMS.

Desenvolva o plano de resposta a incidentes

O plano de resposta a incidentes foi projetado para ser a base de seu programa e estratégia de resposta a incidentes. O plano de resposta a incidentes deve estar em um documento formal. Um plano de resposta a incidentes geralmente inclui as seguintes seções:

- Visão geral da equipe de resposta a incidentes: descreve as metas e funções da equipe de resposta a incidentes.
- Papéis e responsabilidades: lista as partes interessadas na resposta a incidentes e detalha seus papéis quando um incidente ocorre.
- Plano de comunicação: detalha as informações de contato e como você se comunica durante um incidente.
- Métodos de comunicação de backup: é prática recomendada ter comunicação fora de banda como backup para a comunicação de incidentes. Um exemplo de aplicação que fornece um canal seguro de comunicação fora de banda é AWS Wickr.
- Fases da resposta a incidentes e ações necessárias: enumera as fases da resposta a incidentes (por exemplo, detectar, analisar, erradicar, conter e recuperar), incluindo ações de alto nível a serem realizadas nessas fases.
- Definições de severidade e priorização de incidentes: detalha como classificar a severidade de um incidente, como priorizar o incidente e, depois, como as definições de severidade afetam os procedimentos de escalação.

Embora essas seções sejam comuns em empresas de diferentes tamanhos e setores, o plano de resposta a incidentes de cada organização é único. Você precisa criar um plano de resposta a incidentes que funcione melhor para a organização.

Recursos

Práticas recomendadas relacionadas:

- SEC04 (Como você detecta e investiga eventos de segurança?)

Documentos relacionados:

- Guia de resposta a incidentes de segurança da AWS
- NIST: Guia de tratamento de incidentes de segurança de computadores

SEC10-BP03 Preparar recursos forenses

Antes de um incidente de segurança, considere o desenvolvimento de recursos forenses para contribuir com as investigações de eventos de segurança.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Os conceitos da análise forense on-premises tradicional se aplicam à AWS. Para obter informações importantes para começar a desenvolver recursos forenses na Nuvem AWS, consulte Estratégias do ambiente de investigação forense na Nuvem AWS.

Depois de configurar o ambiente e a estrutura da Conta da AWS para análise forense, defina as tecnologias necessárias para executar com eficácia metodologias forenses sólidas nas quatro fases:

- Coleta: colete logs relevantes da AWS, como do AWS CloudTrail, do AWS Config, logs de fluxo de VPC e logs em nível de host. Colete snapshots, backups e despejos de memória dos recursos afetados da AWS, quando disponíveis.
- Exame: examine os dados coletados extraindo e avaliando as informações relevantes.
- Análise: analise os dados coletados para entender o incidente e tirar conclusões do ocorrido.
- Relatório: apresente as informações resultantes da fase de análise.

Etapas de implementação

Prepare o ambiente forense

O AWS Organizations ajuda a gerenciar e reger centralmente um ambiente da AWS à medida que você expande e escala os recursos da AWS. Uma organização da AWS consolida suas Contas da AWS para que você possa administrá-las como uma única unidade. Você pode usar unidades organizacionais (UOs) para agrupar contas e administrá-las como uma unidade única.

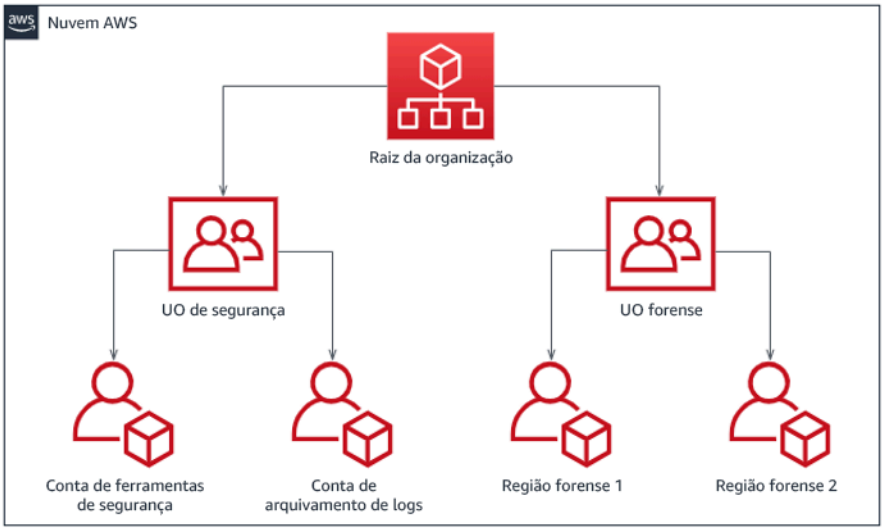
Para resposta a incidentes, é útil ter uma estrutura da Conta da AWS compatível com as funções de resposta a incidentes, que inclui uma UO de segurança e uma UO forense. Dentro da OU de segurança, é necessário ter contas para:

- Arquivamento de logs: agregue logs em uma Conta da AWS de arquivamento de logs com permissões limitadas.
- Ferramentas de segurança: centralize os serviços de segurança em uma Conta da AWS de ferramentas de segurança. Essa conta opera como administrador delegado dos serviços de segurança.

Dentro da UO forense, você tem a opção de implementar uma única conta ou contas forenses para cada região em que opera, dependendo da que funciona melhor para sua empresa e modelo operacional. Se você criar uma conta forense por região, poderá bloquear a criação de recursos da AWS fora dessa região e reduzir o risco de os recursos serem copiados para uma região não pretendida. Por exemplo, se você operasse apenas na região Leste dos EUA (Norte da Virgínia) (`us-east-1`) e Oeste dos EUA (Oregon) (`us-west-2`), você teria duas contas na UO forense: uma para `us-east-1` e outra para `us-west-2`.

É possível criar uma Conta da AWS de análise forense para várias regiões. Você deve ter cuidado ao copiar recursos da AWS para essa conta para verificar se está de acordo com seus requisitos de soberania de dados. Como é preciso tempo para provisionar novas contas, é imperativo criar e instrumentar as contas forenses bem antes de um incidente, para que os respondentes possam estar preparados para usá-las de forma eficaz em suas respostas.

O diagrama a seguir exibe um exemplo de estrutura de contas, incluindo uma UO forense com contas forenses por região:



Estrutura de contas por região para resposta a incidentes

Capture backups e snapshots

Configurar backups dos principais sistemas e bancos de dados é essencial para a recuperação de um incidente de segurança e para fins forenses. Com os backups em vigor, você pode restaurar seus sistemas ao estado seguro anterior. Na AWS, é possível criar snapshots de vários recursos. Os snapshots fornecem backups pontuais desses recursos. Há muitos serviços da AWS que podem ajudar em backup e recuperação. Para obter detalhes sobre esses serviços e abordagens para backup e recuperação, consulte Orientação prescritiva de backup e recuperação e Usar backups para se recuperar de incidentes de segurança.

Especialmente quando se trata de situações como ransomware, é fundamental que os backups estejam bem protegidos. Para obter orientações sobre como proteger backups, consulte As dez principais práticas recomendas de segurança para proteger backups na AWS. Além de proteger os backups, você deve testar regularmente seus processos de backup e restauração para verificar se a tecnologia e os processos implementados funcionam conforme o esperado.

Automatize a análise forense

Durante um evento de segurança, sua equipe de resposta a incidentes deve ser capaz de coletar e analisar evidências rapidamente, mantendo a precisão durante o período em torno do evento (como capturar registros relacionados a um evento ou recurso específico ou coletar o despejo de memória de uma instância do Amazon EC2). É desafiador e demorado para a equipe de resposta a incidentes coletar manualmente as evidências relevantes, especialmente em um grande número de instâncias

e contas. Além disso, a coleta manual pode estar sujeita a erros humanos. Por esses motivos, você deve desenvolver e implementar a automação para perícia forense o máximo possível.

A AWS oferece vários recursos de automação para análise forense, os quais são listados na seção de recursos a seguir. Esses recursos são exemplos de padrões forenses que desenvolvemos e que os clientes implementaram. Embora possam ser uma arquitetura de referência útil para começar, considere modificá-las ou criar padrões de automação forense com base em seu ambiente, requisitos, ferramentas e processos forenses.

Recursos

Documentos relacionados:

- Guia de resposta a incidentes de segurança da AWS: Desenvolver recursos forenses
- Guia de resposta a incidentes de segurança da AWS: Recursos forenses
- Estratégias do ambiente de investigação forense na Nuvem AWS
- Como automatizar a coleta forense de discos na AWS
- Recomendações da AWS: Automatizar a resposta a incidentes e a análise forense

Vídeos relacionados:

- Automatizar a resposta a incidentes e a análise forense

Exemplos relacionados:

- Framework de resposta a incidentes e análise forense automatizadas
- Orquestrador forense automatizado para Amazon EC2

SEC10-BP04 Desenvolver e testar playbooks de resposta a incidentes de segurança

Uma parte fundamental da preparação de seus processos de resposta a incidentes é desenvolver playbooks. Os playbooks de resposta a incidentes fornecem uma série de recomendações e etapas a serem seguidas quando um evento de segurança ocorre. Ter uma estrutura e etapas claras simplifica a resposta e reduz a probabilidade de erro humano.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Os playbooks devem ser criados para cenários de incidentes, como:

- Incidentes esperados: os playbooks devem ser criados para os incidentes previstos. Isso inclui ameaças como negação de serviço (DoS), ransomware e comprometimento de credenciais.
- Descobertas ou alertas de segurança conhecidos: os playbooks devem ser criados para descobertas e alertas de segurança conhecidos, como descobertas do GuardDuty. Você pode receber uma descoberta do GuardDuty e pensar: "E agora?" Para evitar que você trate incorretamente ou ignore uma descoberta do GuardDuty, crie um playbook para cada possível descoberta do GuardDuty. Alguns detalhes e orientações sobre a correção podem ser encontrados na documentação do GuardDuty. É importante notar que o GuardDuty não está habilitado por padrão e seu uso gera custos. Para obter mais detalhes sobre o GuardDuty, consulte o Apêndice A: Definições de capacidade da nuvem: visibilidade e alertas.

Os playbooks devem conter etapas técnicas a serem concluídas por um analista de segurança para investigar e responder adequadamente a um possível incidente de segurança.

Etapas de implementação

Os itens a serem incluídos em um playbook incluem:

- Visão geral do playbook: qual cenário de risco ou incidente esse playbook aborda? Qual é o objetivo do playbook?
- Pré-requisitos: quais logs, mecanismos de detecção e ferramentas automatizadas são necessários para esse cenário de incidente? Qual é a notificação esperada?
- Informações de comunicação e escalação: quem está envolvido e quais são suas informações de contato? Quais são as responsabilidades de cada parte interessada?
- Etapas da resposta: em todas as fases da resposta a incidentes, quais etapas táticas devem ser seguidas? Que consultas um analista deve executar? Que código deve ser executado para alcançar o resultado desejado?
  - Detectar: como o incidente será detectado?
  - Analisar: como o escopo do impacto será determinado?
  - Conter: como o incidente será isolado para limitar o escopo?
  - Erradicar: como a ameaça será removida do ambiente?
  - Recuperar: como o sistema ou o recurso afetado voltará à produção?

- Resultados esperados: depois que as consultas e o código forem executados, qual é o resultado esperado do playbook?

Recursos

Práticas recomendadas do Well-Architected relacionadas:

- SEC10-BP02 Desenvolver planos de gerenciamento de incidentes

Documentos relacionados:

- Framework para playbooks de resposta a incidentes
- Como desenvolver seus próprios playbooks de resposta a incidentes
- Exemplos de playbook de resposta a incidentes
- Criar um runbook de resposta a incidentes da AWS using playbooks do Jupyter e o CloudTrail Lake

SEC10-BP05 Provisionar acesso previamente

Verifique se os respondedores a incidentes têm o acesso correto pré-provisionado na AWS para reduzir o tempo de investigação necessário até a recuperação.

Práticas comuns que devem ser evitadas:

- Uso da conta raiz para a resposta a incidentes.
- Alterar contas de usuário existentes.
- Manipular permissões do IAM diretamente ao fornecer elevação de privilégios just-in-time.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A AWS recomenda reduzir ou eliminar a dependência de credenciais de longa duração sempre que possível, dando preferência a credenciais temporárias e a mecanismos de escalação de privilégios just-in-time. As credenciais de longa duração são propensas a riscos de segurança e aumentam a sobrecarga operacional. Para a maioria das tarefas de gerenciamento, bem como para as tarefas

de resposta a incidentes, recomendamos implementar a federação de identidades junto com a escalação temporária para acesso administrativo. Nesse modelo, um usuário solicita elevação para um nível de privilégio superior (como um perfil de resposta a incidentes) e, considerando que ele seja elegível para a elevação, a solicitação é enviada a um aprovador. Se a solicitação for aprovada, o usuário receberá um conjunto de credenciais da AWS temporárias que podem ser usadas para concluir suas tarefas. Quando essas credenciais expirarem, o usuário deverá enviar uma nova solicitação de elevação.

Recomendamos usar a escalação de privilégio temporária para a maioria dos cenários de resposta a incidentes. A maneira correta de fazer isso é usar o AWS Security Token Service e políticas de sessão para definir o escopo do acesso.

Há cenários em que as identidades federadas não estão disponíveis, como:

- Interrupção relacionada a um provedor de identidades (IdP) comprometido.
- Erro de configuração ou erro humano causando uma falha no sistema de gerenciamento de acesso federado.
- Atividade mal-intencionada, como um evento de negação de serviço distribuído (DDoS) ou que causa a indisponibilidade do sistema.

Nos casos anteriores, deve haver um acesso emergencial de "vidro quebrado" configurado para permitir a investigação e a correção rápida de incidentes. Recomendamos usar um usuário, grupo ou perfil com as permissões apropriadas para realizar tarefas e acessar recursos da AWS. Use as credenciais de usuário-raiz somente para realizar tarefas que exijam as credenciais de usuário-raiz. Para verificar se os respondedores de um incidente têm o nível de acesso correto à AWS e a outros sistemas relevantes, recomendamos provisionar previamente contas dedicadas. As contas exigem acesso privilegiado e devem ser estritamente controladas e monitoradas. As contas devem ser criadas com os privilégios mínimos exigidos para realizar as tarefas necessárias e o nível de acesso deve ser baseado nos playbooks criados como parte do plano de gerenciamento de incidentes.

Como prática recomendada, utilize perfis e usuários dedicados e com propósito específico. Escalar temporariamente o acesso de usuários ou perfis por meio da adição de políticas do IAM não deixa claro qual é o acesso que os usuários tinham durante o incidente, e há um risco de que os privilégios escalados não sejam revogados.

É importante remover o máximo de dependências possível para verificar se o acesso pode ser obtido com o maior número possível de cenários de falha. Como forma de auxiliar esse processo, crie um playbook para verificar se os usuários de resposta a incidentes são criados como usuários em uma

conta de segurança dedicada e não são gerenciados por nenhuma solução de autenticação única (SSO) ou federação existente. Cada respondedor individual deve ter sua própria conta nomeada. A configuração da conta deve aplicar uma política de senha forte e autenticação multifator (MFA). Se os playbooks de resposta a incidentes só exigem acesso ao AWS Management Console, o usuário não deve ter chaves de acesso configuradas e deve ser proibido explicitamente de criar chaves de acesso. Isso pode ser configurado com políticas do IAM ou políticas de controle de serviços (SCPs), conforme mencionado nas Práticas recomendadas de segurança da AWS para SCPs do AWS Organizations Os usuários não devem ter privilégios além da capacidade de assumir perfis de resposta a incidentes em outras contas.

Durante um incidente, talvez seja necessário conceder acesso a outros indivíduos internos ou externos para apoiar a investigação, a correção ou as atividades de recuperação. Nesse caso, use o mecanismo do playbook mencionado anteriormente, e deve haver um processo para verificar se qualquer acesso adicional foi revogado imediatamente após a conclusão do incidente.

Para verificar se o uso de perfis de resposta a incidentes pode ser monitorado e auditado corretamente, é essencial que as contas de usuário do IAM criadas para esse fim não sejam compartilhadas entre indivíduos e que o Usuário raiz da conta da AWS não seja utilizado, a menos que isso seja necessário para uma tarefa específica. Se o usuário-raiz for necessário (por exemplo, quando o acesso do IAM a uma conta específica estiver indisponível), use um processo separado com um playbook disponível para verificar a disponibilidade das credenciais de início de sessão e do token de MFA do usuário-raiz.

Para configurar as políticas do IAM para os perfis de resposta a incidentes, considere usar o IAM Access Analyzer para gerar políticas com base em logs do AWS CloudTrail. Para fazer isso, conceda acesso de administrador ao perfil de resposta a incidentes em uma conta de não produção e execute as etapas descritas nos playbooks. Concluído o processo, uma política que permita somente as ações realizadas pode ser criada. Essa política pode ser então aplicada a todos os perfis de resposta a incidentes em todas as contas. Você pode criar uma política do IAM separada para cada playbook a fim de facilitar o gerenciamento e a auditoria. Exemplos de playbook podem incluir planos de resposta para ransomware, violações de dados, perda de acesso à produção, entre outros cenários.

Use as contas de resposta a incidentes para assumir perfis do IAM de resposta a incidentes dedicados em outras Contas da AWS Esses perfis também devem ser configurados para poder ser assumidos somente por usuários na conta de segurança, e o relacionamento de confiança deve exigir que a entidade principal que está fazendo a chamada seja autenticada com MFA. Os perfis devem usar políticas do IAM com escopo estritamente definido para controlar o acesso. Certifique-se de que todas as solicitações de `AssumeRole` para esses perfis sejam registradas em

log no CloudTrail e acionem alertas, e que quaisquer ações realizadas usando esses perfis sejam registradas em log.

É altamente recomendável que as contas do IAM e os perfis do IAM sejam claramente nomeados para permitir que sejam encontrados com facilidade nos logs do CloudTrail. Um exemplo disso seria nomear as contas do IAM como *<USER_ID>*-BREAK-GLASS e os perfis do IAM como BREAK-GLASS-ROLE.

O CloudTrail é usado para registrar a atividade da API em suas contas da AWS e deve ser usado para configurar alertas sobre o uso dos perfis de resposta a incidentes. Consulte a publicação do blog sobre como configurar alertas quando as chaves-raiz são usadas. As instruções podem ser modificadas para configurar o filtro métrico do Amazon CloudWatch para filtrar eventos AssumeRole relacionados ao perfil do IAM de resposta a incidentes:

```
{ $.eventName = "AssumeRole" && $.requestParameters.roleArn =
 "<INCIDENT_RESPONSE_ROLE_ARN>" && $.userIdentity.invokedBy NOT EXISTS && $.eventType !
= "AwsServiceEvent" }
```

Como é provável que os perfis de resposta a incidentes tenham um alto nível de acesso, é importante que esses alertas sejam transmitidos a um grupo amplo e que as atitudes necessárias sejam tomadas rapidamente.

Durante um incidente, é possível que um respondedor possa exigir acesso a sistemas que não são protegidos diretamente pelo IAM. Isso pode incluir instâncias do Amazon Elastic Compute Cloud, bancos de dados do Amazon Relational Database Service ou plataformas de software como serviço (SaaS). É altamente recomendável que, em vez de usar protocolos nativos, como SSH ou RDP, o AWS Systems Manager Session Manager seja usado para todo o acesso administrativo às instâncias do Amazon EC2. Esse acesso pode ser controlado usando o IAM, que é seguro e auditado. Talvez também seja possível automatizar partes de seus playbook usando documentos de execução de comandos do AWS Systems Manager, o que pode reduzir os erros do usuário e melhorar o tempo de recuperação. Para acesso aos bancos de dados e a ferramentas de terceiros, recomendamos armazenar as credenciais de acesso no AWS Secrets Manager e conceder acesso aos perfis de respondedores a incidentes.

Por fim, o gerenciamento das contas do IAM de resposta a incidentes deve ser adicionado aos seus processos de Joiners, Movers e Leavers e revisado e testado periodicamente para verificar se somente o acesso pretendido é permitido.

Recursos

Documentos relacionados:

- Gerenciar o acesso elevado temporário ao seu ambiente da AWS
- Guia de resposta a incidentes de segurança da AWS
- AWS Elastic Disaster Recovery
- AWS Systems Manager Incident Manager
- Definir uma política de senhas de contas para usuários do IAM
- Usar a autenticação multifator (MFA) na AWS
- Configurar o acesso entre contas com MFA
- Usar o IAM Access Analyzer para gerar políticas do IAM
- Práticas recomendadas para Políticas de controle de serviços do AWS Organizations em um ambiente com várias contas
- Como receber notificações quando as chaves de acesso raiz da sua conta da AWS são usadas
- Criar permissões de sessão refinadas usando políticas gerenciadas pelo IAM

Vídeos relacionados:

- Automatizar a resposta a incidentes e a análise forense na AWS
- Guia de faça você mesmo para runbooks, relatórios de incidentes e resposta a incidentes
- Como se preparar e responder a incidentes de segurança no ambiente da AWS

Exemplos relacionados:

- Laboratório: configuração da conta da AWS e usuário-raiz
- Laboratório: resposta a incidentes com o console da AWS e a CLI

SEC10-BP06 Implantar ferramentas previamente

Verifique se o pessoal de segurança tem as ferramentas certas pré-implantadas para reduzir o tempo de investigação até a recuperação.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Para automatizar as funções de resposta e operações de segurança, é possível usar um conjunto abrangente de APIs e ferramentas da AWS. Você pode automatizar totalmente os recursos de gerenciamento de identidade, segurança de rede, proteção de dados e monitoramento e disponibilizá-los com métodos populares de desenvolvimento de software já em vigor. Quando você cria a automação da segurança, seu sistema pode monitorar, analisar e iniciar uma resposta, em vez de fazer com que as pessoas monitorem a sua posição de segurança e reajam manualmente a eventos.

Se as equipes de resposta a incidentes continuarem a responder aos alertas da mesma forma, haverá o risco de se acostumarem aos alertas. Com o passar do tempo, a equipe pode se tornar dessensibilizada para alertas e cometer erros ao lidar com situações comuns ou perder alertas incomuns. A automação ajuda a evitar a exaustão de alertas usando funções que processam alertas repetitivos e comuns, permitindo que as pessoas lidem com incidentes confidenciais e exclusivos. A integração de sistemas de detecção de anomalias, como Amazon GuardDuty, AWS CloudTrail Insights e Amazon CloudWatch Anomaly Detection, pode reduzir a carga de alertas baseados em limites comuns.

Você pode melhorar os processos manuais com a automatização programática das etapas do processo. Depois de definir o padrão de correção para um evento, você poderá decompor esse padrão em lógica acionável e desenvolver o código para executar essa lógica. Os respondedores podem executar esse código para corrigir o problema. Com o passar do tempo, você pode automatizar mais e mais etapas e, por fim, lidar automaticamente com classes inteiras de incidentes comuns.

Durante uma investigação de segurança, você precisa ser capaz de revisar os logs relevantes para registrar e compreender o escopo completo e o cronograma do incidente. Os logs também são necessários para geração de alertas indicando que determinadas ações de interesse ocorreram. É essencial selecionar, ativar, armazenar e configurar mecanismos de consulta e recuperação, bem como definir alertas. Além disso, uma forma eficaz de fornecer ferramentas para pesquisar dados de log é o Amazon Detective.

A AWS oferece mais de 200 serviços em nuvem e milhares de recursos. Recomendamos que você analise os serviços que podem apoiar e simplificar sua estratégia de resposta a incidentes.

Além do registro, você deve desenvolver e implementar uma estratégia de marcação. A marcação pode ajudar a fornecer contexto sobre a finalidade de um recurso da AWS. A marcação também pode ser usada para automação.

Etapas de implementação

Selecione e configure logs para análise e alertas

Consulte a documentação a seguir sobre como configurar logs para resposta a incidentes:

- Estratégias de log para resposta a incidentes de segurança
- SEC04-BP01 Configurar o registro em log de serviços e aplicações

Habilite serviços de segurança para oferecer suporte a detecção e resposta

A AWS fornece recursos nativos de detecção, prevenção e resposta, e outros serviços podem ser usados para arquitetar soluções de segurança personalizadas. Para obter uma lista dos serviços mais relevantes para resposta a incidentes de segurança, consulte Definições de capacidade de nuvem.

Desenvolva e implemente uma estratégia de marcação

Obter informações contextuais sobre o caso de uso empresarial e as partes interessadas internas relevantes em torno de um recurso da AWS pode ser difícil. Uma forma de fazer isso é na forma de tags, que atribuem metadados aos recursos da AWS e consistem em uma chave e um valor definidos pelo usuário. Você pode criar tags para categorizar os recursos por finalidade, proprietário, ambiente, tipo de dados processados e outros critérios de sua escolha.

Ter uma estratégia de marcação consistente pode acelerar os tempos de resposta e minimizar o tempo gasto no contexto organizacional, permitindo identificar e discernir rapidamente as informações contextuais sobre um recurso da AWS. As tags também podem servir como um mecanismo para iniciar automações de resposta. Para obter mais detalhes sobre o que marcar, consulte Como marcar seus recursos da AWS. Primeiro, você deve definir as tags que deseja implementar em toda a sua organização. Depois disso, você implementará e aplicará essa estratégia de marcação. Para obter mais detalhes sobre implementação e fiscalização, consulte Implementar a estratégia de marcação de recursos da AWS usando políticas de tags e políticas de controle de serviços (SCPs) da AWS.

Recursos

Práticas recomendadas do Well-Architected relacionadas:

- SEC04-BP01 Configurar o registro em log de serviços e aplicações
- SEC04-BP02 Capturar logs, descobertas e métricas em locais padronizados

Documentos relacionados:

- Estratégias de log para resposta a incidentes de segurança
- Definições de recursos de nuvem para resposta a incidentes

Exemplos relacionados:

- Detecção e resposta a ameaças com o Amazon GuardDuty e o Amazon Detective
- Workshop do Security Hub
- Gerenciamento de vulnerabilidades com o Amazon Inspector

SEC10-BP07 Executar simulações

À medida que as organizações crescem e evoluem com o tempo, o mesmo acontece com o cenário de ameaças, o que torna importante analisar continuamente seus recursos de resposta a incidentes. Executar simulações (também conhecidas como "game days") é um método que pode ser usado para realizar essa avaliação. As simulações usam cenários de eventos de segurança do mundo real projetados para imitar as táticas, as técnicas e os procedimentos (TTPs) de um agente de ameaças e permitir que uma organização exercite e avalie seus recursos de resposta a incidentes respondendo a esses eventos cibernéticos simulados da mesma forma que em uma situação real.

Benefícios de implantar esta prática recomendada: as simulações trazem vários benefícios:

- Validar a prontidão cibernética e desenvolver a confiança de seus socorristas.
- Testar a precisão e a eficiência de ferramentas e fluxos de trabalho.
- Refinar os métodos de comunicação e escalação alinhados ao seu plano de resposta a incidentes.
- Proporcionar uma oportunidade de responder a vetores menos comuns.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Existem três tipos principais de simulações:

- Simulações teóricas: a abordagem de simulações teóricas é uma sessão baseada em discussões que envolvem as várias partes interessadas na resposta a incidentes para exercer funções e responsabilidades e usar ferramentas de comunicação e playbooks estabelecidos. A facilitação

das simulações normalmente pode ser realizada em um dia inteiro em um local virtual, local físico ou uma combinação de ambos. Por ser baseada em discussões, a simulação teórica se concentra em processos, pessoas e colaboração. A tecnologia é parte integrante da discussão, mas o uso real de ferramentas ou scripts de resposta a incidentes geralmente não faz parte da simulação teórica.

- Simulações da equipe roxa: As simulações da equipe roxa aumentam o nível de colaboração entre os respondedores ao incidente (equipe azul) e os agentes de ameaças simuladas (equipe vermelha). A equipe azul é composta por membros do centro de operações de segurança (SOC), mas também pode incluir outras partes interessadas que estariam envolvidas durante um evento cibernético real. A equipe vermelha é composta por uma equipe de testes de penetração ou pelas principais partes interessadas treinadas em segurança ofensiva. A equipe vermelha trabalha em colaboração com os facilitadores da simulação ao projetar um cenário que seja preciso e viável. Durante as simulações da equipe roxa, o foco principal está nos mecanismos de detecção, nas ferramentas e nos procedimentos operacionais padrão (SOPs) que apoiam os esforços de resposta a incidentes.
- Simulações da equipe vermelha: durante uma simulação da equipe vermelha, o infrator (equipe vermelha) realiza uma simulação para atingir um determinado objetivo ou conjunto de objetivos a partir de um escopo predeterminado. Os defensores (equipe azul) não necessariamente terão conhecimento do escopo e da duração da simulação, o que oferece uma avaliação mais realista de como eles responderiam a um incidente real. Como as simulações da equipe vermelha podem ser testes invasivos, tenha cuidado e implemente controles para verificar se a simulação não causa danos reais ao ambiente.

Considere facilitar as simulações cibernéticas em intervalos regulares. Cada tipo de simulação pode oferecer benefícios exclusivos aos participantes e à organização como um todo. Portanto, você pode optar por começar com tipos de simulação menos complexos (como simulações teóricas) e avançar para tipos de simulação mais complexos (simulações da equipe vermelha). Você deve selecionar um tipo de simulação com base em sua maturidade de segurança, recursos e resultados desejados. Alguns clientes podem não optar por realizar simulações da equipe vermelha devido à complexidade e ao custo.

Etapas de implementação

Independentemente do tipo de simulação que você escolher, as simulações geralmente seguem estas etapas de implementação:

1. Defina os elementos fundamentais da simulação: defina o cenário e os objetivos da simulação. Ambos devem ter aceitação da equipe de liderança.
2. Identifique as principais partes interessadas: no mínimo, a simulação precisa de facilitadores e participantes. Dependendo do cenário, outras partes interessadas, como departamento jurídico, de comunicação ou liderança executiva, podem estar envolvidos.
3. Crie e teste o cenário: talvez o cenário precise ser redefinido durante a criação se elementos específicos não forem viáveis. Espera-se um cenário finalizado como resultado dessa etapa.
4. Facilite a simulação: o tipo de simulação determina a facilitação usada (um cenário impresso em comparação a um cenário simulado altamente técnico). Os facilitadores devem alinhar suas táticas de facilitação aos objetos da simulação e envolver todos os participantes sempre que possível para proporcionar o máximo benefício.
5. Desenvolva o relatório pós-ação (AAR): identifique as áreas de sucesso, aquelas que podem ser melhoradas e possíveis lacunas. O AAR deve medir a eficácia da simulação, bem como a resposta da equipe ao evento simulado, para que o progresso possa ser monitorado ao longo do tempo com simulações futuras.

Recursos

Documentos relacionados:

- Guia de resposta a incidentes da AWS

Vídeos relacionados:

- AWS GameDay: edição de segurança

SEC10-BP08 Estabelecer um framework para aprender com os incidentes

Implementar um framework de lições aprendidas e o recurso de análise da causa-raiz não só ajudará a melhorar os recursos de resposta a incidentes, mas também a evitar que o incidente se repita. Ao aprender com cada incidente, você pode ajudar a evitar a repetição dos mesmos erros, exposições ou configurações incorretas, não apenas melhorando seu procedimento de segurança, mas também minimizando o tempo perdido em situações evitáveis.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

É importante implementar um framework de lições aprendidas que estabeleça e atinja, em alto nível, os seguintes pontos:

- Quando um processo de lições aprendidas é realizado?
- O que está envolvido no processo de lições aprendidas?
- Como um processo de lições aprendidas é realizado?
- Quem está envolvido no processo e como?
- Como as áreas de melhoria serão identificadas?
- Como você garantirá que as melhorias sejam monitoradas e implementadas de forma eficaz?

O framework não deve se concentrar em culpar os indivíduos, mas sim na melhoria de ferramentas e processos.

Etapas de implementação

Além dos resultados de alto nível listados acima, é importante garantir que você faça as perguntas certas para obter o máximo valor (informações que levem a melhorias práticas) do processo. Considere estas perguntas para ajudar você a começar a promover discussões sobre lições aprendidas:

- Como foi o incidente?
- Quando o incidente foi identificado pela primeira vez?
- Como ele foi identificado?
- Que sistemas alertaram sobre a atividade?
- Que sistemas, serviços e dados estiveram envolvidos?
- O que ocorreu especificamente?
- O que funcionou bem?
- O que não funcionou bem?
- Que processos ou procedimentos falharam ou não tiveram a escala ajustada para responder ao incidente?
- O que pode ser melhorado nas seguintes áreas:
  - Pessoas

    - As pessoas que precisavam ser contatadas estavam realmente disponíveis e a lista de contatos estava atualizada?
    - As pessoas estavam perdendo treinamentos ou não tinham os recursos necessários para responder e investigar o incidente de forma eficaz?
    - Os recursos apropriados estavam prontos e disponíveis?
- Processo
    - Os processos e procedimentos foram seguidos?
    - Os processos e procedimentos foram documentados e estavam disponíveis para esse (tipo de) incidente?
    - Havia processos e procedimentos necessários faltando?
    - Os respondedores conseguiram obter acesso oportuno às informações necessárias para responder ao problema?
- Tecnologia
    - Os sistemas de alerta existentes identificaram e alertaram efetivamente sobre a atividade?
    - Como poderíamos ter reduzido o tempo de detecção em 50%?
    - Os alertas existentes precisam ser aprimorados ou novos alertas precisam ser criados para esse (tipo de) incidente?
    - As ferramentas existentes permitiram uma investigação (pesquisa/análise) eficaz do incidente?
    - O que pode ser feito para ajudar a identificar esse (tipo de) incidente mais cedo?
    - O que pode ser feito para ajudar a evitar que esse (tipo de) incidente ocorra novamente?
    - Quem é o proprietário do plano de melhoria e como você testará se ele foi implementado?
    - Qual é o cronograma para que os controles e processos adicionais de monitoramento ou prevenção sejam implementados e testados?

Essa lista não inclui tudo, mas serve como ponto de partida para identificar quais são as necessidades da organização e da empresa e como você pode analisá-las para aprender com os incidentes de forma mais eficaz e melhorar constantemente seu procedimento de segurança. O mais importante é começar incorporando as lições aprendidas como parte padrão do processo de resposta a incidentes, da documentação e das expectativas das partes interessadas.

Recursos

Documentos relacionados:

- Guia de resposta a incidentes da AWS: estabelecer um framework para aprender com os incidentes
- Orientações do NCSC CAF: lições aprendidas

Segurança de aplicações

Pergunta

- SEC 11. Como incorporar e validar as propriedades de segurança de aplicações durante o ciclo de vida de design, desenvolvimento e implantação?

SEC 11. Como incorporar e validar as propriedades de segurança de aplicações durante o ciclo de vida de design, desenvolvimento e implantação?

Treinar a equipe, testar por meio da automação, entender as dependências e validar as propriedades de segurança de ferramentas e aplicações ajuda a diminuir a probabilidade de problemas de segurança em workloads de produção.

Práticas recomendadas

- SEC11-BP01 Treinar para segurança de aplicações
- SEC11-BP02 Automatizar o teste durante o ciclo de vida de desenvolvimento e lançamento
- SEC11-BP03 Realizar teste de penetração regular
- SEC11-BP04 Análises manuais de código
- SEC11-BP05 Centralizar serviços para pacotes e dependências
- SEC11-BP06 Implantar software programaticamente
- SEC11-BP07 Avaliar regularmente as propriedades de segurança dos pipelines
- SEC11-BP08 Criar um programa que incorpore a propriedade de segurança nas equipes de workload

SEC11-BP01 Treinar para segurança de aplicações

Forneça treinamento aos criadores em sua organização sobre práticas comuns para promover a segurança no desenvolvimento e na operação de aplicações. A adoção de práticas de desenvolvimento com foco na segurança ajuda a diminuir a probabilidade de problemas que são detectados somente no estágio de avaliação da segurança.

Resultado desejado: o software deve ser projetado e construído com a segurança em mente. Quando os criadores em uma organização são treinados em práticas de desenvolvimento seguras que começam com um modelo de ameaças, isso melhora a qualidade e a segurança gerais do software produzido. Essa abordagem pode reduzir o tempo de entrega do software ou de recursos porque não é necessário tanto retrabalho após o estágio de avaliação da segurança.

Para fins dessa prática recomendada, desenvolvimento seguro refere-se ao software que está sendo escrito e às ferramentas ou sistemas que suportam o ciclo de vida de desenvolvimento de software (SDLC).

Práticas comuns que devem ser evitadas:

- Aguardar uma avaliação da segurança e, depois, considerar as propriedades de segurança de um sistema.
- Deixar todas as decisões de segurança para a equipe de segurança.
- Não comunicar como as decisões tomadas no SDLC se relacionam às expectativas ou as políticas de segurança gerais da organização.
- Iniciar o processo de avaliação da segurança muito tardiamente.

Benefícios de implementar esta prática recomendada:

- Melhor conhecimento dos requisitos organizacionais para a segurança na fase inicial do ciclo de desenvolvimento.
- Ser capaz de identificar e solucionar possíveis problemas de segurança com maior rapidez, promovendo uma entrega de recursos mais rápida.
- Maior qualidade do software e dos sistemas.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Ofereça treinamento aos criadores em sua organização. Começar com um curso sobre modelagem de ameaças é uma boa base para ajudar no treinamento em segurança. Preferencialmente, os criadores devem ser capazes de acessar de forma independente as informações relevantes às respectivas workloads. Esse acesso os ajuda a tomar decisões embasadas sobre as propriedades de segurança dos sistemas criados por eles sem a necessidade de solicitar outra equipe. O processo para envolver a equipe de segurança para avaliações deve ser claramente definido e simples de

seguir. As etapas do processo de avaliação devem ser incluídas no treinamento de segurança. Quando houver padrões ou modelos de implementação disponíveis, eles deverão ser simples de encontrar e vincular aos requisitos de segurança gerais. Considere usar o AWS CloudFormation, construtos do AWS Cloud Development Kit (AWS CDK), o Service Catalog ou outras ferramentas de modelagem para reduzir a necessidade de configuração personalizada.

Etapas de implementação

- Inicie os criadores com um curso sobre modelagem de ameaças para criar uma boa base e ajude a treiná-los em como pensar em segurança.
- Forneça acesso a treinamentos da Treinamento da AWS and Certification, do setor ou de parceiros da AWS.
- Forneça treinamento sobre o processo de avaliação da segurança de sua organização, que esclarece a divisão de responsabilidades entre a equipe de segurança, as equipes de workload e outras partes interessadas.
- Publique orientações de autoatendimento sobre como atender aos seus requisitos de segurança, inclusive códigos de exemplo e modelos, se disponíveis.
- Obtenha feedback regularmente de equipes de criadores sobre a experiência deles com o processo e o treinamento de processo de avaliação da segurança e usar esse feedback para promover melhorias.
- Utilize game days ou campanhas de bug bash para ajudar a reduzir o número de problemas e aumentar as habilidades de seus criadores.

Recursos

Práticas recomendadas relacionadas:

- SEC11-BP08 Criar um programa que incorpore a propriedade de segurança nas equipes de workload

Documentos relacionados:

- Treinamento da AWS e certificação
- Como pensar sobre a governança da segurança na nuvem
- Como abordar a modelagem de ameaças
- Acelerar o treinamento — AWS Skills Guild

Vídeos relacionados:

- Segurança proativa: considerações e abordagens

Exemplos relacionados:

- Workshop sobre modelagem de ameaças
- Conscientização do setor para desenvolvedores

Serviços relacionados:

- AWS CloudFormation
- Constructos do AWS Cloud Development Kit (AWS CDK) (AWS CDK)
- Service Catalog
- AWS BugBust

SEC11-BP02 Automatizar o teste durante o ciclo de vida de desenvolvimento e lançamento

Automatize o teste das propriedades de segurança durante o ciclo de vida de desenvolvimento e lançamento. Com a automação, é mais fácil identificar de forma consistente e repetível possíveis problemas no software antes do lançamento, o que reduz o risco de problemas de segurança no software que está sendo fornecido.

Resultado desejado: o objetivo dos testes automatizados é fornecer uma forma programática de detectar possíveis problemas com antecedência e frequência durante todo o ciclo de vida do desenvolvimento. Ao automatizar o teste de regressão, você pode executar novamente testes funcionais e não funcionais para verificar se o software testado anteriormente ainda funciona da forma esperada após uma alteração. Ao definir testes de unidade de segurança para conferir configurações incorretas comuns, como uma autenticação ausente ou danificada, é possível identificar e resolver esses problemas logo no início do processo de desenvolvimento.

A automação de testes utiliza casos de teste para um propósito específico para validação de aplicações, com base nos requisitos e na funcionalidade desejada da aplicação. O resultado dos testes automatizados baseia-se na comparação da saída do teste gerado com a respectiva saída esperada, o que acelera o ciclo de vida dos testes em geral. As metodologias de teste, como teste de regressão e pacotes de teste de unidade, são mais adequadas para automação. A automação dos testes de propriedades de segurança possibilita aos criadores receber feedback automatizado sem

precisar esperar por uma avaliação da segurança. Os testes automatizados em forma de análise de código estático ou dinâmico podem melhorar a qualidade do código e ajudar a detectar possíveis problemas de software no ciclo de vida de desenvolvimento.

Práticas comuns que devem ser evitadas:

- Não comunicar os casos de teste e os resultados dos testes automatizados.
- Realizar os testes automatizados somente antes de um lançamento.
- Automatizar casos de teste com requisitos que mudam com frequência.
- Não fornecer orientações sobre como abordar os resultados dos testes de segurança.

Benefícios de implementar esta prática recomendada:

- Redução da dependência de pessoas que avaliam as propriedades de segurança dos sistemas.
- Descobertas consistentes em vários fluxos de trabalho que melhoram a consistência.
- Redução da probabilidade de introduzir problemas de segurança no software de produção.
- Redução do período de tempo entre a detecção e a correção devido à detecção precoce dos problemas de software.
- Maior visibilidade do problema sistêmico ou repetido entre os vários fluxos de trabalho, o que pode ser utilizado para promover melhorias em toda a organização.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Ao criar um software, adote vários mecanismos de teste para garantir que você esteja testando os requisitos funcionais da aplicação com base na respectiva lógica de negócios e em requisitos não funcionais, os quais se focam a confiabilidade, a performance e a segurança da aplicação.

O teste de segurança de aplicação estática (SAST) analisa padrões de segurança anômalos no código-fonte e fornece indicações de código propenso a defeitos. O SAST depende de entradas estáticas, como documentação (especificação de requisitos, documentação e especificações de design) e código-fonte da aplicação, para testar uma série de problemas de segurança conhecidos. Os analisadores de código estático podem ajudar a acelerar a análise de grandes volumes de código. O NIST Quality Group fornece uma comparação de analisadores de segurança de código-

fonte, que inclui ferramentas de código aberto para scanners de código em bytes e scanners de código binário.

Complemente seu teste estático com metodologias de teste de segurança de análise dinâmica (DAST), as quais realizam testes na aplicação em execução a fim de identificar comportamento possivelmente inesperado. O teste dinâmico pode ser utilizado para detectar possíveis problemas que não são detectáveis por meio de análise estática. Por meio dos testes nos estágios de repositório de código, compilação e pipeline, é possível impedir que diferentes tipos de problema em potencial ocorram no código. O Amazon CodeWhisperer fornece recomendações de código, incluindo verificação de segurança, no IDE do construtor. O Amazon CodeGuru Reviewer pode identificar problemas críticos, problemas de segurança e bugs difíceis de encontrar durante o desenvolvimento da aplicação e fornece recomendações para melhorar a qualidade do código.

O workshop Segurança para desenvolvedores usa ferramentas para desenvolvedores da AWS, como, AWS CodeBuild, AWS CodeCommit e AWS CodePipeline para automação do pipeline de lançamento que inclui metodologias de teste SAST e DAST.

À medida que você avançar no SDLC, estabeleça um processo iterativo que inclua avaliações de aplicação periódicas com sua equipe de segurança. O feedback coletado dessas avaliações de segurança deve ser abordado e validado como parte de sua revisão de prontidão do lançamento. Essas avaliações estabelecem um procedimento de segurança robusto de aplicações e fornecem aos criadores feedback útil para resolver possíveis problemas.

Etapas de implementação

- Implemente um IDE consistente, análise de código e ferramentas de CI/CD que incluam teste de segurança.
- Considere quando no SDLC é adequado bloquear pipelines em vez de apenas notificar os criadores de que problemas precisam ser corrigidos.
- O workshop Segurança para desenvolvedores fornece um exemplo de integração de testes estáticos e dinâmicos em um pipeline de lançamento.
- Realizar testes ou análises de código usando ferramentas automatizadas, como o Amazon CodeWhisperer integrado aos IDEs de desenvolvedores e o Amazon CodeGuru Reviewer para verificar o código na confirmação, ajuda os criadores a obter feedback no momento certo.
- Ao compilar usando o AWS Lambda, é possível usar o Amazon Inspector para verificar o código da aplicação em suas funções.
- Quando testes automatizados são incluídos em pipelines de CI/CD, é necessário usar um sistema de emissão de tíquetes para rastrear a notificação e a correção de problemas de software.

- Para testes de segurança que podem gerar descobertas, a vinculação com orientações para correção ajuda os criadores a melhorar a qualidade do código.
- Analise regularmente as descobertas das ferramentas automatizadas para priorizar a próxima automação, o treinamento de criadores ou a campanha de conscientização.

Recursos

Documentos relacionados:

- Entrega e implantação contínuas

- Parceiros de competência DevOps da AWS
- Parceiros de competência Segurança da AWS para segurança de aplicações
- Escolher uma abordagem de CI/CD do Well-Architected
- Monitorar eventos do CodeCommit no Amazon EventBridge e no Amazon CloudWatch Events
- Detecção de segredos no Amazon CodeGuru Reviewer
- Acelerar as implantações na AWS com uma governança efetiva
- Como a AWS aborda a automatização de implantações seguras e sem intervenção manual

Vídeos relacionados:

- Sem intervenção manual: como automatizar os pipelines de entrega contínua na Amazon
- Automatizar pipelines de CI/CD entre contas

Exemplos relacionados:

- Conscientização do setor para desenvolvedores
- Governança da AWS CodePipeline (GitHub)
- Workshop Segurança para desenvolvedores

SEC11-BP03 Realizar teste de penetração regular

Realize um teste de penetração regular do software. Esse mecanismo ajuda a identificar possíveis problemas de software que não podem ser detectados pelo teste automatizado ou por uma revisão

manual do código. Ele também ajuda você a entender a eficácia dos controles de detecção. O teste de penetração deve tentar determinar se o software pode ser executado de formas inesperadas, por exemplo, expondo dados que devem ser protegidos ou concedendo permissões mais amplas que o esperado.

Resultado desejado: o teste de penetração é usado para detectar, corrigir e validar as propriedades de segurança da sua aplicação. O teste de penetração regular e agendado deve ser realizado como parte do ciclo de vida de desenvolvimento de software (SDLC). As descobertas do teste de penetração devem ser abordadas antes do lançamento do software. As descobertas do teste de penetração devem ser analisaras para identificar se há problemas que podem ser encontrados usando a automação. Ter um processo de teste de penetração regular e repetível que inclua um mecanismo de feedback ativo ajuda a transmitir as orientações aos criadores e melhora a qualidade do software.

Práticas comuns que devem ser evitadas:

- Realizar um teste de penetração somente para problemas de segurança conhecidos ou prevalentes.
- Realizar um teste de penetração em aplicações sem ferramentas e bibliotecas de terceiros dependentes.
- Realizar um teste de penetração em aplicações em busca de problemas de segurança de pacote e não avaliar a lógica de negócios implementada.

Benefícios de implementar esta prática recomendada:

- Maior confiança nas propriedades de segurança do software antes do lançamento.
- Oportunidade de identificar padrões de aplicação preferenciais, o que aumenta a qualidade do software.
- Um ciclo de feedback que identifica mais cedo no ciclo de desenvolvimento quando a automação ou treinamento adicional pode melhorar as propriedades de segurança do software.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

O teste de penetração é um exercício de teste de segurança estruturado em que você executa cenários de violação de segurança planejados a fim de detectar, corrigir e validar controles de

segurança. Os testes de penetração começam com o reconhecimento, fase durante a qual os dados são coletados com base no design atual da aplicação e nas respectivas dependências. Uma lista selecionada de cenários de teste específicos de segurança é criada e executada. A principal finalidade desses testes é revelar problemas de segurança em sua aplicação, os quais podem ser explorados para obter acesso não intencional ao seu ambiente ou acesso não autorizado aos dados. É necessáraio realizar o teste de penetração ao lançar novos recursos ou sempre que sua aplicação passar por alterações importantes na implementação técnica ou de funções.

É necessário identificar o estágio mais apropriado do ciclo de vida de desenvolvimento para realizar o teste de penetração. Esse teste deve ocorrer em uma fase tardia o suficiente para que a funcionalidade do sistema esteja próxima ao estado de lançamento pretendido, mas com tempo suficiente para corrigir todos os problemas.

Etapas de implementação

- Tenha um processo estruturado de como o teste de penetração é definido. Basear esse processo no modelo de ameaças é uma boa maneira de manter o contexto.
- Identifique o estágio apropriado do ciclo de vida de desenvolvimento para realizar o teste de penetração, o qual deverá ocorrer quando houver o mínimo de alterações esperadas na aplicação e tempo suficiente para realizar a correção.
- Treine os criadores sobre o que esperar das descobertas do teste de penetração e como ter informações sobre correção.
- Utilize ferramentas para acelerar o processo de testes de penetração automatizando testes comuns ou repetíveis.
- Analise as descobertas do teste de penetração para identificar problemas de segurança sistêmicos e utilize esses dados para embasar testes automatizados adicionais e a instrução contínua dos criadores.

Recursos

Práticas recomendadas relacionadas:

- SEC11-BP01 Treinar para segurança de aplicações
- SEC11-BP02 Automatizar o teste durante o ciclo de vida de desenvolvimento e lançamento

Documentos relacionados:

- O Teste de penetração da AWS fornece orientação detalhada para testes de penetração na AWS
- Acelerar as implantações na AWS com uma governança efetiva
- Parceiros de competência Segurança da AWS
- Modernizar sua arquitetura de testes de penetração no AWS Fargate
- AWS Fault Injection Simulator

Exemplos relacionados:

- Automatizar os testes de API com o AWS CodePipeline (GitHub)
- Auxiliar de segurança automatizado (GitHub)

SEC11-BP04 Análises manuais de código

Realize uma análise manual do código do software que você produzir. Esse processo ajuda a verificar se a pessoa que escreveu o código não é a única que está conferindo sua qualidade.

Resultado desejado: incluir uma etapa manual de revisão de código durante o desenvolvimento aumenta a qualidade do software que está sendo escrito, ajuda a capacitar membros menos experientes da equipe e oferece a oportunidade de identificar locais onde a automação pode ser usada. É possível oferecer compatibilidade com as análises de código manuais com ferramentas e testes automatizados.

Práticas comuns que devem ser evitadas:

- Não realizar análises de código antes da implantação.
- Usar mesma pessoa para escrever e analisar o código.
- Não utilizar a automação para auxiliar ou orquestrar as análises de código.
- Não treinar os criadores em segurança de aplicações antes de analisarem o código.

Benefícios de implementar esta prática recomendada:

- Código de melhor qualidade.
- Maior consistência do desenvolvimento do código por meio da reutilização de abordagens comuns.
- Redução no número de problemas descobertos durante o teste de penetração e em estágios posteriores.

- Maior transferência de conhecimentos na equipe.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A etapa de análise deve ser implementada como parte do fluxo de gerenciamento de código geral. Os detalhes dependem da abordagem utilizada para ramificação, solicitações de pull e mesclagem. Você pode utilizar o AWS CodeCommit ou soluções de terceiros, como GitHub, GitLab ou Bitbucket. Seja qual for o método utilizado, é importante verificar se seus processos precisam de análise de código antes da implantação em um ambiente de produção. O uso de ferramentas como o Amazon CodeGuru Reviewer pode facilitar a orquestração do processo de revisão de código.

Etapas de implementação

- Implemente uma etapa de análise manual como parte do fluxo de gerenciamento de código e realize essa análise antes de prosseguir.
- Considere o Amazon CodeGuru Reviewer para gerenciar e auxiliar nas análises de código.
- Implemente um fluxo de aprovação que exija a realização de uma análise de código antes de avançá-lo para o próximo estágio.
- Verifique se há um processo para identificar problemas encontrados durante as análises de código manuais que possam ser detectados automaticamente.
- Integre a etapa de análise de código manual de forma que se alinhe às suas práticas de desenvolvimento de código.

Recursos

Práticas recomendadas relacionadas:

- SEC11-BP02 Automatizar o teste durante o ciclo de vida de desenvolvimento e lançamento

Documentos relacionados:

- Trabalhar com solicitações pull em repositórios do AWS CodeCommit
- Trabalhar com modelos de regra de aprovação no AWS CodeCommit
- Sobre solicitações pull no GitHub

- Automatizar as revisões de código com o Amazon CodeGuru Reviewer
- Automatizar a detecção de vulnerabilidades e bugs de segurança em pipelines de CI/CD usando a CLI do Amazon CodeGuru Reviewer

Vídeos relacionados:

- Melhoria contínua da qualidade do código com o Amazon CodeGuru

Exemplos relacionados:

- Workshop Segurança para desenvolvedores

SEC11-BP05 Centralizar serviços para pacotes e dependências

Forneça serviços centralizados a equipes de criadores para obter pacotes de software e outras dependências. Isso permite validar pacotes antes que eles sejam incluídos no software que você escreve e fornece uma fonte de dados para a análise do software que está sendo usado na sua organização.

Resultado desejado: o software é composto por um conjunto de outros pacotes de software, além do código que está sendo escrito. Isso simplifica o consumo de implementações de funcionalidades que são utilizadas repetidamente, como um analisador JSON ou uma biblioteca de criptografia. A centralização lógica das fontes desses pacotes e dependências oferece um mecanismo para as equipes de segurança validarem as propriedades dos pacotes antes de eles serem utilizados. Essa abordagem também reduz o risco de um problema inesperado ser provocado por uma alteração em um pacote existente ou pela inclusão de pacotes arbitrários diretamente da Internet pelas equipes de criadores. Utilize essa abordagem em conjunto com os fluxos de testes manuais e automatizados para aumentar a confiança na qualidade do software que está sendo desenvolvido.

Práticas comuns que devem ser evitadas:

- Extrair pacotes de repositórios arbitrários na Internet.
- Não testar novos pacotes antes de disponibilizá-los aos criadores.

Benefícios de implementar esta prática recomendada:

- Melhor entendimento de quais pacotes estão sendo utilizados no software que está sendo criado.

- Capacidade de notificar as equipes de workload quando um pacote precisa ser atualizado com base no entendimento de quem está usando o quê.
- Redução do risco de um pacote com problemas ser incluído em seu software.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Forneça serviços centralizados para pacotes e dependências de uma forma simples para os criadores consumirem. Serviços centralizados podem ser centralizados logicamente em vez de implementados como um sistema monolítico. Essa abordagem possibilita fornecer serviços de uma forma que atenda às necessidades dos criadores. Você deve implementar uma forma eficiente de adicionar pacotes ao repositório quando ocorrerem atualizações ou surgirem novos requisitos. Serviços da AWS como o AWS CodeArtifact ou soluções similares de parceiros da AWS oferecem uma maneira de fornecer esse recurso.

Etapas de implementação:

- Implemente um serviço de repositório centralizado logicamente disponível em todos os ambientes onde o software é desenvolvido.
- Inclua acesso ao repositório como parte do processo de provisionamento da Conta da AWS.
- Crie automação para testar pacotes antes de serem publicados em um repositório.
- Mantenha métricas dos pacotes mais utilizados, das linguagens e das equipes com a maior quantidade de alterações.
- Forneça um mecanismo automatizado para as equipes de criadores solicitarem novos pacotes e enviarem feedback.
- Verifique regularmente os pacotes em seu repositório para identificar o possível impacto de problemas recém-descobertos.

Recursos

Práticas recomendadas relacionadas:

- SEC11-BP02 Automatizar o teste durante o ciclo de vida de desenvolvimento e lançamento

Documentos relacionados:

- Acelerar as implantações na AWS com uma governança efetiva
- Aumentar a segurança do seu pacote com o kit de ferramentas CodeArtifact Package Origin Control
- Detectar problemas de segurança no log com o Amazon CodeGuru Reviewer
- Níveis da cadeia de suprimentos para artefatos de software (SLSA)

Vídeos relacionados:

- Segurança proativa: considerações e abordagens
- A filosofia de segurança da AWS (re:Invent 2017)
- Quando a segurança, a proteção e a urgência são importantes: como lidar com o Log4Shell

Exemplos relacionados:

- Pipeline de publicação de pacotes multirregionais (GitHub)
- Publicar módulos Node.js no AWS CodeArtifact usando o AWS CodePipeline (GitHub)
- Exemplo do Java CodeArtifact Pipeline no AWS CDK (GitHub)
- Distribuir pacotes .NET NuGet privados com o AWS CodeArtifact (GitHub)

SEC11-BP06 Implantar software programaticamente

Faça implantações de software de forma programática quando possível. Essa abordagem diminui a probabilidade de falha em uma implantação ou da introdução de um problema inesperado devido a erro humano.

Resultado desejado: manter as pessoas afastadas dos dados é um princípio fundamental para criar com segurança no. Nuvem AWS Esse princípio inclui como implantar seu software.

Os benefícios de não contar com pessoas para implantar software é a maior confiança de que o componente testado é o que será implantado e de que a implantação sempre é realizada de forma consistente. O software não deve precisar de alterações para funcionar em diferentes ambientes. O uso dos princípios de desenvolvimento de aplicações de 12 fatores, especificamente a externalização da configuração, possibilita implantar o mesmo código em vários ambientes sem a necessidade de alterações. Assinar de forma criptográfica os pacotes de software é uma boa maneira de garantir que nada tenha sido alterado entre os ambientes. O resultado geral dessa

abordagem é reduzir o risco em seu processo de alterações e melhorar a consistência das versões do software.

Práticas comuns que devem ser evitadas:

- Implantar software manualmente em produção.
- Realizar alterações manualmente no software para suprir diferentes ambientes.

Benefícios de implementar esta prática recomendada:

- Maior confiança no processo de lançamento de software.
- Redução do risco de uma alteração com falha afetar a funcionalidade dos negócios.
- Maior cadência de lançamentos devido ao menor risco de alterações.
- Recurso de reversão automática para eventos inesperados durante a implantação.
- Capacidade de comprovar de forma criptográfica que o software testado é o software implantado.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Crie a infraestrutura de sua Conta da AWS para remover o acesso humano persistente dos ambientes e use ferramentas de CI/CD para realizar implantações. Arquitete suas aplicações para que os dados de configuração específicos do ambiente sejam obtidos de uma fonte externa, como o AWS Systems Manager Parameter Store. Assine pacotes depois de testados e valide essas assinaturas durante a implantação. Configure seus pipelines de CI/CD para enviar código da aplicação e use canários para confirmar a implantação bem-sucedida. Use ferramentas como o AWS CloudFormation ou o AWS CDK para definir sua infraestrutura e, em seguida, use o AWS CodeBuild e o AWS CodePipeline para executar operações de CI/CD.

Etapas de implementação

- Crie pipelines de CI/CD bem definidos para simplificar o processo de implantação.
- Usar o AWS CodeBuild e um AWS Code Pipeline para fornecer capacidade de CI/CD simplifica a integração de testes de segurança em seus pipelines.
- Siga as orientações sobre separação de ambientes no whitepaper Como organizar seu ambiente da AWS usando várias contas.

- Verifique se não há nenhum acesso humano persistente aos ambientes nos quais as workloads de produção estão em execução.
- Projete as aplicações para oferecer compatibilidade com a externalização de dados de configuração.
- Considere implantar com o uso do modelo de implantação azul/verde.
- Implemente canários para validar a implantação bem-sucedida do software.
- Use ferramentas criptográficas como o AWS Signer ou i AWS Key Management Service (AWS KMS) para assinar e verificar os pacotes de software que estão sendo implantandos.

Recursos

Práticas recomendadas relacionadas:

- SEC11-BP02 Automatizar o teste durante o ciclo de vida de desenvolvimento e lançamento

Documentos relacionados:

- Workshop de CI/CD da AWS
- Acelerar as implantações na AWS com uma governança efetiva
- Automatizar implantações seguras e sem intervenção manual
- Assinatura de código usando CA privada do AWS Certificate Manager e chaves assimétricas do AWS Key Management Service
- Assinatura de código, um controle de confiança e integridade para AWS Lambda

Vídeos relacionados:

- Sem intervenção manual: como automatizar os pipelines de entrega contínua na Amazon

Exemplos relacionados:

- Implantações azuis/verdes com o AWS Fargate

SEC11-BP07 Avaliar regularmente as propriedades de segurança dos pipelines

Aplique os princípios do pilar Segurança do Well-Architected aos seus pipelines, com atenção especial à separação das permissões. Avalie as propriedades de segurança de sua infraestrutura de pipelines. O gerenciamento eficaz da segurança dos pipelines permite fornecer segurança ao software que passa pelos pipelines.

Resultado desejado: os pipelines usados para criar e implantar seu software devem seguir as mesmas práticas recomendadas de qualquer outra workload em seu ambiente. Os testes implementados nos pipelines não devem ser editáveis pelos criadores que os estão utilizando. Os pipelines só devem ter as permissões necessárias para as implantações que eles estão realizando e devem implementar proteções para evitar a implantação em ambientes errados. Os pipelines não devem contar com credenciais de longo prazo e devem ser configurados para emitir o estado de forma que a integridade dos ambientes de compilação possa ser validada.

Práticas comuns que devem ser evitadas:

- Testes de segurança que podem ser ignorados pelos criadores.
- Permissões excessivamente amplas para pipelines de implantação.
- Pipelines não configurados para validar entradas.
- Ausência de análise regular das permissões associadas à infraestrutura de CI/CD.
- Uso de credenciais de longo prazo ou codificadas.

Benefícios de implementar esta prática recomendada:

- Maior confiança na integridade do software que está sendo criado e implantado pelos pipelines.
- Capacidade de interromper uma implantação quando há atividade suspeita.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Começar com serviços de CI/CD gerenciados que ofereçam compatibilidade com perfis do IAM reduz o risco de vazamento de credenciais. Aplicar os princípios do pilar Segurança à sua infraestrutura de pipeline de CI/CD pode ajudar a determinar onde é possível realizar melhorias de segurança. Seguir a arquitetura de referência do AWS Deployment Pipelines é um bom ponto de partida para criar seus ambientes de CI/CD. Analisar regularmente a implementação de pipelines e analisar

comportamentos inesperados nos logs pode ajudar você a entender os padrões de uso dos pipelines que estão sendo utilizados para implantar o software.

Etapas de implementação

- Comece com a arquitetura de referência dos pipelines de implantação da AWS.
- Considere usar o AWS IAM Access Analyzer para gerar programaticamente políticas do IAM com privilégios mínimos para os pipelines.
- Integre seus pipelines com monitoramento e alertas para receber notificações sobre atividades inesperadas ou anormais. Para serviços gerenciados da AWS, o Amazon EventBridge permite direcionar dados para destinos como o AWS Lambda ou o Amazon Simple Notification Service (Amazon SNS).

Recursos

Documentos relacionados:

- Arquitetura de referência dos pipelines de implantação da AWS
- Monitorar o AWS CodePipeline
- Práticas recomendadas de segurança para o AWS CodePipeline

Exemplos relacionados:

- Painel de monitoramento do DevOps (GitHub)

SEC11-BP08 Criar um programa que incorpore a propriedade de segurança nas equipes de workload

Crie um programa ou mecanismo que capacite as equipes de criadores a tomar decisões de segurança sobre o software que elas estão criando. Ainda é necessário que sua equipe de segurança valide essas decisões durante uma revisão, mas a incorporação da propriedade de segurança nas equipes de criadores aumenta a velocidade e segurança do processo de criação de workloads. Esse mecanismo também promove uma cultura de propriedade que afeta de forma positiva a operação dos sistemas que você cria.

Resultado desejado: para incorporar a propriedade de segurança e a tomada de decisões nas equipes de construtores, é possível treinar os construtores em como pensar em segurança ou

aumentar o treinamento deles com pessoas de segurança incorporadas ou associadas às equipes de construtores. As duas abordagens são válidas e possibilitam à equipe tomar decisões de segurança de melhor qualidade logo no início do ciclo de desenvolvimento. Esse modelo de propriedade é baseado em treinamento para segurança de aplicações. Iniciar com o modelo de ameaças para a workload específica ajuda a direcionar o design thinking (pensamento de design) para o contexto apropriado. Outro benefício de ter uma comunidade de criadores concentrados em segurança ou um grupo de engenheiros de segurança que trabalhem com equipes de criadores é que você pode entender mais profundamente como o software é escrito. Esse entendimento ajuda você a determinar as próximas áreas de melhoria em seu recurso de automação.

Práticas comuns que devem ser evitadas:

- Deixar todas as decisões de design de segurança para a equipe de segurança.
- Não abordar os requisitos de segurança cedo o suficiente no processo de desenvolvimento.
- Não obter feedback dos criadores e do pessoal de segurança sobre a operação do programa.

Benefícios de implementar esta prática recomendada:

- Redução do tempo para concluir as avaliações de segurança.
- Redução dos problemas de segurança que são detectados apenas no estágio de avaliação da segurança.
- Melhoria da qualidade geral do software que está sendo escrito.
- Oportunidade de identificar e entender problemas sistêmicos ou áreas de melhoria de alto valor.
- Redução da quantidade de revisão necessária devido às descobertas da avaliação da segurança.
- Melhoria da percepção da função de segurança.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Comece com a orientação em SEC11-BP01 Treinar para segurança de aplicações. Depois, identifique o modelo operacional para o programa que você acredita ser o melhor para a sua organização. Os dois padrões principais são treinar os criadores ou incorporar o pessoal de segurança às equipes de criadores. Depois de decidir sobre a abordagem inicial, é necessário criar um piloto com uma equipe de workload ou um grupo pequeno de equipes de workload para comprovar que o modelo funciona para sua organização. O apoio de liderança dos criadores e da

segurança da organização contribui para a entrega e o sucesso do programa. À medida que você criar esse programa, é importante selecionar as métricas que podem ser utilizadas para mostrar seu valor. Saber como a AWS resolveu esse problema é uma boa experiência de aprendizado. A prática recomendada é muito concentrada na mudança e cultura organizacionais. As ferramentas que você utiliza devem ser compatíveis com a colaboração entre as comunidades de criadores e de segurança.

Etapas de implementação

- Comece com o treinamento dos criadores para segurança de aplicações.
- Crie uma comunidade e um programa de integração para instruir os criadores.
- Selecione um nome para o programa. Guardiões, patrocinadores ou defensores são utilizados com frequência.
- Identifique o modelo a ser utilizado: treinar criadores, incorporar engenheiros de segurança e ter perfis de segurança de afinidade.
- Identifique patrocinadores do projeto em grupos de segurança e de criadores e possivelmente em outros grupos relevantes.
- Rastreie as métricas do número de pessoas envolvidas no programa, o tempo gasto em avaliações e o feedback dos criadores e do pessoal de segurança. Utilize essas métricas para realizar melhorias.

Recursos

Práticas recomendadas relacionadas:

- SEC11-BP01 Treinar para segurança de aplicações
- SEC11-BP02 Automatizar o teste durante o ciclo de vida de desenvolvimento e lançamento

Documentos relacionados:

- Como abordar a modelagem de ameaças
- Como pensar sobre a governança da segurança na nuvem

Vídeos relacionados:

- Segurança proativa: considerações e abordagens

Confiabilidade

O pilar Confiabilidade abrange a capacidade de uma workload de executar a função pretendida correta e consistentemente quando esperado. Recomendações sobre implementação estão disponíveis no whitepaper Pilar Confiabilidade.

Áreas de práticas recomendadas

- Fundamentos
- Arquitetura da workload
- Gerenciamento de alterações
- Gerenciamento de falhas

Fundamentos

Perguntas

- REL 1. Como você gerencia cotas de serviço e restrições?
- REL 2. Como você planeja sua topologia de rede?

REL 1. Como você gerencia cotas de serviço e restrições?

Para arquiteturas de workload baseadas na nuvem, existem cotas de serviço (que também são chamadas de limites de serviço). Essas cotas existem para evitar o provisionamento acidental de mais recursos do que você precisa e para limitar as taxas de solicitação nas operações de API, a fim de proteger os serviços contra uso abusivo. Também há restrições de recursos, por exemplo, a taxa em que você pode propagar bits por um cabo de fibra óptica ou a quantidade de armazenamento em um disco físico.

Práticas recomendadas

- REL01-BP01 Conhecer as cotas e restrições de serviços
- REL01-BP02 Gerenciar cotas de serviço em várias contas e regiões
- REL01-BP03 Acomodar restrições e cotas de serviço fixo por meio de arquitetura
- REL01-BP04 Monitorar e gerenciar cotas
- REL01-BP05 Automatizar o gerenciamento de cotas

- REL01-BP06 Garantir que existe uma lacuna suficiente entre as cotas atuais e o uso máximo para acomodar o failover

REL01-BP01 Conhecer as cotas e restrições de serviços

Esteja ciente das suas cotas padrão e das solicitações de aumento de cota referentes à sua arquitetura da workload. Saiba quais restrições de recursos, como disco ou rede, podem gerar impactos.

Resultado desejado: os clientes podem evitar a degradação ou interrupção do serviço em suas Contas da AWS implementando diretrizes adequadas para monitorar as principais métricas, análises de infraestrutura e etapas de correção de automação para verificar se as cotas e restrições de serviços não foram atingidas, o que poderia causar degradação ou interrupção do serviço.

Práticas comuns que devem ser evitadas:

- Implantar uma workload sem compreender as cotas flexíveis ou fixas e seus limites para os serviços utilizados.
- Implantar uma workload de substituição sem analisar e reconfigurar as cotas necessárias ou entrar em contato com o suporte com antecedência.
- Pressupor que os serviços em nuvem não têm limites e os serviços podem ser usados sem considerar taxas, limites, contagens e quantidades.
- Pressupor que as cotas aumentarão automaticamente.
- Não saber o processo e a linha de tempo das solicitações de cota.
- Pressupor que a cota de serviço em nuvem padrão é idêntica para todos os serviços em comparação entre as regiões.
- Pressupor que as restrições do serviço podem ser violadas e os sistemas vão ser escalados automaticamente ou aumentar o limite além das restrições do recurso
- Não testar a aplicação em tráfego de pico a fim de aplicar tensão na utilização de seus recursos.
- Provisionar o recurso sem analisar o respectivo tamanho necessário.
- Provisionar capacidade em excesso selecionando tipos de recurso que superam em muito a necessidade real ou os picos esperados.
- Não avaliar os requisitos de capacidade para novos níveis de tráfego antes de um novo evento de cliente ou implantação de uma nova tecnologia.

Benefícios de implementar esta prática recomendada: o monitoramento e o gerenciamento automatizado de cotas de serviço e restrições de recursos podem reduzir proativamente as falhas. As alterações nos padrões de tráfego do serviço de um cliente poderão causar interrupção ou degradação se as práticas recomendadas não forem seguidas. Ao monitorar e gerenciar esses valores em todas as regiões e contas, as aplicações podem ter uma resiliência aprimorada em eventos adversos ou não planejados.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

O Service Quotas é um serviço da AWS que ajuda a gerenciar em um único local suas cotas para mais de 250 serviços da AWS. Além de pesquisar os valores de cotas, você também pode solicitar e rastrear aumentos de cota no console do Service Quotas ou por meio do AWS SDK. O AWS Trusted Advisor oferece uma verificação de cotas de serviço que exibe o uso e as cotas para certos aspectos de alguns serviços. As cotas de serviço padrão por serviço também estão na documentação da AWS por respectivo serviço (por exemplo, consulte Cotas da Amazon VPC).

Alguns limites de serviço, como os limites de taxa para APIs com controle de utilização, são definidos no próprio Amazon API Gateway por meio da configuração de um plano de uso. Alguns limites definidos como configuração em seus respectivos serviços incluem IOPS provisionadas, armazenamento do Amazon RDS alocado e alocações de volume do Amazon EBS. O Amazon Elastic Compute Cloud tem seu próprio painel de limites de serviço que pode ajudar você a gerenciar sua instância, o Amazon Elastic Block Store e limites de endereços IP elásticos. Se você tiver um caso de uso em que as cotas de serviço afetam a performance de sua aplicação e elas não forem ajustadas às suas necessidades, entre em contato com o AWS Support para ver se há mitigações.

As cotas de serviço podem ser específicas da região e também pode ser globais por natureza. O uso de um serviço da AWS com a cota atingida fará com que o comportamento dele não seja o esperado e poderá causar interrupção ou degradação do serviço. Por exemplo, uma cota de serviço limita o número de instâncias do Amazon EC2 DL usadas em uma região. Esse limite pode ser alcançado durante um evento de escalação de tráfego usando grupos do Auto Scaling (ASG).

As cotas de serviço de cada conta devem ser avaliadas regularmente quanto ao uso a fim de determinar quais são os limites de serviço apropriados para a conta em questão. Essas cotas de serviço existem como barreiras de proteção operacionais a fim de impedir o provisionamento acidental de recursos além do necessário. Elas também servem para limitar as taxas de solicitação em operações de API para proteger os serviços contra abuso.

Restrições de serviço são diferentes de cotas de serviço. As restrições de serviço representam os limites de um recurso específico conforme definido pelo tipo de recurso em questão. Elas podem ser a capacidade de armazenamento (por exemplo, gp2 tem um limite de tamanho de 1 GB a 16 TB) ou o throughput de disco. É essencial que a restrição de um tipo de recurso seja projetada e avaliada constantemente quanto a tipos de uso que podem atingir o limite. Se uma restrição for atingida de modo inesperado, as aplicações da conta ou os serviços poderão sofrer degradação ou interrupção.

Se houver um caso de uso em que as cotas de serviço afetem a performance de uma aplicação e elas não puderem ser ajustadas às necessidades, entre em contato com o AWS Support para ver se há mitigações. Para obter mais detalhes sobre o ajuste de cotas fixas, consulte REL01-BP03 Acomodar restrições e cotas de serviço fixo por meio de arquitetura.

Há uma série de serviços e ferramentas da AWS para ajudar a monitorar e gerenciar o Service Quotas. O serviço e as ferramentas devem ser utilizados para oferecer verificações automatizadas ou manuais dos níveis de cota.

- O AWS Trusted Advisor oferece uma verificação de cotas de serviço que exibe o uso e cotas para alguns aspectos de alguns serviços. Ele pode ajudar na identificação de serviços que estão próximos da cota.
- O AWS Management Console oferece métodos para exibir valores de cota de serviço, gerenciar, solicitar novas cotas, monitorar o status das solicitações de cota e exibir o histórico de cotas.
- A AWS CLI e os CDKs oferecem métodos programáticos para gerenciar e monitorar automaticamente os níveis e o uso de cotas de serviço.

Etapas de implementação

Para o Service Quotas:

- Revise o AWS Service Quotas.
- Para saber suas cotas de serviço existentes, determine os serviços (como o IAM Access Analyzer) utilizados. Há cerca de 250 serviços da AWS controlados pelo Service Quotas. Depois, determine o nome específico da cota de serviço que pode estar sendo usada em cada conta e região. Há cerca de 3 mil nomes de cota de serviço por região.
- Aumente essa análise de cotas com o AWS Config para encontrar todos os recursos da AWS usados em suas Contas da AWS.
- Use dados do AWS CloudFormation para determinar seus recursos da AWS usados. Veja os recursos que foram criados no AWS Management Console ou com o comando `list-stack-`

`resources` da AWS CLI. Também é possível ver no próprio modelo os recursos configurados para implantação.

- Examine o código da implantação para determinar todos os serviços necessários à sua workload.
- Determine as cotas de serviço aplicáveis. Use as informações acessíveis programaticamente por meio do Trusted Advisor e do Service Quotas.
- Estabeleça um método de monitoramento automatizado (consulte REL01-BP02 Gerenciar cotas de serviço em várias contas e regiões e REL01-BP04 Monitorar e gerenciar cotas) para alertar e informar se as cotas de serviços estão próximas ou atingiram seu limite.
- Estabeleça um método automatizado e programático para verificar se uma cota de serviço foi alterada em uma região, mas não em outras regiões da mesma conta (consulte REL01-BP02 Gerenciar cotas de serviço em várias contas e regiões e REL01-BP04 Monitorar e gerenciar cotas).
- Automatize as verificações de logs e métricas de aplicações para determinar se há erros de restrição de serviço ou cota. Se houver esses erros, envie alertas ao sistema de monitoramento.
- Estabeleça procedimentos de engenharia para calcular a mudança necessária na cota (consulte REL01-BP05 Automatizar o gerenciamento de cotas) depois de identificar que cotas maiores são necessárias para serviços específicos.
- Crie um fluxo de trabalho de provisionamento e aprovação para solicitar alterações na cota de serviço. Isso deve incluir um fluxo de trabalho de exceção em caso de negação de solicitação ou aprovação parcial.
- Crie um método de engenharia para analisar cotas de serviço antes de provisionar e usar novos serviços da AWS antes de distribuir na produção ou carregar ambientes (por exemplo, conta de teste de carga).

Para restrições de serviço:

- Estabeleça métodos de monitoramento e métricas para alertar sobre recursos que estejam próximos de suas restrições de recurso. Utilize o CloudWatch conforme apropriado para métricas ou monitoramento de logs.
- Estabeleça limites de alerta para cada recurso que tenha uma restrição significativa para a aplicação ou o sistema.
- Crie um fluxo de trabalho e procedimentos de gerenciamento de infraestrutura para alterar o tipo de recurso se a restrição estiver próxima da utilização. Esse fluxo de trabalho deve incluir testes de

carga como prática recomendada para verificar se o novo tipo de recurso é o correto com as novas restrições.

- Migre o recurso identificado para o novo tipo de recurso usando procedimentos e processos existentes.

Recursos

Práticas recomendadas relacionadas:

- REL01-BP02 Gerenciar cotas de serviço em várias contas e regiões
- REL01-BP03 Acomodar restrições e cotas de serviço fixo por meio de arquitetura
- REL01-BP04 Monitorar e gerenciar cotas
- REL01-BP05 Automatizar o gerenciamento de cotas
- REL01-BP06 Garantir que existe uma lacuna suficiente entre as cotas atuais e o uso máximo para acomodar o failover
- REL03-BP01 Escolher como segmentar a workload
- REL10-BP01 Implantar a workload em vários locais
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas
- REL11-BP03 Automatizar a reparação em todas as camadas
- REL12-BP05 Testar a resiliência via engenharia do caos

Documentos relacionados:

- Pilar Confiabilidade do AWS Well-Architected Framework: Disponibilidade
- AWS Service Quotas (antigamente conhecido como Limites de serviço)
- Verificações de práticas recomendadas do AWS Trusted Advisor (consulte a seção Limites de serviço)
- Monitor de limites da AWS em respostas da AWS
- Limites de serviço do Amazon EC2
- O que é o Service Quotas?
- Como solicitar um aumento da cota
- Endpoints e cotas de serviço

- Guia do usuário do Service Quotas
- Monitor de cotas para AWS
- Limites de isolamento de falhas da AWS
- Disponibilidade com redundância
- AWS para dados
- O que é integração contínua?
- O que é entrega contínua?
- Parceiro da APN: parceiros que podem ajudar no gerenciamento de configuração
- Gerenciar o ciclo de vida da conta em ambientes SaaS de conta por locatário na AWS
- Gerenciar e monitorar o controle de utilização de APIs em suas workloads
- Visualizar recomendações do AWS Trusted Advisor em grande escala com o AWS Organizations
- Automatizar aumentos de limites de serviço e suporte corporativo com o AWS Control Tower

Vídeos relacionados:

- AWS Live re:Inforce 2019: Service Quotas
- Visualizar e gerenciar cotas para serviços da AWS com o Service Quotas
- Demonstração de cotas do AWS IAM

Ferramentas relacionadas:

- Amazon CodeGuru Reviewer
- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK

- AWS Systems Manager
- AWS Marketplace

REL01-BP02 Gerenciar cotas de serviço em várias contas e regiões

Se você estiver usando várias contas ou regiões, solicite as cotas adequadas em todos os ambientes nos quais suas workloads de produção são executadas.

Resultado desejado: serviços e aplicações não devem ser afetados pelo esgotamento da cota de serviço para configurações que abrangem contas ou regiões ou que tenham projetos de resiliência usando failover de zona, região ou conta.

Práticas comuns que devem ser evitadas:

- Permitir que a utilização de recursos em uma região de isolamento aumente sem nenhum mecanismo para manter a capacidade das demais.
- Configurar manualmente todas as cotas nas regiões de isolamento de forma independente.
- Não considerar o efeito das arquiteturas de resiliência (como ativa ou passiva) em necessidades futuras de cota durante a degradação na região que não é a principal.
- Não avaliar as cotas regularmente e fazer alterações necessárias em cada região e conta nas quais a workload é executada.
- Não reutilizar modelos de solicitação de cotas para solicitar aumentos em várias regiões e contas.
- Não atualizar as cotas de serviço por imaginar incorretamente que aumentar as cotas tem implicações de custo, como solicitações de reserva computacional.

Benefícios de implementar esta prática recomendada: verificar se você pode lidar com sua carga atual em regiões ou contas secundárias se os serviços regionais ficarem indisponíveis. Isso pode ajudar a reduzir o número de erros ou níveis de degradações que ocorrem durante a perda da região.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Cotas de serviço são rastreadas por conta. A menos que especificado de outra forma, cada cota é específica da Região da AWS. Além dos ambientes de produção, gerencie também as cotas em todos os ambientes aplicáveis que não são de produção para que os testes e o desenvolvimento não

sejam dificultados. Manter um alto grau de resiliência exige que as cotas de serviço sejam avaliadas de forma contínua (sejam elas automatizadas ou manuais).

Com cada vez mais workloads abrangendo regiões devido à implementação de projetos usando as abordagens Ativo/Ativo, Ativo/Passivo – Quente, Ativo/Passivo – Frio e Ativo/Passivo – Luz piloto, é essencial entender todos os níveis de cota da região e da conta. Padrões de tráfego passados nem sempre são um bom indicador de que a cota de serviço está definida corretamente.

Igualmente importante, o limite do nome da cota de serviço nem sempre é o mesmo para cada região. Em uma região, o valor pode ser cinco e em outra região pode ser dez. O gerenciamento dessas cotas deve abranger todos os mesmos serviços, contas e regiões para fornecer resiliência consistente sob carga.

Reconcilie todas as diferenças de cota de serviço em todas as diferentes regiões (Região ativa ou Região passiva) e crie processos para reconciliar essas diferenças de forma contínua. Os planos de teste de failovers de região passiva raramente são escalados para a capacidade ativa de pico, o que significa que os exercícios de simulações teóricas e dias de teste podem não encontrar diferenças em cotas de serviço entre regiões e também depois manter os limites corretos.

Monitorar e avaliar o desvio da cota de serviço, a condição em que os limites da cota de serviço para uma cota específica são alterados em uma região e não em todas as regiões, é muito importante. É necessário pensar em alterar a cota em regiões com tráfego ou que possam ter tráfego.

- Selecione as contas e as regiões relevantes conforme seus requisitos de serviço, de latência, regulatórios e de recuperação de desastres.
- Identifique as cotas de serviço de todas as contas, regiões e zonas de disponibilidade relevantes. O escopo dos limites é definido para conta e região. Esses valores devem ser comparados em relação a diferenças.

Etapas de implementação

- Analise os valores do Service Quotas que podem ter ultrapassado um nível de risco de uso. O AWS Trusted Advisor oferece alertas para violações de limite de 80% e 90%.
- Analise os valores de cotas de serviço em todas as regiões passivas (em um design Ativo/Passivo). Verifique se a carga será executada com êxito em regiões secundárias em caso de falha na região principal.
- Automatize a avaliação caso qualquer desvio de cota de serviço tenha ocorrido entre as regiões na mesma conta e aja adequadamente para alterar os limites.

- Se as unidades organizações (UO) do cliente estiverem estruturadas da forma compatível, os modelos de cota de serviço deverão ser atualizados para refletir alterações em todas as cotas que devem ser aplicadas a várias regiões e contas.
  - Crie um modelo e associe regiões à alteração de cota.
  - Analise todos os modelos de cota de serviço existentes para todas as alterações necessárias (região, limites e contas).

Recursos

Práticas recomendadas relacionadas:

- REL01-BP01 Conhecer as cotas e restrições de serviços
- REL01-BP03 Acomodar restrições e cotas de serviço fixo por meio de arquitetura
- REL01-BP04 Monitorar e gerenciar cotas
- REL01-BP05 Automatizar o gerenciamento de cotas
- REL01-BP06 Garantir que existe uma lacuna suficiente entre as cotas atuais e o uso máximo para acomodar o failover
- REL03-BP01 Escolher como segmentar a workload
- REL10-BP01 Implantar a workload em vários locais
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas
- REL11-BP03 Automatizar a reparação em todas as camadas
- REL12-BP05 Testar a resiliência via engenharia do caos

Documentos relacionados:

- Pilar Confiabilidade do AWS Well-Architected Framework: Disponibilidade
- AWS Service Quotas (antigamente conhecido como Limites de serviço)
- Verificações de práticas recomendadas do AWS Trusted Advisor (consulte a seção Limites de serviço)
- Monitor de limites da AWS em respostas da AWS
- Limites de serviço do Amazon EC2
- O que é o Service Quotas?
- Como solicitar um aumento da cota

- Endpoints e cotas de serviço
- Guia do usuário do Service Quotas
- Monitor de cotas para AWS
- Limites de isolamento de falhas da AWS
- Disponibilidade com redundância
- AWS para dados
- O que é integração contínua?
- O que é entrega contínua?
- Parceiro da APN: parceiros que podem ajudar no gerenciamento de configuração
- Gerenciar o ciclo de vida da conta em ambientes SaaS de conta por locatário na AWS
- Gerenciar e monitorar o controle de utilização de APIs em suas workloads
- Visualizar recomendações do AWS Trusted Advisor em grande escala com o AWS Organizations
- Automatizar aumentos de limites de serviço e suporte corporativo com o AWS Control Tower

Vídeos relacionados:

- AWS Live re:Inforce 2019: Service Quotas
- Visualizar e gerenciar cotas para serviços da AWS com o Service Quotas
- Demonstração de cotas do AWS IAM

Serviços relacionados:

- Amazon CodeGuru Reviewer
- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK

- AWS Systems Manager
- AWS Marketplace

REL01-BP03 Acomodar restrições e cotas de serviço fixo por meio de arquitetura

Esteja ciente das cotas de serviço, das restrições de serviço e dos limites de recursos físicos que não podem ser alterados. Projete arquiteturas para aplicações e serviços visando evitar que esses limites afetem a confiabilidade.

Os exemplos incluem largura de banda da rede, tamanho da carga útil da invocação da função sem servidor, taxa de intermitência de controle de utilização para um gateway da API e conexões simultâneas de usuários com um banco de dados.

Resultado desejado: a aplicação ou o serviço funciona conforme o esperado em condições normais e de alto tráfego. Eles foram desenvolvidos para funcionar com as limitações referentes às restrições ficas ou cotas de serviço do recurso.

Práticas comuns que devem ser evitadas:

- Escolher um design que usa um recurso de um serviço sem saber que há restrições de design que causarão falha à medida que você escala.
- Usar parâmetros de comparação fora da realidade e que atingirão as cotas fixas do serviço durante os testes. Por exemplo, executar testes em um limite de intermitência mas por um período estendido.
- Escolher um design que não possa ser escalado nem modificado caso seja necessário ultrapassar as cotas fixas do serviço. Por exemplo, um tamanho de carga útil do SQS de 256 KB.
- A observabilidade não foi projetada nem implementada para monitorar e alertar sobre os limites das cotas de serviço que podem estar em risco durante eventos com tráfego alto.

Benefícios de implementar esta prática recomendada: verificar se a aplicação será executada em todos os níveis de carga de serviços projetados sem interrupção ou degradação.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Ao contrário das cotas de serviço flexíveis ou de recursos que são substituídos com unidades de capacidade mais altas, as cotas fixas dos serviços da AWS não podem ser alteradas. Isso significa

que todos esses tipos de serviços da AWS devem ser avaliados com relação a possíveis limites de capacidade rígidos quando usados no design de uma aplicação.

Os limites rígidos são mostrados no console do Service Quotas. Se as colunas mostrarem `ADJUSTABLE = No`, o serviço tem um limite rígido. Os limites rígidos também são mostrados em algumas páginas de configuração de recursos. Por exemplo, o Lambda tem limites rígidos específicos que não podem ser ajustados.

Como exemplo, ao projetar uma aplicação Python para ser executada em uma função do Lambda, a aplicação deve ser avaliada para determinar se há alguma chance de o Lambda ser executado por mais de 15 minutos. Se código puder ser executado mais do que esse limite de cota de serviço, tecnologias ou designs alternativos devem ser considerados. Se esse limite for atingido depois da implantação na produção, a aplicação sofrerá uma degradação e interrupção até que isso possa ser corrigido. Ao contrário das cotas flexíveis, não há um método para alterar esses limites mesmo sob eventos de emergência de severidade 1.

Depois que a aplicação for implantada em um ambiente de teste, estratégias devem ser usadas para descobrir se algum limite rígido pode ser atingido. Testes de estresse, testes de carga e testes de caos devem fazer parte do plano de teste de introdução.

Etapas de implementação

- Revise a lista completa de serviços da AWS que poderiam ser usados na fase de design da aplicação.
- Revise os limites da cota flexível e os da cota rígida para todos esses serviços. Nem todos os limites são mostrados no console do Service Quotas. Alguns serviços descrevem esses limites em locais alternativos.
- À medida que você planeja a aplicação, revise os fatores que impulsionam a tecnologia e os negócios da workload, como resultados empresariais, casos de uso, sistemas dependentes, destinos de disponibilidade e objetos de recuperação de desastres. Permita que os fatores que impulsionam a tecnologia e os negócios orientem o processo para identificar o sistema distribuído certo para sua workload.
- Analise a carga do serviço nas regiões e contas. Muitos limites rígidos são regionais para os serviços. No entanto, alguns limites são baseados em conta.
- Analise arquiteturas de resiliência quanto ao uso de recursos durante uma falha de zona e de região. Na progressão de designs de várias regiões usando as abordagens ativo/ativo, ativo/passivo – quente, ativo/passivo – frio e ativo/passivo – luz-piloto, esses casos de falha resultarão em maior uso. Isso cria um possível caso de uso para atingir limites rígidos.

Recursos

Práticas recomendadas relacionadas:

- REL01-BP01 Conhecer as cotas e restrições de serviços
- REL01-BP02 Gerenciar cotas de serviço em várias contas e regiões
- REL01-BP04 Monitorar e gerenciar cotas
- REL01-BP05 Automatizar o gerenciamento de cotas
- REL01-BP06 Garantir que existe uma lacuna suficiente entre as cotas atuais e o uso máximo para acomodar o failover
- REL03-BP01 Escolher como segmentar a workload
- REL10-BP01 Implantar a workload em vários locais
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas
- REL11-BP03 Automatizar a reparação em todas as camadas
- REL12-BP05 Testar a resiliência via engenharia do caos

Documentos relacionados:

- Pilar Confiabilidade do AWS Well-Architected Framework: Disponibilidade
- AWS Service Quotas (antigamente conhecido como Limites de serviço)
- Verificações de práticas recomendadas do AWS Trusted Advisor (consulte a seção Limites de serviço)
- Monitor de limites da AWS em respostas da AWS
- Limites de serviço do Amazon EC2
- O que é o Service Quotas?
- Como solicitar um aumento da cota
- Endpoints e cotas de serviço
- Guia do usuário do Service Quotas
- Monitor de cotas para AWS
- Limites de isolamento de falhas da AWS
- Disponibilidade com redundância
- AWS para dados
- O que é integração contínua?

- O que é entrega contínua?
- Parceiro da APN: parceiros que podem ajudar no gerenciamento de configuração
- Gerenciar o ciclo de vida da conta em ambientes SaaS de conta por locatário na AWS
- Gerenciar e monitorar o controle de utilização de APIs em suas workloads
- Visualizar recomendações do AWS Trusted Advisor em grande escala com o AWS Organizations
- Automatizar aumentos de limites de serviço e suporte corporativo com o AWS Control Tower
- Ações, recursos e chaves de condição para o Service Quotas

Vídeos relacionados:

- AWS Live re:Inforce 2019: Service Quotas
- Visualizar e gerenciar cotas para serviços da AWS com o Service Quotas
- Demonstração de cotas do AWS IAM
- AWS re:Invent 2018: Fechar loops e abrir mentes: como assumir o controle de sistemas grandes e pequenos

Ferramentas relacionadas:

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP04 Monitorar e gerenciar cotas

Avalie seu uso potencial e aumente suas cotas adequadamente para possibilitar o crescimento planejado do uso.

Resultado desejado: sistemas ativos e automatizados que gerenciam e monitoram foram implantados. Essas soluções de operações garantem que os limites de uso da cota estejam próximos de ser atingidos. Isso seria corrigido proativamente por mudanças solicitadas na cota.

Práticas comuns que devem ser evitadas:

- Não configurar o monitoramento para verificar os limites da cota de serviço.
- Não configurar o monitoramento de limites rígidos, embora esses valores não possam ser alterados.
- Presumir que o tempo necessário para solicitar e proteger uma mudança de cota flexível seja imediato ou período curto.
- Configurar alarmes para quando as cotas de serviço estiverem sendo atingidas, mas não ter um processo de resposta a um alerta.
- Configurar apenas alarmes para serviços compatíveis com o AWS Service Quotas e não monitorar outros serviços da AWS.
- Não considerar o gerenciamento da cota para designs com resiliência de várias regiões, como as abordagens Ativo/Ativo, Ativo/Passivo – Quente, Ativo/Passivo – Frio e Ativo/Passivo – Luz piloto.
- Não avaliar as diferenças de cota entre regiões.
- Não avaliar as necessidades de cada região com relação a uma solicitação de aumento de cota específica.
- Não utilizar modelos para gerenciamento de cotas em várias regiões.

Benefícios de implementar esta prática recomendada: o rastreamento automático das cotas de serviço da AWS e o monitoramento do seu uso em relação a essas cotas permitirão que você veja quando está perto de atingir uma cota. Também é possível usar esse monitoramento de dados para ajudar a limitar qualquer dano resultante da exaustão da cota.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Para dispositivos compatíveis, é possível monitorar as cotas configurando vários serviços diferentes que podem avaliar e, então, enviar alertas ou alarmes. Isso pode auxiliar o monitoramento do uso e alertar quando você estiver se aproximando das cotas. Esses alarmes podem ser invocados via AWS Config, funções do Lambda, Amazon CloudWatch ou AWS Trusted Advisor. Você também pode usar filtros de métrica no CloudWatch Logs para pesquisar e extrair padrões nos logs para determinar se o uso está se aproximando dos limites de cota.

Etapas de implementação

Para monitoramento:

- Capture o consumo atual de recursos (por exemplo, buckets ou instâncias). Use as operações de API de serviço, como a API `DescribeInstances` do Amazon EC2, para coletar o consumo atual de recursos.
- Capture as cotas atuais que são essenciais e aplicáveis aos serviços usando:
  - AWS Service Quotas
  - AWS Trusted Advisor
  - Documentação AWS
  - Páginas específicas de serviços da AWS
  - AWS Command Line Interface (AWS CLI)
  - AWS Cloud Development Kit (AWS CDK)
- Use o AWS Service Quotas, um serviço da AWS que ajuda a gerenciar em um único local suas cotas para mais de 250 serviços da AWS.
- Use os limites de serviço do Trusted Advisor para monitorar os limites de serviço atuais em vários limites.
- Use o histórico da cota de serviço (console ou AWS CLI) para verificar os aumentos regionais.
- Compare as alterações na cota de serviço em cada região e cada conta para criar equivalência, se necessário.

Para gerenciamento:

- Automático: configure uma regra personalizada do AWS Config para verificar as cotas de serviço nas regiões e comparar as diferenças.
- Automático: configure uma função do Lambda agendada para verificar as cotas de serviço nas regiões e comparar as diferenças.
- Manual: verifique as cotas de serviço via AWS CLI, API ou Console da AWS para conferir as cotas de serviço nas regiões e comparar as diferenças. Informe as diferenças.
- Se diferenças nas cotas entre as regiões forem identificadas, solicite uma mudança na cota, se necessário.
- Revise o resultado de todas as solicitações.

Recursos

Práticas recomendadas relacionadas:

- REL01-BP01 Conhecer as cotas e restrições de serviços
- REL01-BP02 Gerenciar cotas de serviço em várias contas e regiões
- REL01-BP03 Acomodar restrições e cotas de serviço fixo por meio de arquitetura
- REL01-BP05 Automatizar o gerenciamento de cotas
- REL01-BP06 Garantir que existe uma lacuna suficiente entre as cotas atuais e o uso máximo para acomodar o failover
- REL03-BP01 Escolher como segmentar a workload
- REL10-BP01 Implantar a workload em vários locais
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas
- REL11-BP03 Automatizar a reparação em todas as camadas
- REL12-BP05 Testar a resiliência via engenharia do caos

Documentos relacionados:

- Pilar Confiabilidade do AWS Well-Architected Framework: Disponibilidade
- AWS Service Quotas (antigamente conhecido como Limites de serviço)
- Verificações de práticas recomendadas do AWS Trusted Advisor (consulte a seção Limites de serviço)
- Monitor de limites da AWS em respostas da AWS
- Limites de serviço do Amazon EC2
- O que é o Service Quotas?
- Como solicitar um aumento da cota
- Endpoints e cotas de serviço
- Guia do usuário do Service Quotas
- Monitor de cotas para AWS
- Limites de isolamento de falhas da AWS
- Disponibilidade com redundância
- AWS para dados
- O que é integração contínua?

- O que é entrega contínua?
- Parceiro da APN: parceiros que podem ajudar no gerenciamento de configuração
- Gerenciar o ciclo de vida da conta em ambientes SaaS de conta por locatário na AWS
- Gerenciar e monitorar o controle de utilização de APIs em suas workloads
- Visualizar recomendações do AWS Trusted Advisor em grande escala com o AWS Organizations
- Automatizar aumentos de limites de serviço e suporte corporativo com o AWS Control Tower
- Ações, recursos e chaves de condição para o Service Quotas

Vídeos relacionados:

- AWS Live re:Inforce 2019: Service Quotas
- Visualizar e gerenciar cotas para serviços da AWS com o Service Quotas
- Demonstração de cotas do AWS IAM
- AWS re:Invent 2018: Fechar loops e abrir mentes: como assumir o controle de sistemas grandes e pequenos

Ferramentas relacionadas:

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP05 Automatizar o gerenciamento de cotas

Implemente ferramentas para alertar quando os limites estiverem perto de ser atingidos. É possível automatizar as solicitações de aumento de cota usando as APIs do AWS Service Quotas.

Se você integrar o Configuration Management Database (CMDB) ou sistema de emissão de tíquetes ao Service Quotas, poderá automatizar o rastreamento de solicitações de aumento de cota e as cotas atuais. Além do AWS SDK, o Service Quotas oferece automação usando a AWS Command Line Interface (AWS CLI).

Práticas comuns que devem ser evitadas:

- rastrear as cotas e o uso em planilhas.
- Executar relatórios sobre o uso diário, semanal ou mensal e comparar o uso com as cotas.

Benefícios de implementar esta prática recomendada: o rastreamento automático das cotas de serviço da AWS e o monitoramento do seu uso em relação a essas cotas permitirão que você veja quando está perto de atingir uma cota. Você pode configurar a automação para ajudar a solicitar um aumento de cota quando necessário. Talvez seja interessante considerar a redução de algumas cotas quando seu uso estiver na direção oposta para aproveitar os benefícios do risco reduzido (no caso de credenciais comprometidas) e da redução de custos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

- Configure o monitoramento automatizado e implemente ferramentas que usem SDKs para alertar você quando os limites estiverem perto de ser atingidos.
  - Use o Service Quotas e complemente o serviço com uma solução automatizada de monitoramento de cotas, como o AWS Limit Monitor ou uma oferta do AWS Marketplace.
    - O que é o Service Quotas?
    - Monitor de cotas na AWS – Solução da AWS
  - Configure respostas automatizadas com base nos limites de cota via APIs do Amazon SNS e do AWS Service Quotas.
  - Teste a automação.
    - Configure os limites.
    - Integre-se a eventos de alteração do AWS Config, de pipelines de implantação, do Amazon EventBridge ou de terceiros.
    - Defina artificialmente limites baixos de cota para testar as respostas.
    - Configure operações automatizadas para tomar as medidas apropriadas ao receber notificações e entre em contato com o AWS Support quando necessário.

- Inicie manualmente os eventos de alteração.
- Realize um game day para testar o processo de alteração de aumento de cota.

Recursos

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudar no gerenciamento da configuração
- AWS Marketplace: produtos CMDB que ajudam a rastrear os limites
- AWS Service Quotas (antigamente conhecido como Limites de serviço)
- Verificações de práticas recomendadas do AWS Trusted Advisor (consulte a seção Limites de serviço)
- Monitor de cotas na AWS – Solução da AWS
- Limites de serviço do Amazon EC2
- O que é o Service Quotas?

Vídeos relacionados:

- AWS Live re:Inforce 2019: Service Quotas

REL01-BP06 Garantir que existe uma lacuna suficiente entre as cotas atuais e o uso máximo para acomodar o failover

Este artigo explica como manter uma distância entre a cota do recurso e seu uso e como isso pode beneficiar sua organização. Quando você termina de usar um recurso, a cota de uso pode continuar contabilizando esse recurso. Isso pode resultar em falha ou em um recurso inacessível. Previna a falha do recurso verificando se as cotas abrangem a sobreposição de recursos inacessíveis e suas substituições. Considere casos de uso como falha de rede, falha na zona de disponibilidade ou falhas regionais ao calcular essa lacuna.

Resultado desejado: falhas pequenas ou grandes nos recursos ou na acessibilidade dos recursos podem ser cobertas dentro dos limites atuais do serviço. As falhas de zona, falhas de rede ou até mesmo falhas regionais foram consideradas no planejamento de recursos.

Práticas comuns que devem ser evitadas:

- Configurar cotas de serviço com base nas necessidades atuais sem considerar os cenários de failover.
- Não considerar as entidades principais de estabilidade estática ao calcular a cota de pico de um serviço.
- Não considerar o potencial de recursos inacessíveis no cálculo da cota total necessária para cada região.
- Não considerar os limites de isolamento de falhas de serviço da AWS para alguns serviços e seus padrões de uso possivelmente anormais.

Benefícios de implementar esta prática recomendada: quando eventos de interrupção do serviço afetam a disponibilidade da aplicação, use a nuvem para implementar estratégias para se recuperar desses eventos. Um exemplo de estratégia é criar recursos adicionais para substituir recursos inacessíveis e acomodar condições de failover sem esgotar seu limite de serviço.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Ao avaliar os limites de cota, considere casos de failover que podem ocorrer devido a algum dano. Considere os seguintes casos de falha:

- Uma VPC interrompida ou inacessível.
- Uma sub-rede inacessível.
- Uma zona de disponibilidade degradada que afeta a acessibilidade dos recursos.
- Rotas de rede ou pontos de ingresso e egresso são bloqueados ou alterados.
- Uma região degradada que afeta a acessibilidade dos recursos.
- Um subconjunto de recursos afetados por uma falha em uma região ou zona de disponibilidade.

A decisão de fazer o failover é única para cada situação, já que o impacto na empresa pode variar drasticamente. Aborde o planejamento da capacidade dos recursos no local de failover e as cotas dos recursos antes de decidir fazer o failover de uma aplicação ou serviço.

Considere picos de atividade acima do normal ao revisar as cotas de cada serviço. Esses picos podem estar relacionados a recursos que estão inacessíveis por questões de rede ou permissões, mas ainda estão ativos. Os recursos ativos não encerrados são contabilizados no limite de cota do serviço.

Etapas de implementação

- Mantenha distância suficiente entre a cota de serviço e o uso máximo para acomodar um failover ou uma perda de acessibilidade.
- Determine suas cotas de serviço. Considere os padrões típicos de implantação, os requisitos de disponibilidade e o crescimento do consumo.
- Solicite aumentos de cota, se necessário. Preveja um tempo de espera para a solicitação de aumento de cota.
- Determine os requisitos de confiabilidade (também conhecidos como "número de noves").
- Entenda possíveis cenários de falha, como perda de um componente, zona de disponibilidade ou região.
- Estabeleça a metodologia de implantação (por exemplo, canário, azul/verde, vermelho/preto ou gradual).
- Inclua uma reserva adequada do limite atual. Um exemplo de buffer pode ser de 15%.
- Inclua cálculos para estabilidade estática (por zona e região), quando apropriado.
- Planeje o aumento do consumo (por exemplo, monitore suas tendências de consumo).
- Considere o impacto da estabilidade estática das suas workloads mais críticas. Avalie os recursos em conformidade com um sistema estaticamente estável em todas as regiões e zonas de disponibilidade.
- Considere usar reservas de capacidade sob demanda para programar a capacidade à frente de qualquer failover. Isso pode ser uma estratégia útil durante os cronogramas empresariais mais críticos a fim de reduzir possíveis riscos de obter a quantidade e o tipo certo de recursos durante o failover.

Recursos

Práticas recomendadas relacionadas:

- REL01-BP01 Conhecer as cotas e restrições de serviços
- REL01-BP02 Gerenciar cotas de serviço em várias contas e regiões
- REL01-BP03 Acomodar restrições e cotas de serviço fixo por meio de arquitetura
- REL01-BP04 Monitorar e gerenciar cotas
- REL01-BP05 Automatizar o gerenciamento de cotas
- REL03-BP01 Escolher como segmentar a workload

- REL10-BP01 Implantar a workload em vários locais
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas
- REL11-BP03 Automatizar a reparação em todas as camadas
- REL12-BP05 Testar a resiliência via engenharia do caos

Documentos relacionados:

- Pilar Confiabilidade do AWS Well-Architected Framework: Disponibilidade
- AWS Service Quotas (antigamente conhecido como Limites de serviço)
- Verificações de práticas recomendadas do AWS Trusted Advisor (consulte a seção Limites de serviço)
- Monitor de limites da AWS em respostas da AWS
- Limites de serviço do Amazon EC2
- O que é o Service Quotas?
- Como solicitar um aumento da cota
- Endpoints e cotas de serviço
- Guia do usuário do Service Quotas
- Monitor de cotas para AWS
- Limites de isolamento de falhas da AWS
- Disponibilidade com redundância
- AWS para dados
- O que é integração contínua?
- O que é entrega contínua?
- Parceiro da APN: parceiros que podem ajudar no gerenciamento de configuração
- Gerenciar o ciclo de vida da conta em ambientes SaaS de conta por locatário na AWS
- Gerenciar e monitorar o controle de utilização de APIs em suas workloads
- Visualizar recomendações do AWS Trusted Advisor em grande escala com o AWS Organizations
- Automatizar aumentos de limites de serviço e suporte corporativo com o AWS Control Tower
- Ações, recursos e chaves de condição para o Service Quotas

Vídeos relacionados:

- AWS Live re:Inforce 2019: Service Quotas
- Visualizar e gerenciar cotas para serviços da AWS com o Service Quotas
- Demonstração de cotas do AWS IAM
- AWS re:Invent 2018: Fechar loops e abrir mentes: como assumir o controle de sistemas grandes e pequenos

Ferramentas relacionadas:

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL 2. Como você planeja sua topologia de rede?

Geralmente, existem workloads em vários ambientes. Isso inclui vários ambientes de nuvem (acessíveis ao público e privados) e, possivelmente, a infraestrutura de datacenter existente. Os planos devem incluir considerações de rede, como conectividade dentro dos sistemas e, entre eles, gerenciamento de endereços IP públicos e privados e resolução de nomes de domínio.

Práticas recomendadas

- REL02-BP01 Usar conectividade de rede altamente disponível em endpoints públicos de workloads
- REL02-BP02 Provisionar conectividade redundante entre redes privadas na nuvem e ambientes on-premises
- REL02-BP03 Garantir contas de alocação de sub-rede IP para expansão e disponibilidade
- REL02-BP04 Preferir topologias hub-and-spoke em vez de malha muitos para muitos

- REL02-BP05 Aplicar intervalos de endereços IP privados não sobrepostos a todos os espaços de endereços privados onde estão conectados

REL02-BP01 Usar conectividade de rede altamente disponível em endpoints públicos de workloads

Construir conectividade de rede altamente disponível nos endpoints públicos das workloads pode ajudar a reduzir o tempo de inatividade devido à perda de conectividade e melhorar a disponibilidade e o SLA da workload. Para que isso seja possível, use DNS altamente disponível, redes de entrega de conteúdo (CDNs), API gateways, balanceamento de carga ou proxies reversos.

Resultado desejado: é fundamental planejar, criar e operacionalizar a conectividade de rede altamente disponível para seus endpoints públicos. Se a workload se tornar inacessível devido a uma perda de conectividade, mesmo se ela estiver em execução e indisponível, os clientes verão o sistema como inativo. Ao combinar a conectividade de rede altamente disponível e resiliente para os endpoints públicos da workload, junto com uma arquitetura resiliente para a própria workload, é possível fornecer o melhor nível possível de serviço e disponibilidade possível aos clientes.

AWS Global Accelerator, Amazon CloudFront, Amazon API Gateway, URLs de funções do AWS Lambda, APIs da AWS AppSync e Elastic Load Balancing (ELB): todos fornecem endpoints públicos altamente disponíveis. O Amazon Route 53 fornece um serviço de DNS altamente disponível para resolução de nomes de domínio para verificar se seus endereços de endpoints públicos podem ser resolvidos.

Também é possível avaliar os appliances de software do AWS Marketplace com relação ao proxy e ao balanceamento de carga.

Práticas comuns que devem ser evitadas:

- Projetar uma workload altamente disponível sem planejar a alta disponibilidade do DNS e da conectividade de rede.
- Usar endereços de internet públicos em instâncias ou contêineres individuais e gerenciar a conectividade com eles por meio de DNS.
- Usar endereços IP em vez de nomes de domínio para localizar serviços.
- Não testar cenários em que a conectividade com os endpoints públicos é perdida.
- Não analisar as necessidades de throughput de rede e os padrões de distribuição.
- Não testar nem se planejar para cenários em que a conectividade de rede da internet com os endpoints públicos da workload possam ser interrompidos.

- Fornecer conteúdo (como páginas da web, ativos estáticos ou arquivos de mídia) para uma grande área geográfica e não usar uma rede de entrega de conteúdo.
- Não se planejar para ataques de negação distribuída de serviços (DDoS). Ataques de DDoS representam um risco de obstruir o tráfego legítimo e reduzir a disponibilidade para os usuários.

Benefícios de implementar esta prática recomendada: projetar para uma conectividade de rede altamente disponível e resiliente garante que sua workload permaneça acessível e disponível para seus usuários.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

No centro da criação de conectividade de rede altamente disponível com os endpoints públicos está o roteamento do tráfego. Para verificar se o tráfego consegue acessar os endpoints, o DNS deve poder resolver os nomes de domínio para os endereços IP correspondentes. Use um Sistema de Nomes de Domínio (DNS) altamente disponível e dimensionável, como o Amazon Route 53 para gerenciar os registros de DNS do seu domínio. Também é possível usar verificações de integridade fornecidas pelo Amazon Route 53. As verificações de integridade conferem se a aplicação está acessível, disponível e funcional e podem ser configuradas de uma maneira que imitem o comportamento do usuário, como solicitar uma página da web ou um URL específico. Em caso de falha, o Amazon Route 53 responde às solicitações de resolução de DNS e direciona o tráfego somente aos endpoints íntegros. Também é possível considerar o uso dos recursos DNS GEO e Roteamento baseado em latência oferecidos pelo Amazon Route 53.

Para verificar se sua workload está altamente disponível, use o Elastic Load Balancing (ELB). O Amazon Route 53 pode ser usado para direcionar o tráfego para o ELB, que distribui o tráfego para as instâncias computacionais de destino. Também é possível usar o Amazon API Gateway com o AWS Lambda para obter uma solução sem servidor. Os clientes também podem executar workloads em várias Regiões da AWS. Com o padrão ativo/ativo de vários sites, a workload pode atender ao tráfego de várias regiões. Com um padrão ativo/passivo de vários sites, a workload atende ao tráfego da região ativa enquanto os dados são replicados para a região secundária e se tornam ativos no caso de uma falha na região primária. As verificações de integridade do Route 53 podem então ser usadas para controlar o failover de DNS de qualquer endpoint em uma região primária para um endpoint em uma região secundária, verificando se sua workload está acessível e disponível para seus usuários.

O Amazon CloudFront fornece uma API simples para distribuir o conteúdo com baixa latência e altas taxas de transferência de dados atendendo a solicitações usando uma rede de locais de borda ao redor do mundo. As redes de entrega de conteúdo (CDNs) atendem os clientes fornecendo conteúdo localizado ou armazenado em cache em um local próximo ao usuário. Isso também melhora a disponibilidade da aplicação à medida que a carga do conteúdo é transferida dos seus servidores para os locais da borda do CloudFront. Os locais da borda e os caches de borda regionais armazenam cópias em cache do conteúdo próximo aos visualizadores, resultando em recuperação rápida e aumentando a acessibilidade e a disponibilidade da workload.

Para workloads com usuários distribuídos geograficamente, o AWS Global Accelerator ajuda a melhorar a disponibilidade e a performance das aplicações. O AWS Global Accelerator fornece endereços IP estáticos anycast que servem como um ponto de entrada fixo para a aplicação hospedada em uma ou mais Regiões da AWS. Isso permite que o tráfego entre na rede global da AWS o mais próximo possível dos usuários, melhorando a acessibilidade e a disponibilidade da workload. O AWS Global Accelerator também monitora a integridade dos endpoints da aplicação usando as verificações de integridade de TCP, HTTP e HTTPS. Qualquer mudança na integridade ou na configuração dos endpoints aciona o redirecionamento do tráfego de usuários para endpoints íntegros que oferecem a melhor performance e disponibilidade aos usuários. Além disso, o AWS Global Accelerator tem um design de isolamento de falhas que usa dois endereços IPv4 estáticos que são fornecidos por zonas de rede independentes, aumentando a disponibilidade das aplicações.

Para ajudar a proteger os clientes contra ataques de DDoS, a AWS oferece o AWS Shield Standard. O Shield Standard é ativado automaticamente e protege contra ataques comuns de infraestrutura (camadas 3 e 4), como inundações SYN/UDP e ataques de reflexão, para oferecer suporte à alta disponibilidade de aplicações na AWS. Para obter mais proteções contra ataques maiores e mais sofisticados (como inundações de UDP), ataques de exaustão de estado (como inundações de TCP SYN) e para ajudar a proteger as aplicações executadas nos serviços Amazon Elastic Compute Cloud (Amazon EC2), Elastic Load Balancing (ELB), Amazon CloudFront, AWS Global Accelerator e Route 53, considere usar o AWS Shield Advanced. Para proteção contra ataques de camada de aplicação como inundações de HTTP POST e GET, use o AWS WAF. O AWS WAF pode usar condições de scripts entre sites, endereços IP, cabeçalhos HTTP, corpo HTTP, strings de URI e injeção de SQL para determinar se uma solicitação deve ser bloqueada ou permitida.

Etapas de implementação

1. Configure o DNS altamente disponível: o Amazon Route 53 é um serviço Web de Sistema de Nomes de Domínio (DNS) altamente disponível e escalável. O Route 53 conecta solicitações

de usuários a aplicações da Internet executadas na AWS ou on-premises. Para obter mais informações, consulte Como configurar o Amazon Route 53 como seu serviço de DNS.

2. Configure verificações de integridade: ao usar o Route 53, verifique se somente os destinos íntegros podem ser resolvidos. Comece criando verificações de integridade do Amazon Route 53 e configurando o failover de DNS. É importante levar em consideração os seguintes aspectos ao configurar verificações de integridade:
   a. Como o Amazon Route 53 determina a integridade de uma verificação de integridade
   b. Criar, atualizar e excluir verificações de integridade
   c. Monitorar o status da verificação de integridade e receber notificações
   d. Práticas recomendadas do Amazon Route 53 DNS
3. Conectar o serviço de DNS aos endpoints.
   a. Ao usar o Elastic Load Balancing como destino para seu tráfego, crie um registro de alias usando o Amazon Route 53 que aponta para o endpoint regional do seu balanceador de carga. Durante a criação do registro de alias, defina a opção "Avaliar integridade do destino" como "Sim".
   b. Para workloads sem servidor ou APIs privadas quando o API Gateway é usado, use o Route 53 para direcionar o tráfego para o API Gateway.
4. Decida sobre uma rede de entrega de conteúdo.
   a. Para entregar conteúdo usando pontos de presença mais próximos do usuário, comece entendendo como o CloudFront entrega conteúdo.
   b. Aprenda os conceitos básicos de uma distribuição simples do CloudFront. O CloudFront então sabe de onde você quer que o conteúdo seja entregue e os detalhes sobre como rastrear e gerenciar a entrega de conteúdo. É importante entender e considerar os aspectos a seguir ao configurar uma distribuição do CloudFront:
      i. Como o armazenamento em cache funciona com os pontos de presença do CloudFront
      ii. Aumentar a taxa de solicitações fornecidas diretamente de caches do CloudFront (taxa de acertos do cache)
      iii. Usar o escudo de origem do Amazon CloudFront
      iv. Otimizar a alta disponibilidade com o failover de origem do CloudFront
5. Configure a proteção da camada da aplicação: o AWS WAF ajuda você a se proteger contra explorações e bots comuns da web que podem afetar a disponibilidade, comprometer a segurança ou consumir recursos em excesso. Para obter uma compreensão mais profunda, analise como o AWS WAF funciona e, quando você estiver pronto para implementar proteções contra inundações

de HTTP POST E GET na camada de aplicação, consulte Conceitos básicos do AWS WAF. Você também pode usar o AWS WAF com o CloudFront, consulte a documentação sobre como o AWS WAF funciona com os recursos do Amazon CloudFront.

6. Configure proteção adicional contra DDoS: por padrão, todos os clientes da AWS recebem proteção contra ataques de DDoS da camada de transporte e rede comuns e que ocorrem com mais frequência que visam seu site ou sua aplicação com o AWS Shield Standard sem custo adicional. Para proteção adicional de aplicações voltadas para a Internet executados no Amazon EC2, Elastic Load Balancing, Amazon CloudFront, AWS Global Accelerator e Amazon Route 53, você pode considerar o AWS Shield Advanced e revisar exemplos de arquiteturas resilientes a DDoS. Para proteger sua workload e seus endpoints públicos contra ataques de DDoS, consulte Conceitos básicos do AWS Shield Advanced.

Recursos

Práticas recomendadas relacionadas:

- REL10-BP01 Implantar a workload em vários locais
- REL10-BP02 Selecionar os locais apropriados para sua implantação de vários locais
- REL11-BP04 Confiar no plano de dados, e não no ambiente de gerenciamento, durante a recuperação
- REL11-BP06 Enviar notificações quando os eventos afetarem a disponibilidade

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudar a planejar sua rede
- AWS Marketplace para infraestrutura de rede
- O que é o AWS Global Accelerator?
- O que é o Amazon CloudFront?
- O que é o Amazon Route 53?
- O que é Elastic Load Balancing?
- Recurso de conectividade de rede: estabelecer suas bases da nuvem
- O que é o Amazon API Gateway?
- O que são AWS WAF, AWS Shield e AWS Firewall Manager?
- What is Amazon Application Recovery Controller?

- Configurar verificações de integridade personalizadas para failover de DNS

Vídeos relacionados:

- AWS re:Invent 2022: Aprimorar a performance e a disponibilidade com o AWS Global Accelerator
- AWS re:Invent 2020: Gerenciamento de tráfego global com o Amazon Route 53
- AWS re:Invent 2022: Operar aplicações Multi-AZ altamente disponíveis
- AWS re:Invent 2022: Mergulho profundo na infraestrutura de rede da AWS
- AWS re:Invent 2022: Construir redes resilientes

Exemplos relacionados:

- Disaster Recovery with Amazon Application Recovery Controller (ARC)
- Workshops de confiabilidade
- Workshop do AWS Global Accelerator

REL02-BP02 Provisionar conectividade redundante entre redes privadas na nuvem e ambientes on-premises

Implemente redundância nas conexões entre redes privadas na nuvem e ambientes on-premises a fim de obter resiliência de conectividade. Isso pode ser feito por meio da implantação de dois ou mais links e caminhos de tráfego, preservando a conectividade em caso de falhas na rede.

Práticas comuns que devem ser evitadas:

- Você depende de apenas uma conexão de rede, o que cria um ponto único de falha.
- Você usa somente um túnel VPN ou vários túneis que terminam na mesma zona de disponibilidade.
- Você depende de um ISP para conectividade VPN, o que pode levar a falhas completas durante interrupções do ISP.
- Não implementar protocolos de roteamento dinâmico, como o BGP, que são cruciais para redirecionar o tráfego durante interrupções na rede.
- Você ignora as limitações de largura de banda dos túneis VPN e superestima as respectivas capacidades de backup.

Benefícios de implementar esta prática recomendada: ao implementar conectividade redundante entre seu ambiente de nuvem e o ambiente corporativo ou on-premises, você pode garantir que os serviços dependentes entre os dois ambientes possam se comunicar de forma confiável.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Ao usar o AWS Direct Connect para conectar sua rede on-premises à AWS, é possível atingir a resiliência máxima da rede (SLA de 99,99%) utilizando conexões separadas que terminam em dispositivos distintos em mais de um ambiente on-premises e em mais de um local do AWS Direct Connect. Essa topologia oferece resiliência contra falhas de dispositivos, problemas de conectividade e interrupções de locais inteiros. Como alternativa, você pode obter alta resiliência (SLA de 99,9%) usando duas conexões individuais com vários locais (cada local on-premises conectado a um único local do Direct Connect). Essa abordagem oferece proteção contra interrupções de conectividade causadas por cortes na fibra ou falhas de dispositivos, além de ajudar a mitigar falhas de locais inteiros. O kit de ferramentas de resiliência do AWS Direct Connect pode ajudar a projetar sua topologia do AWS Direct Connect.

Você também pode considerar a terminação do AWS Site-to-Site VPN em um AWS Transit Gateway como um backup econômico para sua conexão primária do AWS Direct Connect. Essa configuração possibilita o roteamento multicaminho de custo igual (ECMP) em vários túneis VPN, permitindo um throughput de até 50 Gbps, mesmo que cada túnel VPN tenha um limite de 1,25 Gbps. No entanto, é importante observar que o AWS Direct Connect ainda é a opção mais eficaz para minimizar as interrupções na rede e oferecer conectividade estável.

Ao usar VPNs pela internet para conectar o ambiente de nuvem ao data center on-premises, configure dois túneis VPN como parte de uma única conexão do Site-to-Site VPN. Cada túnel deve terminar em uma zona de disponibilidade diferente para proporcionar alta disponibilidade, usar hardware redundante e evitar falhas no dispositivo on-premises. Além disso, considere várias conexões à internet de diversos provedores de serviços de Internet (ISPs) em seu local on-premises a fim de evitar a interrupção completa da conectividade VPN devido à interrupção de um único ISP. A seleção de ISPs com roteamento e infraestrutura diversos, especialmente aqueles com caminhos físicos separados até endpoints da AWS, fornece alta disponibilidade de conectividade.

Além da redundância física com várias conexões do AWS Direct Connect e vários túneis VPN (ou uma combinação de ambos), a implementação do roteamento dinâmico do Protocolo de Gateway da Borda (BGP) também é fundamental. O BGP dinâmico fornece redirecionamento automático do tráfego de um caminho para outro com base nas condições de rede em tempo real e nas

políticas configuradas. Esse comportamento dinâmico é especialmente benéfico para manter a disponibilidade da rede e a continuidade do serviço em caso de falhas no link ou na rede. Ele seleciona rapidamente caminhos alternativos, aumentando a resiliência e a confiabilidade da rede.

Etapas de implementação

- Adquira conectividade de alta disponibilidade entre a AWS e seu ambiente on-premises.
  - Use várias conexões do AWS Direct Connect ou túneis VPN entre as redes privadas implantadas separadamente.
  - Use vários locais do AWS Direct Connect para gerar alta disponibilidade.
  - Se estiver usando várias Regiões da AWS, crie redundância em pelo menos duas delas.
- Use AWS Transit Gateway, quando possível, para encerrar sua conexão VPN.
- Avalie os appliances do AWS Marketplace para acabar com as VPNs ou estender sua SD-WAN para a AWS. Se você usa appliances do AWS Marketplace, implante instâncias redundantes em zonas de disponibilidade diferentes para alta disponibilidade.
- Forneça uma conexão redundante com o ambiente on-premises.
  - Você pode precisar de conexões redundantes com várias Regiões da AWS para atender às suas necessidades de disponibilidade.
  - Use o Kit de ferramentas de resiliência do AWS Direct Connect para começar.

Recursos

Documentos relacionados:

- Recomendações de resiliência da AWS Direct Connect
- Usar conexões VPN site a site redundantes para realizar failover
- Políticas de roteamento e comunidades BGP
- Configurações Ativas/Ativas e Ativas/Passivas no AWS Direct Connect
- Parceiro da APN: parceiros que podem ajudar a planejar sua rede
- AWS Marketplace para infraestrutura de rede
- Whitepaper de opções de conectividade da Amazon Virtual Private Cloud
- Construção de uma infraestrutura de rede da AWS Multi-VPC escalável e segura
- Usar conexões VPN site a site redundantes para realizar failover
- Usar o Kit de ferramentas de resiliência do AWS Direct Connect para começar

- Endpoints da VPC e serviços de endpoint da VPC (AWS PrivateLink)
- O que é Amazon VPC?
- O que é um gateway de trânsito?
- O que é AWS Site-to-Site VPN?
- Trabalhar com gateways Direct Connect

Vídeos relacionados:

- AWS re:Invent 2018: Design de VPCs avançadas e novos recursos para Amazon VPC
- AWS re:Invent 2019: Arquiteturas de referência do AWS Transit Gateway para muitas VPCs

REL02-BP03 Garantir contas de alocação de sub-rede IP para expansão e disponibilidade

Os intervalos de endereços IP da Amazon VPC devem ser grandes o suficiente para acomodar os requisitos da workload, incluindo a futura expansão e alocação de endereços IP para sub-redes nas zonas de disponibilidade. Isso inclui balanceadores de carga, instâncias do EC2 e aplicações baseadas em contêiner.

Ao planejar sua topologia de rede, a primeira etapa é definir o espaço do endereço IP em si. Intervalos de endereços IP privados (seguindo as diretrizes RFC 1918) devem ser alocados para cada VPC. Atenda aos seguintes requisitos como parte desse processo:

- Permitir espaço de endereço IP para mais de uma VPC por região.
- Em uma VPC, deixe espaço para várias sub-redes para cobrir várias zonas de disponibilidade.
- Considere deixar o espaço de bloco CIDR não utilizado em uma VPC para expansão futura.
- Verifique se há espaço de endereço IP para atender às necessidades de qualquer frota transitória de instâncias do Amazon EC2 que você use, como frotas spot para machine learning, clusters do Amazon EMR ou clusters do Amazon Redshift. Consideração semelhante deve ser dada aos clusters do Kubernetes, como o Amazon Elastic Kubernetes Service (Amazon EKS), pois cada pod do Kubernetes recebe um endereço roteável do bloco CIDR da VPC por padrão.
- Observe que os primeiros quatro endereços IP e o último endereço IP em cada bloco CIDR da sub-rede estão reservados e não estão disponíveis para seu uso.
- Observe que o bloco CIDR inicial da VPC alocado para sua VPC não pode ser alterado ou excluído, mas você pode adicionar blocos CIDR não sobrepostos à VPC. Os CIDRs IPv4 da sub-rede não podem ser alterados, mas os CIDRs IPv6 podem.

- O maior bloco CIDR de VPC possível é /16 e o menor é /28.
- Considere outras redes conectadas (VPC, on-premises ou outros provedores de nuvem) e garanta que o espaço de endereço IP não se sobreponha. Para obter mais informações, consulte REL02-BP05 Aplicar intervalos de endereços IP privados não sobrepostos a todos os espaços de endereços privados onde estão conectados.

Resultado desejado: uma sub-rede IP escalável pode ajudar você a se adaptar ao crescimento futuro e evitar desperdícios desnecessários.

Práticas comuns que devem ser evitadas:

- Deixar de considerar o crescimento futuro, o que resulta em blocos CIDR muito pequenos que exigem reconfiguração e causando um possível tempo de inatividade.
- Estimar incorretamente quantos endereços IP um Elastic Load Balancer pode usar.
- Implantar muitos balanceadores de carga de alto tráfego nas mesmas sub-redes
- Usar mecanismos de ajuste de escala automático automatizados sem monitorar o consumo de endereços IP.
- Definir intervalos CIDR excessivamente grandes muito além das expectativas de crescimento futuro, o que pode dificultar o emparelhamento com outras redes com intervalos de endereços sobrepostos.

Benefícios de implementar esta prática recomendada: isso garante que você possa acomodar o crescimento das suas workloads e continuar a fornecer disponibilidade à medida que elas se expandem.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Planeje sua rede para acomodar crescimento, conformidade regulatória e integração com outras pessoas. O crescimento pode ser subestimado, a conformidade regulatória pode mudar e as aquisições ou conexões de rede privada podem ser difíceis de implementar sem o planejamento adequado.

- Selecione as Contas da AWS e regiões relevantes conforme seus requisitos de serviço, de latência, regulatórios e de recuperação de desastres (DR).
- Identifique suas necessidades para implantações regionais de VPC.

- Identifique o tamanho das VPCs.
  - Determine se você pretende implantar conectividade com várias VPCs.
    - O que é um gateway de trânsito?
    - Conectividade com várias VPCs de região única
  - Determine se você precisa de rede segregada por questões regulatórias.
  - Crie VPCs com blocos CIDR de tamanho adequado para acomodar suas necessidades atuais e futuras.
    - Se você tiver projeções de crescimento desconhecidas, talvez prefira arriscar blocos CIDR maiores para reduzir a possibilidade de reconfiguração futura
  - Considere usar o endereçamento IPv6 para sub-redes como parte de uma VPC de pilha dupla. O IPv6 é adequado para sub-redes privadas contendo frotas de instâncias ou contêineres efêmeros que, de outra forma, exigiriam um grande número de endereços IPv4.

Recursos

Práticas recomendadas do Well-Architected relacionadas:

- REL02-BP05 Aplicar intervalos de endereços IP privados não sobrepostos a todos os espaços de endereços privados onde estão conectados

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudar a planejar sua rede
- AWS Marketplace para infraestrutura de rede
- Whitepaper de opções de conectividade da Amazon Virtual Private Cloud
- Conectividade de rede de alta disponibilidade de vários datacenters
- Conectividade com várias VPCs de região única
- O que é Amazon VPC?
- IPv6 na AWS
- IPv6 em arquiteturas de referência
- Amazon Elastic Kubernetes Service lança suporte a IPv6
- Recommendations for your VPC: Classic Load Balancers
- Sub-redes de zona de disponibilidade: Application Load Balancers

- Zonas de disponibilidade: Network Load Balancers

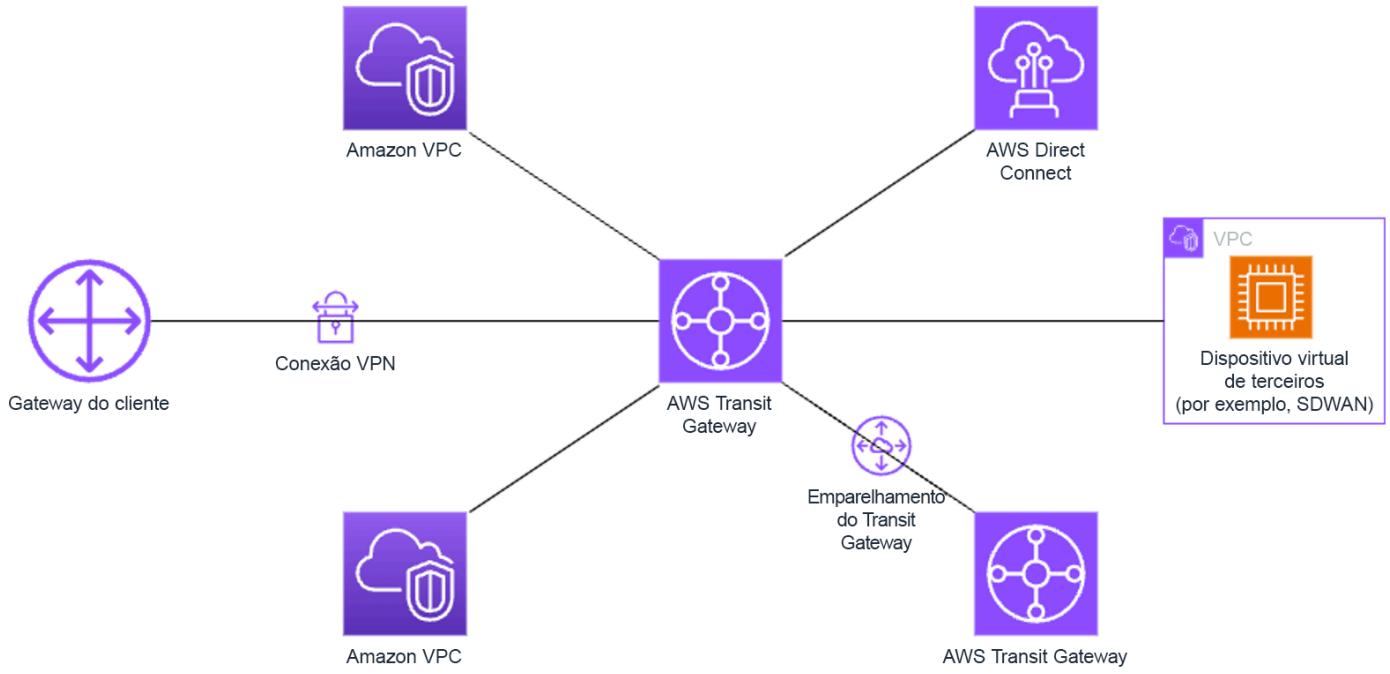
Vídeos relacionados:

- AWS re:Invent 2018: Design de VPCs avançadas e novos recursos para Amazon VPC (NET303)
- AWS re:Invent 2019: Arquiteturas de referência do AWS Transit Gateway para muitas VPCs (NET406-R1)
- AWS re:Invent 2023: AWS pronto para o que vem a seguir? Desenvolver redes para crescimento e flexibilidade (NET310)

REL02-BP04 Preferir topologias hub-and-spoke em vez de malha muitos para muitos

Ao conectar várias redes privadas, como nuvens privadas virtuais (VPCs) e redes on-premises, opte por uma topologia hub-and-spoke em vez de uma em malha. Diferentemente das topologias em malha, em que cada rede se conecta diretamente às outras e aumenta a complexidade e as despesas indiretas de gerenciamento, a arquitetura hub-and-spoke centraliza as conexões por meio de um único hub. Essa centralização simplifica a estrutura da rede e aprimora a operabilidade, a escalabilidade e o controle.

O AWS Transit Gateway é um serviço gerenciado, escalável e de alta disponibilidade projetado para a construção de redes hub-and-spoke na AWS. Ele serve como o hub central da rede que fornece segmentação de rede, roteamento centralizado e conexão simplificada com ambientes on-premises e na nuvem. A figura a seguir ilustra como você pode usar o AWS Transit Gateway para criar a topologia hub-and-spoke.



Práticas comuns que devem ser evitadas:

- Complicar demais as políticas de roteamento em uma arquitetura hub-and-spoke, o que reduz a eficiência da rede e complica tanto a solução de problemas quanto o gerenciamento proativo.
- A segmentação insuficiente baseada em roteamento dentro do hub pode causar vulnerabilidades, o que possivelmente expõe a rede ao acesso não autorizado.
- Sem uma otimização cuidadosa, o tráfego roteado pelo hub pode gerar maiores custos de transferência de dados, especialmente para tráfego entre zonas de disponibilidade e regiões. Estratégias eficazes de gerenciamento de tráfego são essenciais para controlar as despesas.

Benefícios de implementar esta prática recomendada: à medida que o número de redes conectadas aumenta, o gerenciamento e a expansão da conectividade em malha se tornam cada vez mais desafiadores. O AWS Transit Gateway oferece um hub gerenciado escalável e confiável para construção e operação de suas topologias hub-and-spoke. Ao usar o AWS Transit Gateway, é possível estabelecer conexões e centralizar o roteamento de tráfego em várias redes.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

- Planeje sua rede.

- Crie o AWS Transit Gateway.
- Conecte as VPCs.
- Se necessário, crie conexões VPN ou gateways Direct Connect e associe-os ao gateway de trânsito.
- Defina como o tráfego é direcionado entre as VPCs conectadas e outras conexões por meio da configuração das tabelas de rotas do gateway de trânsito.
- Use o Amazon CloudWatch para monitorar e ajustar as configurações conforme necessário para otimizar a performance e os custos.

Recursos

Documentos relacionados:

- O que é um gateway de trânsito?
- Construção de uma infraestrutura de rede da AWS Multi-VPC escalável e segura
- Criar uma rede global usando o emparelhamento entre regiões do AWS Transit Gateway
- Opções de conectividade da Amazon Virtual Private Cloud
- Parceiro da APN: parceiros que podem ajudar a planejar sua rede
- AWS Marketplace para infraestrutura de rede

Vídeos relacionados:

- AWS re:Invent 2023: Fundamentos de rede na AWS
- AWS re:Invent 2023: Designs e novos recursos de VPCs avançadas

REL02-BP05 Aplicar intervalos de endereços IP privados não sobrepostos a todos os espaços de endereços privados onde estão conectados

Os intervalos de endereços IP de cada uma das VPCs não devem se sobrepor quando emparelhados, conectados via Transit Gateway ou conectados por VPN. Evite conflitos de endereço IP entre uma VPC e ambientes on-premises ou com outros provedores de nuvem que você usa. Você também deve ter uma maneira de alocar intervalos de endereços IP privados quando necessário. Um sistema de gerenciamento de endereços IP (IPAM) pode ajudar a automatizar isso.

Resultado desejado:

- Não há nenhum conflito de intervalo de endereços IP entre VPCs, ambientes on-premises ou outros provedores de nuvem.
- O gerenciamento adequado de endereços IP facilita o ajuste de escala da infraestrutura de rede para atender ao crescimento e às mudanças nos requisitos de rede.

Práticas comuns que devem ser evitadas:

- Usar o mesmo intervalo de IPs na VPC que você tem on-premises, na rede corporativa ou em outros provedores de nuvem
- Não rastrear os intervalos IPs das VPCs usadas para implantar suas workloads.
- Depender de processos manuais de gerenciamento de endereços IP, como planilhas.
- Superdimensionar ou subdimensionar blocos CIDR, o que resulta em desperdício de endereços IP ou espaço de endereços insuficiente para a workload.

Benefícios de implementar esta prática recomendada: o planejamento ativo da rede garantirá que você não tenha várias ocorrências do mesmo endereço IP em redes interconectadas. Isso evita que problemas de roteamento ocorram em partes da workload que usam as diferentes aplicações.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Use um IPAM, como o Gerenciador de endereços IP da Amazon VPC, para monitorar e gerenciar seu uso do CIDR. Vários IPAMs estão disponíveis no AWS Marketplace. Avalie seu uso potencial na AWS, adicione intervalos de CIDR às VPCs existentes e crie VPCs para permitir um crescimento planejado no uso.

Etapas de implementação

- Colete o consumo atual de CIDR (por exemplo, VPCs e sub-redes).
  - Use as operações de API de serviço para coletar o consumo atual de CIDR.
  - Use o Gerenciador de endereços IP da Amazon VPC para descobrir recursos.
- Capture seu uso atual de sub-rede.
  - Use as operações de API de serviço para coletar sub-redes por VPC em cada região.
  - Use o Gerenciador de endereços IP da Amazon VPC para descobrir recursos.
- Registre o uso atual.

- Determine se você criou intervalos de IPs sobrepostos.
- Calcule a capacidade não utilizada.
- Identifique intervalos de IP sobrepostos. Você pode migrar para um novo intervalo de endereços ou considerar o uso de técnicas como gateway NAT privado ou AWS PrivateLink se precisar conectar os intervalos sobrepostos.

Recursos

Práticas recomendadas relacionadas:

- Proteção de redes

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudar a planejar sua rede
- AWS Marketplace para infraestrutura de rede
- Whitepaper de opções de conectividade da Amazon Virtual Private Cloud
- Conectividade de rede de alta disponibilidade de vários datacenters
- Conectar redes com intervalos de IP sobrepostos
- O que é Amazon VPC?
- O que é IPAM?

Vídeos relacionados:

- AWS re:Invent 2023: Designs e novos recursos de VPCs avançadas
- AWS re:Invent 2019: Arquiteturas de referência do AWS Transit Gateway para muitas VPCs
- AWS re:Invent 2023: Pronto para o que vem a seguir? Desenvolver redes para crescimento e flexibilidade
- AWS re:Invent 2021: {Novo lancamento} Gerenciar seus endereços IP em grande escala na AWS

Arquitetura da workload

Perguntas

- REL 3. Como projetar sua arquitetura de serviços de workload?

- REL 4. Como projetar interações em um sistema distribuído para evitar falhas?
- REL 5. Como você projeta interações em um sistema distribuído para mitigar ou resistir a falhas?

REL 3. Como projetar sua arquitetura de serviços de workload?

Use uma arquitetura orientada a serviços (SOA) ou uma arquitetura de microsserviços para criar workloads altamente escaláveis e confiáveis. A arquitetura orientada a serviços (SOA) é a prática de tornar componentes de software reutilizáveis por meio de interfaces de serviço. A arquitetura de microsserviços vai além para tornar os componentes menores e mais simples.

Práticas recomendadas

- REL03-BP01 Escolher como segmentar a workload
- REL03-BP02 Criar serviços voltados para domínios e funcionalidades de negócios específicos
- REL03-BP03 Fornecer contratos de serviço por API

REL03-BP01 Escolher como segmentar a workload

A segmentação de workloads é importante ao determinar os requisitos de resiliência da sua aplicação. Uma arquitetura monolítica deve ser evitada sempre que possível. Em vez disso, considere cuidadosamente quais componentes da aplicação podem ser distribuídos em microsserviços. Dependendo dos requisitos de sua aplicação, isso pode acabar sendo uma combinação de uma arquitetura orientada a serviços (SOA) com microsserviços sempre que possível. Workloads com capacidade para serem do tipo sem estado têm maior chance de ser implantadas como microsserviços.

Resultado desejado: as workloads devem ser compatíveis, escaláveis e o mais vagamente agrupadas possível.

Ao tomar decisões sobre como segmentar uma workload, pondere os benefícios e as complexidades. O que é ideal para um novo produto a caminho do seu primeiro lançamento não se aplica a uma workload que foi criada para ajuste de escala a partir das necessidades iniciais. Ao refatorar um monólito existente, será necessário considerar o quanto a aplicação poderá oferecer um bom suporte a uma decomposição em direção à condição sem estado. A divisão dos serviços em pedaços menores permite que equipes pequenas e bem definidas os desenvolvam e gerenciem. No entanto, serviços menores podem introduzir complexidades que incluem maior latência potencial, depuração mais complexa e carga operacional aumentada.

Práticas comuns que devem ser evitadas:

- O microsserviço Death Star é uma situação em que os componentes atômicos se tornam tão altamente interdependentes que a falha de um resulta em uma falha muito maior, o que torna os componentes tão rígidos e frágeis quanto um monólito.

Benefícios de estabelecer esta prática:

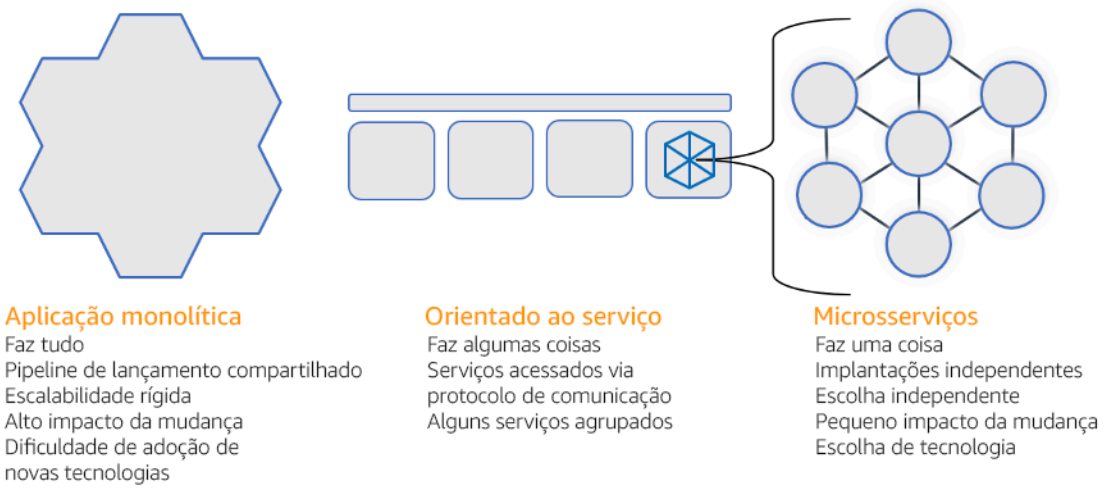
- Mais segmentos específicos geram maior agilidade, flexibilidade organizacional e escalabilidade.
- Redução do impacto das interrupções do serviço.
- Os componentes da aplicação podem ter requisitos de disponibilidade diferentes, aos quais uma segmentação mais atômica pode oferecer suporte.
- Responsabilidades bem definidas para as equipes que oferecem suporte à workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Escolha o tipo de arquitetura com base no modo como você segmentará a workload. Escolha uma SOA ou arquitetura de microsserviços (ou, em alguns casos, uma arquitetura monolítica). Mesmo que você opte por começar com uma arquitetura monolítica, é necessário garantir que ela seja modular e tenha a capacidade de evoluir para SOA ou microsserviços à medida que o produto escala com a adoção do usuário. A SOA e os microsserviços oferecem, respectivamente, segmentação menor, que é preferida como uma arquitetura moderna escalável e confiável, mas há compensações a serem consideradas, especialmente ao implantar uma arquitetura de microsserviços.

Uma compensação primária é que você agora tem uma arquitetura de computação distribuída que pode tornar mais difícil alcançar requisitos de latência do usuário final, e há complexidade adicional na depuração e no rastreamento de interações com o usuário. Use o AWS X-Ray para ajudar você a resolver esse problema. Outro efeito a ser considerado é o aumento da complexidade operacional à medida que você aumenta o número de aplicações que está gerenciando, o que requer a implantação de vários componentes de independência.



Arquiteturas monolítica, orientada a serviços e de microsserviços

Etapas de implementação

- Determine a arquitetura adequada para refatorar ou desenvolver sua aplicação. A SOA e os microsserviços oferecem respectivamente segmentação menor, que é preferida por ser uma arquitetura moderna escalável e confiável. A SOA pode ser o meio-termo ideal para alcançar uma segmentação menor e também evitar algumas das complexidades dos microsserviços. Para obter mais detalhes, consulte Compensações de microsserviços.
- Se sua workload aceitá-la e sua organização puder sustentá-la, use uma arquitetura de microsserviços para obter a melhor agilidade e confiabilidade. Para obter mais informações, consulte Implementar microsserviços na AWS.
- Considere seguir o padrão Strangler Fig para refatorar um monólito em componentes menores. Isso envolve a substituição gradual de componentes específicos da aplicação por novos serviços e aplicações. O AWS Migration Hub Refactor Spaces atua como ponto de partida para a refatoração incremental. Para obter mais detalhes, consulte Migração simplificada de workloads on-premises herdadas usando um padrão strangler.
- A implementação de microsserviços pode exigir um mecanismo de descoberta de serviços para permitir que esses serviços distribuídos se comuniquem entre si. O AWS App Mesh pode ser usado com arquiteturas orientadas a serviços para fornecer descoberta e acesso confiáveis aos serviços. O AWS Cloud Map também pode ser usado para descoberta dinâmica de serviços baseada em DNS.
- Se você estiver migrando de um monólito para SOA, o Amazon MQ poderá ajudar a preencher a lacuna como um barramento de serviço ao redesenhar aplicações herdadas na nuvem.

- Para monólitos existentes com um único banco de dados compartilhado, escolha como reorganizar os dados em segmentos menores. Isso pode acontecer por unidade de negócios, padrão de acesso ou estrutura de dados. A esta altura no processo de refatoração, escolha se deseja prosseguir com um banco de dados relacional ou não relacional (NoSQL). Para obter mais detalhes, consulte Do SQL ao NoSQL.

Nível de esforço do plano de implementação: Alto

Recursos

Práticas recomendadas relacionadas:

- REL03-BP02 Criar serviços voltados para domínios e funcionalidades de negócios específicos

Documentos relacionados:

- Amazon API Gateway: configurar uma API REST usando a OpenAPI
- O que é arquitetura orientada a serviços?
- Contexto delimitado (um padrão central no design orientado por domínio)
- Implementar microsserviços na AWS
- Compensações de microsserviços
- Microsserviços: uma definição desse novo termo de arquitetura
- Microsserviços na AWS
- O que é AWS App Mesh?

Exemplos relacionados:

- Workshop Modernização iterativa de aplicações

Vídeos relacionados:

- Como entregar excelência com microsserviços na AWS

REL03-BP02 Criar serviços voltados para domínios e funcionalidades de negócios específicos

A arquitetura orientada a serviços (SOA) define serviços com funções bem delineadas estabelecidas pelas necessidades dos negócios. Os microsserviços usam modelos de domínio e contexto delimitado para traçar limites de serviço ao longo dos limites do contexto de negócios. O foco nos domínios de negócios e na funcionalidade ajuda as equipes a definir requisitos independentes de confiabilidade para seus serviços. Contextos delimitados isolam e encapsulam a lógica de negócios, permitindo que as equipes raciocinem melhor sobre como lidar com falhas.

Resultado desejado: em conjunto, engenheiros e partes interessadas do negócio definem contextos delimitados e os usam para projetar sistemas como serviços que cumprem funções empresariais específicas. Essas equipes usam práticas estabelecidas, como Event Storming, para definir os requisitos. As novas aplicações são projetadas como serviços, limites bem definidos e acoplamento fraco. Os monólitos existentes são decompostos em contextos limitados, e os projetos de sistemas migram para arquiteturas SOA ou de microsserviços. Quando os monólitos são refatorados, abordagens estabelecidas, como contextos de bolha e padrões de decomposição de monólitos, são aplicadas.

Os serviços orientados por domínios são executados como um ou mais processos que não compartilham o estado. Eles respondem de forma independente às flutuações na demanda e lidam com cenários de falha à luz dos requisitos específicos do domínio.

Práticas comuns que devem ser evitadas:

- As equipes são formadas em torno de domínios técnicos específicos, como UI e UX, middleware ou banco de dados, em vez de domínios empresariais específicos.
- As aplicações abrangem as responsabilidades do domínio. Serviços que abrangem contextos delimitados podem ser mais difíceis de manter, exigir maiores esforços de teste e que várias equipes de domínio participem das atualizações de software.
- As dependências de domínio, como as bibliotecas de entidades de domínio, são compartilhadas entre serviços de uma forma que as alterações em um domínio de serviço exijam alterações em outros domínios de serviço.
- Os contratos de serviço e a lógica de negócios não expressam entidades em uma linguagem de domínio comum e consistente, ocasionando camadas de tradução que complicam os sistemas e aumentam os esforços de depuração.

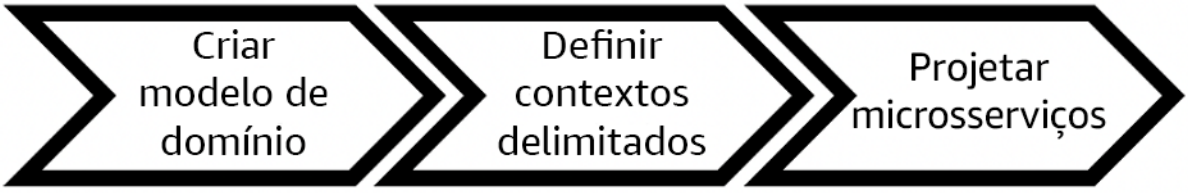
Benefícios de implementar esta prática recomendada: as aplicações são projetadas como serviços independentes delimitados por domínios de negócios e usam uma linguagem comercial comum.

Os serviços podem ser testados e implantados de forma independente. Os serviços atendem aos requisitos de resiliência específicos do domínio implementado.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

O design orientado por domínio (DDD) é a abordagem fundamental para projetar e criar software em torno de domínios empresariais. É útil trabalhar com um framework existente ao criar serviços voltados para domínios empresariais. Ao trabalhar com aplicações monolíticas existentes, você pode utilizar os padrões de decomposição que fornecem técnicas estabelecidas para modernizar aplicações em serviços.



Design orientado por domínio

Etapas de implementação

- As equipes podem realizar workshops de Event Storming a fim de identificar rapidamente eventos, comandos, agregados e domínios em um formato leve de notas adesivas.
- Depois que as entidades e funções de domínio forem formadas em um contexto de domínio, você poderá dividir seu domínio em serviços usando contexto limitado em que entidades que compartilham características e atributos semelhantes são agrupadas. Com o modelo dividido em contextos, surge um modelo de como delimitar microsserviços.
  - Por exemplo, as entidades do site Amazon.com podem incluir pacote, entrega, cronograma, preço, desconto e moeda.
  - Pacote, entrega e cronograma são agrupados no contexto de envio, enquanto preço, desconto e moeda são agrupados no contexto de preços.
- A decomposição de monólitos em microsserviços descreve padrões para refatorar microsserviços. O uso de padrões para decomposição por capacidade comercial, subdomínio ou transação se alinha bem às abordagens orientadas por domínio.
- Técnicas táticas como o contexto de bolha permitem introduzir o DDD em aplicações existentes ou legadas sem reformulações antecipadas e compromissos totais com o DDD. Em uma abordagem

de contexto de bolha, um pequeno contexto limitado é estabelecido usando um mapeamento e coordenação de serviços, ou camada corrompimento, que protege o modelo de domínio recém-definido contra influências externas.

Depois que as equipes realizarem a análise de domínio e definirem entidades e contratos de serviço, elas podem utilizar os serviços da AWS para implementar o design orientado por domínio como serviços baseados em nuvem.

- Comece o desenvolvimento definindo testes que simulem as regras de negócios do seu domínio. O desenvolvimento orientado por testes (TDD) e o desenvolvimento orientado por comportamento (BDD) ajudam as equipes a manter os serviços voltados para a solução de problemas de negócios.
- Selecione os serviços da AWS que melhor atendem aos requisitos de domínio da sua empresa e à arquitetura de microsserviços:
  - A tecnologia sem servidor da AWS permite que sua equipe enfoque a lógica de domínio específica em vez de gerenciar servidores e infraestrutura.
  - Os contêineres na AWS simplificam o gerenciamento de sua infraestrutura para que você possa enfocar nos requisitos de domínio.
  - Os bancos de dados com propósito específico ajudam você a adequar seus requisitos de domínio ao tipo de banco de dados mais adequado.
- Criar arquiteturas hexagonais na AWS descreve uma framework para criar lógica de negócios em serviços que funcionam retroativamente a partir de um domínio empresarial para atender aos requisitos funcionais e, depois, conectar adaptadores de integração. Os padrões que separam os detalhes da interface da lógica de negócios com serviços da AWS ajudam as equipes a enfocar na funcionalidade do domínio e melhorar a qualidade do software.

Recursos

Práticas recomendadas relacionadas:

- REL03-BP01 Escolher como segmentar a workload
- REL03-BP03 Fornecer contratos de serviço por API

Documentos relacionados:

- Microsserviços da AWS

- Implementar microsserviços na AWS
- Como dividir um monólito em microsserviços
- Conceitos básicos do DDD quando cercado por sistemas herdados
- Design orientado por domínio: como lidar com a complexidade no núcleo do software
- Criação de arquiteturas hexagonais na AWS
- Decompor monólitos em microsserviços
- Event Storming
- Mensagens entre contextos delimitados
- Microsserviços
- Desenvolvimento orientado por testes
- Desenvolvimento orientado por comportamento

Exemplos relacionados:

- Como projetar microsserviços nativos da nuvem na AWS (do DDD/EventStormingWorkshop)

Ferramentas relacionadas:

- Bancos de dados na Nuvem AWS
- Tecnologia sem servidor na AWS
- Contêineres na AWS

REL03-BP03 Fornecer contratos de serviço por API

Os contratos de serviço são acordos documentados entre produtores e consumidores de API estabelecidos em uma definição de API legível por máquina. Uma estratégia de versionamento de contrato permite que os consumidores continuem usando a API existente e migrem suas aplicações para uma API mais recente quando estiverem prontos. A implantação do produtor pode acontecer a qualquer momento, desde que o contrato seja cumprido. A equipe de serviços pode usar a pilha de tecnologia de sua preferência para cumprir o contrato de API.

Resultado desejado: as aplicações criadas com arquiteturas orientadas a serviços ou de microsserviços podem operar de forma independente e, ao mesmo tempo, ter uma dependência de tempo de execução integrada. As alterações implantadas em um consumidor ou produtor de API não interrompem a estabilidade do sistema geral quando os dois lados seguem um contrato

de API comum. Os componentes que se comunicam por meio de APIs de serviço podem realizar lançamentos funcionais independentes, atualizações para dependências de runtime ou fazer failover em um site de recuperação de desastres (DR) com pouco ou nenhum impacto entre si. Além disso, serviços diferentes são capazes de escalar de forma independente a absorção da demanda de recursos sem exigir que outros serviços escalem simultaneamente.

Práticas comuns que devem ser evitadas:

- Criação de APIs de serviço sem esquemas altamente tipificados. Isso resulta em APIs que não podem ser usadas para gerar vinculações de API e payloads que não podem ser validadas de maneira programática.
- Não adotar uma estratégia de versionamento, o que força os consumidores de API a atualizarem e lançarem ou falharem com a evolução dos contratos de serviço.
- Mensagens de erro que vazam detalhes da implementação do serviço subjacente em vez de descreverem falhas de integração no contexto e no idioma do domínio.
- Não usar contratos de API para desenvolver casos de teste e simular implementações de API para permitir testes independentes dos componentes do serviço.

Benefícios de implementar esta prática recomendada: sistemas distribuídos compostos por componentes que se comunicam por meio de contratos de serviço de API podem aumentar a confiabilidade. Os desenvolvedores podem detectar possíveis problemas no início do processo de desenvolvimento com a verificação de tipo durante a compilação a fim de verificar se as solicitações e as respostas seguem o contrato da API e se os campos obrigatórios estão presentes. Os contratos de API oferecem uma interface clara de autodocumentação de APIs e oferecem melhor interoperabilidade entre diferentes sistemas e linguagens de programação.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Depois de identificar os domínios de negócios e determinar a segmentação da workload, você pode desenvolver suas APIs de serviço. Primeiro, defina contratos de serviço legíveis por máquina para APIs e, depois, implemente uma estratégia de versionamento de API. Quando estiver pronto para integrar serviços em protocolos comuns, como REST, GraphQL ou eventos assíncronos, você poderá incorporar serviços da AWS à sua arquitetura para integrar seus componentes com contratos de API altamente tipificados.

Serviços da AWS para contratos de API de serviços

Incorpore serviços da AWS, incluindo o Amazon API Gateway, o AWS AppSync e o Amazon EventBridge, em sua arquitetura para usar contratos de serviços de API em sua aplicação. O Amazon API Gateway ajuda você a se integrar diretamente com serviços da AWS nativos e outros serviços Web. O API Gateway é compatível com a especificação da OpenAPI e versionamento. O AWS AppSync é um endpoint gerenciado do GraphQL que você configura definindo um esquema do GraphQL para definir uma interface de serviço para consultas, mutações e assinaturas. O Amazon EventBridge usa esquemas de eventos para definir eventos e gerar vinculações de código para seus eventos.

Etapas de implementação

- Primeiro, defina um contrato para sua API. Um contrato expressará os recursos de uma API, bem como definirá objetos e campos de dados altamente tipificados para a entrada e a saída da API.
- Ao configurar APIs no API Gateway, você pode importar e exportar especificações da OpenAPI para seus endpoints.
  - Importar uma definição da OpenAPI simplifica a criação de sua API e pode ser integrada a ferramentas de infraestrutura como código da AWS, como o AWS Serverless Application Model e o AWS Cloud Development Kit (AWS CDK).
  - Exportar uma definição de API simplifica a integração a ferramentas de teste de API e oferece ao consumidor de serviços uma especificação de integração.
- Você pode definir e gerenciar as APIs do GraphQL com o AWS AppSync definindo um arquivo de esquema do GraphQL para gerar sua interface de contrato e simplificar a interação com modelos REST complexos, várias tabelas de banco de dados ou serviços legados.
- Os projetos do AWS Amplify integrados ao AWS AppSync geram arquivos de consulta JavaScript altamente tipificados para uso em sua aplicação, bem como uma biblioteca cliente do AWS AppSync GraphQL para tabelas do Amazon DynamoDB.
- Quando você consome eventos de serviço do Amazon EventBridge, eles seguem os esquemas já existentes no registro do esquema ou os definidos com a especificação da OpenAPI. Com um esquema definido no registro, também é possível gerar vinculações de cliente a partir do contrato de esquema para integrar seu código aos eventos.
- Estender ou realizar o versionamento de sua API. Estender uma API é uma opção mais simples ao adicionar campos que podem ser configurados com campos opcionais ou valores padrão para campos obrigatórios.
  - Contratos baseados em JSON para protocolos, como REST e GraphQL, podem ser uma boa opção para a extensão do contrato.

    - Contratos baseados em XML para protocolos, como SOAP, devem ser testados com consumidores de serviços para determinar a viabilidade da extensão do contrato.
- Ao realizar o versionamento de uma API, considere implementar o controle de versão por procuração em que uma fachada é usada para oferecer compatibilidade com versões para que a lógica possa ser mantida em uma única base de código.
    - Com o API Gateway, você pode usar mapeamentos de solicitação e resposta para simplificar a absorção de alterações no contrato estabelecendo uma fachada para fornecer valores padrão para novos campos ou para retirar os campos removidos de uma solicitação ou resposta. Com essa abordagem, o serviço subjacente pode manter uma única base de código.

Recursos

Práticas recomendadas relacionadas:

- REL03-BP01 Escolher como segmentar a workload
- REL03-BP02 Criar serviços voltados para domínios e funcionalidades de negócios específicos
- REL04-BP02 Implementar dependências com acoplamento fraco
- REL05-BP03 Controlar e limitar chamadas de novas tentativas
- REL05-BP05 Definir tempos limite do cliente

Documentos relacionados:

- O que é uma API (interface de programação de aplicações)?
- Implementar microsserviços na AWS
- Compensações de microsserviços
- Microsserviços: uma definição desse novo termo de arquitetura
- Microsserviços na AWS
- Trabalhar com extensões do API Gateway para OpenAPI
- Especificação da OpenAPI
- GraphQL: esquemas e tipos
- Vinculações de código do Amazon EventBridge

Exemplos relacionados:

- Amazon API Gateway: configurar uma API REST usando a OpenAPI
- Amazon API Gateway para a aplicação CRUD do Amazon DynamoDB usando a OpenAPI
- Padrões modernos de integração de aplicações em uma era sem servidor: integração do serviço do API Gateway
- Implementar o versionamento do API Gateway baseado em cabeçalho com o Amazon CloudFront
- AWS AppSync: como criar uma aplicação cliente

Vídeos relacionados:

- Usar OpenAPI na AWS SAM para gerenciar o API Gateway

Ferramentas relacionadas:

- Amazon API Gateway
- AWS AppSync
- Amazon EventBridge

REL 4. Como projetar interações em um sistema distribuído para evitar falhas?

Os sistemas distribuídos dependem de redes de comunicação para interconectar componentes, como servidores ou serviços. A workload deve operar de forma confiável, apesar da perda de dados ou da latência nessas redes. Os componentes do sistema distribuído devem operar de uma maneira que não afete negativamente outros componentes ou a workload. Essas práticas recomendadas evitam falhas e melhoram o tempo médio entre falhas (MTBF).

Práticas recomendadas

- REL04-BP01 Identificar qual tipo de sistema distribuído é necessário
- REL04-BP02 Implementar dependências com acoplamento fraco
- REL04-BP03 Fazer um trabalho constante
- REL04-BP04 Fazer com que todas as respostas sejam idempotentes

REL04-BP01 Identificar qual tipo de sistema distribuído é necessário

Os sistemas distribuídos podem ser síncronos, assíncronos ou em lote. Os sistemas síncronos devem processar solicitações o mais rápido possível e se comunicar uns com os outros fazendo

chamadas síncronas de solicitação e resposta usando protocolos HTTP/S, REST ou de chamada de procedimento remoto (RPC). Os sistemas assíncronos se comunicam uns com os outros trocando dados de forma assíncrona por meio de um serviço intermediário sem acoplar sistemas individuais. Os sistemas em lote recebem um grande volume de dados de entrada, executam processos de dados automatizados sem intervenção humana e geram dados de saída.

Resultado desejado: crie uma workload que interaja efetivamente com dependências síncronas, assíncronas e em lote.

Práticas comuns que devem ser evitadas:

- A workload espera indefinidamente por uma resposta de suas dependências, o que pode fazer com que os clientes da workload esgotem o tempo limite, sem saber se a solicitação foi recebida.
- A workload usa uma cadeia de sistemas dependentes que chamam um ao outro de forma síncrona. Para que toda a cadeia tenha êxito, isso exige primeiro que cada sistema esteja disponível e consiga processar uma solicitação, possivelmente fragilizando o comportamento e a disponibilidade geral.
- A workload comunica-se com as dependências de forma assíncrona e depende do conceito de entrega de mensagens garantida exatamente uma vez, quando muitas vezes ainda é possível receber mensagens duplicadas.
- A workload não usa ferramentas adequadas de agendamento em lote e permite a execução simultânea do mesmo trabalho em lotes.

Benefícios de implementar esta prática recomendada: é comum que uma determinada workload implemente um ou mais estilos de comunicação entre síncrono, assíncrono e em lote. Essa prática recomendada ajuda você a identificar as diferentes vantagens e desvantagens associadas a cada estilo de comunicação para tornar a workload capaz de tolerar interrupções em qualquer uma das dependências.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

As seções a seguir contêm diretrizes de implementação gerais e específicas de cada tipo de dependência.

Orientações gerais

- Certifique-se de que os objetivos de nível de serviço (SLOs) de performance e confiabilidade que suas dependências oferecem atendam aos requisitos de performance e confiabilidade da workload.
- Use serviços de observabilidade da AWS para monitorar os tempos de resposta e as taxas de erro para garantir que sua dependência esteja fornecendo serviços nos níveis necessários para sua workload.
- Identifique os possíveis desafios que a workload pode enfrentar ao se comunicar com as dependências. Os sistemas distribuídos apresentam uma ampla variedade de desafios que podem aumentar a complexidade arquitetônica, a carga operacional e o custo. Os desafios comuns são: latência, interrupções na rede, perda de dados, ajuste de escala e atraso na replicação de dados.
- Implemente gerenciamento e registro de erros robustos para obter ajuda para solucionar problemas quando sua dependência apresentar problemas.

Dependência síncrona

Nas comunicações síncronas, a workload envia uma solicitação para a dependência e bloqueia a operação à espera de uma resposta. Quando a dependência recebe a solicitação, ela tenta tratá-la o mais rápido possível e envia uma resposta de volta à workload. Um desafio significativo da comunicação síncrona é que ela causa o acoplamento temporal, o que exige que a workload e as respectivas dependências estejam disponíveis ao mesmo tempo. Quando a workload precisar se comunicar de forma síncrona com as dependências, pense na seguinte orientação:

- Sua workload não deve depender de várias dependências síncronas para realizar uma única função. Essa cadeia de dependências aumenta a fragilidade geral porque todas as dependências no caminho precisam estar disponíveis para que a solicitação seja concluída com êxito.
- Quando uma dependência não estiver íntegra ou estiver indisponível, determine suas estratégias de tratamento de erros e de novas tentativas. Evite usar comportamento bimodal. O comportamento bimodal ocorre quando a workload exibe um comportamento diferente nos modos normal e de falha. Para obter mais detalhes sobre o comportamento bimodal, consulte REL11-BP05 Usar estabilidade estática para evitar comportamento bimodal.
- Lembre-se que antecipar-se à falha é melhor do que fazer a workload esperar. Por exemplo, o Guia do desenvolvedor do AWS Lambda descreve como lidar com novas tentativas e falhas ao invocar funções do Lambda.
- Defina tempos limite quando a workload chamar sua dependência. Essa técnica evita esperas muito longas ou indefinidas por uma resposta. Para ver uma discussão útil sobre esse problema,

consulte Ajustar as configurações de solicitação HTTP do AWS SDK Java para aplicações do Amazon DynamoDB com reconhecimento de latência.

- Minimize o número de chamadas feitas da workload para a dependência para atender a uma única solicitação. Ter chamadas interativas entre elas aumenta o acoplamento e a latência.

Dependência assíncrona

Para dissociar temporariamente a workload de sua dependência, elas devem se comunicar de forma assíncrona. Usando uma abordagem assíncrona, a workload pode continuar com qualquer outro processamento sem precisar esperar que a dependência, ou cadeia de dependências, envie uma resposta.

Quando a workload precisar se comunicar de forma assíncrona com a dependência, pense na seguinte orientação:

- Determine se deseja usar mensagens ou streaming de eventos com base no caso de uso e requisitos. O sistema de mensagens permite que sua workload se comunique com sua dependência enviando e recebendo mensagens por meio de um agente de mensagens. O streaming de eventos permite que sua workload e sua dependência usem um serviço de streaming para publicar e assinar eventos, entregues como fluxos contínuos de dados, que precisam ser processados o mais rápido possível.

- O sistema de mensagens e o streaming de eventos gerenciam as mensagens de forma diferente, então é necessário tomar decisões sobre concessão com base em:
  - Prioridade da mensagem: os agentes de mensagens podem processar mensagens de alta prioridade antes das mensagens normais. No streaming de eventos, todas as mensagens têm a mesma prioridade.
  - Consumo de mensagens: os agentes de mensagens garantem que os consumidores recebam a mensagem. Os consumidores de streaming de eventos devem rastrear a última mensagem que leram.
  - Ordenação das mensagens: com o sistema de mensagens, não é garantido receber mensagens na ordem exata em que elas são enviadas, a menos que você use a abordagem FIFO (primeira a entrar, primeira a sair). O streaming de eventos sempre preserva a ordem na qual os dados foram produzidos.
  - Exclusão de mensagens: com o sistema de mensagens, o consumidor deve excluir a mensagem após processá-la. O serviço de streaming de eventos anexa a mensagem a um fluxo e

permanece lá até que o período de retenção da mensagem expire. Essa política de exclusão torna o streaming de eventos adequado para reproduzir mensagens.

- Defina como a workload sabe quando a dependência conclui o trabalho. Por exemplo, quando sua workload invoca uma função do Lambda de forma assíncrona, o Lambda coloca o evento em uma fila e retorna uma resposta informando êxito, sem informações adicionais. Após a conclusão do processamento, a função do Lambda pode enviar o resultado para um destino, configurável com base no sucesso ou na falha.
- Crie a workload para lidar com mensagens duplicadas utilizando a idempotência. Idempotência significa que os resultados da workload não mudam, mesmo que ela seja gerada mais de uma vez para a mesma mensagem. É importante ressaltar que os serviços de mensagens ou streaming reenviarão uma mensagem se ocorrer uma falha na rede ou se uma confirmação não for recebida.
- Se a workload não receber uma resposta da dependência, ela precisará reenviar a solicitação. Considere limitar o número de novas tentativas para preservar a CPU, a memória e os recursos de rede da workload para lidar com outras solicitações. A documentação do AWS Lambda mostra como lidar com erros de invocação assíncrona.
- Utilize as ferramentas adequadas de observabilidade, depuração e rastreamento para gerenciar e operar a comunicação assíncrona da workload com a dependência. É possível usar o Amazon CloudWatch para monitorar serviços de mensagens e streaming de eventos. Você também pode instrumentar sua workload com o AWS X-Ray para obter insights rapidamente para solucionar problemas.

Dependência de lote

Os sistemas em lote utilizam dados de entrada, iniciam uma série de trabalhos para processá-los e produzem alguns dados de saída, sem intervenção manual. Dependendo do tamanho dos dados, os trabalhos podem ser executados de minutos a, em alguns casos, vários dias. Quando a workload se comunica com a dependência em lote, pense na seguinte orientação:

- Defina a janela de tempo em que a workload deve executar o trabalho em lote. A workload pode configurar um padrão de recorrência para invocar um sistema em lote, por exemplo, a cada hora ou no final de cada mês.
- Determine a localização da entrada de dados e da saída de dados processados. Escolha um serviço de armazenamento, como o Amazon Simple Storage Services (Amazon S3), o Amazon Elastic File System (Amazon EFS) e o Amazon FSx para Lustre, que permita que sua workload leia e grave arquivos em grande escala.

- Se sua workload precisar invocar vários trabalhos em lote, você poderá usar o AWS Step Functions para simplificar a orquestração de trabalhos em lote executados na AWS ou on-premises. Este projeto de exemplo demonstra a orquestração de trabalhos em lote usando Step Functions, o AWS Batch e o Lambda.
- Monitore trabalhos em lote para procurar anormalidades, como um trabalho que leva mais tempo do que deveria para ser concluído. Você pode usar ferramentas como o CloudWatch Container Insights para monitorar ambientes e trabalhos em AWS Batch. Nesse caso, a workload impediria o início do próximo trabalho e informaria a equipe relevante sobre a exceção.

Recursos

Documentos relacionados:

- Operações da Nuvem AWS: monitoramento e observabilidade
- Amazon Builders' Library: desafios com sistemas distribuídos
- REL11-BP05 Usar estabilidade estática para evitar comportamento bimodal
- Guia do desenvolvedor do AWS Lambda: tratamento de erros e novas tentativas automáticas no AWS Lambda
- Ajustar as configurações de solicitação HTTP do AWS SDK Java para aplicações do Amazon DynamoDB com reconhecimento de latência
- Sistema de mensagens da AWS
- O que é streaming de dados?
- Guia do desenvolvedor do AWS Lambda: invocação assíncrona
- Perguntas frequentes do Amazon Simple Queue Service: filas FIFO
- Guia do desenvolvedor do Amazon Kinesis Data Streams: tratar registros duplicados
- Guia do desenvolvedor do Amazon Simple Queue Service: métricas do CloudWatch disponíveis para Amazon SQS
- Guia do desenvolvedor do Amazon Kinesis Data Streams: monitorar o serviço Amazon Kinesis Data Streams com o Amazon CloudWatch
- Guia do desenvolvedor do AWS X-Ray: conceitos do AWS X-Ray
- Exemplos da AWS no GitHub: AWS Step functions Complex Orchestrator App
- Guia do usuário do AWS Batch: AWS Batch CloudWatch Container Insights

Vídeos relacionados:

- AWS Summit SF 2022: Observabilidade full-stack e monitoramento de aplicações com a AWS (COP310)

Ferramentas relacionadas:

- Amazon CloudWatch
- Amazon CloudWatch Logs
- AWS X-Ray
- Amazon Simple Storage Service (Amazon S3)
- Amazon Elastic File System (Amazon EFS)
- Amazon FSx para Lustre
- AWS Step Functions
- AWS Batch

REL04-BP02 Implementar dependências com acoplamento fraco

As dependências, como sistemas de enfileiramento, sistemas de streaming, fluxos de trabalho e balanceadores de carga, têm acoplamento fraco. O acoplamento fraco ajuda a isolar o comportamento de um componente de outros componentes que dependem dele, aumentando a resiliência e a agilidade.

Dependências de desacoplamento, como sistemas de filas, sistemas de streaming e fluxos de trabalho, ajudam a minimizar o impacto de alterações ou falhas em um sistema. Essa separação impede o comportamento de um componente de afetar outros que dependem dele, melhorando a resiliência e a agilidade.

Em sistemas fortemente acoplados, alterações em um componente podem exigir mudanças em outros componentes que dependem dele, o que resulta em performance degradada em todos eles. O acoplamento fraco interrompe essa dependência para que os componentes dependentes só precisem saber a interface versionada e publicada. A implementação de um acoplamento fraco entre dependências isola uma falha em uma dependência para não afetar a outra.

O acoplamento fraco permite modificar o código ou adicionar recursos a um componente, minimizando o risco para outros componentes que dependem dele. Ele também permite resiliência granular em nível de componente, caso em que é possível aumentar a escala horizontalmente ou até mesmo alterar a implementação subjacente da dependência.

Para melhorar ainda mais a resiliência por meio do acoplamento fraco, torne as interações de componentes assíncronas sempre que possível. Esse modelo é adequado para qualquer interação que não precise de uma resposta imediata e em que uma confirmação de que uma solicitação foi registrada será suficiente. Envolve um componente que gera eventos e outro que os consome. Os dois componentes não se integram por meio de interação direta ponto a ponto, mas geralmente por meio de uma camada de armazenamento durável intermediária, como uma fila do Amazon SQS, uma plataforma de dados de streaming, como o Amazon Kinesis, ou o AWS Step Functions.

Figura 4: Dependências como sistemas de enfileiramento e balanceadores de carga têm acoplamento fraco

As filas do Amazon SQS e os AWS Step Functions são apenas duas maneiras de adicionar uma camada intermediária para acoplamento fraco. As arquiteturas orientadas a eventos também podem ser criadas na Nuvem AWS com o Amazon EventBridge, que pode abstrair clientes (produtores de eventos) dos serviços dos quais eles dependem (consumidores de eventos). O Amazon Simple Notification Service (Amazon SNS) é uma solução eficaz quando você precisa de mensagens de alto throughput, baseadas em push e muitos para muitos. Usando tópicos do Amazon SNS, seus sistemas de publicadores podem enviar mensagens para um grande número de endpoints assinantes para processamento paralelo.

Embora as filas ofereçam várias vantagens, na maioria dos sistemas complexos em tempo real, as solicitações mais antigas do que um tempo limite (geralmente segundos) devem ser consideradas obsoletas (o cliente desistiu e não está mais esperando por uma resposta) e não devem ser processadas. Dessa forma, as solicitações mais recentes (e provavelmente ainda válidas) podem ser processadas.

Resultado desejado: a implementação de dependências com acoplamento fraco permite minimizar a área de superfície de falha em um nível de componente, o que ajuda a diagnosticar e resolver problemas. Ela também simplifica os ciclos de desenvolvimento, permitindo que as equipes implementem mudanças em um nível modular sem impactar a performance de outros componentes que dependem delas. Essa abordagem fornece a capacidade de aumentar a escala horizontalmente em nível de componente com base nas necessidades dos recursos, bem como na utilização de um componente que contribui para a redução de custos.

Práticas comuns que devem ser evitadas:

- Implantar uma workload monolítica.

- Invocar diretamente as APIs entre níveis de workload sem recurso de failover ou processamento assíncrono da solicitação.
- Acoplamento forte usando dados compartilhados. Sistemas com acoplamento fraco devem evitar o compartilhamento de dados por meio de bancos de dados compartilhados ou outras formas de armazenamento de dados com acoplamento forte, o que pode reintroduzir o acoplamento forte e impedir a escalabilidade.
- Ignorar a pressão contrária. A workload deve ter a capacidade de diminuir ou interromper a entrada de dados quando um componente não puder processá-los na mesma velocidade.

Benefícios de implementar esta prática recomendada: o acoplamento fraco ajuda a isolar o comportamento de um componente de outros componentes que dependem dele, aumentando a resiliência e a agilidade. Uma falha em um componente é isolada dos demais.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Implemente dependências com acoplamento fraco. Existem várias soluções que permitem criar aplicações com acoplamento fraco. Isso inclui serviços para implementar filas totalmente gerenciadas, fluxos de trabalho automatizados, reação a eventos e APIs, entre outros, que podem ajudar a isolar o comportamento de componentes de outros componentes e, dessa forma, aumentar a resiliência e a agilidade.

- Crie arquiteturas orientadas a eventos: o Amazon EventBridge ajuda você a criar arquiteturas orientadas a eventos distribuídas e com acoplamento fraco.
- Implemente filas em sistemas distribuídos: é possível usar o Amazon Simple Queue Service (Amazon SQS) para integrar e desacoplar sistemas distribuídos.
- Containerize componentes na forma de microsserviços: os microsserviços permitem que as equipes criem aplicações formadas por pequenos componentes independentes que se comunicam por meio de APIs bem definidas. O Amazon Elastic Container Service (Amazon ECS) e o Amazon Elastic Kubernetes Service (Amazon EKS) podem ajudar você a começar a usar contêineres mais rápido.
- Gerencie fluxos de trabalho com Step Functions: o Step Functions ajuda você a coordenar vários serviços da AWS em fluxos de trabalho flexíveis.
- Utilize as arquiteturas de mensagens publicador-assinante (pub/sub): o Amazon Simple Notification Service (Amazon SNS) fornece entrega de mensagens de publicadores para assinantes (também conhecidos como produtores e consumidores).

Etapas de implementação

- Os componentes em uma arquitetura orientada a eventos são iniciados por eventos. Eventos são ações que ocorrem em um sistema, como um usuário que adiciona um item a um carrinho. Quando uma ação é bem-sucedida, um evento que aciona o próximo componente do sistema é gerado.
  - Criar aplicações orientadas por eventos com o Amazon EventBridge
  - AWS re:Invent 2022: Desenvolver integrações orientadas por eventos com o Amazon EventBridge
- Os sistemas de mensagens distribuídos têm três partes principais que precisam ser implementadas para uma arquitetura baseada em fila. Eles incluem componentes do sistema distribuído, a fila usada para desacoplamento (distribuída em servidores do Amazon SQS) e as mensagens na fila. Um sistema típico tem produtores que iniciam a mensagem na fila e o consumidor que recebe a mensagem da fila. A fila armazena as mensagens em vários servidores do Amazon SQS para fins de redundância.
  - Arquitetura básica do Amazon SQS
  - Envie mensagens entre aplicações distribuídas com o Amazon Simple Queue Service
- Os microsserviços, quando bem utilizados, melhoram a capacidade de manutenção e aumentam a escalabilidade, pois os componentes com acoplamento fraco são gerenciados por equipes independentes. Isso também permite o isolamento de comportamentos em um único componente em caso de alterações.
  - Implementar microsserviços na AWS
  - Vamos arquitetar! Arquitetar microsserviços com contêineres
- Com o AWS Step Functions é possível criar aplicações distribuídas, automatizar processos, orquestrar microsserviços, entre outras coisas. A orquestração de vários componentes em um fluxo de trabalho automatizado permite desacoplar as dependências na aplicação.
  - Criar um fluxo de trabalho sem servidor com o AWS Step Functions e o AWS Lambda
  - Conceitos básicos do AWS Step Functions

Recursos

Documentos relacionados:

- Amazon EC2: garantia da idempotência
- Amazon Builders' Library: desafios com sistemas distribuídos

- Amazon Builders' Library: confiabilidade, trabalho constante e uma boa xícara de café
- O que é o Amazon EventBridge?
- O que é o Amazon Simple Queue Service?
- Romper com seu monólito
- Organizar microsserviços baseados em filas com o AWS Step Functions e o Amazon SQS
- Arquitetura básica do Amazon SQS
- Arquitetura baseada em fila

Vídeos relacionados:

- AWS New York Summit 2019: Introdução a arquiteturas orientadas por eventos e ao Amazon EventBridge (MAD205)
- AWS re:Invent 2018: Fechar loops e abrir mentes: como assumir o controle de sistemas grandes e pequenos ARC337 (inclui acoplamento fraco, trabalho constante, estabilidade estática)
- AWS re:Invent 2019: Migrar para arquiteturas orientadas por eventos (SVS308)
- AWS re:Invent 2019: Aplicações sem servidor orientadas por eventos e escaláveis usando o Amazon SQS e o Lambda
- AWS re:Invent 2022: Desenvolver integrações orientadas por eventos com o Amazon EventBridge
- AWS re:Invent 2017: Mergulho profundo e práticas recomendadas do Elastic Load Balancing

REL04-BP03 Fazer um trabalho constante

Os sistemas podem falhar quando há alterações grandes e rápidas na carga. Por exemplo, se a sua workload está realizando uma verificação de integridade que monitora a integridade de milhares de servidores, ela deve sempre enviar a carga útil com o mesmo tamanho (um snapshot completo do estado atual). Independentemente de nenhum servidor falhar ou todos eles, o sistema de verificação de integridade está realizando um trabalho constante sem alterações grandes e rápidas.

Por exemplo, se o sistema de verificação de integridade estiver monitorando 100 mil servidores, a carga nele será nominal a uma taxa de falha do servidor normalmente leve. No entanto, se um evento importante deixar metade desses servidores com problemas de integridade, o sistema de verificação de integridade ficará sobrecarregado tentando atualizar os sistemas de notificação e comunicar o estado com seus clientes. Portanto, em vez disso, o sistema de verificação de integridade deve enviar o snapshot completo do estado atual a cada vez. 100.000 estados de integridade do servidor, cada um representado por um bit, seriam apenas uma carga útil de 12,5

KB. Independentemente de nenhum servidor ou falhar, ou se todos eles falharem, o sistema de verificação de integridade está realizando um trabalho constante, e alterações grandes e rápidas não são uma ameaça para a estabilidade do sistema. Na verdade, é assim que o Amazon Route 53 lida com verificações de integridade de endpoints (como endereços IP) para determinar como os usuários finais são roteados para eles.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

- Faça um trabalho constante para que os sistemas não falhem quando houver mudanças rápidas e grandes na carga.
- Implemente dependências com acoplamento fraco. As dependências, como sistemas de enfileiramento, sistemas de streaming, fluxos de trabalho e balanceadores de carga, têm acoplamento fraco. O acoplamento fraco ajuda a isolar o comportamento de um componente de outros componentes que dependem dele, aumentando a resiliência e a agilidade.
  - Amazon Builders' Library: confiabilidade, trabalho constante e uma boa xícara de café
  - AWS re:Invent 2018: Fechar loops e abrir mentes: como assumir o controle de sistemas grandes e pequenos ARC337 (inclui trabalho constante)
    - Para o exemplo de um sistema de verificação de integridade monitorando 100.000 servidores, crie workloads para que os tamanhos das cargas permaneçam constantes, independentemente do número de sucessos ou falhas.

Recursos

Documentos relacionados:

- Amazon EC2: garantia da idempotência
- Amazon Builders' Library: desafios com sistemas distribuídos
- Amazon Builders' Library: confiabilidade, trabalho constante e uma boa xícara de café

Vídeos relacionados:

- AWS New York Summit 2019: Introdução a arquiteturas orientadas por eventos e ao Amazon EventBridge (MAD205)
- AWS re:Invent 2018: Fechar loops e abrir mentes: como assumir o controle de sistemas grandes e pequenos ARC337 (inclui trabalho constante)

- AWS re:Invent 2018: Fechar loops e abrir mentes: como assumir o controle de sistemas grandes e pequenos ARC337 (inclui acoplamento fraco, trabalho constante, estabilidade estática)
- AWS re:Invent 2019: Migrar para arquiteturas orientadas por eventos (SVS308)

REL04-BP04 Fazer com que todas as respostas sejam idempotentes

Um serviço idempotente garante que cada solicitação seja concluída exatamente uma vez, de modo que fazer várias solicitações idênticas causa o mesmo efeito que uma única solicitação. Um serviço idempotente facilita para um cliente implementar novas tentativas sem o receio de que uma solicitação seja processada erroneamente várias vezes. Para fazer isso, os clientes podem emitir solicitações de API com um token de idempotência. O mesmo token é usado sempre que a solicitação é repetida. Uma API de serviço idempotente usa o token para retornar uma resposta idêntica à resposta que foi retornada na primeira vez que a solicitação foi concluída.

Em um sistema distribuído, é fácil executar uma ação no máximo uma vez (o cliente faz apenas uma solicitação) ou pelo menos uma vez (continue solicitando até o cliente receber a confirmação do sucesso). Mas é difícil garantir que uma ação seja idempotente, o que significa que ela é executada exatamente uma vez, de modo que fazer várias solicitações idênticas tem o mesmo efeito que fazer uma única solicitação. Usando tokens de idempotência em APIs, os serviços podem receber uma solicitação mutante uma vez ou mais sem a criação de registros duplicados nem efeitos colaterais.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

- Faça com que todas as respostas sejam idempotentes. Um serviço idempotente garante que cada solicitação seja concluída exatamente uma vez, de modo que fazer várias solicitações idênticas causa o mesmo efeito que uma única solicitação.
  - Os clientes podem emitir solicitações de API com um token de idempotência; o mesmo token é usado sempre que a solicitação é repetida. Uma API de serviço idempotente usa o token para retornar uma resposta idêntica à resposta que foi retornada na primeira vez que a solicitação foi concluída.
    - Amazon EC2: garantia da idempotência

Recursos

Documentos relacionados:

- Amazon EC2: garantia da idempotência
- Amazon Builders' Library: desafios com sistemas distribuídos
- Amazon Builders' Library: confiabilidade, trabalho constante e uma boa xícara de café

Vídeos relacionados:

- AWS New York Summit 2019: Introdução a arquiteturas orientadas por eventos e ao Amazon EventBridge (MAD205)
- AWS re:Invent 2018: Fechar loops e abrir mentes: como assumir o controle de sistemas grandes e pequenos ARC337 (inclui acoplamento fraco, trabalho constante, estabilidade estática)
- AWS re:Invent 2019: Migrar para arquiteturas orientadas por eventos (SVS308)

REL 5. Como você projeta interações em um sistema distribuído para mitigar ou resistir a falhas?

Os sistemas distribuídos dependem de redes de comunicação para interconectar componentes (como servidores ou serviços). Sua workload deve operar de forma confiável, apesar da perda de dados ou da latência nessas redes. Os componentes do sistema distribuído devem operar de uma maneira que não afete negativamente outros componentes ou a workload. Essas práticas recomendadas permitem que as workloads resistam a tensões ou falhas, recuperem-se mais rapidamente delas e reduzam o impacto de tais prejuízos. Como resultado, o tempo médio para recuperação (MTTR) é melhorado.

Práticas recomendadas

- REL05-BP01 Implementar uma degradação normal para transformar dependências rígidas aplicáveis em dependências flexíveis
- REL05-BP02 Controlar a utilização de solicitações
- REL05-BP03 Controlar e limitar chamadas de novas tentativas
- REL05-BP04 Antecipar-se à falha e limitar filas
- REL05-BP05 Definir tempos limite do cliente
- REL05-BP06 Criar serviços sem estado sempre que possível
- REL05-BP07 Implementar medidas emergenciais

REL05-BP01 Implementar uma degradação normal para transformar dependências rígidas aplicáveis em dependências flexíveis

Os componentes da aplicação devem continuar desempenhando sua função principal mesmo que as dependências se tornem indisponíveis. Eles podem estar fornecendo dados um pouco obsoletos, dados alternativos ou até mesmo nenhum dado. Isso garante que o funcionamento geral do sistema seja minimamente impedido por falhas localizadas e, ao mesmo tempo, ofereça o valor empresarial central.

Resultado desejado: quando as dependências de um componente não estão íntegras, o próprio componente ainda pode funcionar, embora de maneira prejudicada. Os modos de falha dos componentes devem ser vistos como operação normal. Os fluxos de trabalho devem ser projetados de forma que essas falhas não ocasionem à falha total ou, pelo menos, a estados previsíveis e recuperáveis.

Práticas comuns que devem ser evitadas:

- Não identificar a principal funcionalidade empresarial necessária. Não testar se os componentes estão funcionando mesmo durante falhas de dependência.
- Não fornecer dados sobre erros ou quando apenas uma das várias dependências não está disponível e resultados parciais ainda podem ser retornados.
- Criar um estado inconsistente quando uma transação falha parcialmente.
- Não ter uma forma alternativa de acessar um armazenamento de parâmetros central.
- Invalidar ou esvaziar o estado local como resultado de uma falha na atualização sem levar em conta as consequências de fazer isso.

Benefícios de implementar esta prática recomendada: a degradação gradual melhora a disponibilidade do sistema como um todo e mantém as funções mais importantes em execução mesmo durante falhas.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

A implementação de uma degradação gradual ajuda a minimizar o impacto das falhas de dependência na função do componente. Preferencialmente, um componente detecta falhas de dependência e as contorna de uma maneira que afeta minimamente outros componentes ou clientes.

Arquitetar para uma degradação gradual significa considerar possíveis modos de falha durante o projeto de dependência. Para cada modo de falha, tenha uma maneira de fornecer a maior parte ou pelo menos a funcionalidade mais crítica do componente para chamadores ou clientes. Essas considerações podem se tornar requisitos adicionais que podem ser testados e verificados. Preferencialmente, um componente é capaz de realizar sua função principal de maneira aceitável, mesmo quando uma ou várias dependências falham.

Trata-se tanto de uma discussão empresarial quanto técnica. Todos os requisitos comerciais são importantes e devem ser atendidos, se possível. No entanto, ainda faz sentido perguntar o que deve acontecer quando nem todos eles podem ser cumpridos. Um sistema pode ser projetado para estar disponível e ser consistente, mas em circunstâncias em que um requisito deve ser descartado, qual deles é mais importante? Para o processamento de pagamentos, pode ser a consistência. Para uma aplicação em tempo real, pode ser a disponibilidade. Para um site voltado para o cliente, a resposta pode depender das expectativas do cliente.

O que isso significa depende dos requisitos do componente e do que deve ser considerado sua função principal. Por exemplo:

- Um site de comércio eletrônico pode exibir dados de vários sistemas diferentes, como recomendações personalizadas, produtos mais bem classificados e status dos pedidos dos clientes na página de pouso. Quando um sistema upstream falha, ainda faz sentido exibir todo o resto em vez de mostrar uma página de erro para um cliente.
- Um componente que executa gravações em lote ainda poderá continuar processando um lote se ocorrer uma falha em uma das operações individuais. Deve ser simples implementar um mecanismo de novas tentativas. Isso pode ser feito retornando informações sobre quais operações foram bem-sucedidas, quais falharam e por que falharam para o chamador, ou colocando solicitações com falha em uma fila de mensagens não entregues para implementar novas tentativas assíncronas. As informações sobre operações com falha também devem ser registradas em log.
- Um sistema que processa transações deve verificar se todas ou nenhuma atualização individual foi executada. Para transações distribuídas, o padrão saga pode ser usado para reverter operações anteriores caso ocorra uma falha em uma operação posterior da mesma transação. Aqui, a função principal é manter a consistência.
- Sistemas essenciais devem ser capazes de lidar com dependências não correspondentes em tempo hábil. Nesses casos, o padrão de disjuntor pode ser usado. Quando as respostas de uma dependência começam a atingir o tempo limite, o sistema pode mudar para um estado fechado em que nenhuma chamada adicional é realizada.

- Uma aplicação pode ler parâmetros de um armazenamento de parâmetros. Pode ser útil criar imagens de contêiner com um conjunto padrão de parâmetros e usá-las caso o armazenamento de parâmetros não esteja disponível.

Observe que as vias percorridas em caso de falha do componente precisam ser testadas e devem ser significativamente mais simples do que a via principal. Em geral, estratégias de fallback devem ser evitadas.

Etapas de implementação

Identifique dependências externas e internas. Leve em conta quais tipos de falhas podem ocorrer nelas. Pense em maneiras de minimizar o impacto negativo nos sistemas upstream e downstream e nos clientes durante essas falhas.

Veja a seguir uma lista de dependências e como degradar normalmente quando elas falham:

1. Falha parcial das dependências: um componente pode fazer várias solicitações para sistemas downstream, como várias solicitações para um sistema ou uma solicitação para vários sistemas cada. Dependendo do contexto empresarial, diferentes maneiras de lidar com isso podem ser apropriadas (para obter mais detalhes, consulte exemplos anteriores em Orientações de implementação).
2. Um sistema downstream não consegue processar solicitações devido à alta carga: se as solicitações para um sistema downstream falharem constantemente, não fará sentido continuar tentando novamente. Isso pode criar carga adicional em um sistema já sobrecarregado e dificultar a recuperação. O padrão de disjuntor pode ser utilizado aqui, o qual monitora as chamadas com falha para um sistema downstream. Se ocorrer uma falha em um grande número de chamadas, ele deixará de enviar mais solicitações para o sistema downstream e só ocasionalmente permitirá que as chamadas passem para testar se o sistema downstream está disponível novamente.
3. Uma loja de parâmetros não está disponível: para transformar um armazenamento de parâmetros, é possível usar o armazenamento em cache flexível de dependências ou padrões razoáveis incluídos nas imagens do contêiner ou da máquina. Observe que esses padrões precisam ser mantidos atualizados e incluídos nos pacotes de testes.
4. Um serviço de monitoramento ou outra dependência não funcional não está disponível: se um componente não conseguir enviar logs, métricas ou rastreamentos de forma intermitente para um serviço de monitoramento central, geralmente é melhor continuar executando as funções empresariais normalmente. Não registrar em log nem enviar métricas silenciosamente por um

longo período geralmente não é aceitável. Além disso, alguns casos de uso podem exigir entradas de auditoria completas para atender aos requisitos de conformidade.

5. Uma instância primária de um banco de dados relacional pode estar indisponível: o Amazon Relational Database Service, como quase todos os bancos de dados relacionais, só pode ter uma instância de gravador principal. Isso cria um único ponto de falha para workloads de gravação e dificulta o ajuste de escala. Isso pode ser parcialmente reduzido com o uso de uma configuração Multi-AZ para alta disponibilidade ou do Amazon Aurora Sem Servidor para melhor ajuste de escala. Para requisitos de disponibilidade muito altos, pode fazer sentido não confiar no gravador principal. Para consultas que são somente leitura, é possível usar réplicas de leitura que fornecem redundância e a capacidade de aumentar a escala horizontalmente, e não apenas verticalmente. As gravações podem ser armazenadas em buffer, por exemplo, em uma fila do Amazon Simple Queue Service, para que as solicitações de gravação dos clientes ainda possam ser aceitas mesmo que a principal esteja temporariamente indisponível.

Recursos

Documentos relacionados:

- Amazon API Gateway: controlar as solicitações de API para um melhor throughput
- CircuitBreaker (resume "Disjuntor" do livro "Release It!")
- Novas tentativas em caso de erro e recuo exponencial na AWS
- Michael Nygard "Release It! Design and Deploy Production-Ready Software"
- Amazon Builders' Library: evitar fallback em sistemas distribuídos
- Amazon Builders' Library: evitar backlogs de fila insuperáveis
- Amazon Builders' Library: desafios e estratégias de armazenamento em cache
- Amazon Builders' Library: tempos limite, novas tentativas e recuo com jitter

Vídeos relacionados:

- Novas tentativas, recuo e jitter: AWS re:Invent 2019: Introdução à Amazon Builders' Library (DOP328)

Exemplos relacionados:

- Laboratório do Well-Architected: Nível 300: implementar verificações de integridade e gerenciar dependências para melhorar a confiabilidade

REL05-BP02 Controlar a utilização de solicitações

Controle a utilização das solicitações para reduzir o esgotamento de recursos devido a aumentos inesperados na demanda. Solicitações abaixo das taxas de controle de utilização são processadas, enquanto aquelas acima do limite definido são rejeitadas com uma mensagem de retorno indicando que o uso da solicitação foi controlado.

Resultado desejado: grandes picos de volume, sejam causados por aumentos repentinos de tráfego de clientes, ataques de inundação ou tempestades de novas tentativas, são reduzidos pelo controle de utilização de solicitações, permitindo que as workloads continuem com o processamento normal do volume de solicitações compatível.

Práticas comuns que devem ser evitadas:

- Os controles de utilização de endpoint da API não são implementados ou são mantidos em valores padrão sem considerar os volumes esperados.
- Não há teste de carregamento nem limites de controle de utilização para os endpoints da API.
- Controlar a utilização de taxas de solicitações sem considerar o tamanho ou a complexidade da solicitação.
- Testar as taxas máximas de solicitação ou o tamanho máximo da solicitação, mas não testar os dois juntos.
- Os recursos não são provisionados nos mesmos limites estabelecidos nos testes.
- Os planos de uso não foram configurados nem considerados para consumidores de API de aplicação para aplicação (A2A).
- Os consumidores da fila que escalam horizontalmente não têm as configurações máximas de simultaneidade configuradas.
- A limitação de taxas por endereço IP não foi implementada.

Benefícios de implementar esta prática recomendada: as workloads que definem limites de controle de utilização podem operar normalmente e processar a carga de solicitações aceitas com êxito em picos de volume inesperados. Os picos repentinos ou contínuos de solicitações para APIs e filas têm controle de utilização e não esgotam os recursos de processamento de solicitações. Os limites de

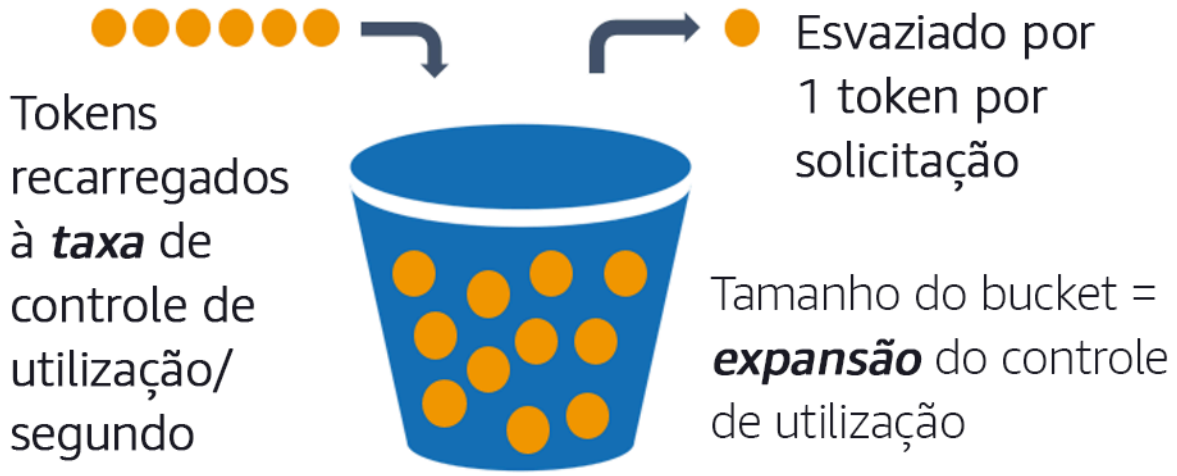
taxas controlam a utilização de solicitantes individuais para que grandes volumes de tráfego de um único endereço IP ou consumidor de API não esgotem os recursos e afetem outros consumidores.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Os serviços devem ser projetados para processar uma capacidade conhecida de solicitações; essa capacidade pode ser estabelecida por meio de testes de carga. Se as taxas de chegada de solicitações excederem os limites, a resposta apropriada sinalizará que uma solicitação teve controle de utilização. Isso permite que o consumidor resolva o erro e tente novamente mais tarde.

Quando seu serviço exigir uma implementação de controle de utilização, considere implementar o algoritmo de bucket de token, em que um token é contabilizado para uma solicitação. Os tokens são recarregados a uma taxa de controle de utilização por segundo e esvaziados de forma assíncrona por meio de um token por solicitação.



O algoritmo do bucket de token.

O Amazon API Gateway implementa o algoritmo do bucket de token de acordo com os limites da conta e da região e pode ser configurado por cliente com planos de uso. Além disso, o Amazon Simple Queue Service (Amazon SQS) e o Amazon Kinesis podem armazenar solicitações em buffer para suavizar a taxa de solicitações e permitir taxas de limitação mais altas para solicitações que podem ser atendidas. Por fim, é possível implementar a limitação de taxa com o AWS WAF para limitar consumidores de API específicos que geram uma carga excepcionalmente alta.

Etapas de implementação

É possível configurar o API Gateway com limites de limitação para suas APIs e retornar erros `429 Too Many Requests` quando os limites são excedidos. É possível usar o AWS WAF com seus endpoints do AWS AppSync e do APÌ Gateway para habilitar o limite de taxa por endereço IP. Além disso, se seu sistema tolerar o processamento assíncrono, será possível colocar mensagens em uma fila ou em um fluxo para acelerar as respostas aos clientes do serviço, o que permite que você atinja taxas de controle de utilização mais altas.

Com o processamento assíncrono, ao configurar o Amazon SQS como fonte de eventos para o AWS Lambda, você pode configurar a simultaneidade máxima para evitar que altas taxas de eventos consumam a cota de execução simultânea disponível da conta necessária para outros serviços em sua workload ou conta.

Embora o API Gateway ofereça uma implementação gerenciada do bucket de token, em casos em que não é possível usar o API Gateway, é possível utilizar as implementações de código aberto específicas da linguagem (veja exemplos relacionados em Recursos) do bucket de token para seus serviços.

- Entenda e configure os limites de controle de utilização do API Gateway no nível da conta por região, API por estágio e chave de API por nível do plano de uso.
- Aplique regras de limitação de taxa do AWS WAF ao API Gateway e aos endpoints do AWS AppSync para se proteger contra inundações e bloquear IPs maliciosos. As regras de controle de utilização de taxas também podem ser configuradas em chaves de API do AWS AppSync para consumidores A2A.
- Decida se você precisa de mais controle de limitação do que limitação de taxas para APIs do AWS AppSync e, em caso afirmativo, configure um API Gateway na frente do seu endpoint do AWS AppSync.
- Quando as filas do Amazon SQS são configuradas como acionadores para consumidores de filas do Lambda, defina a simultaneidade máxima para um valor que processe o suficiente para atender aos seus objetivos de nível de serviço, mas não consuma limites de simultaneidade que afetem outras funções do Lambda. Considere definir a simultaneidade reservada em outras funções do Lambda na mesma conta e região ao consumir filas com o Lambda.
- Use o API Gateway com integrações de serviços nativos ao Amazon SQS ou Kinesis para armazenar solicitações em buffer.

- Se você não puder usar o API Gateway, consulte bibliotecas específicas de linguagens para implementar o algoritmo do bucket de token para sua workload. Confira a seção de exemplos e faça sua própria pesquisa para encontrar uma biblioteca adequada.
- Teste os limites que você planeja definir ou permitir que sejam aumentados e documente os limites testados.
- Não aumente os limites além do que foi estabelecido nos testes. Ao aumentar um limite, verifique se os recursos provisionados já são equivalentes ou maiores do que os dos cenários de teste antes de aplicar o aumento.

Recursos

Práticas recomendadas relacionadas:

- REL04-BP03 Fazer um trabalho constante
- REL05-BP03 Controlar e limitar chamadas de novas tentativas

Documentos relacionados:

- Amazon API Gateway: controlar as solicitações de API para um melhor throughput
- AWS WAF: declaração de regra baseada em intervalos
- Introduzir simultaneidade máxima do AWS Lambda ao usar o Amazon SQS como fonte de eventos
- AWS Lambda: simultaneidade máxima

Exemplos relacionados:

- As três regras mais importantes baseadas em taxas do AWS WAF
- Java Bucket4j
- Bucket de tokens do Python
- Bucket de tokens do Node
- Limitação da taxa de segmentação do .NET System

Vídeos relacionados:

- Implementar as práticas recomendadas de segurança da API GraphQL com o AWS AppSync

Ferramentas relacionadas:

- Amazon API Gateway
- AWS AppSync
- Amazon SQS
- Amazon Kinesis
- AWS WAF

REL05-BP03 Controlar e limitar chamadas de novas tentativas

Use o recuo exponencial para tentar as solicitações novamente em intervalos progressivamente maiores entre cada nova tentativa. Introduza jitter entre as novas tentativas para tornar os intervalos de repetição aleatórios. Limite o número máximo de novas tentativas.

Resultado desejado: os componentes típicos em um sistema de software distribuído incluem servidores, balanceadores de carga, bancos de dados e servidores DNS. Durante a operação normal, esses componentes podem responder a solicitações com erros temporários ou limitados, além de erros que seriam persistentes, independentemente de repetições. Quando os clientes fazem solicitações aos serviços, elas consomem recursos, incluindo memória, threads, conexões, portas ou quaisquer outros recursos limitados. Controlar e limitar as repetições é uma estratégia para liberar e minimizar o consumo de recursos para que os componentes do sistema sob pressão não fiquem sobrecarregados.

Quando as solicitações do cliente atingem o tempo limite ou recebem respostas de erro, ele deve determinar se deve ou não tentar novamente. Se tentar novamente, ele o fará com um recuo exponencial com jitter e um valor máximo de nova tentativa. Como resultado, os serviços e os processos de backend recebem alívio da carga e do tempo de recuperação automática, ocasionando uma recuperação mais rápida e atendimento bem-sucedido das solicitações.

Práticas comuns que devem ser evitadas:

- Implementar novas tentativas sem adicionar recuo exponencial, jitter e valores máximos de novas tentativas. O recuo e o jitter ajudam a evitar picos artificiais de tráfego devido a novas tentativas coordenadas involuntariamente em intervalos comuns.
- Implementar novas tentativas sem testar seus efeitos ou presumir que as novas tentativas já estejam incorporadas a um SDK sem testar cenários de repetição.

- Não entender os códigos de erro publicados das dependências, ocasionando novas tentativas de todos os erros, inclusive aqueles com uma causa clara que indica falta de permissão, erro de configuração ou outra condição que, previsivelmente, não será resolvida sem intervenção manual.
- Não abordar práticas de observabilidade, incluindo monitoramento e alertas sobre falhas repetidas de serviço para que os problemas subjacentes sejam divulgados e possam ser resolvidos.
- Desenvolver mecanismos de novas tentativas personalizados quando os recursos de novas tentativas integrados ou de terceiros são suficientes.
- Tentar novamente em várias camadas da pilha de aplicações de uma forma que agrava as novas tentativas, consumindo ainda mais recursos em uma tempestade de repetições. Entenda como esses erros afetam sua aplicação, as dependências nas quais você confia e implemente novas tentativas em apenas um nível.
- Tentar novamente chamadas de serviço que não são idempotentes, causando efeitos colaterais inesperados, como resultados duplicados.

Benefícios de implementar esta prática recomendada: as novas tentativas ajudam os clientes a obter os resultados desejados quando as solicitações falham, mas também consomem mais tempo do servidor para obter as respostas bem-sucedidas que eles desejam. Quando as falhas são raras ou transitórias, as novas tentativas funcionam bem. Quando as falhas são causadas pela sobrecarga de recursos, as novas tentativas podem piorar as coisas. Adicionar um recuo exponencial com jitter às novas tentativas do cliente permite que os servidores se recuperem quando as falhas são causadas pela sobrecarga de recursos. O jitter evita o alinhamento das solicitações em picos, e o recuo diminui a escalação de carga causado pela adição de repetições à carga normal da solicitação. Por fim, é importante configurar um número máximo de novas tentativas ou o tempo decorrido para evitar a criação de backlogs que produzam falhas metaestáveis.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Controle e limite as chamadas de novas tentativas. Use o recuo exponencial para tentar novamente após intervalos progressivamente mais longos. Introduza jitter para tornar esses intervalos de novas tentativas aleatórios e limite o número máximo de repetições.

Alguns AWS SDKs implementam novas tentativas e recuo exponencial por padrão. Use essas implementações integradas da AWS quando aplicável em sua workload. Implemente uma lógica semelhante em sua workload ao chamar serviços que sejam idempotentes e em que repetições melhorem a disponibilidade do cliente. Decida quais são os tempos limite e quando parar de tentar

novamente com base no seu caso de uso. Crie e simule cenários de teste para esses casos de uso de novas tentativas.

Etapas de implementação

- Determine a camada ideal em sua pilha de aplicações para implementar novas tentativas para os serviços dos quais sua aplicação depende.
- Conheça os SDKs existentes que implementam estratégias comprovadas de novas tentativas com retrocesso exponencial e jitter para a linguagem de sua escolha e dê preferência a esses SDKs em vez de escrever suas próprias implementações de repetição.
- Verifique se os serviços são idempotentes antes de implementar novas tentativas. Depois que as novas tentativas forem implementadas, elas deverão ser testadas e simuladas regularmente na produção.
- Ao chamar as APIs de serviço da AWS, use os AWS SDKs e a AWS CLI e entenda as opções de configuração de nova tentativa. Determine se os padrões funcionam para seu caso de uso, teste e ajuste conforme necessário.

Recursos

Práticas recomendadas relacionadas:

- REL04-BP04 Fazer com que todas as respostas sejam idempotentes
- REL05-BP02 Controlar a utilização de solicitações
- REL05-BP04 Antecipar-se à falha e limitar filas
- REL05-BP05 Definir tempos limite do cliente
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas

Documentos relacionados:

- Novas tentativas em caso de erro e recuo exponencial na AWS
- Amazon Builders' Library: tempos limite, novas tentativas e recuo com jitter
- Recuo exponencial e jitter
- Tornar as tentativas seguras com APIs idempotentes

Exemplos relacionados:

- Spring Retry
- Resilience4j Retry

Vídeos relacionados:

- Novas tentativas, recuo e jitter: AWS re:Invent 2019: Introdução à Amazon Builders' Library (DOP328)

Ferramentas relacionadas:

- AWS SDKs e ferramentas: comportamento de novas tentativas
- AWS Command Line Interface: Novas tentativas via AWS CLI

REL05-BP04 Antecipar-se à falha e limitar filas

Quando um serviço não consegue responder com êxito a uma solicitação, antecipe-se à falha. Isso permite a liberação dos recursos associados a uma solicitação e possibilita que o serviço se recupere se estiver ficando sem recursos. Antecipar-se à falha é um padrão de design de software bem estabelecido que pode ser utilizado para criar workloads altamente confiáveis na nuvem. As filas também correspondem a um padrão de integração empresarial bem estabelecido que pode facilitar o carregamento e permitir que os clientes liberem recursos quando o processamento assíncrono pode ser tolerado. Quando um serviço consegue responder com êxito em condições normais, mas falha quando a taxa de solicitações é muito alta, use uma fila para armazenar solicitações em buffer. No entanto, não permita a formação de backlogs de filas longas que possam ocasionar o processamento de solicitações antigas das quais um cliente já desistiu.

Resultado desejado: quando os sistemas enfrentam contenção de recursos, tempos limite, exceções ou falhas de causa desconhecida que tornam os objetivos de nível de serviço inatingíveis, as estratégias de antecipação a falhas permitem uma recuperação mais rápida do sistema. Sistemas que precisam absorver picos de tráfego e acomodar o processamento assíncrono podem melhorar a confiabilidade ao permitir que os clientes liberem solicitações rapidamente usando filas para armazenar solicitações em buffer para serviços de backend. Ao armazenar solicitações em filas, estratégias de gerenciamento de filas são implementadas para evitar backlogs intransponíveis.

Práticas comuns que devem ser evitadas:

- Implementar filas de mensagens, mas não configurar filas de mensagens não entregues (DLQ) ou alarmes em volumes DLQ para detectar quando um sistema está em falha.
- Não medir a idade das mensagens em uma fila, uma medida de latência para entender quando os consumidores da fila estão ficando para trás ou cometendo erros, ocasionando repetições.
- Não limpar mensagens pendentes de uma fila, quando não há utilidade em processar essas mensagens se a necessidade empresarial deixar de existir.
- Configurar filas do tipo “first in first out” (FIFO) quando filas do tipo “last in first out” (LIFO) atenderia melhor às necessidades do cliente, por exemplo, quando a ordenação rigorosa não é necessária e o processamento de backlog está atrasando todas as solicitações novas e urgentes, ocasionando violação dos níveis de serviço de todos os clientes.
- Expor filas internas aos clientes em vez de expor APIs que gerenciem a entrada de trabalho e coloquem as solicitações em filas internas.
- Combinar muitos tipos de solicitações de trabalho em uma única fila, o que pode agravar as condições de backlog ao distribuir a demanda de recursos entre os tipos de solicitação.
- Processar solicitações complexas e simples na mesma fila, apesar da necessidade de monitoramento, tempos limite e alocação de recursos diferentes.
- Não validar entradas ou usar afirmações para implementar mecanismos de antecipação à falha em software que agreguem exceções a componentes de nível superior que podem lidar com erros sem problemas.
- Não remover recursos com defeito do roteamento de solicitações, principalmente quando as falhas estão emitindo êxitos e falhas em decorrência de travamento e reinicialização, falha de dependência intermitente, capacidade reduzida ou perda de pacotes de rede.

Benefícios de implementar esta prática recomendada: sistemas que se antecipam às falhas são mais fáceis de depurar e corrigir e geralmente expõem problemas de codificação e configuração antes que as versões sejam publicadas em produção. Os sistemas que incorporam estratégias eficazes de filas oferecem maior resiliência e confiabilidade a picos de tráfego e às condições intermitentes de falha do sistema.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

As estratégias de antecipação à falha podem ser codificadas em soluções de software e configuradas em infraestrutura. Além de se anteciparem à falha, as filas são uma técnica arquitetônica simples, mas poderosa, para dissociar os componentes do sistema e facilitar o

carregamento. O Amazon CloudWatch oferece recursos para monitorar e alertar sobre falhas. Quando se sabe que um sistema está falhando, estratégias de mitigação podem ser invocadas, inclusive evitar recursos afetados. Quando os sistemas implementam filas com o Amazon SQS e outras tecnologias de fila para facilitar o carregamento, eles devem considerar como gerenciar os backlogs de filas, bem como as falhas no consumo de mensagens.

Etapas de implementação

- Implemente afirmações programáticas ou métricas específicas em seu software e use-as para alertar explicitamente sobre problemas do sistema. O Amazon CloudWatch ajuda você a criar métricas e alarmes com base no padrão de log da aplicação e na instrumentação do SDK.
- Use métricas e alarmes do CloudWatch para eliminar recursos danificados que estão aumentando a latência no processamento ou falhando repetidamente no processamento das solicitações.
- Use o processamento assíncrono criando APIs para aceitar e anexar solicitações às filas internas usando o Amazon SQS e, em seguida, responder ao cliente que produz a mensagem com uma mensagem de êxito para que o cliente possa liberar recursos e prosseguir com outros trabalhos enquanto os consumidores da fila de backend processam as solicitações.
- Avalie e monitore a latência do processamento da fila produzindo uma métrica do CloudWatch sempre que retirar uma mensagem de uma fila, comparando o momento presente com o carimbo de data/hora da mensagem.
- Quando falhas impedem o processamento bem-sucedido de mensagens ou geram picos de tráfego em volumes que não podem ser processados de acordo com acordos de serviço, deixe de lado o tráfego antigo ou excedente para uma fila de transbordamento. Isso permite o processamento prioritário de trabalhos novos e antigos quando há capacidade disponível. Essa técnica é uma aproximação do processamento LIFO e permite o processamento normal do sistema para todos os novos trabalhos.
- Use filas de mensagens não entregues ou de redirecionamento para mover mensagens que não podem ser processadas do backlog para um local que possa ser pesquisado e resolvido posteriormente.
- Tente novamente ou, quando possível, elimine as mensagens antigas comparando o momento presente com o carimbo de data/hora da mensagem e descartando as mensagens que não são mais relevantes para o cliente solicitante.

Recursos

Práticas recomendadas relacionadas:

- REL04-BP02 Implementar dependências com acoplamento fraco
- REL05-BP02 Controlar a utilização de solicitações
- REL05-BP03 Controlar e limitar chamadas de novas tentativas
- REL06-BP02 Definir e calcular métricas (agregação)
- REL06-BP07 Monitorar o rastreamento completo das solicitações por meio de seu sistema

Documentos relacionados:

- Evitar backlogs de fila intransponíveis
- Antecipar-se à falha
- Como posso evitar um aumento no atraso das mensagens na minha fila Amazon SQS?
- Elastic Load Balancing: mudança de zona
- Amazon Application Recovery Controller: controle de roteamento para failover de tráfego

Exemplos relacionados:

- Padrões de integração empresarial: canal de mensagens não entregues

Vídeos relacionados:

- AWS re:Invent 2022: Operar aplicações Multi-AZ altamente disponíveis

Ferramentas relacionadas:

- Amazon SQS
- Amazon MQ
- AWS IoT Core
- Amazon CloudWatch

REL05-BP05 Definir tempos limite do cliente

Defina tempos limite adequados para conexões e solicitações, verifique-os sistematicamente e não confie nos valores padrão, pois eles não estão cientes das especificações da workload.

Resultado desejado: os tempos limite do cliente devem considerar o custo para o cliente, o servidor e a workload associados à espera por solicitações que levam um tempo anormal para serem concluídas. Como não é possível saber a causa exata de nenhum tempo limite, os clientes devem usar o conhecimento dos serviços para desenvolver expectativas de causas prováveis e prazos apropriados.

As conexões do cliente atingem o tempo limite com base nos valores configurados. Depois de encontrar um tempo limite, os clientes tomam a decisão de recuar e tentar novamente ou abrir um disjuntor. Esses padrões evitam a emissão de solicitações que podem exacerbar uma condição de erro subjacente.

Práticas comuns que devem ser evitadas:

- Não estar ciente dos tempos limite do sistema ou dos tempos limite padrão.
- Não estar ciente do tempo normal de conclusão da solicitação.
- Não estar ciente das possíveis causas das solicitações levarem muito tempo para serem concluídas ou dos custos de performance do cliente, do serviço ou da workload associados à espera por essas conclusões.
- Não estar ciente da probabilidade de uma rede danificada fazer com que uma solicitação falhe somente quando o tempo limite é atingido e dos custos para a performance do cliente e da workload por não adotar um tempo limite mais curto.
- Não testar cenários de tempo limite tanto para conexões quanto para solicitações.
- Definir tempos limite muito altos, o que pode resultar em longos tempos de espera e aumentar a utilização de recursos.
- Definir tempos limite muito baixos, gerando falhas artificiais.
- Ignorar padrões para lidar com erros de tempo limite para chamadas remotas, como disjuntores e novas tentativas.
- Não considerar o monitoramento de taxas de erro de chamadas de serviço, objetivos de nível de serviço para latência e valores atípicos de latência. Essas métricas podem fornecer informações sobre tempos limite agressivos ou permissivos.

Benefícios de implementar esta prática recomendada: os tempos limite de chamadas remotas são configurados e os sistemas são projetados para lidar com os tempos limite normalmente de forma que os recursos sejam conservados quando as chamadas remotas respondem de forma anormalmente lenta e os erros de tempo limite sejam tratados normalmente pelos clientes do serviço.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Defina um tempo limite de conexão e um tempo limite de solicitação em qualquer chamada de dependência de serviço e, geralmente, em qualquer chamada entre processos. Muitas frameworks oferecem recursos de tempo limite integrados, mas tenha cuidado, pois algumas têm valores padrão infinitos ou superiores ao aceitável para seus objetivos de serviço. Um valor muito alto reduz a utilidade do tempo limite porque os recursos continuam a ser consumidos enquanto o cliente aguarda o decorrer do tempo limite. Um valor muito baixo pode gerar maior tráfego no backend e maior latência, porque muitas solicitações são repetidas. Em alguns casos, isso pode levar a interrupções completas porque todas as solicitações estão sendo repetidas.

Considere o seguinte ao determinar as estratégias de tempo limite:

- As solicitações podem levar mais tempo do que o normal para serem processadas devido ao conteúdo, a deficiências em um serviço de destino ou a uma falha na partição de rede.
- Solicitações com conteúdo anormalmente caro podem consumir recursos desnecessários do servidor e do cliente. Nesse caso, reduzir o tempo limite dessas solicitações e não tentar novamente pode preservar os recursos. Os serviços também devem se proteger de conteúdo anormalmente caro com limitações e tempos limite do servidor.
- Solicitações que demoram muito devido a uma falha no serviço podem expirar e ser repetidas. Deve-se considerar os custos do serviço para a solicitação e a nova tentativa, mas se a causa for uma deficiência localizada, uma nova tentativa provavelmente não será cara e reduzirá o consumo de recursos do cliente. O tempo limite também pode liberar recursos do servidor, dependendo da natureza da deficiência.
- Solicitações que demoram muito para serem concluídas porque a solicitação ou a resposta não foi entregue pela rede podem expirar e ser repetidas. Como a solicitação ou a resposta não foi entregue, a falha teria sido o resultado, independentemente da duração do tempo limite. Nesse caso, o tempo limite não liberará recursos do servidor, mas liberará recursos do cliente e melhorará a performance da workload.

Aproveite os padrões de design bem estabelecidos, como novas tentativas e disjuntores, para lidar com os tempos de espera de forma eficiente e oferecer compatibilidade com abordagens de antecipação à falha. AWS Os SDKs e a AWS CLI permitem a configuração de tempos limite de conexão e solicitação e novas tentativas com recuo exponencial e jitter. As funções do AWS Lambda são compatíveis com a configuração de tempos limite. E, com o AWS Step Functions, você pode

criar disjuntores com pouco código que aproveitam as integrações pré-construídas com os serviços e SDKs da AWS. AWS App Mesh O Envoy oferece recursos de tempo limite e disjuntor.

Etapas de implementação

- Configure tempos limite em chamadas de serviço remoto e utilize os recursos de tempo limite de linguagem integrados ou as bibliotecas de tempo limite de código aberto.
- Quando sua workload fizer chamadas com um AWS SDK, revise a documentação para saber a configuração de tempo limite específica da linguagem.
  - Python
  - PHP
  - .NET
  - Ruby
  - Java
  - Go
  - Node.js
  - C++
- Ao usar AWS SDKs ou comandos da AWS CLI em sua workload, configure os valores de tempo limite padrão definindo os padrões de configuração da AWS para `connectTimeoutInMillis` e `tlsNegotiationTimeoutInMillis`.
- Aplique opções de linha de comando `cli-connect-timeout` e `cli-read-timeout` para controlar comandos da AWS CLI únicos para serviços da AWS.
- Monitore o tempo limite de chamadas de serviço remoto e defina alarmes para erros persistentes para que você possa lidar proativamente com cenários de erro.
- Implemente métricas do CloudWatch e detecção de anomalias do CloudWatch em taxas de erro de chamada, objetivos de nível de serviço para latência e valores atípicos de latência para fornecer informações sobre o gerenciamento de tempos limite excessivamente agressivos ou permissivos.
- Configure tempos limite nas funções do Lambda.
- Os clientes do API Gateway devem implementar suas próprias repetições ao lidar com os tempos limite. O API Gateway oferece suporte a um tempo limite de integração de 50 milissegundos a 29 segundos para integrações downstream e não tenta novamente quando a integração solicita o tempo limite.

- Implemente o padrão de disjuntor para evitar fazer chamadas remotas quando o tempo limite está prestes a ser atingido. Abra o circuito para evitar falhas nas chamadas e feche-o quando as chamadas estiverem respondendo normalmente.
- Para workloads baseadas em contêineres, revise os recursos do App Mesh Envoy para aproveitar os tempos limite e os disjuntores integrados.
- Use o AWS Step Functions para criar disjuntores de pouco uso de código para chamadas de serviço remoto, especialmente ao chamar SDKs nativos da AWS e integrações do Step Functions compatíveis para simplificar sua workload.

Recursos

Práticas recomendadas relacionadas:

- REL05-BP03 Controlar e limitar chamadas de novas tentativas
- REL05-BP04 Antecipar-se à falha e limitar filas
- REL06-BP07 Monitorar o rastreamento completo das solicitações por meio de seu sistema

Documentos relacionados:

- AWS SDK: novas tentativas e tempos limite
- Amazon Builders' Library: tempos limite, novas tentativas e recuo com jitter
- Cotas do Amazon API Gateway e notas importantes
- Opções de linha de comando do AWS Command Line Interface
- AWS SDK for Java 2.x: configurar tempos limite de API
- AWS Botocore usando o objeto de configuração e a referência de configuração
- AWS SDK for .NET: novas tentativas e tempos limite
- AWS Lambda: configurar as opções da função do Lambda

Exemplos relacionados:

- Usar o padrão do disjuntor com o AWS Step Functions e o Amazon DynamoDB
- Martin Fowler: CircuitBreaker

Ferramentas relacionadas:

- AWS SDKs
- AWS Lambda
- Amazon SQS
- AWS Step Functions
- AWS Command Line Interface

REL05-BP06 Criar serviços sem estado sempre que possível

Os sistemas não devem exigir estado ou devem descarregar o estado de modo que não haja dependência entre solicitações de clientes diferentes em relação aos dados armazenados localmente no disco ou na memória. Isso permite que os servidores sejam substituídos quando necessário sem prejudicar a disponibilidade.

Quando os usuários ou serviços interagem com uma aplicação, eles geralmente executam uma série de interações que formam uma sessão. Uma sessão são dados exclusivos para usuários que persistem entre solicitações enquanto usam a aplicação. Uma aplicação sem estado é uma aplicação que não precisa de conhecimento de interações anteriores e não armazena informações da sessão.

Depois de projetados para serem sem estado, você pode usar serviços de computação com tecnologia sem servidor, como o AWS Lambda ou o AWS Fargate.

Além da substituição do servidor, outro benefício das aplicações sem estado é que elas podem escalar horizontalmente, pois qualquer um dos recursos de computação disponíveis (como instâncias do EC2 e funções do AWS Lambda) pode atender a qualquer solicitação.

Benefícios de implementar esta prática recomendada: os sistemas projetados para serem sem estado são mais adaptáveis ao dimensionamento horizontal, possibilitando a adição ou remoção de capacidade com base na flutuação do tráfego e da demanda. Eles também são inerentemente resilientes a falhas e oferecem flexibilidade e agilidade no desenvolvimento de aplicações.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Crie aplicações sem estado. As aplicações sem estado permitem o ajuste de escala horizontal e são tolerantes a falhas de um nó individual. Analise e compreenda os componentes da aplicação que mantêm estado dentro da arquitetura. Isso ajuda você a avaliar o impacto potencial da transição para um design sem estado. Uma arquitetura sem estado dissocia os dados de usuários e descarrega os

dados de sessões. Isso oferece a flexibilidade de escalar cada componente de forma independente para atender às diferentes demandas de workload e otimizar a utilização de recursos.

Etapas de implementação

- Identifique e compreenda os componentes com estado na aplicação.
- Dissocie os dados, separando e gerenciando os dados de usuários da lógica principal da aplicação.
  - O Amazon Cognito pode dissociar os dados do usuário do código da aplicação usando recursos, como bancos de identidades, grupos de usuários e o Amazon Cognito Sync.
  - É possível usar o AWS Secrets Manager para desacoplar dados do usuário armazenando segredos em um local seguro e centralizado. Isso significa que o código da aplicação não precisa armazenar segredos, o que a torna mais segura.
  - Considere usar o Amazon S3 para armazenar dados grandes e não estruturados, como imagens e documentos. Sua aplicação poderá recuperar esses dados quando necessário, eliminando a necessidade de armazená-los na memória.
  - Use o Amazon DynamoDB para armazenar informações, como perfis de usuário. Sua aplicação poderá consultar esses dados praticamente em tempo real.
- Descarregue os dados de sessões em um banco de dados, cache ou arquivos externos.
  - O Amazon ElastiCache, o Amazon DynamoDB, o Amazon Elastic File System (Amazon EFS) e o Amazon MemoryDB são exemplos de serviços da AWS que você pode usar para descarregar dados da sessão.
- Crie uma arquitetura sem estado depois de identificar quais dados de estado e de usuários precisam ser mantidos com sua solução de armazenamento preferida.

Recursos

Práticas recomendadas relacionadas:

- REL11-BP03 Automatizar a reparação em todas as camadas

Documentos relacionados:

- Amazon Builders' Library: evitar fallback em sistemas distribuídos
- Amazon Builders' Library: evitar backlogs de fila insuperáveis
- Amazon Builders' Library: desafios e estratégias de armazenamento em cache

- Práticas recomendadas para níveis na Web sem estado na AWS

REL05-BP07 Implementar medidas emergenciais

Medidas emergenciais são processos rápidos que podem atenuar o impacto da disponibilidade na workload.

As medidas emergenciais funcionam com a desativação, o controle de utilização ou a alteração do comportamento dos componentes ou das dependências com o uso de mecanismos conhecidos e testados. Isso pode aliviar as deficiências da workload decorrentes da exaustão dos recursos provocada por aumentos inesperados na demanda e reduzir o impacto de falhas em componentes não essenciais da workload.

Resultado desejado: ao implementar medidas de emergência, você pode estabelecer processos em boas condições para manter a disponibilidade de componentes essenciais em sua workload. A workload deve se degradar normalmente e continuar desempenhando suas funções essenciais aos negócios durante a ativação de uma medida emergencial. Para obter detalhes sobre a degradação normal, consulte REL05-BP01 Implementar uma degradação normal para transformar dependências rígidas aplicáveis em dependências flexíveis.

Práticas comuns que devem ser evitadas:

- A falha de dependências não essenciais afeta a disponibilidade da workload principal.
- Não testar ou verificar o comportamento dos componentes essenciais durante a deterioração de componentes não essenciais.
- Não há critérios claros e determinísticos definidos para ativação ou desativação de uma medida emergencial.

Benefícios de implementar esta prática recomendada: a implementação de medidas emergenciais pode melhorar a disponibilidade dos componentes críticos em sua workload, fornecendo aos seus resolvedores processos estabelecidos para responder a picos inesperados na demanda ou falhas de dependências não críticas.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

- Identifique os componentes essenciais na workload.

- Projete e arquitete os componentes essenciais na workload para resistirem à falha de componentes não essenciais.
- Conduza testes para validar o comportamento dos componentes essenciais durante a falha de componentes não essenciais.
- Defina e monitore métricas ou acionadores relevantes para iniciar procedimentos de medida emergencial.
- Defina os procedimentos (manuais ou automatizados) que compõem a medida emergencial.

Etapas de implementação

- Identifique os componentes essenciais aos negócios na workload.
  - Cada componente técnico na workload deve ser mapeado para a função de negócios relevante e classificado como essencial ou não essencial. Para exemplos de funcionalidades críticas e não críticas na Amazon, consulte Qualquer dia pode ser o Prime Day: Como a pesquisa da Amazon.com usa a engenharia do caos para lidar com mais de 84 mil solicitações por segundo.
  - Essa é uma decisão técnica e de negócios e varia de acordo com a organização e a workload.
- Projete e arquitete os componentes essenciais na workload para resistirem à falha de componentes não essenciais.
  - Durante a análise de dependências, considere todos os possíveis modos de falha e verifique se os mecanismos de medida emergencial fornecem a funcionalidade essencial aos componentes subsequentes.
- Conduza testes para validar o comportamento dos componentes essenciais durante a ativação das medidas emergenciais.
  - Evite comportamento bimodal. Para obter mais detalhes, consulte REL11-BP05 Usar estabilidade estática para evitar comportamento bimodal
- Defina, monitore e emita alertas sobre as métricas relevantes para iniciar o procedimento de medida emergencial.
  - A descoberta das métricas certas a serem monitoradas depende da workload. Alguns exemplos de métricas são a latência ou o número de solicitações com falha feitas para uma dependência.
- Defina os procedimentos, manuais ou automatizados, que compõem a medida emergencial.
  - Isso pode incluir mecanismos como redução de carga, controle de utilização de solicitações ou implementação de degradação normal.

Recursos

Práticas recomendadas relacionadas:

- REL05-BP01 Implementar uma degradação normal para transformar dependências rígidas aplicáveis em dependências flexíveis
- REL05-BP02 Controlar a utilização de solicitações
- REL11-BP05 Usar estabilidade estática para evitar comportamento bimodal

Documentos relacionados:

- Automatizar implantações seguras e sem intervenção
- Qualquer dia pode ser o Prime Day: como a pesquisa da Amazon.com usa a engenharia do caos para lidar com mais de 84 mil solicitações por segundo

Vídeos relacionados:

- AWS re:Invent 2020: Confiabilidade, consistência e confiança por meio da imutabilidade

Gerenciamento de alterações

Perguntas

- REL 6. Como você monitora os recursos da workload?
- REL 7. Como você projeta a workload para se adaptar às alterações na demanda?
- REL 8. Como você implementa a alteração?

REL 6. Como você monitora os recursos da workload?

Logs e métricas são ferramentas avançadas para obter informações sobre a integridade da workload. Você pode configurar a workload para monitorar logs e métricas e enviar notificações quando os limites forem ultrapassados ou ocorrerem eventos significativos. O monitoramento permite que sua workload reconheça quando os limites de baixa performance são ultrapassados ou quando há falhas para que ela possa se recuperar automaticamente em resposta.

Práticas recomendadas

- REL06-BP01 Monitorar todos os componentes da workload (geração)

- REL06-BP02 Definir e calcular métricas (agregação)
- REL06-BP03 Enviar notificações (processamento e alarmes em tempo real)
- REL06-BP04 Automatizar respostas (processamento e alarmes em tempo real)
- REL06-BP05 Analisar logs
- REL06-BP06 Realizar revisões regularmente
- REL06-BP07 Monitorar o rastreamento completo das solicitações por meio de seu sistema

REL06-BP01 Monitorar todos os componentes da workload (geração)

Monitore os componentes da workload com o Amazon CloudWatch ou ferramentas de terceiros. Monitore os serviços da AWS com o AWS Health Dashboard.

Todos os componentes da workload devem ser monitorados, incluindo frontend, lógica de negócios e níveis de armazenamento. Defina as principais métricas e como extraí-las dos logs (se necessário) e defina limites para invocação de eventos de alarme correspondentes. Garanta que as métricas sejam relevantes para os indicadores-chave de performance (KPIs) da sua workload e use métricas e logs para identificar sinais precoces de degradação do serviço. Por exemplo, uma métrica relacionada aos resultados comerciais, como o número de pedidos processados com sucesso por minuto, pode indicar problemas de workload mais rapidamente do que uma métrica técnica, como a utilização da CPU. Use o AWS Health Dashboard para obter uma visualização personalizada da performance e da disponibilidade dos serviços da AWS subjacentes aos recursos da AWS.

O monitoramento na nuvem oferece novas oportunidades. A maioria dos provedores de nuvem desenvolveu hooks personalizáveis e pode fornecer informações para ajudar você a monitorar várias camadas da sua workload. Serviços da AWS como o Amazon CloudWatch aplicam algoritmos estatísticos e de machine learning para analisar continuamente métricas de sistemas e aplicações, determinar linhas de base normais e apontar anomalias com intervenção mínima do usuário. Os algoritmos de detecção de anomalias consideram a sazonalidade e as mudanças de tendência das métricas.

A AWS disponibiliza uma enorme quantidade de informações de monitoramento e log para consumo que podem ser usadas para definir métricas específicas da workload e processos de alteração sob demanda e adotar técnicas de machine learning, independentemente da experiência em ML.

Além disso, monitore todos os seus endpoints externos para garantir que eles sejam independentes de sua implementação de base. Esse monitoramento ativo pode ser feito com transações sintéticas (às vezes chamadas de canários do usuário, mas que não devem ser confundidas com implantações

canários) que executam periodicamente diversas tarefas comuns correspondentes a ações executadas pelos consumidores da workload. Mantenha essas tarefas com curta duração e certifique-se de não sobrecarregar sua workload durante o teste. O Amazon CloudWatch Synthetics permite criar canários sintéticos para monitorar endpoints e APIs. Você também pode combinar os nós sintéticos do cliente canário com o console do AWS X-Ray para identificar quais canários sintéticos estão enfrentando problemas com erros, falhas ou taxas de controle de utilização para o período selecionado.

Resultado desejado:

Colete e use métricas críticas de todos os componentes da workload para garantir a confiabilidade da workload e a experiência ideal do usuário. Detectar que uma workload não está alcançando resultados comerciais permite que você declare rapidamente um desastre e se recupere de um incidente.

Práticas comuns que devem ser evitadas:

- Monitorar apenas as interfaces externas com sua workload.
- Não gerar nenhuma métrica específica da workload nem depender apenas das métricas fornecidas pelos serviços da AWS usados por sua workload.
- Usar apenas métricas técnicas em sua workload e não monitorar nenhuma métrica relacionada a KPIs não técnicos para os quais a workload contribui.
- Contar com o tráfego de produção e com verificações de saúde simples para monitorar e avaliar o estado da workload.

Benefícios de implementar esta prática recomendada: o monitoramento em todos os níveis de sua workload permite que você antecipe e resolva problemas mais rapidamente nos componentes que fazem parte da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

1. Ative o registro quando disponível. Os dados de monitoramento devem ser obtidos de todos os componentes das workloads. Ative o log adicional, como os logs de acesso do S3, e permita que sua workload registre dados específicos da workload. Colete métricas para médias de CPU, E/S de rede e E/S de disco de serviços como Amazon ECS, Amazon EKS, Amazon EC2, Elastic Load Balancing, AWS Auto Scaling e Amazon EMR. Consulte Serviços da AWS que publicam métricas do CloudWatch para obter uma lista dos serviços da AWS que publicam métricas no CloudWatch.

2. Analise todas as métricas padrão e explore quaisquer lacunas na coleta de dados. Cada serviço gera métricas padrão. A coleta de métricas padrão permite que você entenda melhor as dependências entre os componentes da workload e como a confiabilidade e a performance dos componentes afetam a workload. Você também pode publicar suas próprias métricas no CloudWatch usando a AWS CLI ou uma API.
3. Avalie todas as métricas para decidir quais delas alertar para cada serviço da AWS em sua workload. Você pode optar por selecionar um subconjunto de métricas que tenham um grande impacto na confiabilidade da workload. Concentrar-se em métricas e limites críticos permite refinar o número de alertas e pode ajudar a minimizar os falsos positivos.
4. Defina alertas e o processo de recuperação para sua workload após a chamada do alerta. A definição de alertas permite que você notifique, escale e siga rapidamente as etapas necessárias para se recuperar de um incidente e atingir seu objetivo de tempo de recuperação (RTO) prescrito. Você pode usar os alarmes do Amazon CloudWatch para invocar fluxos de trabalho automatizados e iniciar procedimentos de recuperação com base em limites definidos.
5. Explore o uso de transações sintéticas para coletar dados relevantes sobre o estado das workloads. O monitoramento sintético segue as mesmas rotas e executa as mesmas ações que um cliente, o que possibilita verificar continuamente a experiência do cliente, mesmo quando você não tem nenhum tráfego de cliente em suas workloads. Ao usar transações sintéticas, é possível descobrir problemas antes que seus clientes o façam.

Recursos

Práticas recomendadas relacionadas:

- REL11-BP03 Automatizar a reparação em todas as camadas

Documentos relacionados:

- Conceitos básicos do AWS Health Dashboard: integridade da conta
- Serviços da AWS que publicam métricas do CloudWatch
- Logs de acesso do Network Load Balancer
- Logs de acesso para seu Application Load Balancer
- Acessar o Amazon CloudWatch Logs para AWS Lambda
- Registro em log de acesso ao servidor do Amazon S3
- Habilitar logs de acesso para seu Classic Load Balancer

- Exportar dados de log para o Amazon S3
- Instalar o agente do CloudWatch em uma instância do Amazon EC2
- Publicar métricas personalizadas
- Usar painéis do Amazon CloudWatch
- Usar métricas do Amazon CloudWatch
- Usar canários (Amazon CloudWatch Synthetics)
- O que são Amazon CloudWatch Logs?

Guias do usuário:

- Criar uma trilha
- Monitorar métricas de memória e disco para instâncias Linux do Amazon EC2
- Usar o CloudWatch Logs com Instâncias de contêiner
- VPC Flow Logs
- O que é o Amazon DevOps Guru?
- O que é AWS X-Ray?

Blogs relacionados:

- Depurar com o Amazon CloudWatch Synthetics e o AWS X-Ray

Exemplos e workshops relacionados:

- Laboratórios do AWS Well-Architected: Excelência operacional – Monitoramento de dependências
- Amazon Builders' Library: instrumentação de sistemas distribuídos para visibilidade operacional
- Workshop Observability

REL06-BP02 Definir e calcular métricas (agregação)

Armazene os dados de log e aplique filtros quando necessário para calcular métricas como contagens de um evento de log específico ou latência calculada com base na data e hora dos eventos de log

O Amazon CloudWatch e o Amazon S3 servem como as camadas de armazenamento e agregação primárias. Para alguns serviços, como o AWS Auto Scaling e o Elastic Load Balancing, métricas padrão são fornecidas por padrão para a carga da CPU ou a latência média da solicitação em um

cluster ou instância. Para serviços de streaming, como os Logs de fluxo da VPC e o AWS CloudTrail, os dados de evento são encaminhados ao CloudWatch Logs e você precisa definir e aplicar filtros de métricas para extraí-las dos dados do evento. Isso fornece dados de séries temporais, que podem servir como entradas para alarmes do CloudWatch que você define para invocar alertas.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

- Defina e calcule as métricas (agregação) Armazene os dados de log e aplique filtros quando necessário para calcular métricas como contagens de um evento de log específico ou latência calculada com base na data e hora dos eventos de log
  - Os filtros de métrica definem os termos e os padrões a serem procurados nos dados de log à medida que são enviados ao CloudWatch Logs. O CloudWatch Logs usa esses filtros de métrica para transformar os dados de log em métricas numéricas do CloudWatch que podem ser usadas para criar um gráfico ou definir um alarme.
    - Pesquisar e filtrar dados de log
  - Use um terceiro confiável para agregar logs
    - Siga as instruções do terceiro. A maioria dos produtos de terceiros integra-se ao CloudWatch e ao Amazon S3.
  - Alguns serviços da AWS podem publicar logs diretamente no Amazon S3. Se seu principal requisito de logs for o armazenamento no Amazon S3, você poderá facilmente fazer com que o serviço que produz os logs os envie diretamente para o Amazon S3 sem configurar uma infraestrutura adicional
    - Enviar logs diretamente para o Amazon S3

Recursos

Documentos relacionados:

- Consultas de exemplo do Amazon CloudWatch Logs Insights
- Depurar com o Amazon CloudWatch Synthetics e o AWS X-Ray
- Workshop One Observability
- Pesquisar e filtrar dados de log
- Enviar logs diretamente para o Amazon S3
- Amazon Builders' Library: instrumentação de sistemas distribuídos para visibilidade operacional

REL06-BP03 Enviar notificações (processamento e alarmes em tempo real)

Quando as organizações detectam possíveis problemas, elas enviam notificações e alertas em tempo real para a equipe e os sistemas apropriados para responder de forma rápida e eficaz a esses problemas.

Resultado desejado: respostas rápidas a eventos operacionais são possíveis por meio da configuração de alarmes relevantes com base em métricas de serviços e aplicações. Quando os limites do alarme são violados, a equipe e os sistemas apropriados são notificados para que possam resolver os problemas subjacentes.

Práticas comuns que devem ser evitadas:

- Configurar alarmes com um limite excessivamente alto, resultando em falha no envio de notificações vitais.
- Configurar alarmes com um limite muito baixo, ocasionando inatividade diante de alertas importantes devido ao ruído de notificações excessivas.
- Não atualizar os alarmes e seu limite quando o uso muda.
- Para alarmes mais bem abordados por meio de ações automatizadas, enviar a notificação ao pessoal em vez de gerar a ação automatizada gera o envio excessivo de notificações.

Benefícios de implementar esta prática recomendada: o envio de notificações e alertas em tempo real para o pessoal e os sistemas apropriados permite a detecção precoce de problemas e respostas rápidas aos incidentes operacionais.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

As workloads devem ser equipadas com processamento e alarmes em tempo real para melhorar a capacidade de detecção de problemas que possam afetar a disponibilidade da aplicação e servir como gatilhos para respostas automatizadas. As organizações podem realizar processamento e alarmes em tempo real criando alertas com métricas definidas para receber notificações sempre que eventos significativos ocorrerem ou quando uma métrica ultrapassar um limite.

O Amazon CloudWatch permite criar alarmes métricos e compostos usando alarmes do CloudWatch com base em limite estático, detecção de anomalias e outros critérios. Para obter mais detalhes sobre os tipos de alarmes que é possível configurar usando o CloudWatch, consulte a seção de alarmes da documentação do CloudWatch.

É possível criar visualizações personalizadas de métricas e alertas dos recursos da AWS para as equipes com os painéis do CloudWatch. As páginas iniciais personalizáveis no console do CloudWatch permitem que você monitore seus recursos em uma única visualização em várias regiões.

Os alarmes podem executar uma ou mais ações, como enviar uma notificação a um tópico do Amazon SNS, executar uma ação do Amazon EC2 ou uma ação do Amazon EC2 Auto Scaling ou criar um OpsItem ou incidente no AWS Systems Manager.

O Amazon CloudWatch usa o Amazon SNS para enviar notificações quando o alarme muda de estado, fornecendo a entrega de mensagens dos publicadores (produtores) para os assinantes (consumidores). Para obter mais detalhes sobre como configurar as notificações do Amazon SNS, consulte Configurar o Amazon SNS.

O CloudWatch envia eventos do EventBridge sempre que um alarme do CloudWatch é criado, atualizado, excluído ou muda de estado. É possível usar o EventBridge com esses eventos para criar regras que realizam ações, como enviar uma notificação sempre que o estado de um alarme mudar ou acionar eventos automaticamente na conta usando a automação do Systems Manager.

Quando você deve usar o EventBridge ou o Amazon SNS?

Tanto o EventBridge quanto o Amazon SNS podem ser usados para desenvolver aplicações orientadas a eventos, e sua escolha dependerá de suas necessidades específicas.

O Amazon EventBridge é recomendado quando você deseja criar uma aplicação que reaja a eventos das suas próprias aplicações, aplicações SaaS e serviços da AWS. O EventBridge é o único serviço baseado em eventos que se integra diretamente com parceiros SaaS terceirizados. O EventBridge também ingere automaticamente eventos de mais de 200 serviços da AWS sem exigir que os desenvolvedores criem recursos em suas contas.

O EventBridge usa uma estrutura definida baseada em JSON para eventos e ajuda você a criar regras que são aplicadas em todo o corpo do evento para selecionar eventos a serem encaminhados para um destino. No momento, o EventBridge oferece suporte a mais de 20 serviços da AWS como destino, incluindo o AWS Lambda, Amazon SQS, Amazon SNS, Amazon Kinesis Data Streams e Amazon Data Firehose.

O Amazon SNS é recomendado para aplicações que precisam de alta distribuição (milhares ou milhões de endpoints). Um padrão comum que observamos é que os clientes usam o Amazon SNS como destino para a regra para filtrar os eventos de que precisam e distribuí-los para vários endpoints.

As mensagens não são estruturadas e podem estar em qualquer formato. O Amazon SNS oferece suporte ao encaminhamento de mensagens para seis tipos diferentes de destinos, incluindo Lambda, Amazon SQS, endpoints do HTTP/S, SMS, push móvel e e-mail. No Amazon SNS, a latência típica é inferior a 30 milissegundos. Uma ampla variedade de serviços da AWS enviam mensagens do Amazon SNS configurando o serviço para fazer isso (mais de 30, incluindo o Amazon EC2, o Amazon S3 e o Amazon RDS).

Etapas de implementação

1. Crie um alarme usando os alarmes do Amazon CloudWatch.
    a. Um alarme de métrica monitora uma única métrica do CloudWatch ou uma expressão dependente de métricas do CloudWatch. O alarme inicia uma ou mais ações com base no valor da métrica ou expressão em comparação com um limite em vários intervalos de tempo. A ação pode ser enviar uma notificação a um tópico do Amazon SNS, executar uma ação do Amazon EC2 ou uma ação do Amazon EC2 Auto Scaling ou criar um OpsItem ou incidente no AWS Systems Manager.
    b. Um alarme composto consiste em uma expressão de regra que considera as condições de alarme de outros alarmes que você criou. O alarme composto só entrará no estado de alarme se todas as condições da regra forem atendidas. Os alarmes especificados na expressão da regra de um alarme composto podem incluir alarmes de métricas e outros alarmes compostos. Os alarmes compostos podem enviar notificações do Amazon SNS quando mudam de estado e podem criar OpsItems ou incidentes do Systems Manager quando entram no estado de alarme, mas não podem executar ações do EC2 ou ações do Auto Scaling.
2. Configure notificações do Amazon SNS. Ao criar um alarme do CloudWatch, é possível incluir um tópico do Amazon SNS para enviar uma notificação quando o alarme mudar de estado.
3. Crie regras no EventBridge que correspondam aos alarmes especificados do CloudWatch. Cada regra é compatível com vários destinos, incluindo funções do Lambda. Por exemplo, você pode definir um alarme que é iniciado quando o espaço disponível em disco está acabando, o que aciona uma função do Lambda por meio de uma regra do EventBridge para limpar o espaço. Para obter mais detalhes sobre os alvos do EventBridge, consulte Destinos do EventBridge.

Recursos

Práticas recomendadas do Well-Architected relacionadas:

- REL06-BP01 Monitorar todos os componentes da workload (geração)
- REL06-BP02 Definir e calcular métricas (agregação)

- REL12-BP01 Usar playbooks para investigar falhas

Documentos relacionados:

- Amazon CloudWatch
- Insights do CloudWatch Logs
- Usar alarmes do Amazon CloudWatch
- Usar painéis do Amazon CloudWatch
- Usar métricas do Amazon CloudWatch
- Configurar notificações do Amazon SNS
- Detecção de anomalias do CloudWatch
- Proteção de dados do CloudWatch Logs
- Amazon EventBridge
- Amazon Simple Notification Service

Vídeos relacionados:

- Vídeos sobre observabilidade do re:Invent 2022
- AWS re:Invent 2022: Práticas recomendadas de observabilidade na Amazon

Exemplos relacionados:

- Workshop One Observability
- Amazon EventBridge para AWS Lambda com controle de feedback por alarmes do Amazon CloudWatch

REL06-BP04 Automatizar respostas (processamento e alarmes em tempo real)

Use a automação para executar uma ação quando um evento é detectado, por exemplo, para substituir componentes com falha.

O processamento automatizado de alarmes em tempo real é implementado para que os sistemas possam tomar medidas corretivas rapidamente e tentar evitar falhas ou degradação dos serviços quando os alarmes são acionados. As respostas automatizadas a alarmes podem incluir a

substituição de componentes com falha, o ajuste da capacidade computacional, o redirecionamento do tráfego para hosts, zonas de disponibilidade ou outras regiões íntegras e a notificação dos operadores.

Resultado desejado: os alarmes em tempo real são identificados e o processamento automatizado dos alarmes é configurado para invocar as ações apropriadas tomadas para manter os objetivos de nível de serviço e os contratos de nível de serviço (SLAs). A automação pode variar de atividades de autorrecuperação de componentes individuais a failover de todo o site.

Práticas comuns que devem ser evitadas:

- Não ter um inventário ou catálogo claro dos principais alarmes em tempo real.
- Não haver respostas automatizadas para alarmes essenciais (por exemplo, quando a computação está quase esgotada, ocorre o ajuste de escala automático).
- Usar ações contraditórias de resposta a alarmes.
- Não haver procedimentos operacionais padrão (SOPs) para os operadores seguirem ao receberem notificações de alerta.
- Não monitorar as alterações da configuração, pois alterações não detectadas podem causar tempo de inatividade nas workloads.
- Não haver uma estratégia para desfazer alterações não intencionais da configuração.

Benefícios de implementar esta prática recomendada: automatizar o processamento de alarmes pode melhorar a resiliência do sistema. O sistema executa ações corretivas automaticamente, reduzindo as atividades manuais que permitem intervenções humanas sujeitas a erros. A workload opera, atende às metas de disponibilidade e reduz a interrupção do serviço.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Para gerenciar alertas com eficiência e automatizar as respectivas respostas, categorize os alertas com base em sua criticidade e impacto, documente os procedimentos de resposta e planeje as respostas antes das tarefas de classificação.

Identifique tarefas que exigem ações específicas (geralmente detalhadas em runbooks) e examine todos os runbooks e playbooks para determinar as tarefas que podem ser automatizadas. Geralmente, se for possível definir ações, elas poderão ser automatizadas. Se não for possível automatizar as ações, documente as etapas manuais em um SOP e treine os operadores a respeito.

Conteste continuamente os processos manuais em busca de oportunidades de automação em que seja possível estabelecer e manter um plano para automatizar as respostas a alertas.

Etapas de implementação

1. Crie um inventário de alarmes: para obter uma lista de todos os alarmes, é possível usar a AWS CLI com o comando `describe-alarms` do Amazon CloudWatch. Dependendo de quantos alarmes você configurou, talvez seja necessário usar a paginação para recuperar um subconjunto de alarmes para cada chamada ou, alternativamente, você pode usar o AWS SDK para obter os alarmes usando uma chamada de API.
2. Documente todas as ações de alarme: atualize um runbook com todos os alarmes e suas ações, independentemente de serem manuais ou automatizados. O AWS Systems Manager fornece runbooks predefinidos. Para obter informações sobre como usar runbooks, consulte Trabalhado com runbooks. Para obter detalhes sobre como visualizar o conteúdo do runbook, consulte Visualizar o conteúdo do runbook.
3. Configure e gerencie ações de alarme: para qualquer um dos alarmes que exijam uma ação, especifique a ação automatizada usando o SDK do CloudWatch. Por exemplo, é possível alterar o estado das instâncias do Amazon EC2 automaticamente com base em um alarme do CloudWatch criando e ativando ações em um alarme ou desativando ações em um alarme.

   Também é possível usar o Amazon EventBridge para responder automaticamente a eventos do sistema, como problemas de disponibilidade de aplicações ou alterações de recursos. É possível criar regras para indicar os eventos de seu interesse e quais ações deverão ser executadas quando um evento corresponder a uma regra. As ações que podem ser iniciadas automaticamente incluem invocar uma função do AWS Lambda, invocar o `Run Command` do Amazon EC2, retransmitir o evento para o Amazon Kinesis Data Streams e consultar Automatizar o Amazon EC2 usando o EventBridge.
4. Procedimentos operacionais padrão (SOPs): com base nos componentes da aplicação, o AWS Resilience Hubrecomenda vários modelos de SOP. É possível usar esses SOPs para documentar todos os processos que um operador deve seguir caso um alerta seja emitido. Você também pode criar um SOP com base nas recomendações do Resilience Hub, onde você precisa de uma aplicação do Hub de Resiliência com uma política de resiliência associada, bem como uma avaliação histórica da resiliência em relação a essa aplicação. As recomendações para o SOP são produzidas pela avaliação de resiliência.

   O Hub de Resiliência trabalha com o Systems Manager para automatizar as etapas de seus SOPs, fornecendo vários documentos do SSM que você pode usar como base para esses SOPs.

Por exemplo, o Hub de Resiliência pode recomendar um SOP para adicionar espaço em disco com base em um documento de automação do SSM existente.

5. Execute ações automatizadas com o Amazon DevOps Guru: é possível usar o Amazon DevOps Guru para monitorar automaticamente recursos de aplicações em busca de comportamento anômalo e entregar recomendações direcionadas para acelerar os tempos de identificação e correção do problema. Com o DevOps Guru, você pode monitorar fluxos de dados operacionais quase em tempo real de várias fontes, incluindo métricas do Amazon CloudWatch, AWS Config, AWS CloudFormation e AWS X-Ray. Você também pode usar o DevOps Guru para criar automaticamente OpsItems no OpsCenter e enviar eventos para o EventBridge para automação adicional.

Recursos

Práticas recomendadas relacionadas:

- REL06-BP01 Monitorar todos os componentes da workload (geração)
- REL06-BP02 Definir e calcular métricas (agregação)
- REL06-BP03 Enviar notificações (processamento e alarmes em tempo real)
- REL08-BP01 Usar runbooks para atividades padrão, como implantação

Documentos relacionados:

- Automação do AWS Systems Manager
- Criar uma regra do EventBridge que seja acionada por um evento de um recurso da AWS
- Workshop One Observability
- Amazon Builders' Library: instrumentação de sistemas distribuídos para visibilidade operacional
- O que é o Amazon DevOps Guru?
- Trabalhar com documentos de automação (playbooks)

Vídeos relacionados:

- AWS re:Invent 2022: Práticas recomendadas de observabilidade na Amazon
- AWS re:Invent 2020: automatizar qualquer coisa com o AWS Systems Manager
- Introdução ao AWS Resilience Hub

- Criar sistemas de tickets personalizados para notificações do Amazon DevOps Guru
- Habilitar a agregação de insights de várias contas com o Amazon DevOps Guru

Exemplos relacionados:

- Workshops de confiabilidade
- Workshop do Amazon CloudWatch e Systems Manager

REL06-BP05 Analisar logs

Colete arquivos de log e históricos de métricas e analise-os para obter tendências mais abrangentes e informações sobre a workload.

O Amazon CloudWatch Logs Insights oferece suporte a uma linguagem de consulta simples, porém poderosa, que você pode usar para analisar dados de log. O Amazon CloudWatch Logs também oferece suporte a assinaturas que permitem que os dados fluam perfeitamente para o Amazon S3, onde você pode usar o ou o Amazon Athena para consultar os dados. Ele também oferece suporte a consultas em uma grande variedade de formatos. Para obter mais informações, consulte SerDes e formatos de dados compatíveis no Guia do usuário do Amazon Athena. Para análise de conjuntos enormes de arquivos de log, você pode executar um cluster do Amazon EMR para executar análises em escala de petabytes.

Existem várias ferramentas fornecidas por parceiros da AWS e terceiros que permitem agregação, processamento, armazenamento e estudo analítico. Essas ferramentas incluem New Relic, Splunk, Loggly, Logstash, CloudHealth e Nagios. Porém, a geração fora dos registros da aplicação e do sistema é única para cada provedor de nuvem e costuma ser única para cada serviço.

Uma parte do processo de monitoramento que costuma ser negligenciada é o gerenciamento de dados. Você precisa determinar os requisitos de retenção para monitorar os dados e então aplicar as políticas de ciclo de vida de acordo. O Amazon S3 é compatível com gerenciamento de ciclo de vida no nível do bucket do S3. Esse gerenciamento de ciclo de vida pode ser aplicado de modo diferente a diferentes caminhos no bucket. Mais perto do fim do ciclo de vida, você pode fazer a transição dos dados para o Amazon S3 Glacier para armazenamento de longo prazo e posterior expiração após o fim do período de retenção. A classe de armazenamento S3 Intelligent-Tiering foi projetada para otimizar custos movendo automaticamente dados para o nível de acesso mais econômico, sem impacto na performance ou sobrecarga operacional.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

- O CloudWatch Logs Insights permite pesquisar e analisar dados de log de modo interativo no Amazon CloudWatch Logs.
  - Analisar logs de dados com o CloudWatch Logs Insights
  - Consultas de exemplo do Amazon CloudWatch Logs Insights
- Use o Amazon CloudWatch Logs para enviar logs ao Amazon S3, no qual você pode usar o Amazon Athena para consultar os dados.
  - Como analiso meus logs de acesso ao servidor do Amazon S3 usando o Athena?
    - Crie uma política de ciclo de vida do S3 para o bucket de logs de acesso ao seu servidor. Configure a política de ciclo de vida para remover periodicamente os arquivos de log. Fazer isso reduz a quantidade de dados que o Athena analisa para cada consulta.
      - Como faço para criar uma política de ciclo de vida para um bucket do S3?

Recursos

Documentos relacionados:

- Consultas de exemplo do Amazon CloudWatch Logs Insights
- Analisar logs de dados com o CloudWatch Logs Insights
- Depurar com o Amazon CloudWatch Synthetics e o AWS X-Ray
- Como faço para criar uma política de ciclo de vida para um bucket do S3?
- Como analiso meus logs de acesso ao servidor do Amazon S3 usando o Athena?
- Workshop One Observability
- Amazon Builders' Library: instrumentação de sistemas distribuídos para visibilidade operacional

REL06-BP06 Realizar revisões regularmente

Revise frequentemente o modo como o monitoramento da workload está implementado e atualize-o com base em eventos e alterações significativos.

O monitoramento eficaz é orientado pelas principais métricas de negócios. Certifique-se de que essas métricas sejam acomodadas em sua workload à medida que as prioridades de negócios mudarem.

Auditar seu monitoramento ajuda a garantir que você saiba quando uma aplicação está atingindo as respectivas metas de disponibilidade. A análise de causa-raiz requer a capacidade de descobrir o que aconteceu quando as falhas ocorrem. A AWS fornece serviços que permitem rastrear o estado dos seus serviços durante um incidente:

- Amazon CloudWatch Logs: armazene seus logs neste serviço e inspecione seu conteúdo.
- Amazon CloudWatch Logs Insights: um serviço totalmente gerenciado que permite analisar logs massivos em segundos. Ele oferece consultas e visualizações rápidas e interativas.
- AWS Config: permite ver qual infraestrutura da AWS estava em uso em diferentes instantes.
- AWS CloudTrail: permite ver quais APIs da AWS foram invocadas a que horas e por qual entidade principal.

Na AWS, fazemos uma reunião semanal para revisar a performance operacional e compartilhar aprendizado entre as equipes. Como há inúmeras equipes na AWS, criamos o The Wheel para escolher aleatoriamente uma workload a ser analisada. Estabelecer um ritmo regular para análises de performance operacional e compartilhamento de conhecimento aprimora sua capacidade de obter uma performance superior de suas equipes operacionais.

Práticas comuns que devem ser evitadas:

- Coletar apenas as métricas padrão.
- Definir uma estratégia de monitoramento e nunca revisá-la.
- Não analisar o monitoramento quando alterações importantes são implantadas.

Benefícios de implementar esta prática recomendada: a revisão regular do monitoramento permite a antecipação de possíveis problemas, em vez de reagir a notificações quando um problema previsto realmente ocorrer.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

- Crie vários painéis para a workload. Você deve ter um painel superior com as principais métricas de negócios e as métricas técnicas identificadas como as mais relevantes à integridade projetada da workload conforme a variação do uso. Você também deve ter painéis para vários níveis e dependências da aplicação que podem ser inspecionados.
  - Usar painéis do Amazon CloudWatch

- Programe e realize revisões regulares dos painéis da workload. Realize uma inspeção regular dos painéis. É possível ter cadências diferentes para a profundidade de inspeção.
  - Inspecione as tendências nas métricas. Compare os valores das métricas com os valores históricos para ver se há tendências que possam indicar algo que precise ser investigado. Exemplos incluem: aumento da latência, diminuição da função principal de negócios e aumento das respostas a falhas.
  - Verifique se há pontos fora da curva ou anomalias em suas métricas. As médias ou os valores medianos podem mascarar pontos fora da curva e anomalias. Examine os valores mais altos e mais baixos durante o período e investigue as causas das pontuações extremas. À medida que você continua a eliminar essas causas, a redução da definição de extremo permite melhorar cada vez mais a consistência da performance da workload.
  - Procure mudanças bruscas no comportamento. Uma mudança imediata na quantidade ou na direção de uma métrica pode indicar que houve uma alteração na aplicação ou talvez você precise de fatores externos para adicionar outras métricas para rastrear.

Recursos

Documentos relacionados:

- Consultas de exemplo do Amazon CloudWatch Logs Insights
- Depurar com o Amazon CloudWatch Synthetics e o AWS X-Ray
- Workshop One Observability
- Amazon Builders' Library: instrumentação de sistemas distribuídos para visibilidade operacional
- Usar painéis do Amazon CloudWatch

REL06-BP07 Monitorar o rastreamento completo das solicitações por meio de seu sistema

Rastreie as solicitações à medida que elas são processadas por meio de componentes de serviço para que as equipes de produto possam analisar e depurar problemas com maior facilidade e melhorar a performance.

Resultado desejado: workloads com rastreamento abrangente em todos os componentes são fáceis de depurar, melhorando o tempo médio de resolução (MTTR) de erros e a latência ao simplificar a descoberta da causa-raiz. O rastreamento completo reduz o tempo necessário para descobrir os componentes afetados e detalhar as causas-raiz dos erros ou da latência.

Práticas comuns que devem ser evitadas:

- O rastreamento é usado para alguns componentes, mas não para todos. Por exemplo, sem rastrear o AWS Lambda, as equipes podem não entender claramente a latência causada por partidas a frio em uma workload com picos.
- Os canários sintéticos ou o monitoramento de usuários reais (RUM) não são configurados com rastreamento. Sem canários ou RUM, a telemetria de interação com o cliente é omitida da análise de rastreamento, gerando um perfil de performance incompleto.
- As workloads híbridas incluem ferramentas de rastreamento nativas da nuvem e de terceiros, mas ainda não foram tomadas medidas eletivas e integram totalmente uma única solução de rastreamento. Com base na solução de rastreamento escolhida, os SDKs de rastreamento nativos de nuvem devem ser usados para instrumentar componentes que não são nativos de nuvem ou ferramentas de terceiros devem ser configuradas para ingerir a telemetria de rastreamento nativa de nuvem.

Benefícios de implementar esta prática recomendada: quando as equipes de desenvolvimento são alertadas sobre problemas, elas podem ter uma visão completa das interações dos componentes do sistema, incluindo a correlação componente por componente com registros em log, performance e falhas. Como o rastreamento facilita a identificação visual das causas-raiz, menos tempo é gasto investigando-as. As equipes que entendem detalhadamente as interações dos componentes tomam decisões melhores e mais rápidas ao resolver problemas. Decisões como quando invocar o failover de recuperação de desastres (DR) ou onde melhor implementar estratégias de autorrecuperação podem ser aprimoradas com a análise de rastreamentos de sistemas e aumentar a satisfação do cliente com seus serviços.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

As equipes que operam aplicações distribuídas podem usar ferramentas de rastreamento para estabelecer um identificador de correlação, coletar rastreamentos de solicitações e criar mapas de serviço dos componentes conectados. Todos os componentes da aplicação devem ser incluídos nos rastreamentos de solicitações, incluindo clientes de serviços, gateways de middleware e barramentos de eventos, componentes computacionais e armazenamento, incluindo armazenamentos de chave-valor e bancos de dados. Inclua canários sintéticos e monitoramento de usuários reais em sua configuração de rastreamento completo a fim de medir as interações remotas com clientes e a latência para que você possa avaliar com precisão a performance de seus sistemas em relação aos seus objetivos e acordos de serviço.

Você pode usar os serviços de instrumentação do AWS X-Ray e do Monitoramento de aplicações do Amazon CloudWatch para fornecer uma visão completa das solicitações à medida que elas percorrem sua aplicação. O X-Ray coleta telemetria da aplicação e permite que você a visualize e filtre em cargas, funções, rastreamentos, serviços, APIs. Além disso, ele pode ser ativado para componentes do sistema sem código ou com pouco código. O monitoramento de aplicações do CloudWatch inclui o ServiceLens para integrar seus rastreamentos com métricas, logs e alarmes. O monitoramento de aplicações do CloudWatch também inclui sintéticos para monitorar seus endpoints e APIs, bem como monitoramento de usuários reais para instrumentar seus clientes de aplicações Web.

Etapas de implementação

- Use o AWS X-Ray em todos os serviços nativos compatíveis, como Amazon S3, AWS Lambda e Amazon API Gateway. Esses serviços da AWS permitem ao X-Ray alternar a configuração usando infraestrutura como código, AWS SDKs ou o AWS Management Console.
- Instrumente aplicações AWS Distro para Open Telemetry e X-Ray ou agentes de coleção de terceiros.
- Consulte o Guia do desenvolvedor da AWS X-Ray para obter informações sobre implementações específicas de linguagens de programação. Essas seções da documentação detalham como instrumentar solicitações HTTP, consultas SQL e outros processos específicos de sua linguagem de programação de aplicações.
- Use o rastreamento do X-Ray para canários do Amazon CloudWatch Synthetic e Amazon CloudWatch RUM para analisar o caminho da solicitação do cliente do usuário final ao longo de sua infraestrutura da AWS downstream.
- Configure métricas e alarmes do CloudWatch com base na integridade dos recursos e na telemetria canário para que as equipes sejam alertadas sobre problemas rapidamente e, depois, possam se aprofundar em rastreamentos e mapas de serviços com o ServiceLens.
- Ative a integração do X-Ray para ferramentas de rastreamento de terceiros, como Datadog, New Relic ou Dynatrace, se você estiver usando ferramentas de terceiros para sua solução de rastreamento principal.

Recursos

Práticas recomendadas relacionadas:

- REL06-BP01 Monitorar todos os componentes da workload (geração)
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas

Documentos relacionados:

- O que é AWS X-Ray?
- Amazon CloudWatch: monitoramento de aplicações
- Depurar com o Amazon CloudWatch Synthetics e o AWS X-Ray
- Amazon Builders' Library: instrumentação de sistemas distribuídos para visibilidade operacional
- Integrar o AWS X-Ray a outros serviços da AWS
- AWS Distro para OpenTelemetry e AWS X-Ray
- Amazon CloudWatch: usar monitoramento sintético
- Amazon CloudWatch: usar o CloudWatch RUM
- Configurar o canário do Amazon CloudWatch Synthetics e o alarme do Amazon CloudWatch
- Disponibilidade e além: como entender e melhorar a resiliência de sistemas distribuídos na AWS

Exemplos relacionados:

- Workshop One Observability

Vídeos relacionados:

- AWS re:Invent 2022: Como monitorar aplicações em várias contas
- Como monitorar suas aplicações da AWS

Ferramentas relacionadas:

- AWS X-Ray
- Amazon CloudWatch
- Amazon Route 53

REL 7. Como você projeta a workload para se adaptar às alterações na demanda?

Uma workload escalável fornece elasticidade para adicionar ou remover recursos automaticamente, de modo que eles correspondam perfeitamente à demanda atual em determinado momento.

Práticas recomendadas

- REL07-BP01: Usar automação ao obter ou escalar recursos

- REL07-BP02 Obter recursos após a detecção de danos em uma workload
- REL07-BP03 Obter recursos após determinar que mais recursos são necessários para uma workload
- REL07-BP04 Fazer o teste de carga da workload

REL07-BP01: Usar automação ao obter ou escalar recursos

Ao substituir recursos danificados ou escalar sua workload, automatize o processo por meio dos serviços gerenciados da AWS, como o Amazon S3 e o AWS Auto Scaling. Também é possível usar ferramentas de terceiros e os AWS SDKs para automatizar o ajuste de escala.

Os serviços gerenciados da AWS incluem Amazon S3, Amazon CloudFront, AWS Auto Scaling, AWS Lambda, Amazon DynamoDB, AWS Fargate e Amazon Route 53.

O AWS Auto Scaling permite detectar e substituir instâncias danificadas. Ele também permite criar planos de ajuste de escala para recursos, incluindo instâncias do Amazon EC2 e frotas spot, tarefas do Amazon ECS, tabelas e índices do Amazon DynamoDB e réplicas do Amazon Aurora.

Ao escalar instâncias do EC2, use várias zonas de disponibilidade (de preferência, pelo menos três) e adicione ou remova capacidade para manter o equilíbrio entre essas zonas de disponibilidade. As tarefas do ECS ou pods do Kubernetes (quando o Amazon Elastic Kubernetes Service é usado) também devem ser distribuídos em várias zonas de disponibilidade.

Ao usar o AWS Lambda, as instâncias são escaladas automaticamente. Sempre que uma notificação de evento é recebida para sua função, o AWS Lambda localiza rapidamente a capacidade livre dentro de sua frota de computação e executa seu código até a simultaneidade alocada. Você precisa garantir que a simultaneidade necessária esteja configurada no Lambda específico e em suas cotas de serviço.

O Amazon S3 escala automaticamente para lidar com altas taxas de solicitação. Por exemplo, a aplicação pode atingir, pelo menos, 3.500 solicitações PUT/POST/DELETE ou 5.500 solicitações GET/HEAD por segundo por prefixo em um bucket. Não há limite para o número de prefixos em um bucket. Você pode aumentar sua performance de leitura ou gravação paralelizando as leituras. Por exemplo, se você criar 10 prefixos em um bucket do Amazon S3 para paralelizar leituras, poderá escalar a performance de leitura para 55.000 solicitações de leitura por segundo.

Configure e use o Amazon CloudFront ou uma rede de entrega de conteúdo (CDN) confiável. Uma CDN pode fornecer tempos de resposta mais rápidos para o usuário final e atender às solicitações de conteúdo do cache, reduzindo assim a necessidade de escalar sua workload.

Práticas comuns que devem ser evitadas:

- Implementar grupos do Auto Scaling para autocorreção, mas não implementar elasticidade.
- Usar o ajuste de escala automático para responder a grandes aumentos no tráfego.
- Implantar aplicações com nível elevado de estado, eliminando a opção de elasticidade.

Benefícios de implementar esta prática recomendada: a automação elimina a possibilidade de erros manuais na implantação e na desativação de recursos. A automação remove o risco de custos excedentes e de negação de serviço decorrentes da lentidão na resposta às necessidades de implantação ou de desativação.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

- Configure e use o AWS Auto Scaling. Ele monitora suas aplicações e ajusta automaticamente a capacidade para manter uma performance estável e previsível com o menor custo possível. Ao usar o AWS Auto Scaling, você pode configurar o ajuste de escala da aplicação para vários recursos em diversos serviços.
  - O que é AWS Auto Scaling?
    - Configure o Auto Scaling nas instâncias e frotas Spot do Amazon EC2, nas tarefas do Amazon ECS, nas tabelas e nos índices do Amazon DynamoDB, nas réplicas do Amazon Aurora e nos appliances do AWS Marketplace, conforme aplicável.
      - Gerenciar a capacidade de throughput automaticamente com o ajuste de escala automático do DynamoDB
        - Use as operações de API de serviço para especificar alarmes, políticas de ajuste de escala e tempos de aquecimento e de resfriamento.
- Use o Elastic Load Balancing. Os balanceadores de carga podem distribuir a carga por caminho ou por conectividade de rede.
  - O que é Elastic Load Balancing?
    - Os Application Load Balancers podem distribuir a carga por caminho.
      - O que é um Application Load Balancer?
        - Configure um Application Load Balancer para distribuir o tráfego para workloads diferentes com base no caminho sob o nome de domínio.

        - Os Application Load Balancers podem ser usados para distribuir as cargas de maneira integrada ao AWS Auto Scaling a fim de gerenciar a demanda.
            - Usar um balanceador de carga com um grupo do Auto Scaling
    - Os Network Load Balancers podem distribuir a carga por conexão.
        - O que é um Network Load Balancer?
            - Configure um Network Load Balancer para distribuir o tráfego para workloads diferentes por meio do TCP ou para ter um conjunto constante de endereços IP para a workload.
            - Os Network Load Balancers podem ser usados para distribuir as cargas de maneira integrada ao AWS Auto Scaling a fim de gerenciar a demanda.
- Use um provedor DNS altamente disponível. Os nomes DNS permitem que os usuários insiram nomes, em vez de endereço IP, para acessar suas workloads e distribuem essas informações a um escopo definido, em geral, globalmente para usuários da workload.
    - Use o Amazon Route 53 ou um provedor de DNS confiável.
        - O que é o Amazon Route 53?
    - Use o Route 53 para gerenciar os balanceadores de carga e as distribuições do CloudFront.
        - Determine os domínios e subdomínios que serão gerenciados.
        - Crie conjuntos de registros adequados com os registros ALIAS ou CNAME.
            - Trabalhar com registros
- Use a rede global da AWS para otimizar o caminho dos seus usuários para suas aplicações. AWS O Global Accelerator monitora continuamente a integridade dos endpoints da aplicação e redireciona o tráfego para endpoints íntegros em menos de 30 segundos.
    - O AWS Global Accelerator é um serviço que melhora a disponibilidade e a performance de suas aplicações com usuários locais ou globais. Ele fornece endereços IP estáticos que atuam como um ponto de entrada fixo para os endpoints de sua aplicação em uma ou várias Regiões da AWS, como Application Load Balancers, Network Load Balancers ou instâncias do Amazon EC2.
        - O que é o AWS Global Accelerator?
- Configure e use o Amazon CloudFront ou uma rede de entrega de conteúdo (CDN) confiável. Uma rede de entrega de conteúdo pode fornecer tempos mais rápidos de resposta ao usuário final e atender a solicitações de conteúdo que podem causar ajuste de escala desnecessário das suas workloads.
    - O que é o Amazon CloudFront?

- Configure as distribuições do Amazon CloudFront para suas workloads ou use uma CDN de terceiros.
  - É possível limitar o acesso às suas workloads somente pelo CloudFront com o uso de intervalos de IPs para o CloudFront em seus grupos de segurança ou suas políticas de acesso de endpoint.

Recursos

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudar a criar soluções de computação automatizadas
- AWS Auto Scaling: como os planos de ajuste de escala funcionam
- AWS Marketplace: produtos que podem ser usados com o Auto Scaling
- Gerenciar a capacidade de throughput automaticamente com o ajuste de escala automático do DynamoDB
- Usar um balanceador de carga com um grupo do Auto Scaling
- O que é o AWS Global Accelerator?
- O que é o Amazon EC2 Auto Scaling?
- O que é AWS Auto Scaling?
- O que é o Amazon CloudFront?
- O que é o Amazon Route 53?
- O que é Elastic Load Balancing?
- O que é um Network Load Balancer?
- O que é um Application Load Balancer?
- Trabalhar com registros

REL07-BP02 Obter recursos após a detecção de danos em uma workload

Escale recursos de modo reativo quando necessário, se a disponibilidade for afetada, para restaurar a disponibilidade da workload.

Primeiro, você deve configurar as verificações de integridade e os critérios nessas verificações para indicar quando a disponibilidade é afetada pela falta de recursos. Notifique o pessoal apropriado para escalar manualmente o recurso ou inicie a automação para escalá-lo automaticamente.

A escala pode ser ajustada manualmente para a workload (por exemplo, alterando o número de instâncias do EC2 em um grupo do Auto Scaling ou modificando o throughput de uma tabela do DynamoDB por meio do AWS Management Console ou da AWS CLI). No entanto, a automação deve ser usada sempre que possível (consulte Usar automação ao obter ou escalar recursos).

Resultado desejado: as atividades de escalação (automática ou manual) são iniciadas para restaurar a disponibilidade após a detecção de uma falha ou degradação da experiência do cliente.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Implemente a observabilidade e o monitoramento em todos os componentes da workload para monitorar a experiência do cliente e detectar falhas. Defina os procedimentos, manuais ou automatizados, que escalam os recursos necessários. Para obter mais informações, consulte REL11-BP01 Monitorar todos os componentes da workload para detectar falhas.

Etapas de implementação

- Defina os procedimentos, manuais ou automatizados, que escalam os recursos necessários.
  - Os procedimentos de ajuste de escala dependem de como os diferentes componentes da workload são projetados.
  - Eles também variam dependendo da tecnologia subjacente utilizada.
    - Os componentes que usam o AWS Auto Scaling podem utilizar planos de ajuste de escala para configurar um conjunto de instruções para escalar os recursos. Se você usa o AWS CloudFormation ou adiciona tags a recursos da AWS, é possível configurar planos de ajuste de escala para diferentes conjuntos de recursos por aplicação. O Auto Scaling faz recomendações de estratégias de ajuste de escala personalizadas para cada recurso. Depois que você criar seu plano de ajuste de escala, o Auto Scaling combinará o ajuste de escala dinâmico com os métodos de escalabilidade preditiva para oferecer suporte à sua estratégia de ajuste de escala. Para obter mais detalhes, consulte Como os planos de ajuste de escala funcionam.
    - O Amazon EC2 Auto Scaling verifica se você tem o número correto de instâncias do Amazon EC2 disponíveis para processar a carga da sua aplicação. Você cria coleções de instâncias EC2, chamadas de grupos de Auto Scaling. É possível especificar o número mínimo e máximo de instâncias em cada grupo do Auto Scaling, e o Amazon EC2 Auto Scaling garantirá que o grupo nunca fique abaixo ou acima desses limites. Para obter mais detalhes, consulte O que é o Amazon EC2 Auto Scaling?

- O Auto Scaling do Amazon DynamoDB usa o serviço Application Auto Scaling para ajustar dinamicamente a capacidade de throughput provisionado em seu nome em resposta aos padrões de tráfego reais. Isso permite que uma tabela ou um índice secundário global aumente a capacidade provisionada de leitura e gravação para processar aumentos repentinos no tráfego, sem limitações. Para obter mais detalhes, consulte Gerenciar a capacidade de throughput automaticamente com o ajuste de escala automático do DynamoDB.

Recursos

Práticas recomendadas relacionadas:

- REL07-BP01 Usar automação ao obter ou escalar recursos
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas

Documentos relacionados:

- AWS Auto Scaling: como os planos de ajuste de escala funcionam
- Gerenciar a capacidade de throughput automaticamente com o ajuste de escala automático do DynamoDB
- O que é o Amazon EC2 Auto Scaling?

REL07-BP03 Obter recursos após determinar que mais recursos são necessários para uma workload

Escale os recursos proativamente para atender à demanda e evitar impacto na disponibilidade.

Muitos serviços da AWS são escalados automaticamente para atender à demanda. Se estiver usando instâncias do Amazon EC2 ou clusters do Amazon ECS, você poderá configurar o ajuste de escala automático desses clusters para que ocorra com base nas métricas de uso que correspondam à demanda da workload. Para o Amazon EC2, a utilização média da CPU, a contagem de solicitações do balanceador de carga ou a largura de banda da rede podem ser usadas para aumentar (ou reduzir) a escala horizontalmente das instâncias do EC2. Para o Amazon ECS, a utilização média da CPU, a contagem de solicitações do balanceador de carga e a utilização da memória podem ser usados para aumentar (ou reduzir) a escala horizontalmente das tarefas do ECS. Usando o ajuste de escala automático do destino na AWS, o Auto Scaler atua como um termostato doméstico, adicionando ou removendo recursos para manter o valor pretendido (por exemplo, 70% de utilização da CPU) que você especificar.

O Amazon EC2 Auto Scaling também pode fazer o ajuste de escala automático preditivo, que usa machine learning para analisar a workload histórica de cada recurso e prevê regularmente a carga futura para os próximos dois dias.

A Lei de Little ajuda a calcular quantas instâncias de computação (instâncias do EC2, funções simultâneas do Lambda, etc.) são necessárias.

$L = \lambda W$

L = número de instâncias (ou simultaneidade média no sistema)

λ = taxa média na qual as solicitações chegam (requisições por segundo)

W = tempo médio que cada solicitação gasta no sistema (s)

Por exemplo, a 100 rps, se cada solicitação demorar 0,5 segundos para ser processada, você precisará de 50 instâncias para acompanhar a demanda.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

- Obtenha recursos após a detecção de que mais recursos são necessários para uma workload. Escale os recursos proativamente para atender à demanda e evitar impacto na disponibilidade.
  - Calcule quantos recursos computacionais serão necessários (simultaneidade de computação) para atender a uma determinada taxa de solicitações.
    - Histórias sobre a Lei de Little
  - Quando você tiver um padrão histórico de uso, configure o ajuste de escala agendado para o Amazon EC2 Auto Scaling.
    - Ajuste de escala agendado para o Amazon EC2 Auto Scaling
  - Usar o ajuste de escala preditivo da AWS
    - Ajuste de escala preditivo para o Amazon EC2 Auto Scaling

Recursos

Documentos relacionados:

- AWS Marketplace: produtos que podem ser usados com o Auto Scaling

- Gerenciar a capacidade de throughput automaticamente com o ajuste de escala automático do DynamoDB
- Ajuste de escala preditivo para o EC2 com Machine Learning
- Ajuste de escala agendado para o Amazon EC2 Auto Scaling
- Histórias sobre a Lei de Little
- O que é o Amazon EC2 Auto Scaling?

REL07-BP04 Fazer o teste de carga da workload

Adote uma metodologia de teste de carga para avaliar se a ação de ajuste de escala atende aos requisitos da workload.

É importante realizar testes de carga sustentada. Os testes de carga devem descobrir o ponto de interrupção e testar a performance da workload. A AWS facilita a configuração de ambientes de teste temporários que modelam a escala de sua workload de produção. Na nuvem, é possível criar um ambiente de teste em escala de produção sob demanda, concluir seus testes e desativar os recursos. Como você paga somente pelo ambiente de teste quando está em execução, é possível simular seu ambiente ativo por uma fração do custo dos testes on-premises.

Os testes de carga em produção também devem ser considerados como parte dos game days em que o sistema de produção é destacado, durante horas de menor utilização do cliente, com todo o pessoal disponível para interpretar os resultados e resolver os problemas que surgirem.

Práticas comuns que devem ser evitadas:

- Executar testes de carga em implantações que não têm a mesma configuração da sua produção.
- Executar testes de carga apenas em componentes individuais da workload, e não nela toda.
- Executar testes de carga com um subconjunto de solicitações, e não com um conjunto representativo de solicitações reais.
- Executar testes de carga para um pequeno fator de segurança acima da carga esperada.

Benefícios de implementar esta prática recomendada: você sabe quais componentes em sua arquitetura falham sob carga e pode identificar as métricas que devem ser observadas para indicar que você está se aproximando dessa carga a tempo de resolver o problema, evitando o impacto dessa falha.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

- Realize testes de carga para identificar qual aspecto da workload indica que é necessário adicionar ou remover capacidade. Os testes de carga devem ter tráfego representativo semelhante ao que você recebe na produção. Aumente a carga enquanto observa as métricas que você preparou para determinar aquelas que indicam quando é necessário adicionar ou remover recursos.
  - Teste de carga distribuída na AWS: simular milhares de usuários conectados
    - Identifique a combinação de solicitações. Diversas combinações de solicitações são possíveis. Portanto, você deve examinar vários períodos ao identificar a combinação de tráfego.
    - Implemente um direcionador de carga. É possível usar código personalizado, código aberto ou um software comercial para implementar um direcionador de carga.
    - Faça o teste de carga inicialmente com uma pequena capacidade. Você percebe alguns efeitos imediatos ao direcionar a carga para uma capacidade menor, possivelmente tão pequena quanto uma instância ou um contêiner.
    - Faça o teste de carga com uma capacidade maior. Os efeitos serão diferentes em uma carga distribuída, portanto, recomenda-se testar o mais próximo possível de um ambiente de produto.

Recursos

Documentos relacionados:

- Teste de carga distribuída na AWS: simular milhares de usuários conectados
- Aplicações de teste de carga

Vídeos relacionados:

- AWS Summit ANZ 2023: Acelerar com confiança com o teste de carga distribuída da AWS

REL 8. Como você implementa a alteração?

As alterações controladas são necessárias para implantar novas funcionalidades e garantir que as workloads e o ambiente operacional executem softwares conhecidos e possam ser corrigidos ou substituídos de maneira previsível. Se essas alterações não forem controladas, será difícil prever o efeito dessas alterações ou resolver os problemas que surgem por causa delas.

Práticas recomendadas

- REL08-BP01 Usar runbooks para atividades padrão, como implantação
- REL08-BP02 Integrar testes funcionais como parte da sua implantação
- REL08-BP03 Integrar testes de resiliência como parte da implantação
- REL08-BP04 Implantar usando infraestrutura imutável
- REL08-BP05 Implantar alterações com automação

REL08-BP01 Usar runbooks para atividades padrão, como implantação

Os runbooks são os procedimentos predefinidos para alcançar um resultado específico. Use-os para executar atividades padrão, sejam elas feitas manual ou automaticamente. Os exemplos incluem a implantação de workloads, aplicação de patches a workloads ou a realização de modificações de DNS.

Por exemplo, coloque processos em vigor para garantir a segurança da reversão durante implantações. Garantir que você possa reverter uma implantação sem qualquer interrupção para seus clientes é essencial para tornar um serviço confiável.

Para procedimentos de runbooks, comece com um processo manual efetivo válido, implemente-o em código e acione a execução automatizada quando adequado.

Mesmo para workloads sofisticadas altamente automatizadas, os runbooks ainda são úteis para executar game days ou atender a requisitos rigorosos de relatórios e auditoria.

Observe que playbooks são usados em resposta a incidentes específicos, e runbooks são usados para alcançar resultados específicos. Muitas vezes, os runbooks são para atividades de rotina, enquanto os playbooks são usados para responder a eventos que não são rotineiras.

Práticas comuns que devem ser evitadas:

- Executar alterações não planejadas na configuração em produção.
- Ignorar as etapas do seu plano para agilizar a implantação, resultando em falha na implantação.
- Fazer alterações sem testar a possibilidade de reversão.

Benefícios de implementar esta prática recomendada: o planejamento eficaz da alteração aumenta sua capacidade de executá-la com êxito porque você está ciente de todos os sistemas afetados. A validação da alteração em ambientes de teste aumenta sua confiança.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

- Habilite respostas consistentes e rápidas para eventos bem conhecidos, documentando procedimentos nos runbooks.
  - AWS Well-Architected Framework: conceitos: runbook
- Use o princípio de infraestrutura como código para definir sua infraestrutura. Ao usar o AWS CloudFormation (ou um terceiro confiável) para definir sua infraestrutura, você pode usar software de controle de versão para controlar as versões e rastrear as alterações.
  - Use o AWS CloudFormation (ou um provedor terceirizado confiável) para definir sua infraestrutura.
    - O que é AWS CloudFormation?
  - Use bons princípios de design de software para criar modelos exclusivos e desacoplados.
    - Determine as permissões, os modelos e as partes responsáveis pela implementação.
      - Como controlar o acesso com o AWS Identity and Access Management
    - Use o controle de código-fonte, como o AWS CodeCommit ou uma ferramenta de terceiros confiável, para implementar o controle de versão.
      - O que é AWS CodeCommit?

Recursos

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudar a criar soluções de implantação automatizada
- AWS Marketplace: produtos que podem ser usados para automatizar suas implantações
- AWS Well-Architected Framework: conceitos: runbook
- O que é AWS CloudFormation?
- O que é AWS CodeCommit?

Exemplos relacionados:

- Automatizar operações com playbooks e runbooks

REL08-BP02 Integrar testes funcionais como parte da sua implantação

Os testes funcionais são executados como parte da implantação automatizada. Se os critérios de êxito não forem atendidos, o pipeline será interrompido ou revertido. Esses testes são executados em um ambiente de pré-produção, que é preparado antes da produção no pipeline. Idealmente, isso é feito como parte de um pipeline de implantação.

Resultado desejado: você usa a automação para realizar testes funcionais, e os dados de teste associados reduzem a duração e as despesas dos testes e melhoram a precisão dos resultados. Você integra testes funcionais como parte do processo de implantação, o que ajuda a automatizar os canais de lançamento para atualizações rápidas e confiáveis de aplicações e infraestrutura.

Práticas comuns que devem ser evitadas:

- Você realiza testes manualmente fora do pipeline de implantação.
- Você ignora as etapas de teste na automação por meio de fluxos de trabalho manuais de emergência.
- Você não segue os planos e os processos de teste estabelecidos em favor de cronogramas acelerados.

Benefícios de implementar esta prática recomendada: os testes funcionais validam que o sistema opera de acordo com os requisitos especificados. Ele é usado para verificar de forma consistente a ordem de funcionamento pretendida dos componentes, como interfaces de usuário, APIs, bancos de dados e código-fonte. Quando você examina esses componentes do sistema, os testes funcionais verificam se cada recurso se comporta conforme o esperado, o que protege as expectativas do usuário e a integridade do software. Integre testes funcionais como parte da implantação regular e use a automação para implantar todas as mudanças, o que reduz a possibilidade de introdução de erros humanos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Integre testes funcionais como parte da implantação. Os testes funcionais são executados como parte da implantação automatizada. Se os critérios de sucesso não forem atendidos, o pipeline será interrompido ou revertido. O AWS CodePipeline fornece um canal de entrega contínua para testes automatizados, o que permite que os testadores automatizem todo o processo de teste e implantação. Ele se integra aos serviços da AWS, como o AWS CodeBuild e o AWS CodeDeploy,

para automatizar as fases de compilação, teste e implantação do ciclo de vida de desenvolvimento de software.

Etapas de implementação

- Configure seu pipeline: configure seus estágios de código-fonte, compilação, teste e implantação usando o console do AWS CodePipeline ou a AWS Command Line Interface (CLI).
  - Defina seu código-fonte: com o AWS CodePipeline, é possível recuperar automaticamente o código-fonte de sistemas de controle de versão como GitHub, AWS CodeCommit ou Bitbucket, o que verifica se o código mais recente é sempre usado para testes.
  - Automatize compilações e testes: o AWS CodeBuild pode compilar e testar automaticamente seu código e gerar relatórios de teste. Ele comporta frameworks de teste conhecidos, como JUnit, NUnit e TestNG.
  - Implante seu código: depois que o código for compilado e testado, o AWS CodeDeploy poderá implantá-lo em seu ambiente de teste, incluindo instâncias do Amazon EC2, funções do AWS Lambda ou servidores on-premises.
  - Monitore os pipelines: o AWS CodePipeline pode rastrear o progresso do seu pipeline e o status de cada estágio. É possível usar verificações de qualidade para bloquear o pipeline de acordo com o status de execução do teste. Você também pode receber notificações sobre qualquer falha ou conclusão do estágio do pipeline.

Recursos

Documentos relacionados:

- Usar o AWS CodePipeline com o AWS CodeBuild para testar código e executar compilações
- Registrar em log e monitorar no AWS CodeBuild
- Indicadores para testes funcionais

REL08-BP03 Integrar testes de resiliência como parte da implantação

Integre os testes de resiliência introduzindo falhas conscientemente no sistema para medir a capacidade em caso de cenários disruptivos. Os testes de resiliência são diferentes dos testes de unidade e de função que geralmente são integrados aos ciclos de implantação, pois focam a identificação de falhas imprevistas no sistema. Embora seja seguro começar com a integração dos testes de resiliência na pré-produção, defina uma meta para implementar esses testes na produção como parte dos game days.

Resultado desejado: o teste de resiliência ajuda a aumentar a confiança na capacidade do sistema de resistir à degradação na produção. Experimentos indicam pontos fracos que podem causar falha, o que ajuda a melhorar o sistema para mitigar falhas e degradações de forma automática e eficiente.

Práticas comuns que devem ser evitadas:

- Falta de observabilidade e monitoramento nos processos de implantação
- Dependência em pessoas para resolver falhas do sistema
- Mecanismos de análise de baixa qualidade
- Foco em problemas conhecidos de um sistema e ausência de experimentação para identificar quaisquer problemas desconhecidos
- Identificação de falhas, mas sem resolução
- Nenhuma documentação de descobertas e runbooks

Benefícios de estabelecer as práticas recomendadas: os testes de resiliência integrados em suas implantações ajudam a identificar problemas desconhecidos no sistema que, de outra forma, passariam despercebidos e poderiam levar à inatividade da produção. A identificação de problemas desconhecidos em um sistema ajuda você a documentar descobertas, integrar testes ao processo de CI/CD e criar runbooks, o que simplifica a mitigação por meio de mecanismos eficientes e reproduzíveis.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

As formas de teste de resiliência mais comuns que podem ser integradas às implantações do sistema são a recuperação de desastres e a engenharia do caos.

- Inclua atualizações nos planos de recuperação de desastres e nos procedimentos operacionais padrão (SOPs) em qualquer implantação significativa.
- Integre testes de confiabilidade aos canais de implantação automatizados. Serviços como AWS Resilience Hub podem ser integrados ao seu pipeline de CI/CD para estabelecer avaliações contínuas de resiliência que são avaliadas automaticamente como parte de cada implantação.
- Defina as aplicações no AWS Resilience Hub. As avaliações de resiliência geram trechos de código que ajudam você a criar procedimentos de recuperação como documentos do AWS Systems Manager para as aplicações e fornecem uma lista de monitores e alarmes recomendados do Amazon CloudWatch.

- Depois que os planos de DR e SOPs forem atualizados, conclua os testes de recuperação de desastres para verificar se eles são eficazes. O teste de recuperação de desastres ajuda a determinar se você pode restaurar o sistema após um evento e retornar às operações normais. Você pode simular várias estratégias de recuperação de desastres e identificar se seu planejamento é suficiente para atender às necessidades de disponibilidade. As estratégias comuns de recuperação de desastres incluem backup e restauração, luz piloto, espera fria, standby passivo, standby a quente e ativo-ativo, e todas elas diferem em custo e complexidade. Antes do teste de recuperação de desastres, recomendamos definir o objetivo de tempo de recuperação (RTO) e o objetivo de ponto de recuperação (RPO) para simplificar a escolha da estratégia a ser simulada. A AWS oferece ferramentas de recuperação de desastres como a AWS Elastic Disaster Recoveryque ajudam você a começar a planejar e testar.
- Os experimentos de engenharia do caos introduzem interrupções no sistema, como interrupções na rede e falhas no serviço. Ao simular com falhas controladas, você pode descobrir as vulnerabilidades do sistema e, ao mesmo tempo, conter os impactos das falhas injetadas. Assim como as outras estratégias, execute simulações de falhas controladas em ambientes de não produção usando serviços como AWS Fault Injection Service para ganhar confiança antes da implantação na produção.

Recursos

Documentos relacionados:

- Experimentar com falhas usando testes de resiliência para aumentar a preparação para a recuperação
- Avaliar continuamente a resiliência da aplicação com o AWS Resilience Hub e o AWS CodePipeline
- Arquitetura de recuperação de desastres (DR) na AWS, Parte I: estratégias de recuperação na nuvem
- Verificar a resiliência de suas workloads via engenharia do caos
- Princípios da engenharia do caos
- Workshop de engenharia do caos

Vídeos relacionados:

- AWS re:Invent 2020: Testar a resiliência via engenharia do caos
- Melhorar a resiliência de aplicações com o AWS Fault Injection Service

- Prepare e proteja suas aplicações contra interrupções com o AWS Resilience Hub

REL08-BP04 Implantar usando infraestrutura imutável

A infraestrutura imutável é um modelo que não requer atualizações, patches de segurança ou alterações na configuração no local nas workloads de produção. Quando uma alteração é necessária, a arquitetura é criada em uma nova infraestrutura e implantada na produção.

Siga uma estratégia de implantação de infraestrutura imutável para aumentar a confiabilidade, a consistência e a reprodutibilidade nas implantações de workload.

Resultado desejado: com uma infraestrutura imutável, nenhuma modificação no local é permitida para executar recursos de infraestrutura em uma workload. Em vez disso, quando uma alteração é necessária, um novo conjunto de recursos atualizados da infraestrutura que contém todas as alterações necessárias é implantado paralelamente aos recursos existentes. Essa implantação é validada automaticamente e, se bem-sucedida, o tráfego é gradualmente transferido para o novo conjunto de recursos.

Essa estratégia de implantação aplica-se a atualizações de software, patches de segurança, alterações na infraestrutura, atualizações de configuração e atualizações de aplicações, entre outros.

Práticas comuns que devem ser evitadas:

- Implementação de mudanças no local em recursos da infraestrutura em execução.

Benefícios de implementar esta prática recomendada:

- Maior consistência em todos os ambientes: como não há diferenças nos recursos de infraestrutura entre os ambientes, a consistência aumenta e os testes são simplificados.
- Redução nos desvios de configuração: ao substituir recursos de infraestrutura por uma configuração conhecida e com controle de versão, a infraestrutura é redefinida para um estado conhecido, testado e confiável, evitando assim desvios de configuração.
- Implantações atômicas confiáveis: as implantações são concluídas com sucesso ou nada muda, aumentando a consistência e a confiabilidade no processo de implantação.
- Implantações simplificadas: as implantações são simplificadas porque não precisam oferecer suporte a atualizações. As atualizações são apenas novas implantações.

- Implantações mais seguras com processos de reversão e recuperação rápidos: as implantações são mais seguras porque a versão de trabalho anterior não foi alterada. Em caso de erros, é possível reverter para ela.
- Postura de segurança aprimorada: ao não permitir alterações na infraestrutura, os mecanismos de acesso remoto (como o SSH) podem ser desabilitados. Isso reduz o vetor de ataque, melhorando a postura de segurança da organização.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Automação

Ao definir uma estratégia de implantação de infraestrutura imutável, é recomendável usar a automação o máximo possível para aumentar a reprodutibilidade e minimizar o potencial de erro humano. Para obter mais detalhes, consulte REL08-BP05 Implantar alterações com automação e Automatizar implantações seguras e sem intervenção manual.

Com a infraestrutura como código (IaC), as etapas de provisionamento, orquestração e implantação da infraestrutura são definidas de forma programática, descritiva e declarativa e armazenadas em um sistema de controle de código-fonte. O uso da infraestrutura como código simplifica a automatização da implantação da infraestrutura e ajuda a obter a imutabilidade da infraestrutura.

Padrões de implantação

Quando uma mudança na workload é necessária, a estratégia de implantação da infraestrutura imutável exige que um novo conjunto de recursos da infraestrutura seja implantado, incluindo todas as alterações necessárias. É importante que esse novo conjunto de recursos siga um padrão de implantação que minimize o impacto sobre o usuário. Há duas estratégias principais para essa implantação:

Implantação canário: a prática de direcionar um pequeno número de seus clientes para a nova versão, geralmente em execução em uma única instância de serviço (o canário). Em seguida, você examina profundamente todas as alterações de comportamento ou erros gerados. O tráfego poderá ser removido da implantação canário se problemas críticos forem encontrados, e os usuários poderão ser enviados de volta para a versão anterior. Se a implantação for bem-sucedida, você poderá continuar implantando na velocidade desejada enquanto monitora as alterações em busca de erros até a implantação ser concluída. O AWS CodeDeploy pode ser configurado com uma configuração de implantação que permita uma implantação canário.

Implantação azul/verde: semelhante à implantação canário, exceto que uma frota completa da aplicação é implantada em paralelo. Você alterna as implantações entre as duas pilhas (azul e verde). Novamente, é possível enviar o tráfego para a nova versão e voltar para a versão antiga se houver problemas na implantação. Normalmente, todo o tráfego é chaveado de uma só vez. No entanto, você também pode usar frações do tráfego para cada versão para aumentar a adoção da nova versão usando os recursos de roteamento de DNS ponderado do Amazon Route 53. O AWS CodeDeploy e o AWS Elastic Beanstalk podem ser definidos com uma configuração de implantação que permitirá uma implantação azul/verde.

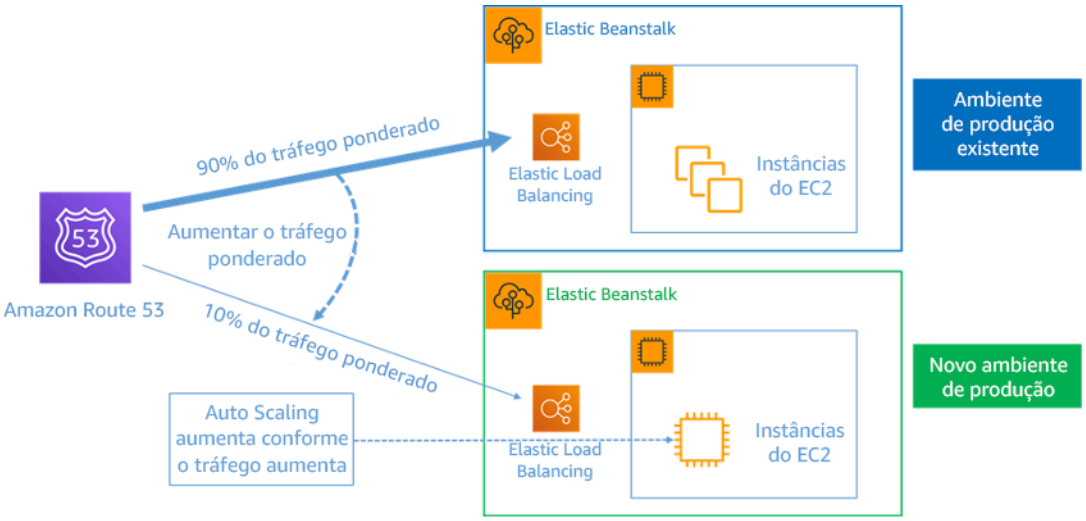


Figura 8: Implantação azul/verde com o AWS Elastic Beanstalk e o Amazon Route 53

Detecção de desvios

Um desvio é definido como qualquer alteração que faz com que um recurso de infraestrutura tenha um estado ou configuração diferente do esperado. Alterações não gerenciadas da configuração, sejam de que tipo for, são contrárias ao conceito de infraestrutura imutável e devem ser detectadas e corrigidas para que a implementação da infraestrutura imutável seja bem-sucedida.

Etapas de implementação

- Proíba a modificação no local dos recursos de infraestrutura em execução.
  - É possível usar o AWS Identity and Access Management (IAM) para especificar quem ou o que pode acessar serviços e recursos na AWS, gerenciar centralmente permissões refinadas e analisar o acesso para refinar permissões na AWS.
- Automatize a implantação dos recursos da infraestrutura para aumentar a reprodutibilidade e minimizar a possibilidade de erro humano.

- Conforme descrito em Introdução a DevOps no whitepaper da AWS, a automação é a base dos serviços da AWS e possui suporte interno em todos os serviços, recursos e ofertas.
- Pré-preparar a imagem de máquina da Amazon (AMI) pode acelerar o tempo de lançamento. O EC2 Image Builder é um serviço da AWS totalmente gerenciado que ajuda a automatizar a criação, a manutenção, a validação, o compartilhamento e a implantação de AMIs personalizadas do Linux ou do Windows.
- Alguns dos serviços compatíveis com automação são:
  - O AWS Elastic Beanstalk é um serviço para implantação e escalação rápidas de aplicações Web desenvolvidas com Java, .NET, PHP, Node.js, Python, Ruby, Go e Docker em servidores familiares, como Apache, NGINX, Passenger e IIS.
  - O AWS Proton ajuda as equipes de plataforma a conectar e coordenar todas as diferentes ferramentas que suas equipes de desenvolvimento precisam para provisionamento de infraestrutura, implantação de código, monitoramento e atualizações. O AWS Proton torna possível a infraestrutura automatizada na forma de provisionamento de código e implantação de aplicações de tecnologia sem servidor baseadas em contêiner.
- A utilização da infraestrutura como código facilita a automatização da implantação da infraestrutura e ajuda a obter a imutabilidade da infraestrutura. A AWS fornece serviços que permitem a criação, a implantação e a manutenção da infraestrutura de forma programática, descritiva e declarativa.
  - O AWS CloudFormation ajuda os desenvolvedores a criar recursos da AWS de forma ordenada e previsível. Os recursos são escritos em arquivos de texto usando o formato JSON ou YAML. Os modelos exigem sintaxe e estrutura específicas que dependem dos tipos de recurso que estão sendo criados e gerenciados. Você cria os recursos em JSON ou YAML com qualquer editor de código, como o AWS Cloud9, e os insere em um sistema de controle de versão, e o CloudFormation cria os serviços especificados de maneira segura e repetível.
  - O AWS Serverless Application Model (AWS SAM) é uma estrutura de código aberto que você pode usar para criar aplicações sem servidor na AWS. O AWS SAM se integra a outros serviços da AWS e é uma extensão do AWS CloudFormation.
  - O AWS Cloud Development Kit (AWS CDK) é um framework de desenvolvimento de software de código aberto para modelar e provisionar recursos de aplicações em nuvem usando linguagens de programação conhecidas. É possível usar o AWS CDK para modelar a infraestrutura de aplicações usando TypeScript, Python, Java e .NET. O AWS CDK usa o AWS CloudFormation em segundo plano para provisionar recursos de forma segura e repetível.

    - O AWS Cloud Control API apresenta um conjunto comum de APIs de criação, leitura, atualização, exclusão e lista (CRUDL) para ajudar os desenvolvedores a gerenciar sua infraestrutura de nuvem de forma fácil e consistente. As APIs comuns do Cloud Control permitem que os desenvolvedores gerenciem de maneira uniforme o ciclo de vida de serviços da AWS e de terceiros.
- Implemente padrões de implantação que minimizem o impacto no usuário.
  - Implantações canário:
    - Configurar uma implantação de versão canário do API Gateway
    - Criar um pipeline com implantações canário para o Amazon ECS usando o AWS App Mesh
  - Implantações azuis/verdes: Implantações azuis/verdes no whitepaper da AWS descreve exemplos de técnicas para implementar estratégias de implantação azul/verde.
- Detecte variações de configuração ou estado. Para obter mais informações, consulte Como detectar alterações de configuração não gerenciadas para pilhas e recursos.

Recursos

Práticas recomendadas relacionadas:

- REL08-BP05 Implantar alterações com automação

Documentos relacionados:

- Automatizar implantações seguras e sem intervenção
- Como utilizar o AWS CloudFormation para criar uma infraestrutura imutável no Nubank
- Infraestrutura como código
- Implementar um alarme para detectar automaticamente desvios nas pilhas do AWS CloudFormation

Vídeos relacionados:

- AWS re:Invent 2020: Confiabilidade, consistência e confiança por meio da imutabilidade

REL08-BP05 Implantar alterações com automação

As implantações e a aplicação de patches são automatizadas para eliminar impactos negativos.

As alterações nos sistemas de produção são uma das maiores áreas de risco para muitas organizações. Consideramos as implantações um problema de primeira classe a ser resolvido junto com os problemas de negócios abordados pelo software. Atualmente, isso significa usar a automação nas operações sempre que for viável, incluindo testar e implantar alterações, adicionar ou remover capacidade e migrar dados.

Resultado desejado: você incorpora segurança de implantação automatizada no processo de lançamento com testes extensivos de pré-produção, reversões automáticas e implantações de produção em etapas. Essa automação minimiza o impacto potencial na produção causado por falhas nas implantações, e os desenvolvedores não precisam mais monitorar ativamente as implantações na produção.

Práticas comuns que devem ser evitadas:

- Você implementa alterações manuais.
- Você ignora etapas na automação por meio de fluxos de trabalho manuais de emergência.
- Você não segue os planos e os processos estabelecidos em favor de cronogramas acelerados.
- Você executa implantações subsequentes rápidas sem permitir o tempo de incorporação.

Benefícios de implementar esta prática recomendada: ao usar a automação para implantar todas as alterações, você remove o potencial de introdução de erros humanos e fornece a capacidade de testar antes de alterar a produção. A execução desse processo antes do início da produção verifica se os planos estão concluídos. Além disso, a reversão automática no processo de liberação pode identificar problemas de produção e retornar a workload ao estado operacional anterior.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Automatize o pipeline de implantação. Os pipelines de implantação permitem invocar testes automatizados e detecção de anomalias. Além disso, eles interrompem o pipeline em uma determinada etapa antes da implantação em produção ou revertem automaticamente uma alteração. Uma parte integral disso é a adoção da cultura de integração contínua e entrega/implantação contínuas (CI/CD) em que uma confirmação ou alteração de código passa por vários estágios automatizados, desde os estágios de construção e teste até a implantação em ambientes de produção.

Embora o bom senso convencional sugira que você mantenha as pessoas informadas para os procedimentos operacionais mais difíceis, sugerimos automatizar esses procedimentos exatamente por isso.

Etapas de implementação

É possível automatizar as implantações para remover as operações manuais seguindo estas etapas:

- Configure um repositório de código para armazenar seu código com segurança: use o AWS CodeCommit para criar um repositório seguro baseado em Git.
- Configure um serviço de integração contínua para compilar seu código-fonte, executar testes e criar artefatos de implantação: para configurar um projeto de compilação para essa finalidade, consulte Introdução ao AWS CodeBuild usando o console.
- Configure um serviço de implantação que automatize as implantações de aplicações e gerencie a complexidade das atualizações de aplicações sem depender de implantações manuais propensas a erros: o AWS CodeDeploy automatiza as implantações de software em toda uma variedade de serviços computacionais, como Amazon EC2, AWS Fargate, AWS Lambda e seus servidores on-premises. Para configurar essas etapas, consulte Introdução ao CodeDeploy.
- Configure um serviço de entrega contínua que automatize os pipelines de lançamento para atualizações rápidas e confiáveis de aplicações e infraestrutura: considere usar o AWS CodePipeline para obter ajuda para automatizar os pipelines de lançamento. Para obter mais detalhes, consulte os tutoriais do CodePipeline.

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP04 Usar sistemas de gerenciamento de compilação e implantação
- OPS05-BP10 Automatizar totalmente a integração e a implantação
- OPS06-BP02 Testar as implantações
- OPS06-BP04 Automatizar os testes e a reversão

Documentos relacionados:

- Entrega contínua de pilhas do AWS CloudFormation aninhadas usando o AWS CodePipeline
- CI/CD completo com AWS CodeCommit, AWS CodeBuild, AWS CodeDeploy e AWS CodePipeline
- Parceiro da APN: parceiros que podem ajudar a criar soluções de implantação automatizada

- AWS Marketplace: produtos que podem ser usados para automatizar suas implantações
- Automatizar mensagens de chat com webhooks
- Amazon Builders' Library: como garantir a segurança da reversão durante implantações
- Amazon Builders' Library: como aumentar a velocidade com a entrega contínua
- O que é o AWS CodePipeline?
- O que é CodeDeploy?
- Gerenciador de patches do AWS Systems Manager
- O que é o Amazon SES?
- O que é o Amazon Simple Notification Service?

Vídeos relacionados:

- AWS Summit 2019: CI/CD na AWS

Gerenciamento de falhas

Perguntas

- REL 9. Como você faz backup dos dados?
- REL 10. Como você usa o isolamento de falhas para proteger a workload?
- REL 11. Como projetar a workload para resistir a falhas de componentes?
- REL 12. Como testar a confiabilidade?
- REL 13. Como planejar para a recuperação de desastres (DR)?

REL 9. Como você faz backup dos dados?

Faça backup de dados, aplicações e configurações para atender às suas necessidades de objetivos de tempo de recuperação (RTO) e objetivos de ponto de recuperação (RPO).

Práticas recomendadas

- REL09-BP01 Identificar e fazer backup de todos os dados que precisam de backup ou reproduzir os dados das fontes
- REL09-BP02 Proteger e criptografar backups
- REL09-BP03 Realizar backups de dados automaticamente

- REL09-BP04 Realizar a recuperação periódica dos dados para verificar a integridade e os processos de backup

REL09-BP01 Identificar e fazer backup de todos os dados que precisam de backup ou reproduzir os dados das fontes

Compreenda e use os recursos de backup dos serviços e recursos de dados usados pela workload. A maioria dos serviços oferece recursos para fazer backup dos dados da workload.

Resultado desejado: as fontes de dados foram identificadas e classificadas com base na criticidade. Depois, estabeleça uma estratégia de recuperação de dados com base no RPO. A estratégia envolve fazer backup dessas fontes de dados ou poder reproduzir dados de outras fontes. Em caso de perda de dados, a estratégia implementada permite a recuperação ou reprodução de dados dentro do RPO e RTO definidos.

Fase de maturidade da nuvem: fundamental

Práticas comuns que devem ser evitadas:

- Não estar ciente de todas as fontes de dados para a workload e sua criticidade.
- Não fazer backups de fontes de dados essenciais.
- Fazer backups apenas de algumas fontes de dados sem usar a criticidade como critério.
- Não ter um RPO definido ou a frequência de backup não atender ao RPO.
- Não avaliar a necessidade de um backup ou se os dados podem ser reproduzidos de outras fontes.

Benefícios de implementar esta prática recomendada: identificar os locais onde os backups são necessários e implementar um mecanismo para criar backups, ou ser capaz de reproduzir os dados de uma fonte externa, melhora a capacidade de restaurar e recuperar dados durante uma interrupção.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Todos os datastores da AWS oferecem recursos de backup. Serviços como o Amazon RDS e o Amazon DynamoDB oferecem suporte adicional ao backup automatizado que possibilita a recuperação para um ponto no tempo (PITR), permitindo restaurar um backup a qualquer momento

até cinco minutos ou menos antes da hora atual. Muitos serviços da AWS oferecem a capacidade de copiar backups para outra Região da AWS. O AWS Backup é uma ferramenta que permite centralizar e automatizar a proteção de dados em todos os serviços da AWS. O AWS Elastic Disaster Recovery permite copiar workloads completas do servidor e manter a proteção contínua dos dados no local, entre AZ ou entre regiões, com um objetivo de ponto de recuperação (RPO) medido em segundos.

O Amazon S3 pode ser usado como um destino de backup para fontes de dados autogerenciadas e gerenciadas pela AWS. Serviços da AWS, como o Amazon EBS, o Amazon RDS e o Amazon DynamoDB, oferecem recursos integrados de criação de backups. É possível também usar um software de backup de terceiros.

Os dados on-premises podem ser copiados para a Nuvem AWS via AWS Storage Gateway ou AWS DataSync. Os buckets do Amazon S3 podem ser usados para armazenar esses dados na AWS. O Amazon S3 oferece vários níveis de armazenamento, como Amazon S3 Glacier ou S3 Glacier Deep Archive para reduzir o custo do armazenamento de dados.

É possível atender às necessidades de recuperação de dados reproduzindo os dados de outras fontes. Por exemplo, os nós de réplica do Amazon ElastiCache ou as réplicas de leitura do Amazon RDS poderão ser usados para reproduzir dados se os dados primários forem perdidos. Nos casos em que fontes como essa podem ser usadas para atender ao objetivo de ponto de recuperação (RPO) e ao objetivo de tempo de recuperação (RTO), talvez um backup não seja necessário. Outro exemplo, se estiver trabalhando com o Amazon EMR, é que talvez não seja necessário fazer backup do seu datastore HDFS, desde que você consiga reproduzir os dados no Amazon EMR a partir do Amazon S3.

Ao selecionar uma estratégia de backup, considere o tempo necessário para recuperar os dados. Ele depende do tipo de backup (no caso de uma estratégia de backup) ou da complexidade do mecanismo de reprodução de dados. O tempo deve respeitar o RTO para a workload.

Etapas de implementação

1. Identifique todas as fontes de dados para a workload. Os dados podem ser armazenados em vários recursos, como bancos de dados, volumes, sistemas de arquivos, sistemas de registro em log e armazenamento de objetos. Consulte a seção Recursos para encontrar Documentos relacionados sobre os diferentes serviços da AWS em que os dados são armazenados e sobre a capacidade de backup oferecida por esses serviços.
2. Classifique as fontes de dados com base na criticidade. Diferentes conjuntos de dados terão diferentes níveis de criticidade para uma workload e, portanto, diferentes requisitos de resiliência.

Por exemplo, alguns dados podem ser críticos e exigir um RPO próximo de zero, enquanto outros dados podem ser menos críticos e tolerar um RPO mais alto e a perda de alguns dados. Da mesma forma, diferentes conjuntos de dados também podem ter diferentes requisitos de RTO.

3. Use serviços da AWS serviços ou de terceiros para criar backups dos dados. O AWS Backup é um serviço gerenciado que permite criar backups de várias fontes de dados na AWS. O AWS Elastic Disaster Recovery cuida da replicação automatizada de dados em menos de um segundo para uma Região da AWS. A maioria dos serviços da AWS também possui recursos nativos para criar backups. O AWS Marketplace tem muitas soluções que também fornecem esses recursos. Consulte os Recursos listados abaixo para obter informações sobre como criar backups de dados de vários serviços da AWS.
4. Para dados sem backup, estabeleça um mecanismo de reprodução de dados. Você pode optar por não fazer backup dos dados que podem ser reproduzidos de outras fontes por vários motivos. Às vezes, pode ser mais barato reproduzir dados de fontes se necessário, em vez de criar um backup, pois pode haver um custo associado ao armazenamento de backups. Outro exemplo é quando a restauração de um backup demora mais do que a reprodução dos dados das fontes, resultando em uma violação no RTO. Nestas situações, considere concessões e estabeleça um processo bem definido de como os dados podem ser reproduzidos dessas fontes quando a recuperação de dados for necessária. Se você tiver carregado dados do Amazon S3 para um data warehouse (como o Amazon Redshift) ou cluster MapReduce (como o Amazon EMR) para fazer análises nesses dados, isso pode ser um exemplo de dados que podem ser reproduzidos de outras fontes. Desde que os resultados dessas análises sejam armazenados em algum lugar ou reproduzíveis, você não sofreria uma perda de dados devido a uma falha no data warehouse ou no cluster do MapReduce. Outros exemplos que podem ser reproduzidos de origens incluem caches (como o Amazon ElastiCache) ou réplicas de leitura do RDS.
5. Estabeleça uma cadência para fazer backup dos dados. A criação de backups de fontes de dados é um processo periódico, e a frequência deve depender do RPO.

Nível de esforço do plano de implementação: Moderado

Recursos

Práticas recomendadas relacionadas:

REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados

REL13-BP02 Usar estratégias de recuperação definidas para cumprir os objetivos de recuperação

Documentos relacionados:

- O que é o AWS Backup?
- O que é o AWS DataSync?
- O que é Gateway de Volumes?
- Parceiro da APN: parceiros que podem ajudar com o backup
- AWS Marketplace: produtos que podem ser usados para backup
- Snapshots do Amazon EBS
- Fazer backup do Amazon EFS
- Fazer backup do Amazon FSx para Windows File Server
- Backup e restauração do ElastiCache para Redis
- Criar um snapshot de cluster de banco de dados no Neptune
- Criar um snapshot de banco de dados
- Criar uma regra do EventBridge que é acionada de acordo com uma programação
- Replicação entre regiões com o Amazon S3
- EFS para AWS Backup de EFS
- Exportar dados de log para o Amazon S3
- Gerenciamento do ciclo de vida de objetos
- Backup e restauração sob demanda para o DynamoDB
- Recuperação para um ponto no tempo para o DynamoDB
- Como trabalhar com snapshots de índices no Amazon OpenSearch Service
- O que é o AWS Elastic Disaster Recovery?

Vídeos relacionados:

- AWS re:Invent 2021: Backup, recuperação de desastres e proteção contra ransomware com a AWS
- Demonstração do AWS Backup: backup entre contas e regiões
- AWS re:Invent 2019: Mergulho profundo no AWS Backup com destaque para o Rackspace (STG341)

Exemplos relacionados:

- Laboratório do Well-Architected: Implementar a replicação bidirecional entre regiões (CRR) para o Amazon S3
- Laboratório do Well-Architected: Testar o backup e a restauração de dados
- Laboratório do Well-Architected: Backup e restauração com failback para workloads analíticas
- Laboratório do Well-Architected: Recuperação de desastres: backup e restauração

REL09-BP02 Proteger e criptografar backups

Controle e detecte o acesso a backups usando autenticação e autorização. Use a criptografia para prevenir e detectar se a integridade dos dados de backups está comprometida.

Práticas comuns que devem ser evitadas:

- Ter o mesmo acesso à automação de backups e restauração que tem aos dados.
- Não criptografar seus backups.

Benefícios de implementar esta prática recomendada: proteger os backups impede a violação dos dados, e a criptografia dos dados impede o acesso a eles caso sejam expostos por engano.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Controle e detecte o acesso a backups usando autenticação e autorização, como o AWS Identity and Access Management (IAM). Use a criptografia para prevenir e detectar se a integridade dos dados de backups está comprometida.

O Amazon S3 oferece suporte a vários métodos de criptografia de dados em repouso. Ao usar a criptografia do lado do servidor, o Amazon S3 aceita os objetos como dados não criptografados e, depois, criptografa-os à medida que são armazenados. Ao usar a criptografia do lado do cliente, a aplicação da workload é responsável por criptografar os dados antes de serem enviados ao Amazon S3. Ambos os métodos permitem que você use o AWS Key Management Service (AWS KMS) para criar e armazenar a chave de dados, ou você pode fornecer sua própria chave, pela qual você é responsável. Usando o AWS KMS, você pode definir políticas usando o IAM sobre quem pode e não pode acessar suas chaves de dados e dados descriptografados.

Para o Amazon RDS, se você tiver optado por criptografar seus bancos de dados, seus backups também serão criptografados. Os backups do DynamoDB sempre são criptografados. Quando o AWS Elastic Disaster Recovery é usado, todos os dados em trânsito e em repouso são

criptografados. Com o Elastic Disaster Recovery, os dados em repouso podem ser criptografados usando a chave de criptografia de volume padrão do Amazon EBS ou uma chave personalizada gerenciada pelo cliente.

Etapas de implementação

1. Use criptografia em cada um dos seus datastores. Se os dados de origem forem criptografados, o backup também será.
   - Use criptografia no Amazon RDS. Você pode configurar a criptografia em repouso usando o AWS Key Management Service ao criar uma instância do RDS.
   - Use criptografia nos volumes do Amazon EBS. Você pode configurar a criptografia padrão ou especificar uma chave exclusiva após a criação do volume.
   - Use a criptografia do Amazon DynamoDB necessária. O DynamoDB criptografa todos os dados em repouso. Você pode usar uma chave do AWS KMS pertencente à AWS ou uma chave do KMS gerenciada pela AWS, especificando uma chave armazenada na sua conta.
   - Criptografe seus dados armazenados no Amazon EFS. Configure a criptografia ao criar seu sistema de arquivos.
   - Configure a criptografia nas regiões de origem e de destino. Você pode configurar a criptografia em repouso no Amazon S3 usando as chaves armazenadas no KMS, mas as chaves são específicas da região. É possível especificar as chaves de destino ao configurar a replicação.
   - Escolha se deseja usar a criptografia padrão ou usar a criptografia do Amazon EBS para Elastic Disaster Recovery. Essa opção criptografa os dados em repouso replicados nos discos da sub-rede da área de preparação e os discos replicados.
2. Implemente permissões de privilégio mínimo para acessar seus backups. Siga as práticas recomendadas para limitar o acesso aos backups, aos snapshots e às réplicas de acordo com as práticas recomendadas de segurança.

Recursos

Documentos relacionados:

- AWS Marketplace: produtos que podem ser usados para backup
- Criptografia do Amazon EBS
- Amazon S3: proteger dados usando criptografia
- Configuração adicional de CRR: replicar objetos criados com a criptografia do lado do servidor (SSE) usando as chaves de criptografia armazenadas no AWS KMS

- Criptografia do DynamoDB em repouso
- Como criptografar recursos do Amazon RDS
- Criptografar dados e metadados no Amazon EFS
- Criptografia para backups no AWS
- Como gerenciar tabelas criptografadas
- Pilar Segurança: AWS Well-Architected Framework
- O que é o AWS Elastic Disaster Recovery?

Exemplos relacionados:

- Laboratório do Well-Architected: Implementar a replicação bidirecional entre regiões (CRR) para o Amazon S3

REL09-BP03 Realizar backups de dados automaticamente

Configure os backups para serem feitos automaticamente com base em uma programação periódica informada pelo objetivo de ponto de recuperação (RPO) ou de acordo com alterações no conjunto de dados. É necessário fazer frequentemente o backup automático de conjuntos de dados críticos com requisitos de baixa perda de dados, enquanto o backup de dados menos críticos, em que alguma perda é aceitável, pode ser feito com menos frequência.

Resultado desejado: um processo automatizado que cria backups das fontes de dados em uma cadência estabelecida.

Práticas comuns que devem ser evitadas:

- Fazer backups manualmente.
- Usar recursos que têm o recurso de backup, mas não incluir o backup em sua automação.

Benefícios de implementar esta prática recomendada: automatizar os backups verifica se eles são feitos regularmente com base no seu RPO e alerta você caso não sejam feitos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

O AWS Backup pode ser usado para criar backups de dados automatizados de várias fontes de dados da AWS. É possível fazer backup de instâncias do Amazon RDS quase continuamente a

cada cinco minutos, e objetos do Amazon S3 podem ser feitos backup quase continuamente a cada quinze minutos, proporcionando recuperação para um ponto no tempo (PITR) dentro do histórico de backup. Para outras fontes de dados da AWS, como volumes do Amazon EBS, tabelas do Amazon DynamoDB ou sistemas de arquivos do Amazon FSx, o AWS Backup pode executar backup automatizado de hora em hora. Esses serviços também oferecem recursos de backup nativos. os serviços da AWS que oferecem backup automatizado com recuperação para um ponto no tempo incluem Amazon DynamoDB, Amazon RDS e Amazon Keyspaces (para Apache Cassandra): esses podem ser restaurados em um momento específico no histórico de backup. A maioria dos outros serviços de armazenamento de dados da AWS permite programar backups periódicos, até de hora em hora.

O Amazon RDS e o Amazon DynamoDB oferecem backup contínuo com recuperação para um ponto no tempo. O versionamento do Amazon S3, uma vez habilitado, é automático. O Amazon Data Lifecycle Manager pode ser usado para automatizar a criação, a retenção e a exclusão de AMIs compatíveis com o EBS. Ele também pode automatizar a criação, a cópia, a suspensão e o cancelamento do registro de imagens de máquina da Amazon (AMIs) com base no Amazon EBS e seus snapshots subjacentes do Amazon EBS.

O AWS Elastic Disaster Recovery fornece replicação contínua no nível de bloco do ambiente de origem (on-premises ou a AWS) para a região de recuperação de destino. Os snapshots do Amazon EBS de um ponto anterior no tempo são criados automaticamente e gerenciados pelo serviço.

Para obter uma visão centralizada da automação e do histórico de backups, o AWS Backup oferece uma solução de backup totalmente gerenciada e baseada em políticas. Ele centraliza e automatiza o backup de dados em vários serviços da AWS, na nuvem e on-premises, usando o AWS Storage Gateway.

Além do controle de versões, o Amazon S3 oferece replicação. Todo o bucket do S3 pode ser replicado automaticamente para outro bucket na mesma Região da AWS ou em uma região diferente.

Etapas de implementação

1. Identifique as fontes de dados que estão sendo copiadas para backup manualmente no momento. Para obter mais detalhes, consulte REL09-BP01 Identificar e fazer backup de todos os dados que precisam de backup ou reproduzir os dados das fontes.
2. Determine o RPO para a workload. Para obter mais detalhes, consulte REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados.

3. Use uma solução automatizada ou um serviço de backup gerenciado. O AWS Backup é um serviço totalmente gerenciado que ajuda você a centralizar e automatizar a proteção de dados nos serviços da AWS, na nuvem e on-premises. Usando planos de backup no AWS Backup, crie regras que definem os recursos para backup e a frequência com que esses backups devem ser criados. A frequência deve ser informada pelo RPO estabelecido na Etapa 2. Para obter orientação prática sobre como criar backups automatizados usando o AWS Backup, consulte Testar o backup e a restauração de dados. A maioria dos serviços da AWS que armazenam dados oferecem recursos de backup nativos. Por exemplo, o RDS pode ser utilizado para fazer backups automatizados com recuperação para um ponto no tempo (PITR).
4. Para fontes de dados sem suporte de uma solução de backup automatizada ou serviço gerenciado, como fontes de dados on-premises ou filas de mensagens, considere usar uma solução terceirizada confiável para criar backups automatizados. Como alternativa, você pode criar automação para fazer isso usando a AWS CLI ou os SDKs. É possível usar o AWS Lambda Functions ou o AWS Step Functions para definir a lógica envolvida na criação de um backup de dados e usar o Amazon EventBridge para executá-la em uma frequência baseada no RPO.

Nível de esforço do plano de implementação: Baixo.

Recursos

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudar com o backup
- AWS Marketplace: produtos que podem ser usados para backup
- Criar uma regra do EventBridge que é acionada de acordo com uma programação
- O que é o AWS Backup?
- O que é o AWS Step Functions?
- O que é o AWS Elastic Disaster Recovery?

Vídeos relacionados:

- AWS re:Invent 2019: Mergulho profundo no AWS Backup com destaque para o Rackspace (STG341)

Exemplos relacionados:

- Laboratório do Well-Architected: Testar o backup e a restauração de dados

REL09-BP04 Realizar a recuperação periódica dos dados para verificar a integridade e os processos de backup

Execute um teste de recuperação para confirmar se a implementação do processo de backup atende aos seus objetivos de tempo de recuperação (RTO) e de ponto de recuperação (RPO).

Resultado desejado: os dados dos backups são recuperados periodicamente usando mecanismos bem definidos para verificar se a recuperação é possível dentro do objetivo de tempo de recuperação (RTO) estabelecido para a workload. Verifique se a restauração de um backup resulta em um recurso contendo os dados originais sem que estejam corrompidos ou inacessíveis e que a perda de dados esteja dentro do objetivo de ponto de recuperação (RPO).

Práticas comuns que devem ser evitadas:

- Restaurar um backup, mas não consultar ou recuperar os dados para garantir que a restauração é utilizável.
- Presumir a existência de um backup.
- Presumir que o backup de um sistema esteja totalmente operacional e que os dados possam ser recuperados.
- Presumir que o tempo para recuperar ou restaurar dados de um backup esteja dentro do RTO para a workload.
- Presumir que os dados contidos no backup estejam dentro do RPO para a workload
- Restaurar ad hoc, sem usar um runbook ou não seguir um procedimento automatizado estabelecido.

Benefícios de estabelecer essa prática recomendada: testar a recuperação dos backups verifica se os dados podem ser restaurados quando necessário, sem a preocupação de que os dados possam estar ausentes ou corrompidos, que a restauração e a recuperação sejam possíveis dentro do RTO da workload e que qualquer perda de dados esteja dentro do RPO da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Testar o recurso de backup e restauração aumenta a confiança na capacidade de realizar essas ações durante uma interrupção. Restaure periodicamente os backups em um novo local e execute

testes para verificar a integridade dos dados. Alguns testes comuns que devem ser realizados são verificar se todos os dados estão disponíveis, se eles não estão corrompidos e se eles estão acessíveis, bem como garantir que toda a perda de dados se enquadre no RPO da workload. Eles também podem ajudar a verificar se os mecanismos de recuperação são rápidos o suficiente para acomodar o RTO da workload.

Ao usar a AWS, você pode criar um ambiente de teste e restaurar os backups para avaliar os recursos de RTO e RPO e executar testes de conteúdo e integridade dos dados.

Além disso, o Amazon RDS e o Amazon DynamoDB permitem a recuperação para um ponto no tempo (PITR). Ao usar o backup contínuo, você pode restaurar o conjunto de dados para o estado em que ele se encontrava em uma data e hora especificadas.

Se todos os dados estão disponíveis, não corrompidos, acessíveis e qualquer perda de dados está de acordo com o RPO da workload. Eles também podem ajudar a verificar se os mecanismos de recuperação são rápidos o suficiente para acomodar o RTO da workload.

O AWS Elastic Disaster Recovery oferece snapshots contínuos de recuperação para um ponto no tempo de volumes do Amazon EBS. À medida que os servidores de origem são replicados, os estados pontuais são registrados ao longo do tempo com base na política configurada. O Elastic Disaster Recovery ajuda você a verificar a integridade desses snapshots lançando instâncias para fins de teste e detalhamento sem redirecionar o tráfego.

Etapas de implementação

1. Identifique as fontes de dados que estão sendo copiadas no momento e onde esses backups estão sendo armazenados. Para obter orientações de implementação, consulte REL09-BP01 Identificar e fazer backup de todos os dados que precisam de backup ou reproduzir os dados das fontes.
2. Estabeleça critérios para validação de dados para cada fonte de dados. Diferentes tipos de dados terão propriedades distintas que podem exigir mecanismos de validação diferentes. Considere como validar esses dados antes de se sentir confiante em usá-los na produção. Algumas maneiras comuns de validar dados são o uso de dados e propriedades de backup, como tipo de dados, formato, soma de verificação, tamanho ou uma combinação deles com lógica de validação personalizada. Por exemplo, pode ser uma comparação dos valores de soma de verificação entre o recurso restaurado e a fonte de dados no momento em que o backup foi criado.
3. Estabeleça o RTO e o RPO para restaurar os dados com base na criticidade dos dados. Para obter orientações de implementação, consulte REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados.

4. Avalie sua capacidade de recuperação. Revise sua estratégia de backup e de restauração para entender se ela pode cumprir o RTO e o RPO e ajuste a estratégia conforme necessário. Usando o Hub de Resiliência da AWS, você pode executar uma avaliação da sua workload. Essa avaliação analisa a configuração da aplicação em relação à política de resiliência e relata se as metas de RTO e RPO podem ser cumpridas.
5. Faça uma restauração de teste usando os processos atualmente estabelecidos usados na produção para restauração de dados. Esses processos dependem de como foi feito o backup da fonte de dados original, do formato e do local de armazenamento do próprio backup ou se os dados são reproduzidos de outras fontes. Por exemplo, se você estiver usando um serviço gerenciado como o AWS Backup, isso poderá ser tão simples quanto restaurar o backup em um novo recurso. Se você usou o AWS Elastic Disaster Recovery, pode iniciar um exercício de recuperação.
6. Valide a recuperação de dados do recurso restaurado com base nos critérios que você estabeleceu anteriormente para validação de dados. Os dados restaurados e recuperados contêm o registro ou item mais recente no momento do backup? Esses dados se enquadram no RPO da workload?
7. Meça o tempo necessário para restauração e recuperação e compare-o com seu RTO estabelecido. Esse processo se enquadra no RTO da workload? Por exemplo, compare o carimbo de data/hora em que o processo de restauração foi iniciado e que a validação da recuperação foi concluída para calcular quanto tempo esse processo demora. Todas as chamadas da API da AWS contêm informações de data e hora, e essas informações estão disponíveis no AWS CloudTrail. Embora essas informações possam fornecer detalhes sobre o início do processo de restauração, o carimbo final de data/hora da conclusão da validação deve ser registrado pela lógica de validação. Se estiver usando um processo automatizado, serviços como o Amazon DynamoDB poderão ser usados para armazenar essas informações. Além disso, muitos serviços da AWS oferecem um histórico de eventos que fornece informações sobre a data e a hora em que determinadas ações ocorreram. No AWS Backup, as ações de backup e restauração são chamadas de trabalhos, e esses trabalhos contêm informações de data e hora como parte de seus metadados que podem ser usadas para medir o tempo necessário para restauração e recuperação.
8. Notifique as partes interessadas se a validação de dados falhar ou se o tempo necessário para restauração e recuperação exceder o RTO estabelecido para a workload. Ao implementar a automação para fazer isso, como neste laboratório, serviços como o Amazon Simple Notification Service (Amazon SNS) podem ser usados para enviar notificações push, como e-mail ou SMS, às partes interessadas. Essas mensagens também podem ser publicadas em aplicações de

mensagens, como Amazon Chime, Slack ou Microsoft Teams ou usadas para criar tarefas como OpsItems usando o AWS Systems Manager OpsCenter.

9. Automatize esse processo para ser executado periodicamente. Por exemplo, serviços como o AWS Lambda ou uma máquina de estado no AWS Step Functions podem ser usados para automatizar os processos de restauração e recuperação, e é possível usar o Amazon EventBridge para invocar esse fluxo de trabalho de automação periodicamente, conforme mostrado no diagrama de arquitetura abaixo. Saiba como automatizar a validação da recuperação de dados com o AWS Backup. Além disso, esse laboratório do Well-Architected fornece uma experiência prática sobre uma forma de automatizar várias das etapas descritas aqui.

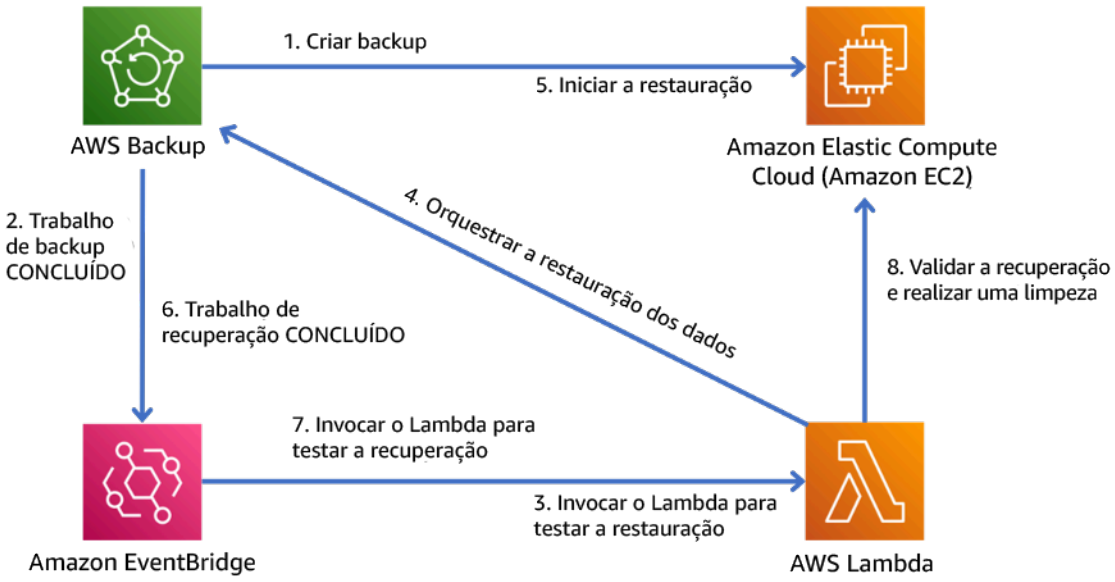


Figura 9. Um processo de backup e restauração automatizado

Nível de esforço para o plano de implementação: Moderado a alto, dependendo da complexidade dos critérios de validação.

Recursos

Documentos relacionados:

- Automatizar a validação da recuperação de dados com o AWS Backup.
- Parceiro da APN: parceiros que podem ajudar com o backup
- AWS Marketplace: produtos que podem ser usados para backup
- Criar uma regra do EventBridge que é acionada de acordo com uma programação
- Backup e restauração sob demanda para o DynamoDB

- O que é o AWS Backup?
- O que é o AWS Step Functions?
- O que é o AWS Elastic Disaster Recovery
- AWS Elastic Disaster Recovery

Exemplos relacionados:

- Laboratório do Well-Architected: Testar o backup e a restauração de dados

REL 10. Como você usa o isolamento de falhas para proteger a workload?

Os limites isolados de falhas limitam o efeito de uma falha em uma workload a um número limitado de componentes. Os componentes fora do limite não são afetados pela falha. Ao usar vários limites isolados de falhas, é possível limitar o impacto na workload.

Práticas recomendadas

- REL10-BP01 Implantar a workload em vários locais
- REL10-BP02 Selecionar os locais apropriados para sua implantação de vários locais
- REL10-BP03 Automatizar a recuperação de componentes restritos a um único local
- REL10-BP04 Usar arquiteturas de anteparo para limitar o escopo de impactos

REL10-BP01 Implantar a workload em vários locais

Distribua os dados e os recursos da workload por várias zonas de disponibilidade ou, quando necessário, por Regiões da AWS. A diversidade dos locais pode variar conforme a necessidade.

Um dos princípios fundamentais do design de serviço na AWS é evitar pontos únicos de falha na infraestrutura física subjacente. Isso nos motiva a criar software e sistemas que usam várias zonas de disponibilidade e são resilientes à falha de uma única zona. De modo similar, os sistemas são criados para serem resilientes à falha de um único nó de computação, volume de armazenamento ou instância de banco de dados. Ao criar um sistema que dependa de componentes redundantes, é importante garantir que os componentes operem de modo independente e, no caso de Regiões da AWS, de modo autônomo. Os benefícios obtidos com cálculos teóricos de disponibilidade com componentes redundantes só serão válidos se isso for verdadeiro.

Zonas de disponibilidade (AZ)

As Regiões da AWS são compostas por várias zonas de disponibilidade projetadas para serem independentes umas das outras. Cada zona de disponibilidade é separada por uma distância física significativa em relação às outras zonas para evitar cenários de falha correlacionados devido a riscos ambientais, como incêndios, enchentes e tornados. Cada zona de disponibilidade tem uma infraestrutura independente: conexões dedicadas à rede elétrica, fontes de alimentação de reserva independentes, serviços mecânicos independentes e conectividade de rede independente dentro e além da zona de disponibilidade. Esse design limita as falhas em qualquer um desses sistemas apenas à AZ afetada. Apesar de serem separadas geograficamente, as zonas de disponibilidade estão localizadas na mesma área regional, o que permite redes de alto throughput e baixa latência. A Região da AWS inteira (em todas as zonas de disponibilidade, consistindo em vários data centers fisicamente independentes) pode ser tratada como um único destino lógico de implantação para sua workload, incluindo a capacidade de replicar dados de forma síncrona (por exemplo, entre bancos de dados). Assim, os clientes podem usar as zonas de disponibilidade em uma configuração ativa/ativa ou ativa/standby.

As zonas de disponibilidade são independentes e, portanto, a disponibilidade da workload aumenta quando ela é projetada para usar várias zonas. Alguns serviços da AWS (incluindo o plano de dados da instância do Amazon EC2) são implantados como serviços estritamente zonais, onde eles têm um destino compartilhado com a zona de disponibilidade como um todo. No entanto, as instâncias do Amazon EC2 nas outras AZs não serão afetadas e continuarão funcionando. Da mesma forma, se uma falha em uma zona de disponibilidade fizer com que um banco de dados Amazon Aurora falhe, uma instância do Aurora de réplica de leitura em uma AZ não afetada pode ser automaticamente promovida para primária. Serviços da AWS regionais, como o Amazon DynamoDB, por outro lado, usam internamente várias zonas de disponibilidade em uma configuração ativa/ativa para atingir as metas de design de disponibilidade desse serviço, sem a necessidade de configurar o posicionamento da AZ.

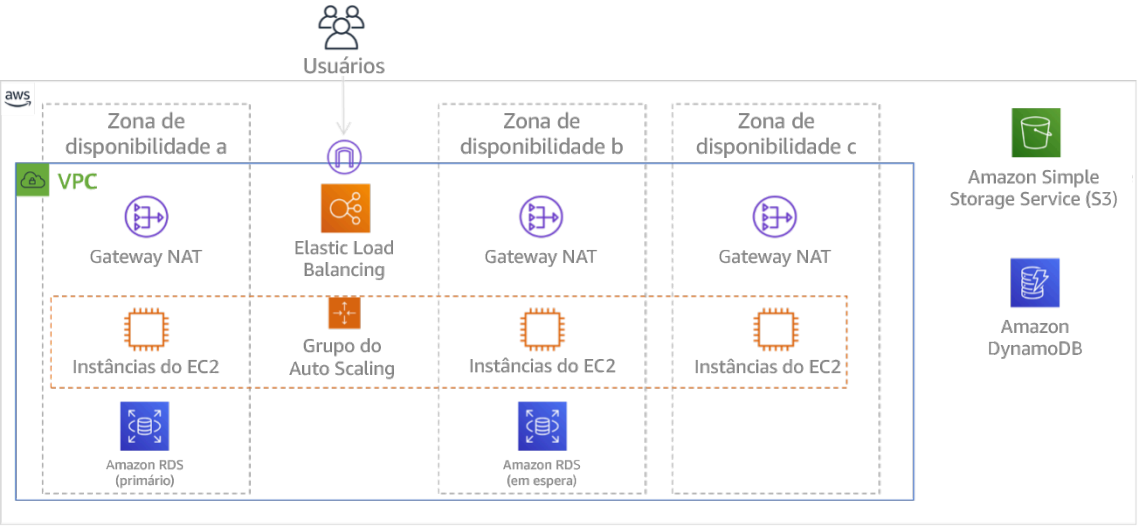


Figura 9: arquitetura multicamadas implantada em três Zonas de disponibilidade. Observe que o Amazon S3 e o Amazon DynamoDB são sempre Multi-AZ automaticamente. O ELB também é implantado em todas as três zonas.

Embora os planos de controle da AWS costumem permitir o gerenciamento de recursos dentro de toda a região (várias zonas de disponibilidade), determinados ambientes de gerenciamento (incluindo Amazon EC2 e Amazon EBS) podem filtrar os resultados para uma única zona de disponibilidade. Quando isso é feito, a solicitação é processada apenas na zona de disponibilidade especificada, o que reduz a exposição a interrupções em outras zonas de disponibilidade. Este exemplo da AWS CLI ilustra a obtenção de informações de instâncias do Amazon EC2 somente da zona de disponibilidade us-east-2c:

```
AWS ec2 describe-instances --filters Name=availability-zone,Values=us-east-2c
```

Zonas locais da AWS

As zonas locais da AWS atuam de forma semelhante às respectivas zonas de disponibilidade em suas respectivas regiões da Região da AWS, pois elas podem ser selecionadas como um local de posicionamento para recursos zonais da AWS, como sub-redes e instâncias do EC2. O que as torna especiais é que elas estão localizadas não na região da Região da AWS associada, mas perto de grandes centros populacionais, industriais e de TI onde não existe nenhuma região da Região da AWS atualmente. No entanto, elas ainda mantêm uma conexão segura e de alta largura de banda entre as Workloads locais na zona local e as executadas na Região da AWS. Você deve usar as zonas locais da AWS para implantar as workloads mais perto de seus usuários para requisitos de baixa latência.

Amazon Global Edge Network

A rede de presença global da Amazon consiste em pontos de presença em cidades em todo o mundo. O Amazon CloudFront usa essa rede para entregar conteúdo aos usuários finais com latência mais baixa. AWS O Global Accelerator permite criar endpoints de workload nesses pontos de presença para oferecer integração à rede global da AWS próxima aos usuários. O Amazon API Gateway habilita endpoints de API otimizados para a borda usando uma distribuição do CloudFront para facilitar o acesso do cliente por meio do ponto de presença mais próximo.

Regiões da AWS

As Regiões da AWS foram projetadas para serem autônomas. Portanto, para usar uma abordagem em várias regiões, você pode implantar cópias dedicadas de serviços em cada uma delas.

Uma abordagem multirregiões é comum para que as estratégias de recuperação de desastres atendam aos objetivos de recuperação quando ocorrem eventos pontuais de grande escala. Consulte Plano de recuperação de desastres (DR) para obter mais informações sobre essas estratégias. Aqui, no entanto, focamos a disponibilidade, que busca oferecer um objetivo médio de tempo de atividade ao longo do tempo. Para objetivos de alta disponibilidade, uma arquitetura multirregiões geralmente será projetada para ser ativa/ativa, em que cada cópia do serviço (em suas respectivas regiões) está ativa (atendendo às solicitações).

> **Recomendação**
>
> Os objetivos de confiabilidade para a maioria das workloads pode ser cumprido usando-se uma estratégia Multi-AZ em uma única Região da AWS. Considere arquiteturas multirregiões somente quando as workloads tiverem requisitos extremos de disponibilidade ou outras metas de negócios que exijam uma arquitetura multirregiões.

A AWS fornece a você os recursos para operar serviços em várias regiões. Por exemplo, a AWS fornece replicação contínua e assíncrona de dados usando a replicação do Amazon Simple Storage Service (Amazon S3), réplicas de leitura do Amazon RDS (incluindo réplicas de leitura do Aurora) e tabelas globais do Amazon DynamoDB. Com a replicação contínua, as versões dos seus dados tornam-se disponíveis para uso quase imediato em cada uma das regiões ativas.

Com o AWS CloudFormation, é possível definir sua infraestrutura e implantá-la de forma consistente nas Contas da AWS e nas Regiões da AWS. O AWS CloudFormation StackSets amplia a funcionalidade das pilhas, permitindo que você crie, atualize ou exclua pilhas do AWS CloudFormation em várias contas e regiões com uma única operação. Para implantações de instâncias do Amazon EC2, uma imagem de máquina da Amazon (AMI) é usada para fornecer informações como configuração de hardware e software instalado. Você pode implementar um pipeline do Amazon EC2 Image Builder que cria as AMIs de que você precisa e as copia para suas regiões ativas. Isso garante que essas AMIs de ouro tenham tudo o que você precisa para implantar e escalar sua workload em cada nova região.

Para rotear o tráfego, tanto o Amazon Route 53 quanto o AWS Global Accelerator permitem a definição de políticas que determinam quais usuários vão para qual endpoint regional ativo. Com o Global Accelerator, você define uma discagem de tráfego para controlar a porcentagem de tráfego que é direcionada para cada endpoint da aplicação. O Route 53 oferece suporte a essa abordagem percentual e também várias outras políticas disponíveis, incluindo políticas baseadas em geoproximidade e latência. O Global Accelerator aproveita automaticamente a extensa rede de

servidores de borda da AWS para integrar o tráfego ao backbone da rede da AWS o mais rápido possível, resultando em menores latências de solicitação.

Todas essas capacidades operam de forma a preservar a autonomia de cada região. Há muito poucas exceções a essa abordagem, incluindo nossos serviços que fornecem entrega de borda global (como o Amazon CloudFront e o Amazon Route 53), junto com o ambiente de gerenciamento para o serviço do AWS Identity and Access Management (IAM). A grande maioria dos serviços opera inteiramente dentro de uma única região.

Datacenter on-premises

Para workloads executadas em um datacenter on-premises, arquitete uma experiência híbrida quando possível. O AWS Direct Connect fornece uma conexão de rede dedicada entre seu ambiente on-premises e a AWS, permitindo que você trabalhe em ambos.

Outra opção é executar a infraestrutura e os serviços da AWS on premises usando o AWS Outposts. O AWS Outposts é um serviço totalmente gerenciado que estende a infraestrutura da AWS, os serviços da AWS, as APIs e as ferramentas para o seu datacenter. A mesma infraestrutura de hardware usada na Nuvem AWS é instalada no seu datacenter. Os AWS Outposts são então conectados à Região da AWS mais próxima. Em seguida, você pode usar o AWS Outposts para oferecer suporte a workloads com baixa latência ou requisitos de processamento de dados locais.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

- Use várias zonas de disponibilidade e Regiões da AWS. Distribua os dados e os recursos da workload por várias zonas de disponibilidade ou, quando necessário, por Regiões da AWS. A diversidade dos locais pode variar conforme a necessidade.
  - Os serviços regionais são inerentemente implantados nas zonas de disponibilidade.
    - Isso inclui o Amazon S3, o Amazon DynamoDB e o AWS Lambda (quando não estão conectados a uma VPC)
  - Implante suas workloads baseadas em contêiner, instância e função em várias zonas de disponibilidade. Use datastores multizona, incluindo caches Use os recursos do Amazon EC2 Auto Scaling, o posicionamento de tarefas do Amazon ECS, a configuração da função do AWS Lambda ao executá-lo em sua VPC e clusters do ElastiCache.
    - Use sub-redes que estão em zonas de disponibilidade separadas ao implantar grupos do Auto Scaling.

        - Exemplo: distribuir instâncias entre zonas de disponibilidade
        - Escolher regiões e zonas de disponibilidade
    - Use os parâmetros de posicionamento de tarefas do ECS, com a especificação de grupos de sub-rede de banco de dados.
        - Estratégias de posicionamento de tarefas do Amazon ECS
    - Use as sub-redes em várias zonas de disponibilidade ao configurar uma função para executar na sua VPC.
        - Configurar uma função do AWS Lambda para acessar recursos em uma Amazon VPC
    - Use várias zonas de disponibilidade com os clusters do ElastiCache.
        - Escolher regiões e zonas de disponibilidade
- Se a workload precisar ser implantada em várias regiões, escolha uma estratégia multirregiões. A maioria das necessidades de confiabilidade pode ser atendida em uma única Região da AWS por meio de uma estratégia de várias zonas de disponibilidade. Use uma estratégia multirregiões quando necessário para atender às suas demandas de negócios.
    - AWS re:Invent 2018: Padrões de arquitetura para aplicações ativas-ativas multirregiões (ARC209-R2)
        - O backup para outra Região da AWS pode servir como mais uma camada visando garantir que os dados estejam disponíveis quando necessário.
        - Algumas workloads têm requisitos normativos que exigem o uso de uma estratégia multirregiões
- Avalie o AWS Outposts para sua workload. Se a workload exigir baixa latência do datacenter on-premises ou tiver requisitos de processamento de dados locais. Então execute a infraestrutura e os serviços da AWS on-premises usando o AWS Outposts.
    - O que é AWS Outposts?
- Determine se as zonas locais da AWS ajudam você a fornecer serviços aos usuários Se você tiver requisitos de baixa latência, veja se as zonas locais da AWS estão próximas dos seus usuários. Se estiverem, use-as para implantar as workloads mais perto desses usuários.
    - Perguntas frequentes sobre zonas locais da AWS

Recursos

Documentos relacionados:

- Infraestrutura global da AWS

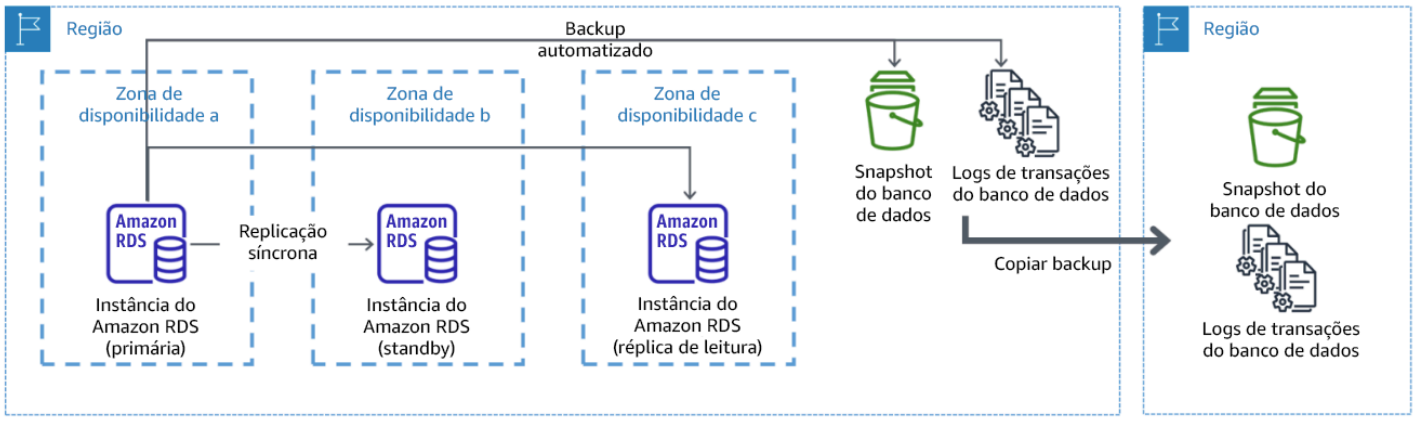
- Perguntas frequentes sobre zonas locais da AWS
- Estratégias de posicionamento de tarefas do Amazon ECS
- Escolher regiões e zonas de disponibilidade
- Exemplo: distribuir instâncias entre zonas de disponibilidade
- Tabelas globais: replicação em várias regiões com o DynamoDB
- Usar bancos de dados globais do Amazon Aurora
- Série de blogs Criar aplicações multirregiões com serviços da AWS
- O que é AWS Outposts?

Vídeos relacionados:

- AWS re:Invent 2018: Padrões de arquitetura para aplicações ativas-ativas multirregiões (ARC209-R2)
- AWS re:Invent 2019: Inovação e operação da infraestrutura de rede global da AWS (NET339)

REL10-BP02 Selecionar os locais apropriados para sua implantação de vários locais

Resultado desejado: para alta disponibilidade, sempre (que possível) implante os componentes da workload em várias zonas de disponibilidade (AZs). Para workloads com requisitos de resiliência extrema, avalie cuidadosamente as opções para uma arquitetura multirregiões.



Uma implantação resiliente de banco de dados Multi-AZ com backup em outra região da AWS

Práticas comuns que devem ser evitadas:

- Optar por projetar uma arquitetura multirregiões quando uma arquitetura Multi-AZ é suficiente para satisfazer os requisitos.

- Não contabilizar as dependências entre os componentes da aplicação se os requisitos de resiliência e de vários locais são diferentes entre esses componentes.

Benefícios de implementar esta prática recomendada: para fins de resiliência, empregue uma abordagem que crie camadas de defesa. Uma camada protege contra interrupções menores e mais comuns criando uma arquitetura altamente disponível usando várias AZs. Outra camada de defesa destina-se a proteger contra eventos raros, como desastres naturais generalizados e interrupções em nível regional. Essa segunda camada envolve arquitetar a aplicação para abranger várias Regiões da AWS.

- A diferença entre uma disponibilidade de 99,5% e uma disponibilidade de 99,99% é superior a 3,5 horas por mês. A disponibilidade esperada de uma workload só pode chegar a "quatro noves" se ela estiver em várias AZs.
- Ao executar sua workload em várias AZs, você pode isolar falhas de energia, resfriamento e rede, bem como a maioria dos desastres naturais, como incêndios e inundações.
- A implementação de uma estratégia multirregiões para a workload ajuda a protegê-la contra desastres naturais generalizados, que afetam uma grande área geográfica de um país, ou falhas técnicas de escopo regional. Esteja ciente de que a implementação de uma arquitetura multirregiões pode ser complexa e, geralmente, não é necessária para a maioria das workloads.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Para um evento de desastre baseado na interrupção ou perda parcial de uma zona de disponibilidade, implementar uma workload altamente disponível em várias zonas de disponibilidade em uma única Região da AWS ajuda a mitigar desastres naturais e técnicos. Cada Região da AWS é composta por várias zonas de disponibilidade, cada uma isolada de falhas nas outras zonas. No entanto, para um evento de desastre que inclua o risco de perder vários componentes da zona de disponibilidade, que estão a uma distância significativa um do outro, você deve implementar opções de recuperação de desastres para mitigar falhas de âmbito regional. Para workloads que exigem extrema resiliência (infraestrutura crítica, aplicações relacionadas à saúde, infraestrutura do sistema financeiro etc.), uma estratégia multirregiões pode ser necessária.

Etapas de implementação

1. Avalie sua workload e determine se as necessidades de resiliência podem ser atendidas por uma abordagem Multi-AZ (Região da AWS única) ou se elas exigem uma abordagem multirregiões. A implementação de uma arquitetura multirregiões para satisfazer esses requisitos introduzirá complexidade adicional, portanto, considere cuidadosamente seu caso de uso e seus requisitos. Os requisitos de resiliência quase sempre podem ser atendidos com uma única Região da AWS. Considere os seguintes requisitos possíveis ao determinar se você precisa usar várias regiões:
   a. Recuperação de desastres (DR): para um evento de desastre baseado em interrupção ou perda parcial de uma zona de disponibilidade, implementar uma workload altamente disponível em várias zonas de disponibilidade em uma única Região da AWS ajuda a mitigar desastres naturais e técnicos. Para um evento de desastre que inclua o risco de perda de vários componentes da zona de disponibilidade que estão a uma distância significativa um do outro, você deve implementar a recuperação de desastres em várias regiões para mitigar desastres naturais ou falhas técnicas de âmbito regional.
   b. Alta disponibilidade (HA): uma arquitetura multirregiões (com várias AZs em cada região) pode ser usada para obter mais de quatro noves (> 99,99%) de disponibilidade.
   c. Localização de pilhas: ao implantar uma workload para um público global, você pode implantar pilhas localizadas em diferentes Regiões da AWS para atender ao público nessas regiões. A localização pode incluir idioma, moeda e tipos de dados armazenados.
   d. Proximidade com os usuários: ao implantar uma workload para um público global, você pode reduzir a latência implantando pilhas Regiões da AWS perto de onde os usuários finais estão.
   e. Residência de dados: algumas workloads estão sujeitas aos requisitos de residência de dados, em que os dados de determinados usuários devem permanecer dentro das fronteiras de um país específico. Com base na regulamentação em questão, você pode optar por implantar uma pilha inteira, ou apenas os dados, em uma Região da AWS dentro dessas fronteiras.
2. Veja a seguir alguns exemplos da funcionalidade Multi-AZ fornecida pelos serviços da AWS:
   a. Para proteger workloads usando o EC2 ou ECS, implante um Elastic Load Balancer na frente dos recursos computacionais. Em seguida, o Elastic Load Balancing fornece a solução para detectar as instâncias nas zonas com problemas de integridade e rotear o tráfego para as instâncias íntegras.
      i. Conceitos básicos de Application Load Balancers
      ii. Conceitos básicos de Network Load Balancers

b. No caso de instâncias do EC2 executando software comercial pronto para uso que não oferece suporte ao balanceamento de carga, você pode obter uma forma de tolerância a falhas implementando uma metodologia de recuperação de desastres Multi-AZ.

   i. the section called “REL13-BP02 Usar estratégias de recuperação definidas para cumprir os objetivos de recuperação”

c. Para tarefas do Amazon ECS, implante seu serviço uniformemente em três AZs para alcançar um equilíbrio entre disponibilidade e custo.

   i. Práticas recomendadas de disponibilidade do Amazon ECS | Contêineres

d. Para o Amazon RDS não Aurora, você pode escolher Multi-AZ como opção de configuração. Em caso de falha na instância primária do banco de dados, o Amazon RDS promove automaticamente um banco de dados em espera para receber tráfego em outra zona de disponibilidade. Réplicas de leitura em várias regiões também podem ser criadas para melhorar a resiliência.

   i. Implantações multi-AZ do Amazon RDS

   ii. Criar uma réplica de leitura em uma Região da AWS diferente

3. Veja a seguir alguns exemplos de funcionalidades multirregiões fornecidas pelos serviços da AWS:

   a. Para workloads do Amazon S3 em que a disponibilidade Multi-AZ é fornecida automaticamente pelo serviço, considere pontos de acesso multirregiões se uma implantação multirregiões for necessária.

      i. Pontos de acesso multirregiões no Amazon S3

   b. Para tabelas do DynamoDB em que a disponibilidade Multi-AZ é fornecida automaticamente pelo serviço, você pode converter facilmente as tabelas existentes em tabelas globais para aproveitar as vantagens de várias regiões.

      i. Converta suas tabelas de região única do Amazon DynamoDB em tabelas globais

   c. Se sua workload for comandada por Application Load Balancers ou Network Load Balancers, use o AWS Global Accelerator para melhorar a disponibilidade da aplicação direcionando o tráfego para várias regiões que contêm endpoints íntegros.

      i. Endpoints para aceleradores padrão no AWS Global Accelerator: AWS Global Accelerator (amazon.com)

   d. Para aplicações que utilizam o AWS EventBridge, considere os barramentos entre regiões para encaminhar eventos para outras regiões que você selecionar.

      i. Como enviar e receber eventos do Amazon EventBridge entre regiões da Regiões da AWS

e. Para bancos de dados do Amazon Aurora, considere os bancos de dados globais do Aurora, pois eles abrangem várias regiões da AWS. Os clusters existentes também podem ser modificados para adicionar novas regiões.

   i. Conceitos básicos de bancos de dados globais do Amazon Aurora

f. Se sua workload incluir chaves de criptografia do AWS Key Management Service (AWS KMS), considere se as chaves multirregiões são apropriadas para a aplicação.

   i. Chaves de multirregiões no AWS KMS

g. Para outros recursos de serviço da AWS, consulte esta série de blogs sobre Criar aplicações multirregiões com serviços da AWS

Nível de esforço do plano de implementação: Moderado a alto

Recursos

Documentos relacionados:

- Série Criar aplicações multirregiões com serviços da AWS
- Arquitetura de recuperação de desastres (DR) na AWS, Parte IV: ativa/ativa em vários sites
- Infraestrutura global da AWS
- Perguntas frequentes sobre zonas locais da AWS
- Arquitetura de recuperação de desastres (DR) na AWS, Parte I: estratégias de recuperação na nuvem
- A recuperação de desastres é diferente na nuvem
- Tabelas globais: replicação em várias regiões com o DynamoDB

Vídeos relacionados:

- AWS re:Invent 2018: Padrões de arquitetura para aplicações ativas-ativas multirregiões (ARC209-R2)
- Auth0: arquitetura multirregiões de alta disponibilidade com capacidade para escalar a até mais de 1,5 bilhão de logins por mês com failover automático

Exemplos relacionados:

- Arquitetura de recuperação de desastres (DR) na AWS, Parte I: estratégias de recuperação na nuvem
- DTCC obtém resiliência muito além do que é possível fazer em ambiente on-premises
- Expedia Group usa uma arquitetura multirregiões e de várias zonas de disponibilidade com um serviço de DNS proprietário para adicionar resiliência às aplicações
- Uber: recuperação de desastres para Kafka multirregiões
- Netflix: ativo-ativo para resiliência multirregiões
- Como criamos residência de dados para o Atlassian Cloud
- Intuit TurboTax opera em duas regiões

REL10-BP03 Automatizar a recuperação de componentes restritos a um único local

Se os componentes da workload só puderem ser executados em uma única zona de disponibilidade ou datacenter on-premises, será necessário implementar capacidade suficiente para fazer uma recompilação completa da workload em conformidade com os objetivos de recuperação definidos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Se a prática recomendada para implantar a workload em vários locais não for possível devido a restrições tecnológicas, você deverá implementar um caminho alternativo para a resiliência. Você deve automatizar a capacidade de recriar a infraestrutura necessária, reimplantar aplicações e recriar os dados necessários para esses casos.

Por exemplo, o Amazon EMR executa todos os nós de um determinado cluster na mesma zona de disponibilidade porque a execução de um cluster na mesma zona melhora a performance dos fluxos de trabalho, pois fornece uma taxa de acesso a dados mais alta. Se esse componente for necessário para a resiliência da workload, você deverá ter uma maneira de reimplantar o cluster e seus dados. Além disso, para o Amazon EMR, você deve provisionar redundância de maneiras diferentes de usar o Multi-AZ. É possível provisionar vários nós. Usando o EMR File System (EMRFS),Regiões da AWS os dados no EMR podem ser armazenados no Amazon S3, que, por sua vez, podem ser replicados em várias zonas de disponibilidade ou regiões da .

Da mesma forma, o Amazon Redshift, por padrão, provisiona o cluster em uma zona de disponibilidade escolhida aleatoriamente dentro da Região da AWS selecionada. Todos os nós de cluster são provisionados na mesma zona.

Para workloads com estado baseadas em servidor e implantadas em um datacenter on-premises, é possível usar o AWS Elastic Disaster Recovery para proteger as workloads na AWS. Se você já estiver hospedado na AWS, poderá usar o Elastic Disaster Recovery para proteger sua workload em uma zona ou região de disponibilidade alternativa. O Elastic Disaster Recovery usa a replicação contínua em nível de bloco para uma área temporária leve para fornecer recuperação rápida e confiável de aplicações on-premises e baseadas na nuvem.

Etapas de implementação

1. Implemente a autorrecuperação. Quando possível, use o ajuste de escala automático para implantar instâncias ou contêineres. Quando não for possível, use a recuperação automática de instâncias do EC2 ou implemente a automação de autorrecuperação com base nos eventos de ciclo de vida do contêiner do Amazon EC2 ou do ECS.
   - Use os grupos do Amazon EC2 Auto Scaling para instâncias e workloads de contêiner que não têm requisitos de endereço IP de instância única, endereço IP privado, endereço IP elástico e metadados de instância.
     - Os dados do usuário do modelo de execução podem ser usados para implementar uma automação que pode recuperar automaticamente a maioria das workloads.
   - Use a recuperação automática de instâncias do Amazon EC2 para workloads que exigem um endereço do ID de instância única, endereço IP privado, endereço IP elástico e metadados de instância.
     - A recuperação automática enviará alertas de status de recuperação para um tópico do SNS quando a falha na instância for detectada.
   - Use eventos de ciclo de vida da instância do Amazon EC2 ou eventos do Amazon ECS para automatizar a autorrecuperação quando ajuste de escala automático ou a recuperação do EC2 não puderem ser usadas.
     - Use os eventos para invocar a automação que recuperará seu componente de acordo com a lógica do processo necessária.
   - Proteja workloads monitoradas que estão limitadas a um único local usando o AWS Elastic Disaster Recovery.

Recursos

Documentos relacionados:

- Eventos do Amazon ECS

- Ganchos do ciclo de vida do Amazon EC2 Auto Scaling
- Recuperar a instância
- Ajuste de escala automático do serviço
- O que é o Amazon EC2 Auto Scaling?
- AWS Elastic Disaster Recovery

REL10-BP04 Usar arquiteturas de anteparo para limitar o escopo de impactos

Implemente arquiteturas de anteparo (também chamadas de arquiteturas baseadas em células) para restringir o efeito ou a falha em uma workload a um número limitado de componentes.

Resultado desejado: uma arquitetura baseada em células usa várias instâncias isoladas de uma workload em que cada instância é conhecida como célula. Cada célula é independente, não compartilha o estado com outras células e processa um subconjunto das solicitações gerais da workload. Isso reduz o possível impacto de uma falha, como uma atualização de software incorreta, a uma célula individual e às solicitações que ela está processando. Se uma workload usa 10 células para atender a 100 solicitações, quando uma falha ocorrer, 90% das solicitações gerais não serão afetadas pela falha.

Práticas comuns que devem ser evitadas:

- Permitir que as células cresçam sem limites.
- Aplicar implantações ou atualizações de código a todas as células ao mesmo tempo.
- Compartilhar o estado ou os componentes entre as células (com a exceção da camada do roteador).
- Adicionar negócios complexos ou rotear lógica para a camada do roteador.
- Não minimizar as interações entre as células.

Benefícios de implementar esta prática recomendada: com arquiteturas baseadas em células, muitos tipos comuns de falha são contidos na própria célula, o que permite o isolamento adicional das falhas. Esses limites de falha podem fornecer resiliência contra tipos de falha que, de outra forma, seriam difíceis de conter, como implantações de código malsucedidas ou solicitações corrompidas ou que invocam um modo de falha específico (também conhecido como solicitações de pílulas venenosas).

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Em uma embarcação, os anteparos garantem que uma ruptura no casco seja contida em uma seção do casco. Em sistemas complexos, esse padrão costuma ser replicado para permitir o isolamento de falhas. Os limites isolados de falhas restringem o efeito de uma falha em uma workload a um número controlado de componentes. Os componentes fora do limite não são afetados pela falha. Ao usar vários limites isolados de falhas, é possível limitar o impacto na workload. Na AWS, os clientes podem usar várias zonas de disponibilidade e regiões para fornecer o isolamento de falhas, mas o conceito do isolamento de falhas também pode ser estendido à arquitetura da workload.

A workload geral é composta por células particionadas por uma chave de partição. Ela precisa se alinhar à granularidade do serviço, ou da maneira natural que a workload de um serviço pode ser subdividida em interações mínimas entre células. Exemplos de chaves de partição são ID de cliente, ID de recurso ou qualquer outro parâmetro facilmente acessível na maioria das chamadas de API. Uma camada de roteamento de célula distribui solicitações a células individuais com base na chave de partição e apresenta um único endpoint aos clientes.

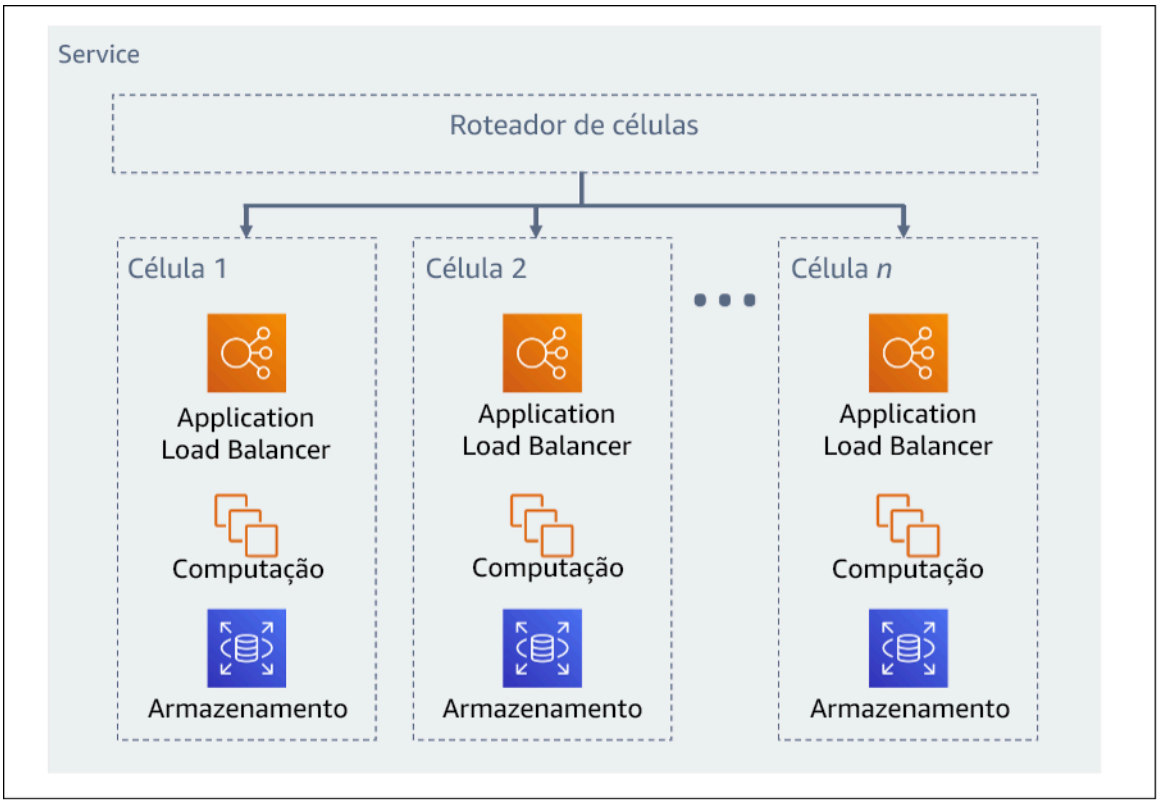


Figura 11: arquitetura baseada em células

Etapas de implementação

Ao projetar uma arquitetura baseada em células, há várias considerações de design a levar em conta:

1. Chave de partição: consideração especial deve ser feita em relação à chave de partição.
   - Ela precisa se alinhar à granularidade do serviço, ou à maneira natural que a workload de um serviço pode ser subdividida em interações mínimas entre células. Os exemplos são `customer ID` ou `resource ID`.
   - A chave de partição deve estar disponível em todas as solicitações, seja diretamente ou de uma maneira que possa ser facilmente inferida de forma determinística por outros parâmetros.
2. Mapeamento celular persistente: os serviços upstream devem interagir somente com uma única célula durante o ciclo de vida de seus recursos.
   - Dependendo da workload, uma estratégia de migração de células pode ser necessária para migrar os dados de uma célula para outra. Um possível cenário de quando é necessário fazer uma migração de célula seria quando um usuário ou recurso específico na workload se torna grande demais e exige uma célula dedicada.
   - As células não devem compartilhar estado ou componentes entre si.
   - Consequentemente, as interações entre as células devem ser evitadas e mantidas no mínimo, já que elas podem criar dependências entre as células e, assim, reduzir as melhorias do isolamento de falhas.
3. Camada do roteador: a camada do roteador é um componente compartilhado entre as células e, portanto, não pode seguir a mesma estratégia de compartimentação das células.
   - É recomendável que a camada do roteador distribua as solicitações para células individuais usando um algoritmo de mapeamento de partição de maneira computacionalmente eficiente, como combinando funções de hash criptográficas e aritmética modular para mapear chaves de partição a células.
   - Para evitar impactos em várias células, a camada de roteamento deve permanecer o mais simples e horizontalmente escalável possível, o que exige evitar uma lógica empresarial complexa nessa camada. Isso traz o benefício adicional de facilitar a compreensão de seu comportamento esperado em todos os momentos, permitindo a realização de testes rigorosos. Conforme explicado por Colm MacCárthaigh em Confiabilidade, trabalho constante e uma boa xícara de café, designs simples e padrões de trabalho constantes produzem sistemas confiáveis e reduzem a antifragilidade.
4. Tamanho da célula: as células devem ter um tamanho máximo e não devem se estender além dele.

- O tamanho máximo deve ser identificado com a realização de testes completos até que os pontos de ruptura sejam atingidos e margens operacionais seguras sejam estabelecidas. Para obter mais detalhes sobre como implementar práticas de testes, consulte REL07-BP04 Fazer o teste de carga da workload
- A workload geral deve crescer com a adição de mais células, permitindo que a workload seja escalada com aumentos na demanda.

5. Estratégias multi-AZ ou multirregiões: utilize várias camadas de resiliência para oferecer proteção contra diferentes domínios de falha.
   - Para resiliência, você deve usar uma abordagem que crie camadas de defesa. Uma camada protege contra interrupções menores e mais comuns criando uma arquitetura altamente disponível usando várias AZs. Outra camada de defesa destina-se a proteger contra eventos raros, como desastres naturais generalizados e interrupções em nível regional. Essa segunda camada envolve arquitetar a aplicação para abranger várias Regiões da AWS. A implementação de uma estratégia multirregiões para a workload ajuda a protegê-la contra desastres naturais generalizados, que afetam uma grande área geográfica de um país, ou falhas técnicas de escopo regional. Esteja ciente de que a implementação de uma arquitetura multirregiões pode ser complexa e, geralmente, não é necessária para a maioria das workloads. Para obter mais detalhes, consulte REL10-BP02 Selecionar os locais apropriados para sua implantação de vários locais.
6. Implantação de código: uma estratégia de implantação de código em etapas deve ser preferida à implantação de alterações de código em todas as células ao mesmo tempo.
   - Isso ajuda a reduzir a possibilidade de falhas em várias células devido a uma implantação incorreta ou a erro humano. Para obter mais detalhes, consulte Automatizar implantações seguras e sem intervenção manual.

Recursos

Práticas recomendadas relacionadas:

- REL07-BP04 Fazer o teste de carga da workload
- REL10-BP02 Selecionar os locais apropriados para sua implantação de vários locais

Documentos relacionados:

- Confiabilidade, trabalho constante e uma boa xícara de café

- AWS e a compartimentalização
- Isolamento de workloads via fragmentação aleatória
- Automatizar implantações seguras e sem intervenção manual

Vídeos relacionados:

- AWS re:Invent 2018: Fechar loops e abrir mentes: como assumir o controle de sistemas grandes e pequenos
- AWS re:Invent 2018: Como a AWS minimiza o raio de ação das falhas (ARC338)
- Fragmentação aleatória: AWS re:Invent 2019: Introdução à Amazon Builders' Library (DOP328)
- AWS Summit ANZ 2021: Tudo falha, o tempo todo: projetando para resiliência

Exemplos relacionados:

- Laboratório do Well-Architected: Isolamento de falhas com fragmentação aleatória

REL 11. Como projetar a workload para resistir a falhas de componentes?

As workloads que exigem alta disponibilidade e baixo tempo médio até a recuperação (MTTR) devem ser projetadas visando a resiliência.

Práticas recomendadas

- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas
- REL11-BP02 Failover para recursos íntegros
- REL11-BP03 Automatizar a reparação em todas as camadas
- REL11-BP04 Confiar no plano de dados, e não no ambiente de gerenciamento, durante a recuperação
- REL11-BP05 Usar estabilidade estática para evitar comportamento bimodal
- REL11-BP06 Enviar notificações quando os eventos afetarem a disponibilidade
- REL11-BP07 Arquitetar o produto para cumprir as metas de disponibilidade e os acordos de serviço (SLAs) de tempo de atividade

REL11-BP01 Monitorar todos os componentes da workload para detectar falhas

Monitore constantemente a integridade da workload para que você e seus sistemas automatizados detectem falhas ou degradações assim que elas ocorrerem. Monitore os indicadores-chave de performance (KPIs) com base no valor empresarial.

Todos os mecanismos de recuperação e correção devem começar com a capacidade de detectar problemas rapidamente. As falhas técnicas devem ser detectadas primeiro para que possam ser resolvidas. No entanto, a disponibilidade é baseada na capacidade da workload de entregar valor empresarial, portanto, os indicadores-chave de performance (KPIs) que medem isso precisam fazer parte da sua estratégia de detecção e remediação.

Resultado desejado: os componentes essenciais de uma workload são monitorados de forma independente para detectar e alertar sobre falhas quando e onde elas acontecem.

Práticas comuns que devem ser evitadas:

- Nenhum alarme foi configurado, portanto as interrupções ocorrem sem notificação.
- Os alarmes existem, mas com limites que não permitem um tempo adequado para reação.
- As métricas não são coletadas com frequência suficiente para atender ao objetivo de tempo de recuperação (RTO).
- Somente as interfaces da workload voltadas para o cliente são monitoradas ativamente.
- Coleta apenas das métricas técnicas, não das métricas de função de negócios.
- Não há métricas que medem a experiência do usuário da workload.
- Monitores em excesso são criados.

Benefícios de implementar esta prática recomendada: o monitoramento adequado de todas as camadas permite reduzir o tempo de recuperação ao reduzir o tempo de detecção.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Identifique todas as workloads que serão analisadas para monitoramento. Depois de identificar todos os componentes da workload que precisarão ser monitorados, será necessário determinar o intervalo de monitoramento. O intervalo de monitoramento terá um impacto direto na rapidez com que a recuperação pode ser iniciada com base no tempo necessário para detectar uma falha. O tempo médio de detecção (MTTD) é a quantidade de tempo entre a ocorrência de uma falha e o início das operações de reparo. A lista de serviços deve ser extensa e completa.

O monitoramento deve abranger todas as camadas da pilha de aplicações, incluindo aplicação, plataforma, infraestrutura e rede.

Sua estratégia de monitoramento deve considerar o impacto de falhas cinzentas. Para obter mais detalhes sobre falhas cinzentas, consulte Falhas cinzentas no whitepaper Padrões de resiliência Multi-AZ avançados.

Etapas de implementação

- O intervalo de monitoramento depende da rapidez com que você precisa fazer a recuperação. O tempo de recuperação é determinado pelo tempo necessário para a recuperação. Desse modo, você deve considerar esse tempo e o objetivo de tempo de recuperação (RTO) para determinar a frequência da coleta.
- Configure o monitoramento detalhado de componentes e serviços gerenciados.
  - Determine se o monitoramento detalhado das instâncias do EC2 e do Auto Scaling é necessário. O monitoramento detalhado fornece métricas de intervalo de um minuto, e o monitoramento padrão fornece métricas de intervalo de cinco minutos.
  - Determine se o monitoramento avançado para RDS é necessário. O monitoramento aprimorado usa um agente nas instâncias do RDS para obter informações úteis sobre processos ou threads diferentes.
  - Determine os requisitos de monitoramento de componentes essenciais sem servidor para Lambda, API Gateway, Amazon EKS, Amazon ECS e todos os tipos de balanceadores de carga.
  - Determine os requisitos de monitoramento dos componentes de armazenamento para Amazon S3, Amazon FSx, Amazon EFS e Amazon EBS.
- Crie métricas personalizadas para medir os indicadores-chave de performance (KPIs) de negócios. As workloads implementam as principais funções empresariais, as quais devem ser usadas como KPIs para ajudar a identificar quando um problema indireto ocorre.
- Utilize os canários de usuário para monitorar a experiência do usuário e verificar se há falhas. O teste de transações sintéticas (também conhecido como teste canário, que não deve ser confundidos com as implantações canário), capaz de executar e simular o comportamento do cliente, está entre os processos de teste mais importantes. Execute esses testes constantemente nos endpoints da workload de diversos locais remotos.
- Crie métricas personalizadas que acompanhem a experiência do usuário Se você puder estabelecer instrumentos de medição da experiência do cliente, conseguirá determinar o momento de degradação da experiência do consumidor.

- Defina alarmes para detectar quando uma parte da workload não estiver funcionando corretamente e indicar quando o ajuste de escala automático dos recursos deve ser feito. Os alarmes podem ser exibidos visualmente em painéis, enviar alertas via Amazon SNS ou e-mail e trabalhar com o Auto Scaling para aumentar ou reduzir a escala dos recursos da workload.
- Crie painéis para visualizar as métricas. É possível usar os painéis para ver as tendências, os casos atípicos e outros indicadores de possíveis problemas ou para obter uma indicação de problemas a serem investigados.
- Crie monitoramento de rastreamento distribuído para seus serviços. Com o monitoramento distribuído, você compreende como está a performance de sua aplicação e seus serviços subjacentes para identificar e solucionar a causa principal de problemas e erros de performance.
- Crie painéis de sistemas de monitoramento (usando CloudWatch ou X-Ray) e coleta de dados em uma região e conta separadas.
- Crie integração com o monitoramento do Amazon Health Aware para permitir a visibilidade de monitoramento de recursos da AWS que possam apresentar degradações. Para workloads essenciais aos negócios, essa solução fornece acesso a alertas proativos e em tempo real para serviços da AWS.

Recursos

Práticas recomendadas relacionadas:

- Definição de disponibilidade
- REL11-BP06 Enviar notificações quando os eventos afetarem a disponibilidade

Documentos relacionados:

- O Amazon CloudWatch Synthetics permite criar canários de usuário
- Habilitar ou desabilitar o monitoramento detalhado da instância
- Monitoramento avançado
- Monitorar grupos do Auto Scaling e instâncias usando o Amazon CloudWatch
- Publicar métricas personalizadas
- Usar alarmes do Amazon CloudWatch
- Usar painéis do CloudWatch
- Usar painéis do CloudWatch entre regiões e contas

- Usar o rastreamento do X-Ray entre regiões e contas
- Noções básicas da disponibilidade
- Implementar o Amazon Health Aware (AHA)

Vídeos relacionados:

- Como mitigar falhas cinzentas

Exemplos relacionados:

- Laboratório do Well-Architected: Nível 300: implementar verificações de integridade e gerenciar dependências para melhorar a confiabilidade
- Workshop One Observability: explorar o X-Ray

Ferramentas relacionadas:

- CloudWatch
- CloudWatch X-Ray

REL11-BP02 Failover para recursos íntegros

Se uma falha ocorrer no recurso, os recursos íntegros deverão continuar atendendo às solicitações. Para falhas de localização (como zona de disponibilidade ou Região da AWS), garanta que você tenha sistemas implementados para realizar failover para recursos íntegros em locais que não apresentam problemas.

Ao projetar um serviço, distribua a carga entre recursos, zonas de disponibilidade ou regiões. Portanto, a falha ou a deficiência de um recurso individual podem ser atenuadas por meio da transferência do tráfego para os recursos íntegros restantes. Pense em como os serviços são descobertos e encaminhados em caso de falha.

Projete seus serviços pensando na recuperação de falhas. Na AWS, projetamos os serviços para minimizar o tempo para recuperação de falhas e o impacto sobre os dados. Nossos serviços usam principalmente datastores que reconhecem solicitações apenas após serem armazenadas de modo durável entre várias réplicas em uma região. Eles são criados para usar isolamento com base em células e usar o isolamento de falhas fornecido por zonas de disponibilidade. Usamos automação

extensivamente em nossos procedimentos operacionais. Também otimizamos nossa funcionalidade de substituir e reiniciar para a recuperação rápida de interrupções.

Os padrões e os designs que permitem o failover variam para cada serviço de plataforma da AWS. Muitos serviços gerenciados nativos da AWS são nativamente várias zonas de disponibilidade (como o Lambda ou o API Gateway). Outros serviços da AWS (como EC2 e EKS) exigem designs específicos de práticas recomendadas para oferecer compatibilidade com o failover de recursos ou armazenamento de dados entre AZs.

O monitoramento deve ser configurado para conferir se o recurso de failover está íntegro, rastrear o andamento do failover dos recursos e monitorar a recuperação do processo empresarial.

Resultado desejado: os sistemas são capazes de usar novos recursos de forma automática ou manual para se recuperarem da degradação.

Práticas comuns que devem ser evitadas:

- Planejar o fracasso não faz parte da fase de planejamento e design.
- O RTO e o RPO não são estabelecidos.
- Monitoramento insuficiente para detectar falhas nos recursos.
- Isolamento adequado dos domínios de falha.
- O failover multirregiões não é considerado.
- A detecção de falhas é sensível ou agressiva demais ao decidir realizar o failover.
- Não testar nem validar o design de failover.
- Executar a automação de autocorreção sem notificar que a reparação era necessária.
- Falta de um período de amortecimento a fim de evitar o failback cedo demais.

Benefícios de implementar esta prática recomendada: é possível criar sistemas mais resilientes que mantenham a confiabilidade em caso de falhas, degradando-se normalmente e se recuperando com rapidez.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Os serviços da AWS como o Elastic Load Balancing e o Amazon EC2 Auto Scaling ajudam a distribuir a carga entre recursos e zonas de disponibilidade. Portanto, a falha de um recurso

individual (como uma instância do EC2) ou o comprometimento de uma zona de disponibilidade podem ser atenuados por meio do desvio do tráfego para os recursos íntegros restantes.

Para workloads multirregiões, os designs são mais complicados. Por exemplo, as réplicas de leitura entre regiões permitem que você implante os dados em várias Regiões da AWS. No entanto, o failover ainda é necessário para promover a réplica de leitura como primária e direcionar seu tráfego para o novo endpoint. O Amazon Route 53, o Amazon Application Recovery Controller (ARC), o Amazon CloudFront e o AWS Global Accelerator podem ajudar a direcionar o tráfego nas Regiões da AWS.

Os serviços da AWS como Amazon S3, Lambda, API Gateway, Amazon SQS, Amazon SNS, Amazon SES, Amazon Pinpoint, Amazon ECR, AWS Certificate Manager, EventBridge ou Amazon DynamoDB são implantados automaticamente em várias zonas de disponibilidade pela AWS. Em caso de falha, esses serviços da AWS direcionam automaticamente o tráfego para locais íntegros. Os dados são armazenados de forma redundante em várias zonas de disponibilidade e permanecem disponíveis.

O Multi-AZ é uma opção de configuração para o Amazon RDS, Amazon Aurora, Amazon Redshift, Amazon EKS ou Amazon ECS. A AWS pode direcionar o tráfego para a instância íntegra se o failover for iniciado. Essa ação de failover pode ser realizada pela AWS ou conforme exigido pelo cliente.

Para instâncias do Amazon EC2, o Amazon Redshift, tarefas do Amazon ECS ou pods do Amazon EKS, escolha em quais zonas de disponibilidade deseja fazer a implantação. Em alguns designs, o Elastic Load Balancing fornece a solução para detectar as instâncias nas zonas com problemas de integridade e rotear o tráfego para as instâncias íntegras. O Elastic Load Balancing também pode rotear tráfego para componentes no seu datacenter on-premises.

No caso de failover de tráfego em várias regiões, o redirecionamento pode utilizar o Amazon Route 53, o Amazon Application Recovery Controller, o AWS Global Accelerator, o Route 53 Private DNS para VPCs ou o CloudFront para oferecer uma maneira de definir domínios da internet e atribuir políticas de roteamento, incluindo verificações de integridade, e rotear o tráfego para regiões íntegras. O AWS Global Accelerator fornece endereços IP estáticos que atuam como um ponto de entrada fixo para a aplicação e, depois, são roteados para os endpoints nas Regiões da AWS de sua escolha, usando a rede global da AWS em vez da internet com o objetivo de melhorar a performance e a confiabilidade.

Etapas de implementação

- Crie designs de failover para todas as aplicações e serviços apropriados. Isole cada componente da arquitetura e crie designs de failover que atendam ao RTO e ao RPO de cada componente.
- Configure ambientes inferiores (como desenvolvimento ou teste) com todos os serviços necessários para ter um plano de failover. Implemente as soluções usando a infraestrutura como código (IaC) para garantir repetibilidade.
- Configure um local de recuperação, como uma segunda região, para implementar e testar os designs de failover. Se necessário, os recursos para testes podem ser configurados temporariamente para limitar os custos adicionais.
- Determine quais planos de failover são automatizados pela AWS, quais podem ser automatizados por um processo de DevOps e quais podem ser manuais. Documente e avalie o RTO e o RPO de cada serviço.
- Crie um playbook de failover e inclua todas as etapas para realizar o failover de cada recurso, aplicação e serviço.
- Crie um playbook de failback e inclua todas as etapas de failback (com tempo) de cada recurso, aplicação e serviço.
- Crie um plano para iniciar e ensaiar o playbook. Use simulações e testes de caos para testar a automação e as etapas do playbook.
- Para falhas de localização (como zona de disponibilidade ou Região da AWS), garanta que você tenha sistemas implementados para realizar failover para recursos íntegros em locais que não apresentam problemas. Confira a cota, os níveis de ajuste de escala automático e os recursos em execução antes do teste de failover.

Recursos

Práticas recomendadas do Well-Architected relacionadas:

- REL13: planejar para DR
- REL10: usar o isolamento de falhas para proteger a workload

Documentos relacionados:

- Definir metas de RTO e RPO
- Failover usando o roteamento ponderado do Route 53

- Disaster Recovery with Amazon Route 53 Application Recovery Controller (ARC)
- EC2 com ajuste de escala automático
- Implantações do EC2: Multi-AZ
- Implantações do ECS: Multi-AZ
- Switch traffic using Amazon Application Recovery Controller
- Lambda com um Application Load Balancer e failover
- Replicação e failover do ACM
- Replicação e failover do repositório de parâmetros
- Replicação entre regiões e failover do ECR
- Configuração da replicação entre regiões do Secrets Manager
- Habilitar a replicação entre regiões para EFS e failover
- Replicação entre regiões e failover do EFS
- Failover de rede
- Failover de endpoint do S3 usando MRAP
- Criar replicação entre regiões para o S3
- Guidance for Cross Region Failover and Graceful Failback on AWS
- Failover com o acelerador global multirregiões
- Failover com DRS
- Creating Disaster Recovery Mechanisms Using Amazon Route 53

Exemplos relacionados:

- Recuperação de desastres na AWS
- Recuperação elástica de desastres na AWS

REL11-BP03 Automatizar a reparação em todas as camadas

Após a detecção de uma falha, use recursos automatizados para executar ações de correção. As degradações podem ser corrigidas automaticamente por meio de mecanismos internos de serviço ou exigir que os recursos sejam reiniciados ou removidos por meio de ações de remediação.

Para aplicações autogerenciadas e reparação entre regiões, os projetos de recuperação e os processos de recuperação automatizados podem ser extraídos de práticas recomendadas existentes.

A capacidade de reiniciar ou remover um recurso é uma ferramenta importante para corrigir falhas. Uma prática recomendada é deixar os serviços sem estado sempre que possível. Isso evita a perda de dados ou disponibilidade na reinicialização do recurso. Na nuvem, você pode (e geralmente deve) substituir todo o recurso (por exemplo, uma instância de computação ou função sem servidor) como parte da reinicialização. A reinicialização em si é uma maneira simples e confiável de se recuperar de falhas. Muitos tipos diferentes de falhas ocorrem em workloads. As falhas podem ocorrer em hardware, software, comunicações e operações.

Reiniciar ou tentar novamente também se aplica às solicitações de rede. Aplique a mesma abordagem de recuperação tanto a um tempo limite de rede quanto a uma falha de dependência em que a dependência retorna um erro. Ambos os eventos têm um efeito similar sobre o sistema. Assim, em vez de tentar tornar qualquer um dos eventos um caso especial, aplique uma estratégia similar de nova tentativa limitada com recuo exponencial e jitter. A capacidade de reiniciar é um mecanismo de recuperação presente na computação orientada para a recuperação e arquiteturas de cluster de alta disponibilidade.

Resultado desejado: ações automatizadas são executadas para corrigir a detecção de uma falha.

Práticas comuns que devem ser evitadas:

- Provisionamento de recursos sem dimensionamento automático.
- Implantação de aplicações em instâncias ou contêineres individualmente.
- Implantação de aplicações que não podem ser implantadas em vários locais sem usar a recuperação automática.
- Reparação manual de aplicações que não são reparadas por meio do ajuste de escala automático e da recuperação automática.
- Sem automação para failover dos bancos de dados.
- Não há métodos automatizados para redirecionar o tráfego para novos endpoints.
- Sem replicação de armazenamento.

Benefícios de implementar esta prática recomendada: a reparação automatizada pode reduzir seu tempo médio de recuperação e melhorar sua disponibilidade.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Os designs para Amazon EKS ou outros serviços do Kubernetes devem incluir conjuntos mínimos e máximos de réplicas ou com estado e tamanho mínimo do cluster e do grupo de nós. Esses mecanismos fornecem uma quantidade mínima de recursos de processamento continuamente disponíveis e, ao mesmo tempo, remediam automaticamente quaisquer falhas usando o ambiente de gerenciamento do Kubernetes.

Os padrões de design que são acessados por meio de um balanceador de carga usando clusters de computação devem utilizar grupos do Auto Scaling. O Elastic Load Balancing (ELB) distribui automaticamente o tráfego de entrada das aplicações entre vários destinos e appliances virtuais em uma ou mais Zonas de Disponibilidade (AZs).

Designs baseados em computação em cluster que não usam balanceamento de carga devem ter seu tamanho projetado para a perda de pelo menos um nó. Isso permitirá que o serviço se mantenha funcionando em uma capacidade potencialmente reduzida enquanto recupera um novo nó. Exemplos de serviços são Mongo, DynamoDB Accelerator, Amazon Redshift, Amazon EMR, Cassandra, Kafka, MSK-EC2, Couchbase, ELK e Amazon OpenSearch Service. Muitos desses serviços podem ser projetados com recursos adicionais de recuperação automática. Algumas tecnologias de cluster devem gerar um alerta sobre a perda de um nó acionando um fluxo de trabalho automatizado ou manual para recriar um novo nó. Esse fluxo de trabalho pode ser automatizado usando o AWS Systems Manager para corrigir problemas rapidamente.

É possível usar o Amazon EventBridge para monitorar e filtrar eventos, como alarmes do CloudWatch ou alterações no estado de outros serviços da AWS. Com base nas informações do evento, ele pode invocar o AWS Lambda, o Systems Manager Automation (ou outros destinos) para executar a lógica de correção personalizada na workload. O Amazon EC2 Auto Scaling pode ser configurado para verificar a integridade da instância do EC2. Se a instância estiver em qualquer estado que não seja em execução, ou se o status do sistema for prejudicado, o Amazon EC2 Auto Scaling considerará a instância como não íntegra e iniciará uma instância de substituição. Para substituições em grande escala (como a perda de uma zona de disponibilidade inteira), a estabilidade estática é preferida para alta disponibilidade.

Etapas de implementação

- Use grupos do Auto Scaling para implantar camadas em uma workload. O Auto Scaling pode executar a autocorreção em aplicações sem estado e adicionar e remover capacidade.

- Para instâncias computacionais mencionadas anteriormente, use o balanceamento de carga e escolha o tipo apropriado de balanceador de carga.
- Considere a possibilidade de reparação do Amazon RDS. Com instâncias em espera, configure o failover automático para a instância em espera. Para a réplica de leitura do Amazon RDS, é necessário um fluxo de trabalho automatizado para transformar uma réplica de leitura em primária.
- Implemente a recuperação automática em instâncias do EC2 que tenham aplicações implantadas que não possam ser implantadas em vários locais e possam tolerar a reinicialização em caso de falhas. É possível usar a recuperação automática para substituir o hardware com falha e reiniciar a instância quando a aplicação não puder ser implantada em vários locais. Os metadados e os endereços IP associados da instância são mantidos, assim como os volumes do EBS e os pontos de montagem para Amazon Elastic File Systems ou File Systems para Lustre e Windows. Com o AWS OpsWorks, é possível configurar a autocorreção das instâncias do EC2 no nível da camada.
- Implemente a recuperação automatizada por meio do AWS Step Functions e do AWS Lambda quando não for possível usar o ajuste de escala automático ou a recuperação automática, ou quando a recuperação automática falhar. Quando não for possível usar o ajuste de escala automático e a recuperação automática ou quando a recuperação automática falhar, você poderá automatizar a reparação usando o AWS Step Functions e o AWS Lambda.
- É possível usar o Amazon EventBridge para monitorar e filtrar eventos, como alarmes do CloudWatch ou alterações no estado de outros serviços da AWS. Com base nas informações do evento, ele pode invocar o AWS Lambda (ou outros destinos) para executar a lógica de correção personalizada na workload.

Recursos

Práticas recomendadas relacionadas:

- Definição de disponibilidade
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas

Documentos relacionados:

- Como o AWS Auto Scaling funciona
- Recuperação automática do Amazon EC2
- Amazon Elastic Block Store (Amazon EBS)
- Amazon Elastic File System (Amazon EFS)

- O que é o Amazon FSx para Lustre?
- O que é o Amazon FSx para Windows File Server?
- AWS OpsWorks: Como usar a reparação automática para substituir instâncias com falha
- O que é AWS Step Functions?
- O que é AWS Lambda?
- O que é o Amazon EventBridge?
- Usar alarmes do Amazon CloudWatch
- Failover do Amazon RDS
- SSM: Systems Manager Automation
- Práticas recomendadas de arquitetura resiliente

Vídeos relacionados:

- Provisionar e escalar automaticamente o serviço OpenSearch
- Failover automático do Amazon RDS

Exemplos relacionados:

- Workshop sobre Auto Scaling
- Workshop Failover do Amazon RDS

Ferramentas relacionadas:

- CloudWatch
- CloudWatch X-Ray

REL11-BP04 Confiar no plano de dados, e não no ambiente de gerenciamento, durante a recuperação

Os ambientes de gerenciamento fornecem as APIs administrativas usadas para criar, ler e descrever, atualizar, excluir e listar recursos (CRUDL), enquanto os planos de dados lidam com o tráfego diário de serviços. Ao implementar respostas de recuperação ou mitigação a eventos potencialmente impactantes na resiliência, concentre-se em usar um número mínimo de operações do ambiente de

gerenciamento para recuperar, redimensionar, restaurar, reparar ou realizar o failover do serviço. A ação do plano de dados deve substituir qualquer atividade durante esses eventos de degradação.

Por exemplo, estas são ações do ambiente de gerenciamento: iniciar uma nova instância de computação, criar armazenamento em bloco e descrever serviços de fila. Quando você executa instâncias de computação, o ambiente de gerenciamento precisa realizar várias tarefas, como encontrar um host físico com capacidade, alocar interfaces de rede, preparar volumes de armazenamento em blocos locais, gerar credenciais e adicionar regras de segurança. Os ambientes de gerenciamento tendem a ser uma orquestração complicada.

Resultado desejado: quando um recurso entra em um estado comprometido, o sistema é capaz de se recuperar automática ou manualmente, transferindo o tráfego de recursos danificados para recursos saudáveis.

Práticas comuns que devem ser evitadas:

- Dependência da alteração dos registros DNS para redirecionar o tráfego.
- Dependência das operações de escalação do ambiente de gerenciamento para substituir componentes danificados devido a recursos insuficientemente provisionados.
- Dependência de ações de ambiente de gerenciamento abrangentes, com vários serviços e várias APIs para remediar qualquer categoria de deficiência.

Benefícios de implementar esta prática recomendada: o aumento da taxa de sucesso da remediação automatizada pode reduzir seu tempo médio de recuperação e melhorar a disponibilidade da workload.

Nível de risco exposto se essa prática recomendada não for estabelecida: Médio. Para certos tipos de degradação do serviço, os ambientes de gerenciamento são afetados. As dependências do uso extensivo do ambiente de gerenciamento para remediação podem aumentar o tempo de recuperação (RTO) e o tempo médio de recuperação (MTTR).

Orientação para implementação

Para limitar as ações do plano de dados, avalie cada serviço quanto às ações necessárias para restaurar o serviço.

Utilize o Amazon Application Recovery Controller para mudar o tráfego de DNS. Esses recursos monitoram continuamente a capacidade da aplicação de se recuperar de falhas, permitindo que

você controle a recuperação da aplicação em várias Regiões da AWS, Zonas de Disponibilidade e ambientes on-premises.

As políticas de roteamento do Route 53 usam o ambiente de gerenciamento. Portanto, não confie nele para recuperação. Os planos de dados do Route 53 respondem às consultas ao DNS, além de realizarem e avaliarem verificações de integridade. Eles são distribuídos globalmente e projetados para um acordo de serviço (SLA) com 100% de disponibilidade.

As APIs e os consoles de gerenciamento do Route 53 usados para criar, atualizar e excluir recursos do Route 53 são executados em ambientes de gerenciamento projetados para priorizar a consistência e a durabilidade necessária para gerenciar o DNS. Para que isso aconteça, os ambientes de gerenciamento estão localizados em uma única região: Leste dos EUA (Norte da Virgínia). Embora ambos os sistemas sejam construídos para serem muito confiáveis, os ambientes de gerenciamento não estão incluídos no SLA. Pode ser que ocorram raros eventos onde o design resiliente do plano de dados permita que ele mantenha a disponibilidade, enquanto os ambientes de gerenciamento não. Para mecanismos de recuperação de desastres e failover, use funções de plano de dados para fornecer a melhor confiabilidade possível.

Projete sua infraestrutura de computação de modo que ela seja estaticamente estável para evitar o uso do ambiente de gerenciamento durante um incidente. Por exemplo, se você estiver usando instâncias do Amazon EC2, evite provisionar novas instâncias manualmente ou instruir grupos do Auto Scaling a adicionar instâncias em resposta. Para obter os níveis mais altos de resiliência, provisione capacidade suficiente no cluster usado para failover. Se essa capacidade precisar ser limitada, defina valores no sistema geral completo para definir com segurança o tráfego total que atinge o conjunto limitado de recursos.

Para serviços como Amazon DynamoDB, Amazon API Gateway, balanceadores de carga e AWS Lambda sem servidor, a utilização desses serviços faz uso do plano de dados. No entanto, criar novas funções, balanceadores de carga, API gateways ou tabelas do DynamoDB é uma ação do ambiente de gerenciamento e deve ser concluída antes da degradação, como preparação para um evento e ensaio das ações de failover. Para o Amazon RDS, as ações do plano de dados permitem o acesso aos dados.

Para obter mais informações sobre planos de dados, ambientes de gerenciamento e como a AWS cria serviços para atender às metas de alta disponibilidade, consulte o whitepaper Estabilidade estática usando zonas de disponibilidade.

Entenda quais operações estão no plano de dados e quais estão no ambiente de gerenciamento.

Etapas de implementação

Para cada workload que precisa ser restaurada após um evento de degradação, avalie o runbook de failover, o projeto de alta disponibilidade, o projeto de recuperação automática ou o plano de restauração de recursos de HA. Identifique cada ação que pode ser considerada uma ação do ambiente de gerenciamento.

Considere alterar a ação de gerenciamento para uma ação do plano de dados:

- Auto Scaling (ambiente de gerenciamento) para recursos do Amazon EC2 pré-escalados (plano de dados)
- Ajuste de escala de instâncias do Amazon EC2 (ambiente de gerenciamento) para ajuste de escala do AWS Lambda (plano de dados)
- Avalie qualquer design usando o Kubernetes e a natureza das ações do ambiente de gerenciamento. Adicionar pods é uma ação do plano de dados no Kubernetes. As ações devem se limitar à adição de pods e não adição de nós. Usar nós superprovisionados é o método preferido para limitar as ações do ambiente de gerenciamento

Considere abordagens alternativas que permitam que as ações do plano de dados afetem a mesma remediação.

- Alteração de registro do Route 53 (ambiente de gerenciamento) ou Amazon Application Recovery Controller (plano de dados)
- Verificações de integridade do Route 53 para atualizações mais automatizadas

Se o serviço for essencial, considere alguns serviços em uma região secundária para permitir mais ações no ambiente de gerenciamento e no plano de dados em uma região não afetada.

- Amazon EC2 Auto Scaling ou Amazon EKS em uma região primária em comparação com Amazon EC2 Auto Scaling ou Amazon EKS em uma região secundária e roteamento de tráfego para região secundária (ação do ambiente de gerenciamento)
- Faça uma réplica de leitura na primária secundária ou tente a mesma ação na região primária (ação do ambiente de gerenciamento).

Recursos

Práticas recomendadas relacionadas:

- Definição de disponibilidade
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudar na automação da tolerância a falhas
- AWS Marketplace: produtos que podem ser usados para tolerância a falhas
- Amazon Builders' Library: evitar a sobrecarga em sistemas distribuídos colocando o menor serviço no controle
- API do Amazon DynamoDB (ambiente de gerenciamento e plano de dados)
- Execuções do AWS Lambda (divididas entre o ambiente de gerenciamento e o plano de dados)
- Plano de dados de AWS Elemental MediaStore
- Building highly resilient applications using Amazon Application Recovery Controller, Part 1: Single-Region stack
- Building highly resilient applications using Amazon Application Recovery Controller, Part 2: Multi-Region stack
- Creating Disaster Recovery Mechanisms Using Amazon Route 53
- What is Amazon Application Recovery Controller
- Ambiente de gerenciamento e plano de dados do Kubernetes

Vídeos relacionados:

- De volta ao básico: uso da estabilidade estática
- Como criar workloads resilientes em vários sites usando serviços globais da AWS

Exemplos relacionados:

- Introducing Amazon Application Recovery Controller
- Amazon Builders' Library: evitar a sobrecarga em sistemas distribuídos colocando o menor serviço no controle
- Building highly resilient applications using Amazon Application Recovery Controller, Part 1: Single-Region stack

- Building highly resilient applications using Amazon Application Recovery Controller, Part 2: Multi-Region stack
- Estabilidade estática com zonas de disponibilidade

Ferramentas relacionadas:

- Amazon CloudWatch
- AWS X-Ray

REL11-BP05 Usar estabilidade estática para evitar comportamento bimodal

As workloads devem ser estaticamente estáveis e operar somente em um único modo normal. O comportamento bimodal ocorre quando a workload exibe um comportamento diferente nos modos normal e de falha.

Por exemplo, você pode tentar se recuperar de uma falha na zona de disponibilidade iniciando novas instâncias em uma zona de disponibilidade diferente. Isso pode resultar em uma resposta bimodal durante um modo de falha. Em vez disso, você deve criar workloads que sejam estaticamente estáveis e que operem em apenas um modo. Neste exemplo, essas instâncias deveriam ter sido provisionadas na segunda zona de disponibilidade antes da falha. Esse design de estabilidade estática verifica se a workload opera somente em um único modo.

Resultado desejado: as workloads não apresentam comportamento bimodal durante os modos normal e de falha.

Práticas comuns que devem ser evitadas:

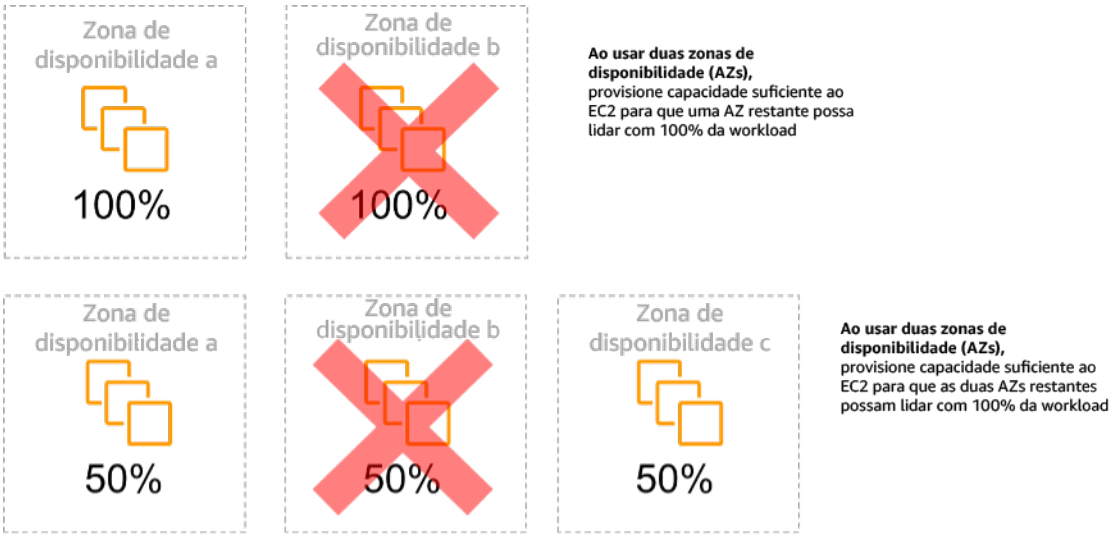
- Supor que os recursos sempre possam ser provisionados, independentemente do escopo da falha.
- Tentar adquirir recursos dinamicamente durante uma falha.
- Não provisionar recursos adequados entre zonas ou regiões até que ocorra uma falha.
- Pensar em projetos estáticos estáveis somente para recursos computacionais.

Benefícios de implementar esta prática recomendada: as workloads executadas com projetos estaticamente estáveis podem ter resultados previsíveis durante eventos normais e de falha.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

O comportamento bimodal ocorre quando a workload apresenta um comportamento diferente nos modos normal e de falha (por exemplo, depender da inicialização de novas instâncias se uma zona de disponibilidade falhar). Um exemplo de comportamento bimodal é quando designs estáveis do Amazon EC2 provisionam instâncias suficientes em cada zona de disponibilidade para lidar com a workload se uma AZ for removidas. O Elastic Load Balancing ou o Amazon Route 53 verificariam a integridade para afastar a carga das instâncias danificadas. Depois que o tráfego for deslocado, use o AWS Auto Scaling para substituir de forma assíncrona instâncias da zona com falha e executá-las nas zonas íntegras. A estabilidade estática para implantação de computação (como instâncias ou contêineres do EC2) resulta na mais alta confiabilidade.



Estabilidade estática de instâncias do EC2 em várias zonas de disponibilidade

Isso deve ser comparado ao custo desse modelo e ao valor comercial de manter a workload em todos os casos de resiliência. É mais barato provisionar menos capacidade computacional e depender da inicialização de novas instâncias em caso de falha. No entanto, para falhas em grande escala (como um dano regional ou da zona de disponibilidade), essa abordagem é menos eficaz porque depende tanto de um plano operacional quanto de recursos suficientes disponíveis nas zonas ou regiões não afetadas.

A solução também deve comparar a confiabilidade com os custos necessários para a workload. As arquiteturas de estabilidade estática se aplicam a uma variedade de arquiteturas, incluindo instâncias de computação espalhadas por zonas de disponibilidade, designs de réplicas de leitura de banco de dados, designs de cluster Kubernetes (EKS) e arquiteturas de failover multirregiões.

Também é possível implementar um design mais estável estaticamente usando mais recursos em cada zona. Ao adicionar mais zonas, você reduz a quantidade de computação adicional necessária para a estabilidade estática.

Um exemplo de comportamento bimodal seria um tempo limite de rede que poderia fazer com que um sistema tentasse atualizar seu próprio estado de configuração por completo. Isso adicionaria uma carga inesperada a outro componente e poderia fazê-lo falhar, resultando em outras consequências inesperadas. Esse ciclo de feedback negativo afeta a disponibilidade da workload. Em vez disso, você pode criar sistemas estaticamente estáveis e operar em apenas um modo. Um design estático estável faria um trabalho constante e sempre atualizaria o estado da configuração em um ritmo fixo. Quando uma chamada falha, a workload usa o valor previamente armazenado em cache e inicia um alarme.

Outro exemplo de comportamento bimodal é permitir que os clientes ignorem o cache da workload em caso de falhas. Essa pode parecer uma solução que acomoda as necessidades do cliente, mas pode alterar significativamente as demandas da workload e provavelmente resultar em falhas.

Avalie workloads importantes para determinar quais workloads exigem esse tipo de projeto de resiliência. Para as que são consideradas críticas, cada componente da aplicação deve ser revisado. Exemplos de tipos de serviço que exigem avaliações de estabilidade estática são:

- Computação: Amazon EC2, EKS-EC2, ECS-EC2, EMR-EC2
- Bancos de dados: Amazon Redshift, Amazon RDS, Amazon Aurora
- Armazenamento: Amazon S3 (zona única), Amazon EFS (montagens), Amazon FSx (montagens)
- Balanceadores de carga: sob determinados designs

Etapas de implementação

- Crie sistemas que sejam estaticamente estáveis e que operem em apenas um modo. Nesse caso, provisione instâncias suficientes em cada zona de disponibilidade ou região para lidar com a capacidade da workload se uma zona de disponibilidade ou região for removida. Diversos serviços podem ser usados para roteamento a recursos íntegros, como:
  - Roteamento de DNS entre regiões
  - Roteamento multirregiões MRAP do Amazon S3
  - AWS Global Accelerator
  - Amazon Application Recovery Controller

- Configure réplicas de leitura de banco de dados para contabilizar a perda de uma única instância principal ou de uma réplica de leitura. Se o tráfego estiver sendo servido por réplicas de leitura, a quantidade em cada zona de disponibilidade e cada região deve ser igual à necessidade geral em caso de falha na zona ou região.
- Configure dados críticos no armazenamento do S3, projetado para ser estaticamente estável para dados armazenados em caso de falha na zona de disponibilidade. Se a classe de armazenamento One Zone-IA do Amazon S3 for usada, ela não deverá ser considerada estaticamente estável, pois a perda dessa zona minimiza o acesso a esses dados armazenados.
- Os balanceadores de carga algumas vezes são configurados incorreta ou intencionalmente para atender a uma zona de disponibilidade específica. Nesse caso, o design estaticamente estável pode envolver a distribuição de uma workload entre várias zonas de disponibilidade em um design mais complexo. O design original pode ser usado para reduzir o tráfego entre zonas por motivos de segurança, latência ou custo.

Recursos

Práticas recomendadas do Well-Architected relacionadas:

- Definição de disponibilidade
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas
- REL11-BP04 Confiar no plano de dados, e não no ambiente de gerenciamento, durante a recuperação

Documentos relacionados:

- Minimizar dependências em um plano de recuperação de desastres
- Amazon Builders' Library: estabilidade estática usando zonas de disponibilidade
- Limites de isolamento de falhas
- Estabilidade estática usando zonas de disponibilidade
- RDS Multi-AZ
- Minimizar dependências em um plano de recuperação de desastres
- Roteamento de DNS entre regiões
- Roteamento multirregiões MRAP do Amazon S3
- AWS Global Accelerator

- Amazon Application Recovery Controller
- Amazon S3 de zona única
- Balanceamento de carga entre zonas

Vídeos relacionados:

- Estabilidade estática na AWS: AWS re:Invent 2019: introdução à Amazon Builders' Library (DOP328)

REL11-BP06 Enviar notificações quando os eventos afetarem a disponibilidade

As notificações são enviadas após a detecção de limites violados, mesmo que o evento causado pelo problema tenha sido resolvido automaticamente.

A correção automatizada permite que a workload seja confiável. No entanto, ele também pode ocultar problemas subjacentes que precisam ser resolvidos. Implemente eventos e monitoramento apropriados para que você possa detectar padrões de problemas, incluindo aqueles abordados pela autocorreção, e consiga resolver problemas de causa-raiz.

Os sistemas resilientes são projetados para que os eventos de degradação sejam comunicados imediatamente às equipes apropriadas. Essas notificações devem ser enviadas por meio de um ou vários canais de comunicação.

Resultado desejado: os alertas são enviados imediatamente às equipes de operações quando os limites são violados. Esses alertas podem incluir taxas de erro, latência ou outras métricas importantes de indicadores-chave de performance (KPI), permitindo que esses problemas sejam resolvidos o mais rápido possível e o impacto do usuário seja evitado ou minimizado.

Práticas comuns que devem ser evitadas:

- Enviar muitos alarmes.
- Enviar alarmes não acionáveis.
- Definir limites de alarme muito altos (supersensíveis) ou muito baixos (subsensíveis).
- Não enviar alarmes para dependências externas.
- Não considerar falhas cinzentas ao projetar monitoramento e alarmes.
- Executar a automação da correção, mas sem notificar a equipe apropriada de que a correção era necessária.

Benefícios de implementar esta prática recomendada: as notificações de recuperação alertam as equipes operacionais e empresariais sobre as degradações do serviço para que elas possam reagir imediatamente a fim de minimizar o tempo médio de detecção (MTTD) e o tempo médio de reparo (MTTR). As notificações de eventos de recuperação também garantem que você não ignore problemas que ocorrem com pouca frequência.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio A falha na implementação de mecanismos adequados de monitoramento e notificação de eventos pode resultar em falha na detecção de padrões de problemas, incluindo aqueles resolvidos pela recuperação automática. Uma equipe só será informada da degradação do sistema quando os usuários entrarem em contato com o atendimento ao cliente ou por acaso.

Orientação para implementação

Ao definir uma estratégia de monitoramento, um alarme acionado é um evento comum. Esse evento provavelmente conteria um identificador para o alarme, o estado do alarme (como `IN ALARM` e `OK`) e detalhes sobre o que o acionou. Em muitos casos, um evento de alarme deve ser detectado e uma notificação por e-mail deve ser enviada. Este é um exemplo de uma ação em um alarme. A notificação do alarme é fundamental para a observabilidade, pois informa às pessoas certas que há um problema. No entanto, quando a ação sobre eventos amadurece em sua solução de observabilidade, ela pode corrigir automaticamente o problema sem a necessidade de intervenção humana.

Depois que os alarmes de monitoramento de KPI forem estabelecidos, os alertas deverão ser enviados às equipes apropriadas quando os limites forem excedidos. Esses alertas também podem ser usados para acionar processos automatizados que tentarão remediar a degradação.

Para um monitoramento de limites mais complexo, considere usar alarmes compostos. Eles usam vários alarmes de monitoramento de KPI para criar um alerta com base na lógica operacional de negócios. Os alarmes do CloudWatch podem ser configurados para enviar e-mails ou registrar incidentes em sistemas de rastreamento de incidentes de terceiros usando a integração com o Amazon SNS ou Amazon EventBridge.

Etapas de implementação

Crie vários tipos de alarme com base na forma como as workloads são monitoradas, por exemplo:

- Os alarmes de aplicações são usados para detectar quando alguma parte da workload não está funcionando adequadamente.

- Os alarmes de infraestrutura indicam quando escalar os recursos. Os alarmes podem ser exibidos visualmente em painéis, enviar alertas via Amazon SNS ou e-mail e trabalhar com o Auto Scaling para aumentar ou reduzir a escala dos recursos da workload.
- Alarmes estáticos de exemplo podem ser criados para monitorar quando uma métrica ultrapassa um limite estático durante um número específico de períodos de avaliação.
- Os alarmes compostos podem representar alarmes complexos de várias origens.
- Depois que o alarme for criado, crie eventos de notificação apropriados. Você pode invocar diretamente uma API do Amazon SNS para enviar notificações e vincular qualquer automação para remediação ou comunicação.
- Integre o monitoramento do Amazon Health Aware para permitir visibilidade de monitoramento de recursos da AWS que possam apresentar degradações. Para workloads essenciais aos negócios, essa solução fornece acesso a alertas proativos e em tempo real para serviços da AWS.

Recursos

Práticas recomendadas do Well-Architected relacionadas:

- Definição de disponibilidade

Documentos relacionados:

- Criar um alarme do CloudWatch com base em um limite estático
- O que é o Amazon EventBridge?
- O que é o Amazon Simple Notification Service?
- Publicar métricas personalizadas
- Usar alarmes do Amazon CloudWatch
- Amazon Health Aware (AHA)
- Configurar alarmes do CloudWatch Composite
- Novidades no AWS Observability na re:Invent 2022

Ferramentas relacionadas:

- CloudWatch
- CloudWatch X-Ray

REL11-BP07 Arquitetar o produto para cumprir as metas de disponibilidade e os acordos de serviço (SLAs) de tempo de atividade

Arquitete o produto para cumprir as metas de disponibilidade e os acordos de serviço (SLAs) de tempo de atividade. Se você publicar ou concordar de forma privada com as metas de disponibilidade ou SLAs de tempo de atividade, verifique se sua arquitetura e seus processos operacionais foram projetados para comportá-los.

Resultado desejado: cada aplicação tem uma meta definida de disponibilidade e um SLA para métricas de performance, as quais podem ser monitoradas e mantidas para atingir os resultados comerciais.

Práticas comuns que devem ser evitadas:

- Planejar e implantar workloads sem definir SLAs.
- As métricas de SLA são definidas muito altas sem justificativas ou requisitos comerciais.
- Definir SLAs sem considerar as dependências e o SLA subjacente.
- Os designs das aplicações são criados sem considerar o modelo de responsabilidade compartilhada para resiliência.

Benefícios de implementar esta prática recomendada: desenvolver aplicações com base nas principais metas de resiliência ajuda a atingir os objetivos de negócios e as expectativas dos clientes. Esses objetivos ajudam a orientar o processo de design da aplicação que avalia diferentes tecnologias e considera as vantagens e desvantagens.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Os designs da aplicação precisam levar em conta um conjunto de requisitos diversos que são derivados de objetivos empresariais, operacionais e financeiros. Nos requisitos operacionais, as workloads precisam ter metas de métricas de resiliência específicas para que possam ser monitorados e comportados adequadamente. As métricas de resiliência não devem ser definidas nem derivadas depois de implantar a workload. Elas devem ser definidas durante a fase de design e ajudar a orientar as diversas decisões e concessões.

- Cada workload deve ter seu próprio conjunto de métricas de resiliência. Essas métricas podem ser diferentes de outras aplicações empresariais.

- Reduzir as dependências pode ter um impacto positivo na disponibilidade. Cada workload deve considerar suas dependências e seus SLAs. Em geral, escolha dependências com metas de disponibilidade iguais ou maiores que as metas da workload.
- Considere designs com acoplamento fraco para que a workload possa operar corretamente apesar do comprometimento da dependência, quando possível.
- Reduza as dependências do ambiente de gerenciamento, especialmente durante uma recuperação ou degradação. Avalie os designs estaticamente estáveis com relação às workloads essenciais à missão. Use a economia de recursos para aumentar a disponibilidade dessas dependências em uma workload.
- A capacidade de observação e a instrumentalização são críticas para cumprir os SLAs reduzindo o tempo médio de detecção (MTTD) e o tempo médio de reparo (MTTR).
- Falhas menos frequentes (MTBF mais longo), tempos de detecção de falhas mais curtos (MTTD mais curto) e tempos de reparo mais curtos (MTTR mais curto) são os três fatores usados para melhorar a disponibilidade em sistemas distribuídos.
- Estabelecer e cumprir métricas de resiliência para uma workload é fundamental para qualquer design eficaz. Esses designs devem levar em consideração as vantagens e desvantagens da complexidade de design, as dependências do serviço, a performance, o ajuste de escala e os custos.

Etapas de implementação

- Analise e documente o design da workload considerando as seguintes questões:
  - Onde os ambientes de gerenciamento são usados na workload?
  - Como a workload implementa tolerância a falhas?
  - Quais são os padrões de design para componentes de ajuste de escala, ajuste de escala automático, redundância e alta disponibilidade?
  - Quais são os requisitos para disponibilidade e consistência de dados?
  - Há considerações quanto à economia de recursos ou estabilidade estática de recursos?
  - Quais são as dependências do serviço?
- Defina métricas de SLA com base na arquitetura da workload enquanto trabalha com as partes interessadas. Considere os SLAs de todas as dependências usadas pela workload.
- Quando a meta de SLA for definida, otimize a arquitetura para cumprir o SLA.
- Quando o design que cumprirá o SLA for definido, implemente mudanças operacionais, automação do processo e runbooks que também terão como foco uma redução de MTTD e MTTR.

- Depois da implantação, monitore e informe sobre o SLA.

Recursos

Práticas recomendadas relacionadas:

- REL03-BP01 Escolher como segmentar a workload
- REL10-BP01 Implantar a workload em vários locais
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas
- REL11-BP03 Automatizar a reparação em todas as camadas
- REL12-BP05 Testar a resiliência via engenharia do caos
- REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados
- Como a integridade da workload funciona

Documentos relacionados:

- Disponibilidade com redundância
- Pilar Confiabilidade: disponibilidade
- Como medir a disponibilidade
- Limites de isolamento de falhas da AWS
- Modelo de responsabilidade compartilhada para resiliência
- Estabilidade estática com zonas de disponibilidade
- Acordos de serviço (SLAs) da AWS
- Orientação para arquitetura baseada em células na AWS
- Infraestrutura da AWS
- Whitepaper Padrões avançados de resiliência Multi-AZ

Serviços relacionados:

- Amazon CloudWatch
- AWS Config
- AWS Trusted Advisor

REL 12. Como testar a confiabilidade?

Depois de projetar a workload para resiliência à pressão da produção, o teste é a única maneira de garantir que ela opere conforme projetado e com a resiliência esperada.

Práticas recomendadas

- REL12-BP01 Usar playbooks para investigar falhas
- REL12-BP02 Realizar análises pós-incidentes
- REL12-BP03 Testar os requisitos funcionais
- REL12-BP04 Testar os requisitos de ajuste de escala e performance
- REL12-BP05 Testar a resiliência via engenharia do caos
- REL12-BP06 Realizar game days regularmente

REL12-BP01 Usar playbooks para investigar falhas

Documente o processo de investigação em playbooks para permitir respostas consistentes e rápidas em cenários de falha. Os playbooks consistem em etapas predefinidas executadas para identificar os fatores que contribuem para um cenário de falha. Os resultados de qualquer etapa do processo são usados para determinar as próximas etapas a serem seguidas até que o problema seja identificado ou escalado.

O playbook é um planejamento proativo que deve ser feito para poder executar ações reativas com eficácia. Quando cenários de falha não cobertos pelo playbook forem encontrados na produção, aborde o problema primeiro ("apague o fogo"). Em seguida, volte e veja as etapas que você seguiu para resolver o problema e use-as para adicionar uma nova entrada no playbook.

Observe que os playbooks são usados em resposta a incidentes específicos, enquanto runbooks são usados para alcançar resultados específicos. Muitas vezes, os runbooks são usados para atividades de rotina e os playbooks são usados para responder a eventos que não são rotineiros.

Práticas comuns que devem ser evitadas:

- Planejar a implantação de uma workload sem conhecer os processos para diagnosticar problemas ou responder a incidentes.
- Decisões não planejadas de quais sistemas coletar logs e métricas ao investigar um evento.
- Não armazenar as métricas e os eventos por tempo suficiente para recuperar os dados.

Benefícios de implementar esta prática recomendada: capturar playbooks garante que os processos possam ser seguidos de forma consistente. A codificação dos seus playbooks limita a introdução de erros por atividades manuais. A automação dos playbooks reduz o tempo de resposta a um evento ao eliminar a necessidade de intervenção de membros da equipe ou ao fornecer a eles informações adicionais desde o início da intervenção.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

- Use playbooks para identificar problemas. Os playbooks são processos documentados para investigar problemas. Documente os processos em playbooks para permitir respostas consistentes e rápidas em cenários de falha. Os playbooks devem incluir as informações e as diretrizes necessárias para que uma pessoa com as devidas qualificações colete as informações aplicáveis, identifique possíveis fontes de falha, isole as falhas e determine os fatores contribuintes (ou seja, faça uma análise pós-incidente).
  - Implemente playbooks como código. Execute suas operações como código ao criar scripts de seus playbooks para garantir a consistência e reduzir os erros causados por processos manuais. Os playbooks podem ser compostos por vários scripts representando as diferentes etapas que podem ser necessárias para identificar os fatores que contribuem para um problema. As atividades do runbook podem ser acionadas ou executadas como parte das atividades do playbook, ou podem solicitar a execução de um playbook em resposta a eventos identificados.
    - Automatizar seus playbooks operacionais com o AWS Systems Manager
    - AWS Systems Manager Run Command
    - AWS Systems Manager Automation
    - O que é AWS Lambda?
    - O que é o Amazon EventBridge?
    - Usar alarmes do Amazon CloudWatch

Recursos

Documentos relacionados:

- AWS Systems Manager Automation
- AWS Systems Manager Run Command
- Automatizar seus playbooks operacionais com o AWS Systems Manager

- Usar alarmes do Amazon CloudWatch
- Usar canários (Amazon CloudWatch Synthetics)
- O que é o Amazon EventBridge?
- O que é AWS Lambda?

Exemplos relacionados:

- Automatizar operações com playbooks e runbooks

REL12-BP02 Realizar análises pós-incidentes

Analise os eventos que afetam o cliente e identifique os fatores contribuintes e os itens de ação preventiva. Use essas informações para desenvolver mitigações e limitar ou evitar recorrência. Desenvolva procedimentos para respostas rápidas e eficazes. Comunique os fatores contribuintes e as ações corretivas conforme apropriado, de acordo com o público-alvo. Tenha um método para comunicar essas causas a outras pessoas, conforme necessário.

Avalie por que os testes existentes não encontraram o problema. Adicione testes para esse caso se os testes ainda não existirem.

Resultado desejado: suas equipes têm uma abordagem consistente e consensual para lidar com a análise pós-incidente. Um mecanismo é o processo de correção de erros (COE). O processo de COE ajuda as equipes a identificar, compreender e abordar as causas básicas dos incidentes, ao mesmo tempo que cria mecanismos e barreiras de proteção para limitar a probabilidade do mesmo incidente ocorrer novamente.

Práticas comuns que devem ser evitadas:

- Encontrar fatores contribuintes, mas não continuar buscando mais profundamente outros possíveis problemas e abordagens de mitigação.
- Identificar apenas as causas de erros humanos e não oferecer nenhum treinamento ou automação que possa evitar erros humanos.
- Concentrar-se em atribuir a culpa em vez de compreender a causa-raiz, criando uma cultura de medo e impedindo a comunicação aberta.
- Não compartilhar insights, o que mantém as descobertas da análise de incidentes em um pequeno grupo e impede que outras pessoas se beneficiem das lições aprendidas.

- Não ter um mecanismo para capturar conhecimento institucional e, dessa forma, perder insights valiosos por não preservar as lições aprendidas na forma de práticas recomendadas atualizadas e resultando em incidentes repetidos com a mesma causa-raiz ou similar.

Benefícios de implementar esta prática recomendada: a realização de análises pós-incidentes e o compartilhamento dos resultados permitem que outras workloads atenuem o risco caso tenham implementado os mesmos fatores contribuintes, além de permitir que elas implementem a mitigação ou a recuperação automatizada antes que ocorra um incidente.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Uma boa análise pós-incidente oferece oportunidades para propor soluções comuns a problemas com padrões de arquitetura usados em outros locais nos sistemas.

A base do processo da COE é documentar e resolver problemas. É recomendável definir uma forma padronizada de documentar as causas-raiz essenciais e garantir que elas sejam analisadas e abordadas. Atribua uma propriedade clara ao processo de análise pós-incidente. Atribua uma equipe ou uma pessoa responsável para supervisionar as investigações e o rastreamento de incidentes.

Incentive uma cultura que se concentre no aprendizado e na melhoria, em vez de na atribuição de culpas. Enfatize que a meta é evitar futuros incidentes, não penalizar pessoas.

Desenvolva procedimentos bem definidos para conduzir análises pós-incidentes. Esses procedimentos devem descrever as etapas a serem seguidas, as informações a serem coletadas e as principais questões a serem abordadas durante a análise. Investigue os incidentes minuciosamente, indo além das causas imediatas para identificar as causas-raiz e os fatores contribuintes. Use técnicas como os cinco porquês para se aprofundar nos problemas subjacentes.

Mantenha um repositório das lições aprendidas com as análises dos incidentes. Esse conhecimento institucional pode servir como referência para futuros incidentes e iniciativas de prevenção. Compartilhe descobertas e insights de análises pós-incidentes e considere realizar reuniões abertas sobre a revisão pós-incidente para discutir as lições aprendidas.

Etapas de implementação

- Ao conduzir a análise pós-incidente, verifique se o processo está livre de culpabilização. Isso permite que as pessoas envolvidas no incidente sejam imparciais com as ações corretivas propostas e promovam uma autoavaliação honesta e a colaboração entre as equipes.

- Defina uma forma padronizada de documentar problemas essenciais. Um exemplo de estrutura para esse documento é o seguinte:
  - O que aconteceu?
  - Qual foi o impacto nos clientes e em sua empresa?
  - Qual foi a causa-raiz?
  - Quais dados você tem para apoiar isso?
    - Por exemplo, métricas e grafos
  - Quais foram as implicações críticas nos pilares, especialmente em relação à segurança?
    - Ao arquitetar workloads, você faz concessões entre os pilares com base no contexto da sua empresa. Essas decisões comerciais podem determinar suas prioridades de engenharia. Você pode reduzir custos e assim diminuir a confiabilidade em ambientes de desenvolvimento, ou otimizar a confiabilidade e aumentar os custos para soluções importantes. A segurança é sempre prioritária, porque você precisa proteger seus clientes.
  - Que lições você aprendeu?
  - Que ações corretivas você está adotando?
    - Itens de ação
    - Itens relacionados
- Crie procedimentos operacionais padrão bem definidos para conduzir análises pós-incidentes.
- Configure um processo padronizado de relatórios de incidentes. Documente todos os incidentes de forma abrangente, incluindo o relatório inicial do incidente, logs, comunicações e ações tomadas durante o incidente.
- Lembre-se de que um incidente não exige uma interrupção. Por exemplo, uma quase falha ou um sistema que, embora esteja funcionando de forma inesperada, cumpre sua função de negócios.
- Melhore continuamente o processo de análise pós-incidente com base no feedback e nas lições aprendidas.
- Capture as principais descobertas em um sistema de gerenciamento de conhecimento e considere os padrões que devem ser adicionados aos guias de desenvolvedor ou às listas de verificação de pré-implantação.

Recursos

Documentos relacionados:

- Por que você deve desenvolver uma correção de erro (COE)

Vídeos relacionados:

- A abordagem da Amazon para falhar com sucesso
- AWS re:Invent 2021: Amazon Builders' Library: excelência operacional da Amazon

REL12-BP03 Testar os requisitos funcionais

Use técnicas como testes de unidade e testes de integração que validem a funcionalidade necessária.

Os melhores resultados são obtidos quando esses testes são executados automaticamente como parte das ações de compilação e implantação. Por exemplo, usando o AWS CodePipeline, os desenvolvedores confirmam alterações em um repositório de origem onde o CodePipeline detecta automaticamente as alterações. Essas alterações são compiladas e os testes são executados. Após a conclusão dos testes, o código criado será implantado nos servidores de preparação para a realização de testes. No servidor de preparação, o CodePipeline executa mais testes, como testes de integração ou de carregamento. Após a conclusão bem-sucedida desses testes, o CodePipeline implanta o código testado e aprovado nas instâncias de produção.

Além disso, a experiência mostra que o teste de transações sintéticas (também conhecido como teste canário, mas que não deve ser confundido com implantações canário), capaz de executar e simular o comportamento do cliente, está entre os processos de teste mais importantes. Execute esses testes constantemente nos endpoints da workload de diversos locais remotos. O Amazon CloudWatch Synthetics permite criar canários para monitorar endpoints e APIs.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

- Teste os requisitos funcionais. Esse procedimento inclui testes de unidade e de integração que validam a funcionalidade necessária.
  - Usar o CodePipeline com o AWS CodeBuild para testar código e executar compilações
  - O AWS CodePipeline adiciona suporte a testes de integração unitários e personalizados com o AWS CodeBuild
  - Integração e entrega contínuas
  - Usar canários (Amazon CloudWatch Synthetics)
  - Automação de teste de software

Recursos

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudar na implementação de um pipeline de integração contínua
- O AWS CodePipeline adiciona suporte a testes de integração unitários e personalizados com o AWS CodeBuild
- AWS Marketplace: produtos que podem ser usados para integração contínua
- Integração e entrega contínuas
- Automação de teste de software
- Usar o CodePipeline com o AWS CodeBuild para testar código e executar compilações
- Usar canários (Amazon CloudWatch Synthetics)

REL12-BP04 Testar os requisitos de ajuste de escala e performance

Use técnicas como teste de carga para validar se a workload atende aos requisitos de ajuste de escala e performance.

É possível criar na nuvem um ambiente de teste em escala de produção sob demanda para sua workload. Se você executar esses testes na infraestrutura reduzida, deverá escalar os resultados observados para o que você acredita que acontecerá na produção. Os testes de carga e performance também podem ser realizados na produção se você tiver cuidado para não afetar os usuários reais e marcar seus dados de teste para que eles não se sintam com dados reais do usuário e estatísticas de uso corrompidas ou relatórios de produção.

Com os testes, certifique-se de que seus recursos básicos, configurações de ajuste de escala, cotas de serviço e design de resiliência operem conforme o esperado sob carga.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

- Teste os requisitos de ajuste de escala e performance. Execute o teste de carga para validar se a workload atende aos requisitos de ajuste de escala e performance.
  - Teste de carga distribuída na AWS: simular milhares de usuários conectados
  - Apache JMeter

- Implante a aplicação em um ambiente idêntico ao seu ambiente de produção e execute um teste de carga.
  - Use os conceitos de infraestrutura como código para criar um ambiente que seja o mais semelhante possível ao ambiente de produção.

Recursos

Documentos relacionados:

- Teste de carga distribuída na AWS: simular milhares de usuários conectados
- Apache JMeter

REL12-BP05 Testar a resiliência via engenharia do caos

Execute experimentos de caos regularmente em ambientes que estão em produção ou muito próximos de entrar em produção para entender como seu sistema responde a condições adversas.

Resultado desejado:

A resiliência da workload é verificada regularmente por meio da aplicação de engenharia do caos na forma de experimentos de injeção de falha ou injeção de carga inesperada, além de testes de resiliência que validam o comportamento conhecido esperado da workload durante um evento. Combine engenharia do caos e testes de resiliência para ter confiança de que sua workload poderá sobreviver à falha de componentes e se recuperar de interferências inesperadas com pouco ou nenhum impacto.

Práticas comuns que devem ser evitadas:

- Projetar para resiliência, mas não verificar como a workload funciona como um todo quando falhas ocorrem.
- Nunca realizar experimentos sob condições reais e de carga esperada.
- Não tratar seus experimentos como código nem mantê-los ao longo do ciclo de desenvolvimento.
- Não realizar experimentos de caos tanto como parte do pipeline de CI/CD quanto fora das implantações.
- Negar o uso de análises pós-incidentes passadas ao determinar quais falhas usar para realizar experimentos.

Benefícios de implementar esta prática recomendada: a injeção de falhas para verificar a resiliência de uma workload permite que você obtenha confiança de que os procedimentos de recuperação de seu design resiliente vão funcionar em caso de falha real.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A engenharia do caos proporciona à sua equipe os recursos para injetar continuamente interferências (simulações) reais de maneira controlada no provedor de serviço, na infraestrutura, na workload e no componente, com pouco ou nenhum impacto para os clientes. Ela permite que as equipes aprendam com as falhas e observem, mensurem e aumentem a resiliência das workloads, além de validar o acionamento de alertas e a notificação das equipes em caso de evento.

Quando realizada continuamente, a engenharia do caos pode destacar deficiências nas workloads que, se não respondidas, podem afetar negativamente a disponibilidade e a operação.

> **Note**
>
> A engenharia do caos é a disciplina de experimentar um sistema distribuído para aumentar a confiança na capacidade do sistema de resistir a condições turbulentas na produção. Princípios da engenharia do caos

Se um sistema é capaz de suportar essas interferências, os experimentos de caos devem ser mantidos como testes de regressão automatizados. Dessa forma, os experimentos de caos devem ser realizados como parte do ciclo de vida de desenvolvimento dos sistemas (SDLC) e como parte do pipeline de CI/CD.

Para garantir que sua workload possa sobreviver à falha de componentes, injete eventos reais como parte dos experimentos. Por exemplo, realize experimentos com perda de instâncias do Amazon EC2 ou failover da instância de banco de dados primária do Amazon RDS e verifique se a workload não é afetada (ou apenas minimamente afetada). Use uma combinação de falhas de componentes para simular eventos que podem ser causados por uma interferência em uma zona de disponibilidade.

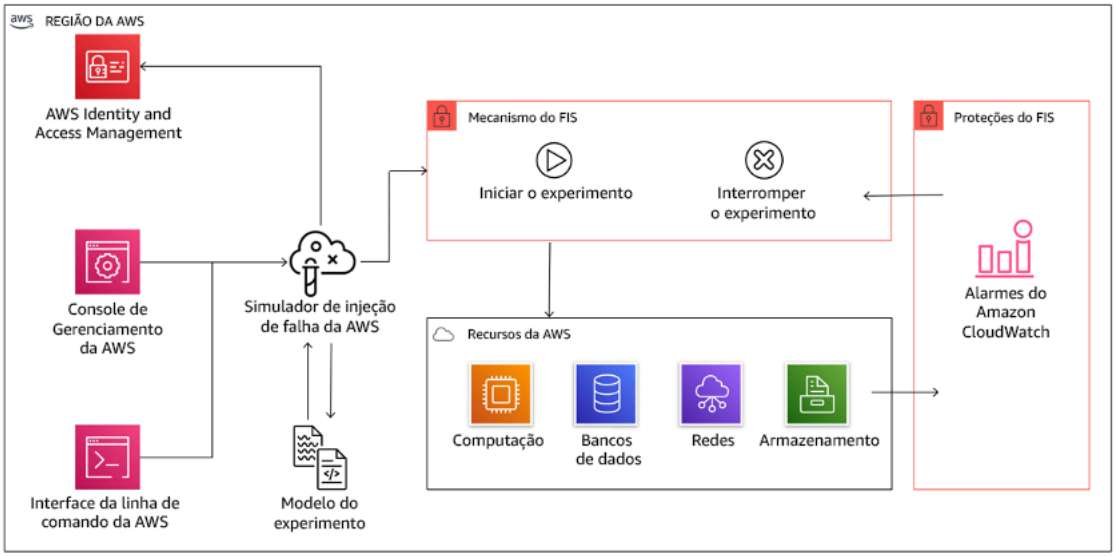
Para falhas no nível da aplicação (como travamentos), você pode começar com fatores de estresse, como exaustão de memória e CPU.

Para validar os mecanismos de fallback ou failover para dependências externas devido a interferências intermitentes na rede, os componentes devem simular esse tipo de evento bloqueando o acesso aos provedores externos durante um período especificado, que pode variar de segundos a horas.

Outros modos de degradação podem levar a uma redução nas funcionalidades e a respostas lentas, muitas vezes levando a uma interrupção dos serviços. Essa degradação costuma resultar de um aumento na latência de serviços críticos e comunicação de rede não confiável (pacotes abandonados). Experimentos com essas falhas, incluindo efeitos de rede como latência, mensagens perdidas e falhas de DNS, podem incluir a incapacidade de resolver um nome, alcançar o serviço de DNS ou estabelecer conexões com serviços dependentes.

Ferramentas de engenharia do caos:

O AWS Fault Injection Service (AWS FIS) é um serviço totalmente gerenciado para a execução de experimentos de injeção de falha que podem ser usados como parte do pipeline de CD, ou fora do pipeline. O AWS FIS é uma boa opção para ser usado durante game days de engenharia de caos. Ele oferece suporte à introdução simultânea de falhas em diferentes tipos de recursos, incluindo Amazon EC2, Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS) e Amazon RDS. Essas falhas incluem encerramento de recursos, failovers forçados, esgotamento de CPU ou memória, controle de utilização, latência e perda de pacotes. Por ser integrado a alarmes do Amazon CloudWatch, é possível definir condições de parada como barreiras de proteção para reverter um experimento se ele causar impacto inesperado.



O AWS Fault Injection Service se integra a recursos da AWS para permitir a execução de experimentos de injeção de falha para as workloads.

Existem também várias opções de terceiros para experimentos de injeção de falhas. Isso inclui ferramentas de código aberto, como Chaos Toolkit, Chaos Mesh e Litmus Chaos, além de opções comerciais, como o Gremlin. Para expandir o escopo de falhas que podem ser injetadas na AWS, o AWS FIS integra-se ao Chaos Mesh e ao Litmus Chaos, permitindo que você coordene fluxos de trabalho de injeção de falhas entre várias ferramentas. Por exemplo, você pode executar um teste de estresse na CPU de um pod usando falhas do Chaos Mesh ou Litmus enquanto encerra uma porcentagem selecionada aleatoriamente de nós de cluster usando ações de falha do AWS FIS.

Etapas de implementação

1. Determine quais falhas usar em seus experimentos.

   Avalie o design da workload quanto à resiliência. Esses designs (criados usando as práticas recomendadas do Well-Architected Framework) consideram os riscos com base em dependências críticas, eventos passados, problemas conhecidos e requisitos de conformidade. Liste cada elemento do design destinado a manter a resiliência e as falhas para o qual foi projetado para mitigar. Para obter mais informações sobre a criação dessas listas, consulte o whitepaper Revisões de prontidão operacional, o qual orienta você sobre como criar um processo para evitar a recorrência de incidentes anteriores. O processo de modos de falhas e análises de efeitos (FMEA) proporciona um framework para realização de análise de falhas em nível de componente e como elas afetam a workload. O FMEA é descrito com mais detalhes por Adrian Cockcroft em Modos de falha e resiliência contínua.
2. Atribua uma prioridade a cada falha.

   Comece com uma categorização bruta, como alta, média e baixa. Para avaliar a prioridade, considere a frequência da falha e o impacto da falha na workload total.

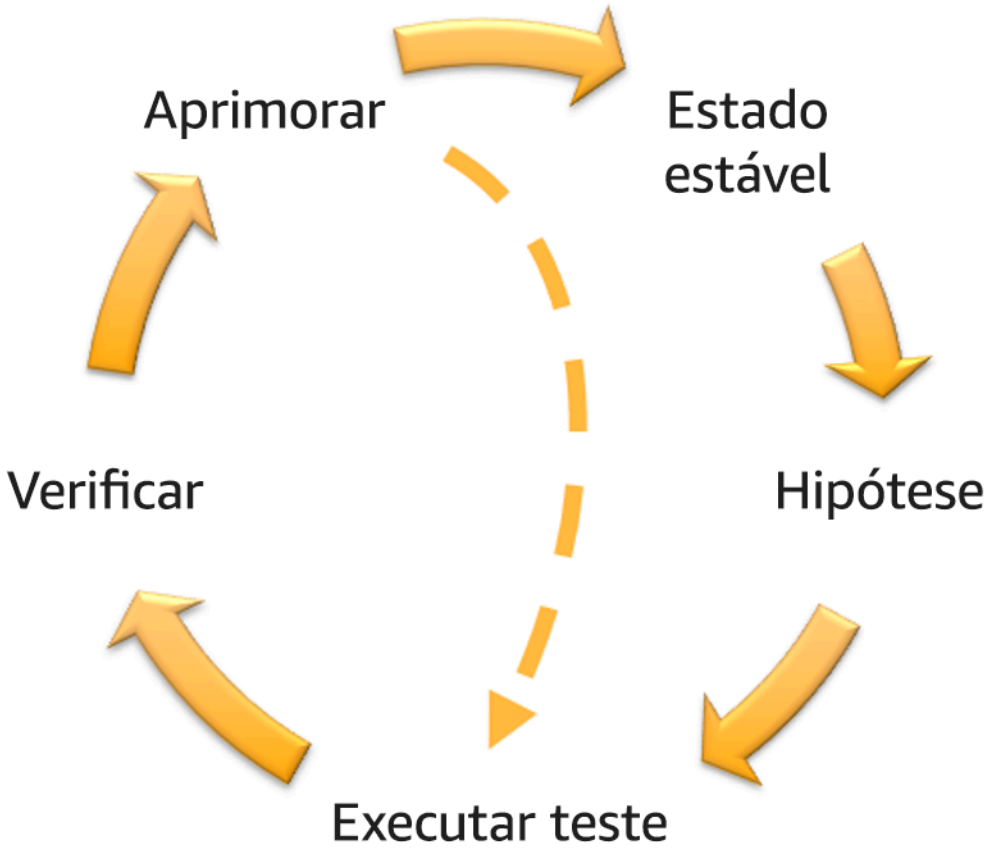
   Ao considerar a frequência de determinada falha, analise os dados passados para essa workload sempre que disponíveis. Caso contrário, use os dados de outras workloads executadas em ambientes semelhantes.

   Ao considerar o impacto de determinada falha, em geral, quanto maior o escopo da falha, maior o impacto. Considere também o design e a finalidade da workload. Por exemplo, a capacidade de acessar os datastores de origem é essencial para uma workload que executa análise e transformação de dados. Nesse caso, priorize experimentos de falhas de acesso, além de acesso controlado e inserção de latência.

   Análises pós-incidente são boas fontes de dados para entender a frequência e o impacto dos modos de falha.

Use a prioridade atribuída para determinar quais falhas escolher para experimentar primeiro e a sequência para desenvolver novos experimentos de injeção de falhas.

3. Para cada experimento realizado, siga o flywheel de engenharia do caos e resiliência contínua na figura a seguir.



Flywheel de engenharia do caos e resiliência contínua usando o método científico, por Adrian Hornsby.

a. Defina o estado estável como uma saída mensurável de uma workload que indica comportamento normal.

   Sua workload apresentará estado estável se estiver operando de maneira confiável e conforme o esperado. Portanto, valide a integridade da workload antes de definir o estado estável. O estado estável nem sempre significa que não há nenhum impacto à workload quando ocorre uma falha, já que determinada porcentagem de falhas pode estar dentro de limites aceitáveis.

O estado estável é a linha de base que você poderá observar durante o experimento, o que irá destacar anomalias se a hipótese definida na próxima etapa não sair conforme o esperado.

Por exemplo, um estado estável de um sistema de pagamentos pode ser definido como o processamento de 300 TPS com taxa de sucesso de 99% e tempo de ida e volta de 500 ms.

b. Formule uma hipótese sobre como a workload irá reagir à falha.

Uma boa hipótese baseia-se em como se espera que a workload mitigue a falha para manter o estado estável. A hipótese afirma que para determinado tipo de falha, o sistema ou a workload permanecerá em estado estável, pois a workload foi projetada com mitigações específicas. O tipo específico de falhas e mitigações deve ser especificado na hipótese.

O modelo a seguir pode ser usado para a hipótese (mas uma redação diferente também é aceitável):

> **Note**
>
> Se uma *falha específica* ocorrer, o *nome da workload descreverá os controles de mitigação* para manter o *impacto da métrica técnica ou comercial*.

Por exemplo:

- Se 20% dos nós no grupo de nós do Amazon EKS forem desativados, a API Transaction Create continuará atendendo ao 99º percentil das solicitações em menos de 100 ms (estado estável). Os nós do Amazon EKS se recuperarão em cinco minutos e os pods serão agendados e processarão o tráfego oito minutos depois do início do experimento. Os alertas serão acionados em três minutos.
- Se uma única falha de instância do Amazon EC2 ocorrer, a verificação de integridade do Elastic Load Balancing do sistema de ordem fará com que o Elastic Load Balancing envie solicitações apenas para as instâncias íntegras restantes, enquanto o Amazon EC2 Auto Scaling substitui a instância com falha, mantendo um aumento inferior a 0,01% na quantidade de erros no servidor (5xx) (estado estável).
- Se a instância de banco de dados primária do Amazon RDS falhar, a workload de coleta de dados da cadeia de suprimentos vai entrar em failover e se conectará à instância de banco de dados de espera do Amazon RDS para manter menos de um minuto de erros de leitura ou gravação de banco de dados (estado estável).

c. Execute o experimento injetando a falha.

Um experimento deve, por padrão, ser seguro contra falhas e tolerado pela workload. Se você sabe que a workload irá falhar, não execute o experimento. A engenharia do caos deve ser usada para encontrar incertezas conhecidas ou desconhecidas. Incertezas conhecidas são coisas que você conhece, mas não entende completamente, enquanto incertezas desconhecidas são coisas das quais você não está ciente nem compreende totalmente. Realizar experimentos em uma workload que você sabe que não funciona não oferecerá novos insights. Seu experimento deve ser cuidadosamente planejado, ter um escopo claro do impacto e fornecer um mecanismo de reversão que possa ser aplicado em caso de turbulência inesperada. Se sua devida diligência mostrar que a workload sobreviverá ao experimento, prossiga com o teste. Há diversas opções para injetar as falhas. Para workloads na AWS, o AWS FIS fornece muitas simulações de falhas predefinidas chamadas ações. Você também pode definir ações personalizadas que são executadas no AWS FIS usando documentos do AWS Systems Manager.

Recomendamos não usar scripts personalizados para experimentos de caos, a menos que os scripts tenham a capacidade de entender o estado atual da workload, sejam capazes de emitir logs e ofereçam mecanismos para rollbacks e condições de parada sempre que possível.

Um conjunto de ferramentas ou framework eficaz que ofereça suporte à engenharia do caos deve monitorar o estado atual de um experimento, emitir logs e fornecer mecanismos de rollback para acomodar à execução controlada de um experimento. Comece com um serviço estabelecido, como o AWS FIS, que permita que você realize experimentos com um escopo claramente definido e mecanismos de segurança que reverterão o experimento se ele introduzir turbulência inesperada. Para saber mais sobre uma variedade maior de experimentos usando o AWS FIS, consulte também o laboratório Aplicações resilientes e bem arquitetadas com engenharia do caos. Além disso, o AWS Resilience Hub analisará sua workload e criará experimentos que podem ser escolhidos para implementação e execução no AWS FIS.

> **Note**
>
> Para cada experimento, entenda claramente o escopo e seu impacto. Recomendamos que as falhas sejam simuladas primeiro em um ambiente de não produção, antes de serem executadas em produção.

Os experimentos devem ser executados em produção sob carga real usando implantações canário que ativam a implantação de um sistema de controle e experimental, sempre que possível. A realização de experimentos durante horários fora de pico é uma boa prática para mitigar o impacto potencial durante o primeiro experimento na produção. Além disso, se o uso de tráfego real de clientes for algo muito arriscado, você poderá executar experimentos usando tráfego sintético na infraestrutura de produção nas implantações de controle e experimentais. Quando não for possível usar a produção, realize os experimentos em ambientes de pré-produção que sejam o mais parecido possível com a produção.

Estabeleça e monitore barreiras de proteção para garantir que o experimento não afete o tráfego de produção ou outros sistemas além dos limites aceitáveis. Estabeleça condições de parada para interromper um experimento se ele atingir um limite definido de uma métrica de barreira de proteção. Isso deve incluir as métricas de estado estável da workload, bem como a métrica em relação aos componentes em que você está injetando a falha. Um monitor sintético (também conhecido como canário de usuário) é uma métrica que geralmente deve ser incluída como proxy de usuário. As condições de parada para AWS FIS são aceitas como parte do modelo de experimento, permitindo até cinco condições de parada por modelo.

Um dos princípios de caos é minimizar o escopo do experimento e seu impacto:

Embora deva existir uma provisão para algum impacto negativo de curto prazo, é responsabilidade e obrigação do engenheiro de caos garantir que as perdas dos experimentos sejam minimizadas e contidas.

Um método para verificar o escopo e o impacto potencial é realizar o experimento primeiro em um ambiente de não produção, verificando se os limites para as condições de parada são ativados conforme o esperado durante o experimento e se há observabilidade em vigor para identificar uma exceção, em vez de testar diretamente em produção.

Ao executar experimentos de injeção de falhas, verifique se todas as partes responsáveis estão bem informadas. Comunique-se com as equipes adequadas, como equipes de operações, equipes de confiabilidade do serviço e atendimento ao cliente, para avisá-las sobre quando os experimentos serão realizados e o que esperar. Ofereça a essas equipes ferramentas de comunicação para que informem os responsáveis pela execução do experimento caso percebam algum efeito adverso.

Você deve restaurar a workload e seus sistemas subjacentes de volta para o estado íntegro original. Normalmente, o design resiliente da workload irá se restaurar automaticamente. No entanto, alguns designs de falhas ou experimentos malsucedidos podem deixar a workload em um estado de falha inesperado. Ao final do experimento, você deverá estar ciente disso e restaurar a workload e os sistemas. Com o AWS FIS, é possível definir uma configuração de reversão (também chamada de ação posterior) nos parâmetros de ação. Uma ação posterior retorna o destino ao estado em que estava antes da execução da ação. Independentemente de serem automatizadas (como as que usam o AWS FIS) ou manuais, essas ações posteriores devem fazer parte de um playbook que descreve como detectar e lidar com falhas.

d. Verifique a hipótese.

   Os Princípios da engenharia do caos oferecem a seguinte orientação sobre como verificar o estado estável de sua workload:

   Concentre-se na saída mensurável de um sistema, em vez de atributos internos do sistema. As medições dessa saída durante um curto período constituem um proxy do estado estável do sistema. O throughput total do sistema, as taxas de erros e os percentis de latência podem ser métricas de interesse que representam o comportamento do estado estável. Ao focar em padrões de comportamento sistêmicos durante os experimentos, a engenharia de caos verifica se o sistema de fato funciona em vez de tentar validar como ele funciona.

   Nos dois exemplos anteriores, incluímos métricas de estado estável de menos de 0,01% de aumento na quantidade de erros no servidor (5xx) e menos de um minuto de erros de leitura ou gravação de banco de dados.

   Os erros 5xx são uma boa métrica, pois são consequência do modo de falha que um cliente da workload vivenciará diretamente. A medição dos erros do banco de dados é boa como consequência direta da falha, mas também deve ser complementada com uma medição de impacto para o cliente, como solicitações malsucedidas ou erros apresentados ao cliente. Além disso, inclua um monitor sintético (também conhecido como canário de usuário) em todas as APIs ou URIs acessadas pelo cliente da workload.

e. Melhore o design da workload para agregar resiliência.

   Se o estado estável não tiver sido mantido, investigue como o design da workload pode ser melhorado para mitigar a falha, aplicando as práticas recomendadas do pilar Confiabilidade do AWS Well-Architected. Orientações e recursos adicionais podem ser encontrados na

AWS Builder's Library, a que qual contém artigos sobre como melhorar suas verificações de interidade ou empregar novas tentativas com atraso no código da aplicação, entre outros.

Depois de implementar essas mudanças, execute o experimento novamente (mostrado pela linha pontilhada no flywheel de engenharia de caos) para determinar a eficácia. Se a etapa de verificação indicar que a hipótese é verdadeira, a workload estará em estado estável e o ciclo continuará.

4. Execute experimentos regularmente.

   Um experimento de caos é um ciclo, e os experimentos devem ser realizados regularmente como parte da engenharia de caos. Depois que uma workload cumprir a hipótese do teste, o experimento deverá ser automatizado para ser executado continuamente como parte de regressão do pipeline de CI/CD. Para saber como fazer isso, consulte este blog sobre como realizar experimentos do AWS FIS usando o AWS CodePipeline. Este laboratório sobre experimentos do AWS FIS recorrentes em um pipeline de CI/CD permite que você trabalhe de forma prática.

   Os experimentos de injeção de falhas também fazem parte dos game days (consulte REL12-BP06 Realizar game days regularmente). Os game days simulam uma falha ou um evento para verificar sistemas, processos e respostas das equipes. O objetivo é executar de fato as ações que a equipe executaria como se um evento excepcional acontecesse.

5. Capture e armazene os resultados do experimento.

   Os resultados dos experimentos de injeção de falhas devem ser capturados e persistidos. Inclua todos os dados necessários (como tempo, workload e condições) para poder analisar os resultados e as tendências do experimento posteriormente. Exemplos de resultados podem incluir capturas de tela de painéis, despejos em CSV do banco de dados da métrica ou um registro manual dos eventos e das observações do experimento. O registro em log de experimentos com o AWS FIS pode fazer parte dessa captura de dados.

Recursos

Práticas recomendadas relacionadas:

- REL08-BP03 Integrar testes de resiliência como parte da implantação
- REL13-BP03 Testar a implementação da recuperação de desastres para validá-la

Documentos relacionados:

- O que é AWS Fault Injection Service?
- O que é AWS Resilience Hub?
- Princípios da engenharia do caos
- Engenharia de caos: como planejar seu primeiro experimento
- Engenharia de resiliência: como aprender a aceitar falhas
- Histórias sobre engenharia do caos
- Como evitar fallback em sistemas distribuídos
- Implantação canário para experimentos de caos

Vídeos relacionados:

- AWS re:Invent 2020: Testar a resiliência via engenharia do caos (ARC316)
- AWS re:Invent 2019: Melhorar a resiliência com engenharia do caos (DOP309-R1)
- AWS re:Invent 2019: Aplicar a engenharia do caos em um universo de tecnologia sem servidor (CMY301)

Exemplos relacionados:

- Laboratório do Well-Architected: Nível 300: testes de resiliência do Amazon EC2, Amazon RDS e Amazon S3
- Laboratório Engenharia do caos na AWS
- Laboratório Aplicações resilientes e bem-arquitetadas com engenharia do caos
- Laboratório Caos na tecnologia sem servidor
- Laboratório Mensurar e aumentar a resiliência da sua aplicação com o AWS Resilience Hub

Ferramentas relacionadas:

- AWS Fault Injection Service
- AWS Marketplace: Plataforma de engenharia de caos Gremlin
- Chaos Toolkit
- Chaos Mesh

- Litmus

REL12-BP06 Realizar game days regularmente

Use os game days para simular regularmente seus procedimentos de resposta a eventos e falhas o mais próximo possível da produção (inclusive em ambientes de produção) e com as pessoas que estarão envolvidas nos cenários de falha reais. Os game days aplicam medidas para garantir que os eventos de produção não afetem os usuários.

Os game days simulam uma falha ou um evento para verificar sistemas, processos e respostas das equipes. O objetivo é executar de fato as ações que a equipe executaria como se um evento excepcional acontecesse. Isso ajudará a compreender onde é possível aplicar melhorias e pode ajudar a desenvolver experiência organizacional ao lidar com eventos. Eles devem ser realizados regularmente para que sua equipe desenvolva memória muscular sobre como responder.

Depois que o projeto de resiliência estiver em vigor e testado em ambientes que não sejam de produção, um game day será a maneira de garantir que tudo funcione conforme o planejado na produção. O game day, especialmente o primeiro, é uma atividade para "todos os funcionários" em que engenheiros e operações são informados quando ele acontecerá e o que ocorrerá. Os runbooks estão prontos. Eventos simulados são executados, incluindo possíveis eventos de falha, nos sistemas de produção da maneira prescrita, e o impacto é avaliado. Se todos os sistemas operarem conforme projetado, a detecção e a recuperação automática ocorrerão com pouco ou nenhum impacto. No entanto, se houver impacto negativo, o teste será revertido e os problemas da workload serão corrigidos manualmente, se necessário (usando o runbook). Como os game days ocorrem na produção, todas as precauções devem ser tomadas para garantir que não haja impacto sobre a disponibilidade dos seus clientes.

Práticas comuns que devem ser evitadas:

- Documentar seus procedimentos, mas nunca os por em prática.
- Não incluir os tomadores de decisão de negócios nas simulações de teste.

Benefícios de implementar esta prática recomendada: a realização frequente dos game days garante que toda a equipe siga as políticas e os procedimentos quando um incidente real ocorrer e valida se essas políticas e esses procedimentos são apropriados.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

- Programe os game days para praticar regularmente os runbooks e os playbooks. Os game days devem incluir todas as pessoas envolvidas em um evento de produção: proprietário da empresa, equipe de desenvolvimento, equipe operacional e equipes de resposta a incidentes.
  - Execute os testes de carga ou de performance e, em seguida, execute a injeção de falha.
  - Procure anomalias nos runbooks e oportunidades de praticar os playbooks.
    - Se você se desviar dos runbooks, refine-os ou corrija o comportamento. Se você praticar o playbook, identifique o runbook que deveria ter sido usado ou crie um.

Recursos

Vídeos relacionados:

- AWS re:Invent 2019: Melhorar a resiliência com engenharia do caos (DOP309-R1)

  Exemplos relacionados:

- Laboratórios do AWS Well-Architected: Testar a resiliência

REL 13. Como planejar para a recuperação de desastres (DR)?

Implementar backups e componentes redundantes de workload é o ponto de partida da sua estratégia de DR. O RTO e o RPO são os objetivos para restaurar a workload. Defina-os de acordo com suas necessidades de negócios. Implemente uma estratégia para atender a esses objetivos, considerando os locais e a função dos recursos e dos dados da workload. A probabilidade de interrupção e o custo de recuperação também são fatores principais que ajudam a determinar o valor empresarial de fornecer a recuperação de desastres para uma workload.

Práticas recomendadas

- REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados
- REL13-BP02 Usar estratégias de recuperação definidas para cumprir os objetivos de recuperação
- REL13-BP03 Testar a implementação da recuperação de desastres para validá-la
- REL13-BP04 Gerenciar o desvio de configuração no local ou na região de recuperação de desastres
- REL13-BP05 Automatizar a recuperação

REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados

A workload tem um objetivo de tempo de recuperação (RTO) e um objetivo de ponto de recuperação (RPO).

O objetivo de tempo de recuperação (RTO) é o atraso aceitável entre a interrupção e a restauração do serviço. Ele determina o que é considerado uma janela de tempo aceitável quando o serviço não está disponível.

O objetivo de ponto de recuperação (RPO) é o tempo máximo aceitável desde o último ponto de recuperação de dados. Isso determina o que é considerado uma perda aceitável de dados entre o último ponto de recuperação e a interrupção do serviço.

Os valores de RTO e RPO são considerações importantes ao selecionar uma estratégia de recuperação de desastres (DR) apropriada para sua workload. Esses objetivos são determinados pela empresa e, em seguida, usados pelas equipes técnicas para selecionar e implementar uma estratégia de DR.

Resultado desejado:

Cada workload tem um RTO e um RPO atribuídos, definidos com base no impacto nos negócios. A workload é atribuída a um nível predefinido, definindo a disponibilidade do serviço e a perda aceitável de dados, com um RTO e um RPO associados. Se essa hierarquização não for possível, ela poderá ser atribuída sob medida por workload, com a intenção de criar camadas posteriormente. O RTO e o RPO são usados como uma das principais considerações para a seleção de uma implementação de estratégia de recuperação de desastres para a workload. Outras considerações ao escolher uma estratégia de DR são restrições de custo, dependências de workload e requisitos operacionais.

Para o RTO, entenda o impacto com base na duração de uma interrupção. Ele linear ou há implicações não lineares? (por exemplo, depois de quatro horas, você desliga uma linha de fabricação até o início do próximo turno).

Uma matriz de recuperação de desastres, como a mostrada a seguir, pode ajudar você a entender como a criticidade da workload se relaciona aos objetivos de recuperação. (Observe que os valores reais dos eixos X e Y devem ser personalizados de acordo com as necessidades da sua organização).

Matriz de recuperação de desastres						
		Objetivo do ponto de recuperação				
		< 1 minuto	< 1 hora	< 6 horas	< 1 dia	+ 1 dia
Objetivo do tempo de recuperação	< 10 minutos	Crítica	Crítica	Alto	Médio	Médio
	< 2 horas	Crítica	Alto	Médio	Médio	Baixo
	< 8 horas	Alto	Médio	Médio	Baixo	Baixo
	< 24 horas	Médio	Médio	Baixo	Baixo	Baixo
	+ de 24 horas	Médio	Baixo	Baixo	Baixo	Baixo

Figura 16: Matriz de recuperação de desastres

Práticas comuns que devem ser evitadas:

- Não há objetivos de recuperação definidos.
- Seleção de objetivos de recuperação arbitrários.
- Seleção de objetivos de recuperação que são muito permissivos e não atendem aos objetivos de negócios.
- Não entender o impacto do tempo de inatividade e da perda de dados.
- Selecionar objetivos de recuperação irreais, como tempo zero para recuperação e zero perda de dados, o que pode não ser possível para sua configuração de workload.
- Seleção de objetivos de recuperação mais rigorosos do que os objetivos de negócios reais. Isso força implementações de DR mais caras e complicadas do que as necessidades da workload.
- Selecionar objetivos de recuperação incompatíveis com os de uma workload dependente.
- Seus objetivos de recuperação não consideram os requisitos de conformidade regulatória.
- RTO e RPO definidos para uma workload, mas nunca testados.

Benefícios de implementar esta prática recomendada: os objetivos de recuperação referentes a tempo e perda de dados são necessários para orientar a implementação de DR.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Para determinada workload, você deve entender o impacto do tempo de inatividade e da perda de dados em sua empresa. O impacto geralmente aumenta com maior tempo de inatividade ou perda de dados, mas a forma desse crescimento pode ser diferente com base no tipo de workload. Por exemplo, você pode tolerar o tempo de inatividade por até uma hora com pouco impacto, mas depois disso o impacto aumenta rapidamente. O impacto nos negócios se manifesta de várias formas, incluindo custo monetário (como perda de receita), confiança do cliente (e impacto na reputação), problemas operacionais (como ausência de folha de pagamento ou diminuição da produtividade) e risco regulatório. Use as etapas a seguir para entender esses impactos e definir RTO e RPO para sua workload.

Etapas de implementação

1. Determine as partes interessadas da sua empresa para essa workload e interaja com elas para implementar essas etapas. Os objetivos de recuperação de uma workload são uma decisão comercial. As equipes técnicas então trabalham com as partes interessadas da empresa para usar esses objetivos para selecionar uma estratégia de DR.

   > **Note**
   >
   > Para as etapas 2 e 3, você pode usar o the section called “Planilha de implementação”.

2. Reúna as informações necessárias para tomar uma decisão respondendo às perguntas abaixo.
3. Você tem categorias ou níveis de criticidade para o impacto da workload em sua organização?
   a. Se sim, atribua essa workload a uma categoria
   b. Se não, estabeleça essas categorias. Crie cinco ou menos categorias e refine o alcance do seu objetivo de tempo de recuperação para cada uma. As categorias de exemplo incluem: crítica, alta, média e baixa. Para entender como as workloads são mapeadas em categorias, considere se a workload é essencial, importante para os negócios ou não.
   c. Defina o RTO e o RPO da workload com base na categoria. Sempre escolha uma categoria mais restrita (menor RTO e RPO) do que os valores brutos calculados ao entrar nessa etapa. Se isso resultar em uma mudança de valor inadequadamente grande, considere criar uma nova categoria.
4. Com base nessas respostas, atribua valores de RTO e RPO à workload. Isso pode ser feito diretamente ou atribuindo-se a workload a um nível predefinido de serviço.

5. Documente o plano de recuperação de desastres (DRP) para essa workload, que faz parte do plano de continuidade de negócios (BCP) da sua organização, em um local acessível à equipe da workload e pelas demais partes interessadas
   a. Registre o RTO e o RPO e as informações usadas para determinar esses valores. Inclua a estratégia usada para avaliar o impacto da workload nos negócios
   b. Registre outras métricas além do RTO e do RPO que você está monitorando ou planeja monitorar para os objetivos de recuperação de desastres
   c. Você adicionará detalhes da sua estratégia de DR e do seu runbook a esse plano ao criá-los.
6. Ao examinar a criticidade da workload em uma matriz como a da Figura 15, você pode começar a estabelecer níveis predefinidos de serviço definidos para sua organização.
7. Depois de implementar uma estratégia de DR (ou uma prova de conceito para uma estratégia de DR) conforme descrito em the section called “REL13-BP02 Usar estratégias de recuperação definidas para cumprir os objetivos de recuperação”, teste essa estratégia para determinar o RTC (capacidade de tempo de recuperação) e o RPC (capacidade de ponto de recuperação) reais da workload. Se eles não atenderem aos objetivos de recuperação desejados, trabalhe com as partes interessadas da empresa para ajustar esses objetivos ou faça alterações na estratégia de DR para atingir os objetivos desejados.

Perguntas primárias

1. Qual é o tempo máximo em que a workload pode permanecer inativa antes que um impacto severo nos negócios ocorra?
   a. Determine o custo monetário (impacto financeiro direto) para a empresa por minuto se a workload for interrompida.
   b. Considere que o impacto nem sempre é linear. O impacto pode ser limitado no início e depois aumentar rapidamente após um momento crítico.
2. Qual é a quantidade máxima de dados que podem ser perdidos antes que um impacto severo nos negócios ocorra?
   a. Considere esse valor para seu datastore mais importante. Identifique a respectiva criticidade para outros datastores.
   b. Os dados da workload podem ser recriados em caso de perda? Se isso for operacionalmente mais fácil do que fazer backup e restauração, escolha o RPO com base na criticidade dos dados de origem usados para recriar os dados da workload.

3. Quais são os objetivos de recuperação e as expectativas de disponibilidade das workloads das quais esta depende (downstream) ou das workloads que dependem dela (upstream)?
    a. Escolha objetivos de recuperação que permitam que essa workload atenda aos requisitos das dependências upstream
    b. Escolha objetivos de recuperação que sejam alcançáveis considerando os recursos de recuperação das dependências posteriores. Dependências downstream não críticas (aquelas que é possível "contornar") podem ser excluídas. Ou trabalhe com dependências downstream críticas para melhorar suas capacidades de recuperação quando necessário.

Alguma outra dúvida?

Considere essas questões e como elas podem se aplicar a essa workload:

4. Você tem RTO e RPO diferentes dependendo do tipo de interrupção (região versus AZ, etc.)?
5. Há um horário específico (sazonalidade, eventos de vendas, lançamentos de produtos) em que seu RTO/RPO pode mudar? Em caso afirmativo, quais são os diferentes limites de tempo e medições?
6. Quantos clientes serão afetados se a workload for interrompida?
7. Qual será o impacto na reputação se a workload for interrompida?
8. Que outros impactos operacionais poderão ocorrer se a workload for interrompida? Por exemplo, impacto na produtividade dos funcionários se os sistemas de e-mail não estiverem disponíveis ou se os sistemas de folha de pagamento não conseguirem enviar transações.
9. Como o RTO e o RPO da workload se alinham à estratégia de DR da linha de negócios e da organização?
10. Há alguma obrigação contratual interna para a prestação de um serviço? Há alguma penalidade por não cumpri-las?
11. Quais são as restrições regulatórias ou de conformidade com os dados?

Planilha de implementação

A planilha pode ser usada para as etapas de implementação 2 e 3. É possível ajustá-la para atender às suas necessidades específicas, como adicionar perguntas adicionais.

Etapa 2: Perguntas principais	Aplicável à workload?	RTO da workload	RPO da workload	Ajuste do RTO.	Ajuste do RPO.	Instruções
[1] tempo máximo em que a workload pode ficar inativa						medido com o tempo desde o início da interrupção da recuperação
[2] quantidade máxima de dados que podem ser perdidos						medido com o tempo desde o conjunto de dados bom mais recente restaurável
[3a] dependências upstream						insira os objetivos mais estritos de recuperação upstream
[3b] dependências downstream						insira os objetivos menos estritos de recuperação downstream
[3a] dependências upstream reconciliadas						Se o valor upstream for menor que os valores atuais e o valor downstream for maior,
[3b] dependências downstream reconciliadas						trabalhe com as dependências para fazer a reconciliação e insira os valores reconciliados aqui
[3] dependências						valores menores para atender às dependências upstream ou aumentá-las com base nas capacidades das dependências downstream
Etapa 2: Perguntas adicionais						Indique se a pergunta é aplicável. Se ela não for aplicável, ignore-a
RTO/RPO de base						Carregue os valores de RTO e de RPO acima para baixo, aqui
[4] tipo de interrupção	[ ]S/[ ]N					Insira os objetivos de recuperação para o tipo de evento com os requisitos mais estritos
[5] objetivos baseados em tempo específico	[ ]S/[ ]N					Insira os objetivos de recuperação para momentos com os requisitos mais estritos
[6] clientes interrompidos	[ ]S/[ ]N					Faça um gráfico dos clientes afetados como uma função de tempo de inatividade ou de perda de dados. Use isso para inserir o RTO e o RPO máximos permissíveis com base no impacto no cliente.
[7] impacto na reputação	[ ]S/[ ]N					Trabalhe com a empresa para determinar o RTO e o RPO máximos com base no impacto na reputação
[8] impacto operacional	[ ]S/[ ]N					Insira o RTO e o RPO máximos com base no impacto operacional
[9] alinhamento organizacional	[ ]S/[ ]N					Insira o RTO e o RPO máximos para workloads desse tipo de acordo com as necessidades da LOB e da organização
[10] obrigações contratuais	[ ]S/[ ]N					Insira o RTO e o RPO máximos com base nas obrigações contratuais
[11] conformidade normativa	[ ]S/[ ]N					Insira o RTO e o RPO máximos com base na conformidade normativa aplicável
alvo baseado em questões adicionais						Use o valor mínimo (valor mais estrito) das perguntas 4 a 11 e insira-o aqui
alvo ajustado						Se os objetivos na linha acima não puderem ser acomodados, trabalhe com as partes interessadas para flexibilizar as restrições e insira o novo mínimo aqui
RTO/RPO ajustado						Insira os valores do RPO/RTO de base ou ajuste o alvo, o que for menor
Etapa 3						
Mapear para categoria ou camada predefinida						Ajuste os dois valores para baixo (mais estritos) para que se alinhem com a camada mais próxima definida

Planilha

Nível de esforço do plano de implementação: Baixo.

Recursos

Práticas recomendadas relacionadas:

- the section called “REL09-BP04 Realizar a recuperação periódica dos dados para verificar a integridade e os processos de backup”
- the section called “REL13-BP02 Usar estratégias de recuperação definidas para cumprir os objetivos de recuperação”
- the section called “REL13-BP03 Testar a implementação da recuperação de desastres para validá-la”

Documentos relacionados:

- Blog de arquitetura da AWS: série de recuperação de desastres
- Recuperação de desastres de workloads na AWS: recuperação na nuvem (whitepaper da AWS)

- Gerenciar políticas de resiliência com o Hub de Resiliência da AWS
- Parceiro da APN: parceiros que podem ajudar com a recuperação de desastres
- AWS Marketplace: produtos que podem ser usados para recuperação de desastres

Vídeos relacionados:

- AWS re:Invent 2018: Padrões de arquitetura para aplicações ativas-ativas multirregiões (ARC209-R2)
- Recuperação de desastres de workloads na AWS

REL13-BP02 Usar estratégias de recuperação definidas para cumprir os objetivos de recuperação

Defina uma estratégia de recuperação de desastres (DR) que cumpra os objetivos de recuperação da workload. Escolha uma estratégia como backup e restauração, standby (ativa/passiva) ou ativa/ativa.

Resultado desejado: para cada workload, há uma estratégia de DR definida e implementada que permite que a workload alcance os objetivos de DR. As estratégias de DR entre workloads fazem uso de padrões reutilizáveis (como as estratégias descritas anteriormente).

Práticas comuns que devem ser evitadas:

- Implementar procedimentos de recuperação inconsistentes para workloads com objetivos de DR semelhantes.
- Deixar que a estratégia de DR seja implementada ad hoc quando um desastre ocorrer.
- Não ter um plano para a recuperação de desastres.
- Depender das operações do ambiente de gerenciamento durante a recuperação.

Benefícios de implementar esta prática recomendada:

- O uso de estratégias de recuperação definidas permite que você adote ferramentas comuns e procedimentos de teste.
- Usar estratégias de recuperação definidas melhora o compartilhamento de conhecimento entre as equipes e a implementação da DR nas workloads pertencentes a elas.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto. Sem uma estratégia de DR planejada, implementada e testada, é improvável que você cumpra os objetivos de recuperação em caso de desastre.

Orientação para implementação

Uma estratégia de DR depende da capacidade de manter a workload em um site de recuperação se o local primário não puder executar a workload. Os objetivos de recuperação mais comuns são o RTO e o RPO, conforme discutido em REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados.

Uma estratégia de DR em várias zonas de disponibilidade (AZs) em uma única Região da AWS pode fornecer mitigação contra eventos de desastre, como incêndios, inundações e grandes interrupções de energia. Se implementar proteção contra um evento improvável que impeça a execução da workload em determinada Região da AWS for um requisito, você poderá optar por uma estratégia de DR que use várias regiões.

Ao arquitetar uma estratégia de DR em várias regiões, é necessário escolher uma das estratégias a seguir. Elas são listadas em ordem crescente de custo e complexidade e em ordem decrescente de RTO e RPO. Região de recuperação se refere a uma Região da AWS diferente da principal usada para sua workload.

Figura 17: Estratégias de recuperação de desastres (DR)

- Backup e restauração (RPO em horas, RTO em 24 horas ou menos): faça backup de seus dados e aplicações na região de recuperação. O uso de backups automatizados ou contínuos permitirá a recuperação para um ponto no tempo (PITR), o que pode reduzir o RPO para até cinco minutos em alguns casos. Em caso de desastre, você implantará a infraestrutura (usando a infraestrutura como código para reduzir o RTO), implantará o código e restaurará os dados salvos para se recuperar de um desastre na região de recuperação.
- Luz piloto (RPO em minutos, RTO em dezenas de minutos): provisione uma cópia da sua infraestrutura principal de workload na região de recuperação. Replique seus dados na região de recuperação e crie backups deles nessa região. Os recursos necessários para permitir a replicação e o backup, como bancos de dados e armazenamento de objetos, estão sempre ativos. Outros elementos, como servidores de aplicações ou computação com tecnologia sem servidor, não são implantados. No entanto, eles podem ser criados com a configuração e o código da aplicação necessários.
- Standby passivo (RPO em segundos, RTO em minutos): mantenha uma versão em escala vertical reduzida, mas totalmente funcional, da workload sempre em execução na região de recuperação. Os sistemas críticos para os negócios são totalmente duplicados e estão sempre ativados, mas com uma frota reduzida. Os dados são replicados e residem na região de recuperação. No momento da recuperação, a escala do sistema sistema é aumentada vertical e rapidamente para processar a carga de produção. Quanto mais a escala do standby passivo for aumentada verticalmente, menor será a dependência do RTO e do ambiente de gerenciamento. Quando totalmente dimensionado, isso é conhecido como standby a quente.
- Ativo-ativo em várias regiões (vários sites) (RPO próximo de zero, RTO potencialmente zero): sua workload é implantada e atende ativamente ao tráfego de várias Regiões da AWS. Essa estratégia exige que você sincronize os dados entre regiões. É necessário evitar ou lidar com possíveis conflitos causados por gravações no mesmo registro em duas réplicas regionais diferentes, o que pode ser complexo. A replicação de dados é útil para a sincronização de dados e protegerá você contra alguns tipos de desastre, mas não contra corrupção ou destruição de dados, a menos que sua solução também inclua opções para recuperação a um ponto anterior no tempo.

> **Note**
>
> Às vezes, a diferença entre luz-piloto e standby passivo pode ser difícil de entender. Ambos incluem um ambiente na região de recuperação com cópias dos ativos da região primária. A diferença é que a luz piloto não pode processar solicitações sem primeiro realizar uma ação adicional, enquanto o standby passivo pode processar o tráfego (em níveis de capacidade reduzidos) imediatamente. A abordagem de luz piloto exigirá que você ative

os servidores, possivelmente implante infraestrutura adicional (não essencial) e aumente a escala verticalmente. Já o standby passivo exige apenas que você aumente a escala verticalmente (tudo já está implantado e em execução). Escolha entre ambas as opções com base nas suas necessidades de RTO e RPO.
Quando o custo é uma preocupação e você deseja alcançar objetivos de RPO e RTO semelhantes, conforme definido na estratégia de standby passivo, é possível considerar soluções nativas da nuvem, como AWS Elastic Disaster Recovery, que adota a abordagem de luz piloto e oferece metas de RPO e RTO aprimoradas.

Etapas de implementação

1. Determine uma estratégia de recuperação de desastres que satisfaça os requisitos de recuperação dessa workload.

   Escolher uma estratégia de recuperação de desastres é uma troca entre reduzir o tempo de inatividade e a perda de dados (RTO e RPO) e o custo e a complexidade da implementação da estratégia. Você deve evitar implementar uma estratégia mais rigorosa do que necessário, pois isso resulta em custos desnecessários.

   Por exemplo, no diagrama a seguir, a empresa determinou o RTO máximo permitido e o orçamento limite da estratégia de restauração de serviço. Considerando os objetivos empresariais, as estratégias de recuperação de desastres de luz-piloto e standby passivo atenderão tanto ao RTO quanto aos critérios de custo.

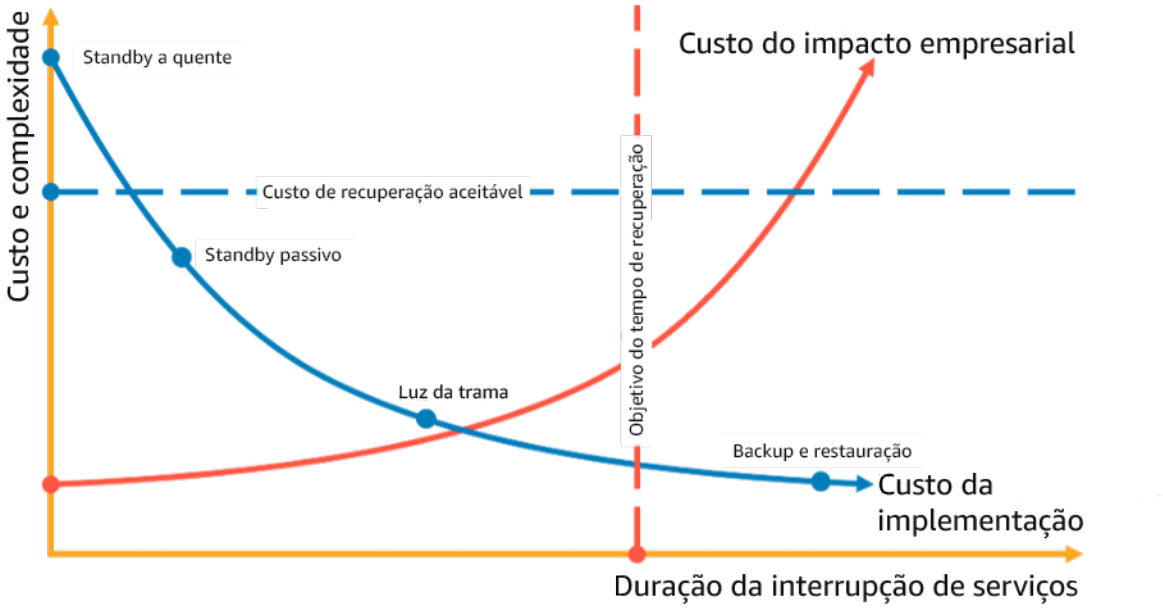


Figura 18: Escolher uma estratégia de recuperação de desastres com base no RTO e no custo

Para saber mais, consulte Plano de continuidade de negócios (BCP).

2. Analise os padrões de como a estratégia de recuperação de desastres selecionada pode ser implementada.

   O objetivo dessa etapa é entender como implementar a estratégia selecionada. As estratégias são explicadas usando as Regiões da AWS como locais primários e de recuperação. No entanto, também é possível optar por usar as zonas de disponibilidade em uma única região como sua estratégia de recuperação de desastres, a qual faz uso de elementos de várias dessas estratégias.

   Nas etapas a seguir, é possível aplicar a estratégia para sua workload específica.

   Backup e restauração

   Backup e restauração é a estratégia menos complexa de implementar, mas exigirá mais tempo e esforços para restaurar a workload, resultando em maior RTO e RPO. É uma boa prática sempre fazer backups dos dados e copiá-los para outro local (como outra Região da AWS).

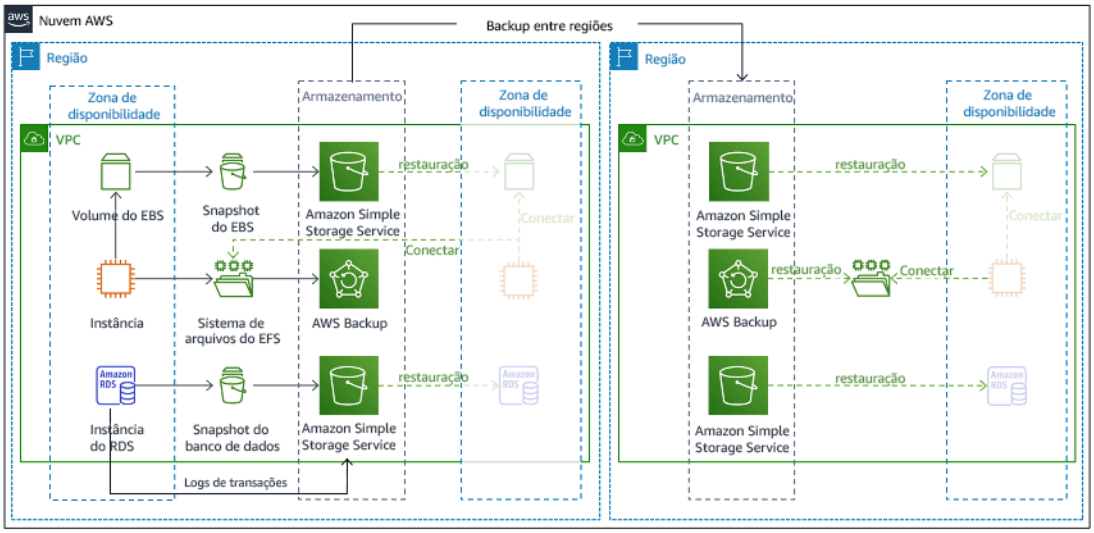


Figura 19: Arquitetura de backup e restauração

Para obter mais detalhes sobre essa estratégia, consulte Arquitetura de recuperação de desastres (DR) na AWS, Parte II: backup e restauração com recuperação rápida.

Luz piloto

Com a abordagem luz piloto, você replica seus dados da região principal para a região de recuperação. Os principais recursos usados para a infraestrutura da workload são implantados na região de recuperação. No entanto, recursos adicionais e as dependências ainda são necessários para tornar a pilha funcional. Por exemplo, na Figura 20, nenhuma instância de computação é implantada.

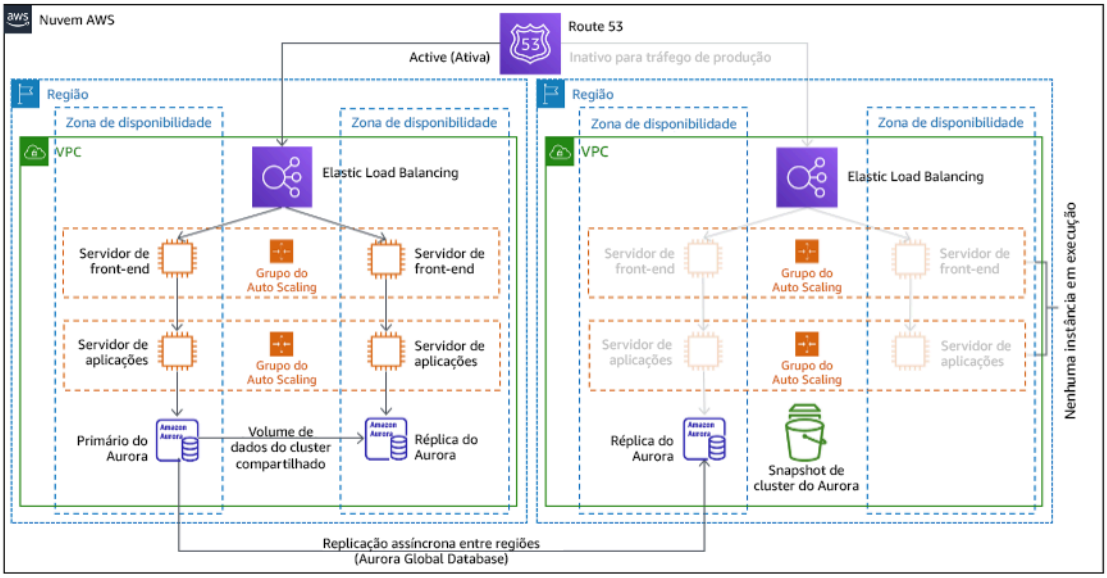


Figura 20: Arquitetura de luz piloto

Para obter mais detalhes sobre essa estratégia, consulte Arquitetura de recuperação de desastres (DR) na AWS, parte III: luz piloto e standby passivo.

Standby passivo

A abordagem de standby passivo envolve garantir que haja uma cópia reduzida, mas totalmente funcional, do seu ambiente de produção em outra região. Essa abordagem estende o conceito de luz piloto e diminui o tempo de recuperação, já que a workload está sempre ativa em outra região. Se a região de recuperação for implantada com capacidade total, isso será conhecido como standby a quente.

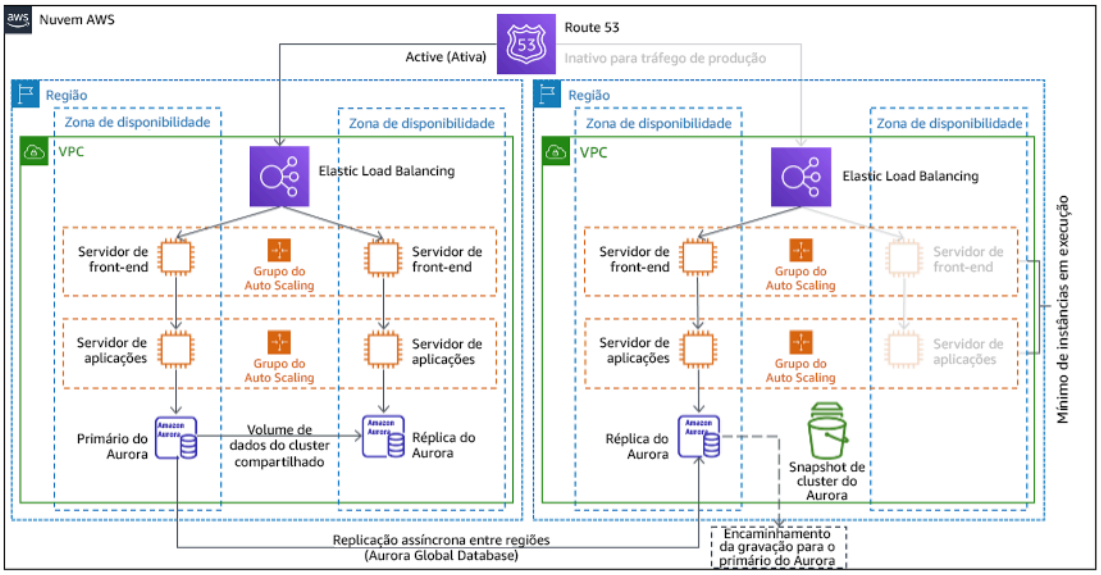


Figura 21: Arquitetura de standby passivo

O uso de standby passivo ou luz piloto requer que a escala dos recursos seja aumentada verticalmente na região de recuperação. Para verificar se a capacidade está disponível quando necessário, considere o uso de reservas de capacidade para instâncias do EC2. Quando o AWS Lambda é usado, a simultaneidade provisionada pode fornecer ambientes de runtime para que eles estejam preparados para responder imediatamente às invocações da sua função.

Para obter mais detalhes sobre essa estratégia, consulte Arquitetura de recuperação de desastres (DR) na AWS, parte III: luz piloto e standby passivo.

Multissite ativa/ativa

Você pode executar sua workload simultaneamente em várias regiões como parte de uma estratégia multissite ativa/ativa. A estratégia multissite ativa-ativa atende ao tráfego de todas as regiões onde está implantada. Os clientes podem selecionar essa estratégia para outros fins além da recuperação de desastres. Ela pode ser usada para aumentar a disponibilidade ou ao implantar uma workload para um público global (a fim de aproximar o endpoint dos usuários e/ou implantar pilhas localizadas para o público nessa região). Como uma estratégia de DR, se a workload não for compatível com uma das Regiões da AWS onde está implantada, essa região será evacuada e as regiões restantes serão usadas para manter a disponibilidade. A multissite ativa/ativa é a estratégia de DR mais complexa operacionalmente e deve ser selecionada apenas quando os requisitos empresariais exigirem.

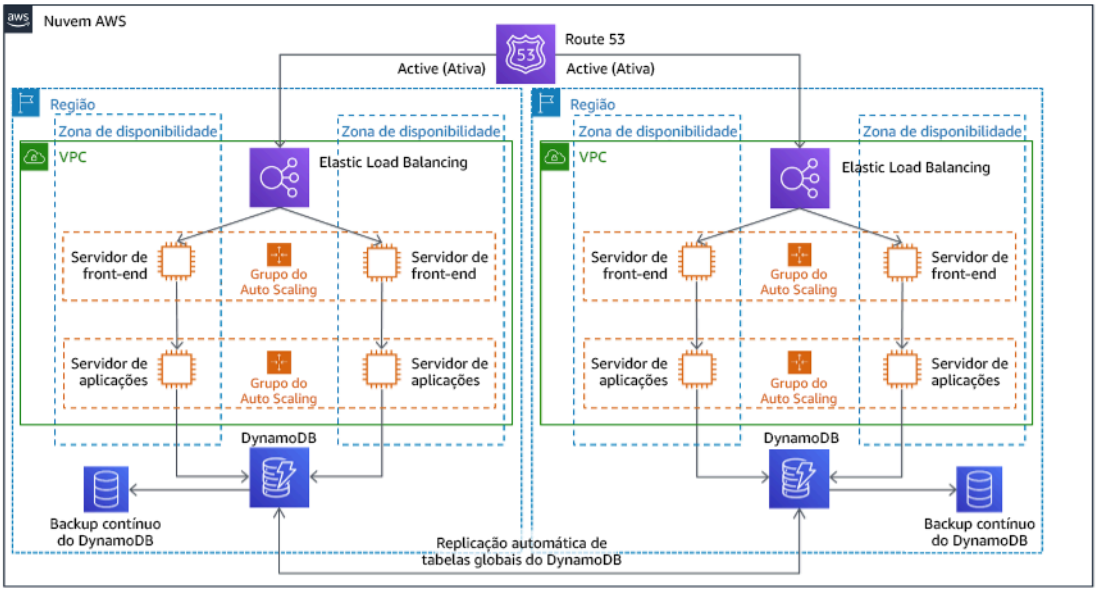


Figura 22: Arquitetura multissite ativa/ativa

Para obter mais detalhes sobre essa estratégia, consulte [Arquitetura de recuperação de desastres (DR) na AWS, Parte IV: multissite ativa/ativa]

AWS Elastic Disaster Recovery

Se você está considerando a estratégia de luz piloto ou standby passivo para recuperação de desastres, o AWS Elastic Disaster Recovery pode fornecer uma abordagem alternativa com benefícios aprimorados. O Elastic Disaster Recovery pode oferecer uma meta de RPO e RTO semelhante ao standby passivo, mas mantendo a abordagem de baixo custo da luz piloto. O Elastic Disaster Recovery replica os dados da sua região primária para sua região de recuperação usando proteção contínua de dados para obter um RPO medido em segundos e um RTO que pode ser medido em minutos. Somente os recursos necessários para replicar os dados são implantados na região de recuperação, o que mantém os custos baixos, semelhante à estratégia de luz piloto. Quando o Elastic Disaster Recovery é usado, o serviço coordena e orquestra a recuperação de recursos de computação quando iniciado como parte de um failover ou de uma simulação.

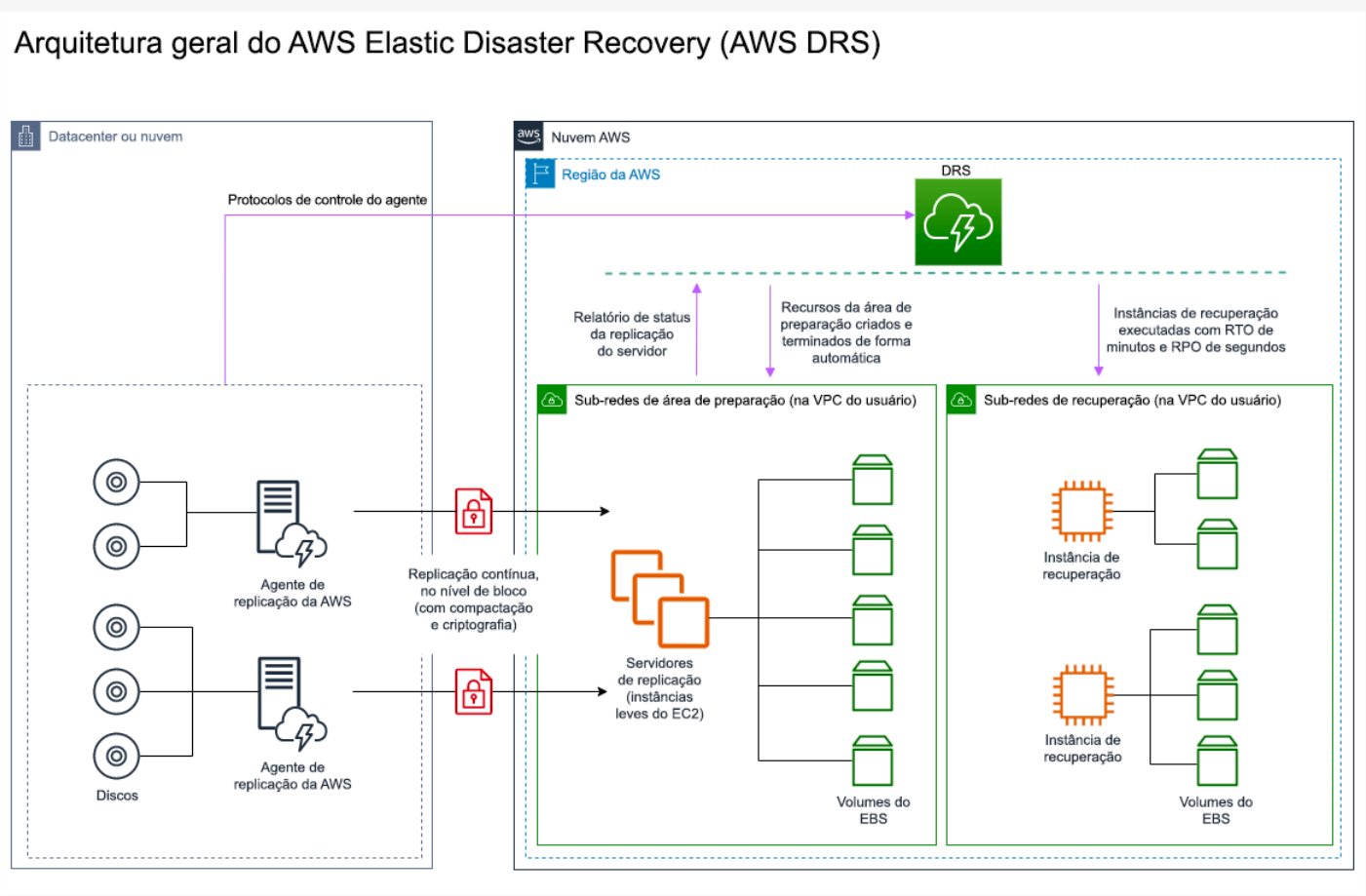


Figura 23: Arquitetura do AWS Elastic Disaster Recovery

Práticas adicionais para proteger dados

Com todas as estratégias, você também deve mitigar as consequências de um desastre de dados. A replicação contínua de dados protege você contra alguns tipos de desastre, mas não contra corrompimento ou destruição de dados, a menos que sua solução também inclua o versionamento de dados armazenados ou opções para recuperação a um ponto anterior no tempo. Você também deve fazer backup dos dados replicados no local de recuperação para criar backups pontuais além das réplicas.

Usar várias zonas de disponibilidade (AZs) em uma única Região da AWS

Ao utilizar várias AZs em uma única região, a implementação de DR usa vários elementos das estratégias acima. Primeiro, você deve criar uma arquitetura de alta disponibilidade (HA), usando várias AZs, conforme mostrado na Figura 23. Essa arquitetura faz uso de uma multissite ativa/ativa, já que as instâncias do Amazon EC2 e o Elastic Load Balancer têm recursos implantados

em várias AZs, processando ativamente as solicitações. A arquitetura também demonstra o standby a quente, em que, se a instância primária do Amazon RDS falhar (ou a própria AZ falhar), a instância em espera será promovida para primária.

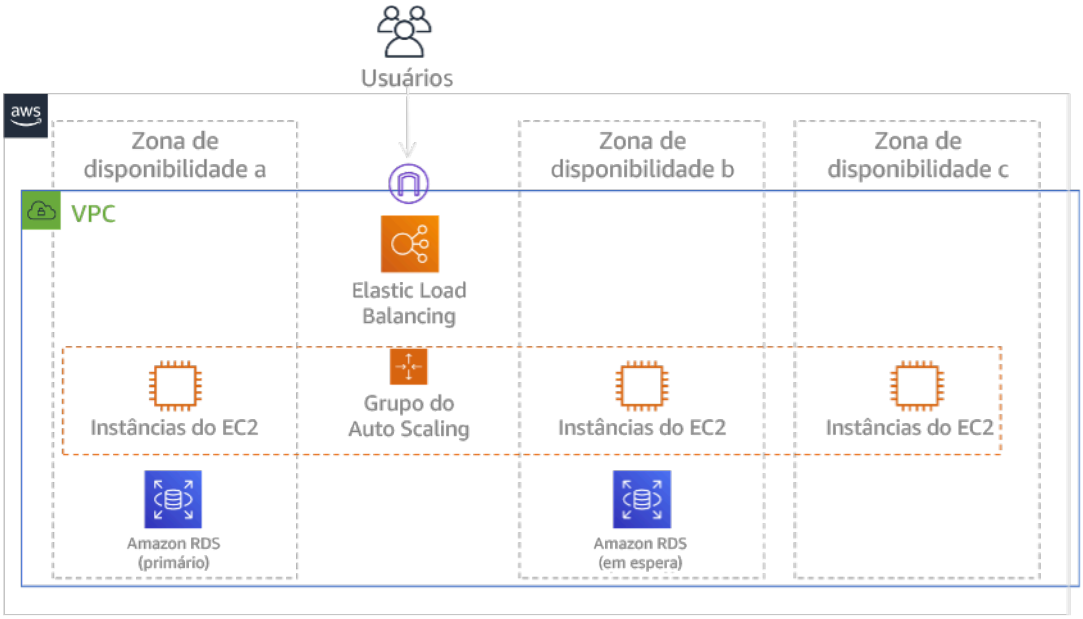


Figura 24: Arquitetura Multi-AZ

Além da arquitetura de alta disponibilidade, é necessário adicionar backups de todos os dados necessários para executar a workload. Isso é especialmente importante para dados restritos a uma única zona, como volumes do Amazon EBS ou clusters do Amazon Redshift. Se uma AZ falhar, você precisará restaurar esses dados para outra AZ. Sempre que possível, copie os backups de dados para outra Região da AWS como uma camada adicional de proteção.

Uma abordagem alternativa menos comum para uma única região, a recuperação de desastres multi-AZ é ilustrada na publicação do blog Building highly resilient applications using Amazon Application Recovery Controller, Part 1: Single-Region stack. Aqui, a estratégia é manter o máximo de isolamento possível entre as AZs, da mesma forma que as regiões operam. Ao usar esta estratégia alternativa, você pode escolher uma abordagem ativa/ativa ou ativa/passiva.

> **Note**
>
> Algumas workloads têm requisitos regulatórios de residência de dados. Se isso se aplicar à sua workload em uma localidade que atualmente tem apenas uma Região da AWS, a opção de multirregiões não atenderá às suas necessidades empresariais. As estratégias Multi-AZ fornecem boa proteção contra a maioria dos desastres.

3. Avalie os recursos da sua workload e qual será sua configuração na região de recuperação antes do failover (durante a operação normal).

   Para infraestrutura e recursos da AWS, use infraestrutura como código como o AWS CloudFormation ou ferramentas de terceiros como o Hashicorp Terraform. Para implantar em várias contas e regiões com uma única operação, é possível usar o AWS CloudFormation StackSets. Para estratégias multissite ativa/ativa e standby a quente, a infraestrutura implantada na região de recuperação tem os mesmos recursos que a região primária. Para as estratégias de luz piloto e standby passivo, a infraestrutura implantada exigirá ações adicionais para ficar pronta para produção. Usando os parâmetros e a lógica condicional do CloudFormation, é possível controlar se uma pilha implantada está ativa ou em espera com um único modelo. Quando o Elastic Disaster Recovery é usado, o serviço replica e orquestra a restauração de configurações da aplicação e os recursos de computação.

   Todas as estratégias de recuperação de desastres exigem que as fontes de dados sejam copiadas para backup dentro da Região da AWS e que esses backups sejam copiados para a região de recuperação. O AWS Backup fornece uma visão centralizada na qual é possível configurar, programar e monitorar backups desses recursos. Para luz piloto, standby passivo e multissite ativa/ativa, também é necessário replicar dados da região principal para recursos de dados na região de recuperação, como instâncias de banco de dados do Amazon Relational Database Service (Amazon RDS) ou tabelas do Amazon DynamoDB. Esses recursos de dados estão ativos e prontos para atender a solicitações na região de recuperação.

   Para saber mais sobre como os serviços da AWS operam entre regiões, consulte esta série de blogs sobre como Criar aplicações multirregiões com serviços da AWS.

4. Determine e implemente como você preparará sua região de recuperação para failover quando necessário (durante um evento de desastre).

   Para multissite ativa/ativa, failover significa evacuar uma região e confiar nas regiões ativas restantes. No geral, essas regiões estão prontas para aceitar tráfego. Para as estratégias de luz piloto e standby passivo, as ações de recuperação precisarão implantar os recursos ausentes, como as instâncias do EC2 na Figura 20, além de quaisquer outros recursos ausentes.

   Para todas as estratégias acima, talvez seja necessário promover instâncias somente leitura de bancos de dados para transformá-las na instância primária de leitura/gravação.

   Para backup e restauração, a restauração de dados do backup cria recursos para esses dados, como volumes do EBS, instâncias de banco de dados do RDS e tabelas do DynamoDB. Você

também precisa restaurar a infraestrutura e implantar o código. É possível usar o AWS Backup para restaurar dados na região de recuperação. Consulte REL09-BP01 Identificar e fazer backup de todos os dados que precisam de backup ou reproduzir os dados das fontes para obter mais detalhes. A reconstrução da infraestrutura inclui a criação de recursos como instâncias do EC2, além da Amazon Virtual Private Cloud (Amazon VPC), sub-redes e grupos de segurança necessários. É possível automatizar grande parte do processo de restauração. Para saber como, consulte esta postagem do blog.

5. Determine e implemente como você preparará sua região de recuperação para failover quando necessário (durante um evento de desastre).

   Essa operação de failover pode ser iniciada de forma manual ou automática. O failover iniciado automaticamente com base em verificações de integridade ou alarmes deve ser usado com cautela, pois um failover desnecessário (alarme falso) resulta em custos como indisponibilidade e perda de dados. Portanto, o failover iniciado manualmente é usado com frequência. Nesse caso, você ainda deve automatizar as etapas para failover para que a inicialização manual ocorra com o apertar um botão.

   Há várias opções de gerenciamento de tráfego a serem consideradas ao usar os serviços da AWS. Uma opção é usar o Amazon Route 53. Ao usar o Amazon Route 53, você pode associar vários endpoints de IP em uma ou mais Regiões da AWS a um nome de domínio do Route 53. Para implementar o failover iniciado manualmente, é possível usar o Amazon Application Recovery Controller, que fornece uma API de plano de dados altamente disponível destinada a redirecionar o tráfego para a região de recuperação. Ao implementar o failover, use as operações do plano de dados e evite as do ambiente de gerenciamento, conforme descrito em REL11-BP04 Confiar no plano de dados, e não no ambiente de gerenciamento, durante a recuperação.

   Para saber mais sobre essa e outras opções, consulte esta seção do whitepaper de recuperação de desastres.

6. Crie um plano de como será failback da workload.

   O failback é quando a operação da workload retorna para a região primária após o término de um evento de desastre. O provisionamento de infraestrutura e código para a região primária geralmente segue as mesmas etapas que foram usadas inicialmente, contando com a infraestrutura como código e pipelines de implantação de código. O desafio com o failback é restaurar os datastores e garantir sua consistência com a região de recuperação em operação.

No estado de failover, os bancos de dados na região de recuperação estão ativos e têm dados atualizados. O objetivo é ressincronizar da região de recuperação para a região primária, garantindo que ela permaneça atualizada.

Alguns serviços da AWS fazem isso automaticamente. Se estiver usando tabelas globais do Amazon DynamoDB, mesmo que a tabela na região primária tenha se tornado indisponível, o DynamoDB retomará a propagação de todas as gravações pendentes assim que ela retornar online. Se estiver usando o Amazon Aurora Global Database e usando failover planejado gerenciado, a topologia de replicação existente do banco de dados global do Aurora será mantida. Portanto, a antiga instância de leitura/gravação na região primária se tornará uma réplica e receberá atualizações da região de recuperação.

Nos casos em que isso não é automático, será necessário restabelecer o banco de dados na região primária como uma réplica do banco de dados na região de recuperação. Em muitos casos, isso envolverá a exclusão do banco de dados primário antigo e a criação de outras réplicas.

Após um failover, se você puder continuar a execução na região de recuperação, considere torná-la a nova região primária. Você ainda seguiria todas as etapas acima para transformar a antiga região primária em uma região de recuperação. Algumas organizações praticam uma rotação agendada, trocando as regiões primárias e de recuperação periodicamente (por exemplo, a cada três meses).

Todas as etapas necessárias para failover e failback devem ser mantidas em um playbook disponível para todos os membros da equipe e que seja revisado periodicamente.

Ao usar o Elastic Disaster Recovery, o serviço auxiliará na orquestração e automatização do processo de failback. Para obter mais detalhes, consulte Como executar um failback.

Nível de esforço do plano de implementação: Alto

Recursos

Práticas recomendadas relacionadas:

- the section called “REL09-BP01 Identificar e fazer backup de todos os dados que precisam de backup ou reproduzir os dados das fontes”
- the section called “REL11-BP04 Confiar no plano de dados, e não no ambiente de gerenciamento, durante a recuperação”

- the section called “REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados”

Documentos relacionados:

- Blog de arquitetura da AWS: série de recuperação de desastres
- Recuperação de desastres de workloads na AWS: recuperação na nuvem (whitepaper da AWS)
- Opções de recuperação de desastres na nuvem
- Criar uma solução de backend multirregiões ativa-ativa sem servidor em uma hora
- Backend multirregiões sem servidor: recarregado
- RDS: como replicar uma réplica de leitura entre regiões
- Route 53: configurar o failover de DNS
- S3: replicação entre regiões
- O que é o AWS Backup?
- What is Amazon Application Recovery Controller?
- AWS Elastic Disaster Recovery
- HashiCorp Terraform: introdução - AWS
- Parceiro da APN: parceiros que podem ajudar com a recuperação de desastres
- AWS Marketplace: produtos que podem ser usados para recuperação de desastres

Vídeos relacionados:

- Recuperação de desastres de workloads na AWS
- AWS re:Invent 2018: Padrões de arquitetura para aplicações ativas-ativas multirregiões (ARC209-R2)
- Introdução ao AWS Elastic Disaster Recovery | Amazon Web Services

Exemplos relacionados:

- Laboratório do Well-Architected: Recuperação de desastres: série de workshops que ilustram estratégias de recuperação de desastres

REL13-BP03 Testar a implementação da recuperação de desastres para validá-la

Teste regularmente o failover para o local de recuperação para verificar se a operação está correta e se o RTO e o RPO são cumpridos.

Práticas comuns que devem ser evitadas:

- Nunca praticar failovers na produção.

Benefícios de implementar esta prática recomendada: testar regularmente seu plano de recuperação de desastres garantirá que ele funcione quando necessário e que sua equipe saiba como executar a estratégia.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Um padrão que deve ser evitado é o desenvolvimento de caminhos de recuperação que raramente são executados. Por exemplo, você pode ter um datastore secundário utilizado para consultas somente leitura. Quando você grava em um datastore e o datastore primário falha, pode ser necessário fazer o failover para o repositório de dados secundário. Se você não testar esse failover com frequência, poderá descobrir que suas suposições sobre as capacidades do datastore secundário são incorretas. A capacidade do secundário, que talvez tenha sido suficiente quando testado pela última vez, pode não conseguir mais tolerar a carga nesse cenário. Nossa experiência mostrou que a única recuperação de erro que funciona é o caminho testado com frequência. É por isso que é melhor ter um pequeno número de caminhos de recuperação. Você pode estabelecer padrões de recuperação e testá-los regularmente. Se você tiver um caminho de recuperação complexo ou crítico, ainda precisará praticar regularmente essa falha na produção para se convencer de que o caminho de recuperação funciona. No exemplo que acabamos de discutir, você deve realizar o failover para o standby regularmente, não importa a necessidade.

Etapas de implementação

1. Projete suas workloads para recuperação. Teste regularmente seus caminhos de recuperação. A computação orientada para a recuperação identifica as características em sistemas que aprimoram a recuperação: isolamento e redundância, capacidade de reverter alterações em todo o sistema, capacidade de monitorar e determinar a integridade, capacidade de realizar diagnósticos, recuperação automatizada, design modular e capacidade de reinicialização. Pratique o caminho da recuperação para verificar se é possível realizá-la no tempo especificado para o

estado determinado. Use seus runbooks durante essa recuperação para documentar problemas e encontrar soluções para eles antes do próximo teste.

2. Para workloads baseadas no Amazon EC2, use o AWS Elastic Disaster Recovery para implementar e lançar instâncias de simulação para sua estratégia de DR. A AWS Elastic Disaster Recovery fornece a capacidade de realizar exercícios com eficiência, o que ajuda você a se preparar para um evento de failover. Também é possível iniciar frequentemente as instâncias usando o Elastic Disaster Recovery para fins de teste e simulação sem redirecionar o tráfego.

Recursos

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudar com a recuperação de desastres
- Blog de arquitetura da AWS: série de recuperação de desastres
- AWS Marketplace: produtos que podem ser usados para recuperação de desastres
- AWS Elastic Disaster Recovery
- Recuperação de desastres de workloads na AWS: recuperação na nuvem (whitepaper da AWS)
- AWS Elastic Disaster Recovery: como se preparar para o failover
- O projeto de computação orientado por recuperação de Berkeley/Stanford
- O que é o AWS Fault Injection Simulator?

Vídeos relacionados:

- AWS re:Invent 2018: Padrões de arquitetura para aplicações ativas-ativas multirregiões (ARC209-R2)
- AWS re:Invent 2019: Soluções de backup e restauração e recuperação de desastres com a AWS

Exemplos relacionados:

- Laboratórios do Well-Architected: Testar a resiliência

REL13-BP04 Gerenciar o desvio de configuração no local ou na região de recuperação de desastres

Certifique-se de que a infraestrutura, os dados e a configuração estejam disponíveis conforme necessário no local ou na região de DR. Por exemplo, verifique se as AMIs e as cotas de serviço estão atualizadas.

O AWS Config monitora e registra de forma contínua suas configurações de recursos da AWS. Ele pode detectar desvios e invocar o AWS Systems Manager Automation para corrigi-los e acionar alarmes. O AWS CloudFormation também pode detectar desvios nas pilhas que você implantou.

Práticas comuns que devem ser evitadas:

- Deixar de fazer atualizações em seus locais de recuperação quando você faz alterações na configuração ou na infraestrutura em seus locais principais.
- Não considerar possíveis limitações (como diferenças de serviço) em seus locais primário e de recuperação.

Benefícios de implementar esta prática recomendada: garantir que o ambiente de DR seja consistente com seu ambiente existente para assegurar a recuperação completa.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

- Garanta que seus pipelines de entrega consigam entregar em seus locais primário e de backup. Os pipelines de entrega para implantação de aplicações em produção devem ser distribuídos para todos os locais de estratégia de recuperação de desastres especificados, incluindo os ambientes de desenvolvimento e de teste.
- Permita que o AWS Config rastreie possíveis locais de desvio. Use as regras do AWS Config para criar sistemas que aplicam suas estratégias de recuperação de desastres e geram alertas ao detectar desvios.
  - Como remediar recursos não compatíveis da AWS pelo Regras do AWS Config
  - AWS Systems Manager Automation
- Use o AWS CloudFormation para implantar sua infraestrutura. O AWS CloudFormation pode detectar desvios entre o que seus modelos do CloudFormation especificam e o que é realmente implantado.
  - AWS CloudFormation: detectar desvios em uma pilha inteira do CloudFormation

Recursos

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudar com a recuperação de desastres
- Blog de arquitetura da AWS: série de recuperação de desastres
- AWS CloudFormation: detectar desvios em uma pilha inteira do CloudFormation
- AWS Marketplace: produtos que podem ser usados para recuperação de desastres
- AWS Systems Manager Automation
- Recuperação de desastres de workloads na AWS: recuperação na nuvem (whitepaper da AWS)
- Como faço para implementar uma solução de gerenciamento de configuração de infraestrutura na AWS?
- Como remediar recursos não compatíveis da AWS pelo Regras do AWS Config

Vídeos relacionados:

- AWS re:Invent 2018: Padrões de arquitetura para aplicações ativas-ativas multirregiões (ARC209-R2)

REL13-BP05 Automatizar a recuperação

Use ferramentas da AWS ou de terceiros para automatizar a recuperação do sistema e rotear o tráfego para o local ou a região de DR.

Com base em verificações de integridade configuradas, os serviços da AWS, como o Elastic Load Balancing e o AWS Auto Scaling, podem distribuir a carga para zonas de disponibilidade íntegras, enquanto serviços como o Amazon Route 53 e o AWS Global Accelerator podem rotear a carga para regiões íntegras da Regiões da AWS. O Amazon Application Recovery Controller ajuda a gerenciar e coordenar o failover usando recursos de verificação de prontidão e controle de roteamento. Esses recursos monitoram continuamente a capacidade da aplicação se recuperar de falhas, permitindo que você controle a recuperação da aplicação em várias Regiões da AWS, zonas de disponibilidade e ambientes on-premises.

Para workloads em data centers físicos ou virtuais existentes ou em nuvens privadas, o AWS Elastic Disaster Recovery permite que as organizações configurem uma estratégia automatizada de recuperação de desastres na AWS. O Elastic Disaster Recovery também oferece suporte à recuperação de desastres entre regiões e zonas de disponibilidade na AWS.

Práticas comuns que devem ser evitadas:

- A implementação de failover e failback automatizados idênticos pode causar oscilação quando uma falha ocorre.

Benefícios de implementar esta prática recomendada: a recuperação automatizada reduz o tempo de recuperação ao eliminar a oportunidade de erros manuais.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

- Automatizar caminhos de recuperação Para tempos de recuperação curtos, siga seu plano de recuperação de desastres para recolocar seus sistemas de TI online rapidamente no caso de uma interrupção.
  - Use o Elastic Disaster Recovery para failover e failback automatizados O Elastic Disaster Recovery replica continuamente suas máquinas (incluindo o sistema operacional, a configuração do estado do sistema, bancos de dados, aplicações e arquivos) em uma área de armazenamento de baixo custo em sua região preferida e Conta da AWS de destino. No caso de um desastre, depois de escolher se recuperar usando o Elastic Disaster Recovery, o Elastic Disaster Recovery automatiza a conversão de seus servidores replicados em workloads totalmente provisionadas em sua região de recuperação na AWS.
    - Usar o Elastic Disaster Recovery para failover e failback
    - Recursos do AWS Elastic Disaster Recovery

Recursos

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudar com a recuperação de desastres
- Blog de arquitetura da AWS: série de recuperação de desastres
- AWS Marketplace: produtos que podem ser usados para recuperação de desastres
- AWS Systems Manager Automation
- AWS Elastic Disaster Recovery
- Recuperação de desastres de workloads na AWS: recuperação na nuvem (whitepaper da AWS)

Vídeos relacionados:

- AWS re:Invent 2018: Padrões de arquitetura para aplicações ativas-ativas multirregiões (ARC209-R2)

Eficiência de performance

O pilar Eficiência de performance inclui a capacidade de usar recursos de nuvem de maneira eficiente para atender aos requisitos de performance e manter essa eficiência à medida que a demanda muda e as tecnologias evoluem. Recomendações sobre implementação estão disponíveis no whitepaper Pilar Eficiência de performance.

Áreas de práticas recomendadas

- Seleção de arquitetura
- Computação e hardware
- Gerenciamento de dados
- Rede e entrega de conteúdo
- Processo e cultura

Seleção de arquitetura

Perguntas

- PERF 1. Como selecionar os recursos e a arquitetura de nuvem apropriados para sua workload?

PERF 1. Como selecionar os recursos e a arquitetura de nuvem apropriados para sua workload?

A solução ideal para uma workload específica pode variar e, muitas vezes, as soluções combinam várias abordagens. As workloads do Well-Architected usam várias soluções e permitem diferentes recursos para aprimorar a performance.

Práticas recomendadas

- PERF01-BP01 Conhecer e compreender os serviços e recursos de nuvem disponíveis
- PERF01-BP02 Usar a orientação do provedor de nuvem ou de um parceiro apropriado para aprender sobre padrões de arquitetura e práticas recomendadas

- PERF01-BP03 Inclua o custo nas decisões de arquitetura
- PERF01-BP04 Avaliar como certas trocas (trade-offs) afetam os clientes e a eficiência da arquitetura
- PERF01-BP05 Usar políticas e arquiteturas de referência
- PERF01-BP06 Usar testes comparativos para orientar decisões de arquitetura
- PERF01-BP07 Usar uma abordagem baseada em dados para escolhas de arquitetura

PERF01-BP01 Conhecer e compreender os serviços e recursos de nuvem disponíveis

Continue a descobrir e aprender sobre serviços e configurações disponíveis que ajudam a tomar decisões e melhorar a eficiência de performance na arquitetura da workload.

Práticas comuns que devem ser evitadas:

- Usar a nuvem como um datacenter colocalizado.
- Não modernizar a aplicação após a migração para a nuvem.
- Usar somente um tipo de armazenamento para tudo que precisa ser mantido.
- Usar tipos de instância que atendem melhor aos seus padrões atuais, mas que sejam maiores quando necessário.
- Você implanta e gerencia tecnologias disponíveis como serviços gerenciados.

Benefícios de implementar esta prática recomendada: ao pensar em novos serviços e configurações, você poderá melhorar consideravelmente a performance, reduzir custos e otimizar o esforço necessário para manter as workloads. Isso também pode ajudar a acelerar o tempo para valorização dos produtos habilitados para a nuvem.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

A AWS lança constantemente novos serviços e recursos que podem melhorar a performance e reduzir o custo das workloads na nuvem. Atualizar-se em relação a esses novos serviços e atributos é crucial para manter a eficácia da performance na nuvem. Modernizar a arquitetura da workload também ajuda a acelerar a produtividade, impulsionar a inovação e ter acesso a mais oportunidades de crescimento.

Etapas de implementação

- Faça um inventário do software e da arquitetura usados para serviços relacionados a suas workloads. Decida sobre qual categoria de produtos você quer saber mais.
- Explore as ofertas da AWS para identificar e aprender sobre os serviços e as opções de configuração relevantes que podem ajudar você a melhorar a performance e reduzir os custos e a complexidade operacional.
  - Nuvem Amazon Web Services
  - AWS Academy
  - Quais são as novidades da AWS?
  - Blog da AWS
  - AWS Skill Builder
  - Eventos e webinars da AWS
  - Treinamento da AWS and Certifications
  - Canal da AWS no Youtube
  - Workshops da AWS
  - Comunidades da AWS
- Use o Amazon Q para obter informações e conselhos relevantes sobre serviços.
- Use ambientes sandbox (sem produção) para aprender e experimentar novos serviços sem incorrer em custos adicionais.
- Aprenda constantemente sobre novos serviços e recursos de nuvem.

Recursos

Documentos relacionados:

- Visão geral da Amazon Web Services
- Recursos do Amazon EC2
- Aprenda passo a passo com um plano de aprendizado de parceiro da AWS
- AWS Training and Certification
- Meu caminho de aprendizado para me tornar um arquiteto de soluções da AWS
- Centro de Arquitetura da AWS
- AWS Partner Network

- Biblioteca de Soluções da AWS
- Centro de Conhecimentos da AWS
- Criar aplicações modernas na AWS

Vídeos relacionados:

- AWS re:Invent 2023: Novidades do Amazon EC2
- AWS re:Invent 2022: Reduzir os custos operacionais e de infraestrutura com o Amazon ECS
- AWS re:Invent 2023: Compilar com a eficiência, a agilidade e a inovação da nuvem com o AWS
- AWS re:Invent 2022: Implantar modelos de ML para inferência com alta performance e baixo custo
- Esta é a minha arquitetura

Exemplos relacionados:

- Exemplos da AWS
- Exemplos do AWS SDK

PERF01-BP02 Usar a orientação do provedor de nuvem ou de um parceiro apropriado para aprender sobre padrões de arquitetura e práticas recomendadas

Use recursos disponibilizados pelo fornecedor de nuvem, como documentação, arquitetos de soluções, serviços profissionais ou parceiros apropriados, para orientar suas decisões durante a escolha da arquitetura. Eles ajudarão a analisar e melhorar sua arquitetura para alcançar a performance ideal.

Práticas comuns que devem ser evitadas:

- Você usa a AWS como um provedor de nuvem comum.
- Você usa os serviços da AWS de uma maneira para a qual eles não foram projetadas.
- Você segue todas as orientações sem considerar seu contexto de negócios.

Benefícios de implementar esta prática recomendada: usar a orientação de um provedor de nuvem ou de um parceiro apropriado pode ajudar a fazer as escolhas de arquitetura certas para as workloads e a conquistar confiança em suas decisões.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A AWS oferece uma ampla variedade de orientações, documentações e recursos que podem ajudar a criar e gerenciar workloads eficientes na nuvem. A documentação da AWS fornece exemplos de código, tutoriais e explicações detalhadas do serviço. Além da documentação, a AWS fornece programas de treinamento e certificação, arquitetos de soluções e serviços profissionais que podem ajudar os clientes a explorar diferentes aspectos dos serviços em nuvem e implementar uma arquitetura de nuvem eficiente na AWS.

Aproveite esses recursos para obter informações sobre conhecimentos valiosos e práticas recomendadas, economizar tempo e obter resultados melhores na Nuvem AWS.

Etapas de implementação

- Analise a documentação e as orientações da AWS e siga as práticas recomendadas. Esses recursos podem ajudar a escolher e configurar serviços com eficiência e obter melhor performance.
  - Documentação da AWS (como guias do usuário e whitepapers)
  - Blog da AWS
  - Treinamento da AWS and Certifications
  - Canal da AWS no Youtube
- Participe de eventos de parceiros da AWS (como os AWS Global Summits, AWS re:Invent, grupos de usuários e workshops) para ouvir dos próprios especialistas da AWS quais são as práticas recomendadas para usar os serviços da AWS.
  - Aprenda passo a passo com um plano de aprendizado de parceiro da AWS
  - Eventos e webinars da AWS
  - Workshops da AWS
  - Comunidades da AWS
- Entre em contato com a AWS para obter assistência quando precisar de mais orientações ou informações sobre produtos. AWS Os arquitetos de soluções e a AWS Professional Services fornecem orientação para a implementação de soluções. AWS Os parceiros oferecem experiência na AWS para ajudar você a desbloquear agilidade e inovação para os negócios.
- Use o AWS Support se precisar de suporte técnico para otimizar o uso de um serviço. Nossos planos de suporte são projetados a fim de oferecer a combinação certa de ferramentas e acesso ao conhecimento especializado para ter sucesso com a AWS e melhorar a performance, gerenciar riscos e manter os custos sob controle.

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de Soluções da AWS
- Centro de Conhecimentos da AWS
- AWS Enterprise Support

Vídeos relacionados:

- Esta é a minha arquitetura
- AWS re:Invent 2023: Padrões avançados orientados a eventos com o Amazon EventBridge
- AWS re:Invent 2023: Implementar padrões de design distribuídos na AWS
- AWS re:Invent 2023: Arquitetura de aplicações como código

Exemplos relacionados:

- Exemplos da AWS
- Exemplos do AWS SDK
- Arquitetura de referência de análise da AWS

PERF01-BP03 Inclua o custo nas decisões de arquitetura

Considere o custo em suas decisões de arquitetura para melhorar a utilização de recursos e a eficiência da performance de suas workloads na nuvem. Quando você está ciente das implicações de custo das suas workloads na nuvem, é mais provável que utilize recursos eficientes e reduza práticas ineficazes.

Práticas comuns que devem ser evitadas:

- Usar somente uma família de instâncias.
- Não avaliar soluções licenciadas em relação a soluções de código aberto.
- Não definir políticas de ciclo de vida de armazenamento.
- Não analisar os novos serviços e recursos da Nuvem AWS.

- Usar somente armazenamento em bloco.

Benefícios de implementar esta prática recomendada: levar em conta o custo em sua tomada de decisão permite que você use recursos mais eficientes e examine outros investimentos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Otimizar as workloads em função do custo pode melhorar a utilização dos recursos e evitar o desperdício em uma workload na nuvem. A consideração do custo nas decisões de arquitetura geralmente inclui o dimensionamento correto dos componentes da workload e a viabilização da elasticidade, o que resulta em maior eficiência da sua performance na nuvem.

Etapas de implementação

- Estabeleça objetivos de custo, como limites orçamentários para a workload na nuvem.
- Identifique os principais componentes (como instâncias e armazenamento) que impulsionam o custo da workload. É possível usar o AWS Pricing Calculator e o AWS Cost Explorer para identificar os principais fatores de custo na workload.
- Entenda os modelos de preços na nuvem, como instâncias sob demanda, instâncias reservadas, Savings Plans e instâncias spot.
- Use as Práticas recomendadas de otimização de custos do Well-Architected para otimizar esses principais componentes em termos de custo.
- Monitore e analise constantemente os custos para identificar oportunidades de otimizar as workloads e economizar.
  - Use o AWS Budgets para receber alertas quando os custos forem inaceitáveis.
  - Use o AWS Compute Optimizer ou o AWS Trusted Advisor para receber recomendações de otimização de custos.
  - Use a Detecção de Anomalias em Custos da AWS para fazer a detecção automática de anomalias de custo e análise de causa-raiz.

Recursos

Documentos relacionados:

- O que é o Gerenciamento de Faturamento e Custos da AWS?

- Otimização de custos com a AWS
- Escolher uma estratégia de gerenciamento de custos na AWS
- Guia do iniciante em gerenciamento de custos na AWS
- Uma visão geral detalhada do Cost Intelligence Dashboard
- Centro de Arquitetura da AWS
- Biblioteca de Soluções da AWS
- Centro de Conhecimentos da AWS

Vídeos relacionados:

- Esta é a minha arquitetura
- AWS re:Invent 2023: Novidades da otimização de custos com a AWS
- AWS re:Invent 2023: Otimizar custos e performance e rastrear o progresso rumo à mitigação
- AWS re:Invent 2023: Práticas recomendadas de otimização de custos de armazenamento na AWS
- AWS re:Invent 2023: Otimizar custos em seus ambientes com várias contas

Exemplos relacionados:

- Código de demonstração do AWS Compute Optimizer
- Workshop de otimização de custos
- Playbooks de implementação técnica de gerenciamento financeiro na nuvem
- Otimização de startups: ajustar a performance da aplicação para obter a máxima eficiência
- Workshop Otimização sem servidor (performance e custo)
- Escalar arquiteturas econômicas

PERF01-BP04 Avaliar como certas trocas (trade-offs) afetam os clientes e a eficiência da arquitetura

Ao avaliar melhorias relacionadas à performance, determine quais escolhas afetam os clientes e a eficiência das workloads. Por exemplo, se o uso de um datastore de chave-valor aumentar a performance do sistema, é importante avaliar como a alteração afetará os clientes após se tornar permanente

Práticas comuns que devem ser evitadas:

- Você pressupõe que todos os ganhos de performance devem ser implementados, mesmo que seja preciso fazer certas trocas para implementação.
- Você só avalia alterações nas workloads quando um problema de performance atinge um ponto crítico.

Benefícios de implementar esta prática recomendada: ao avaliar possíveis melhorias relacionadas à performance, você deve decidir se as concessões para as alterações são aceitáveis com os requisitos da workload. Em alguns casos, talvez seja necessário implementar controles adicionais para compensar as compensações.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Identifique áreas críticas na arquitetura em termos de performance e impacto para o cliente. Determine como é possível promover aprimoramentos, quais compromissos esses aprimoramentos exigem e como eles afetam o sistema e a experiência do usuário. Por exemplo, a implementação de armazenamento de dados em cache pode ajudar a aprimorar drasticamente a performance, mas requer uma estratégia clara de como e quando atualizar ou invalidar dados em cache a fim de prevenir comportamentos incorretos do sistema.

Etapas de implementação

- Entenda os SLAs e requisitos das suas workloads.
- Defina claramente os fatores de avaliação. Os fatores podem estar relacionados a custo, confiabilidade, segurança e performance das workloads.
- Selecione arquitetura e serviços que possam atender às suas necessidades.
- Realize experiências e provas de conceitos (POCs) para avaliar os fatores e o impacto de certas trocas para os clientes e para a eficiência da arquitetura. Normalmente, workloads de alta disponibilidade, com boa performance e seguras consomem mais recursos da nuvem e, ao mesmo tempo, proporcionam uma melhor experiência ao cliente. Entenda as vantagens e desvantagens da complexidade, da performance e do custo da workload. Normalmente, priorizar dois dos fatores inviabiliza o terceiro.

Recursos

Documentos relacionados:

- Amazon Builders' Library
- KPIs do Amazon QuickSight
- Amazon CloudWatch RUM
- Documentação do X-Ray
- Entender padrões de resiliência e compromissos para arquitetar de forma eficiente na nuvem

Vídeos relacionados:

- Otimizar aplicações com o Amazon CloudWatch RUM
- AWS re:Invent 2023: Capacidade, disponibilidade, eficiência de custos: escolha três
- AWS re:Invent 2023: padrões de integração avançados e compromissos para sistemas com acoplamento fraco

Exemplos relacionados:

- Medir o tempo de carregamento da com o Amazon CloudWatch Synthetics
- Cliente Web do Amazon CloudWatch RUM

PERF01-BP05 Usar políticas e arquiteturas de referência

Use políticas internas e arquiteturas de referência existentes ao selecionar serviços e configurações para ser mais eficiente ao projetar e implementar a workload.

Práticas comuns que devem ser evitadas:

- Você permite uma ampla variedade de tecnologias que podem afetar os custos de gerenciamento da empresa.

Benefícios de implementar esta prática recomendada: estabelecer uma política para opções de arquitetura, tecnologia e fornecedor permite que as decisões sejam tomadas rapidamente.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Adotar políticas internas na seleção de recursos e arquitetura fornece padrões e diretrizes a serem seguidos ao fazer escolhas arquitetônicas. Essas diretrizes simplificam o processo de tomada

de decisão ao escolher o serviço de nuvem certo e podem ajudar a melhorar a eficiência da performance. Implante a workload usando políticas ou arquiteturas de referência. Integre os serviços à implantação na nuvem e, depois, use testes de performance para verificar se você pode continuar a atender aos seus requisitos de performance.

Etapas de implementação

- Entenda claramente os requisitos da sua workload na nuvem.
- Revise as políticas internas e externas para identificar as mais relevantes.
- Use as arquiteturas de referência apropriadas fornecidas pela AWS ou as práticas recomendadas do seu setor.
- Crie um continuum que consiste em políticas, padrões, arquiteturas de referência e diretrizes prescritivas para situações comuns. Isso permite que suas equipes ajam mais rapidamente. Adapte os ativos para sua vertical, se aplicável.
- Valide essas políticas e arquiteturas de referência para sua workload em ambientes de sandbox.
- Atualize-se com relação aos padrões do setor e atualizações da AWS para garantir que suas políticas e arquiteturas de referência ajudem a otimizar sua workload na nuvem.

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de Soluções da AWS
- Centro de Conhecimentos da AWS
- Blog de arquitetura da AWS

Vídeos relacionados:

- Esta é a minha arquitetura
- AWS re:Invent 2022: Acelerar a geração de valor para seus negócios com o SAP e a arquitetura de referência da AWS

Exemplos relacionados:

- Exemplos da AWS
- Exemplos do AWS SDK

PERF01-BP06 Usar testes comparativos para orientar decisões de arquitetura

Compare a performance de uma workload existente para entender sua performance na nuvem e orientar decisões de arquitetura com base nesses dados.

Práticas comuns que devem ser evitadas:

- Você depende de testes comparativos comuns que não são indicativos das características da workload.
- Você conta com o feedback e as percepções de clientes como seu único teste comparativo.

Benefícios de implementar esta prática recomendada: o benchmarking da sua implementação atual permite medir melhorias de performance.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Use testes comparativos com testes sintéticos para avaliar a performance dos componentes da workload. O benchmarking é usado na avaliação da tecnologia para um componente específico e geralmente é mais simples de configurar do que testes de carga. Muitas vezes o benchmarking é usado no início de um novo projeto, quando ainda não há uma solução completa para o teste de carga.

Você pode criar seus próprios testes comparativos personalizados ou usar um teste padrão do setor, como o TPC-DS, para comparar suas workloads. Os benchmarks do setor são úteis ao comparar ambientes. Já os benchmarks personalizados são úteis para direcionar a tipos específicos de operações que você espera realizar em sua arquitetura.

Ao realizar testes comparativos, é importante "preaquecer" o ambiente de teste para obter resultados válidos. Execute o mesmo teste comparativo várias vezes para verificar a captura de qualquer variação ao longo do tempo.

Como normalmente é mais rápido executar testes comparativos do que testes de carga, eles podem ser usados mais cedo no pipeline de implantação e fornecer um feedback mais rápido sobre

desvios de performance. Ao avaliar uma alteração significativa em um componente ou serviço, o teste comparativo pode ser uma maneira rápida de verificar se é possível justificar a iniciativa para concretizar a alteração. O uso de testes comparativos em conjunto com testes de carga é importante porque o teste de carga informa como é a performance da workload no ambiente de produção.

Etapas de implementação

- Planeje e defina:
  - Defina os objetivos, o parâmetro de referência, os cenários de teste, as métricas (como utilização da CPU, latência ou throughput) e os KPIs para o teste comparativo.
  - Concentre-se nos requisitos do usuário em termos de experiência do usuário e em outros fatores, como tempo de resposta e acessibilidade.
  - Identifique uma ferramenta de testes comparativos adequada à workload. Você pode usar serviços da AWS como o Amazon CloudWatch ou uma ferramenta de terceiros que seja compatível com a workload.
- Configure e instrumente:
  - Prepare o ambiente e configure os recursos.
  - Implemente monitoramento e registro em log para capturar os resultados dos testes.
- Compare e monitore:
  - Execute testes comparativos e monitore as métricas durante o teste.
- Analise e documente:
  - Documente o processo de comparação e as descobertas.
  - Analise os resultados para identificar gargalos, tendências e áreas para melhoria.
  - Use os resultados do teste para tomar decisões de arquitetura e ajustar a workload. Isso pode incluir a mudança de serviços ou a adoção de novos recursos.
- Otimize e repita:
  - Ajuste as configurações e alocações de recursos com base nos testes comparativos.
  - Teste novamente a workload depois do ajuste para validar as melhorias.
  - Documente seu aprendizado e repita o processo para identificar outras áreas para melhoria.

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS

- AWS Partner Network
- Biblioteca de Soluções da AWS
- Centro de Conhecimentos da AWS
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Fluxos de trabalho de genômica, Parte 5: benchmarking automatizado
- Comparar e otimizar a implantação de endpoints no Amazon SageMaker JumpStart

Vídeos relacionados:

- AWS re:Invent 2023: Testes comparativos de partida a frio do AWS Lambda
- Testes comparativos de serviços com estado na nuvem
- Esta é a minha arquitetura
- Otimizar aplicações com o Amazon CloudWatch RUM
- Demonstração do Amazon CloudWatch Synthetics

Exemplos relacionados:

- Exemplos da AWS
- Exemplos do AWS SDK
- Testes de carga distribuídos
- Medir o tempo de carregamento da com o Amazon CloudWatch Synthetics
- Cliente Web do Amazon CloudWatch RUM

PERF01-BP07 Usar uma abordagem baseada em dados para escolhas de arquitetura

Defina uma abordagem clara e baseada em dados para escolhas de arquitetura a fim de verificar se os serviços e configurações de nuvem corretos são usados para atender às suas necessidades comerciais específicas.

Práticas comuns que devem ser evitadas:

- Você pressupõe que sua arquitetura atual é estática e não deve ser atualizada ao longo do tempo.
- Suas escolhas de arquitetura são baseadas em suposições.

- Você apresenta alterações de arquitetura ao longo do tempo sem justificativa.

Benefícios de implementar esta prática recomendada: ao aplicar uma abordagem bem definida para fazer escolhas de arquitetura, você usa dados para influenciar o projeto das workloads e tomar decisões conscientes ao longo do tempo.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Use a experiência interna e o conhecimento da nuvem ou de recursos externos, como casos de uso publicados ou whitepapers, para escolher recursos e serviços em sua arquitetura. Você deve ter um processo bem definido que incentive a experimentação e os testes comparativos com os serviços que podem ser usados em suas workloads.

Os atrasos de workloads críticas devem consistir não apenas em histórias de usuários que venham a oferecer funcionalidades relevantes para os negócios e usuários, mas também em histórias técnicas que formem uma base de arquitetura para as workloads. Essa base é formada por novos avanços em tecnologia e novos serviços e os adota em função de dados e justificativas adequadas. Isso verifica se a arquitetura permanece preparada para o futuro e não se torna estagnada.

Etapas de implementação

- Interaja com as principais partes interessadas para definir os requisitos das workloads, incluindo considerações de performance, disponibilidade e custo. Considere fatores como o número de usuários e o padrão de uso das workloads.
- Crie uma base de arquitetura ou uma lista de pendências de tecnologia que seja priorizada junto com a lista de pendências funcional.
- Avalie diferentes serviços em nuvem (para obter mais detalhes, consulte PERF01-BP01 Conhecer e compreender os serviços e recursos de nuvem disponíveis).
- Explore diferentes padrões de arquitetura, como microsserviços ou tecnologia sem servidor, que atendem aos requisitos de performance (para obter mais detalhes, consulte PERF01-BP02 Usar a orientação do provedor de nuvem ou de um parceiro apropriado para aprender sobre padrões de arquitetura e práticas recomendadas).

- Consulte outras equipes, diagramas de arquitetura e recursos, como arquitetos de soluções da AWS, o Centro de Arquitetura da AWS e o AWS Partner Network, para obter ajuda para escolher a arquitetura certa para sua workload.

- Defina métricas de performance, como throughput e tempo de resposta, que podem ajudar você a avaliar a performance das workloads.
- Experimente e use métricas definidas para validar a performance da arquitetura selecionada.
- Monitore e faça ajustes contínuos conforme necessário para manter a performance ideal da arquitetura.
- Documente a arquitetura e as decisões selecionadas como referência para futuras atualizações e aprendizados.
- Revise e atualize constantemente a abordagem para seleção de arquitetura com base em aprendizados, novas tecnologias e métricas. Esses parâmetros podem indicar que é necessário mudar ou que há algum problema na abordagem atual.

Recursos

Documentos relacionados:

- Biblioteca de Soluções da AWS
- Centro de Conhecimentos da AWS
- Padrões de arquitetura para criar aplicações orientadas a dados fim a fim na AWS

Vídeos relacionados:

- Esta é a minha arquitetura
- AWS re:Invent 2021: Empresa orientada a dados: da visão ao valor
- AWS re:Invent 2022: Como entregar arquiteturas sustentáveis e de alta performance
- AWS re:Invent 2023: Otimizar custos e performance e rastrear o progresso rumo à mitigação
- AWS re:Invent 2022: Otimização da AWS: etapas acionáveis para resultados imediatos

Exemplos relacionados:

- Exemplos da AWS
- Exemplos do AWS SDK

Computação e hardware

Perguntas

- PERF 2. Como selecionar e usar recursos computacionais em sua workload?

PERF 2. Como selecionar e usar recursos computacionais em sua workload?

A opção ideal de computação para uma workload específica pode variar de acordo com o design, os padrões de uso e as definições de configuração da aplicação. As arquiteturas podem usar diferentes opções de computação para vários componentes e permitir diferentes recursos para aprimorar a performance. A seleção da opção de computação incorreta para uma arquitetura pode levar a uma menor eficiência de performance.

Práticas recomendadas

- PERF02-BP01 Selecionar as melhores opções de computação para as workloads
- PERF02-BP02 Entender a configuração e os recursos de computação disponíveis
- PERF02-BP03 Coletar métricas relacionadas à computação
- PERF02-BP04 Configurar e dimensionar corretamente os recursos de computação
- PERF02-BP05 Dimensionar recursos de computação dinamicamente
- PERF02-BP06 Usar aceleradores de computação baseados em hardware otimizados

PERF02-BP01 Selecionar as melhores opções de computação para as workloads

Selecionar a opção de computação mais adequada para suas workloads permite melhorar a performance, reduzir os custos desnecessários de infraestrutura e reduzir os esforços operacionais necessários para mantê-las.

Práticas comuns que devem ser evitadas:

- A mesma opção de computação utilizada on-premises é usada.
- Você não tem conhecimento das opções, dos atributos e das soluções de computação em nuvem e de como essas soluções podem melhorar a performance computacional.
- Uma opção de computação existente é provisionada de forma excessiva para atender aos requisitos de ajuste de escala ou performance quando uma opção alternativa de computação se alinharia às características da workload com mais precisão.

Benefícios de implementar esta prática recomendada: ao identificar os requisitos de computação e avaliar as opções disponíveis, você pode tornar a workload mais eficiente em termos de recursos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Para otimizar as workloads na nuvem quanto à eficiência de performance, é importante selecionar as opções de computação mais apropriadas para seu caso de uso e requisitos de performance. A AWS fornece uma variedade de opções de computação que atendem a diferentes workloads na nuvem. Por exemplo, você pode usar o Amazon EC2 para iniciar e gerenciar servidores virtuais, o AWS Lambda para executar código sem precisar provisionar ou gerenciar servidores, o Amazon ECS ou o Amazon EKS para executar e gerenciar contêineres ou o AWS Batch para processar grandes volumes de dados em paralelo. Com base em sua escala e necessidades de computação, você deve escolher e configurar a solução ideal para sua situação. Você também pode considerar o uso de vários tipos de soluções de computação em uma única workload, pois cada uma tem suas próprias vantagens e desvantagens.

As etapas a seguir orientam você na seleção das opções de computação certas para atender às características da workload e aos requisitos de performance.

Etapas de implementação

- Entenda os requisitos de computação das workloads. Os principais requisitos a serem considerados incluem necessidades de processamento, padrões de tráfego, padrões de acesso a dados, necessidades de ajuste de escala e requisitos de latência.
- Saiba mais sobre os diferentes serviços de computação da AWS para sua workload. Para ter mais informações, consulte PERF01-BP01 Conhecer e compreender os serviços e recursos de nuvem disponíveis. Veja algumas das principais opções de computação da AWS, as características e casos de uso comuns:

Serviço da AWS	Características principais	Casos de uso comuns
Amazon Elastic Compute Cloud (Amazon EC2)	Oferece opção dedicada para hardware, requisitos de licença, grande seleção de diferentes famílias de instâncias, tipos de	Migrações do tipo mover sem alterações (lift-and-shift), aplicações monolíticas, ambientes híbridos, aplicaçõe s empresariais

Serviço da AWS	Características principais	Casos de uso comuns
	processadores e acelerado res de computação.	
Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS)	Implantação fácil, ambientes consistentes, escaláveis	Microsserviços, ambientes híbridos
AWS Lambda	Serviço de computação sem servidor que executa código em resposta a eventos e gerencia automaticamente os recursos computacionais subjacentes.	Microsserviços, aplicações orientadas a eventos
AWS Batch	Provisiona e escala de forma eficiente e dinâmica os recursos de computação do Amazon Elastic Container Service (Amazon ECS), do Amazon Elastic Kubernete s Service (Amazon EKS) e do AWS Fargate, oferecend o a opção de usar instância s sob demanda ou spot com base em seus requisitos de trabalho	HPC, treinamento de modelos de ML.
Amazon Lightsail	Aplicação Linux e Windows pré-configurada para executar pequenas workloads .	Aplicações Web simples, site personalizado.

- Avalie o custo (como cobrança por hora ou transferência de dados) e as despesas gerais de gerenciamento (como aplicação de patches e ajuste de escala) associados a cada opção de computação.

- Faça experimentos e análises comparativas em um ambiente de não produção para identificar qual opção de computação pode atender melhor às necessidades da workload.
- Depois de experimentar e identificar sua nova solução de computação, planeje a migração e valide as métricas de performance.
- Use ferramentas de monitoramento da AWS, como o Amazon CloudWatch, e serviços de otimização, como o AWS Compute Optimizer, para otimizar constantemente a computação com base em padrões de uso real.

Recursos

Documentos relacionados:

- Computação em nuvem com a AWS
- Tipos de instância do Amazon EC2
- Contêineres do Amazon EKS: nós de processamento do Amazon EKS
- Contêineres do Amazon ECS: instâncias de contêiner do Amazon ECS
- Funções: configuração da função do Lambda
- Recomendações para contêineres
- Recomendações para tecnologia sem servidor

Vídeos relacionados:

- AWS re:Invent 2023: AWS Graviton: a melhor performance de preços para suas workloads da AWS
- AWS re:Invent 2023: Novos recursos de IA generativa do Amazon Elastic Compute Cloud no AMS
- AWS re:Invent 2023: Novidades do Amazon Elastic Compute Cloud
- AWS re:Invent 2023: Economias inteligentes: estratégias de otimização de custos no Amazon Elastic Compute Cloud
- AWS re:Invent 2021: Habilitar o Amazon Elastic Compute Cloud da próxima geração: mergulho profundo no Nitro System
- AWS re:Invent 2019: Otimizar a performance e os custos para sua computação na AWS
- AWS re:Invent 2019: Fundamentos da Amazon Elastic Compute Cloud
- AWS re:Invent 2022: Implantar modelos de ML para inferência com alta performance e baixo custo

- AWS re:Invent 2019: Otimizar a performance e os custos para sua computação na AWS
- Fundamentos do Amazon EC2
- Implemente modelos de ML para inferência com alta performance e baixo custo

Exemplos relacionados:

- Migrar aplicações Web para contêineres
- Executar uma aplicação Hello World sem servidor
- Workshop do Amazon EKS
- Workshop do Amazon EC2
- Workloads eficientes e resilientes com o Amazon Elastic Compute Cloud Auto Scaling
- Migrar para o AWS Graviton com serviços de contêiner

PERF02-BP02 Entender a configuração e os recursos de computação disponíveis

Entenda as opções de configuração e os recursos disponíveis para seu serviço de computação a fim de ajudar a provisionar a quantidade certa de recursos e melhorar a eficiência de performance.

Práticas comuns que devem ser evitadas:

- Não avaliar as opções de computação ou as famílias de instâncias disponíveis em relação às características da workload.
- Provisionar recursos de computação em excesso para atender aos requisitos de pico de demanda.

Benefícios de implementar esta prática recomendada: familiarizar-se com os atributos e as configurações de computação da AWS a fim de poder usar uma solução de computação otimizada para atender às características e às necessidades da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Cada solução de computação tem configurações e recursos exclusivos disponíveis para acomodar diferentes características e requisitos das workloads. Saiba como essas opções complementam sua workload e determine quais opções de configuração são melhores para sua aplicação. Exemplos dessas opções são famílias de instâncias, tamanhos, recursos (GPU, E/S), expansão, tempos

limite, tamanhos de função, instâncias de contêineres e simultaneidade. Se a workload estiver usando a mesma opção de computação há mais de quatro semanas, e se a previsão for de que as características permanecerão as mesmas no futuro, você poderá usar o AWS Compute Optimizer para descobrir se sua opção de computação atual é adequada para as workloads de uma perspectiva de CPU e memória.

Etapas de implementação

- Entenda os requisitos da workload (como necessidade de CPU, memória e latência).
- Analise a documentação e as práticas recomendadas da AWS para saber mais sobre as opções de configuração indicadas que podem ajudar a melhorar a performance da computação. Aqui estão algumas das principais opções de configuração a serem consideradas:

Opção de configuração	Exemplos
Tipo de instância	• As instâncias otimizadas para computação são ideais para workloads que exigem uma proporção maior de vCPU/memória. • As instâncias otimizadas para memória entregam grandes quantidades de memória para oferecer compatibilidade com as workloads com uso intenso de memória. • As instâncias otimizadas para armazenamento são projetadas para workloads que exigem alta leitura sequencial e acesso de gravação (IOPS) no armazenamento local.
Modelo de definição de preços	• As instâncias sob demanda permitem usar a capacidade de computação por hora ou segundo sem uma confirmação de longo prazo. Essas instâncias são ideais para expansões acima das necessidades de performance da linha de base. • Os Savings Plans oferecem economias significativas em relação às instâncias sob demanda em troca do compromisso de usar uma quantidade específica de potência

Opção de configuração	Exemplos
	computacional por um período de um ou três anos. • As instâncias spot permitem que você aproveite a capacidade de instância não utilizada com um desconto para as workloads sem estado e tolerantes a falhas.
Auto Scaling	Use a configuração de Auto Scaling para combinar recursos computacionais com padrões de tráfego.
Dimensionamento	• Use o Compute Optimizer para obter uma recomendação de machine learning sobre a configuração de computação que correspon de melhor às características da computaçã o. • Use o AWS Lambda Power Tuning para selecionar a melhor configuração para a função do Lambda.
Aceleradores de computação baseados em hardware	• As instâncias com computação acelerada executam funções como processamento gráfico ou correspondência de padrões de dados com mais eficiência do que as alternativas baseadas em CPU. • Para workloads de machine learning, utilize hardware específico para sua workload, como AWS Trainium, AWS Inferentia e Amazon EC2 DL1

Recursos

Documentos relacionados:

- Computação em nuvem com a AWS
- Tipos de instância do Amazon EC2
- Controle do estado do processador para sua instância do Amazon EC2
- Contêineres do Amazon EKS: nós de processamento do Amazon EKS
- Contêineres do Amazon ECS: instâncias de contêiner do Amazon ECS
- Funções: configuração da função do Lambda

Vídeos relacionados:

- AWS re:Invent 2023: AWS Graviton: a melhor performance de preço para suas workloads da AWS
- AWS re:Invent 2023: Novos recursos de IA generativa do Amazon EC2 no AWS Management Console
- AWS re:Invent 2023: Novidades do Amazon EC2
- AWS re:Invent 2023: Economias inteligentes: estratégias de otimização de custos com o Amazon EC2
- AWS re:Invent 2021: Capacitar o Amazon EC2 da próxima geração: mergulho profundo no Nitro System
- AWS re:Invent 2019: Fundamentos do Amazon EC2
- AWS re:Invent 2022: Otimizar o Amazon EKS para performance e custo na AWS

Exemplos relacionados:

- Código de demonstração do Compute Optimizer
- Workshop sobre instâncias spot do Amazon EC2
- Workloads eficientes e resilientes com o Amazon EC2 AWS Auto Scaling
- Workshop de desenvolvedores para Graviton
- Dia de imersão em workloads da AWS para Microsoft
- Dia de imersão em workloads da AWS para Linux
- Código de demonstração do AWS Compute Optimizer
- Workshop do Amazon EKS

PERF02-BP03 Coletar métricas relacionadas à computação

Registre e acompanhe métricas relacionadas à computação para entender melhor a performance dos seus recursos e melhorar sua performance e utilização.

Práticas comuns que devem ser evitadas:

- Você só usa a pesquisa manual de arquivos de log para métricas.
- Você só usa as métricas padrão registradas pelo software de monitoramento.
- Você só revisa as métricas quando há um problema.

Benefícios de implementar esta prática recomendada: a coleta de métricas relacionadas à performance ajudará você a alinhar a performance da aplicação aos requisitos empresariais para garantir que você atenda às necessidades da workload. Isso também pode ajudar a melhorar constantemente a performance e a utilização dos recursos na workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

As workloads na nuvem podem gerar grandes volumes de dados, como métricas, logs e eventos. Na Nuvem AWS, coletar métricas é uma etapa essencial para melhorar a segurança, a eficiência de custos, a performance e a sustentabilidade. A AWS oferece uma ampla variedade de métricas relacionadas à performance usando serviços de monitoramento, por exemplo, o Amazon CloudWatch, para fornecer insights valiosos. Métricas como utilização de CPU, utilização de memória, E/S de disco e entrada e saída da rede podem fornecer informações sobre os níveis de utilização ou gargalos de performance. Use essas métricas como parte de uma abordagem impulsionada por dados para ajustar e otimizar ativamente os recursos de sua workload. Em um caso ideal, você deve coletar todas as métricas relacionadas aos recursos de computação em uma única plataforma com políticas de retenção implementadas para apoiar as metas operacionais e de custo.

Etapas de implementação

- Identifique quais métricas relacionadas à performance são relevantes para a workload. Você deve coletar métricas sobre a utilização de recursos e a forma como a workload na nuvem está operando (como tempo de resposta e throughput).
  - Métricas padrão do Amazon EC2

  - Métricas padrão do Amazon ECS
  - Métricas padrão do Amazon EKS
  - Métricas padrão do Lambda
  - Métricas de memória e disco do Amazon EC2
- Escolha e configure a solução certa de registro e monitoramento para a workload.
  - Observabilidade nativa da AWS
  - AWS Distro para OpenTelemetry
  - Amazon Managed Service for Prometheus
- Defina o filtro e a agregação necessários para as métricas com base nos requisitos da workload.
  - Quantificar méticas de aplicações personalizadas com o Amazon CloudWatch Logs e filtros métricos
  - Coletar métricas personalizadas com a marcação com tags estratégica do Amazon CloudWatch
- Configure políticas de retenção de dados para que as métricas correspondam às metas operacionais e de segurança.
  - Retenção de dados padrão para métricas do CloudWatch
  - Retenção de dados padrão para CloudWatch Logs
- Se necessário, crie alarmes e notificações para as métricas a fim de ajudar a reagir proativamente a problemas relacionados à performance.
  - Criar alarmes para métricas personalizadas usando a detecção de anomalias do Amazon CloudWatch
  - Criar métricas e alarmes para páginas da Web específicas com o Amazon CloudWatch RUM
- Use a automação para implantar os agentes de agregação de métricas e logs.
  - Automação do AWS Systems Manager
  - Coletor do OpenTelemetry

Recursos

Documentos relacionados:

- Monitoramento e observabilidade
- Práticas recomendadas: implementar a observabilidade com a AWS
- Documentação do Amazon CloudWatch

- Coletar métricas e logs de instâncias do Amazon EC2 e servidores on-premises com o agente do CloudWatch
- Acessar o Amazon CloudWatch Logs para AWS Lambda
- Usar o CloudWatch Logs com Instâncias de contêiner
- Publicar métricas personalizadas
- AWS Answers: log centralizado
- Serviços da AWS que publicam métricas do CloudWatch
- Monitorar o Amazon EKS no AWS Fargate

Vídeos relacionados:

- AWS re:Invent 2023 [LANÇAMENTO]: Monitoramento de aplicações para workloads modernas
- AWS re:Invent 2023: Como implementar a observabilidade de aplicações
- AWS re:Invent 2023: Criar uma estratégia efetiva de observabilidade
- AWS re:Invent 2023: Observabilidade direta com o AWS Distro para OpenTelemetry
- Gerenciamento da performance de aplicações na AWS

Exemplos relacionados:

- Dia de imersão em workloads da AWS para Linux - Amazon CloudWatch
- Monitorar clusters e contêineres do Amazon ECS
- Monitorar com painéis do Amazon CloudWatch
- Workshop do Amazon EKS

PERF02-BP04 Configurar e dimensionar corretamente os recursos de computação

Configure e dimensione corretamente os recursos de computação para atender aos requisitos de performance das workloads e evitar que recursos sejam subutilizados ou usados em excesso.

Práticas comuns que devem ser evitadas:

- Ignorar os requisitos de performance das workloads, o que ocasiona recursos computacionais superprovisionados ou subprovisionados.
- Você escolhe somente a maior ou a menor instância disponível para todas as workloads.

- Você usa apenas uma família de instâncias para facilitar o gerenciamento.
- Você ignora as recomendações do AWS Cost Explorer ou do Compute Optimizer para o dimensionamento correto.
- Você não reavalia a workload quanto à adequação dos novos tipos de instância.
- Você certifica apenas um pequeno número de configurações de instâncias para sua organização.

Benefícios de implementar esta prática recomendada: o dimensionamento correto dos recursos computacionais garante a operação ideal na nuvem, evitando o provisionamento excessivo e o subprovisionamento de recursos. O dimensionamento adequado dos recursos de computação normalmente resulta em melhor performance e melhor experiência do cliente, além de reduzir custos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

O dimensionamento correto permite que as organizações operem a infraestrutura de nuvem de forma eficiente e econômica ao mesmo tempo que atendem às suas necessidades comerciais. O provisionamento excessivo de recursos na nuvem pode gerar custos extras, enquanto o provisionamento insuficiente pode resultar em baixa performance e em uma experiência negativa para o cliente. A AWS fornece ferramentas como AWS Compute Optimizer e AWS Trusted Advisor que usam dados históricos para fornecer recomendações para dimensionar corretamente seus recursos computacionais.

Etapas de implementação

- Escolha um tipo de instância que melhor atenda às suas necessidades:
  - Como faço para escolher o tipo de instância do Amazon EC2 apropriado para minha workload?
  - Seleção de tipo de instância baseada em atributos para o Amazon EC2 Fleet
  - Criar um grupo do Auto Scaling usando seleção de tipo de instância baseada em atributos
  - Otimizar seus custos de computação do Kubernetes com a consolidação do Karpenter
- Analise as várias características de performance da sua workload e como elas se relacionam a uso de memória, rede e CPU. Use esses dados para escolher os recursos que melhor correspondam ao perfil e às metas de performance da workload.
- Monitore o uso de recursos usando ferramentas de monitoramento da AWS, como o Amazon CloudWatch.
- Selecione a configuração correta para os recursos computacionais.

  - Para workloads efêmeras, avalie as métricas da instância do Amazon CloudWatch, como `CPUUtilization`, para identificar se a instância está subutilizada ou superutilizada.
  - Para workloads estáveis, verifique as ferramentas de dimensionamento correto da AWS, como AWS Compute Optimizer e AWS Trusted Advisor em intervalos regulares para identificar oportunidades de otimizar e dimensionar corretamente o recurso de computação.
- Teste as alterações na configuração em um ambiente de não produção antes de implementá-las em um ambiente ativo.
- Reavalie constantemente novas ofertas de computação e compare-as com as necessidades da workload.

Recursos

Documentos relacionados:

- Computação na nuvem com a AWS
- Tipos de instância do Amazon EC2
- Contêineres do Amazon ECS: instâncias de contêiner do Amazon ECS
- Contêineres do Amazon EKS: nós de processamento do Amazon EKS
- Funções: configuração da função do Lambda
- Controle do estado do processo para sua instância do Amazon EC2

Vídeos relacionados:

- Fundamentos do Amazon EC2
- AWS re:Invent 2023: AWS Graviton: a melhor performance de preço para suas workloads da AWS
- AWS re:Invent 2023: Novos recursos de IA generativa do Amazon EC2 no AWS Management Console
- AWS re:Invent 2023: Novidades do Amazon EC2
- AWS re:Invent 2023: Economias inteligentes: estratégias de otimização de custos com o Amazon EC2
- AWS re:Invent 2021: Capacitar o Amazon EC2 da próxima geração: mergulho profundo no Nitro System
- AWS re:Invent 2019: Fundamentos do Amazon EC2

Exemplos relacionados:

- Código de demonstração do AWS Compute Optimizer
- Workshop do Amazon EKS
- Recomendações de dimensionamento correto

PERF02-BP05 Dimensionar recursos de computação dinamicamente

Use a elasticidade da nuvem para aumentar ou diminuir os recursos de computação dinamicamente a fim de atender às suas necessidades e evitar provisionamento excessivo ou insuficiente da capacidade para a workload.

Práticas comuns que devem ser evitadas:

- Reagir a alarmes aumentando a capacidade manualmente.
- Usar as mesmas diretrizes de dimensionamento (geralmente infraestrutura estática) do ambiente on-premises.
- Manter a capacidade aumentada após um evento de ajuste de escala, em vez de reduzi-la novamente.

Benefícios de implementar esta prática recomendada: configurar e testar a elasticidade dos recursos computacionais pode ajudar você a economizar dinheiro, manter os benchmarks de performance e melhorar a confiabilidade à medida que o tráfego muda.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

A AWS oferece a flexibilidade de aumentar ou diminuir seus recursos dinamicamente por meio de uma variedade de mecanismos de ajuste de escala a fim de atender às mudanças na demanda. Combinado com métricas relacionadas à computação, um ajuste de escala dinâmico permite que as workloads respondam automaticamente às mudanças e usem o conjunto ideal de recursos computacionais para atingir sua meta.

Você pode usar diversas abordagens diferentes para corresponder a oferta de recursos com a demanda.

- Abordagem de monitoramento de meta: monitore a métrica de ajuste de escala e aumente ou diminua automaticamente a capacidade conforme necessário.

- Ajuste de escala preditivo: aumente ou reduza a escala em antecipação às tendências diárias e semanais.
- Abordagem baseada em cronograma: defina seu próprio cronograma de ajuste de escala de acordo com as mudanças de carga previsíveis.
- Ajuste de escala de serviços: escolha serviços (como de tecnologia sem servidor) que sejam escalados automaticamente de acordo com o projeto.

É necessário garantir que as implantações de workload possam lidar com eventos de expansão e redução da escala.

Etapas de implementação

- Instâncias, contêineres e funções de computação oferecem mecanismos para elasticidade, seja em combinação com o ajuste de escala automático ou como um recurso do serviço. Veja alguns exemplos de mecanismos de ajuste de escala automático:

Mecanismo de ajuste de escala automático	Onde usar
Amazon EC2 Auto Scaling	Ajuda a garantir que você tenha o número correto de instâncias do Amazon EC2 disponíveis para processar a carga da aplicação.
Application Auto Scaling	Para escalar automaticamente os recursos para serviços da AWS individuais além do Amazon EC2, como funções do AWS Lambda ou serviços do Amazon Elastic Container Service (Amazon ECS).
Kubernetes Cluster Autoscaler/Karpenter	Para escalar automaticamente os clusters do Kubernetes.

- O ajuste de escala geralmente é discutido em relação a serviços de computação, como instâncias do Amazon EC2 ou funções do AWS Lambda. Não se esqueça de considerar também a configuração de serviços não computacionais, como AWS Glue, para atender à demanda.
- Verifique se as métricas de ajuste de escala correspondem às características da workload que está sendo implantada. Se você estiver implantando uma aplicação de transcodificação de vídeo,

espera-se que a utilização da CPU seja de 100%, e essa não deve ser sua métrica principal. Use a profundidade da fila de trabalhos de transcodificação. Você pode usar uma métrica personalizada para sua política de ajuste de escala, se necessário. Para escolher as métricas certas, considere a seguinte orientação para o Amazon EC2:

  - A métrica deve ser uma métrica de utilização válida e descrever o quanto uma instância está ocupada.
  - O valor da métrica deve aumentar e diminuir em proporção ao número das instâncias no grupo do Auto Scaling.
- Certifique-se de usar o ajuste de escala dinâmico em vez do ajuste de escala manual para seu grupo do Auto Scaling. Também recomendamos usar políticas de ajuste de escala de rastreamento de metas em seu ajuste de escala dinâmico.
- Verifique se as implantações da workload podem lidar com os dois eventos de ajuste de escala (aumento e redução). Como exemplo, você pode usar o Histórico de atividades para verificar uma atividade de ajuste de escala em um grupo do Auto Scaling.
- Avalie sua workload em relação a padrões previsíveis e, ao antecipar alterações previstas e planejadas na demanda, ajuste a escala proativamente. Com o ajuste de escala preditivo, é possível eliminar a necessidade de superprovisionar a capacidade. Para obter mais informações, consulte Ajuste de escala preditivo com o Amazon EC2 Auto Scaling.

Recursos

Documentos relacionados:

- Computação na nuvem com a AWS
- Tipos de instância do Amazon EC2
- Contêineres do Amazon ECS: instâncias de contêiner do Amazon ECS
- Contêineres do Amazon EKS: nós de processamento do Amazon EKS
- Funções: configuração da função do Lambda
- Controle do estado do processo para sua instância do Amazon EC2
- Mergulho profundo no ajuste de escala automático de clusters do Amazon ECS
- Introdução ao Karpenter: um dimensionador automático de clusters do Kubernetes de código aberto e alta performance

Vídeos relacionados:

- AWS re:Invent 2023: AWS Graviton: a melhor performance de preço para suas workloads da AWS
- AWS re:Invent 2023: Novos recursos de IA generativa no Amazon EC2 no console de gerenciamento da AWS
- AWS re:Invent 2023: Novidades do Amazon EC2
- AWS re:Invent 2023: Economias inteligentes: estratégias de otimização de custos com o Amazon EC2
- AWS re:Invent 2021: Capacitar o Amazon EC2 da próxima geração: mergulho profundo no Nitro System
- AWS re:Invent 2019: Fundamentos do Amazon EC2

Exemplos relacionados:

- Exemplos de grupos do Amazon EC2 Auto Scaling
- Workshop do Amazon EKS
- Escalar suas workloads do Amazon EKS executando-as em IPv6

PERF02-BP06 Usar aceleradores de computação baseados em hardware otimizados

Use aceleradores de hardware para executar determinadas funções com mais eficiência do que as alternativas baseadas em CPU.

Práticas comuns que devem ser evitadas:

- Em sua workload, você não compara uma instância de uso geral com uma instância criada para um propósito específico capaz de oferecer maior performance e menor custo.
- Você está usando aceleradores de computação baseados em hardware para tarefas que podem ser eficientes com o uso de alternativas baseadas em CPU.
- Você não está monitorando o uso da GPU.

Benefícios de implementar esta prática recomendada: ao usar aceleradores baseados em hardware, como unidades de processamento gráfico (GPUs) e matrizes de portas programáveis em campo (FPGAs), você pode executar determinadas funções de processamento com mais eficiência.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

As instâncias com computação acelerada fornecem acesso a aceleradores de computação baseados em hardware, como GPUs e FPGAs. Esses aceleradores de hardware executam certas funções, como processamento gráfico ou correspondência de padrões de dados, com mais eficiência do que alternativas baseadas em CPU. Muitas workloads aceleradas, como renderização, transcodificação e machine learning, são altamente variáveis em termos de uso de recursos. Execute esse hardware apenas pelo tempo necessário e desative-o com automação quando não precisar mais dele para melhorar a eficiência da performance geral.

Etapas de implementação

- Identifique quais instâncias com computação acelerada podem atender às suas necessidades.
- Para workloads de machine learning, utilize hardware específico para sua workload, como AWS Trainium, AWS Inferentia e Amazon EC2 DL1. AWS Instâncias do Inferentia, como instâncias Inf2, oferecem performance até 50% melhor por watt em relação a instâncias comparáveis do Amazon EC2.
- Colete métricas de uso para as instâncias com computação acelerada. Por exemplo, você pode usar o agente do CloudWatch para coletar métricas como `utilization_gpu` e `utilization_memory` e para suas GPUs, conforme mostrado em Coletar métricas de GPU NVIDIA com o Amazon CloudWatch.
- Otimize o código, a operação de rede e as configurações dos aceleradores de hardware para garantir que o hardware subjacente seja totalmente utilizado.
  - Otimizar as configurações da GPU
  - Monitoramento e otimização da GPU na AMI de aprendizado profundo
  - Otimizar a E/S para ajuste de performance da GPU de treinamento de aprendizado profundo no Amazon SageMaker
- Use as mais recentes bibliotecas de alta performance e drivers de GPU.
- Use automação para liberar instâncias de GPU quando não estiverem em uso.

Recursos

Documentos relacionados:

- Trabalhar com GPUs no Amazon Elastic Container Service
- Instâncias de GPU

- Instâncias com AWS Trainium
- Instâncias com o AWS Inferentia
- Vamos arquitetar! Como arquitetar com chips e aceleradores personalizados

- Computação acelerada
- Instâncias VT1 do Amazon EC2
- Como faço para escolher o tipo de instância do Amazon EC2 apropriado para minha workload?
- Escolher o melhor acelerador de IA e compilação de modelos para inferência de visão computacional com o Amazon SageMaker

Vídeos relacionados:

- AWS re:Invent 2021: Como selecionar instâncias de GPU do Amazon Elastic Compute Cloud para aprendizado profundo
- AWS re:Invent 2022 [NOVO LANÇAMENTO!]: Introdução as instâncias Inf2 do Amazon EC2 baseadas no AWS Inferentia2
- AWS re:Invent 2022: Acelerar o aprendizado profundo e inovar com mais rapidez com o AWS Trainium
- AWS re:Invent 2022: Aprendizado profundo na AWS com a NVIDIA: do treinamento à implantação

Exemplos relacionados:

- Amazon SageMaker e NVIDIA GPU Cloud (NGC)
- Use o SageMaker com Trainium e Inferentia para workloads otimizadas de treinamento e inferência em aprendizado profundo
- Otimizar modelos de PLN com instâncias Inf1 do Amazon Elastic Compute Cloud no Amazon SageMaker

Gerenciamento de dados

Perguntas

- PERF 3. Como armazenar, gerenciar e acessar dados em sua workload?

PERF 3. Como armazenar, gerenciar e acessar dados em sua workload?

A solução de gerenciamento de dados ideal para um sistema específico varia conforme o tipo de dados (bloco, arquivo ou objeto), os padrões de acesso (aleatório ou sequencial), o throughput necessário, a frequência de acesso (online, offline, arquivamento), a frequência de atualização (WORM, dinâmica) e as restrições de disponibilidade e durabilidade. As workloads do Well-Architected usam datastores específicos que permitem que recursos diferentes melhorem a performance.

Práticas recomendadas

- PERF03-BP01 Usar um datastore com propósitos específicos que melhor atenda aos requisitos de acesso e armazenamento de dados
- PERF03-BP02 Avaliar as opções de configuração disponíveis para o datastore
- PERF03-BP03 Coletar e registrar métricas de performance do datastore
- PERF03-BP04 Implementar estratégias para melhorar a performance da consulta no datastore
- PERF03-BP05 Implementar padrões de acesso a dados que utilizam cache

PERF03-BP01 Usar um datastore com propósitos específicos que melhor atenda aos requisitos de acesso e armazenamento de dados

Entenda as características dos dados (como possibilidade de compartilhamento, tamanho, tamanho do cache, padrões de acesso, latência, throughput e persistência dos dados) a fim de selecionar os datastores com propósito específico (armazenamento ou banco de dados) para sua workload.

Práticas comuns que devem ser evitadas:

- Utilizar um único datastore porque há experiência e conhecimento internos de um tipo específico de solução de banco de dados.
- Você pressupõe que todas as workloads têm requisitos de acesso e armazenamento de dados semelhantes.
- Você não implementou um catálogo de dados para criar um inventário de seus ativos de dados.

Benefícios de implementar esta prática recomendada: entender as características e os requisitos de dados permite que você determine a tecnologia de armazenamento mais eficiente e com melhor performance adequada às necessidades da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Ao selecionar e implementar o armazenamento de dados, certifique-se de que as características de consulta, ajuste de escala e armazenamento atendam aos requisitos de dados da workload. A AWS fornece várias tecnologias de armazenamento de dados e banco de dados, incluindo armazenamento em blocos, armazenamento de objetos, armazenamento de streaming, sistema de arquivos, bancos de dados relacionais, de chave-valor, de documentos, na memória, de grafos, de séries temporais e ledger. Cada solução de gerenciamento de dados tem opções e configurações disponíveis para compatibilidade com seus casos de uso e modelos de dados. Ao compreender as características e os requisitos dos dados, você pode se separar da tecnologia de armazenamento monolítico e das abordagens restritivas e únicas para se concentrar no gerenciamento adequado dos dados.

Etapas de implementação

- Realize um inventário dos vários tipos de dados que existem na workload.
- Entenda e documente as características e os requisitos dos dados, incluindo:
  - Tipo de dados (não estruturados, semiestruturados, relacionais)
  - Volume e crescimento de dados
  - Durabilidade dos dados: persistentes, efêmeros, transitórios
  - Requisitos de ACID (atomicidade, consistência, isolamento, durabilidade)
  - Padrões de acesso a dados (com muita leitura ou gravação)
  - Latência
  - Throughput
  - IOPS (operações de entrada/saída por segundo)
  - Período de retenção de dados
- Conheça os diferentes datastores (serviços de armazenamento e banco de dados) disponíveis para a workload na AWS que podem atender às características dos dados (conforme descrito em PERF01-BP01 Conhecer e compreender os serviços e recursos de nuvem disponíveis. Alguns exemplos de tecnologias de armazenamento da AWS e suas principais características incluem:

Tipo	Serviços da AWS	Características principais
Armazenamento de objetos	Amazon S3	Escalabilidade ilimitada, alta disponibilidade e várias

Tipo	Serviços da AWS	Características principais
		opções de acessibilidade. A transferência e o acesso a objetos dentro e fora do Amazon S3 podem usar um serviço, como o Transfer Acceleration ou Access Points, para oferecer suporte a sua localização, necessidades de segurança e padrões de acesso.
Armazenamento de arquivamento	Amazon S3 Glacier	Desenvolvido para arquivamento de dados.
Armazenamento de streaming	Amazon Kinesis Amazon Managed Streaming for Apache Kafka (Amazon MSK)	Ingestão e armazenamento eficientes de dados de streaming.
Sistema de arquivos compartilhado	Amazon Elastic File System (Amazon EFS)	Sistema de arquivos montável que pode ser acessado por vários tipos de soluções de computação.

Tipo	Serviços da AWS	Características principais
Sistema de arquivos compartilhado	Amazon FSx	Baseia-se nas soluções de computação mais recentes da AWS para oferecer compatibilidade com quatro sistemas de arquivos usados com frequência: NetApp ONTAP, OpenZFS, Windows File Server e Lustre. A latência, o throughput e as IOPS do Amazon FSx variam de acordo com o sistema de arquivos e devem ser consideradas ao selecionar o sistema de arquivos certo para as necessidades da sua workload.
Armazenamento em bloco	Amazon Elastic Block Store (Amazon EBS)	Serviço de armazenamento em blocos fácil de usar, escalável e de alta performance projetado para o Amazon Elastic Compute Cloud (Amazon EC2). O Amazon EBS inclui armazenamento baseado em SSD para workloads transacionais de alto throughput e em HDD para workloads trabalho de alto throughput.

Tipo	Serviços da AWS	Características principais
Banco de dados relacional	Amazon Aurora, Amazon RDS, Amazon Redshift.	Projetados para oferecer compatibilidade com transações ACID (atomicidade, consistência, isolamento, durabilidade) e manter a integridade referencial e uma forte consistência de dados. Muitas aplicações tradicionais, planejamento de recursos empresariais (ERP), gerenciamento de relacionamentos com o cliente (CRM) e comércio eletrônico usam bancos de dados relacionais para armazenar os dados.
Banco de dados de chave-valor	Amazon DynamoDB	Otimizados para padrões de acesso comuns, normalmente visando armazenar e recuperar grandes volumes de dados. Aplicações web de alto tráfego, sistemas de comércio eletrônico e aplicações de jogos são os casos de uso habituais para bancos de dados de chave-valor.

Tipo	Serviços da AWS	Características principais
Banco de dados de documentos	Amazon DocumentDB	Projetado para armazenar dados semiestruturados, como documentos do tipo JSON. Esses bancos de dados ajudam os desenvolvedores a criar e atualizar rapidamente aplicações de gerenciamento de conteúdo, catálogos e perfis de usuário, por exemplo.
Banco de dados na memória	Amazon ElastiCache , Amazon MemoryDB for Redis	Usados para aplicações que exigem acesso em tempo real aos dados, latência mais baixa e throughput mais alto. É possível usar bancos de dados na memória para armazenamento em cache de aplicações, gerenciamento de sessões, tabelas de classificação de jogos, arquivo de atributos de ML de baixa latência, sistema de mensagens de microsserviços e um mecanismo de streaming de alto throughput.

Tipo	Serviços da AWS	Características principais
Banco de dados de grafos	Amazon Neptune	Utilizado para aplicações que precisam navegar e consultar milhões de relacionamentos entre conjuntos de dados de grafos altamente conectados com latência de milissegundos em grande escala. Muitas empresas usam bancos de dados gráficos para detecção de fraudes, redes sociais e mecanismos de recomendação.
Banco de dados de séries temporais	Amazon Timestream	Utilizado para coletar, sintetizar e gerar com eficiência insights de dados que mudam ao longo do tempo. Aplicações de IoT, DevOps e telemetria industrial podem utilizar bancos de dados de séries temporais.
Coluna ampla	Amazon Keyspaces (for Apache Cassandra)	Usa tabelas, linhas e colunas, mas ao contrário de um banco de dados relacional, os nomes e o formato das colunas podem variar de linha para linha na mesma tabela. Normalmente, você vê um repositório de coluna ampla em aplicações industriais de alta escala para manutenção de equipamentos, gerenciamento de frotas e otimização de rotas.

Tipo	Serviços da AWS	Características principais
ledger	Amazon Quantum Ledger Database (Amazon QLDB)	Oferece uma autoridade centralizada e confiável para manter um registro escalável, imutável e criptograficamente verificável de transações para cada aplicação. Vemos os bancos de dados de livro-razão empregados em sistemas de registro, cadeia de suprimentos, inscrições e até mesmo transações bancárias.

- Se você estiver criando uma plataforma de dados, utilize a arquitetura de dados moderna na AWS para integrar seu data lake, data warehouse e datastores específicos.
- As principais questões que você precisa considerar ao escolher um datastore para sua workload são as seguintes:

Pergunta	Fatos a serem considerados
Como os dados são estruturados?	• Se os dados não forem estruturados, considere um armazenamento de objetos, como o Amazon S3, ou um banco de dados NoSQL, como o Amazon DocumentDB. • Para dados de valor-chave, considere o DynamoDB, o Amazon ElastiCache (Redis OSS) ou o Amazon MemoryDB.
Que nível de integridade referencial é necessário?	• Para restrições de chave estrangeira, bancos de dados relacionais como Amazon RDS e Aurora podem fornecer esse nível de integridade. • Normalmente, em um modelo de dados NoSQL, você desnormalizaria os dados

Pergunta	Fatos a serem considerados
	em um único documento ou coleção de documentos para serem recuperados em uma única solicitação em vez de unir documentos ou tabelas de diferentes locais.
A conformidade com ACID (atomicidade, consistência, isolamento, durabilidade) é necessária?	• Se as propriedades ACID associadas aos bancos de dados relacionais forem necessárias, pense em um banco de dados relacional, como o Amazon RDS e o Aurora. • Se uma consistência forte for necessária para o banco de dados NoSQL, você pode usar leituras altamente consistentes com o DynamoDB.
Como as necessidades de armazenamento serão alteradas ao longo do tempo? Como isso afeta a escalabilidade?	• Os bancos de dados sem servidor, como o DynamoDB e o Amazon Quantum Ledger Database (Amazon QLDB), serão escalados dinamicamente. • Os bancos de dados relacionais têm limites superiores em armazenamento provisionado e devem ser particionados horizontalmente usando mecanismos, como fragmentação, quando atingem esses limites.
Qual é a proporção de consultas de leitura em relação a consultas de gravação? O armazenamento em cache melhoraria a performance?	• Workloads com muitas operações de leitura poderão se beneficiar de uma camada de cache, como ElastiCache ou DAX, se o banco de dados for o DynamoDB. • As leituras também podem ser descarreg adas em réplicas de leitura com bancos de dados relacionais, como o Amazon RDS.

Pergunta	Fatos a serem considerados
O armazenamento e a modificação (OLTP – Processamento de transações on-line) ou a recuperação e a geração de relatórios (OLAP – Processamento analítico on-line) têm uma prioridade mais alta?	• Para um processamento transacional de throughput alto de leitura no estado em que se encontra, considere um banco de dados NoSQL, como o DynamoDB. • Para padrões de leitura complexos e de throughput alto (como junção) com consistência use o Amazon RDS. • Para consultas analíticas, considere usar um banco de dados colunar, como o Amazon Redshift, ou exportar os dados para o Amazon S3 e realizar análises usando o Athena ou o Amazon QuickSight.
Que nível de durabilidade os dados exigem?	• O Aurora replica automaticamente os dados entre três zonas de disponibilidade em uma região, o que significa que seus dados terão mais durabilidade com menos chance de serem perdidos. • O DynamoDB é automaticamente replicado entre várias zonas de disponibilidade, fornecendo alta disponibilidade e durabilid ade aos dados. • O Amazon S3 fornece 11 noves de durabilid ade. Muitos serviços de banco de dados, como o Amazon RDS e o DynamoDB, são compatíveis com a exportação de dados para o Amazon S3 para retenção de longo prazo e arquivamento.

Pergunta	Fatos a serem considerados
Você quer se livrar de mecanismos de bancos de dados comerciais ou custos de licenças?	• Considere usar mecanismos de código aberto, como PostgreSQL e MySQL no Amazon RDS ou Aurora. • Utilize o AWS Database Migration Service e o AWS Schema Conversion Tool para realizar migrações de mecanismos de bancos de dados comerciais para código aberto
Qual é a expectativa operacional para o banco de dados? A migração para serviços gerenciados é uma preocupação importante?	• Utilizar o Amazon RDS em vez do Amazon EC2 e o DynamoDB ou o Amazon DocumentDB em vez de um host automático de um banco de dados NoSQL pode reduzir a sobrecarga operacional.
Como o banco de dados é acessado no momento? Ele é acessado apenas por aplicações ou há usuários de inteligência de negócios (BI) e outras aplicações prontas para uso conectadas?	• Se houver dependências de ferramentas externas, talvez seja necessário manter a compatibilidade com os bancos de dados que elas suportam. O Amazon RDS é totalmente compatível com as diferentes versões do mecanismo a que ele oferece suporte, incluindo Microsoft SQL Server, Oracle, MySQL e PostgreSQL.

- Faça experimentos e testes comparativos em um ambiente de não produção para identificar qual datastore pode atender às necessidades da workload.

Recursos

Documentos relacionados:

- Tipos de volume do Amazon EBS
- Armazenamento do Amazon EC2
- Amazon EFS: performance do Amazon EFS
- Performance do Amazon FSx para Lustre

- Performance do Amazon FSx para Windows File Server
- Amazon S3 Glacier: documentação do S3 Glacier
- Amazon S3: considerações sobre performance e taxa de solicitações
- Armazenamento na nuvem com a AWS
- Características de E/S do Amazon EBS
- Bancos de dados na nuvem com a AWS
- Cache de banco de dados da AWS
- DynamoDB Accelerator
- Práticas recomendadas do Amazon Aurora
- Performance do Amazon Redshift
- As 10 melhores dicas de performance para Amazon Athena
- Práticas recomendadas do Amazon Redshift Spectrum
- Práticas recomendadas do Amazon DynamoDB
- Escolher entre o Amazon EC2 e o Amazon RDS
- Práticas recomendadas de implementação do Amazon ElastiCache

Vídeos relacionados:

- AWS re:Invent 2023: Melhorar a eficiência do Amazon Elastic Block Store e ser mais econômico
- AWS re:Invent 2023: Otimizar o preço e a performance do armazenamento com o Amazon Simple Storage Service
- AWS re:Invent 2023: Criar e otimizar data lakes no Amazon Simple Storage Service
- AWS re:Invent 2022: Construir arquiteturas de dados modernos na AWS
- AWS re:Invent 2022: Construir arquiteturas de data mesh na AWS
- AWS re:Invent 2023: Mergulho profundo no Amazon Aurora e suas inovações
- AWS re:Invent 2023: Modelagem de dados com o Amazon DynamoDB
- AWS re:Invent 2022: Modernizar aplicações com bancos de dados com propósito específico
- Mergulho profundo no Amazon DynamoDB: padrões de design avançados (DAT403-R1)

Exemplos relacionados:

- Workshop de bancos de dados com propósito específico na AWS

- Bancos de dados para desenvolvedores
- Dia de imersão na arquitetura de dados moderna na AWS
- Criar um data mesh na AWS
- Exemplos do Amazon S3
- Otimizar o padrão de dados usando o compartilhamento de dados do Amazon Redshift
- Migrações de bancos de dados
- MS SQL Server: demonstração da replicação do AWS Database Migration Service (AWS DMS)
- Workshop prático de modernização de bancos de dados
- Exemplos do Amazon Neptune

PERF03-BP02 Avaliar as opções de configuração disponíveis para o datastore

Entenda e avalie os vários atributos e opções de configuração disponíveis para seus datastores a fim de otimizar o espaço de armazenamento e a performance da workload.

Práticas comuns que devem ser evitadas:

- Você só usa um tipo de armazenamento, como o Amazon EBS, para todas as workloads.
- Você usa as IOPS provisionadas para todas as workloads sem testes reais em todos os níveis de armazenamento.
- Você não sabe quais são as opções de configuração da solução de gerenciamento de dados escolhida.
- Você conta somente com o aumento do tamanho da instância sem examinar outras opções de configuração.
- Você não testa as características de ajuste de escala do datastore.

Benefícios de implementar esta prática recomendada: a exploração e a experimentação das configurações de datastore permitem que você reduza o custo da infraestrutura, melhore a performance e diminua o esforço necessário para manter as workloads.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Uma workload pode ter um ou mais datastores usados com base nos requisitos de armazenamento e acesso aos dados. Para otimizar a eficiência de performance e custos, é necessário avaliar

os padrões de acesso aos dados para determinar as configurações apropriadas do datastore. Ao explorar as opções de datastore, leve em consideração vários aspectos, como opções de armazenamento, memória, computação, réplica de leitura, requisitos de consistência, grupo de conexões e opções de armazenamento em cache. Experimente essas várias opções de configuração para melhorar as métricas de eficiência de performance.

Etapas de implementação

- Entenda as configurações atuais (como tipo de instância, tamanho do armazenamento ou versão do mecanismo de banco de dados) do datastore.
- Analise a documentação e as práticas recomendadas da AWS para saber mais sobre as opções de configuração indicadas que podem ajudar a melhorar a performance do datastore. As principais opções de datastore a serem consideradas são:

Opção de configuração	Exemplos
Descarregar leituras (como réplicas de leitura e cache)	• Nas tabelas do DynamoDB, é possível descarregar leituras usando o DAX para armazenamento em cache. • É possível criar um cluster do Amazon ElastiCache (Redis OSS) e configurar a aplicação para ler primeiro do cache e voltar para o banco de dados caso o item solicitado não esteja presente. • Todos os bancos de dados relacionais, como Amazon RDS e Aurora, e bancos de dados NoSQL provisionados, como Neptune e Amazon DocumentDB, permitem adicionar réplicas de leitura para descarregar as partes de leitura da workload. • Os bancos de dados de tecnologia sem servidor, como o DynamoDB, ajustarão a escala automaticamente. Verifique se você tem unidades de capacidade de leitura (RCU) suficientes provisionadas para processar a workload.

Opção de configuração	Exemplos
Escalar gravações (como a fragmentação da chave da partição ou a introdução de uma fila)	• No caso de bancos de dados relacionais, é possível aumentar o tamanho da instância para acomodar uma workload maior, ou aumentar as IOPs provisionadas para permitir um throughput mais alto no armazenamento subjacente. • Também é possível introduzir uma fila na frente do banco de dados, em vez de gravar diretamente nele. Esse padrão permite desacoplar a ingestão do banco de dados e controlar a taxa de fluxo para que o banco de dados não fique sobrecarregado. • Usar solicitações de gravação em lote em vez de criar muitas transações de curta duração pode ajudar a melhorar o throughput em bancos de dados relacionais de alto volume de gravação. • Os bancos de dados com tecnologia sem servidor, como o DynamoDB, podem ajustar a escala do throughput de gravação automaticamente ou ajustar as unidades da capacidade de gravação (WCU) provisionadas, dependendo do modo da capacidade. • Você ainda pode ter problemas com partições ativas ao atingir os limites de throughput de determinada chave de partição. Isso pode ser mitigado com a escolha de uma chave de partição mais uniformemente distribuída ou por meio da fragmentação da gravação da chave de partição.

Opção de configuração	Exemplos
Políticas para gerenciar o ciclo de vida dos seus conjuntos de dados	• O Amazon S3 Lifecycle pode ser usado para gerenciar seus objetos durante todo o ciclo de vida de cada um. Se os padrões de acesso forem desconhecidos, variáveis ou imprevisíveis, você poderá usar o Amazon S3 Intelligent-Tiering para monitorar os padrões de acesso e mover automaticamente os objetos que não foram acessados para níveis de acesso de baixo custo. Você pode aproveitar as métricas da Lente de Armazenamento do Amazon S3 para identificar oportunidades de otimizaçã o e lacunas no gerenciamento do ciclo de vida. • O gerenciamento do ciclo de vida do Amazon EFS gerencia automaticamente o armazenamento de arquivos para seus sistemas de arquivos.
Gerenciamento e agrupamento de conexões	• O Amazon RDS Proxy pode ser usado com o Amazon RDS e o Aurora para gerenciar as conexões com o banco de dados. • Os bancos de dados com tecnologia sem servidor, como o DynamoDB, não têm conexões associadas a eles, mas considere a capacidade provisionada e as políticas de ajuste de escala automático para lidar com picos na carga.

- Realize experimentos e testes comparativos em um ambiente de não produção para identificar qual opção de configuração pode atender aos requisitos da workload.
- Depois de experimentar, planeje a migração e valide as métricas de performance.

- Use ferramentas de monitoramento da AWS (como o Amazon CloudWatch) e de otimização (como a Lente de Armazenamento do Amazon S3) para otimizar constantemente o datastore usando um padrão de uso real.

Recursos

Documentos relacionados:

- Armazenamento na nuvem com a AWS
- Tipos de volume do Amazon EBS
- Armazenamento do Amazon EC2
- Amazon EFS: performance do Amazon EFS
- Performance do Amazon FSx para Lustre
- Performance do Amazon FSx para Windows File Server
- Amazon S3 Glacier: documentação do S3 Glacier
- Amazon S3: considerações sobre performance e taxa de solicitações
- Características de E/S do Amazon EBS
- Bancos de dados na nuvem com a AWS
- Cache de banco de dados da AWS
- DynamoDB Accelerator
- Práticas recomendadas do Amazon Aurora
- Performance do Amazon Redshift
- As 10 melhores dicas de performance para Amazon Athena
- Práticas recomendadas do Amazon Redshift Spectrum
- Práticas recomendadas do Amazon DynamoDB

Vídeos relacionados:

- AWS re:Invent 2023: Melhorar a eficiência do Amazon Elastic Block Store e ser mais econômico
- AWS re:Invent 2023: Otimizar o preço e a performance do armazenamento com o Amazon Simple Storage Service
- AWS re:Invent 2023: Criar e otimizar data lakes no Amazon Simple Storage Service

- AWS re:Invent 2023: Novidades do armazenamento de arquivos da AWS
- AWS re:Invent 2023: Mergulho profundo no Amazon DynamoDB

Exemplos relacionados:

- Workshop de bancos de dados com propósito específico na AWS
- Bancos de dados para desenvolvedores
- Dia de imersão na arquitetura de dados moderna na AWS
- Ajuste de escala automático do Amazon EBS
- Exemplos do Amazon S3
- Exemplos do Amazon DynamoDB
- Exemplos de migração de banco de dados da AWS
- Workshop de modernização de bancos de dados
- Trabalhar com parâmetros no Amazon RDS para Postgress DB

PERF03-BP03 Coletar e registrar métricas de performance do datastore

Acompanhe e registre métricas de performance relevantes para o datastore a fim de entender a performance das suas soluções de gerenciamento de dados. Essas métricas podem ajudar você a otimizar o datastore, verificar se os requisitos da workload foram atendidos e fornecer uma visão geral clara da performance da workload.

Práticas comuns que devem ser evitadas:

- Você só usa a pesquisa manual de arquivos de log para métricas.
- Você só publica métricas em ferramentas internas usadas pela equipe e não tem uma imagem abrangente da workload.
- Você só usa as métricas comuns registradas pelo software de monitoramento selecionado.
- Você só revisa as métricas quando há um problema.
- Você só monitora as métricas no sistema e não captura as métricas de uso e acesso aos dados.

Benefícios de implementar esta prática recomendada: o estabelecimento de uma linha de base de performance ajuda a compreender o comportamento normal e os requisitos das workloads. Padrões

anormais podem ser identificados e depurados mais rapidamente, melhorando a performance e a confiabilidade do datastore.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Para monitorar a performance dos datastores, é necessário registrar várias métricas de performance ao longo de um período. Isso permite detectar anomalias e avaliar a performance em relação às métricas de negócios para verificar se as necessidades da workload estão sendo atendidas.

As métricas devem incluir as do sistema subjacente que oferece suporte ao datastore e as do banco de dados. As métricas do sistema subjacente podem incluir métricas de utilização de CPU, memória, armazenamento em disco disponível, E/S de disco, taxa de acertos do cache e entrada e saída da rede, enquanto as métricas do datastore devem incluir transações por segundo, tempos de resposta, uso de índice, bloqueios de tabela, tempos limite de consultas e número de conexões abertas. Esses dados são essenciais para compreender a performance da workload e como a solução de gerenciamento de dados é usada. Use essas métricas como parte de uma abordagem orientada por dados para ajustar e otimizar os recursos da workload.

Use ferramentas, bibliotecas e sistemas que registram as medidas de performance relacionadas ao banco de dados.

Etapas de implementação

- Identifique as principais métricas de performance que o datastore deve monitorar.
  - Métricas e dimensões do Amazon S3
  - Monitorar métricas em uma instância do Amazon RDS
  - Monitorar a workload de banco de dados com o Performance Insights no Amazon RDS
  - Visão geral do monitoramento avançado
  - Métricas e dimensões do DynamoDB
  - Monitorar o DynamoDB Accelerator
  - Monitorar o Amazon MemoryDB com o Amazon CloudWatch
  - Que métricas devo monitorar?
  - Monitorar a performance do cluster do Amazon Redshift
  - Métricas e dimensões do Timestream
  - Métricas do Amazon CloudWatch para o Amazon Aurora

  - Registrar em log e monitorar o Amazon Keyspaces (para Apache Cassandra)
  - Monitorar recursos do Amazon Neptune
- Use uma solução aprovada de registro em log e monitoramento para coletar essas métricas. O Amazon CloudWatch pode coletar métricas nos recursos em sua arquitetura. Você também pode coletar e publicar métricas personalizadas para descobrir métricas de negócio ou derivadas. Use o CloudWatch ou soluções de terceiros para definir alarmes que indicam quando os limites são violados.
- Confira se o monitoramento do datastore pode se beneficiar de uma solução de machine learning que detecta anomalias de performance.
  - O Amazon DevOps Guru para Amazon RDS fornece visibilidade dos problemas de performance e faz recomendações de ações corretivas.
- Configure a retenção de dados em sua solução de monitoramento e de log para corresponder às suas metas operacionais e de segurança.
  - Retenção de dados padrão para métricas do CloudWatch
  - Retenção de dados padrão para CloudWatch Logs

Recursos

Documentos relacionados:

- Cache de banco de dados da AWS
- As 10 melhores dicas de performance para Amazon Athena
- Práticas recomendadas do Amazon Aurora
- DynamoDB Accelerator
- Práticas recomendadas do Amazon DynamoDB
- Práticas recomendadas do Amazon Redshift Spectrum
- Performance do Amazon Redshift
- Bancos de dados na nuvem com a AWS
- Insights de Performance do Amazon RDS

Vídeos relacionados:

- AWS re:Invent 2022: Monitoramento de performance com o Amazon RDS e o Aurora, com destaque para Autodesk

- Monitoramento e ajuste de performance de banco de dados com o Amazon DevOps Guru para Amazon RDS
- AWS re:Invent 2023: Novidades do armazenamento de arquivos na AWS
- AWS re:Invent 2023: Mergulho profundo no Amazon DynamoDB
- AWS re:Invent 2023: Criar e otimizar data lakes no Amazon S3
- AWS re:Invent 2023: Novidades do armazenamento de arquivos na AWS
- AWS re:Invent 2023: Mergulho profundo no Amazon DynamoDB
- Práticas recomendadas para monitorar workloads do Redis no Amazon ElastiCache

Exemplos relacionados:

- Framework de coleta de métricas de ingestão de conjunto de dados na AWS
- Workshop de monitoramento do Amazon RDS
- Workshop de bancos de dados com propósito específico na AWS

PERF03-BP04 Implementar estratégias para melhorar a performance da consulta no datastore

Implemente estratégias para otimizar os dados e melhorar a consulta de dados a fim de permitir mais escalabilidade e performance eficiente para a workload.

Práticas comuns que devem ser evitadas:

- Você não particiona dados no datastore.
- Você armazena dados em apenas um formato de arquivo no datastore.
- Você não usa índices no datastore.

Benefícios de implementar esta prática recomendada: a otimização da performance dos dados e das consultas ocasiona mais eficiência, menor custo e melhor experiência do usuário.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A otimização de dados e o ajuste de consultas são aspectos essenciais da eficiência de performance em um datastore, pois afetam não só a performance, mas também a capacidade de resposta de toda

a workload na nuvem. Consultas não otimizadas podem ocasionar maior uso de recursos e gargalos, o que reduz a eficiência geral de um datastore.

A otimização de dados inclui várias técnicas para garantir o armazenamento e o acesso eficientes aos dados. Esse processo também ajuda a melhorar a performance da consulta em um datastore. As principais estratégias incluem particionamento, compactação e desnormalização de dados, o que ajuda a otimizá-los para armazenamento e acesso.

Etapas de implementação

- Entenda e analise as consultas de dados críticos que são realizadas no datastore.
- Identifique as consultas com execução lenta no datastore e use planos de consulta para entender o estado atual delas.
  - Analisar o plano de consulta no Amazon Redshift
  - Usar EXPLAIN e EXPLAIN ANALYZE no Athena
- Implemente estratégias para melhorar a performance da consulta. Algumas das principais estratégias incluem:
  - Usar um formato de arquivo colunar (como Parquet ou ORC).
  - Compactar os dados no datastore para reduzir o espaço de armazenamento e as operações de E/S.
  - Particionar os dados para dividi-los em partes menores e reduzir o tempo de verificação dos dados.
    - Particionamento de dados no Athena
    - Partições e distribuição de dados
  - Indexação de dados nas colunas comuns na consulta.
  - Use visões materializadas para consultas frequentes.
    - Entender as visões materializadas
    - Criar visões materializadas no Amazon Redshift
  - Escolha a operação de junção correta para consulta. Ao unir duas tabelas, especifique a tabela maior no lado esquerdo da junção e a tabela menor no lado direito.
  - Solução de cache distribuído para melhorar a latência e reduzir o número de operações de E/S do banco de dados.
  - Manutenção regular, como aspiração, reindexação e estatísticas de execução.
- Experimente e teste estratégias em um ambiente de não produção.

Recursos

Documentos relacionados:

- Práticas recomendadas do Amazon Aurora
- Performance do Amazon Redshift
- As 10 melhores dicas de performance para Amazon Athena
- Cache de banco de dados da AWS
- Práticas recomendadas para implementar o Amazon ElastiCache
- Particionamento de dados no Athena

Vídeos relacionados:

- AWS re:Invent 2023: Práticas recomendadas de otimização de custos de armazenamento na AWS
- AWS re:Invent 2022: Monitoramento de performance com o Amazon RDS e o Aurora, com destaque para Autodesk
- Otimizar consultas do Amazon Athena com novas ferramentas de análise de consultas

Exemplos relacionados:

- Amazon S3 Select: consultar dados sem servidores ou bancos de dados
- Workshop de bancos de dados com propósito específico na AWS

PERF03-BP05 Implementar padrões de acesso a dados que utilizam cache

Implemente padrões de acesso que possam se beneficiar do armazenamento em cache de dados para recuperação rápida de dados acessados com frequência.

Práticas comuns que devem ser evitadas:

- Armazenar em cache dados que mudam com frequência.
- Depender dos dados em cache como se estivessem armazenados de forma durável e sempre disponíveis.
- Não levar em conta a consistência dos seus dados em cache.
- Não monitorar a eficiência da sua implementação de cache.

Benefícios de implementar esta prática recomendada: armazenar dados em um cache pode melhorar a latência de leitura, o throughput de leitura, a experiência do usuário e a eficiência geral, além de reduzir custos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Um cache é um componente de software ou hardware destinado a armazenar dados para que futuras solicitações dos mesmos dados possam ser atendidas com maior rapidez e eficiência. Os dados armazenados em um cache podem ser reconstruídos se perdidos, repetindo um cálculo anterior ou obtendo-os de outro datastore.

O armazenamento de dados em cache pode ser uma das estratégias mais eficazes para melhorar a performance geral da aplicação e reduzir a carga sobre as fontes de dados primárias subjacentes. Os dados podem ser armazenados em vários níveis na aplicação, como dentro da aplicação e fazendo chamadas remotas, o que é conhecido como cache do lado do cliente, ou usando um serviço secundário rápido para armazenar os dados, conhecido como cache remoto.

Armazenamento em cache no lado do cliente

Com o armazenamento em cache no lado do cliente, cada cliente (uma aplicação ou serviço que consulta o datastore de backend) pode armazenar os resultados de suas consultas exclusivas localmente por um período especificado. Isso pode reduzir o número de solicitações na rede para um datastore ao verificar primeiro o cache do cliente local. Se os resultados não estiverem presentes, a aplicação poderá então consultar o datastore e armazenar esses resultados localmente. Esse padrão permite que cada cliente armazene dados no local mais próximo (o próprio cliente), resultando na menor latência possível. Os clientes também podem continuar a atender algumas consultas quando o datastore de backend não está disponível, aumentando a disponibilidade geral do sistema.

Uma desvantagem dessa abordagem é que, quando vários clientes estão envolvidos, eles podem armazenar os mesmos dados em cache localmente. Isso resulta no uso de armazenamento duplicado e na inconsistência de dados entre esses clientes. Um cliente pode armazenar em cache os resultados de uma consulta e, um minuto depois, outro cliente pode executar a mesma consulta e obter um resultado diferente.

Armazenamento em cache remoto

Para resolver o problema de dados duplicados entre clientes, um serviço externo rápido ou cache remoto pode ser usado para armazenar os dados consultados. Em vez de verificar um

datastore local, cada cliente verificará o cache remoto antes de consultar o datastore de backend. Essa estratégia permite respostas mais consistentes entre clientes, melhor eficiência nos dados armazenados e um volume maior de dados em cache, pois o espaço de armazenamento é dimensionado independentemente dos clientes.

A desvantagem de um cache remoto é que o sistema geral pode ter uma latência maior, pois é necessário um salto de rede adicional para verificar o cache remoto. O cache do lado do cliente pode ser usado junto com o armazenamento em cache remoto para o armazenamento em vários níveis para melhorar a latência.

Etapas de implementação

- Identifique bancos de dados, APIs e serviços de rede que poderiam se beneficiar do armazenamento em cache. Serviços que têm workloads de leitura pesadas, uma alta taxa de leitura e gravação ou que são caros para escalar são candidatos ao armazenamento em cache.
  - Armazenamento em cache de banco de dados
  - Habilitar o armazenamento em cache de APIs para melhorar a capacidade de resposta
- Identifique o tipo apropriado de estratégia de armazenamento em cache que melhor se adapte ao seu padrão de acesso.
  - Estratégias de armazenamento em cache
  - Soluções de armazenamento em cache da AWS
- Siga as práticas recomendadas de armazenamento em cache para seu datastore.
- Configure uma estratégia de invalidação de cache, como um time-to-live (TTL), para todos os dados que equilibre a atualização dos dados e reduza a pressão sobre o datastore de backend.
- Habilite recursos como novas tentativas automáticas de conexão, recuo exponencial, tempos limite no lado do cliente e pool de conexões no cliente, se disponíveis, pois eles podem melhorar a performance e a confiabilidade.
  - Práticas recomendadas: clientes Redis e Amazon ElastiCache (Redis OSS)
- Monitore a taxa de acertos de cache com uma meta de 80% ou mais. Valores mais baixos podem indicar tamanho insuficiente do cache ou um padrão de acesso que não se beneficia do armazenamento em cache.
  - Que métricas devo monitorar?
  - Práticas recomendadas para monitorar workloads do Redis no Amazon ElastiCache
  - Monitorar as práticas recomendadas com o Amazon ElastiCache (Redis OSS) usando o Amazon CloudWatch

- Implemente a replicação de dados para descarregar as leituras em várias instâncias e melhorar a performance e a disponibilidade da leitura de dados.

Recursos

Documentos relacionados:

- Usar a Lente do Well-Architected para o Amazon ElastiCache
- Monitorar as práticas recomendadas com o Amazon ElastiCache (Redis OSS) usando o Amazon CloudWatch
- Que métricas devo monitorar?
- Whitepaper Performance em grande escala com o Amazon ElastiCache
- Desafios e estratégias de armazenamento em cache

Vídeos relacionados:

- Plano de aprendizado do Amazon ElastiCache
- Design para o sucesso com as práticas recomendadas do Amazon ElastiCache
- AWS re:Invent 2020: Design para o sucesso com as práticas recomendadas do Amazon ElastiCache
- AWS re:Invent 2023 [LANÇAMENTO]: Introdução ao Amazon ElastiCache sem servidor
- AWS re:Invent 2022: Cinco excelentes formas de reimaginar sua camada de dados com o Redis
- AWS re:Invent 2021: Mergulho profundo no Amazon ElastiCache (Redis OSS)

Exemplos relacionados:

- Como aumentar a performance de bancos de dados MySQL com o Amazon ElastiCache (Redis OSS)

Rede e entrega de conteúdo

Perguntas

- PERF 4. Como selecionar e configurar os recursos de rede em sua workload?

PERF 4. Como selecionar e configurar os recursos de rede em sua workload?

A solução de rede ideal para uma workload varia com base em latência, requisitos de throughput, jitter e largura de banda. Restrições físicas, como recursos de usuário ou on-premises, determinam as opções de localização. Essas restrições podem ser compensadas com locais de borda ou posicionamento de recursos.

Práticas recomendadas

- PERF04-BP01 Compreender como as redes afetam a performance
- PERF04-BP02 Avaliar os recursos de rede disponíveis
- PERF04-BP03 Escolher a conectividade dedicada ou VPN apropriada para a workload
- PERF04-BP04 Usar o balanceamento de carga para distribuir o tráfego em vários recursos
- PERF04-BP05 Escolher protocolos de rede para melhorar a performance
- PERF04-BP06 Escolher o local da workload com base nos requisitos de rede
- PERF04-BP07 Otimizar a configuração da rede com base em métricas

PERF04-BP01 Compreender como as redes afetam a performance

Analise e entenda como as decisões relacionadas à rede afetam sua workload para fornecer performance eficiente e uma melhor experiência do usuário.

Práticas comuns que devem ser evitadas:

- Todo o tráfego flui por meio dos data centers existentes.
- Você direciona todo o tráfego por meio de firewalls centrais em vez de usar ferramentas de segurança de rede nativas da nuvem.
- Você provisiona conexões do AWS Direct Connect sem entender os requisitos reais de uso.
- Você não considera as características da workload e a sobrecarga da criptografia ao definir suas soluções de redes.
- Você usa conceitos e estratégias de on-premises para soluções de redes na nuvem.

Benefícios de implementar esta prática recomendada: a compreensão de como as redes afetam a performance da workload ajuda a identificar gargalos potenciais, a melhorar a experiência dos usuários, a aumentar a confiabilidade e a reduzir a manutenção operacional à medida que a workload muda.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

A rede é responsável pela conectividade entre os componentes da aplicação, os serviços de nuvem, as redes de borda e os dados on-premises. Portanto, ela pode afetar significativamente a performance da workload. Além da performance da workload, a experiência dos usuários também é afetada por latência de rede, largura de banda, protocolos, localização, congestão de rede, jitter, throughput e regras de roteamento.

Ter uma lista documentada dos requisitos de rede da workload, incluindo latência, tamanho de pacotes, regras de roteamento, protocolos e padrões de tráfego compatíveis. Analise as soluções de redes disponíveis e identifique os serviços que atendem às características de rede da sua workload. É possível recriar as redes baseadas na nuvem rapidamente. Portanto, é necessário evoluir sua arquitetura de rede ao longo do tempo para melhorar a eficiência da performance.

Etapas de implementação:

- Defina e documente os requisitos de performance da rede, incluindo métricas como latência da rede, largura de banda, protocolos, locais, padrões de tráfego (picos e frequência), throughput, criptografia, inspeção e regras de roteamento.
- Saiba mais sobre os principais serviços de rede da AWS, como VPCs, AWS Direct Connect, Elastic Load Balancing (ELB) e Amazon Route 53.
- Capture as seguintes características principais da rede:

Características	Ferramentas e métricas
Características básicas de rede	• Logs de fluxo da VPC • Logs de fluxo do AWS Transit Gateway • Métricas de AWS Transit Gateway do • Métricas de AWS PrivateLink do
Características da rede da aplicação	• Adaptador de malha elástica • Métricas de AWS App Mesh do • Métricas do Amazon API Gateway
Características da rede da borda	• Métricas do Amazon CloudFront • Métricas do Amazon Route 53

Características	Ferramentas e métricas
	• Métricas de AWS Global Accelerator do
Características da rede híbrida	• Métricas de AWS Direct Connect do • Métricas de AWS Site-to-Site VPN do • Métricas de AWS Client VPN do • Métricas da WAN da Nuvem AWS
Características da rede de segurança	• Métricas do AWS Shield, AWS WAF e AWS Network Firewall
Características de rastreamento	• AWS X-Ray • VPC Reachability Analyzer • Analisador de Acesso à Rede • Amazon Inspector • Amazon CloudWatch RUM

- Teste comparativo e de performance da rede:
  - Faça o teste comparativo do throughput da rede, pois alguns fatores podem afetar a performance da rede do Amazon EC2 quando as instâncias estão na mesma VPC. Meça a largura de banda da rede entre as instâncias Linux do Amazon EC2 na mesma VPC.
  - Faça testes de carga para experimentar soluções e opções de redes.

Recursos

Documentos relacionados:

- Application Load Balancer
- Rede avançada do EC2 no Linux
- Rede avançada do EC2 no Windows
- Grupos de posicionamento do EC2
- Como habilitar a rede avançada com o Adaptador de Rede Elástica (ENA) em instâncias Linux
- Network Load Balancer
- Produtos de redes com a AWS

- Gateway de trânsito
- Passar para o roteamento baseado em latência no Amazon Route 53
- VPC Endpoints

Vídeos relacionados:

- AWS re:Invent 2023: Fundamentos de rede na AWS
- AWS re:Invent 2023: O que rede pode fazer por sua aplicação?
- AWS re:Invent 2023: Designs e novos recursos de VPCs avançadas
- AWS re:Invent 2023: Um guia do desenvolvedor para redes na nuvem
- AWS re:Invent 2019: Conectividade à AWS e arquiteturas de rede na AWS híbridas
- AWS re:Invent 2019: Otimizar a performance da rede para instâncias do Amazon EC2
- AWS Summit Online: Melhorar a performance de rede global para aplicações
- AWS re:Invent 2020: Dicas e práticas recomendadas de rede com o Well-Architected Framework
- AWS re:Invent 2020: Práticas recomendadas de rede na AWS em migrações de grande escala

Exemplos relacionados:

- AWS Transit Gateway e soluções de segurança escaláveis
- Workshops de redes da AWS
- Workshop prático de firewall de rede
- Observar e diagnosticar sua rede na AWS
- Como encontrar e lidar com configurações de rede incorretas na AWS

PERF04-BP02 Avaliar os recursos de rede disponíveis

Avalie recursos de rede na nuvem que possam melhorar a performance. Meça o impacto desses recursos por meio de testes, métricas e análises. Por exemplo, utilize os recursos de rede disponíveis para reduzir a latência, a distância ou o jitter da rede.

Práticas comuns que devem ser evitadas:

- Você permanece em uma região, pois é onde a sede da sua empresa ou organização está fisicamente localizada.

- Você usa firewalls em vez de grupos de segurança para filtrar o tráfego.
- Você quebra o TLS para inspeção de tráfego em vez de confiar em grupos de segurança, políticas de endpoint e outras funcionalidades nativas da nuvem.
- Você só usa segmentação baseada em sub-rede em vez de grupos de segurança.

Benefícios de implementar esta prática recomendada: avaliar todos os recursos e opções de serviços pode aumentar a performance da workload, reduzir o custo da infraestrutura, diminuir o esforço necessário para manter sua workload e aumentar sua postura geral de segurança. É possível utilizar o backbone global da AWS para garantir a experiência ideal de rede para os clientes.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

A AWS oferece serviços como o AWS Global Accelerator e o Amazon CloudFront que podem ajudar a melhorar a performance da rede, enquanto a maioria dos serviços da AWS tem recursos de produto (como o recurso Amazon S3 Transfer Acceleration) para otimizar o tráfego da rede.

Analise quais opções de configuração de rede estão disponíveis e como elas poderiam afetar a workload. A otimização da performance depende da compreensão de como essas opções interagem com sua arquitetura e do impacto que elas terão na performance medida e na experiência do usuário.

Etapas de implementação

- Crie uma lista de componentes da workload.
  - Considere usar a WAN da Nuvem AWS para criar, gerenciar e monitorar a rede da sua organização ao criar uma rede global unificada.
  - Monitore suas redes globais e centrais com as métricas do Amazon CloudWatch Logs. Utilize o Amazon CloudWatch RUM, que fornece insights para ajudar a identificar, entender e aprimorar a experiência digital dos usuários.
  - Visualize a latência agregada da rede entre as Regiões da AWS e as zonas de disponibilidade, bem como dentro de cada zona de disponibilidade, usando o AWS Network Manager para obter informações sobre como a performance da aplicação se relaciona com a performance da rede da AWS subjacente.
  - Use uma ferramenta existente de banco de dados de gerenciamento de configuração (CMDB) ou um serviço como o AWS Config para criar um inventário da sua workload e de como ela está configurada.

- Se for uma workload existente, identifique e documente a referência para suas métricas de performance, focando os gargalos e nas áreas de melhoria. As métricas de rede associadas à performance irão variar de acordo com a workload com base nos requisitos comerciais e nas características da workload. Como ponto de partida, a análise dessas métricas pode ser importante para sua workload: largura de banda, latência, perda de pacotes, jitter e retransmissões.
- Se essa for uma nova workload, realize testes de carga para identificar gargalos de performance.
- Para os gargalos de performance que identificar, revise as opções de configuração para suas soluções a fim de identificar oportunidades de melhoria da performance. Confira os seguintes recursos de rede e opções importantes:

Oportunidade de melhoria	Solução
Caminho ou rotas de rede	Use o Analisador de Acesso à Rede para identificar caminhos ou rotas.
Protocolos de rede	Consulte PERF04-BP05 Escolher protocolos de rede para melhorar a performance
Topologia de rede	Avalie seus compromissos operacionais e de performance entre o emparelhamento de VPC e o AWS Transit Gateway ao conectar várias contas. O AWS Transit Gateway simplific a a forma como você interconecta todas as suas VPCs, as quais podem se estender por milhares de Contas da AWS e até redes on-premises. Compartilhe seu AWS Transit Gateway entre várias contas usando o AWS Resource Access Manager. Consulte PERF04-BP03 Escolher a conectividade dedicada ou VPN apropriada para a workload
Serviços de rede	O AWS Global Accelerator é um serviço de rede que melhora a performance do tráfego

Oportunidade de melhoria	Solução
	dos usuários em até 60% usando a infraestr utura de rede global da AWS. O Amazon CloudFront pode melhorar a performance da entrega e da latência de conteúdo da workload globalmente. Use o Lambda@Edge para executar funções que personalizam o conteúdo que o CloudFront entrega mais perto dos usuários, reduzir a latência e melhorar a performance. O Amazon Route 53 oferece opções de roteamento baseado em latência, roteamento por geolocalização, roteamento por geoproxim idade e roteamento baseado em IP para ajudar você a melhorar a performance da sua workload para um público global. Identifiq ue qual opção de roteamento otimizaria a performance da workload analisando o respectivo tráfego e a localização do usuário quando ela for distribuída globalmente.

Oportunidade de melhoria	Solução
Recursos do atributo de armazenamento	O Amazon S3 Transfer Acceleration é um recurso que permite aos usuários externos beneficiarem-se de otimizações de rede do CloudFront para o upload de dados para o Amazon S3. Isso melhora a capacidade de transferir grandes quantidades de dados com origem em locais remotos que não têm conectividade dedicada com a Nuvem AWS. O Amazon S3 Multi-Region Access Points replica conteúdo para várias regiões e simplifica a workload ao fornecer um ponto de acesso. Quando um ponto de acesso multirregiões é usado, você pode solicitar ou gravar dados no Amazon S3 com o serviço identificando o bucket de menor latência.

Oportunidade de melhoria	Solução
Atributos dos recursos computacionais	As interfaces de rede elásticas (ENI) usadas por instâncias do Amazon EC2, contêineres e funções do Lambda são limitadas por fluxo. Revise seus grupos de posicionamento para otimizar seu throughput de rede do EC2. Para evitar gargalos em uma abordagem por fluxo, projete sua aplicação para usar vários fluxos. Para monitorar e obter visibilidade de suas métricas de rede relacionadas à computação, use o CloudWatch Metrics e a ethtool. O comando `ethtool` está incluído no driver da ENA e expõe métricas adicionais relacionadas à rede que podem ser publicadas como uma métrica personalizada no CloudWatch. Os adaptadores de rede elástica (ENA) da Amazon aumentam ainda mais a otimização oferecendo throughput melhor para suas instâncias em um grupo de posicionamento de cluster. O Elastic Fabric Adapter (EFA) é uma interface de rede para instâncias do Amazon EC2 que permite executar workloads que exigem altos níveis de comunicação entre nós em grande escala na AWS. As instâncias otimizadas para Amazon EBS usam uma pilha de configuração otimizada e fornecem capacidade adicional e dedicada para aumentar a E/S do Amazon EBS.

Recursos

Documentos relacionados:

- Application Load Balancer
- Rede avançada do EC2 no Linux
- Rede avançada do EC2 no Windows
- Grupos de posicionamento do EC2
- Como habilitar a rede avançada com o Adaptador de Rede Elástica (ENA) em instâncias Linux
- Network Load Balancer
- Produtos de redes com a AWS
- Passar para o encaminhamento por latência no Amazon Route 53
- VPC Endpoints
- VPC Flow Logs

Vídeos relacionados:

- AWS re:Invent 2023: Pronto para o que vem a seguir? Desenvolver redes para crescimento e flexibilidade
- AWS re:Invent 2023: Designs e novos recursos de VPCs avançadas
- AWS re:Invent 2023: Um guia do desenvolvedor para redes na nuvem
- AWS re:Invent 2022: Mergulho profundo na infraestrutura de rede da AWS
- AWS re:Invent 2019: Conectividade à AWS e arquiteturas de rede na AWS híbridas
- AWS re:Invent 2018: Otimizar a performance da rede para instâncias do Amazon EC2
- AWS Global Accelerator

Exemplos relacionados:

- AWS Transit Gateway e soluções de segurança escaláveis
- Workshops de redes da AWS
- Observar e diagnosticar sua rede
- Como encontrar e lidar com configurações de rede incorretas na AWS

PERF04-BP03 Escolher a conectividade dedicada ou VPN apropriada para a workload

Quando a conectividade híbrida é necessária para conectar recursos on-premises e na nuvem, provisione a largura de banda adequada para atender aos requisitos de performance. Estime os

requisitos de largura de banda e de latência para a workload híbrida. Esses números determinarão seus requisitos de dimensionamento.

Práticas comuns que devem ser evitadas:

- Avaliar somente as soluções de VPN para seus requisitos de criptografia de rede.
- Não avaliar as opções de backup ou de conectividade redundante.
- Não identificar todos os requisitos da workload (necessidades de criptografia, protocolo, largura de banda e tráfego).

Benefícios de implementar esta prática recomendada: selecionar e configurar soluções de conectividade apropriadas aumentará a confiabilidade da workload e maximizará a performance. A identificação dos requisitos da workload, o planejamento antecipado e a avaliação das soluções híbridas podem minimizar alterações dispendiosas da rede física e despesas operacionais, e aumentará seu tempo para geração de valor.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Desenvolva uma arquitetura de rede híbrida com base em seus requisitos de largura de banda. O AWS Direct Connect permite a você conectar sua rede on-premises de forma privada com a AWS. Isso é conveniente quando você precisa de largura de banda alta e baixa latência com performance consistente. Uma conexão VPN estabelece uma conexão segura via Internet. Ela é usada quando apenas uma conexão temporária é necessária, quando o custo é um fator ou como uma contingência enquanto se espera que uma conectividade de rede física resiliente seja estabelecida durante o uso do AWS Direct Connect.

Se seus requisitos de largura de banda forem altos, considere vários serviços do AWS Direct Connect ou de VPN. O tráfego pode ser balanceado entre os serviços, embora o balanceamento de carga entre o AWS Direct Connect e a VPN não seja recomendado devido às diferenças de latência e largura de banda.

Etapas de implementação

- Calcule os requisitos de largura de banda e latência de suas aplicações existentes.
  - Para workloads existentes que estão sendo migradas para a AWS, utilize os dados de seus sistemas de monitoramento de rede internos.

  - Para workloads novas ou existentes para as quais não há dados de monitoramento, consulte os proprietários do produto para determinar métricas de performance adequadas e fornecer uma experiência do usuário satisfatória.
- Escolha uma conexão dedicada ou VPN como sua opção de conectividade. Com base em todos os requisitos da workload (necessidades de criptografia, largura de banda e tráfego), é possível escolher o AWS Direct Connect ou o AWS VPN (ou ambos). O diagrama a seguir ajudará você a escolher o tipo de conexão apropriada.
  - O AWS Direct Connect fornece conectividade dedicada ao ambiente da AWS, de 50 Mbps a 100 Gbps, usando conexões dedicadas ou conexões hospedadas. Isso permite que você tenha latência gerenciada e controlada, além de largura de banda provisionada para que a workload possa se conectar de forma eficiente com outros ambientes. Com os parceiros do AWS Direct Connect, é possível ter conectividade completa para vários ambientes, fornecendo uma rede estendida com performance consistente. A AWS oferece ajuste de escala da largura de banda da conexão direta usando o grupo de agregação nativo (LAG) de 100 Gbps ou o BGP equal-cost multipath (ECMP).
  - A AWS Site-to-Site VPN fornece um serviço de VPN gerenciada compatível com o protocolo de segurança da internet (IPsec). Quando uma conexão VPN é criada, cada conexão VPN inclui dois túneis para alta disponibilidade.
- Siga a documentação da AWS para escolher uma opção de conectividade apropriada:
  - Se você decidir usar o AWS Direct Connect, selecione a largura de banda apropriada para sua conectividade.
  - Se você estiver usando um AWS Site-to-Site VPN em vários locais para se conectar a uma Região da AWS, use uma conexão do Site-to-Site VPN acelerada para ter a oportunidade de melhorar a performance da rede.
  - Se o design da sua rede consistir em uma conexão VPN IPsec via AWS Direct Connect, considere usar uma VPN IP privada para melhorar a segurança e obter segmentação. AWS Uma VPN IP privada site a site é implantada por meio da interface virtual de trânsito (VIF).
  - O AWS Direct Connect SiteLink permite criar conexões redundantes e de baixa latência entre seus datacenters em todo o mundo, enviando dados pelo caminho mais rápido entre locais do AWS Direct Connect, ignorando as Regiões da AWS.
- Valide sua configuração de conectividade antes de implantá-la na produção. Execute testes de segurança e performance para garantir que ela atenda aos requisitos de largura de banda, confiabilidade, latência e conformidade.
- Monitore regularmente a performance e o uso da conectividade e otimize, se necessário.

Fluxograma de performance determinística

Recursos

Documentos relacionados:

- Produtos de rede com a AWS
- AWS Transit Gateway
- Endpoints da VPC
- Construção de uma infraestrutura de rede da AWS Multi-VPC escalável e segura
- VPN do cliente

Vídeos relacionados:

- AWS re:Invent 2023: Criar conectividade de rede híbrida com a AWS
- AWS re:Invent 2023: Proteger a conectividade remota com a AWS
- AWS re:Invent 2022: Otimizar a performance com o Amazon CloudFront
- AWS re:Invent 2019: Conectividade à AWS e arquiteturas de rede na AWS híbridas
- AWS re:Invent 2020: Conectar ao AWS Transit Gateway

Exemplos relacionados:

- AWS Transit Gateway e soluções de segurança escaláveis
- Workshops de redes da AWS

PERF04-BP04 Usar o balanceamento de carga para distribuir o tráfego em vários recursos

Distribua o tráfego entre vários recursos e serviços para permitir que sua workload aproveite a elasticidade oferecida pela nuvem. Também é possível usar o balanceamento de carga para descarregar a terminação de criptografia a fim de melhorar a performance, a confiabilidade e gerenciar e rotear o tráfego de maneira eficaz.

Práticas comuns que devem ser evitadas:

- Você não considera os requisitos da workload ao escolher o tipo de balanceador de carga.

- Você não utiliza os recursos do balanceador de carga para otimização da performance.
- A workload é exposta diretamente à internet sem um balanceador de carga.
- Você roteia todo o tráfego da Internet por meio de balanceadores de carga existentes.
- Você usa o balanceamento de carga TCP genérico e faz com que cada nó de computação lide com a criptografia SSL.

Benefícios de implementar esta prática recomendada: um balanceador de carga lida com a carga variável do tráfego da sua aplicação em uma única zona de disponibilidade ou em várias zonas de disponibilidade e permite alta disponibilidade, ajuste de escala automático e melhor utilização da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Os balanceadores de carga atuam como o ponto de entrada para sua workload, ponto a partir do qual distribuem o tráfego para seus destinos de backend, como instâncias de computação ou contêineres, para melhorar a utilização.

Escolher o tipo certo de balanceador de carga é a primeira etapa para otimizar sua arquitetura. Comece listando as características da workload, como protocolo (como TCP, HTTP, TLS ou WebSockets), o tipo de destino (como instâncias, contêineres ou tecnologia sem servidor), requisitos da aplicação (como conexões de execução longa, autenticação de usuários ou adesão) e posicionamento (como região, zona local, Outpost ou isolamento por zona).

A AWS fornece vários modelos para que suas aplicações usem o balanceamento de carga. O Application Load Balancer é mais adequado para balanceamento de carga de tráfego HTTP e HTTPS e oferece roteamento de solicitação avançado direcionado para a entrega de arquiteturas de aplicações modernas, incluindo microsserviços e contêineres.

O Network Load Balancer é mais adequado para o balanceamento de carga de tráfego TCP que exija performance extrema. Ele é capaz de processar milhões de solicitações por segundo enquanto mantém latências ultrabaixas, e também é otimizado para lidar com padrões de tráfego súbitos e voláteis.

O Elastic Load Balancing oferece gerenciamento integrado de certificados e descriptografia SSL/TLS, o que proporciona a flexibilidade de gerenciar centralmente as configurações SSL do load balancer e descarregar de sua workload as interações com uso intenso de CPU.

Após escolher o balanceador de carga certo, você poderá começar a utilizar seus recursos para reduzir a quantidade de esforço que seu backend precisa fazer para atender o tráfego.

Por exemplo, ao usar tanto o Application Load Balancer (ALB) quanto o Network Load Balancer (NLB), é possível realizar o descarregamento de criptografia SSL/TLS, que é uma oportunidade de evitar que o handshake TLS com uso intenso da CPU seja concluído pelos destinos e também melhorar o gerenciamento de certificados.

Ao configurar o descarregamento de SSL/TLS no balanceador de carga, ele se torna responsável pela criptografia do tráfego de e para os clientes enquanto entrega o tráfego não criptografado aos backends, liberando os recursos de backend e melhorando o tempo de resposta para os clientes.

O Application Load Balancer também pode fornecer tráfego HTTP/2 sem precisar acomodá-lo em seus destinos. Essa simples decisão pode melhorar o tempo de resposta da aplicação, já que o HTTP/2 usa conexões TCP de forma mais eficiente.

Os requisitos de latência da workload devem ser considerados ao definir a arquitetura. Como exemplo, se você tiver uma aplicação sensível à latência, poderá decidir usar o Network Load Balancer, que oferece latências extremamente baixas. Como alternativa, você pode decidir aproximar a workload dos clientes utilizando o Application Load Balancer em zonas locais da AWS ou mesmo no AWS Outposts.

Outra consideração para workloads sensíveis à latência é o balanceamento de carga entre zonas. Com o balanceamento de carga entre zonas, cada nó do balanceador de carga distribui o tráfego entre os destinos registrados em todas as Zonas de Disponibilidade habilitadas.

Use o Auto Scaling integrado ao balanceador de carga. Um dos principais aspectos de um sistema com performance eficiente está relacionado ao dimensionamento correto dos recursos de backend. Para fazer isso, é possível utilizar as integrações do balanceador de carga para os recursos de destino de backend. Ao usar a integração do balanceador de carga com os grupos do Auto Scaling, os destinos serão adicionados ou removidos do balanceador de carga conforme exigido em resposta ao tráfego recebido. Os balanceadores de carga também podem ser integrados ao Amazon ECS e ao Amazon EKS para workloads em contêineres.

- Amazon ECS: balanceamento de carga do serviço
- Application Load Balancer no Amazon EKS
- Network Load Balancer no Amazon EKS

Etapas de implementação

- Defina seus requisitos de balanceamento de carga, incluindo volume de tráfego, disponibilidade e escalabilidade de aplicações.
- Escolha o tipo certo de balanceador de carga para sua aplicação.
  - Use o Application Load Balancer para workloads HTTP/HTTPS.
  - Use o Network Load Balancer para workloads não HTTP executadas em TCP ou UDP.
  - Use uma combinação de ambos (ALB como destino do NLB) se quiser aproveitar os recursos de ambos os produtos. Por exemplo, é possível fazer isso se você quiser usar os IPs estáticos do NLB junto com o roteamento baseado em cabeçalho HTTP do ALB, ou se quiser expor a workload HTTP em um AWS PrivateLink.
  - Para ver uma comparação completa dos balanceadores de carga, consulte a comparação de produtos do ELB.
- Use o descarregamento de SSL/TLS, se possível.
  - Configure receptores HTTPS/TLS com o Application Load Balancer e o Network Load Balancer integrados ao AWS Certificate Manager.
  - Observe que algumas workloads podem exigir criptografia completa por motivos de conformidade. Nesse caso, é um requisito para permitir a criptografia nos destinos.
  - Para conhecer as práticas recomendadas de segurança, consulte SEC09-BP02 Aplicar criptografia em trânsito.
- Escolha o algoritmo de roteamento certo (apenas ALB).
  - O algoritmo de roteamento pode fazer a diferença em como os destinos de backend são bem utilizados e, portanto, na forma como afetam a performance. Por exemplo, o ALB fornece duas opções para algoritmos de roteamento:
  - Solicitações menos pendentes: use para obter uma melhor distribuição de carga para seus destinos de backend em casos nos quais as solicitações para a aplicação variam em complexidade ou os destinos variam na capacidade de processamento.
  - Round robin: use quando as solicitações e os destinos forem semelhantes, ou se você precisar distribuir as solicitações igualmente entre os destinos.
- Considere isolamento por zona ou entre zonas.
  - Desative a opção entre zonas (isolamento por zona) para melhorias de latência e domínios com falha de zona. Ela é desativada por padrão no NLB e é possível desativá-la por grupo-alvo no ALB.

  - Ative a opção entre zonas para maior disponibilidade e flexibilidade. Ela é ativada por padrão no ALB e é possível ativá-la por grupo-alvo no NLB.
- Ative as manutenções de funcionamento de HTTP para as workloads HTTP (apenas ALB). Com esse recurso, o balanceador de carga pode reutilizar as conexões de backend até expirar o tempo limite da manutenção de funcionamento, melhorando a solicitação HTTP e o tempo de resposta, além de reduzir a utilização de recursos nos destinos de backend. Para obter detalhes sobre como fazer isso para o Apache e o Nginx, consulte Quais são as configurações ideais para usar o Apache ou o NGINX como servidor de backend para o ELB?
- Ative o monitoramento do balanceador de carga.
  - Ative os logs de acesso para seu Application Load Balancer e Network Load Balancer.
  - Os principais campos a considerar para o ALB são `request_processing_time`, `request_processing_time` e `response_processing_time`.
  - Os principais campos a considerar para o NLB são `connection_time` e `tls_handshake_time`.
  - Esteja pronto para consultar os logs quando precisar deles. É possível usar o Amazon Athena para consultar tanto os logs do ALB quanto os logs do NLB.
  - Crie alarmes para métricas relacionadas à performance, como `TargetResponseTime` para ALB.

Recursos

Documentos relacionados:

- Comparação de produtos de ELB
- Infraestrutura global da AWS
- Melhorar a performance e reduzir os custos usando a afinidade de zona de disponibilidade
- Passo a passo para a análise de logs com o Amazon Athena
- Consultar logs do Application Load Balancer
- Monitorar seus Application Load Balancers
- Monitorar os Network Load Balancers
- Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no grupo do Auto Scaling

Vídeos relacionados:

- AWS re:Invent 2023: O que rede pode fazer por sua aplicação?
- AWS re:Inforce 20: Como usar o Elastic Load Balancing para melhorar seu procedimento de segurança em escala
- AWS re:Invent 2018: Mergulho profundo e práticas recomendadas do Elastic Load Balancing
- AWS re:Invent 2021: Como escolher o balanceador de carga certo para suas workloads da AWS
- AWS re:Invent 2019: Como aproveitar ao máximo o Elastic Load Balancing para diferentes workloads

Exemplos relacionados:

- Gateway Load Balancer
- CDK e exemplos do AWS CloudFormation para análise de logs com o Amazon Athena

PERF04-BP05 Escolher protocolos de rede para melhorar a performance

Tome decisões sobre protocolos de comunicação entre sistemas e redes com base no impacto na performance da workload.

Há uma relação entre latência e largura de banda para alcançar o throughput. Por exemplo, se a transferência de arquivos estiver usando TCP, latências mais altas provavelmente reduzirão o throughput geral. Existem abordagens para corrigir isso com ajuste de TCP e protocolos de transferência otimizados, mas uma solução é usar o protocolo UDP.

Práticas comuns que devem ser evitadas:

- Você usa TCP para todas as workloads, independentemente dos requisitos de performance.

Benefícios de implementar esta prática recomendada: verificar se um protocolo apropriado é usado para comunicação entre usuários e componentes da workload ajuda a melhorar a experiência geral do usuário para as aplicações. Por exemplo, o UDP sem conexão permite alta velocidade, mas não oferece retransmissão ou alta confiabilidade. O TCP é um protocolo completo, mas requer maior sobrecarga para processar os pacotes.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Se você puder escolher protocolos diferentes para sua aplicação e tiver experiência nessa área, otimize sua aplicação e a experiência do usuário final usando um protocolo diferente. Observe que essa abordagem apresenta dificuldades significativas e só deve ser experimentada se você tiver otimizado sua aplicação de outras maneiras primeiro.

Uma consideração primária para melhorar a performance da workload é entender os requisitos de latência e throughput e escolher os protocolos de rede que otimizam essa performance.

Quando considerar o uso do TCP

O TCP oferece entrega de dados confiável e pode ser usado para comunicação entre componentes da workload em que a confiabilidade e a entrega garantida de dados é importante. Muitas aplicações baseadas na Web dependem de protocolos baseados em TCP, como HTTP e HTTPS, para abrir soquetes TCP para comunicação entre componentes da aplicação. As transferências de dados por e-mail e arquivo são aplicações comuns que também usam o TCP, pois é um mecanismo de transferência simples e confiável entre componentes de aplicações. Usar o TLS com TCP pode adicionar sobrecarga à comunicação, o que pode resultar em maior latência e redução de throughput, mas traz a vantagem da segurança. A sobrecarga vem principalmente da sobrecarga adicionada do processo de handshake, que pode exigir várias idas e voltas para ser concluído. Quando o handshake for concluído, a sobrecarga da criptografia e descriptografia de dados será relativamente pequena.

Quando considerar o uso do UDP

O UDP é um protocolo sem conexão e, portanto, é adequado para aplicações que precisam de uma transmissão rápida e eficiente, como log, monitoramento e dados de VoIP. Além disso, considere usar o UDP se você tiver componentes da workload que respondam a pequenas consultas de grandes números de clientes para garantir a performance ideal da workload. O Datagram Transport Layer Security (DTLS) é o equivalente UDP do Transport Layer Security (TLS). Ao usar DTLS com UDP, a sobrecarga vem da criptografia e descriptografia de dados, já que o processo de handshake é simplificado. O DTLS também adiciona uma pequena quantidade de sobrecarga aos pacotes de UDP, já que inclui campos adicionais para indicar os parâmetros de segurança e detectar violações.

Quando considerar o uso do SRD

O SRD (datagrama confiável escalável) é um protocolo de transporte de rede otimizado para workloads de alto throughput devido à sua capacidade de fazer o balanceamento de carga do tráfego

em vários caminhos e de se recuperar rapidamente de quedas de pacote ou falhas no link. Assim, o SRD é melhor nos casos de workloads de computação de alta performance (HPC) que exigem comunicação de alto throughput e baixa latência entre os nós de computação. Isso pode incluir tarefas de processamento paralelas, como simulação, modelagem e análise de dados que envolvem uma grande quantidade de transferência de dados entre os nós.

Etapas de implementação

- Use os serviços AWS Global Accelerator e AWS Transfer Family para melhorar o throughput de suas aplicações de transferência de arquivos online. O serviço AWS Global Accelerator ajuda você a obter baixa latência entre os dispositivos cliente e a workload na AWS. Com o AWS Transfer Family, é possível usar protocolos baseados em TCP, como SFTP e FTPS, para escalar e gerenciar com segurança as transferências de arquivos para os serviços de armazenamento da AWS.
- Use a latência de rede para determinar se o TCP é adequado para comunicação entre os componentes da workload. Se a latência de rede entre a aplicação cliente e o servidor for alta, o handshake de três vias do TCP pode levar um tempo, afetando, assim, a capacidade de resposta da aplicação. Métricas como tempo até o primeiro byte (TTFB) e tempo de ida e volta (RTT) podem ser usadas para medir a latência da rede. Se sua workload serve conteúdo dinâmico para os usuários, considere usar o Amazon CloudFront, que estabelece uma conexão persistente com cada origem de conteúdo dinâmico para remover o tempo de configuração da conexão que, de outra forma, diminuiria a velocidade de cada solicitação do cliente.
- Usar TLS com TCP ou UDP pode resultar em maior latência e menor throughput para a workload devido ao impacto da criptografia e descriptografia. Para workloads desse tipo, considere usar o descarregamento de SSL/TLS no Elastic Load Balancing para melhorar a performance da workload, permitindo que o balanceador de carga lide com o processo de criptografia e descriptografia de SSL/TLS em vez de deixar que as instâncias de backend façam isso. Isso pode ajudar a reduzir a utilização da CPU nas instâncias de backend, o que pode melhorar a performance e aumentar a capacidade.
- Use o Network Load Balancer (NLB) para implantar serviços que dependem do protocolo UDP, como autenticação e autorização, registro em log, DNS, IoT e mídia de streaming, visando melhorar a performance e a confiabilidade da workload. O NLB distribui o tráfego de UDP de entrada em vários destinos, permitindo escalar a workload horizontalmente, aumentar a capacidade e reduzir a sobrecarga de um único destino.
- Para suas workloads de computação de alta performance (HPC), considere usar a funcionalidade de Adaptador de Rede Elástica (ENA) Express, que usa o protocolo SRD para melhorar a

performance da rede, fornecendo uma maior largura de banda de fluxo único (25 Gbps) e menor latência final (99,9 percentil) para tráfego de rede entre instâncias do EC2.

- Use o Application Load Balancer (ALB) para rotear e balancear a carga do tráfego de gRPC (Chamadas de procedimento remoto) entre os componentes da workload ou entre os serviços e clientes com gRPC habilitadas. As gRPC usam o protocolo HTTP/2 baseado em TCP para transporte e oferece benefícios de performance, como pegada de rede mais leve, compactação, serialização binária eficiente, suporte para várias linguagens e streaming bidirecional.

Recursos

Documentos relacionados:

- Como rotear tráfego UDP para o Kubernetes
- Application Load Balancer
- Rede avançada do EC2 no Linux
- Rede avançada do EC2 no Windows
- Grupos de posicionamento do EC2
- Como habilitar a rede avançada com o Adaptador de Rede Elástica (ENA) em instâncias Linux
- Network Load Balancer
- Produtos de redes com a AWS
- Passar para o encaminhamento por latência no Amazon Route 53
- VPC Endpoints

Vídeos relacionados:

- AWS re:Invent 2022: Escalar a performance da rede em instâncias do Amazon Elastic Compute Cloud de última geração
- AWS re:Invent 2022: Fundamentos de rede das aplicações

Exemplos relacionados:

- AWS Transit Gateway e soluções de segurança escaláveis
- Workshops de redes da AWS

PERF04-BP06 Escolher o local da workload com base nos requisitos de rede

Avalie as opções para o posicionamento de recursos visando reduzir a latência da rede e melhorar o throughput, proporcionando uma ótima experiência do usuário ao reduzir os tempos de carregamento da página e de transferência de dados.

Práticas comuns que devem ser evitadas:

- Consolidar todos os recursos da workload em uma única localização geográfica.
- Escolher a região mais próxima ao seu local, mas não ao usuário final da workload.

Benefícios de implementar esta prática recomendada: a experiência do usuário é muito afetada pela latência entre o usuário e sua aplicação. Ao usar Regiões da AWS adequadas e a rede global privada da AWS, é possível reduzir a latência e oferecer uma melhor experiência aos usuários remotos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Recursos, como instâncias do Amazon EC2, são colocados em zonas de disponibilidade em Regiões da AWS, zonas locais da AWS, AWS Outposts ou zonas do AWS Wavelength. A escolha desse local influencia o throughput e a latência da rede de determinado local do usuário. Serviços de borda como o Amazon CloudFront e o AWS Global Accelerator também podem ser usados para melhorar a performance da rede, seja armazenando o conteúdo em cache nos locais da borda ou oferecendo aos usuários um ótimo caminho para a workload por meio da rede global da AWS.

O Amazon EC2 oferece grupos de posicionamento para redes. Um grupo de posicionamento é um agrupamento lógico de instâncias para diminuir a latência. O uso de grupos de posicionamento com tipos de instância compatíveis e um Adaptador de Rede Elástica (ENA) permite que as workloads participem de uma rede de baixa latência e com jitter reduzido e de 25 Gbps. Recomenda-se o uso de grupos de posicionamento para workloads que se beneficiam de baixa latência de rede, alto throughput de rede ou ambos.

Serviços sensíveis à latência são fornecidos em locais de borda usando uma rede global da AWS, como o Amazon CloudFront. Esses locais de borda costumam oferecer serviços, como rede de entrega de conteúdo (CDN) e sistema de nomes de domínio (DNS). Ao ter esses serviços na borda, as workloads podem responder com baixa latência a solicitações de conteúdo ou resolução de DNS. Esses serviços também fornecem serviços geográficos, como direcionamento geográfico de

conteúdo (fornecendo conteúdo diferente conforme o local do usuário final) ou encaminhamento por latência para direcionar os usuários finais à região mais próxima (latência mínima).

Use serviços de borda para reduzir a latência e possibilitar o armazenamento do conteúdo em cache. Configure corretamente o controle de cache para DNS e HTTP/HTTPS a fim de aproveitar ao máximo essas abordagens.

Etapas de implementação

- Capture informações sobre o tráfego IP que entra e sai das interfaces de rede.
  - Como registrar tráfego IP em log com Logs de fluxo da VPC
  - Como o endereço IP do cliente é preservado no AWS Global Accelerator
- Analise os padrões de acesso à rede em sua workload para identificar como os usuários utilizam sua aplicação.
  - Use ferramentas de monitoramento, como o Amazon CloudWatch e o AWS CloudTrail, para coletar dados sobre atividades de rede.
  - Analise os dados para identificar o padrão de acesso à rede.
- Selecione as Regiões para implantação da workload com base nos seguintes elementos fundamentais:
  - Onde seus dados estão localizados: para aplicações com uso intenso de dados (como big data e machine learning), o código da aplicação deve ser executado o mais perto possível dos dados.
  - Onde seus usuários estão localizados: para aplicações voltadas ao usuário, escolha uma ou mais regiões perto dos clientes da workload.
  - Outras restrições: considere restrições como custo e conformidade, conforme explicado em O que considerar ao selecionar uma região para suas workloads.
- Use zonas locais da AWS para executar workloads como renderização de vídeo. As zonas locais permitem que você se beneficie de ter recursos de computação e armazenamento mais próximos dos usuários finais.
- Use o AWS Outposts para workloads que precisam permanecer on-premises e onde você deseja que essa workload seja executada ininterruptamente com o restante de suas workloads na AWS.
- Aplicações como streaming de vídeo ao vivo em alta resolução, áudio de alta fidelidade ou realidade aumentada/realidade virtual (RA/RV) exigem latência ultrabaixa para dispositivos 5G. Para aplicações desse tipo, considere o AWS Wavelength. O AWS Wavelength incorpora serviços de armazenamento e computação da AWS em redes 5G, fornecendo a infraestrutura móvel de computação de borda para desenvolver, implantar e escalar aplicações de latência ultrabaixa.

- Use o armazenamento em cache local ou soluções de armazenamento em cache da AWS para dados usados com frequência a fim de aumentar a performance, reduzir a movimentação de dados e diminuir o impacto ambiental.

Serviço	Quando usar
Amazon CloudFront	Use para armazenar conteúdo estático em cache, como imagens, scripts e vídeos, além de conteúdo dinâmico como respostas de API ou aplicações Web.
Amazon ElastiCache	Use para armazenar conteúdo em cache para aplicações Web.
DynamoDB Accelerator	Use para adicionar aceleração na memória às suas tabelas do DynamoDB.

- Use serviços que podem ajudar você a executar código mais perto dos usuários da workload, como a seguir:

Serviço	Quando usar
Lambda@Edge	Use para operações com uso intenso de computação que são iniciadas quando objetos não estão no cache.
Amazon CloudFront Functions	Use para casos de uso simples, como solicitações HTTP(s) ou manipulações de resposta que podem ser iniciadas por funções de curta duração.
AWS IoT Greengrass	Use para executar computação local, mensagens e armazenamento de dados em cache para dispositivos conectados.

- Algumas aplicações exigem pontos de entrada fixos ou maior performance ao reduzir o jitter e a latência de primeiro byte, além de aumentar o throughput. Essas aplicações podem se beneficiar de serviços de rede que fornecem endereços IP anycast estáticos e terminação TCP em locais

da borda. O AWS Global Accelerator pode melhorar a performance das suas aplicações em até 60% e fornecer failover rápido para arquiteturas multirregiões. O AWS Global Accelerator fornece endereços IP anycast estáticos que servem como um ponto de entrada fixo para suas aplicações hospedadas em uma ou mais Regiões da AWS. Esses endereços IP permitem que o tráfego entre na rede global da AWS o mais próximo possível dos usuários. O AWS Global Accelerator reduz o tempo de configuração da conexão inicial ao estabelecer uma conexão TCP entre o cliente e o local da borda da AWS mais próximo ao cliente. Analise o uso do AWS Global Accelerator para melhorar a performance das workloads de TCP/UDP e forneça failover rápido para arquiteturas de várias Regiões.

Recursos

Práticas recomendadas relacionadas:

- COST07-BP02 Implementar regiões com base nos custos
- COST08-BP03 Implementar serviços para reduzir custos de transferência de dados
- REL10-BP01 Implantar a workload em vários locais
- REL10-BP02 Escolher os locais apropriados para sua implantação de vários locais
- SUS01-BP01 Escolher a região com base nos requisitos empresariais e nas metas de sustentabilidade
- SUS02-BP04 Otimizar o posicionamento geográfico das workloads com base nos respectivos requisitos de rede
- SUS04-BP07 Minimizar a movimentação de dados entre redes

Documentos relacionados:

- Infraestrutura global da AWS
- Zonas locais da AWS e AWS Outposts: como escolher a tecnologia certa para sua workload de borda)
- Grupos de posicionamento
- Zonas locais da AWS
- AWS Outposts
- AWS Wavelength
- Amazon CloudFront

- AWS Global Accelerator
- AWS Direct Connect
- AWS Site-to-Site VPN
- Amazon Route 53

Vídeos relacionados:

- Vídeo explicativo de zonas locais da AWS
- Visão geral do AWS Outposts e como ele funciona
- AWS re:Invent 2023: Uma estratégia de migração para workloads periféricas e on-premises
- AWS re:Invent 2021: AWS Outposts: como trazer a experiência da AWS para ambientes on-premises
- AWS re:Invent 2020: AWS Wavelength: executar aplicações com latência ultrabaixa na borda 5G
- AWS re:Invent 2022: Zonas locais da AWS: como criar aplicações para uma borda distribuída
- AWS re:Invent 2021: Criar sites de baixa latência com o Amazon CloudFront
- AWS re:Invent 2022: Aprimorar a performance e a disponibilidade com o AWS Global Accelerator
- AWS re:Invent 2022: Criar sua rede de longa distância usando a AWS
- AWS re:Invent 2020: Gerenciamento de tráfego global com o Amazon Route 53

Exemplos relacionados:

- Workshop de roteamento personalizado no AWS Global Accelerator
- Como lidar com reescritas e redirecionamentos usando funções da borda

PERF04-BP07 Otimizar a configuração da rede com base em métricas

Use dados coletados e analisados para tomar decisões bem informadas sobre a otimização da configuração da rede.

Práticas comuns que devem ser evitadas:

- Pressupor que todos os problemas relacionados à performance são relacionados à aplicação.
- Testar a performance da rede a partir de um local próximo ao local em que a workload foi implantada.

- Usar configurações-padrão para todos os serviços de rede.
- Provisionar em excesso recursos de rede para fornecer capacidade suficiente.

Benefícios de implementar esta prática recomendada: coletar as métricas necessárias da rede da AWS e implementar ferramentas de monitoramento de rede permite entender a performance da rede e otimizar as respectivas configurações.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Monitorar o tráfego de entrada e saída das VPCs, sub-redes ou interfaces de rede é fundamental para entender como utilizar os recursos de rede da AWS e otimizar as configurações da rede. Ao usar as ferramentas de rede da AWS a seguir, é possível verificar mais informações sobre o uso do tráfego, o acesso à rede e os logs.

Etapas de implementação

- Identifique as principais métricas de performance, como latência ou perda de pacotes. A AWS fornece diversas ferramentas que podem ajudar você a coletar essas métricas. Ao usar as ferramentas a seguir, é possível verificar mais informações sobre o uso do tráfego, o acesso à rede e os logs:

Ferramenta AWS	Onde usar
Gerenciador de endereços IP da Amazon VPC.	Use o IPAM para planejar, rastrear e monitorar endereços IP para workloads da AWS e on-premises. Essa é uma prática recomendada para otimizar o uso e a alocação de endereços IP.
Logs de fluxo da VPC	Use os Logs de fluxo da VPC para obter informações detalhadas sobre o tráfego de entrada e saída das interfaces de rede nas VPCs. Com os logs de fluxo da VPC, é possível diagnosticar regras extremame nte restritivas ou permissivas do grupo de

Ferramenta AWS	Onde usar
	segurança e determinar a direção do tráfego de entrada e saída das interfaces de rede.
Logs de fluxo do AWS Transit Gateway	Use os logs de fluxo do AWS Transit Gateway para capturar informações sobre o tráfego IP que entra e sai dos seus gateways de trânsito.
Registro em log de consultas ao DNS	Registre informações sobre consultas ao DNS, públicas ou privadas recebidas pelo Route 53. Com os logs de DNS, é possível otimizar as configurações de DNS entendend o o domínio ou subdomínio solicitado ou os locais da borda do Route 53 que responderam às consultas ao DNS.
Reachability Analyzer	O Reachability Analyzer ajuda a analisar e depurar a acessibilidade da rede. O Reachabil ity Analyzer é uma ferramenta de análise de configuração que permite realizar testes de conectividade entre um recurso de origem e um recurso de destino em suas VPCs. Essa ferramenta ajuda a verificar se a configura ção da rede corresponde à conectividade pretendida.
Analisador de Acesso à Rede	O Analisador de Acesso à Rede ajuda a entender o acesso via rede aos seus recursos. O Analisador de Acesso à Rede pode ser usado para especificar os requisito s de acesso à rede e identificar possíveis caminhos de rede que não atendam aos requisitos especificados. Ao otimizar a configuração da rede correspondente, é possível entender e verificar o estado da rede e demonstrar se a rede na AWS atende aos seus requisitos de conformidade.

Ferramenta AWS	Onde usar
Amazon CloudWatch	Use o Amazon CloudWatch e ative as métricas apropriadas para as opções de rede. Escolha a métrica de rede certa para sua workload. Por exemplo, é possível habilitar métricas para o uso do endereço de rede da VPC, o gateway NAT da VPC, o AWS Transit Gateway, o túnel da VPN, o AWS Network Firewall, o Elastic Load Balancing e o AWS Direct Connect. Monitorar continuamente as métricas é uma prática recomendada para observar e entender o status e o uso da rede, o que ajuda a otimizar a configuração da rede com base em suas observações.
AWS Network Manager	Com o AWS Network Manager, é possível monitorar a performance histórica e em tempo real da Rede Global da AWS para fins operacionais e de planejamento. O Gerenciador de Rede fornece latência de rede agregada entre as Regiões da AWS e as zonas de disponibilidade e dentro de cada zona de disponibilidade, permitindo que você entenda melhor como a performance da sua aplicação se relaciona à performance da rede da AWS subjacente.
Amazon CloudWatch RUM	Use o Amazon CloudWatch RUM para coletar as métricas que fornecem os insights que ajudam a identificar, entender e melhorar a experiência do usuário.

- Identifique os principais interlocutores e os padrões de tráfego de aplicações usando VPC e logs de fluxo do AWS Transit Gateway.

- Avalie e otimize sua arquitetura de rede atual, incluindo VPCs, sub-redes e roteamento. Como exemplo, você pode avaliar como diferentes emparelhamentos de VPC ou AWS Transit Gateway podem ajudar a melhorar a rede em sua arquitetura.
- Avalie os caminhos de roteamento em sua rede para verificar se o caminho mais curto entre os destinos é sempre usado. O Analisador de Acesso à Rede pode ajudar a fazer isso.

Recursos

Documentos relacionados:

- Log de consultas ao DNS público
- O que é IPAM?
- O que é o Reachability Analyzer?
- O que é o Analisador de Acesso à Rede?
- Métricas do CloudWatch para suas VPCs
- Otimizar a performance e reduzir os custos de análise de rede com os Logs de fluxo da VPC no formato Apache Parquet
- Monitorar suas redes global e básica com métricas do Amazon CloudWatch
- Monitorar continuamente o tráfego e os recursos da rede

Vídeos relacionados:

- AWS re:Invent 2023: Um guia do desenvolvedor para redes na nuvem
- AWS re:Invent 2023: Pronto para o que vem a seguir? Desenvolver redes para crescimento e flexibilidade
- AWS re:Invent 2023: Designs e novos recursos de VPCs avançadas
- AWS re:Invent 2022: Mergulho profundo na infraestrutura de rede da AWS
- AWS re:Invent 2020: Dicas e práticas recomendadas de rede com o AWS Well-Architected Framework
- AWS re:Invent 2020: Monitorar e solucionar problemas de tráfego de rede

Exemplos relacionados:

- Workshops de redes da AWS

- Monitoramento de rede da AWS
- Observar e diagnosticar sua rede na AWS
- Como encontrar e lidar com configurações de rede incorretas na AWS

Processo e cultura

Perguntas

- PERF 5. Como suas práticas e cultura organizacionais contribuem para a eficiência de performance em sua workload?

PERF 5. Como suas práticas e cultura organizacionais contribuem para a eficiência de performance em sua workload?

Ao arquitetar workloads, há princípios e práticas que você pode adotar para ajudar na melhor execução de workloads na nuvem eficientes e de alta performance. Para adotar uma cultura que promova a eficiência de performance das workloads na nuvem, considere estes princípios e práticas fundamentais:

Práticas recomendadas

- PERF05-BP01 Estabelecer indicadores-chave de performance (KPIs) para medir a integridade e a performance da workload
- PERF05-BP02 Usar soluções de monitoramento para entender as áreas em que a performance é mais crítica
- PERF05-BP03 Definir um processo para melhorar a performance da workload
- PERF05-BP04 Fazer o teste de carga da workload
- PERF05-BP05 Usar automação para corrigir proativamente problemas relacionados a performance
- PERF05-BP06 Manter a workload e os serviços atualizados
- PERF05-BP07 Revisar as métricas regularmente

PERF05-BP01 Estabelecer indicadores-chave de performance (KPIs) para medir a integridade e a performance da workload

Identifique os KPIs que medem a performance da workload de forma quantitativa e qualitativa. Os KPIs ajudam você a medir a integridade e a performance de uma workload relacionada a uma meta empresarial.

Práticas comuns que devem ser evitadas:

- Monitorar as métricas somente no nível do sistema para obter informações da workload e não compreende aos impactos dessas métricas nos negócios.
- Pressupor que os KPIs já estejam publicados e compartilhados como dados de métricas comuns.
- Não definir um KPI quantitativo e mensurável.
- Não alinhar os KPIs às metas ou estratégias empresariais.

Benefícios de implementar esta prática recomendada: identificar KPIs específicos que representam a integridade e a performance da workload ajuda a alinhar as equipes em suas prioridades e a definir resultados empresariais bem-sucedidos. O compartilhamento dessas métricas com todos os departamentos fornece visibilidade e alinhamento dos limites, das expectativas e do impacto nos negócios.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Os KPIs permitem que as empresas e as equipes de engenharia alinhem a medição das metas e estratégias de como esses fatores são combinados para produzir resultados comerciais. Por exemplo, a workload de um site pode usar o tempo de carregamento da página como uma indicação da performance geral. Essa métrica seria um dos vários pontos de dados que medem a experiência do usuário. Além de identificar os limites do tempo de carregamento da página, documente o resultado esperado ou o risco da empresa se a performance ideal não for atingida. Um longo tempo de carregamento da página afeta diretamente os usuários finais, diminui a classificação da experiência do usuário e pode resultar em perda de clientes. Ao definir os limites dos KPIs, combine os testes comparativos do setor e as expectativas dos usuários finais. Por exemplo, se o teste comparativo do setor aplicável for o carregamento de uma página da Web em dois segundos, mas os usuários finais esperarem que uma página da Web seja carregada em um segundo, você deverá pensar nos dois pontos de dados ao estabelecer o KPI.

Sua equipe deve avaliar os KPIs da workload usando dados detalhados em tempo real e dados históricos para referência e criar painéis que calculem as métricas nos dados de KPI para derivar informações operacionais e de utilização. Os KPIs devem ser documentados e incluir limites que apoiem as metas e estratégias empresariais, bem como mapeados de acordo com as métricas que estão sendo monitoradas. Os KPIs devem ser revisitados quando as metas e as estratégias da empresa ou os requisitos dos usuários finais mudam.

Etapas de implementação

- Identifique as partes interessadas: identifique e documente as principais partes interessadas da empresa, incluindo as equipes de desenvolvimento e operações.
- Defina objetivos: trabalhe com essas partes interessadas para definir e documentar os objetivos da workload. Considere os aspectos críticos de performance das workloads, como throughput, tempo de resposta e custo, bem como as metas de negócios, como a satisfação dos usuários.
- Revise as práticas recomendadas do setor: revise as práticas recomendadas do setor para identificar KPIs relevantes alinhados aos objetivos da workload.
- Identifique métricas: identifique métricas que estejam alinhadas aos objetivos da sua workload e possam ajudar a medir a performance e as metas de negócios. Estabeleça KPIs com base nessas métricas. Exemplos de métricas são tempo médio de resposta, número de usuários simultâneos, entre outras.
- Defina e documente KPIs: use as práticas recomendadas do setor e os objetivos da workload para definir metas de KPI da workload. Use essas informações para definir limites de KPI no nível de gravidade ou de alarme. Identifique e documente o risco e o impacto no caso de um KPI não ser atendido.
- Implemente monitoramento: use ferramentas de monitoramento como o Amazon CloudWatch ou o AWS Config para coletar métricas e medir KPIs.
- Comunique visualmente os KPIs: use ferramentas de painel como o Amazon QuickSight para visualizar e comunicar os KPIs com as partes interessadas.
- Analise e otimize: revise e analise regularmente as métricas para identificar áreas da workload que precisam ser aprimoradas. Trabalhe com as partes interessadas para implementar essas melhorias.
- Revise e refine: revise regularmente as métricas e os KPIs para avaliar sua eficácia, especialmente quando as metas de negócios ou a performance da workload mudam.

Recursos

Documentos relacionados:

- Documentação do CloudWatch
- AWS Partners de monitoramento, registro em log e performance
- Ferramentas de observabilidade da AWS
- A importância dos indicadores-chave de performance (KPIs) para migrações para a nuvem em grande escala
- Como rastrear KPIs de otimização de custos com o painel de KPI
- Documentação do X-Ray
- Usar painéis do Amazon CloudWatch
- KPIs do Amazon QuickSight

Vídeos relacionados:

- AWS re:Invent 2023: Otimizar custos e performance e rastrear o progresso rumo à mitigação
- AWS re:Invent 2023: Gerenciar eventos do ciclo de vida dos recursos em grande escala com o AWS Health
- AWS re:Invent 2023: Performance e eficiência no Pinterest: otimizando as instâncias mais recentes
- AWS re:Invent 2022: Otimização da AWS: etapas acionáveis para resultados imediatos
- AWS re:Invent 2023: Criar uma estratégia efetiva de observabilidade
- AWS Summit SF 2022: Observabilidade full-stack e monitoramento de aplicações com a AWS
- AWS re:Invent 2023: Escalar na AWS para seus primeiros 10 milhões de usuários
- AWS re:Invent 2022: Como a Amazon usa métricas melhores para aprimorar a performance de sites
- Criar uma estratégia de métricas eficaz para sua empresa | Eventos da AWS

Exemplos relacionados:

- Criar um painel com o Amazon QuickSight

PERF05-BP02 Usar soluções de monitoramento para entender as áreas em que a performance é mais crítica

Entenda e identifique áreas em que aumentar a performance de sua workload causará um impacto positivo sobre a eficiência ou a experiência do cliente. Por exemplo, um site que tenha muita interação com o cliente se beneficiaria do uso de serviços de borda para aproximar a entrega de conteúdo dos clientes.

Práticas comuns que devem ser evitadas:

- Você pressupõe que as métricas de computação padrão, como utilização de CPU ou pressão de memória, são suficientes para detectar problemas de performance.
- Você só usa as métricas comuns registradas pelo software de monitoramento selecionado.
- Você só revisa as métricas quando há um problema.

Benefícios de implementar esta prática recomendada: compreender áreas críticas de performance ajuda os proprietários de workloads a monitorar KPIs e priorizar melhorias de alto impacto.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Configure um rastreamento completo para identificar padrões de tráfego, latência e áreas de performance críticas. Monitore os padrões de acesso aos dados para consultas lentas ou dados particionados e fragmentados incorretamente. Identifique as áreas de restrição da workload usando o teste ou monitoramento de carga.

Aumente a eficiência de performance entendendo sua arquitetura, os padrões de tráfego e os padrões de acesso aos dados, além de identificar os tempos de latência e processamento. Identifique possíveis gargalos que possam afetar a experiência do cliente com o crescimento da workload. Depois de investigar essas áreas, veja qual solução você pode implantar para eliminar esses problemas de performance.

Etapas de implementação

- Configure um monitoramento completo para capturar todos os componentes e as métricas da workload. Aqui estão alguns exemplos de soluções de monitoramento na AWS.

Serviço	Onde usar
Amazon CloudWatch Real-User Monitoring (RUM)	Para capturar as métricas de performance da aplicação de sessões de frontend e do lado do cliente de usuários reais.
AWS X-Ray	Para monitorar o tráfego por meio das camadas de aplicação e identificar a latência entre componentes e dependências. Use os mapas do serviço X-Ray para ver os relaciona mentos e a latência entre os componentes da workload.
Insights de performance do Amazon Relationa l Database Service	Para ver as métricas de performance do banco de dados e identificar melhorias de performance.
Monitoramento avançado do Amazon RDS	Para ver métricas de performance do SO do banco de dados.
Amazon DevOps Guru	Para detectar padrões operacionais anormais a fim de identificar problemas operacionais antes que eles afetem os clientes.

- Realize testes para gerar métricas, identificar padrões de tráfego, gargalos e áreas de performance críticas. Aqui estão alguns exemplos de como realizar testes:
  - Configure os CloudWatch Synthetic Canaries para imitar programaticamente as atividades do usuário baseadas no navegador usando trabalhos cron do Linux ou expressões rate para gerar métricas consistentes ao longo do tempo.
  - Use a solução AWS Distributed Load Testing para gerar tráfego de pico ou testar a workload na taxa de crescimento esperada.
- Avalie as métricas e a telemetria para identificar as áreas de performance críticas. Avalie essas áreas com sua equipe para discutir sobre o monitoramento e as soluções visando evitar gargalos.
- Experimente com melhorias de performance e meça essas alterações com dados. Como exemplo, você pode usar o CloudWatch Evidently para testar novas melhorias e impactos de performance em sua workload.

Recursos

Documentos relacionados:

- Novidades no AWS Observability na re:Invent 2023
- Amazon Builders' Library
- Documentação do X-Ray
- Amazon CloudWatch RUM
- Amazon DevOps Guru

Vídeos relacionados:

- AWS re:Invent 2023: [LANÇAMENTO] Monitoramento de aplicações para workloads modernas
- AWS re:Invent 2023: Como implementar a observabilidade de aplicações
- AWS re:Invent 2023: Criar uma estratégia efetiva de observabilidade
- AWS Summit SF 2022: Observabilidade full-stack e monitoramento de aplicações com a AWS
- AWS re:Invent 2022: Otimização da AWS: etapas acionáveis para resultados imediatos
- AWS re:Invent 2022: Amazon Builders' Library: 25 anos de excelência operacional da Amazon
- AWS re:Invent 2022: Como a Amazon usa métricas melhores para aprimorar a performance de sites
- Monitoramento visual de aplicações com o Amazon CloudWatch Synthetics

Exemplos relacionados:

- Medir o tempo de carregamento da com o Amazon CloudWatch Synthetics
- Cliente Web do Amazon CloudWatch RUM
- X-Ray SDK para Python
- Teste de carga distribuída na AWS

PERF05-BP03 Definir um processo para melhorar a performance da workload

Defina um processo para avaliar novos serviços, padrões de design, tipos de recursos e configurações à medida que eles se tornam disponíveis. Por exemplo, execute testes de

performance existentes em novas ofertas de instância para determinar o potencial delas de aprimorar sua workload.

Práticas comuns que devem ser evitadas:

- Você pressupõe que sua arquitetura atual é estática e não será atualizada ao longo do tempo.
- Você apresenta alterações de arquitetura ao longo do tempo sem justificativa de métrica.

Benefícios de implementar esta prática recomendada: ao definir seu processo para fazer alterações de arquitetura, é possível usar os dados coletados para influenciar o projeto da workload ao longo do tempo.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A performance da sua workload tem algumas restrições importantes. Guarde essas restrições para saber que tipos de inovação podem aumentar a performance da sua workload. Use essas informações enquanto estiver aprendendo sobre novos serviços ou tecnologias à medida que surgem e identificar maneiras de reduzir restrições ou gargalos.

Identifique as principais restrições de performance da workload. Documente suas restrições de performance da workload para que você saiba quais tipos de inovação podem aprimorar a performance da workload.

Etapas de implementação

- Identifique os KPIs: identifique os KPIs de performance da workload conforme descrito em PERF05-BP01 Estabelecer indicadores-chave de performance (KPIs) para medir a integridade e a performance da workload para definir sua workload.
- Implemente monitoramento: use ferramentas de observabilidade da AWS para coletar métricas de performance e medir KPIs.
- Analise: faça uma análise aprofundada para identificar as áreas (como configuração e código da aplicação) na workload que apresentam baixa performance, conforme descrito em PERF05-BP02 Usar soluções de monitoramento para entender as áreas em que a performance é mais crítica. Use suas ferramentas de análise e performance para identificar as estratégias de melhoria de performance.
- Valide as melhorias: use ambientes de sandbox ou de pré-produção para validar a eficácia das estratégias de aperfeiçoamento.

- Implemente mudanças: implemente as mudanças na produção e monitore constantemente a performance da workload. Documente as melhorias e comunique as mudanças às partes interessadas.
- Revise e refine: revise regularmente seu processo de melhoria de performance para identificar áreas a serem aprimoradas.

Recursos

Documentos relacionados:

- Blog da AWS
- Novidades da AWS
- AWS Skill Builder

Vídeos relacionados:

- AWS re:Invent 2022: Como entregar arquiteturas sustentáveis e de alta performance
- AWS re:Invent 2023: Otimizar custos e performance e rastrear o progresso rumo à mitigação
- AWS re:Invent 2022: Otimização da AWS: etapas acionáveis para resultados imediatos
- AWS re:Invent 2022: Otimize suas workloads da AWS com a orientação de práticas recomendadas

Exemplos relacionados:

- GitHub da AWS

PERF05-BP04 Fazer o teste de carga da workload

Teste sua workload para verificar se ela pode lidar com a carga de produção e identificar qualquer gargalo de performance.

Práticas comuns que devem ser evitadas:

- Você faz um teste de carga de partes individuais da workload, mas não de toda ela.
- Você faz um teste de carga em uma infraestrutura que não é igual ao seu ambiente de produção.
- Você só faz testes de carga para a carga esperada, mas para nada além dela, para ajudar a prever onde pode haver problemas futuros.

- Você faz testes de carga sem consultar a política de testes do Amazon EC2 e enviar um formulário de envio de eventos simulados. Isso faz com que o teste não seja executado, pois parece um evento de negação de serviço.

Benefícios de implementar esta prática recomendada: medir sua performance em um teste de carga mostrará onde você será afetado à medida que a carga aumentar. Com isso você terá a capacidade de antecipar as alterações necessárias antes que elas afetem sua workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

O teste de carga na nuvem é um processo para medir a performance da workload na nuvem em condições realistas com a carga esperada do usuário. Esse processo envolve o provisionamento de um ambiente de nuvem semelhante ao de produção, o uso de ferramentas de teste de carga para gerar carga e a análise de métricas para avaliar a capacidade da workload de lidar com cargas realistas. Execute os testes de carga usando versões sintéticas ou limpas dos dados de produção (remova informações confidenciais ou de identificação). Realize testes de carga automaticamente como parte de seu pipeline de entrega e compare os resultados a KPIs e limites predefinidos. Esse processo ajuda você a continuar alcançando a performance necessária.

Etapas de implementação

- Defina seus objetivos de teste: identifique os aspectos de performance da workload que você deseja avaliar, como throughput e tempo de resposta.
- Selecione uma ferramenta de teste: escolha e configure a ferramenta de teste de carga adequada à workload.
- Configure seu ambiente: configure o ambiente de teste com base no ambiente de produção. É possível usar os serviços da AWS para executar ambientes em escala de produção para testar a arquitetura.
- Implemente o monitoramento: use ferramentas de monitoramento como o Amazon CloudWatch para coletar métricas dos recursos em sua arquitetura. Você também pode coletar e publicar métricas personalizadas.
- Defina cenários: defina os cenários e parâmetros do teste de carga (como duração do teste e número de usuários).
- Faça testes de carga: realize cenários de teste em grande escala. Aproveite a Nuvem AWS para testar a workload e descobrir se há uma falha na escala ou se ela está com a escala reduzida

horizontalmente de maneira não linear. Por exemplo, use instâncias spot para gerar cargas a um baixo custo e descobrir gargalos antes que eles ocorram em produção.

- Analise os resultados do teste: analise os resultados para identificar gargalos de performance e áreas para melhorias.
- Documente e compartilhe descobertas: documente e relate as descobertas e recomendações. Compartilhe essas informações com as partes interessadas para ajudá-las a tomar decisões embasadas sobre estratégias de otimização da performance.
- Faça iterações contínuas: o teste de carga deve ser realizado regularmente, especialmente após uma alteração ou atualização do sistema.

Recursos

Documentos relacionados:

- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Teste de carga distribuída na AWS

Vídeos relacionados:

- AWS Summit ANZ 2023: Acelere com confiança com o teste de carga distribuída da AWS
- AWS re:Invent 2022: Escalar na AWS para seus primeiros 10 milhões de usuários
- Resolver com soluções da AWS: teste de carga distribuída

- AWS re:Invent 2021: Otimize aplicações com base em insights do usuário final com o Amazon CloudWatch RUM
- Demonstração do Amazon CloudWatch Synthetics

Exemplos relacionados:

- Teste de carga distribuída na AWS

PERF05-BP05 Usar automação para corrigir proativamente problemas relacionados a performance

Use indicadores-chave de performance (KPIs), aliados a sistemas de monitoramento e alerta, para abordar proativamente problemas relacionados à performance.

Práticas comuns que devem ser evitadas:

- Você só permite que a equipe de operações faça alterações operacionais na workload.
- Você permite todos os filtros de alarmes para a equipe de operações, sem correção proativa.

Benefícios de implementar esta prática recomendada: a correção proativa de ações de alarme permite que a equipe de suporte se concentre nos itens que não são acionáveis automaticamente. Isso ajuda a equipe de operações a lidar com todos os alarmes sem ficar sobrecarregada e, em vez disso, se concentrar apenas nos alarmes críticos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Sempre que possível, use alarmes para desencadear ações automatizadas visando corrigir problemas. Se a resposta automatizada não for possível, encaminhe o alarme para aqueles capazes de responder. Por exemplo, você pode ter um sistema capaz de prever os valores de indicadores-chave de performance (KPI) esperados e emitir um alarme quando eles ultrapassarem determinados limites, ou uma ferramenta capaz de interromper ou reverter automaticamente as implantações caso os KPIs estejam fora dos valores esperados.

Implemente processos que deem visibilidade à performance à medida que a workload estiver sendo executada. Para determinar se a performance da workload é ideal, crie painéis de monitoramento e estabeleça normas de linha de base para as expectativas de performance.

Etapas de implementação

- Identifique o fluxo de trabalho de correção: identifique e compreenda o problema de performance que pode ser corrigido automaticamente. Use soluções de monitoramento da AWS como o Amazon CloudWatch ou o AWS X-Ray para obter ajuda para entender melhor a causa-raiz do problema.
- Defina o processo de automação: crie um plano e um processo de correção detalhados que possam ser usados para corrigir automaticamente o problema.

- Configure o evento de iniciação: configure o evento para iniciar automaticamente o processo de correção. Por exemplo, você pode definir um acionador para reiniciar automaticamente uma instância quando ela atinge determinado limite de utilização da CPU.
- Automatize a correção: use serviços e tecnologias da AWS para automatizar o processo de correção. Por exemplo, o AWS Systems Manager Automation fornece uma maneira segura e escalável de automatizar o processo de correção. Use a lógica de autocorreção para reverter as alterações se elas não conseguirem resolver o problema.
- Teste o fluxo de trabalho: teste o processo de correção automatizado em um ambiente de pré-produção.
- Implemente o fluxo de trabalho: implemente a correção automatizada no ambiente de produção.
- Desenvolva um playbook: desenvolva e documente um playbook que descreva as etapas do plano de correção, incluindo os eventos de iniciação, a lógica de correção e as ações tomadas. Treine as partes interessadas para ajudá-las a responder com eficácia aos eventos de correção automatizada.
- Revise e refine: avalie regularmente a eficácia do fluxo de trabalho automatizado de correção. Ajuste os eventos de iniciação e a lógica de correção, se necessário.

Recursos

Documentos relacionados:

- Documentação do CloudWatch
- Parceiros de monitoramento, log e performance da AWS Partner Network
- Documentação do X-Ray
- Usar alarmes e ações de alarme no CloudWatch
- Criar uma prática de automação de nuvem para excelência operacional: práticas recomendadas do AWS Managed Services
- Automatizar o ajuste de performance do Amazon Redshift com a otimização automática de tabelas

Vídeos relacionados:

- AWS re:Invent 2023: Estratégias para escalação automatizada, correção e autocorreção inteligente
- AWS re:Invent 2023: [LANÇAMENTO] Monitoramento de aplicações para workloads modernas

- AWS re:Invent 2023: Como implementar a observabilidade de aplicações
- AWS re:Invent 2021: Automatizar de forma inteligente as operações na nuvem
- AWS re:Invent 2022: Configurar controles em escala em seu ambiente da AWS
- AWS re:Invent 2022: Automatizar o gerenciamento e a conformidade de patches usando a AWS
- AWS re:Invent 2022: Como a Amazon usa métricas melhores para aprimorar a performance de sites
- AWS re:Invent 2023: Relaxe: diagnostique e resolva problemas de performance com o Amazon RDS
- AWS re:Invent 2021: {Novo lançamento} Detecte e resolva problemas automaticamente com o Amazon DevOps Guru
- AWS re:Invent 2023: Centralize suas operações

Exemplos relacionados:

- O CloudWatch Logs personaliza alarmes

PERF05-BP06 Manter a workload e os serviços atualizados

Fique em dia com os novos serviços e atributos de nuvem para adotar recursos eficientes, remover problemas e melhorar a eficiência geral da performance da workload.

Práticas comuns que devem ser evitadas:

- Você pressupõe que sua arquitetura atual é estática e não será atualizada ao longo do tempo.
- Você não tem nenhum sistema ou ritmo regular para avaliar se software ou pacotes atualizados são compatíveis com sua workload.

Benefícios de implementar esta prática recomendada: ao estabelecer um processo para se atualizar sobre novos serviços e ofertas, você pode adotar novos atributos e recursos, resolver problemas e melhorar a performance da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Avalie maneiras de melhorar a performance à medida que novos serviços, padrões de design e atributos de produtos são disponibilizados. Determine quais deles poderiam aprimorar a performance

ou aumentar a eficiência da workload por meio de avaliações, discussões internas ou análises externas. Defina um processo para avaliar atualizações, novos recursos e serviços relevantes para sua workload. Por exemplo, crie uma prova de conceito que use novas tecnologias ou consulte um grupo interno. Ao testar novas ideias ou serviços, faça testes de performance para medir o impacto causado por eles na performance da workload.

Etapas de implementação

- Faça o inventário da workload: faça o inventário de software e arquitetura da workload e identifique os componentes que precisam ser atualizados.
- Identifique fontes de atualizações: identifique novidades e atualize fontes relacionadas aos componentes da workload. Como exemplo, você pode assinar Novidades no blog da AWS para ver os produtos que correspondem ao componente da sua workload. Você pode assinar o feed RSS ou gerenciar suas assinaturas de e-mail.
- Defina um cronograma de atualizações: defina um cronograma para avaliar novos serviços e atributos para a workload.
  - É possível usar o AWS Systems Manager Inventory para coletar metadados de sistema operacional (SO), aplicação e instância das instâncias do Amazon EC2 e entender rapidamente quais instâncias executam o software e as configurações exigidas pela política de software e quais instâncias precisam ser atualizadas.
- Avalie a nova atualização: entenda como atualizar os componentes da sua workload. Aproveite a agilidade da nuvem para testar rapidamente como novos recursos podem melhorar a workload com o intuito de obter eficiência de performance.
- Use automação: use automação no processo de atualização para reduzir o nível de esforço para implantar novos recursos e limitar erros causados por processos manuais.
  - É possível usar CI/CD para atualizar automaticamente AMIs, imagens de contêiner e outros artefatos relacionados à aplicação de nuvem.
  - Você pode usar ferramentas como o AWS Systems Manager Patch Manager para automatizar o processo de atualizações do sistema e agendar a atividade usando as Janelas de Manutenção do AWS Systems Manager.
- Documente o processo: documente seu processo para avaliar atualizações e novos serviços. Forneça aos proprietários o tempo e o espaço necessários para pesquisar, testar, experimentar e validar atualizações e novos serviços. Consulte novamente os KPIs e requisitos de negócios documentados para ajudar a priorizar qual atualização trará um impacto positivo à empresa.

Recursos

Documentos relacionados:

- Blog da AWS
- Novidades da AWS
- Implementar imagens atualizadas com pipelines automatizados do EC2 Image Builder

Vídeos relacionados:

- AWS re:Inforce 2022: Automatizar o gerenciamento e a conformidade de patches usando a AWS
- All Things Patch: AWS Systems Manager | Eventos da AWS

Exemplos relacionados:

- Gerenciamento de inventário e patches
- Workshop One Observability

PERF05-BP07 Revisar as métricas regularmente

Como parte da manutenção de rotina, ou em resposta a eventos ou incidentes, revise quais métricas são coletadas. Use essas análises para identificar quais métricas foram essenciais para resolver problemas e quais métricas adicionais poderiam ajudar a identificar, resolver ou prevenir problemas se estivessem sendo acompanhadas.

Práticas comuns que devem ser evitadas:

- Você permite que as métricas permaneçam em um estado de alarme por um período prolongado.
- Você cria alarmes que não são acionáveis por um sistema de automação.

Benefícios de implementar esta prática recomendada: analise continuamente as métricas que estão sendo coletadas para garantir que identifiquem, resolvam ou evitem problemas corretamente. As métricas também podem se tornar obsoletas se você permitir que elas permaneçam em um estado de alarme por um período prolongado.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Melhore constantemente a coleta e o monitoramento de métricas. Como parte da resposta a incidentes ou eventos, avalie as métricas que foram úteis para resolver o problema e quais poderiam ter ajudado, mas não estão sendo acompanhadas no momento. Use este método para aprimorar a qualidade das métricas coletadas para prevenir ou resolver incidentes futuros mais rapidamente.

Como parte da resposta a incidentes ou eventos, avalie as métricas que foram úteis para resolver o problema e quais poderiam ter ajudado, mas não estão sendo acompanhadas no momento. Use esses dados para aprimorar a qualidade das métricas coletadas para prevenir ou resolver incidentes futuros mais rapidamente.

Etapas de implementação

- Defina métricas: defina métricas críticas de performance para monitorar que estejam alinhadas aos objetivos da sua workload, incluindo métricas como tempo de resposta e utilização de recursos.
- Estabeleça linhas de base: defina uma linha de base e um valor desejável para cada métrica. A linha de base deve fornecer pontos de referência para a identificação de desvios ou anomalias.
- Defina uma frequência: defina uma frequência (como semanal ou mensal) para revisar as métricas essenciais.
- Identifique problemas de performance: durante cada revisão, avalie as tendências e o desvio dos valores base. Procure gargalos ou anomalias de performance. Para os problemas identificados, realize uma análise aprofundada da causa-raiz para entender o principal motivo do problema.
- Identifique ações corretivas: use sua análise para identificar ações corretivas. Isso pode incluir ajuste de parâmetros, correção de bugs e ajustes na escala dos recursos.
- Documente as descobertas: documente suas descobertas, incluindo problemas identificados, causas-raiz e ações corretivas.
- Itere e aprimore: avalie e melhore constantemente o processo de revisão de métricas. Use a lição aprendida com a análise anterior para aprimorar o processo ao longo do tempo.

Recursos

Documentos relacionados:

- Documentação do CloudWatch
- Coletar métricas e logs de instâncias do Amazon EC2 e servidores on-premises com o CloudWatch Agent

- Consultar métricas com o CloudWatch Metrics Insights
- Parceiros de monitoramento, log e performance da AWS Partner Network
- Documentação do X-Ray

Vídeos relacionados:

- AWS re:Invent 2022: Configurar controles em escala em seu ambiente da AWS
- AWS re:Invent 2022: Como a Amazon usa métricas melhores para aprimorar a performance de sites
- AWS re:Invent 2023: Criar uma estratégia efetiva de observabilidade
- AWS Summit SF 2022: Observabilidade full-stack e monitoramento de aplicações com a AWS
- AWS re:Invent 2023: Relaxe: diagnostique e resolva problemas de performance com o Amazon RDS

Exemplos relacionados:

- Criar um painel com o Amazon QuickSight
- Painéis do CloudWatch

Otimização de custo

O pilar Otimização de custos inclui a capacidade de executar sistemas para proporcionar valor comercial pelo menor preço. Recomendações sobre implementação estão disponíveis no whitepaper Pilar Otimização de custos.

Áreas de práticas recomendadas

- Gerenciamento financeiro na nuvem
- Reconhecimento de despesas e usos
- Recursos economicamente eficientes
- Gerenciar recursos de demanda e fornecimento
- Otimização ao longo do tempo

Gerenciamento financeiro na nuvem

Pergunta

- COST 1. Como implementar o gerenciamento financeiro na nuvem?

COST 1. Como implementar o gerenciamento financeiro na nuvem?

A implementação do gerenciamento financeiro na nuvem ajuda as organizações a obterem valor empresarial e sucesso financeiro à medida que otimizam os custos e o uso e escalam na AWS.

Práticas recomendadas

- COST01-BP01 Estabelecer a propriedade da otimização de custos
- COST01-BP02 Estabelecer uma parceria entre finanças e tecnologia
- COST01-BP03 Estabelecer orçamentos e previsões para a nuvem
- COST01-BP04 Implementar a conscientização de custos em seus processos organizacionais
- COST01-BP05 Relatar e notificar sobre a otimização de custos
- COST01-BP06 Monitorar custos proativamente
- COST01-BP07 Manter-se em dia com os novos lançamentos de serviços
- COST01-BP08 Criar uma cultura de conscientização de custos
- COST01-BP09 Quantificar o valor comercial proveniente da otimização de custos

COST01-BP01 Estabelecer a propriedade da otimização de custos

Crie uma equipe (escritório de negócios na nuvem, Centro de Excelência da Nuvem ou FinOps) responsável por estabelecer e manter a conscientização de custos em toda a organização. O responsável pela otimização de custos pode ser uma pessoa ou uma equipe (requer pessoal das equipes de finanças, tecnologia e negócios) que conheça toda a organização e as finanças da nuvem.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Esta é a introdução de uma função ou equipe de escritório de negócios na nuvem (CBO) ou Centro de Excelência da Nuvem (CCOE) responsável por estabelecer e manter uma cultura de conscientização de custos de computação em nuvem. Em toda a organização, essa função pode ser

exercida por qualquer pessoa ou equipe existente, ou por uma nova equipe com as principais partes interessadas em finanças, tecnologia e organização.

A função (individual ou equipe) prioriza e dedica a porcentagem necessária de seu tempo a atividades de gerenciamento e otimização de custos. Para uma organização pequena, a função pode gastar uma porcentagem de tempo menor em comparação com uma função de tempo integral para uma empresa maior.

A função exige uma abordagem multidisciplinar, com recursos de gerenciamento de projetos, ciência de dados, análise financeira e desenvolvimento de software ou infraestrutura. Ela pode melhorar a eficiência da workload realizando otimizações de custos em três propriedades diferentes:

- Centralizada: por meio de equipes designadas, como a equipe FinOps, a equipe de gerenciamento financeiro na nuvem (CFM), o escritório de negócios na nuvem (CBO) ou o Centro de Excelência da Nuvem (CCoE), os clientes podem projetar e implementar mecanismos de governança e promover as práticas recomendadas em toda a empresa.
- Descentralizada: as equipes de tecnologia são convencidas a realizar otimizações de custos.
- Híbrida: combinação de equipes centralizadas e descentralizadas que podem trabalhar em conjunto para realizar otimizações de custo.

A função pode ser medida ao comparar a sua capacidade de realização e entrega com as metas de otimização de custos (por exemplo, métricas de eficiência da workload).

É necessário garantir que haja patrocínio executivo para essa função, o que é um fator de sucesso fundamental. O patrocinador é considerado defensor do consumo de nuvem econômico e oferece suporte ao encaminhamento para a equipe a fim de garantir que as atividades de otimização de custos sejam tratadas de acordo com o nível de prioridade definido pela organização. Caso contrário, a orientação poderá ser ignorada e as oportunidades de redução de custo não serão priorizadas. Juntos, o patrocinador e a equipe ajudam a organização a consumir a nuvem com eficiência e agregar valor comercial.

Se você tem um plano de suporte Business, Enterprise-On-Ramp ou Enterprise Support e precisa de ajuda para elaborar essa equipe ou função, entre em contato com seus especialistas de gerenciamento financeiro na nuvem (CFM) por meio de sua equipe de conta.

Etapas de implementação

- Defina os membros principais: todas as partes relevantes da organização devem contribuir e ter interesse pelo gerenciamento de custos. As equipes comuns dentro das organizações geralmente

incluem: finanças, proprietários de aplicações ou produtos, gerenciamento e equipes técnicas (DevOps). Alguns são contratados em tempo integral (financeiro ou técnico), enquanto outros são contratados periodicamente, conforme necessário. Pessoas ou equipes encarregadas de executar o CFM precisam dos seguintes conjuntos de habilidades:

  - Desenvolvimento de software: quando ocorre o desenvolvimento de scripts e automação.
  - Engenharia de infraestrutura: para implantar scripts, automatizar processos e entender como os serviços e os recursos são provisionados.
  - Perspicácia de operações: CFM é sobre operar na nuvem de maneira eficiente por meio de medição, monitoramento, modificação, planejamento e dimensionamento do uso eficiente da nuvem.

- Defina objetivos e métricas: a função precisa agregar valor à organização de diferentes formas. Esses objetivos são definidos e evoluem continuamente com a organização. As atividades comuns incluem: criação e execução de programas educacionais sobre otimização de custos em toda a organização, desenvolvimento de padrões em toda a organização (como monitoramento e geração de relatórios para otimização de custos) e definição de metas de workload sobre otimização. Essa função também precisa informar regularmente a organização sobre o recurso de otimização de custos.

  Você pode definir indicadores-chave de performance (KPIs) baseados em valor ou custo. Ao definir os KPIs, você pode calcular o custo esperado em termos de eficiência e o resultado comercial esperado. KPIs baseados em valor vinculam métricas de uso e custo a motivadores de valor empresarial e ajudam a racionalizar mudanças em gastos na AWS. O primeiro passo para derivar KPIs baseados em valor é trabalhar em conjunto, em toda a organização, para selecionar e concordar sobre um conjunto padrão de KPIs.
- Estabeleça uma cadência regular: o grupo (equipes financeira, empresarial e de tecnologia) devem se reunir regularmente para analisar metas e métricas. Um ritmo típico envolve analisar o estado da organização, todos os programas em execução no momento e as métricas financeiras e de otimização gerais. Em seguida, as principais workloads são relatadas em mais detalhes.

  Durante essas revisões regulares, é possível analisar a eficiência (custo) da workload e o resultado empresarial. Por exemplo, um aumento de 20% no custo de uma workload pode ser consequência de um aumento do uso pelos clientes. Neste caso, esse aumento de 20% no custo pode ser interpretado como um investimento. Essas chamadas regulares podem ajudar as equipes a identificar KPIs de valor que ofereçam propósito para toda a organização.

Recursos

Documentos relacionados:

- Blog de CCoE da AWS
- Criar um escritório de negócios na nuvem
- CCoE: Centro de Excelência da Nuvem

Vídeos relacionados:

- História de sucesso de CCoE de vanguarda

Exemplos relacionados:

- Usar um Centro de Excelência da Nuvem (CCoE) para transformar toda a empresa
- Criar um CCoE para transformar toda a empresa
- Sete obstáculos que devem ser evitados ao criar um CCoE

COST01-BP02 Estabelecer uma parceria entre finanças e tecnologia

Envolva equipes financeiras e de tecnologia em discussões sobre custo e uso em todas as etapas da jornada para a nuvem. As equipes se reúnem e discutem regularmente assuntos como objetivos e metas organizacionais, o estado atual de custo e uso e práticas financeiras e contábeis.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

As equipes de tecnologia inovam mais rapidamente na nuvem devido à redução dos ciclos de implantação de aprovação, aquisição e infraestrutura. Isso pode ser um ajuste para organizações financeiras anteriormente usadas para executar processos demorados e com uso intensivo de recursos para aquisição e implantação de capital em ambientes de datacenter on-premises, além de alocação de custos apenas na aprovação do projeto.

Do ponto de vista da organização financeira e de aquisição, o processo de definição orçamentária, solicitações de capital, aprovações, aquisição e instalação de infraestrutura física é algo que levou décadas para ser aprendido e padronizado:

- Equipes de engenharia ou TI costumam ser os solicitantes
- Equipes financeiras diversas atuam como aprovadores e compradores
- Equipes de operação estendem, acumulam e disponibilizam infraestrutura pronta para ser usada

Com a adoção da nuvem, a aquisição e o consumo de infraestrutura deixaram de estar vinculados a uma série de dependências. No modelo de nuvem, as equipes de tecnologia e produto deixam de ser simples desenvolvedoras, passando a ser operadoras e proprietárias de seus produtos, além de responsáveis pela maioria das atividades historicamente associadas às equipes financeiras e de operações, incluindo aquisição e implantação.

Basta uma conta e o conjunto adequado de permissões para provisionar recursos na nuvem. Também é isso que reduz o risco financeiro e de TI, o que significa que as equipes estão sempre a poucos cliques ou chamadas de API de encerrar recursos ociosos ou desnecessários na nuvem. Também é isso que permite que as equipes de tecnologia inovem com mais rapidez: a agilidade e capacidade de aplicar e derrubar experimentos. Embora a natureza variável do consumo na nuvem possa afetar a previsibilidade do ponto de vista de previsão e definição orçamentária, a nuvem oferece às organizações a capacidade de reduzir o custo de provisionamento em excesso, além de reduzir o custo de oportunidade associado ao subprovisionamento conservador.

Estabelecer uma parceria entre as principais partes interessadas em finanças e tecnologia para criar uma compreensão compartilhada dos objetivos organizacionais e desenvolver mecanismos para obter sucesso financeiro no modelo de gastos variáveis da computação em nuvem. As equipes relevantes da sua organização devem estar envolvidas em discussões de custo e uso em todas as fases da jornada para a nuvem, incluindo:

- Líderes financeiros: CFOs, controladores financeiros, planejadores financeiros, analistas de negócios, aquisições, sourcing e contas a pagar devem compreender o modelo de nuvem de consumo, as opções de compra e o processo de faturamento mensal. O departamento financeiro precisa se unir às equipes de tecnologia para criar e socializar uma narrativa de valor de TI, ajudando as equipes comerciais a entender como o gasto com tecnologia está associado aos resultados comerciais. Assim, as despesas com tecnologia são vistas não como custos, mas como investimentos. Devido às diferenças fundamentais entre a nuvem (como a taxa de alteração no uso, definição de preço com pagamento conforme o uso, definição de preço em camadas, modelos de definição de preço e informações detalhadas de faturamento e uso) em comparação à operação on-premises, é essencial que a organização financeira entenda como o uso da nuvem pode afetar aspectos empresariais, incluindo processos de aquisição, rastreamento de incentivos, alocação de custos e demonstrações financeiras.
- Líderes de tecnologia: os líderes de tecnologia (incluindo proprietários de produtos e aplicações) devem estar cientes dos requisitos financeiros (por exemplo, restrições orçamentárias), bem como dos requisitos de negócios (por exemplo, contratos de nível de serviço). Isso permite que a workload seja implementado para atingir os objetivos desejados da organização.

A parceria entre finanças e tecnologia oferece os seguintes benefícios:

- As equipes de finanças e tecnologia têm visibilidade praticamente em tempo real dos custos e do uso.
- As equipes de finanças e tecnologia estabelecem um procedimento operacional padrão para lidar com a variação de gastos na nuvem.
- As partes interessadas em finanças atuam como consultores estratégicos com relação à forma como o capital é usado para comprar descontos de compromissos (por exemplo, instâncias reservadas ou Savings Plans da AWS) e como a nuvem é usada para expandir a organização.
- Contas a pagar e processos de aquisição existentes são usados com a nuvem.
- As equipes de finanças e tecnologia colaboram na previsão de custos e uso futuros da AWS para alinhar e criar orçamentos organizacionais.
- Melhor comunicação entre organizações por meio de uma linguagem compartilhada e entendimento comum dos conceitos financeiros.

As partes interessadas adicionais dentro da sua organização que devem ser envolvidas em discussões de custo e uso incluem:

- Proprietários de unidades de negócios: os proprietários de unidades de negócios devem compreender o modelo de negócios de nuvem para que possam fornecer orientações tanto para as unidades de negócios quanto para toda a empresa. Esse conhecimento de nuvem é essencial quando há necessidade de prever o crescimento e o uso da workload, e ao avaliar opções de compra de longo prazo, como instâncias reservadas ou Savings Plans.
- Equipe de engenharia: uma parceria entre as equipes financeira e de tecnologia é essencial para o desenvolvimento de uma cultura de consciência dos custos que encoraja os engenheiros a agirem em relação ao gerenciamento financeiro na nuvem (CFM). Um dos problemas comuns dos profissionais de CFM ou operações financeiras e das equipes financeiras é fazer com que os engenheiros entendam todos os negócios na nuvem, sigam as práticas recomendadas e adotem as medidas recomendadas.
- Terceiros: se sua organização usa terceiros (por exemplo, consultores ou ferramentas), certifique-se de que eles estejam alinhados aos seus objetivos financeiros e possam demonstrar o alinhamento por meio de seus modelos de engajamento e um retorno sobre o investimento (ROI). Terceiros normalmente contribuirão para o relatório e a análise de qualquer workload que gerenciem e fornecerão análise de custo de qualquer workload que projetem.

Implementar o CFM e obter sucesso requer a colaboração das equipes financeira, comercial e de tecnologia, além de uma mudança na forma como os gastos com nuvem são comunicados e avaliados em toda a organização. Inclua as equipes de engenharia para que façam parte dessas conversas sobre custos e uso em todos os estágios, incentivando-as a seguir as práticas recomendadas e tomar medidas previamente acordadas conforme for apropriado.

Etapas de implementação

- Defina os membros importantes: verifique se todos os membros relevantes de suas equipes de finanças e tecnologia participam da parceria. Os membros financeiros relevantes serão aqueles que interagem com a conta da nuvem. Normalmente serão CFOs, controladores financeiros, planejadores financeiros, analistas de negócios, compras e sourcing. Normalmente, os membros de tecnologia serão proprietários de produtos e aplicações, gerentes técnicos e representantes de todas as equipes que criam na nuvem. Outros membros podem incluir proprietários de unidades de negócios, como marketing, que influenciarão o uso de produtos, e terceiros, como consultores para alcançar o alinhamento com seus objetivos e mecanismos e para auxiliar na geração de relatórios.
- Defina tópicos para discussão: defina os tópicos que são comuns entre as equipes ou que precisarão de um entendimento compartilhado. Siga o custo a partir do momento em que ele

é criado até que a fatura seja paga. Observe todos os membros envolvidos e os processos organizacionais que devem ser aplicados. Compreenda cada etapa ou processo que ele atravessa e as informações associadas, como modelos de preços disponíveis, preços em camadas, modelos de desconto, orçamento e requisitos financeiros.

- Estabeleça uma cadência regular: para criar uma parceria financeira e tecnológica, estabeleça uma comunicação regular para criar e manter o alinhamento. O grupo precisa se reunir regularmente para comparar metas e métricas. Um ritmo típico envolve analisar o estado da organização, todos os programas em execução no momento e as métricas financeiras e de otimização gerais. Em seguida, as workloads principais são relatadas em mais detalhes.

Recursos

Documentos relacionados:

- Notícias do blog da AWS

COST01-BP03 Estabelecer orçamentos e previsões para a nuvem

Ajuste os processos de previsão e orçamento organizacional existentes para que sejam compatíveis com a natureza altamente variável dos custos e uso da nuvem. Os processos devem ser dinâmicos, usando algoritmos baseados em tendências ou em direcionadores de negócios, ou uma combinação de ambos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Nas configurações tradicionais de TI on-premises, os clientes geralmente enfrentam o desafio de planejar custos fixos que mudam apenas ocasionalmente, em geral com a compra de novos serviços e hardware de TI para atender aos picos de demanda. Em contrapartida, a Nuvem AWS adota uma abordagem diferente na qual os clientes pagam pelos recursos que usam, conforme determinado pelas reais necessidades comerciais e de TI. No ambiente de nuvem, a demanda pode variar mensalmente, diariamente ou até mesmo de hora em hora.

O uso da nuvem traz eficiência, velocidade e agilidade, o que ocasiona um padrão de custo e de uso altamente variável. Os custos podem diminuir ou, às vezes, aumentar em resposta à maior eficiência da workload ou à implantação de novas workloads e de recursos. À medida que as workloads escalam para atender a uma base de clientes em expansão, o uso e os custos da nuvem aumentam

em igual proporção em virtude da maior acessibilidade dos recursos. Essa flexibilidade nos serviços de nuvem estende-se aos custos e às previsões, criando um grau de elasticidade.

É essencial haver um estreito alinhamento com essas mudanças nas necessidades comerciais e nos motivadores de demanda e buscar o planejamento mais preciso possível. Os processos de orçamento organizacional tradicionais precisam se adaptar para acomodar essa variabilidade.

Pense na modelagem de custos ao prever o custo de novas workloads. A modelagem de custos cria uma compreensão básica dos custos esperados da nuvem, o que ajuda você a determinar o custo total de propriedade (TCO), o retorno sobre o investimento (ROI) e outras análises financeiras, definir metas e expectativas com as partes interessadas e identificar oportunidades de otimização de custos.

Sua organização precisa entender as definições de custo e os agrupamentos aceitos. O nível de detalhe no qual você prevê pode variar com base na estrutura e nos fluxos de trabalho internos da organização. Selecione um nível de granularidade que atenda aos requisitos específicos e à configuração organizacional. É importante entender em que nível a previsão é realizada:

- Nível de conta de gerenciamento ou AWS Organizations: a conta de gerenciamento é a conta que você usa para criar AWS Organizations. Por padrão, as organizações têm uma conta de gerenciamento.
- Conta-membro ou vinculada: uma conta no Organizations é uma Conta da AWS padrão que contém os seus recursos da AWS e as identidades que podem acessar esses recursos.
- Ambiente: um ambiente é uma coleção de recursos da AWS que executam uma versão da aplicação. Um ambiente pode ser criado com várias contas-membro ou vinculadas.
- Projeto: um projeto é uma combinação de objetivos ou tarefas definidas a serem realizadas dentro de um período fixo. É importante pensar no ciclo de vida do projeto durante a previsão.
- Serviços da AWS: grupos ou categorias, como serviços de computação ou armazenamento, nos quais você pode agrupar serviços da AWS de acordo com sua previsão.
- Agrupamento personalizado: é possível criar grupos personalizados com base nas necessidades da sua organização, como unidades de negócios, centros de custo, equipes, tags de alocação de custos, categorias de custo, contas vinculadas ou uma combinação delas.

Identifique os motivadores empresariais que podem afetar o custo do uso e faça uma previsão para cada um deles separadamente a fim de calcular o uso esperado com antecedência. Alguns dos motivadores podem estar vinculados às equipes de TI e de produtos da organização. Outros

motivadores empresariais, como eventos de marketing, promoções, expansões geográficas, fusões e aquisições, são conhecidos por seus líderes de vendas, de marketing e de negócios, e é importante colaborar e pensar também em todos esses motivadores de demanda.

É possível usar o AWS Cost Explorer para fazer previsões baseadas em tendências em um período futuro definido com base no gasto no passado. O mecanismo de previsão do AWS Cost Explorer segmenta os dados históricos com base em tipos de cobrança (por exemplo, instâncias reservadas) e usa uma combinação de machine learning e modelos baseados em regras com a finalidade de prever os gastos individualmente para todos os tipos de cobrança.

Depois de estabelecer seu processo de previsão e criar modelos, você poderá usar o AWS Budgets para definir orçamentos personalizados em um nível granular especificando o período, a recorrência ou o valor (fixo ou variável) e adicionando filtros como serviço, Região da AWS e tags. Geralmente, o orçamento é preparado para um único ano e permanece fixo, o que exige adesão estrita de todos os envolvidos. Entretanto, a previsão é mais flexível, permitindo reajustes ao longo do ano e fornecendo projeções dinâmicas em um período de um, dois ou três anos. Tanto o orçamento quanto as previsões desempenham um papel fundamental para estabelecer expectativas financeiras entre várias partes interessadas em tecnologia e negócios. A precisão da previsão e implementação também impõe responsabilidade às partes interessadas que já são diretamente responsáveis pelo custo de provisionamento, o que também pode contribuir para o reconhecimento geral de custos.

Para se informar sobre a performance dos orçamentos atuais, é possível criar e programar relatórios do AWS Budgets para serem enviados por e-mail a você e às respectivas partes interessadas regularmente. Também é possível pode criar alertas do AWS Budgets com base nos custos reais, cuja natureza é reativa, ou com base nos custos previstos, que oferecem tempo para mitigar possíveis excessos de custos. Você pode receber um alerta quando o custo ou o uso realmente excederem determinado nível ou se houver previsão de que eles excederão o valor orçado.

Ajuste os processos existentes de orçamento e de previsão para se tornarem mais dinâmicos por meio de algoritmos baseados em tendências (com custos históricos como entradas) e algoritmos baseados em motivadores (por exemplo, lançamentos de novos produtos, expansão regional ou novos ambientes para workloads), que são ideais para um ambiente de gastos dinâmico e variável. Depois de determinar sua previsão baseada em tendências usando o Explorador de Custos ou qualquer outra ferramenta, use o AWS Pricing Calculator para estimar seu caso de uso da AWS e os custos futuros com base no uso esperado (tráfego, solicitações por segundo ou instâncias necessárias do Amazon EC2).

Monitore o quanto precisa é essa previsão, pois os orçamentos devem ser definidos com base nesses cálculos e estimativas. Monitore a precisão e eficácia das previsões de custos de nuvem

integradas. Analise regularmente os gastos reais em comparação com a previsão e ajuste conforme necessário para torná-la mais precisa. Monitore a variação da previsão e analise a causa-raiz da variação relatada para agir e ajustar as previsões.

Conforme mencionado em COST01-BP02 Estabelecer uma parceria entre finanças e tecnologia, é importante estimular parceria e ritmo entre TI, departamento financeiro e outras partes interessadas para verificar se todos usam as mesmas ferramentas e processos para manter a consistência. Nas situações em que os orçamentos precisam sofrer alterações, aumentar o ritmo dos pontos de contato pode ajudar na hora de reagir a essas mudanças com maior rapidez.

Etapas de implementação

- Defina a linguagem de custo dentro da organização: Crie uma linguagem de custo comum da AWS dentro da organização com várias dimensões e agrupamentos. Verifique se as partes interessadas entendem a granularidade das previsões, os modelos de preços e o nível de suas previsões de custos.
- Analise previsões baseadas em tendências: use ferramentas de previsão baseadas em tendências, como o AWS Cost Explorer e o Amazon Forecast. Analise o custo de uso em diferentes dimensões, como serviço, conta, tags e categorias de custos. Se for necessária uma previsão avançada, importe os dados de Custos e Uso (CUR) da AWS para o Amazon Forecast (o qual aplica a regressão linear como uma forma de machine learning para que seja feita a previsão).
- Previsões baseadas em motivadores: identifique o impacto dos motivadores empresariais no uso da nuvem e faça uma previsão para cada um deles separadamente a fim de calcular o custo de uso esperado com antecedência. Trabalhe em estreita colaboração com proprietários de unidades de negócios e partes interessadas a fim de entender o impacto sobre os novos motivadores e calcular as mudanças de custo esperadas para definir orçamentos precisos.
- Atualize os processos existentes de previsão e orçamento: usando como referência métodos de previsão adotados, como baseados em tendências, baseados em motivadores de negócios ou uma combinação de ambos os métodos de previsão, defina seus processos de previsão e orçamento. Os orçamentos devem ser calculados, realistas e baseados em suas previsões.
- Configure alertas e notificações: use alertas do AWS Budgets e a detecção de anomalias em custos para receber alertas e notificações.
- Faça revisões periódicas com as principais partes interessadas: por exemplo, alinhe-se em relação às mudanças na direção e no uso dos negócios com as partes interessadas em TI, finanças, equipes de plataforma e outras áreas de negócios.

Recursos

Documentos relacionados:

- AWS Cost Explorer
- AWS Cost and Usage Report
- Fazer previsões com o Explorador de Custos
- Fazer previsões com o Amazon QuickSight
- Amazon Forecast
- AWS Budgets

Vídeos relacionados:

- Como posso usar o AWS Budgets para monitorar meus gastos e uso?
- Série de otimização de custos da AWS: AWS Budgets

Exemplos relacionados:

- Entender e criar previsões baseadas em motivadores
- Como estabelecer e impulsionar uma cultura de previsão
- Como melhorar sua previsão de custos na nuvem
- Usar as ferramentas certas para prever custos na nuvem

COST01-BP04 Implementar a conscientização de custos em seus processos organizacionais

Implemente a conscientização de custos, crie transparência e contabilize os custos em processos novos ou existentes que afetem o uso e aproveite os processos existentes para conscientização de custos. Implemente a conscientização de custos no treinamento de funcionários.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

A conscientização de custos deve ser implementado em processos organizacionais novos e existentes. Trata-se de um dos recursos fundamentais para outras práticas recomendadas. Recomendamos reutilizar e modificar processos existentes sempre que possível, o que minimiza

o impacto na agilidade e velocidade. Informe os custos da nuvem para as equipes de tecnologia e os responsáveis por decisões nas equipes financeira e comercial para conscientizar sobre os custos, e estabeleça indicadores-chave de performance (KPIs) para as partes interessadas dos departamentos financeiro e comercial. As recomendações a seguir ajudarão a implementar a conscientização de custos em sua workload:

- Verifique se o gerenciamento de alterações inclui uma medição de custo para quantificar o impacto financeiro das mudanças. Isso ajuda a abordar de forma proativa as preocupações relacionadas a custos e a destacar as reduções de custos.
- Verifique se a otimização de custos é um componente essencial de seus recursos operacionais. Por exemplo, você pode aproveitar os processos existentes de gerenciamento de incidentes para investigar e identificar causas-raiz das anomalias de custo e uso ou excessos de custo.
- Acelere a redução de custos e a obtenção de valor empresarial por meio da automação ou de ferramentas. Ao pensar sobre o custo da implementação, enquadre a conversa para incluir um componente de retorno sobre o investimento (ROI) para justificar o investimento de tempo ou dinheiro.
- Aloque os custos de nuvem implementando showbacks ou chargebacks de gastos na nuvem, incluindo gastos com opções de compra baseadas em compromissos, serviços compartilhados e compras de marketplace para impulsionar um consumo da nuvem mais consciente sobre custos.
- Estenda os programas de treinamento e desenvolvimento existentes para incluir treinamento com conscientização de custos em toda a organização. Recomendamos que isso inclua treinamento e certificação contínuos. Isso criará uma organização capaz de autogerenciar custos e uso.
- Aproveite ao máximo as ferramentas nativas e gratuitas da AWS, como AWS Cost Anomaly Detection, AWS Budgets e AWS Budgets Reports.

Quando as organizações adotam consistentemente as práticas de gerenciamento financeiro na nuvem (CFM), esses comportamentos se tornam enraizados na forma de trabalhar e tomar decisões. O resultado é uma cultura mais consciente em relação aos custos, desde os desenvolvedores que arquitetam uma nova aplicação concebida na nuvem até gerentes financeiros que analisam o ROI desses novos investimentos na nuvem.

Etapas de implementação

- Identifique os processos organizacionais relevantes: cada unidade organizacional analisa os processos que possui e identifica aqueles que afetam o custo e o uso. Todos os processos que resultam na criação ou no encerramento de um recurso precisam ser incluídos para análise.

Procure processos que possam sustentar a conscientização de custos na empresa, como gerenciamento de incidentes e treinamento.

- Estabeleça uma cultura com consciência de custos e autossustentável: garanta que todas as partes interessadas relevantes se alinhem ao motivo da mudança e impacto como custo para que entendam os custos da nuvem. Isso permitirá que sua organização estabeleça uma cultura de inovação autossustentável com conscientização de custos.
- Atualize os processos com conscientização de custos: cada processo é modificado para se tornar consciente dos custos. O processo pode exigir pré-verificações adicionais, como avaliação do impacto do custo, ou pós-verificações que validam se as mudanças esperadas no custo e no uso ocorreram. Processos de suporte, como treinamento e gerenciamento de incidentes, podem ser estendidos para incluir itens de custo e uso.

Para obter ajuda, fale com especialistas em CFM por meio de sua equipe de conta, ou explore os recursos e os documentos relacionados abaixo.

Recursos

Documentos relacionados:

- Gerenciamento financeiro na Nuvem AWS

Exemplos relacionados:

- Estratégia para o gerenciamento eficiente dos custos na nuvem
- Série de blogs sobre controle de custos n.º 3: Como lidar com o impacto dos custos
- Um guia de introdução ao AWS Cost Management

COST01-BP05 Relatar e notificar sobre a otimização de custos

Configure orçamentos de nuvem e mecanismos para detectar anomalias no uso. Configure ferramentas relacionadas para alertas de custo e uso em relação a metas predefinidas e receba notificações quando algum uso exceder essas metas. Faça reuniões regulares para analisar a relação custo-benefício das workloads e promover a conscientização de custos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Informe regularmente sobre a otimização de custos e os usos dentro da sua organização. Implemente sessões dedicadas para discutir a relação de custo/performance ou inclua a otimização de custos em seus ciclos regulares de relatórios operacionais para as workloads. Use serviços e ferramentas para monitorar a relação de custo/performance regularmente e implementar oportunidades de redução de custos.

Visualize seus custos e o uso com diversos filtros e granularidades usando o AWS Cost Explorer, que fornece painéis e relatórios, como custos por serviço ou por conta, custos diários ou custos de mercado. Acompanhe o andamento do custo e do uso em relação aos orçamentos configurados com o AWS Budgets Reports.

Use o AWS Budgets para definir orçamentos personalizados para monitorar seus custos e uso e responder rapidamente aos alertas recebidos por e-mail ou notificações do Amazon Simple Notification Service (Amazon SNS) se você exceder seu limite. Defina seu período orçamentário preferido como diário, mensal, trimestral ou anual e crie limites orçamentários específicos para se manter informado sobre como os custos e o uso reais ou previstos progridem em relação ao limite do orçamento. Também é possível configurar alertas e ações em resposta a esses alertas para que sejam executados automaticamente ou por meio de um processo de aprovação quando uma meta de orçamento é excedida.

Implemente notificações sobre custo e uso para garantir que as mudanças no custo e no uso possam ser corrigidas rapidamente se forem inesperadas. O AWS Cost Anomaly Detection permite que você reduza as surpresas de custo e aprimore o controle sem retardar a inovação. O AWS Cost Anomaly Detection identifica gastos anômalos e causas-raiz, o que ajuda a reduzir o risco de surpresas no faturamento. Com três etapas simples, você pode criar seu próprio monitor contextualizado e receber alertas sempre que um gasto anormal for detectado.

Você também pode usar o Amazon QuickSight com dados do AWS Cost and Usage Report CUR) para fornecer relatórios altamente personalizados com dados mais granulares. O Amazon QuickSight permite agendar relatórios e receber e-mails periódicos de relatórios de custos com informações sobre custos e usos históricos ou oportunidades de redução de custos. Confira nossa solução Cost Intelligence Dashboard (CID) criada no Amazon QuickSight para oferecer visibilidade avançada.

Use o AWS Trusted Advisor, que oferece orientação para verificar se os recursos provisionados se alinham às práticas recomendadas da AWS para otimização de custos.

Verifique as recomendações de Savings Plans por meio de grafos visuais em comparação com o custo e uso detalhados. Os grafos por hora mostram os gastos sob demanda com o compromisso

recomendado dos Savings Plans, fornecendo informações sobre economias estimadas, cobertura dos Savings Plans e utilização dos Savings Plans Isso ajuda as organizações a entender como os Savings Plans se aplicam a cada hora de gasto sem precisar investir tempo e recursos na criação de modelos para analisar as despesas.

Crie periodicamente relatórios que contêm um destaque de Savings Plans, instâncias reservadas e recomendações do AWS Cost Explorer para dimensionamento do Amazon EC2 para começar a reduzir o custo associado a workloads estacionárias e recursos ociosos ou subutilizados. Identifique e recupere os gastos associados ao desperdício de recursos implantados na nuvem. O desperdício na nuvem ocorre quando recursos dimensionados incorretamente são criados ou quando se observa padrões de uso diferentes do esperado. Siga as práticas recomendadas da AWS para reduzir o desperdício ou peça ajuda à equipe de contas e parceiro para otimizar e economizar seus gastos na nuvem.

Gere relatórios regularmente para melhorar as opções de compra de recursos a fim de reduzir os custos unitários das workloads. Opções de compra como Savings Plans, instâncias reservadas ou instâncias spot do Amazon EC2 oferecem as maiores economias para workloads tolerantes a falhas e permitem que as partes interessadas (proprietários de negócios e equipes financeiras e de tecnologia) façam parte das conversas sobre comprometimento.

Compartilhe os relatórios que contêm oportunidades ou anúncios de novos lançamentos que possam ajudar você a reduzir o custo total de propriedade (TCO) da nuvem. Adote novos serviços, regiões, recursos, soluções ou maneiras de obter mais reduções de custo.

Etapas de implementação

- Configure o AWS Budgets: configure o AWS Budgets em todas as contas para a sua workload. Defina um orçamento para o gasto total da conta e outro para a workload usando tags.
  - Laboratórios do Well-Architected: Governança de custos e uso
- Informe sobre a otimização de custos: configure um ciclo regular para discutir e analisar a eficiência da workload. Usando as métricas estabelecidas, informe sobre as métricas obtidas e o custo para alcançá-las. Identifique e corrija tendências negativas, bem como tendências positivas que possam ser promovidas em toda a organização. Os relatórios devem envolver representantes das equipes e proprietários de aplicações, bem como do setor financeiro, e os principais tomadores de decisão sobre as despesas com a nuvem.

Recursos

Documentos relacionados:

- AWS Cost Explorer
- AWS Trusted Advisor
- AWS Budgets
- AWS Cost and Usage Report
- Práticas recomendadas do AWS Budgets.
- Análise do Amazon S3

Exemplos relacionados:

- Laboratórios do Well-Architected: Governança de custos e uso
- Principais formas de começar a otimizar seus custos de nuvem da AWS

COST01-BP06 Monitorar custos proativamente

Implemente ferramentas e painéis para monitorar os custos proativamente para a workload. Revise regularmente os custos com ferramentas configuradas ou prontas para usar em vez de apenas analisar os custos e as categorias quando receber notificações. O monitoramento e a análise proativa dos custos ajuda a identificar tendências positivas e permite que você as promova em toda a organização.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Recomenda-se monitorar custos e uso proativamente em sua organização, e não apenas quando há exceções ou anomalias. Painéis altamente visíveis em todo o escritório ou ambiente de trabalho garantem que as pessoas-chave tenham acesso às informações necessárias e indicam o foco da organização na otimização de custos. Os painéis visíveis permitem promover ativamente resultados bem-sucedidos e implementá-los em toda a organização.

Crie uma rotina diária ou frequente para usar o AWS Cost Explorer ou qualquer outro painel, como o Amazon QuickSight, para ver os custos e fazer análises proativas. Analise o uso e os custos dos serviços da AWS na conta da AWS, no nível da workload ou em um serviço específico da AWS com

agrupamento e filtragem, e valide se estão dentro do esperado ou não. Use a granularidade no nível de hora e recurso e as tags para filtrar e identificar os custos incorridos para os principais recursos. Você também pode criar seus próprios relatórios com o Cost Intelligence Dashboard, uma solução do Amazon QuickSight criada por arquitetos de soluções da AWS, e comparar seus orçamentos com o custo e o uso reais.

Etapas de implementação

- Informe sobre a otimização de custos: configure um ciclo regular para discutir e analisar a eficiência da workload. Usando as métricas estabelecidas, informe sobre as métricas obtidas e o custo para alcançá-las. Identifique e corrija quaisquer tendências negativas e identifique tendências positivas a serem promovidas em toda a organização. Os relatórios devem envolver representantes das equipes de aplicações e dos proprietários, de finanças e da gerência.
- Crie e ative a granularidade diária do AWS Budgets para o custo e o uso para adotar medidas rápidas para evitar possíveis excessos de custos: o AWS Budgets permite configurar notificações de alerta para que você fique informado se algum dos seus tipos de orçamento estiver fora dos limites pré-configurados. A melhor forma de aproveitar o AWS Budgets é definir o custo e o uso esperados como limites, para que qualquer coisa acima do seu orçamento seja considerada excesso.
- Crie o AWS Cost Anomaly Detection para monitor de custos: o AWS Cost Anomaly Detection usa tecnologia avançada de machine learning para identificar gastos anormais e causas-raiz para que você possa agir rapidamente. Ele permite configurar monitores de custo que definem os segmentos de gastos que você deseja avaliar (por exemplo, serviços individuais da AWS, contas-membro, tags de alocação de custo e categorias de custo) e permite que você defina quando, onde e como recebe notificações de alerta. Para cada monitor, anexe várias assinaturas de alertas para proprietários de negócios e equipes de tecnologia, incluindo um nome, um limite de impacto do custo e a frequência de alerta (alertas individuais, resumo diário, resumo semanal) para cada assinatura.
- Use o AWS Cost Explorer ou integre seus dados do AWS Cost and Usage Report (CUR) aos painéis do Amazon QuickSight para visualizar os custos da sua organização: o AWS Cost Explorer tem uma interface fácil de usar que permite visualizar, entender e gerenciar seus custos e uso da AWS ao longo do tempo. O Cost Intelligence Dashboard é um painel personalizável e acessível que ajuda a criar a base de sua própria ferramenta de gerenciamento e otimização dos custos.

Recursos

Documentos relacionados:

- AWS Budgets
- AWS Cost Explorer
- Orçamentos de custos e uso diários
- AWS Cost Anomaly Detection

Exemplos relacionados:

- Laboratórios do Well-Architected: Visualização
- Laboratórios do Well-Architected: Visualização avançada
- Laboratórios do Well-Architected: Cloud Intelligence Dashboards
- Laboratórios do Well-Architected: Visualização de custos
- Alerta do AWS Cost Anomaly Detection com Slack

COST01-BP07 Manter-se em dia com os novos lançamentos de serviços

Consulte regularmente especialistas ou parceiros da AWS para considerar quais serviços e recursos oferecem menor custo. Revise os blogs da AWS e outras fontes de informação.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A AWS adiciona novos recursos constantemente para que você possa aproveitar as tecnologias mais recentes a fim de experimentar e inovar com maior rapidez. Você pode implementar novos serviços e recursos da AWS para aumentar a eficiência de custos na workload. Revise regularmente o Gerenciamento de Custos da AWS, o Blog de notícias da AWS, o Blog de gerenciamento de custos da AWS e as Novidades da AWS para obter informações sobre lançamentos de novos serviços e recursos. As postagens Novidades oferecem uma breve visão geral de todos os anúncios de serviços, recursos e expansões de regiões da AWS à medida que são lançados.

Etapas de implementação

- Inscreva-se em blogs: acesse as páginas de blogs da AWS e inscreva-se em Novidades e em outros blogs relevantes. Você pode se inscrever na página de preferências de comunicação com seu endereço de e-mail.

- Inscreva-se em Notícias da AWS: revise regularmente o Blog de notícias da AWS e Novidades da AWS para obter informações sobre novos lançamentos de serviços e recursos. Assine o feed RSS, ou use seu e-mail para ficar por dentro dos anúncios e lançamentos.
- Acompanhe as reduções de preços da AWS: cortes regulares nos preços de todos os nossos serviços são uma prática padrão que a AWS usa para passar os benefícios econômicos obtidos pela nossa escala aos clientes. Em 20 de setembro de 2023, a AWS já havia reduzido os preços 134 vezes desde 2006. Se você tiver qualquer decisão comercial pendente por motivos de preço, poderá reavaliá-la depois das reduções de preços e das novas integrações de serviços. Aprenda sobre os esforços anteriores de redução de preços, incluindo instâncias do Amazon Elastic Compute Cloud (Amazon EC2), na categoria de redução de preços do Blog de notícias da AWS.
- Eventos e reuniões da AWS: participe da conferência local da AWS e de qualquer reunião local com outras organizações da área. Se não puder participar presencialmente, tente participar dos eventos virtuais para ouvir mais de especialistas da AWS e casos de negócios de outros clientes.
- Reuna-se com sua equipe de conta: programe um ritmo regular com a equipe de conta, encontre-se com ela e discuta as tendências do setor e os serviços da AWS. Fale com seu gerente de conta, o arquiteto de soluções e a equipe de suporte.

Recursos

Documentos relacionados:

- Gerenciamento de custos na AWS
- Novidades da AWS
- Notícias do blog da AWS

Exemplos relacionados:

- Amazon EC2: 15 anos otimizando e reduzindo custos de TI
- Blog de notícias da AWS: redução de preços

COST01-BP08 Criar uma cultura de conscientização de custos

Implemente mudanças ou programas em toda a organização para criar uma cultura de conscientização de custos. Recomenda-se começar aos poucos e, à medida que seus recursos aumentarem e o uso da nuvem por sua organização crescer, implementar programas grandes e abrangentes.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Uma cultura de conscientização de custos permite escalar a otimização de custos e o gerenciamento financeiro na nuvem (operações financeiras, centro de excelência da nuvem, equipes de operações na nuvem e assim por diante) por meio de práticas recomendadas executadas de maneira orgânica e descentralizada em toda a organização. A conscientização de custos permite que você crie altos níveis de capacidade em toda a organização com o mínimo de esforço em comparação com uma abordagem centralizada e de cima para baixo.

Provocar a conscientização de custos em computação em nuvem, principalmente para geradores de custos primários na computação em nuvem, permite que as equipes entendam os resultados esperados de quaisquer alterações na perspectiva de custos. As equipes que acessam os ambientes de nuvem devem conhecer os modelos de preços e a diferença entre datacenters on-premises tradicionais e computação em nuvem.

O principal benefício de uma cultura de conscientização de custos é que as equipes de tecnologia otimizam os custos de maneira proativa e contínua (por exemplo, são consideradas um requisito não funcional ao arquitetar novas workloads ou alterar workloads existentes) em vez de realizarem otimizações de custo reativas somente quando necessárias.

Pequenas mudanças na cultura podem ter grandes impactos na eficiência de suas workloads atuais e futuras. Exemplos incluem:

- Oferecer visibilidade e conscientizar as equipes de engenharia para que entendam o que fazem e qual seu impacto em termos de custo.
- Gamificação do custo e uso em toda a organização. Isso pode ser feito por meio de um painel visível publicamente ou de um relatório que compara custos e uso normalizados entre equipes (por exemplo, custo por workload e custo por transação).
- Reconhecimento da eficiência de custos. Recompense realizações de otimização de custos voluntárias ou não solicitadas publicamente ou de forma privada e aprenda com os erros para evitar repeti-los no futuro.
- Criar requisitos organizacionais de cima para baixo para workloads a serem executadas em orçamentos predefinidos.
- Questionar os requisitos comerciais das mudanças e o impacto sobre os custos das mudanças solicitadas na infraestrutura de arquitetura ou configuração de workload para garantir que você pague somente o necessário.

- Garantir que o planejador das mudanças esteja ciente das mudanças esperadas que impactam o custo e que elas sejam confirmadas pelas partes interessadas para que proporcionem resultados comerciais com economia.

Etapas de implementação

- Informe os custos da nuvem às equipes de tecnologia: para conscientizar sobre os custos e estabelecer KPIs de eficiência para partes interessadas financeiras e comerciais.
- Informe as partes interessadas ou membros da equipe sobre mudanças planejadas: crie um item na agenda para discutir mudanças planejadas e o impacto de custo-benefício sobre a workload durante as reuniões semanais de mudanças.
- Reuna-se com sua equipe de conta: defina um cronograma de reuniões regulares com a equipe de conta, encontre-se com ela e discuta as tendências do setor e os serviços da AWS. Fale com o gerente de contas, o arquiteto e a equipe de suporte.
- Compartilhe histórias de sucesso: compartilhe histórias de sucesso sobre redução de custos de qualquer workload, Conta da AWS ou organização para gerar uma atitude positiva e encorajar sobre a otimização dos custos.
- Treinamento: garanta que as equipes técnicas ou os membros da equipe sejam treinados reconhecer os custos dos recursos na Nuvem AWS.
- Eventos e reuniões da AWS: participe de conferências locais da AWS e de quaisquer reuniões locais com outras organizações da área.
- Inscreva-se em blogs: acesse as páginas dos blogs da AWS e assine o blog Novidades e outros blogs relevantes para acompanhar os novos lançamentos, implementações, exemplos e mudanças compartilhados pela AWS.

Recursos

Documentos relacionados:

- Blog da AWS
- Gerenciamento de custos na AWS
- Notícias do blog da AWS

Exemplos relacionados:

- Gerenciamento financeiro na Nuvem AWS
- Laboratórios do AWS Well-Architected: Gerenciamento financeiro na nuvem

COST01-BP09 Quantificar o valor comercial proveniente da otimização de custos

A quantificação do valor empresarial da otimização de custos permite que você entenda todo o conjunto de benefícios da sua organização. Como a otimização de custos é um investimento necessário, quantificar o valor empresarial permite que você explique o retorno sobre o investimento para as partes interessadas. A quantificação do valor empresarial pode ajudar você a ganhar mais participação das partes interessadas em futuros investimentos de otimização de custos e fornece uma estrutura para medir os resultados das atividades de otimização de custos da sua organização.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Quantificar o valor comercial significa avaliar os benefícios que as empresas obtêm com as ações e decisões que tomam. O valor comercial pode ser tangível (como a redução das despesas ou o aumento dos lucros) ou intangível (como a melhoria da reputação da marca ou o aumento da satisfação do cliente).

Quantificar o valor comercial proveniente da otimização de custos significa determinar o valor ou o benefício que você está obtendo de seus esforços para gastar com maior eficiência. Por exemplo, se uma empresa gastar USD 100 mil para implantar uma workload na AWS e depois otimizá-la, o novo custo se tornará apenas USD 80 mil, sem prejudicar a qualidade ou a produção. Nesse cenário, o valor comercial quantificado da otimização de custos seria uma economia de USD 20 mil. Mas, além das economias, a empresa também pode quantificar o valor em termos de prazos de entrega mais rápidos, maior satisfação do cliente ou outras métricas resultantes das iniciativas de otimização de custos. As partes interessadas precisam tomar decisões sobre o valor potencial da otimização de custos, o custo da otimização da workload e o valor de retorno.

Além de relatar economias com base na otimização de custos, recomenda-se quantificar o valor adicional entregue. Os benefícios de otimização de custos normalmente são quantificados em termos de custos mais baixos por resultado comercial. Por exemplo, é possível quantificar a redução de custo do Amazon Elastic Compute Cloud (Amazon EC2) ao comprar Savings Plans, que reduzem os custos e mantêm os níveis de saída da workload. Você pode quantificar reduções de custos nos gastos da AWS quando instâncias ociosas do Amazon EC2 são encerradas ou volumes não vinculados do Amazon Elastic Block Store (Amazon EBS) são excluídos.

No entanto, os benefícios da otimização de custos vão além da redução ou da prevenção de custos. Considere a captura de dados adicionais para medir melhorias de eficiência e valor empresarial.

Etapas de implementação

- Avalie os benefícios comerciais: esse é o processo de analisar e ajustar os custos da Nuvem AWS de forma a maximizar o benefício recebido de cada dólar gasto. Em vez de enfatizar a redução de custos sem valor comercial, considere os benefícios empresarias e o retorno sobre o investimento da otimização de custos, o que pode agregar maior valor ao dispêndio. Isso significa gastar com sabedoria e fazer investimentos e despesas em áreas que geram o melhor retorno.
- Analise os custos de previsão da AWS: a previsão ajuda as partes interessadas financeiras a definir expectativas com outras partes interessadas internas e externas da organização e pode melhorar a previsibilidade financeira da sua organização. O AWS Cost Explorer pode ser usado para realizar previsões de seu custo e uso.

Recursos

Documentos relacionados:

- Fatores econômicos da Nuvem AWS:
- Blog da AWS
- Gerenciamento de custos na AWS
- Notícias do blog da AWS
- Whitepaper Pilar Confiabilidade do Well-Architected
- Explorador de Custos da AWS

Vídeos relacionados:

- Desbloquear o valor comercial com o Windows na AWS

Exemplos relacionados:

- Medir e maximizar o valor empresarial do Cliente 360
- O valor comercial da adoção de bancos de dados gerenciados da Amazon Web Services
- O valor comercial da Amazon Web Services para fornecedores independentes de software
- O valor empresarial da modernização na nuvem

- O valor empresarial da migração para a Amazon Web Services

Reconhecimento de despesas e usos

Perguntas

- COST 2. Como governar o uso?
- COST 3. Como monitorar custos e uso?
- COST 4. Como desativar recursos?

COST 2. Como governar o uso?

Estabeleça políticas e mecanismos para garantir que os custos adequados sejam gerados enquanto os objetivos são alcançados. Ao empregar uma abordagem de verificação e equilíbrio, é possível inovar sem gastar demais.

Práticas recomendadas

- COST02-BP01 Desenvolver políticas com base nos requisitos da sua organização
- COST02-BP02 Implementar metas e objetivos
- COST02-BP03 Implementar uma estrutura de contas
- COST02-BP04 Implementar grupos e perfis
- COST02-BP05 Implementar controles de custos
- COST02-BP06 rastrear o ciclo de vida do projeto

COST02-BP01 Desenvolver políticas com base nos requisitos da sua organização

Desenvolva políticas que definam como os recursos são gerenciados pela sua organização e inspecione-os periodicamente. As políticas devem abranger aspectos de custos de recursos e workloads, incluindo criação, modificação e desativação ao longo da vida útil do recurso.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Entender os custos e os motivadores da sua organização é essencial para gerenciar seus custos e utilização com eficácia e também identificar oportunidades de redução de custos. Normalmente, as

organizações operam várias workloads executadas por várias equipes. Essas equipes podem estar em diferentes unidades da organização, cada uma com o próprio fluxo de receita. A capacidade de atribuir os custos dos recursos a workloads, à uma organização individual ou aos donos do produto propicia um comportamento eficiente de uso e ajuda a reduzir o desperdício. O monitoramento preciso de custos e uso ajuda você a entender a otimização de sua workload, bem como a lucratividade das unidades e produtos da organização. Esse conhecimento permite uma tomada de decisão mais consciente sobre onde alocar recursos em sua organização. A conscientização sobre o uso em todos os níveis da organização é essencial para promover mudanças, pois a mudança no uso gera mudanças no custo. Considere adotar uma abordagem multifacetada para se manter em dia com seus custos e utilização.

O primeiro passo para realizar governança é usar os requisitos da sua organização para desenvolver políticas para o uso da nuvem. Essas políticas definem como a sua organização usa a nuvem e como os recursos são gerenciados. As políticas devem abranger todos os aspectos de recursos e workloads relacionados ao custo ou à utilização, incluindo criação, modificação e desativação durante a vida útil do recurso. Verifique se as políticas e os procedimentos são seguidos e implementados para qualquer alteração em um ambiente de nuvem. Durante as reuniões de gerenciamento de mudanças de TI, questione para descobrir o impacto do custo das alterações planejadas, sejam de aumento ou redução, a justificativa de negócios e o resultado esperado.

As políticas devem ser simples para que sejam facilmente compreendidas e possam ser implementadas com eficácia em toda a organização. As políticas também precisam ser fáceis de seguir e interpretar (para que sejam usadas) e específicas (para evitar erros de interpretação entre as equipes). Além disso, elas precisam ser inspecionadas periodicamente (como nossos mecanismos) e atualizadas à medida que as condições ou as prioridades de negócios dos clientes mudam, o que tornaria a política desatualizada.

Comece com políticas amplas e de alto nível, por exemplo, qual região geográfica usar ou horários do dia em que os recursos devem estar em execução. Refine gradualmente as políticas para as várias unidades organizacionais e workloads. As políticas comuns incluem quais serviços e recursos podem ser usados (por exemplo, armazenamento de dados com menor performance em ambientes de teste e desenvolvimento), quais tipos de recursos podem ser usados por diferentes grupos (por exemplo, o maior tamanho de um recurso em uma conta de desenvolvimento é médio) e por quanto tempo esses recursos ficarão em uso (se temporariamente, em curto prazo ou por um período específico).

Exemplo de política

Veja a seguir um exemplo de política que você pode revisar para criar suas próprias políticas de governança de nuvem, que enfocam a otimização de custos. Ajuste a política com base nos requisitos de sua organização e nas solicitações das partes interessadas.

- Nome da política: defina um nome de política claro, como Política de otimização de recursos e redução de custos.
- Finalidade: explique por que essa política deve ser usada e qual é o resultado esperado. O objetivo dessa política é verificar se há um custo mínimo necessário para implantar e executar a workload desejada para atender aos requisitos de negócios.
- Escopo: defina claramente quem deve usar essa política e quando ela deve ser usada, como o DevOps X Team, para usar essa política em clientes do leste dos EUA para o ambiente X (produção ou não produção).

Declaração da política

1. Selecione us-east-1 ou várias regiões do leste dos EUA com base no ambiente de sua workload e nos requisitos de negócios (desenvolvimento, teste de aceitação do usuário, pré-produção ou produção).
2. Programe instâncias do Amazon EC2 e do Amazon RDS para execução entre 6h e 20h (Horário Padrão do Leste (EST)).
3. Interrompa todas as instâncias do Amazon EC2 não utilizadas após oito horas e as instâncias do Amazon RDS não utilizadas após 24 horas de inatividade.
4. Encerre todas as instâncias do Amazon EC2 não utilizadas após 24 horas de inatividade em ambientes que não sejam de produção. Lembre o proprietário da instância do Amazon EC2 (com base em tags) de revisar suas instâncias do Amazon EC2 paradas na produção e informá-lo de que elas serão encerradas em 72 horas se não forem usadas.
5. Use família de instância e tamanho genéricos, como m5.large, e, depois, redimensione a instância com base na utilização da CPU e da memória usando o AWS Compute Optimizer.
6. Priorize o uso do ajuste de escala automático para ajustar dinamicamente o número de instâncias em execução com base no tráfego.
7. Use instâncias spot para workloads não essenciais.
8. Analise os requisitos de capacidade para comprometer Saving Plans ou instâncias reservadas para workloads previsíveis e informe a equipe de gerenciamento financeiro da nuvem.

9. Use políticas de ciclo de vida do Amazon S3 para mover dados acessados com pouca frequência para níveis de armazenamento mais baratos. Se nenhuma política de retenção for definida, use o Amazon S3 Intelligent Tiering para mover objetos automaticamente para a camada arquivada.
10. Monitore a utilização de recursos e defina alarmes para acionar eventos de ajuste de escala usando o Amazon CloudWatch.
11. Para cada Conta da AWS, use o AWS Budgets para definir orçamentos de custo e uso para sua conta com base no centro de custos e nas unidades de negócios.
12. Usar o AWS Budgets para definir orçamentos de custo e uso para sua conta pode ajudar você a controlar seus gastos e evitar contas inesperadas, proporcionando um melhor controle sobre seus custos.

Procedimento: forneça procedimentos detalhados para implementar essa política ou consulte outros documentos que descrevam como implementar cada declaração de política. Esta seção deve fornecer instruções detalhadas para a elaboração dos requisitos da política.

Para implementar essa política, você pode usar várias ferramentas de terceiros ou regras do AWS Config para conferir a conformidade com a declaração de política e acionar ações de correção automatizadas usando funções do AWS Lambda. Você também pode usar o AWS Organizations para aplicar a política. Além disso, você deve revisar regularmente o uso de recursos e ajustar a política conforme necessário para verificar se ela continua atendendo às suas necessidades comerciais.

Etapas de implementação

- Reuna-se com as partes interessadas: para desenvolver políticas, peça às partes interessadas (escritórios de negócios na nuvem, engenheiros ou tomadores de decisão funcionais para aplicação de políticas) em sua organização que especifiquem seus requisitos e os documentem. Adote uma abordagem iterativa iniciando uma refinação ampla e contínua para as menores unidades em cada etapa. Os membros da equipe incluem aqueles com interesse direto na workload, como unidades da organização ou proprietários de aplicações, além de grupos de apoio como equipes de segurança e finanças.
- Obtenha confirmação: garanta que as equipes concordem com as políticas que determinam quem pode acessar e implantar na Nuvem AWS. Certifique-se de que elas sigam as políticas da sua organização e confirme se o provisionamento de recursos está alinhado com as políticas e procedimentos estabelecidos.

- Crie sessões de treinamento para integração: peça que os novos membros completem cursos de formação de integração para criar conscientização de custo e requisitos da organização. As equipes podem assumir políticas diferentes das suas experiências anteriores ou simplesmente ignorá-las.
- Defina locais para sua workload: defina onde sua workload opera, incluindo o país e a área dentro do país. Essas informações são usadas para mapear as zonas de disponibilidade e Regiões da AWS.
- Defina e agrupe serviços e recursos: defina os serviços que as workloads exigem. Para cada serviço, especifique os tipos, o tamanho e o número de recursos necessários. Defina grupos para os recursos por função, como servidores de aplicações ou armazenamento de banco de dados. Os recursos podem pertencer a vários grupos.
- Defina e agrupe os usuários por função: defina os usuários que interagem com a workload, concentrando-se no que eles fazem e em como usam a workload, e não em quem são nem em suas posições na organização. Agrupe usuários ou funções semelhantes. Você pode usar as políticas gerenciadas da AWS como um guia.
- Defina as ações: usando os locais, recursos e usuários identificados anteriormente, defina as ações que são exigidas por cada um para alcançar os resultados da workload ao longo do tempo de vida (desenvolvimento, operação e desativação). Identifique as ações com base nos grupos, e não nos elementos individuais nos grupos, em cada local. Comece de forma ampla como leitura ou gravação e, em seguida, refine ações específicas para cada serviço.
- Defina o período de análise: as workloads e os requisitos organizacionais podem mudar com o tempo. Defina a programação de análise da workload para garantir que ela permaneça alinhada com as prioridades organizacionais.
- Documente as políticas: verifique se as políticas que foram definidas estão acessíveis conforme exigido pela sua organização. Essas políticas são usadas para implementar, manter e auditar o acesso de seus ambientes.

Recursos

Documentos relacionados:

- Gerenciamento de alterações na nuvem
- Políticas gerenciadas pela AWS para funções de trabalho
- Estratégia de faturamento para várias contas da AWS
- Ações, recursos e chaves de condição para serviços da AWS

- Gerenciamento e governança da AWS
- Controlar o acesso às Regiões da AWS usando as políticas do IAM
- Regiões e AZs de infraestruturas globais

Vídeos relacionados:

- Gerenciamento e governança da AWS em grande escala

Exemplos relacionados:

- VMware: o que são políticas de nuvem?

COST02-BP02 Implementar metas e objetivos

Implemente metas e objetivos de custos e uso para sua workload. As metas fornecem orientação para sua organização quanto aos resultados esperados, e os objetivos oferecem resultados mensuráveis específicos a serem alcançados para suas workloads.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Desenvolva metas e objetivos de custos e uso para a sua organização. Como uma organização em crescimento na AWS, é importante definir e monitorar metas para otimização de custos. Essas metas ou indicadores-chave de performance (KPIs) podem incluir itens como porcentagem de gastos sob demanda ou adoção de certos serviços otimizados, como instâncias do AWS Graviton ou tipos de volume gp3 do EBS. Estabeleça metas mensuráveis e alcançáveis para ajudar a medir as melhorias de eficiência, o que é importante para suas operações comerciais. As metas fornecem orientações e direcionamento para a sua organização quanto aos resultados esperados.

Os objetivos fornecem resultados mensuráveis específicos a serem alcançados. Em suma, uma meta é a direção que você deseja seguir e o objetivo é até que ponto nessa direção a meta deve ir e quando ela deve ser concretizada [usando a orientação do método SMART (específico, mensurável, atribuível, realista e rápido)]. Um exemplo de meta é que o uso da plataforma deve aumentar significativamente, com apenas um pequeno aumento (não linear) no custo. Um exemplo de objetivo é um aumento de 20% no uso da plataforma, com um aumento de menos de 5% nos custos. Outra meta comum é que as workloads precisam ser mais eficientes a cada seis meses. O objetivo

que a acompanha seria o custo de acordo com as métricas empresariais diminuir em 5% a cada seis meses. Use as métricas corretas e defina KPIs calculados para sua organização. Você pode começar com KPIs básicos e expandir posteriormente com base nas necessidades da empresa.

Uma meta para a otimização de custos é aumentar a eficiência da workload, o que corresponde a diminuir o custo por resultado empresarial da workload ao longo do tempo. Implemente essa meta para todas as workloads e defina um objetivo, como um aumento de 5% na eficiência a cada seis meses a um ano. Na nuvem, é possível conseguir isso ao estabelecer a capacidade de otimização de custos, bem como novos lançamentos de serviços e recursos.

Os objetivos são as referências quantificáveis que você deseja alcançar para concretizar suas metas, ao passo que as referências comparam seus resultados reais com um objetivo. Estabeleça referências com KPIs para o custo por unidade de serviços de computação (como adoção de spot, adoção do Graviton, tipos de instância mais recentes e cobertura sob demanda), de serviços de armazenamento (como adoção do EBS GP3, snapshots obsoletos do EBS e armazenamento do Amazon S3 padrão) ou de uso de serviços de banco de dados (como mecanismos de código aberto do RDS, adoção do Graviton e cobertura sob demanda). Essas referências e KPIs podem ajudar a verificar se você usa os serviços da AWS da maneira mais econômica.

A tabela a seguir fornece uma lista de métricas padrão da AWS para referência. Cada organização pode ter valores de objetivo diferentes para esses KPIs.

Categoria	KPI	Descrição
Computação	Cobertura de uso do EC2	Instâncias do EC2 (em custo ou horas) usando SP+RI +spot em comparação com o total (em custo ou horas) de instâncias do EC2
Computação	Utilização de SP/RI de computação	Horas de SP ou RI utilizada s em comparação com o total de horas de SP ou RI disponíveis
Computação	Custo do EC2/hora	Custo do EC2 dividido pelo número de instâncias do EC2 em execução naquela hora

Categoria	KPI	Descrição
Computação	Custo de vCPU	Custo por vCPU para todas as instâncias
Computação	Última geração de instância	Porcentagem de instância s no Graviton (ou em outros tipos de instância de geração moderna)
Banco de dados	Cobertura de RDS	Instâncias do RDS (em custo ou horas) usando RI em comparação com o total (em custo ou horas) de instâncias do RDS
Banco de dados	Utilização do RDS	Horas de RI utilizadas em comparação com o total de horas de RI disponíveis
Banco de dados	Tempo de atividade do RDS	Custo do RDS dividido pelo número de instâncias do RDS em execução naquela hora
Banco de dados	Última geração de instância	Porcentagem de instâncias no Graviton (ou em outros tipos de instância moderna)
Armazenamento	Utilização de armazenamento	Custo do armazenamento otimizado (por exemplo, Glacier, arquivamento profundo ou acesso infrequen te) dividido pelo custo total de armazenamento

Categoria	KPI	Descrição
Tags	Recursos não marcados	Explorador de Custos 1. Filtre créditos, descontos, impostos, reembolsos, marketplace e copie o custo mensal mais recente. 2. Selecione Mostrar somente recursos não marcados no Explorador de Custos 3. Divida o valor em recursos não marcados com seu custo mensal.

Usando essa tabela, inclua valores de metas ou de referência, os quais devem ser calculados com base nas metas organizacionais. É necessário avaliar determinadas métricas para sua empresa e entender os resultados comerciais dessa workload para definir KPIs precisos e realistas. Ao avaliar as métricas de performance em uma organização, faça a distinção entre os diferentes tipos de métrica que servem a propósitos distintos. Essas métricas avaliam principalmente a performance e a eficiência da infraestrutura técnica, e não diretamente o impacto geral nos negócios. Por exemplo, elas podem monitorar os tempos de resposta do servidor, a latência da rede ou o tempo de atividade do sistema. Essas métricas são essenciais para avaliar a capacidade da infraestrutura de comportar as operações técnicas da organização. No entanto, elas não fornecem informações diretas sobre objetivos comerciais mais amplos, como satisfação do cliente, crescimento da receita ou participação de mercado. Para obter uma compreensão abrangente da performance dos negócios, complemente essas métricas de eficiência com métricas estratégicas de negócios que se correlacionem diretamente com os resultados comerciais.

Crie visibilidade quase em tempo real sobre seus KPIs e oportunidades de economia relacionadas e acompanhe seu progresso ao longo do tempo. Para começar a definir e monitorar os objetivos de KPI, recomendamos o painel de KPI dos Cloud Intelligence Dashboards (CID). Com base nos dados do Relatório de Custos e Uso (CUR), o painel de KPI oferece uma série de KPIs de otimização de custos recomendados com a capacidade de definir metas personalizadas e rastrear o progresso ao longo do tempo.

Se você tiver outras soluções que definam e monitorem objetivos de KPI, garanta que esses métodos sejam adotados por todas as partes interessadas de gerenciamento financeiro na nuvem em sua organização.

Etapas de implementação

- Defina os níveis de uso esperados: para começar, concentre-se nos níveis de uso. Interaja com os proprietários de aplicações, a equipe de marketing e as equipes de negócios maiores para entender quais serão os níveis de uso esperados para a workload. Como a demanda do cliente pode mudar com o tempo e o que pode mudar em decorrência de aumentos sazonais ou campanhas de marketing?
- Defina recursos e custos da workload: com os níveis de uso definidos, quantifique as alterações nos recursos da workload necessárias para atender a esses níveis de uso. Talvez seja necessário aumentar o tamanho ou o número de recursos para um componente de workload, aumentar a transferência de dados ou alterar componentes de workload para um serviço diferente em um nível específico. Especifique os custos em cada um desses pontos principais e preveja a mudança no custo quando houver uma mudança no uso.
- Defina metas de negócios: combine o resultado das alterações esperadas no uso e no custo com as alterações esperadas na tecnologia ou qualquer programa que você esteja executando e desenvolva metas para a workload. As metas devem abordar o uso e o custo, bem como a relação entre os dois. As metas devem ser simples e gerais e ajudar as pessoas a entender o que a empresa espera em termos de resultados (por exemplo, garantir que recursos não utilizados sejam mantidos abaixo de determinado nível de custo). Não é necessário definir metas para cada tipo de recurso não utilizado nem definir custos que possam causar perdas em metas e objetivos. Verifique se há programas organizacionais (por exemplo, criação de recursos como treinamento e educação) se houver alterações esperadas no custo sem alterações no uso.
- Definir objetivos: para cada uma das metas definidas, especifique um objetivo mensurável. Se a meta for aumentar a eficiência na workload, o objetivo deverá quantificar a melhoria (normalmente nos resultados de negócios para cada dólar gasto) e quando ela deverá ser entregue. Por exemplo, é possível definir uma meta para minimizar o desperdício devido ao excesso de provisionamento. Com essa meta, seu objetivo pode ser que o desperdício decorrente do superprovisionamento de computação no primeiro nível de workloads de produção não exceda 10% do custo de computação do nível. Além disso, um segundo objetivo pode ser que o desperdício decorrente do provisionamento excessivo de computação no segundo nível de workloads de produção não exceda 5% do custo de computação do nível.

Recursos

Documentos relacionados:

- Políticas gerenciadas pela AWS para funções de trabalho
- Estratégia de faturamento para várias contas da AWS
- Controlar o acesso às Regiões da AWS usando as políticas do IAM
- Objetivos do S.M.A.R.T.
- Como rastrear KPIs de otimização de custos com o painel de KPI do CID

Vídeos relacionados:

- Laboratórios do Well-Architected: Metas e objetivos (Nível 100)

Exemplos relacionados:

- O que é uma métrica unitária?
- Selecionar uma métrica unitária para apoiar sua empresa
- Métricas unitárias na prática: lições aprendidas
- Como as métricas unitárias ajudam a criar alinhamento entre as funções de negócios
- Laboratórios do Well-Architected: Desativar recursos (metas e objetivos)
- Laboratórios do Well-Architected: Tipo, tamanho e número de recursos (metas e objetivos)

COST02-BP03 Implementar uma estrutura de contas

Implemente uma estrutura de contas que mapeie para sua organização. Isso auxilia na alocação e no gerenciamento de custos em toda a organização.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

O AWS Organizations permite criar várias Contas da AWS que podem ajudar você a gerenciar de maneira centralizada seu ambiente à medida que dimensiona suas workloads na AWS. É possível modelar sua hierarquia organizacional agrupando Contas da AWS na estrutura da unidade

organizacional (UO) e criando várias Contas da AWS em cada UO. Para criar uma estrutura de contas, primeiramente, você precisa decidir qual das suas Contas da AWS será a conta de gerenciamento. Depois disso, você pode criar Contas da AWS novas ou selecionar contas existentes como contas-membro com base na estrutura de conta projetada ao seguir as práticas recomendadas para contas de gerenciamento e as práticas recomendadas para contas-membro.

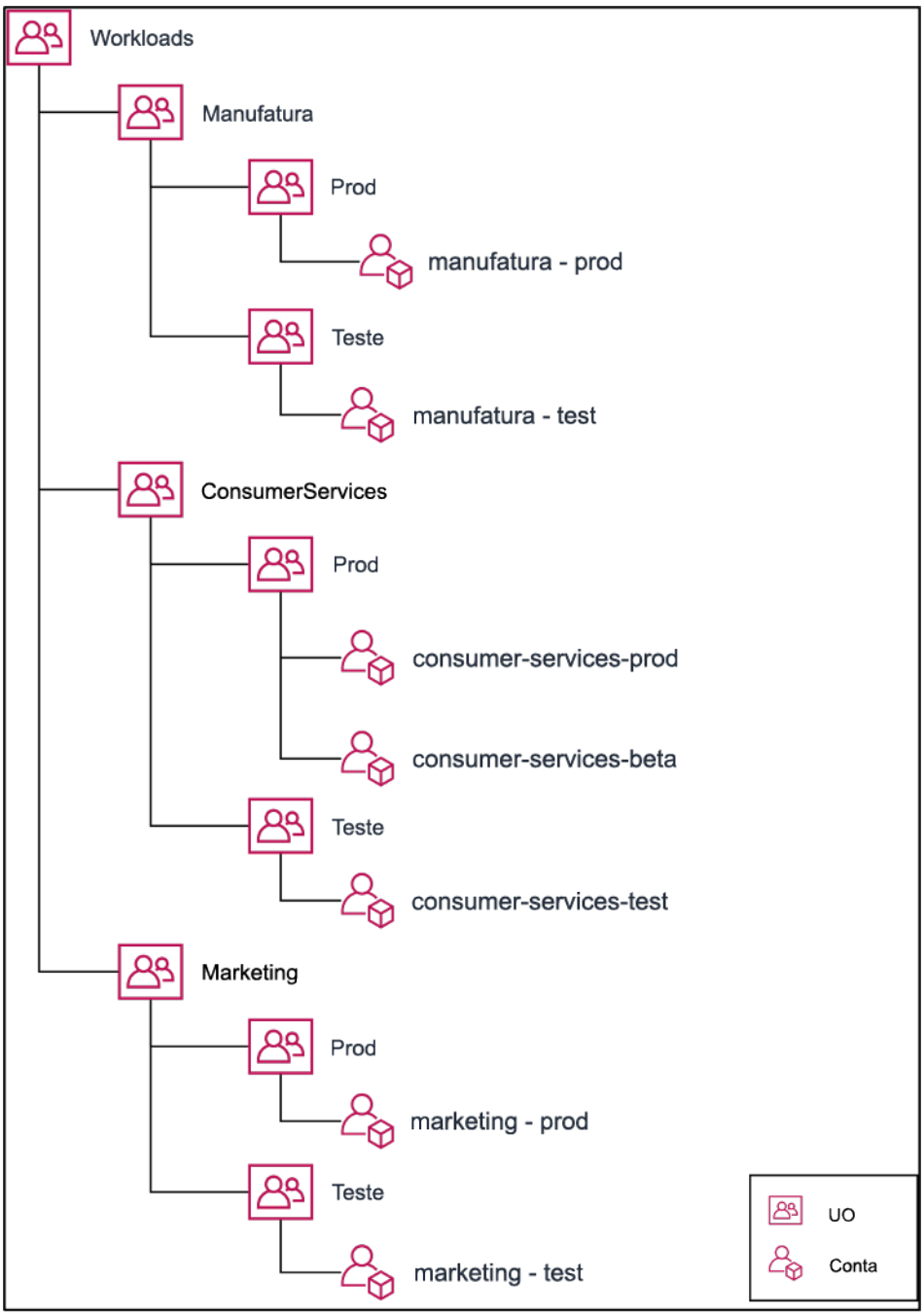
Recomenda-se que sempre haja, pelo menos, uma conta de gerenciamento com uma conta-membro, independentemente do tamanho ou uso da organização. Todos os recursos de workload devem residir somente nas contas-membro, e nenhum recurso deve ser criado na conta de gerenciamento. Não há uma resposta geral para a quantidade de Contas da AWS que você deveria ter. Avalie seus modelos de custo e operacionais atuais e futuros para garantir que a estrutura de suas Contas da AWS reflita os objetivos da sua organização. Algumas empresas criam várias Contas da AWS por motivos de negócios, por exemplo:

- O isolamento administrativo ou fiscal e de faturamento é necessário entre unidades da organização, centros de custo ou workloads específicas.
- Os limites de serviço da AWS são definidos para que sejam específicos a determinadas workloads.
- Há um requisito de isolamento e separação entre workloads e recursos.

Dentro do AWS Organizations, o faturamento consolidado cria a construção entre uma ou mais contas-membro e a conta mestre. As contas-membro permitem que você isole e diferencie seu custo e uso por grupos. Uma prática comum é ter contas-membro separadas para cada unidade da organização (como finanças, marketing e vendas), ou para cada ciclo de vida do ambiente (como desenvolvimento, teste e produção) ou para cada workload (workload a, b e c) e, em seguida, agregar essas contas vinculadas usando o faturamento consolidado.

O faturamento consolidado permite consolidar o pagamento de várias Contas da AWS-membro em uma única conta de gerenciamento, sem deixar de oferecer visibilidade para a atividade de cada conta vinculada. Como os custos e o uso são agregados na conta de gerenciamento, você pode maximizar seus descontos por volume de serviço e maximizar o uso de seus descontos de compromisso (Savings Plans e instâncias reservadas) para obter os maiores descontos possíveis.

O diagrama a seguir mostra como é possível usar o AWS Organizations com unidades organizacionais (UO) para agrupar várias contas e colocar várias Contas da AWS em cada UO. Recomenda-se usar UOs para vários casos de uso e workloads que fornecem padrões para organizar contas.



Exemplo de agrupamento de várias Contas da AWS em unidades organizacionais.

O AWS Control Tower pode instalar e configurar rapidamente várias contas da AWS, garantindo que a governança esteja alinhada com os requisitos da sua organização.

Etapas de implementação

- Defina requisitos de separação: os requisitos de separação são uma combinação de vários fatores, incluindo segurança, confiabilidade e construções financeiras. Trabalhe em cada fator em ordem

e especifique se a workload ou o respectivo ambiente devem ser separados de outras workloads. A segurança promove a adesão aos requisitos de acesso e de dados. A confiabilidade gerencia os limites para que os ambientes e as workloads não afetem os outros. Revise os pilares de segurança e de confiabilidade do Well-Architected Framework periodicamente e siga as práticas recomendadas fornecidas. As estruturas financeiras criam separação financeira rígida (diferentes centros de custo, propriedades de workload e responsabilidades). Exemplos comuns de separação são workloads de produção e de teste executadas em contas separadas ou o uso de uma conta separada para que os dados da fatura e do faturamento possam ser fornecidos às unidades de negócios individuais ou aos departamentos da organização, ou à parte interessada que possui a conta.

- Defina requisitos de agrupamento: os requisitos de agrupamento não modificam os requisitos de separação, mas são usados para auxiliar o gerenciamento. Agrupe ambientes semelhantes ou workloads que não exigem separação. Um exemplo disso é o agrupamento de vários ambientes de teste ou desenvolvimento de uma ou mais workloads.
- Defina a estrutura da conta: usando essas separações e agrupamentos, especifique uma conta para cada grupo e mantenha os requisitos de separação. Essas contas são suas contas-membro ou vinculadas. Ao agrupar essas contas-membro em uma única conta de gerenciamento ou pagante, você combina o uso, o que permite maiores descontos por volume em todas as contas e fornece uma única fatura para todas as contas. É possível separar dados de faturamento e fornecer a cada conta-membro uma visualização individual dos dados de faturamento. Se uma conta-membro não precisar ter os dados de uso ou de faturamento visíveis para nenhuma outra conta, ou se uma fatura separada da AWS for necessária, você deverá definir várias contas de gerenciamento ou pagantes. Nesse caso, cada conta-membro tem a própria conta de gerenciamento ou pagante. Os recursos devem sempre ser colocados em contas-membro ou vinculadas. As contas de gerenciamento ou pagantes devem ser usadas somente para gerenciamento.

Recursos

Documentos relacionados:

- Usar tags de alocação de custos
- Políticas gerenciadas pela AWS para funções de trabalho
- Estratégia de faturamento para várias contas da AWS
- Controlar o acesso às Regiões da AWS usando as políticas do IAM
- AWS Control Tower

- AWS Organizations
- Práticas recomendadas para contas de gerenciamento e contas-membro
- Organizar seu ambiente da AWS usando várias contas
- Ativar descontos compartilhados de instâncias reservadas e Savings Plans
- Faturamento consolidado
- Faturamento consolidado

Exemplos relacionados:

- Dividindo o CUR e compartilhando o acesso

Vídeos relacionados:

- Introdução ao AWS Organizations
- Configurar um ambiente da AWS com várias contas que use práticas recomendadas para o AWS Organizations

Exemplos relacionados:

- Laboratórios do Well-Architected: Criar um AWS Organization (Nível 100)
- Dividir o AWS Cost and Usage Report e compartilhar acesso
- Definir uma estratégia de várias contas da AWS para empresas de telecomunicações
- Práticas recomendadas para otimizar Contas da AWS
- Práticas recomendadas para unidades organizacionais com o AWS Organizations

COST02-BP04 Implementar grupos e perfis

Implemente grupos e funções que se alinhem às políticas e controle quem pode criar, modificar ou desativar instâncias e recursos em cada grupo. Por exemplo, implemente grupos de desenvolvimento, teste e produção. Isso se aplica a serviços da AWS e a soluções de terceiros.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Os perfis e os grupos de usuários são elementos fundamentais no design e na implementação de sistemas seguros e eficientes. Os perfis e os grupos ajudam as organizações a equilibrar a necessidade de controle com a necessidade de flexibilidade e produtividade, e, acima de tudo, apoiando os objetivos organizacionais e as necessidades dos usuários. Conforme recomendado na seção Gerenciamento de identidade e acesso do Pilar Segurança do AWS Well-Architected Framework, um gerenciamento robusto de identidade e permissões é necessário para fornecer acesso aos recursos certos para as pessoas certas nas condições certas. Os usuários recebem somente o acesso necessário para realizar suas tarefas. Isso minimiza o risco associado a acesso não autorizado ou uso indevido.

Depois de desenvolver políticas, é possível criar perfis e grupos lógicos de usuários em sua organização. Isso permite que você atribua permissões, controle o uso e ajude a implementar mecanismos robustos de controle de acesso, impedindo o acesso não autorizado a informações sigilosas. Comece com agrupamentos de pessoas de alto nível. Normalmente, isso se alinha às unidades organizacionais e aos cargos (por exemplo, administrador de sistemas no departamento de TI, controlador financeiro ou analista de negócios). Os grupos categorizam pessoas que realizam tarefas semelhantes e precisam de acesso semelhante. Os perfis definem o que um grupo deve fazer. É mais fácil gerenciar permissões para grupos e perfis do que para usuários individuais. Os perfis e os grupos atribuem permissões de forma consistente e sistemática a todos os usuários, evitando erros e inconsistências.

Quando o perfil de um usuário muda, os administradores podem ajustar o acesso por perfil ou grupo, em vez de reconfigurar as contas de usuários individuais. Por exemplo, um administrador de sistemas em TI requer acesso para criar todos os recursos, mas um membro da equipe de análise só precisa criar recursos de análise.

Etapas de implementação

- Implemente grupos: usando os grupos de usuários definidos em suas políticas organizacionais, implemente os grupos correspondentes, se necessário. Para conhecer as práticas recomendadas para usuários, grupos e autenticação, consulte o Pilar Segurança do AWS Well-Architected Framework.
- Implemente perfis e políticas: usando as ações definidas em suas políticas organizacionais, crie os perfis e as políticas de acesso necessários. Para conhecer as práticas recomendadas para perfis e políticas, consulte o Pilar Segurança do AWS Well-Architected Framework.

Recursos

Documentos relacionados:

- Políticas gerenciadas pela AWS para funções de trabalho
- Estratégia de faturamento para várias contas da AWS
- Pilar Segurança do AWS Well-Architected Framework
- AWS Identity and Access Management (IAM)
- Políticas do AWS Identity and Access Management

Vídeos relacionados:

- Por que usar gerenciamento de identidade e acesso

Exemplos relacionados:

- Laboratório do Well-Architected: Identidade e acesso básicos
- Controlar o acesso às Regiões da AWS usando as políticas do IAM
- Como iniciar sua jornada de gerenciamento financeiro na nuvem: operações de custos na nuvem

COST02-BP05 Implementar controles de custos

Implemente controles baseados nas políticas da organização e nos perfis e grupos definidos. Isso garante que os custos sejam gerados somente conforme definido pelos requisitos da organização, como controle do acesso a regiões ou tipos de recursos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Uma primeira etapa comum na implementação de controles de custo é configurar notificações quando eventos de custo ou de uso ocorrerem fora das políticas. É possível adotar medidas rápidas e verificar se alguma ação corretiva é necessária, sem restringir ou afetar negativamente workloads ou novas atividades. Depois de conhecer os limites da workload e do ambiente, você pode aplicar a governança. O AWS Budgets permite que você defina notificações e orçamentos mensais para seus custos, uso e descontos de compromisso (Savings Plans e Reserved Instances) da AWS. É possível criar orçamentos em um nível de custo agregado (por exemplo, todos os custos) ou em um nível

mais granular, onde você inclui apenas dimensões específicas, como contas vinculadas, serviços, tags ou zonas de disponibilidade.

Depois de definir seus limites de orçamento com o AWS Budgets, use o AWS Cost Anomaly Detection para reduzir seu custo inesperado. O AWS Cost Anomaly Detection é um serviço de gerenciamento de custos que usa machine learning para monitorar continuamente seus custos e uso para detectar gastos incomuns. Ele ajuda a identificar gastos anômalos e causas-raiz para que você possa agir rapidamente. Primeiro, crie um monitor de custos no AWS Cost Anomaly Detection e, em seguida, escolha sua preferência de alerta configurando um limite em dólares (como um alerta sobre anomalias com impacto superior a USD 1 mil). Ao receber um alerta, você poderá analisar a causa-raiz por trás da anomalia e o impacto em seus custos. Também é possível monitorar e realizar sua própria análise de anomalias no AWS Cost Explorer.

Imponha políticas de governança na AWS por meio do AWS Identity and Access Management e de políticas de controle de serviços (SCP) do AWS Organizations. O IAM permite que você gerencie com segurança o acesso aos serviços e recursos da AWS. Com o IAM, você pode controlar quem pode criar ou gerenciar recursos da AWS, os tipos de recursos que podem ser criados e onde eles podem ser criados. Isso minimiza a possibilidade de recursos serem criados fora da política definida. Use as funções e grupos criados anteriormente e atribua políticas do IAM para impor o uso correto. Uma SCP oferece controle central sobre o número máximo de permissões disponíveis para todas as contas na sua organização, garantindo que suas contas permaneçam dentro das diretrizes de controle de acesso. As SCPs estão disponíveis somente em uma organização com todos os recursos habilitados, e você pode configurar as SCPs para negar ou permitir ações para contas-membro por padrão. Para obter mais detalhes sobre a implementação do gerenciamento de acesso, consulte o whitepaper Pilar Segurança do Well-Architected.

A governança também pode ser implementada por meio do gerenciamento de cotas de serviço da AWS. Ao garantir que as cotas de serviço sejam configuradas com o mínimo de sobrecarga e mantidas com precisão, você pode minimizar a criação de recursos fora dos requisitos da sua organização. Para conseguir isso, você deve entender a rapidez com que seus requisitos podem mudar, compreender projetos em andamento (criação e desativação de recursos) e considerar a rapidez com que as alterações de cota podem ser implementadas. As cotas de serviço podem ser usadas para aumentar suas cotas quando necessário.

Etapas de implementação

- Implemente notificações sobre gastos: usando suas políticas organizacionais definidas, crie AWS Budgets para receber notificações quando os gastos estiverem fora de suas políticas. Configure vários orçamentos de custos, um para cada conta, para ser notificado sobre os gastos gerais da

conta. Configure orçamentos de custos adicionais dentro de cada conta para unidades menores dentro da conta. Essas unidades variam de acordo com a estrutura da sua conta. Alguns exemplos comuns são Regiões da AWS, workloads (usando tags) ou serviços da AWS. Configure uma lista de distribuição de e-mails como o destinatário das notificações, e não uma conta de e-mail de uma pessoa. É possível configurar um orçamento real para quando um valor for ultrapassado ou usar um orçamento previsto para notificar sobre o uso previsto. Você também pode pré-configurar ações do AWS Budgets que podem aplicar políticas específicas do IAM ou SCP ou interromper instâncias do Amazon EC2 ou Amazon RDS. As ações de orçamento podem ser executadas automaticamente ou exigir aprovação do fluxo de trabalho.

- Implemente notificações sobre gastos anômalos: use o AWS Cost Anomaly Detection para reduzir custos inesperados em sua organização e analisar a causa-raiz de possíveis gastos anômalos. Depois de criar o monitor de custos para identificar gastos incomuns em sua granularidade especificada e configurar notificações no AWS Cost Anomaly Detection, ele envia um alerta quando um gasto incomum é detectado. Isso permitirá que você analise a causa-raiz por trás da anomalia e entenda o impacto em seu custo. Use Categorias de Custos da AWS durante a configuração do AWS Cost Anomaly Detection para identificar qual equipe de projeto ou equipe de unidade de negócios pode analisar a causa-raiz do custo inesperado e tomar as ações necessárias em tempo hábil.
- Implemente controles de utilização: usando as políticas da organização definidas, implemente políticas e perfis do IAM para especificar quais ações os usuários podem e não podem executar. Várias políticas organizacionais podem ser incluídas em uma política da AWS. Da mesma forma que você definiu políticas, comece de forma mais ampla e, em seguida, aplique controles mais granulares em cada etapa. Os limites de serviço também são um controle eficaz do uso. Implemente os limites de serviço corretos em todas as suas contas.

Recursos

Documentos relacionados:

- Políticas gerenciadas pela AWS para funções de trabalho
- Estratégia de faturamento para várias contas da AWS
- Controlar o acesso às Regiões da AWS usando as políticas do IAM
- AWS Budgets
- AWS Cost Anomaly Detection
- Controlar seus custos da AWS

Vídeos relacionados:

- Como posso usar o AWS Budgets para monitorar meus gastos e uso?

Exemplos relacionados:

- Exemplos de políticas de gerenciamento de acesso do IAM
- Exemplos de políticas de controle de serviços
- AWS Budgets Actions
- Criar uma política do IAM para controlar o acesso aos recursos do Amazon EC2 usando tags
- Restringir o acesso da identidade do IAM a recursos específicos do Amazon EC2
- Criar uma política do IAM para restringir o uso do Amazon EC2 pela família
- Laboratórios do Well-Architected: Governança de custos e uso (Nível 100)
- Laboratórios do Well-Architected: Governança de custos e uso (Nível 200)
- Integrações do Slack para detecção de anomalias de custo usando o AWS Chatbot

COST02-BP06 rastrear o ciclo de vida do projeto

Acompanhe, meça e realize auditorias no ciclo de vida dos projetos, equipes e ambientes para evitar o uso e pagamento de recursos desnecessários.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Ao monitorar com eficácia o ciclo de vida do projeto, as organizações podem ter um melhor controle de custos por meio de planejamento, gerenciamento e otimização de recursos aprimorados. Os insights recebidos por meio do rastreamento são inestimáveis para a tomada de decisões fundamentadas que contribuem para a relação custo-benefício e o sucesso geral do projeto.

O rastreamento de todo o ciclo de vida da workload ajuda a compreender quando as workloads ou os respectivos componentes não são mais necessários. As workloads e os componentes existentes podem parecer estar em uso, mas quando a AWS libera novos serviços ou recursos, eles podem ser desativados ou adotados. Confira os estágios anteriores das workloads. Depois que uma workload entra em produção, os ambientes anteriores podem ser desativados ou ter a capacidade significativamente reduzida até que sejam necessários novamente.

É possível marcar recursos com um prazo ou um lembrete para fixar a hora em que a workload foi analisada. Por exemplo, se o ambiente de desenvolvimento foi analisado pela última vez meses atrás, talvez seja um bom momento para analisá-lo novamente com o objetivo de examinar se novos serviços podem ser adotados ou se o ambiente está em uso. Você pode agrupar e marcar aplicações com o myApplications na AWS para gerenciar e rastrear metadados, como criticidade, ambiente, última revisão e centro de custos. É possível monitorar o ciclo de vida da workload e monitorar e gerenciar o custo, a integridade, o procedimento de segurança e a performance das aplicações.

A AWS fornece uma série de serviços de gerenciamento e de governança que você pode usar para o monitoramento do ciclo de vida da entidade. É possível usar o AWS Config ou o AWS Systems Manager para fornecer um inventário detalhado dos recursos e da configuração da AWS. Recomendamos integrá-lo a seus sistemas existentes de gerenciamento de projetos ou ativos para rastrear projetos e produtos ativos em sua organização. A combinação do seu sistema atual com o conjunto de eventos e métricas avançados fornecido pela AWS permite criar uma visão de eventos de ciclo de vida significativos e gerenciar os recursos proativamente para reduzir os custos desnecessários.

Semelhante ao gerenciamento do ciclo de vida da aplicação (ALM), o rastreamento do ciclo de vida do projeto deve envolver vários processos, ferramentas e equipes trabalhando em conjunto, como design e desenvolvimento, testes, produção, suporte e redundância de workload.

Ao monitorar cuidadosamente cada fase do ciclo de vida de um projeto, as organizações obtêm insights cruciais e um controle aprimorado, o que facilita o sucesso do planejamento, da implementação e da conclusão do projeto. Essa supervisão cuidadosa verifica se os projetos, além de atenderem aos padrões de qualidade, são entregues no prazo e dentro do orçamento, promovendo o custo-benefício de modo geral.

Consulte o whitepaper Pilar Excelência operacional do AWS Well-Architected para obter mais detalhes sobre a implementação do rastreamento do ciclo de vida da entidade.

Etapas de implementação

- Estabeleça o processo de monitoramento do ciclo de vida do projeto: a equipe do Centro de Excelência da Nuvem deve estabelecer o processo de monitoramento do ciclo de vida do projeto. Estabeleça uma abordagem estruturada e sistemática para monitorar as workloads a fim de melhorar o controle, a visibilidade e a performance dos projetos. Torne o processo de monitoramento transparente, colaborativo e dedicado à melhoria contínua para maximizar sua eficácia e valor.

- Realize análises da workload: conforme definido por suas políticas organizacionais, configure uma frequência regular para auditar seus projetos existentes e realizar análises da workload. A quantidade de esforço empregado na auditoria deve ser proporcional ao risco aproximado, ao valor ou ao custo para a organização. As principais áreas a serem incluídas na auditoria seriam riscos para a organização de um incidente ou interrupção, valor ou contribuição para a organização (medidos em receita ou reputação da marca), custo da workload (medido como custo total de recursos e custos operacionais) e uso da workload (medido em número de resultados da organização por unidade de tempo). Se essas áreas mudarem ao longo do ciclo de vida, ajustes serão necessários na workload, como desativação total ou parcial.

Recursos

Documentos relacionados:

- Orientação para marcação com tags na AWS
- O que é gerenciamento do ciclo de vida de aplicações (ALM)?
- Políticas gerenciadas pela AWS para funções de trabalho

Exemplos relacionados:

- Controlar o acesso às Regiões da AWS usando as políticas do IAM

Ferramentas relacionadas

- AWS Config
- AWS Systems Manager
- AWS Budgets
- AWS Organizations
- AWS CloudFormation

COST 3. Como monitorar custos e uso?

Estabeleça políticas e procedimentos para monitorar e alocar adequadamente os custos. Isso permite medir e aprimorar a eficiência de custos dessa workload.

Práticas recomendadas

- COST03-BP01 Configurar fontes de informações detalhadas
- COST03-BP02 Adicionar informações da organização aos custos e ao uso
- COST03-BP03 Identificar categorias de atribuição de custos
- COST03-BP04 Estabelecer métricas da organização
- COST03-BP05 Configurar ferramentas de faturamento e gerenciamento de custos
- COST03-BP06 Alocar custos com base nas métricas de workload

COST03-BP01 Configurar fontes de informações detalhadas

Configure ferramentas de gerenciamento de custos e geração de relatórios para aprimorar as análises e obter transparência dos dados de custo e uso. Configure a workload para criar entradas de log que facilitem o rastreamento e a segmentação de custos e uso.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Informações detalhadas de faturamento, como granularidade por hora nas ferramentas de gerenciamento de custos, permitem que as organizações acompanhem suas taxas de consumo com mais detalhes e as ajudem a identificar alguns dos motivos do aumento de custos. Essas fontes de dados oferecem a visualização mais precisa do custo e do uso em toda a organização.

Você pode usar o Exportações de dados da AWS para criar exportações do AWS Cost and Usage Report (CUR) 2.0. Ele é a nova forma recomendada de receber dados de custo e uso detalhados da AWS. Ele fornece granularidade de uso diário ou por hora, taxas, custos e atributos de uso para todos os serviços da AWS cobráveis (as mesmas informações do CUR), além de algumas melhorias. Todas as dimensões possíveis estão no CUR, incluindo marcação, localização, atributos de recurso e IDs de conta.

Há três tipos de exportação com base no tipo de exportação que você deseja criar: uma exportação de dados padrão, uma exportação para um painel de custos e uso integrado ao Amazon QuickSight ou uma exportação de dados herdados.

- Exportação de dados padrão: exportação personalizada de uma tabela que é entregue ao Amazon S3 de forma recorrente.
- Painel de Custos e Uso: exportação e integração com o Amazon QuickSight que implanta um painel pré-criado de custos e uso.

- Exportação de dados herdados: uma exportação do AWS Cost and Usage Report (CUR) herdado.

É possível criar exportações de dados com as seguintes personalizações:

- Incluir IDs de recurso
- Dados de alocação de custos divididos
- Detalhamento por hora
- Versionamento
- Tipo de compactação e formato de arquivo

Para workloads que executam contêineres no Amazon EC2 ou no Amazon EKS, habilite os dados de alocação de custos divididos para que você possa alocar seus custos de contêiner para unidades de negócios e equipes individuais, com base em como as workloads de contêiner consomem os recursos compartilhados de computação e memória. Os dados de alocação de custos divididos apresentam dados de custos e uso de novos recursos em nível de contêiner ao AWS Cost and Usage Report. Os dados de alocação de custos divididos são calculados computando-se os custos de serviços e tarefas individuais do ECS em execução no cluster.

Um painel de custos e uso exporta a tabela do painel de custos e uso para um bucket do S3 de forma recorrente e implanta um painel de custos e uso predefinido no Amazon QuickSight. Use essa opção se quiser implantar rapidamente um painel de seus dados de custos e uso sem a possibilidade de personalização.

Se desejar, você ainda poderá exportar o CUR no modo herdado, onde é possível integrar outros serviços de processamento, como o AWS Glue, para preparar os dados para análise e realizar análises de dados com o Amazon Athena usando SQL para consultar os dados.

Etapas de implementação

- Crie exportações de dados: crie exportações personalizadas com os dados desejados e controle o esquema das suas exportações. Crie exportações de dados de gerenciamento de custos e cobrança usando SQL básico e visualizar os dados de gerenciamento de custos e faturamento por meio da integração com o Amazon QuickSight. Você também pode exportar seus dados no modo padrão para analisá-los com outras ferramentas de processamento, como o Amazon Athena.
- Configure o relatório de custos e uso: usando o console de faturamento, configure pelo menos um relatório de custos e uso. Configure um relatório com granularidade por hora que inclua todos os

identificadores e IDs de recursos. Você também pode criar outros relatórios com diferentes níveis de detalhamento para fornecer informações resumidas de alto nível.

- Configure a granularidade por hora no Explorador de Custos: para acessar dados de custo e uso com granularidade horária dos últimos 14 dias, considere ativar dados horários e de nível de recursos no console de faturamento.
- Configure o log da aplicação: verifique se a aplicação registra cada resultado comercial entregue para que ele possa ser acompanhado e medido. Verifique se a granularidade desses dados é pelo menos por hora para corresponder aos dados de custo e uso. Para obter mais detalhes sobre registro e monitoramento, consulte Pilar Excelência operacional do Well-Architected.

Recursos

Documentos relacionados:

- Exportações de dados da AWS
- AWS Glue
- Amazon QuickSight
- Preços do gerenciamento de custos da AWS
- Marcando recursos do AWS
- Analisar custos com Explorador de Custos
- Gerenciar AWS Cost and Usage Reports
- Pilar Excelência operacional da Well-Architected

Exemplos relacionados:

- Configuração da conta da AWS
- Exportação de dados para o Gerenciamento de Faturamento e Custos da AWS
- Casos de uso comuns do AWS Cost Explorer

COST03-BP02 Adicionar informações da organização aos custos e ao uso

Defina um esquema de marcação com tags com base na sua organização, atributos da workload e categorias de alocação de custos para que você possa filtrar e pesquisar recursos ou monitorar custos e uso em ferramentas de gerenciamento de custos. Implemente marcação consistente

em todos os recursos, sempre que possível, por finalidade, equipe, ambiente ou outros critérios relevantes ao seu negócio.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Implemente a marcação com tags na AWS para adicionar informações da organização aos seus recursos, as quais serão adicionadas às suas informações de custo e uso. Uma tag é um par de chave-valor. A chave é definida e deve ser exclusiva em toda a organização, e o valor é exclusivo para um grupo de recursos. Um exemplo de par chave-valor é a chave `Environment`, com um valor de `Production`. Todos os recursos no ambiente de produção terão esse par de chave-valor. A marcação com tags permite categorizar e rastrear seus custos com informações relevantes e significativas da organização. Você pode aplicar tags que representem categorias de organização (como centros de custo, nomes de aplicações, projetos ou proprietários) e identificar workloads e características de workloads (como teste ou produção) para atribuir seus custos e uso em toda a organização.

Quando você aplica tags aos seus recursos da AWS (como instâncias do Amazon Elastic Compute Cloud ou buckets do Amazon Simple Storage Service) e as ativa, a AWS adiciona essas informações aos relatórios de custos e uso. Você pode gerar relatórios e realizar análises em recursos marcados e não marcados para permitir maior conformidade com políticas internas de gerenciamento de custos e garantir a atribuição precisa.

Criar e implementar um padrão de marcação da AWS em todas as contas da organização ajuda você a gerenciar e administrar seus ambientes da AWS de maneira consistente e uniforme. Use políticas de tag no AWS Organizations para definir regras de como as tags podem ser usadas em recursos da AWS nas suas contas no AWS Organizations. As políticas de tag permitem adotar facilmente uma abordagem padronizada para marcar os recursos da AWS.

O AWS Tag Editor permite adicionar, excluir e gerenciar tags de vários recursos. Com o Tag Editor, você pode pesquisar os recursos que deseja marcar e gerenciar as tags dos recursos nos resultados da pesquisa.

As Categorias de Custos da AWS permitem que você atribua significado da organização aos seus custos, sem exigir tags nos recursos. É possível mapear suas informações de custo e uso em estruturas internas exclusivas da organização. Você define regras de categoria para mapear e categorizar custos usando dimensões de faturamento, como contas e tags. Isso fornece outro nível de capacidade de gerenciamento, além da marcação. Você também pode mapear contas e tags específicas para vários projetos.

Etapas de implementação

- Defina um esquema de marcação com tags: reúna todas as partes interessadas de toda a sua empresa para definir um esquema. Isso geralmente inclui pessoas dos departamentos técnico, financeiro e de gerenciamento. Defina uma lista de tags que todos os recursos devem obrigatoriamente ter, bem como outra lista com as tags que os recursos poderiam ter. Verifique se os nomes e valores das tags são consistentes em toda a organização.
- Marque recursos com tags: usando suas categorias de atribuição de custo definidas, coloque tags em todos os recursos em suas workloads de acordo com as categorias. Use ferramentas como CLI, Tag Editor ou AWS Systems Manager para aumentar a eficiência.
- Implemente Categorias de Custos da AWS: você pode criar categorias de custos sem implementar a marcação com tags. As categorias de custos usam as dimensões de custo e uso existentes. Crie regras de categoria a partir do esquema e as implemente nas categorias de custos.
- Automatize a marcação com tags: para garantir que você mantenha altos níveis de marcação em todos os recursos, automatize a marcação com tags para que os recursos sejam marcados automaticamente quando forem criados. Use serviços como o AWS CloudFormation para verificar se os recursos são marcados quando criados. Você também pode criar uma solução personalizada para fazer a marcação com tags automaticamente usando funções do Lambda ou usar um microsserviço personalizado que verifica a workload periodicamente e remove todos os recursos que não estão marcados, o que é ideal para ambientes de teste e desenvolvimento.
- Monitore e gere relatórios de tags: para garantir que você mantenha altos níveis de marcação em toda a organização, relate e monitore as tags em todas as workloads. É possível usar o AWS Cost Explorer para visualizar o custo de recursos marcados e não marcados ou usar serviços como o Tag Editor. Analise regularmente o número de recursos não marcados com tags e tome medidas para adicionar etiquetas até atingir o nível desejado de marcação.

Recursos

Documentos relacionados:

- Práticas recomendadas de marcação com tags
- Tag de recurso do AWS CloudFormation
- AWS Cost Categories
- Marcando recursos do AWS
- Analisar os custos com o AWS Budgets

- Analisar custos com Explorador de Custos
- Gerenciar Relatórios de Custos e Uso da AWS

Vídeos relacionados:

- Como posso marcar meus recursos da AWS com tags para dividir minha fatura por centro de custos ou projeto?
- Marcar recursos da AWS com tags

COST03-BP03 Identificar categorias de atribuição de custos

Identifique categorias organizacionais, como unidades de negócios, departamentos ou projetos, que poderiam ser usadas para alocar custos em sua organização às entidades consumidoras internas. Use essas categorias para impor a responsabilidade de gastos, bem como promover a conscientização de custos e comportamentos de consumo eficazes.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

O processo de categorização de custos é crucial em orçamentos, contabilidade, relatórios financeiros, tomada de decisão, benchmarking e gerenciamento de projetos. Ao classificar e categorizar as despesas, as equipes podem entender melhor os tipos de custos que gerarão ao longo da jornada para a nuvem, ajudando-as a tomar decisões conscientes e gerenciar orçamentos de forma eficaz.

A responsabilidade pelos gastos com a nuvem estabelece um forte incentivo para o gerenciamento disciplinado da demanda e dos custos. O resultado é uma economia significativamente maior nos custos da nuvem para organizações que alocam a maior parte de seus gastos com a nuvem para unidades de negócios ou equipes consumidoras. Além disso, a alocação de gastos na nuvem ajuda as organizações a adotar mais práticas recomendadas de governança centralizada da nuvem.

Trabalhe com sua equipe financeira e outras partes interessadas relevantes para entender os requisitos de como os custos devem ser alocados em sua organização durante suas chamadas regulares. Os custos da workload devem ser alocados durante todo o ciclo de vida, incluindo desenvolvimento, teste, produção e desativação. Entenda como os custos incorridos para o aprendizado, o desenvolvimento da equipe e a criação de ideias são atribuídos na organização. Isso pode ser útil para alocar corretamente contas usadas para essa finalidade para orçamentos de treinamento e desenvolvimento, em vez de orçamentos genéricos de custo de TI.

Depois de definir suas categorias de atribuição de custos com as partes interessadas em sua organização, use Categorias de Custos da AWS para agrupar suas informações de custos e uso em categorias significativas na Nuvem AWS, como custo de um projeto específico, ou Contas da AWS para departamentos ou unidades de negócios. É possível criar categorias personalizadas e mapear as informações de custo e uso nessas categorias com base nas regras definidas usando várias dimensões, como conta, tag, serviço ou tipo de cobrança. Assim que as categorias de custos forem definidas, você verá as informações de custos e uso de acordo com elas, permitindo que a organização tome melhores decisões estratégicas e de compras. Também é possível ver essas categorias no AWS Cost Explorer, no AWS Budgets e no AWS Cost and Usage Report.

Por exemplo, é possível criar categorias de custos para suas unidades de negócios (equipe DevOps) e, em cada categoria, criar várias regras (para cada subcategoria) com várias dimensões (Contas da AWS, tags de alocação de custos, serviços ou tipo de cobrança) com base nos seus agrupamentos definidos. Com as categorias de custo, é possível organizar seus custos usando um mecanismo baseado em regras. As regras que você configurar organizarão seus custos em categorias. Dentro dessas regras, é possível aplicar filtros usando várias dimensões para cada categoria, como Contas da AWS, serviços da AWS ou tipos de cobrança específicos. Você pode usar essas categorias em vários produtos do console do Gerenciamento de Faturamento e Custos da AWS Billing and Cost Management. Isso inclui AWS Cost Explorer, AWS Budgets, AWS Cost and Usage Report e AWS Cost Anomaly Detection.

Como exemplo, o diagrama a seguir mostra como agrupar as informações de custos e uso em sua organização, com várias equipes (categoria de custos), vários ambientes (regras) e vários recursos ou ativos em cada ambiente (dimensões).

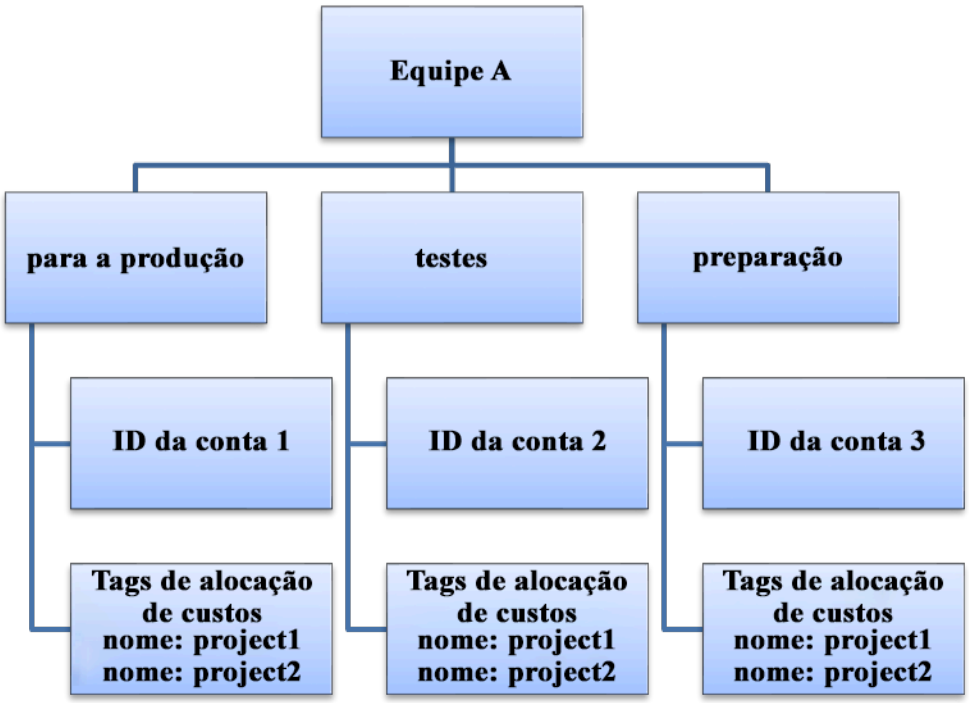


Tabela de organização de custos e uso

Também é possível criar agrupamento de custos usando as categorias de custos. Depois de criar as categorias de custos (aguardando até 24 horas após a criação de uma categoria para que seus registros de uso sejam atualizados com valores), elas aparecem no AWS Cost Explorer, AWS Budgets, AWS Cost and Usage Report e AWS Cost Anomaly Detection. No AWS Cost Explorer e no AWS Budgets, uma categoria de custos aparece como uma dimensão de faturamento adicional. Você pode usar isso para filtrar o valor da categoria de custo específica ou agrupar pela categoria de custo.

Etapas de implementação

- Defina as categorias da sua organização: reúna-se com as unidades de negócios e as partes interessadas internas para definir categorias que reflitam a estrutura e os requisitos da organização. Essas categorias devem ser associadas diretamente à estrutura das categorias financeiras existentes, como unidade de negócios, orçamento, centro de custo ou departamento. Veja os resultados que a nuvem oferece para a sua empresa, como treinamento ou educação, já que também são categorias de organização.
- Defina suas categorias funcionais: reúna-se com as unidades de negócios e as partes interessadas internas para definir categorias que reflitam as funções presentes na empresa. Essas

podem ser os nomes da workload ou da aplicação e o tipo de ambiente, como produção, teste ou desenvolvimento.

- Defina Categorias de Custos da AWS: crie categorias de custos para organizar suas informações de custo e uso usando Categorias de Custos da AWS e mapeie seus custos e uso da AWS em categorias significativas. Várias categorias podem ser atribuídas a um recurso, e um recurso pode estar em várias categorias diferentes. Portanto, defina quantas categorias forem necessárias para gerenciar seus custos dentro da estrutura categorizada usando Categorias de Custos da AWS.

Recursos

Documentos relacionados:

- Marcando recursos do AWS
- Usar tags de alocação de custos
- Analisar custos com o AWS Budgets
- Analisar custos com Explorador de Custos
- Gerenciar AWS Cost and Usage Reports
- AWS Cost Categories
- Gerenciar seus custos com as Categorias de Custos da AWS
- Criar categorias de custos
- Marcar categorias de custos com tags
- Separar cobranças em categorias de custos
- Recursos das Categorias de Custos da AWS

Exemplos relacionados:

- Organizar seus dados de custos e uso com as Categorias de Custos da AWS
- Gerenciar seus custos com as Categorias de Custos da AWS
- Laboratórios do Well-Architected: Visualização de custos e uso
- Laboratórios do Well-Architected: Categorias de custos

COST03-BP04 Estabelecer métricas da organização

Estabeleça as métricas da organização que são necessárias para esta workload. Exemplo de métricas de uma workload são relatórios de clientes produzidos ou páginas da Web veiculadas para os clientes.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Entenda como a saída da workload é medida em relação ao sucesso empresarial. Cada workload normalmente tem um pequeno conjunto de saídas principais que indicam a performance. Se você tiver uma workload complexa com muitos componentes, poderá priorizar a lista ou definir e rastrear métricas para cada componente. Trabalhe com suas equipes para entender quais métricas usar. Essa unidade será usada para compreender a eficiência da workload ou o custo de cada saída de negócios.

Etapas de implementação

- Defina resultados da workload: reúna-se com as partes interessadas da empresa e defina os resultados para a workload. Essas são medidas principais de uso do cliente e devem ser métricas de negócios, e não técnicas. Deve haver um pequeno número de métricas de alto nível (menos de cinco) por workload. Se a workload produzir vários resultados para diferentes casos de uso, agrupe-os em uma única métrica.
- Definir resultados dos componentes da workload: opcionalmente, se você tiver uma workload grande e complexa ou puder dividir facilmente sua workload em componentes (como microsserviços) com entradas e saídas bem definidas, defina métricas para cada componente. O esforço deve refletir o valor e o custo do componente. Comece com os maiores componentes e trabalhe em direção aos componentes menores.

Recursos

Documentos relacionados:

- Marcando recursos do AWS
- Analisar os custos com o AWS Budgets
- Analisar custos com Explorador de Custos
- Gerenciar Relatórios de Custos e Uso da AWS

COST03-BP05 Configurar ferramentas de faturamento e gerenciamento de custos

Configure ferramentas de gerenciamento de custos que atendam às políticas da sua organização para gerenciar e otimizar gastos com a nuvem. Isso inclui serviços, ferramentas e recursos para organizar e monitorar dados de custos e uso, aprimorar o controle por meio de faturamento consolidado e permissão de acesso, melhorar o planejamento por meio de orçamento e previsões, receber notificações ou alertas e reduzir os custos com recursos e otimizações de preços.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Para estabelecer uma responsabilização consistente, pense primeiro na estratégia de conta como parte de sua estratégia de alocação de custos. Faça isso do jeito certo e talvez não precise ir além. Caso contrário, poderá haver inconsciência e outros pontos problemáticos.

Para incentivar a responsabilização pelos gastos com a nuvem, conceda aos usuários acesso a ferramentas que forneçam visibilidade sobre custos e uso. A AWS recomenda configurar todas as workloads e as equipes para as seguintes finalidades:

- Organize: estabeleça a alocação de custos e linha de base de governança com sua própria estratégia de marcação e taxonomia. Crie várias contas da AWS com ferramentas como o AWS Control Tower ou o AWS Organization. Marque os recursos compatíveis da AWS com tags e categorize-os de uma maneira que faça sentido com base na estrutura da sua organização (unidades de negócios, departamentos ou projetos). Marque nomes de contas para centros de custo específicos e associe-os a Categorias de Custos da AWS para agrupar contas de unidades de negócios nos centros de custos de modo que o proprietário da unidade de negócios possa ver o consumo de várias contas em um só lugar.
- Acesso: rastreie as informações de faturamento de toda a organização no faturamento consolidado. Verifique se as partes interessadas e os proprietários de negócios certos têm acesso.
- Controle: crie mecanismos de governança eficazes com as barreiras de proteção certas para evitar cenários inesperados ao usar políticas de controle de serviços (SCPs), políticas de tags, políticas do IAM e alertas de orçamento. Por exemplo, é possível permitir que as equipes criem recursos específicos nas regiões preferidas somente usando mecanismos de controle eficazes, bem como evitar a criação de recursos sem uma tag específica (como centro de custos).
- Estado atual: configure um painel que mostra os níveis atuais de custo e uso. O painel deve estar disponível em um local altamente visível dentro do ambiente de trabalho, de modo semelhante a um painel de operações. É possível exportar dados e usar o Painel de Custos e Uso do Hub de Otimização de Custos da AWS ou qualquer produto compatível para criar essa visibilidade. Talvez

seja necessário criar painéis diferentes para pessoas diferentes. Por exemplo, o painel do gerente pode ser diferente de um painel de engenharia.

- Notificações: forneça notificações quando o custo ou o uso excederem os limites definidos e ocorrerem anomalias com o AWS Budgets ou a Detecção de Anomalias em Custos da AWS.
- Relatórios: resuma todas as informações de custos e uso. Aumente a conscientização e a responsabilização sobre seus gastos com a nuvem com dados de custos detalhados e alocáveis. Crie relatórios que sejam relevantes para a equipe que os consome e contenham recomendações.
- Rastreamento: mostre os custos e o uso atuais em relação a metas ou objetivos configurados.
- Análise: permita que os membros da equipe realizem análises personalizadas e profundas até a granularidade por hora, diária ou mensal com filtros diferentes (recurso, conta, tag etc.).
- Inspeção: mantenha-se em dia com suas oportunidades de otimização de custos e implantação de recursos. Receba notificações usando o Amazon CloudWatch, o Amazon SNS ou o Amazon SES para implantações de recursos na organização. Analise as recomendações de otimização de custos com o AWS Trusted Advisor ou o AWS Compute Optimizer.
- Relatórios de tendências: exiba a variabilidade de custo e uso ao longo do período necessário, com a granularidade necessária.
- Previsões: mostre os custos futuros estimados, faça uma estimativa do uso de recursos e gaste com painéis de previsão criados por você.

Você pode usar o Hub de Otimização de Custos da AWS para entender possíveis oportunidades de redução de custos consolidadas a partir de um local centralizado e criar exportações de dados para integração com o Amazon Athena. Você também pode usar o Hub de Otimização de Custos da AWS para implantar o Painel de Custos e Uso, que utiliza o Amazon QuickSight para análise interativa de custos e compartilhamento seguro de insights de custos.

Se você não conta com as habilidades essenciais ou a largura de banda necessária em sua organização, então poderá trabalhar com o AWS ProServ, o AWS Managed Services (AMS) ou parceiros da AWS. Você também pode usar ferramentas de terceiros, mas não se esqueça de validar a proposta de valor.

Etapas de implementação

- Permita o acesso baseado em equipe às ferramentas: configure suas contas e crie grupos que tenham acesso aos relatórios de custos e uso necessários para seus consumos e use o AWS Identity and Access Managementpara controlar o acesso a ferramentas como o AWS Cost Explorer. Esses grupos devem incluir representantes de todas as equipes que possuem ou

gerenciam uma aplicação. Isso garante que cada equipe tenha acesso às próprias informações de custo e uso para monitorar o consumo.

- Organize tags e categorias de custos: organize seus custos entre equipes, unidades de negócios, aplicações, ambientes e projetos. Use tags de recursos para organizar custos por tags de alocação de custos. Crie categorias de custos com base nas dimensões usando tags, contas, serviços etc. para mapear os custos.
- Configure o AWS Budgets: configure o AWS Budgets em todas as contas para suas workloads. Defina orçamentos para o gasto geral da conta e orçamentos para as workloads usando tags e categorias de custos. Configure notificações no AWS Budgets para receber alertas quando você exceder valores orçados ou quando os custos estimados excederem seus orçamentos.
- Configure a Detecção de Anomalias em Custos da AWS: use a Detecção de Anomalias em Custos da AWS para as contas, os serviços centrais ou as categorias de custos criadas para monitorar os custos e o uso e detectar gastos incomuns. É possível receber alertas individualmente em relatórios agregados, assim como alertas por e-mail ou em um tópico do Amazon SNS, o que permite analisar e determinar a causa-raiz de uma anomalia e identificar o fator que está aumentando o custo.
- Use ferramentas de análise custos: configure o AWS Cost Explorer para sua workload e contas para visualizar seus dados de custos para análise posterior. Crie um painel para a workload que rastreie o gasto geral, as principais métricas de uso da workload e a previsão de custos futuros com base nos seus dados de custo históricos.
- Use ferramentas de análise de redução de custos: use o Hub de Otimização de Custos da AWS para identificar oportunidades de economia com recomendações personalizadas, incluindo exclusão de recursos não utilizados, dimensionamento correto, Savings Plans, reservas e recomendações de otimizadores de computação.
- Configure ferramentas avançadas: opcionalmente, é possível criar recursos visuais para facilitar a análise interativa e o compartilhamento de informações sobre custos. Com as exportações de dados no Hub de Otimização de Custos da AWS, é possível criar um painel de custo e uso viabilizado pelo Amazon QuickSight para sua organização que fornece detalhes adicionais e granularidade. Você também pode implementar recursos avançados de análise usando exportações de dados no Amazon Athena para consultas avançadas e criar painéis no Amazon QuickSight. Trabalhe com Parceiros da AWS para adotar soluções de gerenciamento de nuvem para monitoramento e otimização consolidados de faturas da nuvem.

Recursos

Documentos relacionados:

- O que é o gerenciamento de custos da AWS Billing and Cost Management?
- Estabelecer seu ambiente de práticas recomendadas da AWS
- Práticas recomendadas para marcação de recursos da AWS com tags
- Marcar recursos da AWS
- AWS Cost Categories
- Analisar os custos com o AWS Budgets
- Analisar custos com o AWS Cost Explorer
- O que são exportações de dados da AWS?

Vídeos relacionados:

- Implantar Cloud Intelligence Dashboards
- Receber alertas sobre qualquer métrica ou KPI de FinOps ou otimização de custos

Exemplos relacionados:

- Painel de Custos e Uso baseado no Amazon QuickSight
- Workshop Governança de custos e uso na AWS

COST03-BP06 Alocar custos com base nas métricas de workload

Aloque os custos da workload com base em métricas de uso ou resultados de negócios para medir o custo-benefício da workload. Implemente um processo para analisar os dados de custo e uso com serviços de análise, que podem fornecer informações e capacidade de estorno.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Otimizar os custos significa fornecer resultados de negócios com o menor preço, o que só pode ser alcançado por meio da alocação dos custos da workload com base nas métricas da workload (medidas pela eficiência da workload). Monitore as métricas de workload definidas por meio de

arquivos de log ou outro monitoramento de aplicações. Combine esses dados com os custos da workload, os quais podem ser obtidos examinando-se os custos com um valor de tag ou ID de conta específico. Execute essa análise por hora. Sua eficiência normalmente muda se você tem componentes de custo estático (por exemplo, um banco de dados de backend em execução de maneira permanente) com uma taxa de solicitações variável (por exemplo, picos de uso entre 9h e 17h e poucas solicitações à noite). Entender a relação entre os custos estáticos e variáveis ajuda você a concentrar suas atividades de otimização.

Criar métricas de workload para recursos compartilhados pode ser um desafio em comparação com recursos como aplicações em contêineres no Amazon Elastic Container Service (Amazon ECS) e no Amazon API Gateway. No entanto, existem algumas maneiras de categorizar o uso e rastrear os custos. Se precisar monitorar recursos compartilhados do AWS Batch e do Amazon ECS, você poderá habilitar os dados de alocação de custos divididos no AWS Cost Explorer. Com dados de alocação de custos divididos, você pode entender e otimizar o custo e o uso de suas aplicações em contêineres e alocar os custos das aplicações para entidades comerciais individuais com base na forma como os recursos compartilhados de computação e memória são consumidos.

Etapas de implementação

- Aloque custos a métricas da workload: usando as métricas definidas e a tags configuradas, crie uma métrica que combine a saída e o custo da workload. Use serviços de análise, como o Amazon Athena e o Amazon QuickSight, para criar um painel de eficiência para a workload geral e todos os componentes.

Recursos

Documentos relacionados:

- Marcando recursos do AWS
- Analisar os custos com o AWS Budgets
- Analisar custos com Explorador de Custos
- Gerenciar Relatórios de Custos e Uso da AWS

Exemplos relacionados:

- Melhorar a visibilidade de custos do Amazon ECS e do AWS Batch com dados de alocação de custos divididos da AWS

COST 4. Como desativar recursos?

Implemente o controle de alterações e o gerenciamento de recursos, desde o início do projeto até o fim da vida útil. Isso garante o desligamento ou encerramento dos recursos não utilizados para reduzir o desperdício.

Práticas recomendadas

- COST04-BP01 Rastrear os recursos ao longo da vida útil
- COST04-BP02 Implementar um processo de desativação
- COST04-BP03 Desativar recursos
- COST04-BP04 Desativar recursos automaticamente
- COST04-BP05 Impor políticas de retenção de dados

COST04-BP01 Rastrear os recursos ao longo da vida útil

Defina e implemente um método para rastrear recursos e suas associações com sistemas ao longo da vida útil. A marcação com tags pode ser usada para identificar a workload ou a função do recurso.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Desative recursos de workload que não são mais necessários. Um exemplo comum são os recursos usados para testes: após a conclusão do teste, os recursos podem ser removidos. O rastreamento de recursos com tags (e execução de relatórios sobre essas tags) pode ajudar a identificar ativos para desativação, pois eles não estarão em uso ou a licença deles expirará. Usar tags é uma maneira eficaz de rastrear recursos por meio da rotulagem do recurso com sua função ou uma data conhecida em que ele pode ser desativado. Os relatórios podem ser executados nessas tags. Os valores de exemplo para marcação de recursos são `feature-X testing` para identificar a finalidade do recurso em termos de ciclo de vida da workload. Outro exemplo é usar `LifeSpan` ou TTL para os recursos, como o nome e o valor da chave da tag a ser excluída, para definir o período ou o horário específico para a desativação.

Etapas de implementação

- Implemente um esquema de marcação com tags: implemente um esquema de marcação que identifique a workload à qual o recurso pertence, verificando se todos os recursos dentro da workload estão marcados da maneira apropriada. A marcação ajuda a categorizar os recursos

por finalidade, equipe, ambiente ou outros critérios relevantes para o seu negócio. Para obter mais detalhes sobre casos de uso, estratégias e técnicas de marcação, consulte Práticas recomendadas de marcação com tags da AWS.

- Implemente o monitoramento de throughput ou saída da workload: implemente o monitoramento ou alarme de throughput da workload, iniciando nas solicitações de entrada ou na conclusão da saída. Configure-o para fornecer notificações quando saídas ou solicitações de workload caírem para zero, indicando que os recursos de workload não são mais usados. Incorpore um fator de tempo se a workload cair periodicamente para zero em condições normais. Para obter mais detalhes sobre recursos não utilizados ou subutilizados, consulte Verificações de otimização de custos da AWS Trusted Advisor.
- Agrupe os recursos da AWS: crie grupos de recursos para seus recursos da AWS. Você pode usar o AWS Resource Groups para organizar e gerenciar seus recursos da AWS que fazem parte da mesma Região da AWS. É possível adicionar tags à maioria de seus recursos para ajudar a identificá-los e classificá-los em sua organização. Use o Tag Editor para adicionar tags em massa aos recursos compatíveis. Considere usar o AWS Service Catalog para criar, gerenciar e distribuir portfólios de produtos aprovados para usuários finais e gerenciar o ciclo de vida de seus produtos.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Trusted Advisor
- Verificações de otimização de custos da AWS Trusted Advisor
- Marcando recursos do AWS
- Publicar métricas personalizadas

Vídeos relacionados:

- Como otimizar custos usando o AWS Trusted Advisor

Exemplos relacionados:

- Organizar recursos da AWS
- Otimizar custos usando o AWS Trusted Advisor

COST04-BP02 Implementar um processo de desativação

Implemente um processo para identificar e desativar recursos não utilizados.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Implemente um processo padronizado em toda a organização para identificar e remover recursos não utilizados. O processo deve definir a frequência das pesquisas e os processos para remover o recurso para verificar se todos os requisitos da organização foram atendidos.

Etapas de implementação

- Crie e implemente um processo de desativação: trabalhe com os proprietários e desenvolvedores da workload para criar um processo de desativação para a workload e seus respectivos recursos. O processo deve abranger o método para verificar se a workload está em uso e também se cada um dos recursos da workload está em uso. Detalhe as etapas necessárias para desativar o recurso, removendo-os do serviço e garantindo a conformidade com os requisitos normativos. Todos os recursos associados, como licenças ou armazenamento anexado, devem ser incluídos. Notifique os proprietários da workload de que o processo de desativação foi iniciado.

  Use as seguintes etapas de desativação para obter orientações sobre o que deve ser verificado como parte do seu processo:

  - Identifique os recursos a serem desativados: identifique os recursos que são elegíveis para desativação em sua Nuvem AWS. Registre todas as informações necessárias e agende a desativação. Em sua linha do tempo, certifique-se de considerar se (e quando) problemas inesperados surgirem durante o processo.
  - Coordene e se comunique: trabalhe com os proprietários da workload para confirmar o recurso que será desativado
  - Registre metadados e crie backups: registre metadados (como IPs públicos, região, AZ, VPC, sub-rede e grupos de segurança) e crie backups (como snapshots do Amazon Elastic Block Store ou AMI, exportação de chaves e exportação de certificados) se eles forem necessários para os recursos no ambiente de produção ou se forem recursos essenciais.
  - Valide a infraestrutura como código: determine se os recursos foram implantados com o AWS CloudFormation, Terraform, AWS Cloud Development Kit (AWS CDK) ou qualquer outra ferramenta de implantação de infraestrutura como código para que possam ser reimplantados, se necessário.

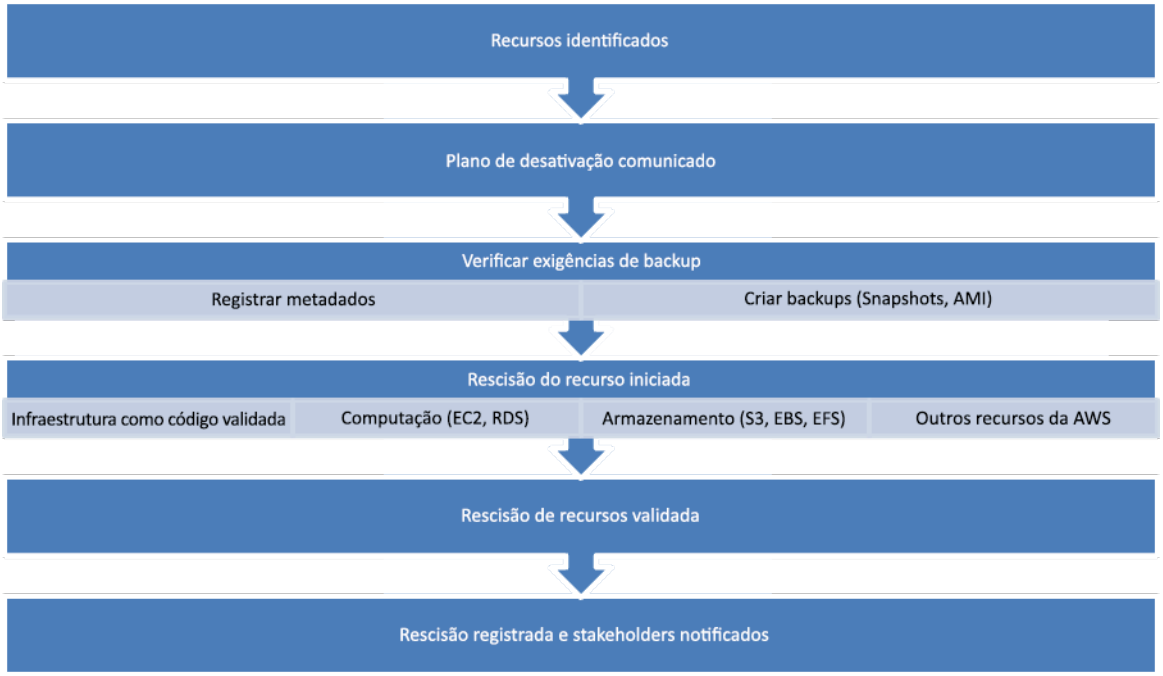
- Impeça o acesso: aplique controles restritivos por um período de tempo para evitar o uso de recursos enquanto você determina se o recurso é necessário. Verifique se o ambiente de recursos pode ser revertido para seu estado original, se necessário.
- Siga seu processo interno de desativação: siga as tarefas administrativas e o processo de desativação praticado por sua organização, como remover o recurso do domínio da organização, remover o registro DNS e remover o recurso das ferramentas de gerenciamento de configuração, de monitoramento, de automação e de segurança.

Se o recurso for uma instância do Amazon EC2, consulte a lista a seguir. Para obter mais detalhes, consulte Como faço para excluir ou terminar meus recursos do Amazon EC2?

- Pare ou encerre todas as suas instâncias e balanceadores de carga do Amazon EC2. Observação: as instâncias do EC2 permanecem visíveis no console por um breve período depois de terminadas. Você não será cobrado por instâncias que não estiverem em estado de execução
- Exclua sua infraestrutura do Auto Scaling.
- Libere todos os hosts dedicados.
- Exclua todos os volumes e snapshots do Amazon EBS.
- Libere todos os endereços IP elásticos.
- Cancele o registro das imagens de máquina da Amazon (AMIs).
- Encerre todos os ambientes do AWS Elastic Beanstalk.

Se o recurso for um objeto armazenado no e se você excluir um arquivo antes de atingir a duração mínima de armazenamento, uma taxa proporcional de exclusão antecipada será cobrada. A duração mínima de armazenamento do Amazon S3 Glacier depende da classe de armazenamento usada. Para obter um resumo da duração mínima de armazenamento para cada classe de armazenamento, consulte Performance nas classes de armazenamento do Amazon S3. Para obter detalhes sobre como as taxas de exclusão antecipada são calculadas, consulte Preços do Amazon S3.

O fluxograma simples do processo de desativação a seguir descreve as etapas de desativação. Antes de desativar recursos, verifique se os recursos que você identificou para desativação não estão sendo usados pela organização.



Fluxo de desativação de recursos.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Trusted Advisor
- AWS CloudTrail

Vídeos relacionados:

- Excluir a pilha do CloudFormation, mas reter alguns recursos
- Descobrir qual usuário iniciou a instância do Amazon EC2

Exemplos relacionados:

- Excluir ou encerrar recursos do Amazon EC2
- Descobrir qual usuário iniciou a instância do Amazon EC2

COST04-BP03 Desativar recursos

Desative recursos iniciados por eventos, como auditorias periódicas ou alterações no uso. Em geral, a desativação pode ser realizada periodicamente e é manual ou automatizada.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A frequência e o esforço para pesquisar recursos não utilizados devem refletir as possíveis economias, portanto, uma conta com custo pequeno deve ser analisada com menos frequência do que uma conta com custos maiores. Pesquisas e eventos de desativação podem ser iniciados por alterações de estado na workload, como um produto que termina a vida útil ou é substituído. Pesquisas e eventos de desativação também podem ser iniciados por eventos externos, como alterações nas condições de mercado ou encerramento do produto.

Etapas de implementação

- Desative recursos: esse é o estágio de depreciação de recursos da AWS que não são mais necessários ou o término de um contrato de licenciamento. Conclua todas as verificações finais antes de avançar para o estágio de descarte e desativação de recursos para evitar interrupções indesejadas, como criar snapshots ou fazer backups. Usando o processo de desativação, desative cada um dos recursos que foram identificados como não utilizados.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Trusted Advisor

Exemplos relacionados:

- Laboratórios do Well-Architected: Desativar recursos (Nível 100)

COST04-BP04 Desativar recursos automaticamente

Projete a workload para lidar normalmente com o encerramento de recursos ao identificar e desativar recursos não críticos, que não são necessários ou com baixa utilização.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Use a automação para reduzir ou remover os custos associados do processo de desativação. Projetar sua workload para executar a desativação automatizada reduzirá os custos gerais da workload durante sua vida útil. Você pode usar o Amazon EC2 Auto Scaling ou o Application Auto Scaling para realizar o processo de desativação. Você também pode implementar código personalizado usando a API ou o SDK para desativar recursos de workload automaticamente.

As aplicações modernas são criadas primeiro sem servidor, uma estratégia que prioriza a adoção de serviços sem servidor. A AWS desenvolveu serviços com tecnologia sem servidor para todas as três camadas da pilha: computação, integração e datastores. O uso da arquitetura sem servidor permitirá que você reduza os custos durante períodos de baixo tráfego com aumento e redução automáticos.

Etapas de implementação

- Implemente o Amazon EC2 Auto Scaling ou o Application Auto Scaling: para recursos compatíveis, configure-os com o Amazon EC2 Auto Scaling ou com o Application Auto Scaling. Esses serviços podem ajudar você a otimizar sua utilização e eficiência de custos ao consumir serviços da AWS. Quando a demanda cair, esses serviços removerão automaticamente qualquer excesso de capacidade de recursos para evitar gastos excessivos.
- Configure o CloudWatch para encerrar instâncias: as instâncias podem ser configuradas para encerrar usando os alarmes do CloudWatch. Usando as métricas do processo de desativação, implemente um alarme com uma ação do Amazon Elastic Compute Cloud. Verifique a operação em um ambiente de não produção antes de implantar.
- Implemente código dentro da workload: você pode usar o AWS SDK ou a AWS CLI para desativar recursos da workload. Implemente código dentro da aplicação que se integre à AWS e encerre ou remova recursos que não são mais usados.
- Use serviços sem servidor: priorize a criação de arquiteturas sem servidor e arquiteturas orientadas a eventos na AWS para criar e suas aplicações. A AWS oferece vários serviços de tecnologia sem servidor que, inerentemente, fornecem automaticamente a utilização otimizada de recursos e a desativação automatizada (expansão e redução). Com aplicações sem servidor, a utilização de recursos é otimizada automaticamente e você nunca paga por provisionamento em excesso.

Recursos

Documentos relacionados:

- Amazon EC2 Auto Scaling
- Conceitos básicos do Amazon EC2 Auto Scaling
- Application Auto Scaling
- AWS Trusted Advisor
- Tecnologia sem servidor na AWS
- Criar alarmes para parar, encerrar, reinicializar ou recuperar uma instância
- Adicionar ações de encerrar para a alarmes do Amazon CloudWatch

Exemplos relacionados:

- Agendar a exclusão automática de pilhas do AWS CloudFormation
- Laboratórios do Well-Architected: Desativar recursos automaticamente (Nível 100)
- AWS Auto Cleanup da Servian

COST04-BP05 Impor políticas de retenção de dados

Defina as políticas de retenção de dados em recursos compatíveis para lidar com exclusão de objetos de acordo com os requisitos de suas organizações. Identifique e exclua recursos e objetos desnecessários ou órfãos que não sejam mais necessários.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Use políticas de retenção de dados e de ciclo de vida para reduzir os custos associados do processo de desativação e de armazenamento dos recursos identificados. A definição de suas políticas de retenção de dados e de ciclo de vida para realizar a exclusão e a migração automatizadas de classe de armazenamento reduzirá os custos gerais de armazenamento durante seu tempo de vida. Você pode usar o Amazon Data Lifecycle Manager para automatizar a criação e a exclusão de snapshots do Elastic Block Store e imagens de máquina (AMIs) baseadas no Amazon EBS, e usar o Amazon S3 Intelligent-Tiering ou uma configuração de ciclo de vida do Amazon S3 para gerenciar o ciclo de vida de seus objetos do Amazon S3. Também é possível pode implementar código personalizado usando a API ou o SDK para criar políticas de ciclo de vida e regras de política para que objetos sejam excluídos automaticamente.

Etapas de implementação

- Use o Amazon Data Lifecycle Manager: use políticas de ciclo de vida no Amazon Data Lifecycle Manager para automatizar a exclusão de snapshots do Amazon EBS e de AMIs baseadas no Amazon EBS.
- Defina a configuração do ciclo de vida em um bucket: use a configuração do ciclo de vida do Amazon S3 em um bucket para definir ações a serem tomadas pelo Amazon S3 durante o ciclo de vida de um objeto, bem como a exclusão no final do ciclo de vida do objeto, com base nos requisitos de sua empresa.

Recursos

Documentos relacionados:

- AWS Trusted Advisor
- Amazon Data Lifecycle Manager
- Como definir uma configuração de ciclo de vida em um bucket do Amazon S3

Vídeos relacionados:

- Automatizar snapshots do Amazon EBS com o Amazon Data Lifecycle Manager
- Esvaziar um bucket do Amazon S3 usando uma regra de configuração de ciclo de vida

Exemplos relacionados:

- Esvaziar um bucket do Amazon S3 usando uma regra de configuração de ciclo de vida
- Laboratório do Well-Architected: Desativar recursos automaticamente (Nível 100)

Recursos economicamente eficientes

Perguntas

- COST 5. Como avaliar o custo ao selecionar serviços?
- COST 6. Como atingir as metas de custo ao selecionar tamanho, número e tipo de recurso?
- COST 7. Como usar os modelos de preços para reduzir custos?
- COST 8. Como planejar as cobranças de transferência de dados?

COST 5. Como avaliar o custo ao selecionar serviços?

O Amazon EC2, Amazon EBS e Amazon S3 são produtos fundamentais da AWS. Os produtos gerenciados, como Amazon RDS e Amazon DynamoDB, são serviços da AWS de nível superior ou de aplicação. Ao selecionar os produtos fundamentais e os serviços gerenciados adequados, é possível otimizar os custos dessa workload. Por exemplo, usando serviços gerenciados, é possível reduzir ou remover grande parte da sobrecarga administrativa e operacional, liberando você para trabalhar em aplicações e atividades relacionadas a negócios.

Práticas recomendadas

- COST05-BP01 Identificar os requisitos de custos da organização
- COST05-BP02 Analisar todos os componentes da workload
- COST05-BP03 Executar uma análise completa de cada componente
- COST05-BP04 Selecionar software com licenciamento econômico
- COST05-BP05 Selecionar os componentes desta workload para otimizar o custo alinhado com as prioridades da organização
- COST05-BP06 Realizar análises de custos para diferentes usos ao longo do tempo

COST05-BP01 Identificar os requisitos de custos da organização

Trabalhe com os membros da equipe para definir o equilíbrio entre otimização de custos e outros pilares, como performance e confiabilidade, para essa workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Na maioria das organizações, o departamento de tecnologia da informação (TI) é composto de várias equipes pequenas, cada uma com sua própria agenda e área de foco, que refletem as especialidades e as habilidades dos respectivos membros. Você precisa compreender os objetivos, as prioridades e as metas gerais da organização e como cada departamento ou projeto contribui para esses objetivos. A categorização de todos os recursos essenciais, incluindo pessoal, equipamentos, tecnologia, materiais e serviços externos, é crucial para alcançar os objetivos organizacionais e um planejamento orçamentário abrangente. A adoção dessa abordagem sistemática para a identificação e a compreensão dos custos é fundamental para estabelecer um plano de custos realista e robusto para a organização.

ao selecionar serviços para a sua workload, é fundamental compreender as prioridades da sua organização. Crie um equilíbrio entre a otimização de custos e outros pilares do AWS Well-Architected Framework, como performance e confiabilidade. Esse processo deve ser conduzido de forma sistemática e regular para refletir as mudanças nos objetivos da organização, nas condições de mercado e na dinâmica operacional. Uma workload totalmente otimizada para custo é a solução mais alinhada aos requisitos da sua organização, mas não necessariamente o menor custo. Reúna-se com todas as equipes da organização, como produtos, negócios, técnicas e finanças, para coletar as informações. Avalie o impacto das compensações entre interesses concorrentes ou abordagens alternativas para ajudar a tomar decisões fundamentadas ao determinar onde concentrar as iniciativas ou escolher um plano de ação.

Por exemplo, a aceleração da velocidade de entrada no mercado de novos recursos pode ser enfatizada em relação à otimização de custos, ou você pode escolher um banco de dados relacional para dados não relacionais para simplificar o esforço de migração de um sistema, em vez de migrar para um banco de dados otimizado para seu tipo de dados e atualizar a aplicação.

Etapas de implementação

- Identifique os requisitos relacionados a custos da organização: reúna-se com membros da equipe da sua organização que incluam gerenciamento de produtos, proprietários de aplicações, equipes de desenvolvimento e operações, gerenciamento e finanças. Priorize os pilares do Well-Architected para essa workload e os respectivos componentes. O resultado deve ser uma lista ordenada dos pilares. Também é possível adicionar um peso a cada pilar para indicar quanto foco adicional ele tem ou o quanto semelhantes são os focos entre dois pilares.
- Resolva a dívida técnica e documente-a: durante a revisão da workload, resolva a dívida técnica. Documente um item de backlog para revisitar a workload no futuro com o objetivo de refatorar ou rearquitetar para otimizá-la ainda mais. É essencial comunicar claramente as compensações feitas para outras partes interessadas.

Recursos

Práticas recomendadas relacionadas:

- REL11-BP07 Arquitetar o produto para cumprir as metas de disponibilidade e os acordos de serviço (SLAs) de tempo de atividade
- OPS01-BP06 Avaliar compensações

Documentos relacionados:

- Calculadora de custo total de propriedade (TCO) da AWS
- Classes de armazenamento do Amazon S3
- Produtos de nuvem

COST05-BP02 Analisar todos os componentes da workload

Verifique se cada componente da workload é analisado, independentemente do tamanho ou dos custos atuais. O trabalho da análise deve refletir o benefício potencial, como os custos atuais e projetados.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Os componentes da workload, projetados para agregar valor comercial à organização, podem abranger vários serviços. Para cada componente, é possível escolher serviços específicos da Nuvem AWS para atender às necessidades dos negócios. Essa seleção pode ser influenciada por fatores como a familiaridade ou a experiência anterior com esses serviços.

Depois de identificar os requisitos de sua organização, conforme mencionado em COST05-BP01 Identificar os requisitos de custos da organização, faça uma análise completa de todos os componentes de sua workload. Analise cada componente levando em conta os custos e os tamanhos atuais e projetados. Pense no custo da análise em relação a qualquer possível economia da workload ao longo do respectivo ciclo de vida. O trabalho despendido na análise de todos os componentes dessa workload deve corresponder às possíveis economias ou melhorias previstas da otimização desse componente específico. Por exemplo, se o custo do recurso proposto for USD 10/mês e, sob as cargas previstas, não exceder USD 15/mês, gastar um dia de trabalho para reduzir os custos em 50% (USD 5 por mês) poderá exceder o benefício potencial durante a vida útil do sistema. Use uma estimativa baseada em dados mais rápida e eficiente para criar o melhor resultado geral para esse componente.

As workloads podem mudar ao longo do tempo, e o conjunto certo de serviços poderá não ser ideal se a arquitetura da workload ou o uso mudarem. A análise para seleção de serviços deve incorporar estados de workload e níveis de uso atuais e futuros. A implementação de um serviço para o estado ou o uso futuro da workload pode reduzir os custos gerais ao reduzir ou remover o esforço necessário para fazer alterações futuras. Por exemplo, usar o EMR sem servidor pode ser

a escolha apropriada inicialmente. No entanto, à medida que o consumo desse serviço aumenta, a transição para o EMR no EC2 pode reduzir os custos desse componente da workload.

O AWS Cost Explorer e os AWS Cost and Usage Reports CUR podem analisar o custo de uma prova de conceito (PoC) ou um ambiente em execução. Você também pode usar o AWS Pricing Calculator para estimar os custos da workload.

Crie um fluxo de trabalho para ser seguido pelas equipes técnicas para analisar as workloads. Mantenha esse fluxo de trabalho simples, mas também abranja todas as etapas necessárias para garantir que as equipes entendam cada componente da workload e seus preços. Sua organização pode então acompanhar e personalizar esse fluxo de trabalho com base nas necessidades específicas de cada equipe.

1. Liste cada serviço em uso para sua workload: esse é um bom ponto de partida. Identifique todos os serviços em uso no momento e a origem dos custos.
2. Entenda como os preços funcionam para esses serviços: entenda o modelo de preços de cada serviço. Diferentes serviços da AWS têm modelos de preço diferentes com base em fatores como volume de uso, transferência de dados e preços específicos de recursos.
3. Concentre-se nos serviços que têm custos inesperados de workload e que não estão alinhados ao uso esperado e ao resultado comercial: identifique valores atípicos ou serviços em que o custo não seja proporcional ao valor ou ao uso utilizando o AWS Cost Explorer ou o AWS Cost and Usage Report. É importante correlacionar os custos com os resultados comerciais para priorizar os esforços de otimização.
4. AWS Cost Explorer, CloudWatch Logs, Logs de fluxo da VPC e Lente de Armazenamento do Amazon S3 para entender a causa-raiz desses altos custos: essas ferramentas são fundamentais no diagnóstico de custos elevados. Cada serviço oferece uma lente diferente para visualizar e analisar o uso e os custos. Por exemplo, o Explorador de Custos ajuda a determinar tendências gerais de custos, o CloudWatch Logs fornece insights operacionais, os Logs de fluxo da VPC exibem o tráfego IP e a Lente de Armazenamento do Amazon S3 é útil para análises de armazenamento.
5. Use o AWS Budgets para definir orçamentos para determinados valores para serviços ou contas: definir orçamentos é uma forma proativa de gerenciar custos. Use o AWS Budgets para definir limites de orçamento personalizados e receber alertas quando os custos excederem esses limites.
6. Configure os alarmes do Amazon CloudWatch para enviar alertas de faturamento e uso: configure o monitoramento e alertas para métricas de custos e uso. Os alarmes do CloudWatch podem notificar você quando determinados limites forem violados, o que melhora o tempo de resposta da intervenção.

Promova melhorias notáveis e economias financeiras ao longo do tempo por meio da análise estratégica de todos os componentes da workload e independentemente de seus atributos atuais. O esforço investido nesse processo de análise deve ser deliberado, com consideração cuidadosa das vantagens que podem ser recebidas.

Etapas de implementação

- Liste os componentes da workload: crie uma lista dos componentes da sua workload. Use essa lista para verificar se cada componente foi analisado. O esforço despendido deve refletir a criticidade da workload conforme definido pelas prioridades da organização. Agrupe recursos de forma funcional para melhorar a eficiência (por exemplo, o armazenamento dos bancos de dados de produção, se houver vários bancos de dados).
- Priorize a lista de componentes: veja a lista de componentes e priorize-a em ordem de esforço. Normalmente, isso é feito por ordem de custos dos componentes, do mais caro para o mais barato, ou da criticidade, conforme definido pelas prioridades da organização.
- Faça a análise: para cada componente na lista, analise as opções e os serviços disponíveis e escolha a opção mais alinhada com suas prioridades organizacionais.

Recursos

Documentos relacionados:

- AWS Pricing Calculator
- AWS Cost Explorer
- Classes de armazenamento do Amazon S3
- Produtos da Nuvem AWS

Vídeos relacionados:

- Série de otimização de custos da AWS: CloudWatch

COST05-BP03 Executar uma análise completa de cada componente

Observe o custo geral de cada componente para a organização. Calcule o custo total de propriedade considerando o custo de operações e gerenciamento, especialmente ao usar serviços gerenciados pelo provedor de nuvem. O esforço de análise deve refletir o benefício potencial (por exemplo, o tempo gasto na análise é proporcional ao custo do componente).

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Considere a economia de tempo que permitirá que sua equipe se concentre na retirada de recursos de endividamento técnico, inovação, agregação de valor e criação de diferenciadores de negócios. Por exemplo, talvez você precise mover sem alterações (lift-and-shift) seu ambiente on-premises para a nuvem (também conhecido como redefinir a hospedagem) e otimizá-lo mais tarde. Vale a pena explorar as possíveis economias obtidas com o uso de serviços gerenciados na AWS que removem ou reduzem os custos de licença. Serviços gerenciados na AWS eliminam a sobrecarga operacional e administrativa da manutenção de um serviço, como aplicação de patches ou atualização do sistema operacional, e permitem que você se concentre na inovação e nos negócios.

Uma vez que os serviços gerenciados operam em escala da nuvem, eles podem oferecer menor custo por transação ou serviço. Você pode realizar possíveis otimizações para alcançar alguns benefícios tangíveis sem alterar a arquitetura principal da aplicação. Por exemplo, talvez você queira reduzir o tempo gasto gerenciando instâncias de banco de dados migrando para uma plataforma de banco de dados como serviço, como o Amazon Relational Database Service (Amazon RDS), ou migrando sua aplicação para uma plataforma totalmente gerenciada, como. AWS Elastic Beanstalk

Geralmente, os serviços gerenciados têm atributos que podem ser definidos para garantir capacidade suficiente. Você deve definir e monitorar esses atributos para que sua capacidade em excesso seja mínima e a performance seja maximizada. Você pode modificar os atributos do AWS Managed Services usando o AWS Management Console ou as APIs e os SDKs da AWS para alinhar as necessidades de recursos à demanda em constante mudança. Por exemplo, é possível aumentar ou diminuir o número de nós em um cluster do Amazon EMR (ou um cluster do Amazon Redshift) para aumentar ou reduzir a escala.

Você também pode unir várias instâncias em um recurso da AWS para ativar usos de maior densidade. Por exemplo, é possível provisionar vários bancos de dados pequenos em uma única instância de banco de dados do Amazon Relational Database Service (Amazon RDS). Conforme o uso aumenta, você pode migrar um dos bancos de dados para uma instância de banco de dados do Amazon RDS dedicada usando um processo de snapshot e restauração.

Ao provisionar workloads em serviços gerenciados, é necessário compreender os requisitos de ajuste da capacidade do serviço. Esses requisitos geralmente são tempo, esforço e qualquer impacto na operação normal da workload. O recurso provisionado deve permitir tempo para que as alterações ocorram. Provisione a sobrecarga necessária para permitir isso. O trabalho contínuo necessário

para modificar os serviços pode ser reduzido a praticamente zero usando APIs e SDKs integrados a ferramentas de sistema e monitoramento como o Amazon CloudWatch.

Por exemplo, o Amazon RDS, o Amazon Redshift e o Amazon ElastiCache fornecem um serviço de banco de dados gerenciado. O Amazon Athena, o Amazon EMR e o Amazon OpenSearch Service oferecem um serviço de análise gerenciado.

O AMS é um serviço que opera a infraestrutura da AWS em nome de clientes e parceiros empresariais. Ele fornece um ambiente seguro e compatível no qual você pode implantar as workloads. O AMS usa modelos operacionais de nuvem empresarial com automação para permitir que você atenda aos requisitos da organização, migre para a nuvem mais rapidamente e reduza os custos de gerenciamento constantes.

Etapas de implementação

- Faça um análise rigorosa: usando a lista de componentes, trabalhe com cada componente da maior prioridade para a menor. Para componentes de prioridade maior e mais caros, execute análises adicionais e avalie todas as opções disponíveis e o impacto a longo prazo. Para componentes de prioridade menor, avalie se alterações no uso alterariam a prioridade do componente e, em seguida, execute uma análise do esforço apropriado.
- Compare recursos gerenciados e não gerenciados: considere o custo operacional dos recursos que você gerencia e compare-os com os recursos gerenciados da AWS. Por exemplo, analise seus bancos de dados em execução em instâncias do Amazon EC2 e compare-os com as opções do Amazon RDS (um serviço gerenciado pela AWS) ou do Amazon EMR em comparação com a execução do Apache Spark no Amazon EC2. Ao migrar de uma workload autogerenciada para uma workload totalmente gerenciada pela AWS, pesquise suas opções com cuidado. Os três fatores mais importantes a serem considerados são o tipo de serviço gerenciado que você deseja usar, o processo que você usará para migrar seus dados e entender o modelo de responsabilidade compartilhada da AWS.

Recursos

Documentos relacionados:

- Calculadora de custo total de propriedade (TCO) da AWS
- Classes de armazenamento do Amazon S3
- Produtos da Nuvem AWS
- Modelo de responsabilidade compartilhada da AWS

Vídeos relacionados:

- Por que migrar para um banco de dados gerenciado?
- O que é o Amazon EMR e como posso usá-lo para processar dados?

Exemplos relacionados:

- Por que migrar para um banco de dados gerenciado?
- Consolide dados de bancos de dados SQL Server idênticos em um único banco de dados do Amazon RDS para SQL Server usando o AWS DMS
- Entregar dados em grande escala ao Amazon Managed Streaming for Apache Kafka (Amazon MSK)
- Migrar uma aplicação Web ASP.NET para o AWS Elastic Beanstalk

COST05-BP04 Selecionar software com licenciamento econômico

Os softwares de código aberto eliminam os custos de licenciamento de software, o que pode contribuir com custos significativos para as workloads. Quando houver necessidade de um software licenciado, evite licenças vinculadas a atributos arbitrários, como CPUs, e procure aquelas que estejam vinculadas à saída ou aos resultados. O custo dessas licenças é mais próximo do benefício que elas oferecem.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

O código aberto originou-se no contexto do desenvolvimento de software para indicar que o software está em conformidade com determinados critérios de distribuição gratuita. O software de código aberto é composto de código-fonte que pode ser inspecionado, modificado e aprimorado por qualquer pessoa. Com base nos requisitos de negócios, nas habilidades dos engenheiros, no uso previsto ou em outras dependências tecnológicas, as organizações podem considerar o uso de software de código aberto na AWS para minimizar os custos de licença. Em outras palavras, o custo das licenças de software pode ser eliminado com o uso de software de código aberto. Isso pode ter impacto significativo nos custos da workload à medida que seu tamanho é dimensionado.

Avalie os benefícios do software licenciado em relação ao custo total para otimizar a workload. Modele todas as alterações no licenciamento e como elas afetariam os custos da workload. Se

um fornecedor alterar o custo da sua licença de banco de dados, investigue como isso afeta a eficiência geral da sua workload. Considere anúncios históricos de preços de seus fornecedores para identificar tendências de alterações de licenciamento em seus produtos. Os custos de licenciamento também podem ser dimensionados independentemente do throughput ou do uso, como licenças que escalam por hardware (licenças vinculadas à CPU). Essas licenças devem ser evitadas porque os custos podem aumentar rapidamente sem resultados correspondentes.

Por exemplo, operar uma instância do Amazon EC2 na região us-east-1 com um sistema operacional Linux permite reduzir os custos em aproximadamente 45% em comparação com a execução de outra instância do Amazon EC2 no Windows.

O AWS Pricing Calculator oferece uma maneira abrangente de comparar os custos de vários recursos com diferentes opções de licença, como instâncias do Amazon RDS e diferentes mecanismos de banco de dados. Além disso, o AWS Cost Explorer fornece uma perspectiva inestimável dos custos das workloads existentes, especialmente daquelas com licenças diferentes. Para gerenciamento de licenças, o AWS License Manager oferece um método simplificado para supervisionar e lidar com licenças de software. Os clientes podem implantar e operacionalizar o software de código aberto preferido na Nuvem AWS.

Etapas de implementação

- Analise as opções de licença: revise os termos de licenciamento do software disponível. Procure versões de código aberto que tenham a funcionalidade necessária e veja se os benefícios do software licenciado superam o custo. Termos favoráveis alinham o custo do software aos benefícios por ele oferecidos.
- Analise o provedor de software: revise todas as alterações históricas de preços ou licenciamento do fornecedor. Procure alterações que não estejam alinhadas aos resultados, como termos punitivos para execução em hardware ou plataformas de fornecedores específicos. Além disso, verifique como eles executam auditorias e as penalidades que poderiam ser impostas.

Recursos

Documentos relacionados:

- Código aberto em AWS
- Calculadora de custo total de propriedade (TCO) da AWS
- Classes de armazenamento do Amazon S3
- Produtos de nuvem

Exemplos relacionados:

- Blogs de código aberto
- Blogs de código aberto da AWS
- Otimização e avaliação do licenciamento

COST05-BP05 Selecionar os componentes desta workload para otimizar o custo alinhado com as prioridades da organização

Considere o custo ao selecionar todos os componentes para sua workload. Isso inclui o uso de serviços gerenciados e em nível de aplicação ou arquitetura sem servidor, contêineres ou orientada a eventos a fim de reduzir o custo geral. Minimize os custos de licença usando um software de código aberto ou que não tenha taxas de licença ou alternativas para reduzir os gastos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Considere o custo de serviços e opções ao selecionar todos os componentes. Isso inclui o uso de serviços gerenciados e em nível de aplicação, como o Amazon Relational Database Service (Amazon RDS), Amazon DynamoDB, Amazon Simple Notification Service (Amazon SNS) e Amazon Simple Email Service (Amazon SES) para reduzir o custo geral da organização.

Use contêineres e tecnologia sem servidor para computação, como o AWS Lambda e o Amazon Simple Storage Service (Amazon S3) para sites estáticos. Se possível, conteinerize sua aplicação e use serviços de contêiner gerenciados da AWS, como Amazon Elastic Container Service (Amazon ECS) ou Amazon Elastic Kubernetes Service (Amazon EKS).

Minimize os custos de licença usando software de código aberto ou software sem taxas de licença: por exemplo, Amazon Linux para workloads de computação ou migração de bancos de dados para o Amazon Aurora.

É possível pode usar serviços sem servidor ou em nível de aplicação, como Lambda, Amazon Simple Queue Service (Amazon SQS), Amazon SNS e Amazon SES. Esses serviços eliminam a necessidade de gerenciar um recurso e fornecem a função de execução de código, serviços de enfileiramento e entrega de mensagens. O outro benefício é que eles escalam a performance e o custo de acordo com o uso, permitindo a alocação e a atribuição eficientes de custos.

O uso da arquitetura orientada a eventos também é possível com serviços sem servidor. Arquiteturas orientadas a eventos são baseadas em push, então, tudo acontece sob demanda à medida que o

evento se apresenta no roteador. Dessa forma, você não paga pela sondagem contínua para conferir um evento. Isso significa um consumo menor de largura de banda de rede, menor utilização de CPU, menor capacidade de frota ociosa e menos handshakes SSL/TLS.

Para obter mais informações sobre o Serverless, consulte o whitepaper Aplicação sem servidor do Well-Architected.

Etapas de implementação

- Selecione cada serviço para otimizar o custo: usando sua análise e lista priorizada, selecione cada opção que fornece a melhor correspondência com suas prioridades organizacionais. Em vez de aumentar a capacidade para atender à demanda, considere outras opções que podem oferecer melhor performance por um custo menor. Por exemplo, se você precisar analisar o tráfego esperado para seus bancos de dados na AWS, considere aumentar o tamanho da instância ou usar serviços do Amazon ElastiCache (Redis ou Memcached) a fim de fornecer mecanismos em cache para seus bancos de dados.
- Avalie a arquitetura orientada a eventos: o uso de uma arquitetura sem servidor também permite criar uma arquitetura orientada a eventos para aplicações distribuídas e baseadas em microsserviço, o que ajuda a criar soluções escaláveis, resilientes, ágeis e econômicas.

Recursos

Documentos relacionados:

- Calculadora de custo total de propriedade (TCO) da AWS
- AWS sem servidor
- O que é arquitetura orientada a eventos
- Classes de armazenamento do Amazon S3
- Produtos de nuvem
- Amazon ElastiCache (Redis OSS)

Exemplos relacionados:

- Conceitos básicos da arquitetura orientada a eventos
- Arquitetura orientada a eventos
- Como a Statsig funciona de forma 100x mais econômica usando o Amazon ElastiCache (Redis OSS)

- Práticas recomendadas para trabalhar com funções do AWS Lambda

COST05-BP06 Realizar análises de custos para diferentes usos ao longo do tempo

As workloads podem mudar ao longo do tempo. Alguns serviços ou recursos são mais econômicos em diferentes níveis de uso. Ao executar a análise em cada componente ao longo do tempo e no uso projetado, a workload continua oferecendo um bom custo-benefício ao longo da vida útil.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

À medida que a AWS lança novos serviços e recursos, os serviços ideais para sua workload podem mudar. O esforço necessário deve refletir possíveis benefícios. A frequência da análise da workload depende dos requisitos da sua organização. Se for uma workload com custo significativo, implementar novos serviços mais cedo maximizará a redução de custos. Portanto, uma revisão mais frequente poderá ser vantajosa. Outro acionador da revisão é a alteração nos padrões de uso. Alterações significativas no uso podem indicar que serviços alternativos seriam opções melhores.

Se precisar mover dados para a Nuvem AWS, você poderá selecionar qualquer variedade de serviços oferecidos pela AWS e ferramentas de parceiros para ajudar a migrar seus conjuntos de dados, sejam eles arquivos, bancos de dados, imagens de máquina, volumes de bloco ou até backups de fita. Por exemplo, para mover um grande volume de dados para a AWS e dela ou processar dados na borda, você pode usar um dos dispositivos com propósito específico da AWS para mover petabytes de dados offline de forma econômica. Outro exemplo é relativo a taxas de transferência de dados mais altas, um serviço de conexão direta pode ser mais barato do que uma VPN, que fornece a conectividade consistente necessária para sua empresa.

Com base na análise de custos para uso diferente no decorrer do tempo, analise sua atividade de ajuste de escala. Analise o resultado para ver se a política de ajuste de escala pode ser ajustada para adicionar instâncias de vários tipos e opções de compra. Analise suas configurações para verificar se é possível reduzir o mínimo para atender às solicitações do usuário, mas com um tamanho de frota menor e adicionar mais recursos para atender à alta demanda esperada.

Realize análises de custos para diferentes usos ao longo do tempo, discutindo com as partes interessadas em sua organização e use o recurso de previsão do AWS Cost Explorer para prever o impacto potencial das mudanças no serviço. Monitore os acionadores de nível de uso utilizando o AWS Budgets, alarmes de faturamento do CloudWatch e o AWS Cost Anomaly Detection para identificar e implementar os serviços mais econômicos com maior rapidez.

Etapas de implementação

- Defina padrões de uso previstos: ao trabalhar com sua organização, como proprietários de produtos e marketing, documente quais serão os padrões de uso previstos e esperados para a workload. Converse com os stakeholders da empresa sobre aumentos de uso e custos históricos e previstos e garanta que os aumentos se alinhem com os requisitos da empresa. Identifique os dias, as semanas ou os meses em que você espera que mais usuários utilizem seus recursos da AWS, o que indica que você deve aumentar a capacidade dos recursos existentes ou adotar serviços adicionais a fim de reduzir o custo e aumentar a performance.
- Execute a análise de custos no uso previsto: usando os padrões de uso definidos, realize a análise em cada um desses pontos. O esforço de análise deve refletir o resultado provável. Por exemplo, se a alteração no uso for grande, uma análise completa deverá ser realizada para verificar quaisquer custos e alterações. Em outras palavras, quando o custo aumenta, o uso também deve aumentar para a empresa.

Recursos

Documentos relacionados:

- Calculadora de custo total de propriedade (TCO) da AWS
- Classes de armazenamento do Amazon S3
- Produtos de nuvem
- Amazon EC2 Auto Scaling
- Migração de dados para nuvem
- AWS Snow Family

Vídeos relacionados:

- AWS OpsHub for Snow Family

COST 6. Como atingir as metas de custo ao selecionar tamanho, número e tipo de recurso?

Escolha o tamanho e o número de recursos apropriados para a tarefa em mãos. Ao selecionar o tipo, tamanho e número mais econômicos, você minimiza o desperdício.

Práticas recomendadas

- COST06-BP01 Realizar modelagem de custos
- COST06-BP02 Selecionar o tipo, o tamanho e o número do recurso com base nos dados
- COST06-BP03 Selecionar o tipo, tamanho e número do recurso automaticamente com base nas métricas
- COST06-BP04 Considerar o uso de recursos compartilhados

COST06-BP01 Realizar modelagem de custos

Identifique os requisitos da organização (como as necessidades de negócios e os compromissos existentes) e realize a modelagem dos custos (custos gerais) da workload e de cada um de seus componentes. Realize atividades de referência para a workload sob diferentes cargas previstas e compare os custos. O esforço de modelagem deve refletir o benefício potencial. Por exemplo, o tempo gasto é proporcional ao custo do componente.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Execute a modelagem de custos para sua workload e cada um de seus componentes para entender o equilíbrio entre recursos e encontrar o tamanho correto para cada recurso na workload, considerando um nível específico de performance. O entendimento das considerações de custo pode embasar seu processo de tomada de decisão e caso de negócios organizacional ao avaliar os resultados da realização de valor para a implantação planejada da workload.

Realize atividades de referência para a workload sob diferentes cargas previstas e compare os custos. O esforço de modelagem deve refletir o benefício potencial. Por exemplo, o tempo gasto é proporcional ao custo do componente ou à economia prevista. Para obter as práticas recomendadas, consulte a seção Revisão do pilar Eficiência de performance do AWS Well-Architected Framework.

Por exemplo, para criar modelagem de custos para uma workload que consiste em recursos computacionais, o AWS Compute Optimizer pode ajudar na modelagem de custos para workloads em execução. Ele fornece recomendações de dimensionamento correto para recursos de computação com base no uso histórico. Implante os CloudWatch Agents nas instâncias do Amazon EC2 para coletar métricas de memória que ajudam você com recomendações mais precisas no AWS Compute Optimizer. Essa é a fonte de dados ideal para recursos de computação, pois é um serviço gratuito e utiliza machine learning para fazer várias recomendações, dependendo dos níveis de risco.

Há vários serviços que você pode usar com logs personalizados como fontes de dados para operações de dimensionamento correto para outros serviços e componentes da workload, como AWS Trusted Advisor, Amazon CloudWatch e Amazon CloudWatch Logs. O AWS Trusted Advisor verifica os recursos e sinaliza aqueles com baixa utilização, o que pode ajudar você a dimensionar corretamente seus recursos e criar modelagem de custos.

Veja a seguir as recomendações para dados e métricas de modelagem de custo:

- O monitoramento deve refletir com precisão a experiência do usuário. Selecione a granularidade correta para o período e escolha com cuidado o máximo ou o 99º percentil, em vez da média.
- Selecione a granularidade correta para o período de análise necessário para cobrir todos os ciclos de workload. Por exemplo, se uma análise de duas semanas for realizada, talvez você esteja deixando passar um ciclo de alta utilização, o que pode levar a subprovisionamento.
- Escolha os serviços da AWS certos para sua workload planejada considerando seus compromissos existentes, modelos de preço selecionados para outras workloads e a capacidade de inovar com maior rapidez e concentrar-se em seu valor comercial principal.

Etapas de implementação

- Faça a modelagem de custos para recursos: implante a workload ou uma prova de conceito em uma conta separada com os tipos e tamanhos de recursos específicos a serem testados. Execute a workload com os dados de teste e registre os resultados de saída, bem como os dados de custo da hora em que o teste foi executado. Depois, reimplante a workload ou altere os tipos e tamanhos de recursos e execute novamente o teste. Inclua taxas de licença para todos os produtos que você pode usar com esses recursos e custos de operações estimados (mão de obra ou engenharia) para implantar e gerenciar esses recursos ao criar a modelagem de custo. Considere a modelagem de custo para um período (por hora, diária, anual ou três anos).

Recursos

Documentos relacionados:

- AWS Auto Scaling
- Identificar oportunidades para dimensionar corretamente
- Recursos do Amazon CloudWatch
- Otimização de custos: dimensionamento correto do Amazon EC2
- AWS Compute Optimizer

- Calculadora de preços da AWS

Exemplos relacionados:

- Execute uma modelagem de custos baseada em dados
- Estime o custo das configurações de recursos da AWS planejados
- Escolher as ferramentas da AWS certas

COST06-BP02 Selecionar o tipo, o tamanho e o número do recurso com base nos dados

Selecione o tamanho ou tipo do recurso com base nos dados sobre a workload e nas características do recurso. Por exemplo, computação, memória, throughput ou gravação intensiva. Essa seleção geralmente é feita usando uma versão anterior (on-premises) da workload, a documentação ou outras fontes de informações sobre a workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

O Amazon EC2 fornece uma ampla seleção de tipos de instância com diferentes níveis de capacidade de CPU, memória, armazenamento e rede para atender a diferentes casos de uso. Esses tipos de instância dispõem de diferentes combinações de capacidade de CPU, memória, armazenamento e rede, oferecendo versatilidade ao selecionar a combinação certa de recursos para os projetos. Eles são disponibilizados em vários tamanhos para que seja possível ajustar os recursos com base nas demandas da workload. Para determinar o tipo de instância necessário, reúna os detalhes dos requisitos do sistema da aplicação ou do software a ser executado na instância. Esses detalhes devem incluir:

- Sistema operacional
- Número de núcleos de CPU
- Núcleos de GPU
- Quantidade de memória do sistema (RAM)
- Tipo e espaço de armazenamento
- Requisito de largura de banda da rede

Identifique a finalidade dos requisitos de computação e a instância necessária e conheça as várias famílias de instâncias do Amazon EC2. A Amazon oferece as seguintes famílias de tipos de instância:

- Finalidade geral
- Otimizadas para computação
- Otimizadas para memória
- Otimizada para armazenamento
- Computação acelerada
- Otimizadas para HPC

Para uma compreensão mais profunda dos propósitos específicos e dos casos de uso que uma família de instâncias específica do Amazon EC2 pode atender, consulte Tipos de instância da AWS.

A coleta dos requisitos do sistema é essencial para selecionar a família e o tipo de instância específicos que melhor atendem às suas necessidades. Os nomes dos tipos de instância são compostos do nome da família e do tamanho da instância. Por exemplo, a instância t2.micro é da família T2 e é de tamanho micro.

Selecione o tamanho ou o tipo de recurso com base na workload e nas características do recurso (por exemplo, computação, memória, throughput ou gravação intensiva). Essa seleção geralmente é feita usando a modelagem de custos, uma versão anterior da workload (como uma versão on-premises), a documentação ou outras fontes de informações sobre a workload (whitepapers ou soluções publicadas). O uso de calculadoras de preços ou de ferramentas de gerenciamento de custos da AWS pode ajudar a tomar decisões fundamentadas sobre tipos, tamanhos e configurações de instância.

Etapas de implementação

- Selecione recursos com base em dados: use seus dados de modelagem de custos para selecionar o nível de uso previsto da workload e escolha o tipo e o tamanho do recurso especificado. Com base nos dados da modelagem de custos, determine o número de CPUs virtuais, a memória total (GiB), o volume de armazenamento de instâncias local (GB), os volumes do Amazon EBS e o nível de performance da rede, levando em consideração a taxa de transferência de dados necessária para a instância. Sempre faça seleções com base em análise detalhada e em dados precisos para otimizar a performance e, ao mesmo tempo, gerenciar os custos de forma eficiente.

Recursos

Documentos relacionados:

- AWS Tipos de instância
- AWS Auto Scaling
- Recursos do Amazon CloudWatch
- Otimização de custos: dimensionamento correto do EC2

Vídeos relacionados:

- Selecionar a instância certa do Amazon EC2 para suas workloads
- Dimensionar seus serviços da maneira certa

Exemplos relacionados:

- Agora é mais fácil descobrir e comparar os tipos de instância do Amazon EC2

COST06-BP03 Selecionar o tipo, tamanho e número do recurso automaticamente com base nas métricas

Use métricas da workload em execução no momento para selecionar o tamanho e o tipo certos para otimizar o custo. Provisione adequadamente o throughput, o dimensionamento e o armazenamento para serviços de computação, armazenamento, dados e rede. Isso pode ser feito com um ciclo de comentários, como ajuste de escala automático ou por código personalizado na workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Crie um loop de comentários dentro da workload que usa métricas ativas da workload em execução para fazer alterações nessa workload. É possível usar um serviço gerenciado, por exemplo, o AWS Auto Scaling, configurado por você para realizar as operações de dimensionamento certas. A AWS também fornece APIs, SDKs e funcionalidades que permitem que os recursos sejam modificados com o mínimo esforço. Você pode programar uma workload para interromper e iniciar uma instância do Amazon EC2 para permitir uma alteração de tamanho ou tipo de instância. Isso fornece os benefícios do dimensionamento correto e, ao mesmo tempo, remove quase todo o custo operacional necessário para fazer a alteração.

Alguns serviços da AWS têm seleção automática de tipo ou tamanho, como o Amazon Simple Storage Service Intelligent-Tiering. O Amazon S3 Intelligent-Tiering move automaticamente seus dados entre dois níveis de acesso: acesso frequente e acesso infrequente, com base em seus padrões de uso.

Etapas de implementação

- Aumente sua observabilidade configurando as métricas da workload: capture as principais métricas da workload. Essas métricas fornecem uma indicação da experiência do cliente, como a saída da workload, e se alinham às diferenças entre tipos e tamanhos de recursos, como uso de CPU e memória. Para recursos de computação, analise os dados de performance para dimensionar corretamente suas instâncias do Amazon EC2. Identifique instâncias ociosas e instâncias subutilizadas. As principais métricas a serem observadas são o uso da CPU e a utilização da memória (por exemplo, 40% de utilização da CPU em 90% do tempo, conforme explicado em Dimensionamento correto com o AWS Compute Optimizer e utilização de memória habilitada). Identifique instâncias com uso máximo de CPU e utilização de memória inferior a 40% em um período de quatro semanas. Essas são as instâncias que devem ser dimensionadas corretamente para reduzir os custos. Para recursos de armazenamento como o Amazon S3, você pode usar a Lente de Armazenamento do Amazon S3, que permite ver 28 métricas em várias categorias no nível do bucket e 14 dias de dados históricos no painel por padrão. Você pode filtrar seu painel da Lente de Armazenamento do Amazon S3 por resumo e otimização de custos ou eventos para analisar métricas específicas.
- Veja as recomendações de dimensionamento correto: use as recomendações de dimensionamento correto no AWS Compute Optimizer e a ferramenta de dimensionamento correto do Amazon EC2 no console de gerenciamento de custos ou revise o dimencionamento correto dos recursos no AWS Trusted Advisor para fazer ajustes em sua workload. É importante usar as ferramentas certas ao dimensionar corretamente diferentes recursos e seguir as diretrizes de dimensionamento correto, seja uma instância do Amazon EC2, classes de armazenamento da AWS ou tipos de instância do Amazon RDS. Para recursos de armazenamento, é possível usar a Lente de Armazenamento do Amazon S3, que oferece visibilidade do uso de armazenamento de objetos e tendências de atividade, bem como faz recomendações acionáveis para otimizar custos e aplicar as práticas recomendadas de proteção de dados. Ao usar as recomendações contextuais que a Lente de Armazenamento do Amazon S3 obtém da análise de métricas em toda a sua organização, você pode tomar medidas imediatas para otimizar seu armazenamento.
- Selecione o tipo e o tamanho do recurso automaticamente com base em métricas: usando as métricas de workload, selecione manual ou automaticamente os recursos da workload. Para recursos de computação, a configuração do AWS Auto Scaling ou a implementação de código

dentro da aplicação pode reduzir o esforço necessário se alterações frequentes forem necessárias e, possivelmente, implementar alterações antes de um processo manual. Você pode iniciar e escalar automaticamente uma frota de instâncias sob demanda e instâncias spot em um único grupo do Auto Scaling. Além de receber descontos pelo uso de instâncias spot, você pode usar instâncias reservadas ou um Savings Plan para receber taxas de desconto da definição de preço normal de instância sob demanda. Todos esses fatores combinados ajudam você a otimizar sua redução de custos para instâncias do Amazon EC2 e determinar a escala e a performance desejadas para a aplicação. Você também pode usar uma estratégia de seleção de tipo de instância baseada em atributos (ABS) em grupos do Auto Scaling (ASG), que permite expressar seus requisitos de instância como um conjunto de atributos, como vCPU, memória e armazenamento. É possível usar automaticamente os tipos de instância de geração mais recente quando eles são lançados e acessar uma variedade mais ampla de capacidade com instâncias spot do Amazon EC2. O Amazon EC2 Fleet e o Amazon EC2 Auto Scaling selecionam e executam instâncias que se ajustam aos atributos especificados, eliminando a necessidade de escolher manualmente os tipos de instância. Para recursos de armazenamento, você pode usar os recursos Amazon S3 Intelligent Tiering e Amazon EFS Infrequent Access, que permitem selecionar automaticamente classes de armazenamento que proporcionam economia automática de custos de armazenamento quando os padrões de acesso aos dados mudam, sem impacto na performance ou sobrecarga operacional.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- Dimensionamento correto da AWS
- AWS Compute Optimizer
- Recursos do Amazon CloudWatch
- Configuração do CloudWatch
- Publicação de métricas personalizadas no CloudWatch
- Conceitos básicos do Amazon EC2 Auto Scaling
- Amazon S3 Storage Lens
- Amazon S3 Intelligent-Tiering
- Amazon EFS Infrequent Access
- Iniciar uma instância do Amazon EC2 usando o SDK

Vídeos relacionados:

- Dimensionar seus serviços da maneira certa

Exemplos relacionados:

- Seleção de tipo de instância baseada em atributos para Auto Scaling para Amazon EC2 Fleet
- Otimizar o Amazon Elastic Container Service para custos usando ajuste de escala agendado
- Ajuste de escala preditivo com o Amazon EC2 Auto Scaling
- Otimizar custos e obter visibilidade do uso com a Lente de Armazenamento do Amazon S3
- Laboratórios do Well-Architected: Recomendações de dimensionamento correto (Nível 100)

COST06-BP04 Considerar o uso de recursos compartilhados

Para serviços já implantados no nível da organização para várias unidades de negócios, considere usar recursos compartilhados para aumentar a utilização e reduzir o custo total de propriedade (TCO). O uso de recursos compartilhados pode ser uma opção econômica para centralizar o gerenciamento e os custos ao usar soluções existentes, compartilhar componentes ou ambos. Gerencie funções comuns, como monitoramento, backups e conectividade, dentro dos limites de uma conta ou em uma conta dedicada. Também é possível reduzir os custos implementando padronização, reduzindo a ocorrência de duplicação e diminuindo a complexidade.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Nas situações em que várias workloads desempenham a mesma função, use soluções existentes e componentes compartilhados para melhorar o gerenciamento e otimizar os custos. Considere usar os recursos existentes (especialmente os compartilhados), como servidores de banco de dados de não produção ou serviços de diretórios, para reduzir os custos de nuvem, seguindo as práticas recomendadas de segurança e as regulamentações organizacionais. Para otimizar a obtenção de valor e a eficiência, é fundamental alocar os custos de volta (usando showback e chargeback) às áreas pertinentes da empresa que geram consumo.

Showback refere-se a relatórios que dividem os custos da nuvem em categorias atribuíveis, como consumidores, unidades de negócios, contas contábeis gerais ou outras entidades responsáveis. O objetivo do showback é mostrar para as equipes, unidades de negócios ou indivíduos os respectivos custos de consumo de recursos da nuvem.

Chargeback significa alocar os gastos com serviços centrais às unidades de custo com base em uma estratégia adequada para um processo específico de gerenciamento financeiro. Para os clientes, o chargeback cobra o custo incorrido de uma conta de serviços compartilhados em diferentes categorias de custos financeiros adequadas para um processo de geração de relatórios de clientes. Ao estabelecer mecanismos de chargeback, você pode relatar os custos incorridos por diferentes unidades de negócios, produtos e equipes.

As workloads podem ser categorizadas como essenciais e não essenciais. Com base nessa classificação, use recursos compartilhados com configurações gerais para workloads menos essenciais. Para otimizar ainda mais os custos, reserve servidores dedicados exclusivamente para workloads essenciais. Compartilhe recursos ou provisione-os em várias contas para gerenciá-los de maneira eficiente. Mesmo em situações com ambientes distintos de desenvolvimento, teste e produção, o compartilhamento seguro é viável e não compromete a estrutura organizacional.

Para melhorar sua compreensão e otimizar os custos e o uso de aplicações em contêineres, utilize dados de alocação de custos divididos que ajudam a alocar os custos para entidades de negócios individuais com base na forma como a aplicação consome recursos compartilhados de computação e memória. Os dados de alocação de custos divididos ajudam você a obter showback e chargeback em nível de tarefa em workloads de contêiner executadas no Amazon Elastic Container Service (Amazon ECS) ou no Amazon Elastic Kubernetes Service (Amazon EKS).

Para arquiteturas distribuídas, crie uma VPC de serviços compartilhados que forneça acesso centralizado aos serviços compartilhados exigidos pelas workloads em cada uma das VPCs. Esses serviços compartilhados podem incluir recursos como serviços de diretório ou endpoints da VPC. Para reduzir as despesas administrativas e os custos, compartilhe os recursos de um local central em vez de criá-los em cada VPC.

Ao usar recursos compartilhados, é possível economizar nos custos operacionais, maximizar a utilização dos recursos e melhorar a consistência. Em um design de várias contas, é posível hospedar alguns serviços da AWS centralmente e acessá-los usando várias aplicações e contas em um hub para reduzir os custos. Você pode usar o AWS Resource Access Manager (AWS RAM) para compartilhar outros recursos comuns, como sub-redes de VPC e anexos do AWS Transit Gateway, AWS Network Firewall ou Amazon SageMaker Pipelines. Em um ambiente de várias contas, use o AWS RAM para criar um recurso uma vez e compartilhá-lo com outras contas.

As organizações devem marcar os custos compartilhados de forma eficaz e verificar se não há uma parte significativa de seus custos sem marcação ou sem alocação. Se você não alocar os custos compartilhados de forma eficaz e ninguém assumir a responsabilidade pelo gerenciamento dos

custos compartilhados, os custos de nuvem compartilhada podem sair do controle. Você deve saber onde custos foram incorridos nos níveis de recurso, workload, equipe ou organização, pois esse conhecimento aprimora sua compreensão do valor fornecido no nível aplicável quando comparado aos resultados comerciais alcançados. Em última análise, as organizações se beneficiam de redução dos custos como resultado do compartilhamento da infraestrutura de nuvem. Incentive a alocação de custos em recursos de nuvem compartilhada para otimizar os gastos com a nuvem.

Etapas de implementação

- Avalie os recursos existentes: analise as workloads existentes que usam serviços semelhantes para sua workload. Dependendo dos componentes da workload, considere usar plataformas existentes se a lógica de negócios ou os requisitos técnicos permitirem.
- Use o compartilhamento de recursos no AWS RAM e restrinja adequadamente: use o AWS RAM para compartilhar recursos com outras contas da AWS em sua organização. Ao compartilhar recursos, você não precisa duplicar recursos em várias contas, o que minimiza a carga operacional da manutenção de recursos. Esse processo também ajuda você a compartilhar com segurança os recursos criados com perfis e usuários em sua conta e em outras Contas da AWS.
- Marque recursos: marque os recursos que são candidatos à geração de relatórios de custos e categorize-os dentro das categorias de custo. Ative essas tags de recursos relacionadas a custos para alocação de custos a fim de fornecer visibilidade do uso de recursos da AWS. Concentre-se em criar um nível adequado de granularidade com relação à visibilidade de custos e uso, além de incentivar comportamentos de consumo na nuvem por meio de relatórios de alocação de custos e rastreamento de KPIs.

Recursos

Práticas recomendadas relacionadas:

- SEC03-BP08 Compartilhar recursos com segurança em sua organização

Documentos relacionados:

- O que é o AWS Resource Access Manager?
- Serviços da AWS que podem ser usados com o AWS Organizations
- Recursos compartilháveis da AWS
- Consultas do AWS Cost and Usage (CUR)

Vídeos relacionados:

- AWS Resource Access Manager: controle de acesso granular com permissões gerenciadas
- Como criar sua estratégia de alocação de custos da AWS
- Categorias de Custos da AWS

Exemplos relacionados:

- Chargeback de serviços compartilhados: um exemplo do AWS Transit Gateway
- Como criar um modelo de chargeback/showback para Savings Plans usando o CUR
- Usar o compartilhamento de VPC para uma arquitetura econômica de microsserviços de várias contas
- Melhorar a visibilidade de custos do Amazon EKS com dados de alocação de custos divididos da AWS
- Melhorar a visibilidade de custos do Amazon ECS e do AWS Batch com dados de alocação de custos divididos da AWS

COST 7. Como usar os modelos de preços para reduzir custos?

Use o modelo de preços mais adequado para seus recursos a fim de minimizar as despesas.

Práticas recomendadas

- COST07-BP01 Executar análise de modelo de preço
- COST07-BP02 Escolher regiões com base no custo
- COST07-BP03 Selecionar contratos de terceiros com termos econômicos
- COST07-BP04 Implementar modelos de preços para todos os componentes da workload
- COST07-BP05 Realizar análise de modelo de preços em nível da conta de gerenciamento

COST07-BP01 Executar análise de modelo de preço

Analise cada componente da workload. Determine se o componente e os recursos serão executados por períodos estendidos (para descontos de compromisso) ou dinâmicos e curtos (para spot ou sob demanda). Execute uma análise da workload usando as recomendações nas ferramentas de gerenciamento de custos e aplique regras de negócios a essas recomendações para alcançar altos retornos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

A AWS tem vários modelos de preços que permitem que você pague pelos seus recursos da maneira mais econômica que atenda às necessidades da sua organização. Trabalhe com suas equipes para determinar o modelo de preço mais apropriado. Com frequência, o modelo de preço consiste em uma combinação de várias opções, tal como determinado por seus requisitos de disponibilidade.

As instâncias sob demanda permitem que você pague pela capacidade computacional ou do banco de dados por hora ou segundo (mínimo de 60 segundos), dependendo das instâncias que você executa, sem compromisso de longo prazo ou pagamentos adiantados.

Os Savings Plans são um modelo de definição de preço flexível que oferece preços baixos no uso do Amazon EC2, Lambda e AWS Fargate em troca de um compromisso com uma quantidade constante de uso (medido em dólares por hora) por um período de vigência de 1 ou 3 anos.

As instâncias spot são um mecanismo de preços do Amazon EC2 que permite que você solicite capacidade computacional extra com desconto por hora (até 90% do preço sob demanda) sem compromisso prévio.

As instâncias reservadas oferecem até 75% de desconto mediante pagamento antecipado pela capacidade. Para obter mais detalhes, consulte Otimização de custos com reservas.

Você pode optar por incluir um Savings Plan para os recursos associados aos ambientes de produção, qualidade e desenvolvimento. Como alternativa, como os recursos de sandbox só são ativados quando necessário, você pode escolher um modelo sob demanda para os recursos desse ambiente. Use instâncias spot da Amazon para reduzir os custos do Amazon EC2 ou use Savings Plans para computação para reduzir os custos do Amazon EC2, Fargate e Lambda. A ferramenta de recomendações do AWS Cost Explorer oferece oportunidades de descontos por compromisso com planos de poupança.

Se você já comprou instâncias reservadas para o Amazon EC2 no passado ou estabeleceu práticas de alocação de custos dentro da sua organização, poderá continuar usando as instâncias reservadas do Amazon EC2 por enquanto. Entretanto, recomendamos elaborar uma estratégia para usar Savings Plans no futuro como um mecanismo de redução de custos mais flexível. Você pode atualizar as recomendações de Savings Plans (SP) no AWS Cost Management para gerar novas recomendações de Savings Plans sempre que quiser. Use instâncias reservadas da para reduzir os custos do Amazon RDS, Amazon Redshift, Amazon ElastiCache e Amazon OpenSearch Service. Os

Saving Plans e as instâncias reservadas estão disponíveis em três opções: pagamento adiantado, pagamento adiantado parcial e sem pagamento adiantado. Use as recomendações de compra de IR e SP fornecidas no AWS Cost Explorer.

Para encontrar oportunidades para workloads spot, use uma visualização por hora do uso geral e procure períodos regulares de uso ou elasticidade variáveis. Você pode usar Instâncias Spot para várias aplicações flexíveis e tolerantes a falhas. Exemplos incluem servidores Web sem estado, endpoints de API, aplicações de big data e análise, workloads conteinerizadas, CI/CD e outras workloads flexíveis.

Analise suas instâncias do Amazon EC2 e do Amazon RDS para ver se elas podem ser desativadas quando não estiverem em uso (após o expediente e nos fins de semana). Essa abordagem permitirá que você reduza os custos em 70% ou mais em comparação a usá-las ininterruptamente. Se você tiver clusters do Amazon Redshift necessários apenas em momentos específicos, poderá pausar o cluster e, posteriormente, retomá-lo. Quando o cluster do Amazon Redshift ou a instância do Amazon EC2 e do Amazon RDS são interrompidos, o faturamento de computação é interrompido e somente se aplica a cobrança de armazenamento.

Observe que as reservas de capacidade sob demanda (ODCR) não são um desconto no preço. A reserva de capacidade é cobrada pela taxa sob demanda equivalente independentemente de você executar instâncias na capacidade reservada ou não. Elas devem ser consideradas quando você precisa fornecer capacidade suficiente para os recursos que pretende executar. As ODCRs não precisam estar atreladas a compromissos de longo prazo, visto que elas podem ser canceladas quando não mais necessárias, mas elas também podem se beneficiar dos descontos que os Savings Plans ou as instâncias reservadas oferecem.

Etapas de implementação

- Analise a elasticidade da workload: usando a granularidade por hora no Explorador de Custos ou um painel personalizado, analise a elasticidade da workload. Procure alterações regulares no número de instâncias em execução. As instâncias de curta duração são candidatas a instâncias spot ou frota spot.
  - Laboratório do Well-Architected: Explorador de Custos
  - Laboratório do Well-Architected: Visualização de custos
- Revise os contratos de preços existentes: revise os contratos ou compromissos atuais quanto a necessidades de longo prazo. Analise o que você tem no momento e quanto esses compromissos estão em uso. Utilize descontos contratuais ou contratos empresariais preexistentes. Os contratos empresariais oferecem aos clientes a opção de personalizar os contratos que melhor atendem às

suas necessidades. Com relação a compromissos de longo prazo, considere descontos de preço reservados, instâncias reservadas ou Savings Plans para o tipo específico de instância, a família de instâncias, a Região da AWS e as zonas de disponibilidade.

- Faça uma análise de desconto por compromisso: usando o Explorador de Custos em sua conta, revise as recomendações de Savings Plans e Reserved Instance. Para verificar se você implementou as recomendações corretas com os descontos e riscos necessários, siga os laboratórios do Well-Architected.

Recursos

Documentos relacionados:

- Como acessar as recomendações de instâncias reservadas
- Opções de compra de instâncias
- AWS Enterprise

Vídeos relacionados:

- Economizar até 90% e executar workloads de produção no spot

Exemplos relacionados:

- Laboratório do Well-Architected: Explorador de Custos
- Laboratório do Well-Architected: Visualização de custos
- Laboratório do Well-Architected: Modelos de preços

COST07-BP02 Escolher regiões com base no custo

Os preços dos recursos podem ser diferentes em cada região. Identifique as diferenças de custo regionais e implante nas regiões com custos mais altos apenas se for necessário atender a requisitos de latência, residência e soberania de dados. A consideração do custo da região ajuda você a pagar o menor preço geral pela workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A infraestrutura da Nuvem AWS é global, hospedada em vários locais em todo o mundo e construída em torno de Regiões da AWS, zonas de disponibilidade, zonas locais, AWS Outposts e zonas do Wavelength. Uma região é um local físico no mundo, e cada região é uma área geográfica separada onde a AWS tem várias zonas de disponibilidade. As zonas de disponibilidade, que são locais isolados em cada região, consistem em um ou mais datacenters discretos, cada um com energia, rede e conectividade redundantes.

Cada Região da AWS opera de acordo com as condições do mercado local, e os preços dos recursos são diferentes em cada região devido às diferenças de custo de imóveis, fibra, eletricidade e impostos, por exemplo. Escolha uma região específica para operar um componente de sua solução completa para que você possa operar no menor preço possível globalmente. Use a Calculadora de Preços da AWS para calcular os custos de sua workload em várias regiões procurando serviços por tipo de local (região, zona do Wavelength e zona local) e região.

Ao projetar suas soluções, uma prática recomendada é buscar colocar os recursos de computação mais perto dos usuários para proporcionar menor latência e forte soberania de dados. Selecione a localização geográfica com base nos requisitos de segurança, performance, privacidade de dados e empresariais. Para aplicações com usuários finais globais, use vários locais.

Use regiões que ofereçam preços mais baixos por serviços da AWS para implantar suas workloads se não houver necessidade de atender a requisitos de privacidade de dados, segurança e empresariais. Por exemplo, se sua região padrão for Ásia-Pacífico (Sydney) (`ap-southwest-2`)e não houver restrições (privacidade de dados, segurança, por exemplo) quanto ao uso de outras regiões, a implantação de instâncias não essenciais (desenvolvimento e teste) do Amazon EC2 na Leste dos EUA (Norte da Virgínia) (`us-east-1`) custará menos.

	Conformidade	Latência	Custos	Serviços/recursos
Região 1	✓	15 ms	$$	✓
Região 2	✓	20 ms	$$$	X
Região 3	✓	80 ms	$	✓
Região 4	✓	15 ms	$$	✓
Região 5	✓	20 ms	$$$	X
Região 6	✓	15 ms	$	✓
Região 7	✓	80 ms	$	✓
Região 8	✓	15 ms	$	X

Tabela de matriz de recursos de regiões

A tabela de matriz anterior mostra que a região 6 é a melhor opção para esse determinado cenário porque a latência é baixa em comparação a outras regiões, o serviço está disponível e é a região mais barata.

Etapas de implementação

- Revise os preços da Região da AWS: analise os custos da workload na região atual. Começando com os custos maiores por serviço e tipo de uso, calcule os custos nas outras regiões que estão disponíveis. Se a economia prevista ultrapassar o custo de mover o componente ou a workload, migre para a nova região.
- Revise os requisitos para implantações multirregionais: analise seus requisitos e obrigações empresariais (privacidade de dados, segurança ou performance) para descobrir se há restrições quanto ao uso de vários regiões. Se não houver obrigações que limitem o uso a uma única região, use várias regiões.
- Analise a transferência de dados necessária: considere os custos de transferência de dados ao selecionar regiões. Mantenha seus dados perto do seu cliente e dos recursos. Selecione Regiões da AWS mais baratas onde os dados fluam e haja transferência de dados mínima. Dependendo dos requisitos de sua empresa para transferência de dados, é possível usar o Amazon CloudFront,

AWS PrivateLink, AWS Direct Connect e AWS Virtual Private Network para reduzir seus custos de rede, melhorar a performance e aprimorar a segurança.

Recursos

Documentos relacionados:

- Como acessar as recomendações de instâncias reservadas
- Definição de preços do Amazon EC2
- Opções de compra de instâncias
- Tabelas de região

Vídeos relacionados:

- Economizar até 90% e executar workloads de produção no spot

Exemplos relacionados:

- Visão geral dos custos de transferência de dados para arquiteturas comuns
- Considerações de custo para implantações globais
- O que considerar ao selecionar uma região para suas workloads
- Laboratórios do Well-Architected: Uso restrito do serviço por região (Nível 200)

COST07-BP03 Selecionar contratos de terceiros com termos econômicos

Contratos e termos econômicos garantem que o custo desses serviços seja dimensionado de acordo com os benefícios oferecidos. Selecione contratos e preços que possam ser escalados ao oferecerem benefícios adicionais à sua organização.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Há vários produtos no mercado que podem ajudar você a gerenciar os custos em ambientes de nuvem. Eles podem ter algumas diferenças em termos de recursos que dependem dos requisitos do cliente, como alguns que enfatizam a governança ou a visibilidade dos custos e outros a otimização

de custos. Um fator muito importante para a eficácia da otimização e da governança de custos é usar a ferramenta certa com os recursos necessários e o modelo de preços correto. Esses produtos têm modelos de preços diferentes. Alguns aplicam determinada porcentagem de cobrança sobre sua fatura mensal, enquanto outros aplicam uma porcentagem sobre as economias obtidas. O ideal é pagar apenas pelo que você precisa.

Ao usar soluções ou serviços de terceiros na nuvem, é importante que as estruturas de preços estejam alinhadas aos resultados desejados. Os preços deve ser dimensionados de acordo com os resultados e o valor fornecido. Por exemplo, em software que leva uma porcentagem das economias que ele fornece, quanto mais você economiza (resultado), mais ele cobra. Os contratos de licença em que você paga mais conforme suas despesas aumentam nem sempre podem ser vantajosos em termos de otimização de custos. No entanto, se o fornecedor oferecer benefícios claros para todos os componentes da sua fatura, talvez essa taxa de ajuste de escala seja aceitável.

Por exemplo, uma solução que fornece recomendações para o Amazon EC2 e que aplica uma porcentagem de cobrança sobre toda a fatura poderá se tornar mais cara se você usar outros serviços que não oferecem nenhum benefício. Outro exemplo é um serviço gerenciado que é cobrado segundo uma porcentagem do custo dos recursos gerenciados. Um tamanho de instância maior pode não exigir necessariamente maior esforço de gerenciamento, mas pode custar mais caro. Verifique se essas disposições de preços de serviços incluem um programa ou recursos de otimização de custos no respectivo serviço para promover a eficiência.

Os clientes podem encontrar produtos mais avançados ou mais fáceis de usar no mercado. Você precisa considerar o custo desses produtos e avaliar possíveis resultados da otimização de custos a longo prazo.

Etapas de implementação

- Analise contratos e termos de terceiros: analise os preços nos contratos de terceiros. Execute modelagem para diferentes níveis de uso e leve em consideração novos custos, como o uso de novos serviços ou aumentos nos serviços atuais, devido ao crescimento da workload. Decida se os custos adicionais fornecem os benefícios necessários para a sua empresa.

Recursos

Documentos relacionados:

- Como acessar as recomendações de instâncias reservadas
- Opções de compra de instâncias

Vídeos relacionados:

- Economizar até 90% e executar workloads de produção no spot

COST07-BP04 Implementar modelos de preços para todos os componentes da workload

Os recursos em execução permanente devem utilizar capacidade reservada, como Savings Plans ou instâncias reservadas. A capacidade de curto prazo está configurada para usar instâncias spot ou frota spot. As instâncias sob demanda são usadas somente para workloads de curto prazo que não podem ser interrompidas e não são executadas por tempo suficiente para a capacidade reservada, entre 25% e 75% do período, dependendo do tipo do recurso.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Para melhorar o custo-benefício, a AWS fornece várias recomendações de compromisso com base no uso anterior. Essas recomendações podem ser usadas para compreender o que você pode economizar e como o compromisso será usado. É possível usar esses serviços como instâncias sob demanda ou spot ou assumir um compromisso por determinado período e reduzir os custos sob demanda com instâncias reservadas (RIs) e Savings Plans (SPs). É necessário compreender, além de cada componente da workload e dos vários serviços da AWS, os descontos de compromisso, as opções de compra e as instâncias spot desses serviços para otimizar a workload.

Considere os requisitos dos componentes da workload e informe-se sobre os diferentes modelos de preços desses serviços. Defina o requisito de disponibilidade desses componentes. Determine se há vários recursos independentes que executam a função na workload e quais são os requisitos da workload ao longo do tempo. Compare o custo dos recursos usando o modelo de preços sob demanda padrão e outros modelos aplicáveis. Leve em consideração possíveis alterações nos recursos ou componentes da workload.

Por exemplo, vamos analisar essa arquitetura de aplicações web na AWS. Esse exemplo de workload consiste em vários serviços da AWS, como Amazon Route 53, AWS WAF, Amazon CloudFront, instâncias do Amazon EC2, instâncias do Amazon RDS, balanceadores de carga, armazenamento do Amazon S3 e Amazon Elastic File System (Amazon EFS). É necessário analisar cada um desses serviços e identificar as possíveis oportunidades de redução de custos com diferentes modelos de preços. Alguns deles podem ser elegíveis para IRs ou SPs, e outros podem estar disponíveis apenas sob demanda. Conforme mostrado na imagem a seguir, alguns dos serviços da AWS podem ser comprometidos via IRs ou SPs.

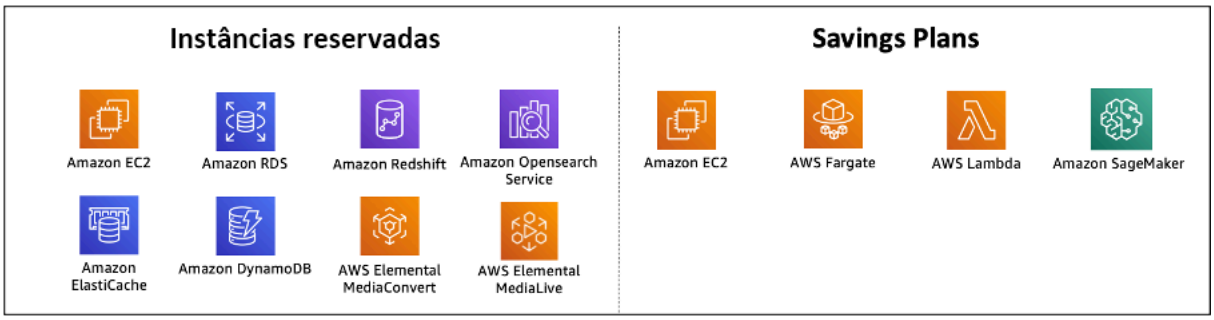


Tabela de serviços AWS da comprometidos via instâncias reservadas e Savings Plans

Etapas de implementação

- Implemente modelos de preços: usando seus resultados de análise, compre Savings Plans, instâncias reservadas ou implemente instâncias spot. Se esta for a sua primeira compra de compromisso, escolha as cinco ou dez principais recomendações da lista, monitore e analise os resultados de um ou dos dois próximos meses. O AWS Cost Management Console fornece orientações durante o processo. Analise as recomendações de IR ou de SP no console, personalize as recomendações (tipo, pagamento e prazo) e analise o compromisso por hora (por exemplo, USD 20 por hora) e adicione ao carrinho. Os descontos se aplicam automaticamente ao uso qualificado. Compre uma pequena quantidade de descontos de compromisso em ciclos regulares (por exemplo, a cada duas semanas ou mensalmente). Implemente instâncias spot para workloads que podem ser interrompidas ou que são sem estado. Finalmente, selecione instâncias sob demanda do Amazon EC2 e aloque recursos para os demais requisitos.
- Ciclo de revisão da workload: implemente um ciclo de análise da workload que analise especificamente a cobertura do modelo de preços. Assim que a workload tiver a cobertura necessária, compre descontos de compromisso adicionais parcialmente (a cada dois meses) ou conforme o uso da sua organização mudar.

Recursos

Documentos relacionados:

- Entender as recomendações de Savings Plans
- Como acessar as recomendações de instâncias reservadas
- Como comprar instâncias reservadas
- Opções de compra de instâncias
- Instâncias spot

- Modelos de reserva para outros serviços da AWS
- Serviços com suporte de Savings Plans

Vídeos relacionados:

- Economizar até 90% e executar workloads de produção no spot

Exemplos relacionados:

- O que você deveria considerar antes de comprar Savings Plans?
- Como posso usar o Explorador de Custos para analisar meus gastos e uso?

COST07-BP05 Realizar análise de modelo de preços em nível da conta de gerenciamento

Confira as ferramentas de gerenciamento de faturamento e de custos e veja os descontos recomendados com compromissos e reservas para realizar uma análise regular no nível da conta de gerenciamento.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Realizar modelagens de custos regularmente ajuda você a implementar oportunidades de otimização em várias workloads. Por exemplo, se várias workloads usarem instâncias sob demanda em um nível agregado, o risco de alteração será menor e a implementação de um desconto baseado em compromisso poderá atingir um custo geral mais baixo. Recomenda-se realizar análises em ciclos regulares de duas semanas a um mês. Isso permite que você faça pequenas compras de ajuste para que a cobertura dos seus modelos de preços continue a evoluir com suas workloads dinâmicas e os respectivos componentes.

Use as ferramentas de recomendação da AWS Cost Explorer para encontrar oportunidades de descontos de compromisso em sua conta de gerenciamento. As recomendações em nível de conta de gerenciamento são calculadas considerando-se o uso em todas as contas da organização da AWS que têm instâncias reservadas (RI) ou Savings Plans (SP). Elas também são calculadas quando o compartilhamento de descontos é ativado para recomendar um compromisso que maximize a economia em todas as contas.

Embora a compra em nível da conta de gerenciamento seja otimizada para obter o máximo de economia em muitos casos, poderá haver situações em que você considera comprar SPs em nível

da conta vinculada, como quando você deseja que os descontos se apliquem primeiro ao uso nessa conta vinculada específica. Conta-membro: as recomendações são calculadas no nível da conta individual ou conta-membro para maximizar a economia de cada conta. Se sua conta tiver compromissos de RI e SP, eles serão aplicados na seguinte ordem:

1. RI de zona
2. RI padrão
3. RI conversível
4. Savings Plans para instâncias
5. Savings Plans para computação

Se você comprar um SP em nível da conta de gerenciamento, a economia será aplicada com base na porcentagem de desconto mais alta para a mais baixa. Os SPs em nível da conta de gerenciamento examinam todas as contas vinculadas e aplicarão as economias sempre que o desconto for maior. Se desejar restringir onde as economias são aplicadas, você poderá comprar um Savings Plans em nível da conta vinculada e, sempre que a conta estiver executando serviços computacionais qualificados, o desconto será aplicado primeiro. Quando a conta não estiver executando serviços computacionais qualificados, o desconto será compartilhado entre as outras contas vinculadas na mesma conta de gerenciamento. O compartilhamento de descontos é ativado por padrão, mas pode ser desativado se necessário.

Em uma Família de Faturamento Consolidado, os Savings Plans são aplicados primeiro à conta do proprietário e depois a outras contas. Isso ocorre somente se você tiver o compartilhamento ativado. Seus Savings Plans são aplicados primeiro à maior porcentagem de economia. Se houver vários usos com porcentagens iguais, os Savings Plans são aplicados ao primeiro uso com a menor taxa. Os Savings Plans continuam em vigor até que não haja mais usos restantes ou que seu compromisso seja esgotado. O uso restante é cobrado na tarifa sob demanda. Você pode atualizar as recomendações de Savings Plans no Gerenciamento de custos da AWS para gerar novas recomendações de Savings Plans sempre que quiser.

Depois de analisar a flexibilidade das instâncias, você pode confirmar de acordo com as recomendações. Crie uma modelagem de custos analisando os custos de curto prazo da workload com possíveis opções de recursos diferentes, analisando os modelos de preço da AWS e alinhando-os aos requisitos empresariais para encontrar o custo total de propriedade e oportunidades de otimização de custos.

Etapas de implementação

Faça uma análise de desconto por compromisso: usando o Explorador de Custos em sua conta, revise as recomendações de Savings Plans e instâncias reservadas. Entenda as recomendações de Savings Plans e estime seus gastos mensais e as economias mensais. Examine as recomendações no nível da conta de gerenciamento, que são calculadas considerando o uso em todas as contas em sua organização da AWS que têm o compartilhamento de descontos de RI ou Savings Plans habilitado, com o intuito de obter o máximo de economia nas contas. É possível verificar se as recomendações corretas foram implementadas com os descontos e riscos necessários nos laboratórios do Well-Architected.

Recursos

Documentos relacionados:

- Como a definição de preços da AWS funciona?
- Opções de compra de instâncias
- Visão geral dos Savings Plans
- Recomendações de Savings Plans
- Como acessar as recomendações de instâncias reservadas
- Conceitos básicos sobre a recomendação de Savings Plans
- Como os Savings Plans se aplicam ao seu uso do AWS
- Saving Plans com faturamento consolidado
- Ativar descontos compartilhados de instâncias reservadas e Savings Plans

Vídeos relacionados:

- Economizar até 90% e executar workloads de produção no spot

Exemplos relacionados:

- Laboratório do AWS Well-Architected: Modelos de preços (Nível 200)
- Laboratórios do AWS Well-Architected: Análise de modelo de preços (Nível 200)
- O que devo considerar antes de comprar um Savings Plans?
- Como posso usar Savings Plans sucessivos para reduzir o risco do compromisso?

- Quando usar instâncias spot

COST 8. Como planejar as cobranças de transferência de dados?

Planeje e monitore as cobranças de transferência de dados para tomar decisões de arquitetura que minimizam custos. Uma mudança arquitetônica pequena, porém eficaz, pode reduzir drasticamente os custos operacionais ao longo do tempo.

Práticas recomendadas

- COST08-BP01 Executar a modelagem de transferência de dados
- COST08-BP02 Selecionar os componentes para otimizar o custo da transferência de dados
- COST08-BP03 Implementar serviços para reduzir custos de transferência de dados

COST08-BP01 Executar a modelagem de transferência de dados

Reúna os requisitos da organização e execute a modelagem de transferência de dados da workload e de cada um dos componentes. Isso identifica o menor ponto de custo para os requisitos atuais de transferência de dados.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Ao projetar uma solução na nuvem, as taxas de transferência de dados geralmente são negligenciadas devido ao hábito de projetar a arquitetura usando datacenters on-premises ou à falta de conhecimento. As taxas de transferência de dados na AWS são determinadas pela origem, pelo destino e pelo volume do tráfego. A consideração dessas taxas durante a fase de projeto pode resultar em redução de custos. É muito importante compreender onde ocorre a transferência de dados na workload, o custo da transferência e os respectivos benefícios associados para estimar com precisão o custo total de propriedade (TCO). Isso permite que você tome uma decisão embasada para modificar ou aceitar a decisão de arquitetura. Por exemplo, você pode ter uma configuração de várias zonas de disponibilidade na qual replica dados entre as zonas de disponibilidade.

Você modela os componentes dos serviços que transferem os dados na workload e conclui que esse é um custo aceitável (de modo semelhante ao pagamento por computação e armazenamento nas duas zonas de disponibilidade) para alcançar a confiabilidade e a resiliência necessárias. Modele os

custos em diferentes níveis de uso. O uso da workload pode mudar ao longo do tempo, e diferentes serviços podem ser mais econômicos em diferentes níveis.

Ao modelar a transferência de dados, considere a quantidade de dados ingeridos e a origem desses dados. Além disso, considere a quantidade de dados processados e a quantidade de armazenamento ou capacidade computacional necessária. Durante a modelagem, siga as práticas recomendadas de rede para sua arquitetura de workload a fim de otimizar os possíveis custos de transferência de dados.

O AWS Pricing Calculator pode ajudar você a ver os custos estimados de serviços específicos da AWS e da transferência de dados esperada. Se você já tiver uma workload em execução (para fins de teste ou em um ambiente de pré-produção), use o AWS Cost Explorer ou o AWS Cost and Usage Report (CUR) para entender e modelar seus custos de transferência de dados. Configure uma prova de conceito (PoC) ou teste sua workload e execute um teste com uma carga simulada realista. É possível modelar seus custos em diferentes demandas de workload.

Etapas de implementação

- Identifique os requisitos: qual é a meta principal e os requisitos de negócios para a transferência planejada de dados entre a origem e o destino? Quais são os resultados comerciais esperados no final? Reúna os requisitos de negócios e defina o resultado esperado.
- Identifique a origem e o destino: quais são a fonte e o destino dos dados para a transferência de dados, como dentro de Regiões da AWS, para serviços da AWS ou para fora da Internet?
  - Transferência de dados dentro de uma Região da AWS
  - Transferência de dados entre Regiões da AWS
  - Transferência de dados para a internet
- Identifique as classificações de dados: qual é a classificação de dados para essa transferência de dados? De que tipo são esses dados? Qual é o tamanho dos dados? Com que frequência os dados devem ser transferidos? Os dados são sigilosos?
- Identifique serviços ou ferramentas da AWS a serem usados: que serviços da AWS são usados para essa transferência de dados? É possível usar um serviço já provisionado para outra workload?
- Calcule os custos de transferência de dados: use o modelo de transferência de dados de Preços da AWS que você criou anteriormente para calcular os custos de transferência de dados para a workload. Calcule os custos da transferência de dados em diferentes níveis de uso, tanto para aumentos quanto para reduções no uso da workload. Quando houver várias opções para a arquitetura da workload, calcule o custo de cada uma delas a título de comparação.

- Vincule os custos aos resultados: para cada custo de transferência de dados incorrido, especifique o resultado que ele atinge para a workload. Se a transferência for entre componentes, poderá ser para desacoplamento; se for entre zonas de disponibilidade, poderá ser para redundância.
- Crie uma modelagem de transferência de dados: depois de reunir todas as informações, crie uma modelagem de transferência de dados de base conceitual para vários casos de uso e workloads diferentes.

Recursos

Documentos relacionados:

- Soluções de armazenamento em cache da AWS
- Definição de preço do AWS
- Definição de preços do Amazon EC2
- Preços da Amazon VPC
- Conceitos básicos das taxas de transferência de dados

Vídeos relacionados:

- Monitorar e otimizar os custos da transferência de dados
- Aceleração de transferências do S3

Exemplos relacionados:

- Visão geral dos custos de transferência de dados para arquiteturas comuns
- Recomendações da AWS para redes

COST08-BP02 Selecionar os componentes para otimizar o custo da transferência de dados

Todos os componentes são selecionados, e a arquitetura é projetada para reduzir os custos de transferência de dados. Isso inclui o uso de componentes como otimização de rede de longa distância (WAN) e configurações de várias zonas de disponibilidade (AZ).

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A arquitetura da transferência de dados minimiza os custos da transferência de dados. Isso pode envolver o uso de redes de entrega de conteúdo para localizar os dados mais perto dos usuários ou o uso de links de rede dedicados do ambiente on-premises para a AWS. Você também pode usar a otimização de WAN e a otimização de aplicações para reduzir a quantidade de dados transferidos entre componentes.

Ao transferir dados para a Nuvem AWS ou dentro dela, é essencial conhecer o destino com base em diversos casos de uso, a natureza dos dados e os recursos de rede disponíveis para selecionar os serviços certos da AWS e otimizar a transferência de dados. A AWS oferece uma variedade de serviços de transferência de dados personalizados para diversos requisitos de migração de dados. Selecione as opções corretas de armazenamento de dados e transferência de dados com base nas necessidades comerciais da sua organização.

Ao planejar ou analisar a arquitetura da workload, considere o seguinte:

- Use endpoints da VPC dentro da AWS: um endpoint da VPC permite conexões privadas entre a VPC e os serviços da AWS compatíveis. Isso permite evitar o uso da internet pública, o que pode resultar em custos de transferência de dados.
- Use um gateway NAT: use um gateway NAT para que as instâncias em uma sub-rede privada possam se conectar a serviços fora da VPC. Verifique se os recursos por trás do gateway NAT que enviam mais tráfego estão na mesma zona de disponibilidade do gateway NAT. Caso contrário, crie novos gateways NAT na mesma zona de disponibilidade do recurso para reduzir as taxas de transferência de dados entre AZs.
- O uso de AWS Direct Connect AWS Direct Connect ignora a Internet pública e estabelece uma conexão direta e privada entre sua rede on-premises e a AWS. Isso pode ser mais econômico e consistente do que transferir grandes volumes de dados pela internet.
- Evite transferir dados entre fronteiras regionais: as transferências de dados entre Regiões da AWS (de uma região para outra) normalmente incorrem em cobranças. A decisão de seguir um caminho multirregional deve ser muito cuidadosa. Para obter mais detalhes, consulte Cenários multirregionais.
- Monitore a transferência de dados: use o Amazon CloudWatch e os Logs de fluxo da VPC para capturar detalhes sobre sua transferência de dados e uso da rede. Analise as informações de tráfego de rede capturadas nas VPCs, como o endereço IP ou o intervalo de entrada e saída das interfaces de rede.

- Analise o uso da rede: use ferramentas de medição e geração de relatóriosAWS Cost Explorer, como CUDOS Dashboards ou CloudWatch, para entender o custo de transferência de dados da sua workload.

Etapas de implementação

- Selecione os componentes para a transferência de dados: usando a modelagem de transferência de dados explicada em COST08-BP01 Executar a modelagem de transferência de dados, concentre-se em onde estão os maiores custos de transferência de dados ou onde eles estariam se o uso da workload mudasse. Procure arquiteturas alternativas ou componentes adicionais que removam ou reduzam a necessidade da transferência de dados (ou que diminuam o custo).

Recursos

Práticas recomendadas relacionadas:

- COST08-BP01 Executar a modelagem de transferência de dados
- COST08-BP03 Implementar serviços para reduzir custos de transferência de dados

Documentos relacionados:

- Migração de dados para nuvem
- Soluções de armazenamento em cache da AWS
- Entregar conteúdo com mais rapidez com o Amazon CloudFront

Exemplos relacionados:

- Visão geral dos custos de transferência de dados para arquiteturas comuns
- Dicas de otimização de rede da AWS
- Otimizar a performance e reduzir os custos de análise de rede com os Logs de fluxo da VPC no formato Apache Parquet

COST08-BP03 Implementar serviços para reduzir custos de transferência de dados

Implemente serviços para reduzir os custos da transferência de dados. Por exemplo, é possível usar locais da borda ou redes de entrega de conteúdo (CDN) para fornecer conteúdo aos usuários finais,

criar camadas de cache na frente de servidores de aplicações ou bancos de dados e usar conexões de rede dedicadas em vez de VPN para conectividade com a nuvem.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Existem vários serviços da AWS que podem ajudar a otimizar o uso da transferência de dados pela rede. Dependendo da arquitetura da nuvem e dos componentes e tipo da workload, esses serviços podem ajudar na compactação, no armazenamento em cache e no compartilhamento e na distribuição do tráfego na nuvem.

- O Amazon CloudFront é uma rede de entrega de conteúdo global que entrega dados com baixa latência e altas velocidades de transferência. Ele armazena dados em cache em pontos de presença no mundo inteiro, o que reduz a carga sobre seus recursos. Ao usar o CloudFront, você pode reduzir o trabalho administrativo para entregar conteúdo a grandes números de usuários globalmente com latência mínima. O pacote promocional de segurança pode ajudar você a economizar até 30% do uso do CloudFront se você planeja aumentar o uso ao longo do tempo.
- O AWS Direct Connect permite a você estabelecer uma conexão de rede dedicada com a AWS. Isso pode reduzir os custos de rede, aumentar a largura de banda e fornecer uma experiência de rede mais consistente do que conexões baseadas na Internet.
- O AWS VPN permite estabelecer uma conexão segura e privada entre a rede privada e a rede global da AWS. Ele é ideal para pequenos escritórios ou parceiros de negócios porque oferece conectividade simplificada, além de ser um serviço totalmente gerenciado e elástico.
- Os endpoints da VPC permitem a conectividade entre os serviços da AWS sobre redes privadas e podem ser usados para reduzir os custos de transferência de dados pública e gateways NAT. Os endpoints da VPC do gateway não têm cobranças por hora e oferecem suporte ao Amazon S3 e ao Amazon DynamoDB. Os endpoints da VPC de interface são fornecidos pelo AWS PrivateLink e têm uma taxa horária e custo de uso por GB.
- Os gateways NAT fornecem ajuste de escala e gerenciamento integrados, reduzindo os custos, em comparação com uma instância NAT independente. Coloque os gateways NAT nas mesmas zonas de disponibilidade das instâncias de alto tráfego e pense no uso de endpoints da VPC para as instâncias que precisam acessar o Amazon DynamoDB ou o Amazon S3 a fim de reduzir os custos de transferência e processamento de dados.
- Use dispositivos AWS Snow Family que tenham recursos de computação para coletar e processar dados na borda. Os dispositivos AWS Snow Family (Snowcone, Snowball e Snowmobile) permitem que você mova petabytes de dados para a Nuvem AWS de forma econômica e offline.

Etapas de implementação

- Implemente os serviços: selecione os serviços de rede aplicáveis da AWS com base no serviço e no tipo de workload usando a modelagem de transferência de dados e revisando os logs de fluxo da VPC. Veja onde estão os maiores custos e os maiores fluxos de volume. Revise os serviços da AWS e avalie se algum deles reduz ou remove a transferência, especificamente a entrega de conteúdo e as redes. Procure também serviços de armazenamento em cache em que há acesso repetido aos dados ou grandes quantidades de dados.

Recursos

Documentos relacionados:

- AWS Direct Connect
- Explorar os produtos da AWS
- Soluções de armazenamento em cache da AWS
- Amazon CloudFront
- AWS Snow Family
- Pacote Promocional de Segurança do Amazon CloudFront

Vídeos relacionados:

- Monitorar e otimizar os custos da transferência de dados
- Série de otimização de custos da AWS: CloudFront
- Como posso reduzir os custos da transferência de dados para meu gateway NAT?

Exemplos relacionados:

- Chargeback de serviços compartilhados: um exemplo do AWS Transit Gateway
- Como entender os detalhes da transferência de dados da AWS com base no relatório de custos e uso utilizando consultas do Athena e o QuickSight
- Visão geral dos custos de transferência de dados para arquiteturas comuns
- Usar o AWS Cost Explorer para analisar custos de transferência de dados
- Otimizar os custos das arquiteturas da AWS utilizando os recursos do Amazon CloudFront
- Como posso reduzir os custos da transferência de dados para meu gateway NAT?

Gerenciar recursos de demanda e fornecimento

Pergunta

- COST 9. Como gerenciar a demanda e fornecer recursos?

COST 9. Como gerenciar a demanda e fornecer recursos?

Para uma workload com gasto e performance equilibrados, verifique se tudo o que você paga está sendo usado e evite instâncias significativamente subutilizadas. Uma métrica de utilização distorcida em ambas as direções tem um impacto adverso sobre a organização, tanto nos custos operacionais (redução na performance em decorrência de utilização excessiva) quanto em despesas desnecessárias na AWS (devido ao excesso de provisionamento).

Práticas recomendadas

- COST09-BP01 Realizar uma análise sobre a demanda da workload
- COST09-BP02 Implementar um buffer ou controle de utilização para gerenciar a demanda
- COST09-BP03 Fornecer recursos dinamicamente

COST09-BP01 Realizar uma análise sobre a demanda da workload

Analise a demanda da workload ao longo do tempo. Garanta que a análise cubra tendências sazonais e represente com precisão as condições operacionais durante toda a vida útil da workload. O trabalho de análise deve refletir o benefício potencial (por exemplo, se o tempo gasto é proporcional ao custo da workload).

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Analisar a demanda de workload para computação em nuvem envolve entender os padrões e as características das tarefas de computação que são iniciadas no ambiente de nuvem. Essa análise ajuda os usuários a otimizar a alocação de recursos, gerenciar custos e verificar se a performance atende aos níveis exigidos.

Conhecer os requisitos da workload. Os requisitos da organização devem indicar os tempos de resposta da workload para solicitações. O tempo de resposta pode ser usado para determinar se a demanda é gerenciada ou se a oferta de recursos deve ser alterada para atender à demanda.

A análise deve incluir a previsibilidade e a repetibilidade da demanda, a taxa de alteração na demanda e a quantidade de alteração na demanda. Realize a análise durante um período longo o suficiente para incorporar qualquer variação sazonal, como processamento de fim de mês ou picos de fim de ano.

O trabalho de análise deve refletir os possíveis benefícios da implementação do ajuste de escala. Observe o custo total esperado do componente e os aumentos ou diminuições no uso e no custo durante a vida útil da workload.

Veja abaixo alguns aspectos importantes a serem considerados ao realizar a análise da demanda de workload para computação em nuvem:

1. Métricas de utilização de recursos e performance: analise como os recursos da AWS estão sendo usados ao longo do tempo. Determine padrões de uso de pico e fora do pico para otimizar as estratégias de alocação e ajuste de escala de recursos. Monitore métricas de performance, como tempos de resposta, latência, throughput e taxas de erro. Essas métricas ajudam a avaliar a integridade geral e a eficiência da infraestrutura de nuvem.
2. Comportamento de ajuste de escala de usuários e aplicações: entenda o comportamento do usuário e como ele afeta a demanda da workload. Examinar os padrões de tráfego de usuários ajuda a aprimorar a entrega de conteúdo e a capacidade de resposta das aplicações. Analise como as workloads escalam com o aumento da demanda. Determine se os parâmetros de ajuste de escala automático estão configurados de forma correta e eficaz para lidar com flutuações de carga.
3. Tipos de workload: identifique os diferentes tipos de workload em execução na nuvem, como processamento em lote, processamento de dados em tempo real, aplicação web, bancos de dados ou machine learning. Cada tipo de workload pode ter requisitos de recursos e perfis de performance diferentes.
4. Acordos de serviço (SLA): compare a performance real com os SLAs para garantir a conformidade e identificar áreas que precisam ser aprimoradas.

Você pode usar o Amazon CloudWatch para coletar e rastrear métricas, coletar e monitorar arquivos de log, definir alarmes e reagir automaticamente a alterações nos seus recursos da AWS. O Amazon CloudWatch pode ser usado para fornecer visibilidade sobre a utilização de recursos, a performance de aplicações e o status operacional em todo o sistema.

Com o AWS Trusted Advisor, é possível provisionar os recursos seguindo as práticas recomendadas para melhorar a performance e a confiabilidade do sistema, aumentar a segurança e procurar

oportunidades de economia. Também é possível desativar o uso e as instâncias de não produção e usar o Amazon CloudWatch e o Auto Scaling para equiparar aumentos ou reduções na demanda.

Finalmente, você pode usar o AWS Cost Explorer ou o Amazon QuickSight com o arquivo do AWS Cost and Usage Report (CUR) ou os logs da aplicação para realizar análises avançadas da demanda de workload.

No geral, uma análise abrangente da demanda da workload permite que as organizações tomem decisões embasadas sobre provisionamento, ajuste de escala e otimização de recursos, o que melhora a performance, o custo-benefício e a satisfação do usuário.

Etapas de implementação

- Analise dados da workload existente: analise dados da workload existente, das versões anteriores da workload ou dos padrões de uso previstos. Use o Amazon CloudWatch, arquivos de log e dados de monitoramento para obter informações sobre como a workload foi usada. Analise um ciclo completo da workload e colete dados para alterações sazonais, como eventos de fim de mês ou de ano. O esforço refletido na análise deve refletir as características da workload. Deve-se concentrar o maior esforço em workloads de alto valor com as maiores alterações na demanda. Por outro lado, deve-se concentrar o menor esforço em workloads de baixo valor que tenham alterações mínimas na demanda.
- Preveja a influência externa: encontre membros da equipe de toda a organização que possam influenciar ou alterar a demanda na workload. Equipes comuns seriam de vendas, marketing ou desenvolvimento de negócios. Trabalhe com elas para saber os ciclos com os quais operam e se há eventos que possam alterar a demanda da workload. Preveja a demanda da workload com esses dados.

Recursos

Documentos relacionados:

- Amazon CloudWatch
- AWS Trusted Advisor
- AWS X-Ray
- AWS Auto Scaling
- Agendador de instâncias da AWS
- Conceitos básicos do Amazon SQS

- AWS Cost Explorer
- Amazon QuickSight

Vídeos relacionados:

Exemplos relacionados:

- Monitorar, rastrear e analisar para fins de otimização de custos
- Pesquisar e analisar logs no CloudWatch

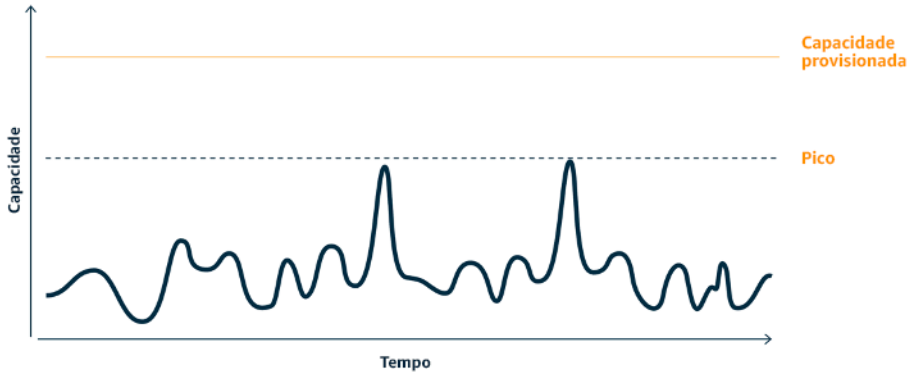
COST09-BP02 Implementar um buffer ou controle de utilização para gerenciar a demanda

O armazenamento em buffer e o controle de utilização modificam a demanda na workload, suavizando todos os picos. Implemente o controle de utilização quando seus clientes realizarem novas tentativas. Implemente o armazenamento em buffer para armazenar a solicitação e adiar o processamento até um momento posterior. Verifique se os controles de utilização e buffers estão projetados para que os clientes recebam uma resposta no tempo necessário.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A implementação de um buffer ou controle de utilização é crucial na computação em nuvem para gerenciar a demanda e reduzir a capacidade provisionada necessária para a workload. Para obter a performance ideal, é essencial avaliar a demanda total, incluindo os picos, o ritmo das mudanças nas solicitações e o tempo de resposta necessário. Quando os clientes têm a capacidade de reenviar solicitações, é prático aplicar o controle de utilização. Entretanto, para clientes que não têm funcionalidades de repetição, a abordagem ideal é implementar uma solução de buffer. Esses buffers agilizam o influxo de solicitações e otimizam a interação de aplicações com velocidades operacionais variadas.



Curva de demanda com dois picos distintos que exigem alta capacidade provisionada.

Considere uma workload com a curva de demanda mostrada na figura anterior. Essa workload tem dois picos e, para lidar com eles, é provisionada a capacidade de recurso mostrada pela linha laranja. Os recursos e a energia usados para essa workload não são indicados pela área abaixo da curva da demanda, mas pela área abaixo da linha da capacidade provisionada, visto que é preciso ter capacidade provisionada para lidar com esses dois picos. Nivelar a curva da demanda pode ajudar você a reduzir a capacidade provisionada para uma workload e a diminuir o respectivo impacto ambiental. Para suavizar o pico, considere implementar uma solução de controle de utilização ou de buffer.

Para entender melhor, vamos examinar o controle de utilização e o buffer.

Controle de utilização: se a origem da demanda tiver capacidade de repetição, você poderá implementar o controle de utilização. O controle de utilização informa à origem que, se não for possível atender à solicitação no momento, ela deverá tentar novamente mais tarde. A origem espera por um período e repete a solicitação. A implementação do controle de utilização tem a vantagem de limitar a quantidade máxima de recursos e custos da workload. Na AWS, o Amazon API Gateway pode ser usado para implementar o controle de utilização.

Buffer: uma abordagem baseada em buffer usa produtores (componentes que enviam mensagens para a fila), consumidores (componentes que recebem mensagens da fila) e uma fila (que contém mensagens) para armazenar as mensagens. As mensagens são lidas pelos consumidores e processadas, permitindo que as mensagens sejam executadas na taxa que atenda aos requisitos de negócios dos consumidores. Usando uma metodologia centrada em buffer, as mensagens dos produtores são armazenadas em filas ou fluxos, prontas para serem acessadas pelos consumidores em um ritmo alinhado às demandas operacionais.

Na AWS, você pode escolher entre vários serviços para implementar uma abordagem de buffering. O Amazon Simple Queue Service (Amazon SQS) é um serviço gerenciado que fornece filas que permitem que um único consumidor leia mensagens individuais. O Amazon Kinesis oferece um fluxo que permite que muitos consumidores leiam as mesmas mensagens.

O buffer e o controle de utilização podem suavizar qualquer pico modificando a demanda da workload. Use o controle de utilização quando os clientes repetirem ações e use o buffer para reter a solicitação e processá-la posteriormente. Ao trabalhar com uma arquitetura com uma abordagem baseada em buffer, arquitete a workload para atender à solicitação no tempo necessário e verifique se é possível lidar com solicitações duplicadas de trabalho. Analise a demanda geral, a taxa de alteração e o tempo de resposta necessário para dimensionar adequadamente o controle ou buffer necessário.

Etapas de implementação

- Analise os requisitos do cliente: analise as solicitações do cliente para determinar se são capazes de executar novas tentativas. Para clientes que não podem realizar novas tentativas, será necessário implementar buffers. Analise a demanda geral, a taxa de alteração e o tempo de resposta necessário para determinar o tamanho do controle de utilização ou do buffer necessário.
- Implemente um buffer ou controle de utilização: implemente um buffer ou controle de utilização na workload. Uma fila como o Amazon Simple Queue Service (Amazon SQS) pode fornecer um buffer para os componentes da workload. O Amazon API Gateway pode oferecer controle de utilização para componentes da workload.

Recursos

Práticas recomendadas relacionadas:

- SUS02-BP06 Implementar armazenamento em buffer ou controle de utilização para nivelar a curva da demanda
- REL05-BP02 Controlar a utilização de solicitações

Documentos relacionados:

- AWS Auto Scaling
- Agendador de instâncias da AWS
- Amazon API Gateway

- Amazon Simple Queue Service
- Conceitos básicos do Amazon SQS
- Amazon Kinesis

Vídeos relacionados:

- Escolher o serviço de mensagens certo para a aplicação distribuída

Exemplos relacionados:

- Gerenciar e monitorar o controle de utilização de APIs em suas workloads
- Controlar a utilização de uma API REST em camadas e multilocatário em grande escala usando o API Gateway
- Habilitar a hierarquização e o controle de utilização em uma solução SaaS multilocatária do Amazon EKS usando o Amazon API Gateway
- Integrar aplicações usando filas e mensagens

COST09-BP03 Fornecer recursos dinamicamente

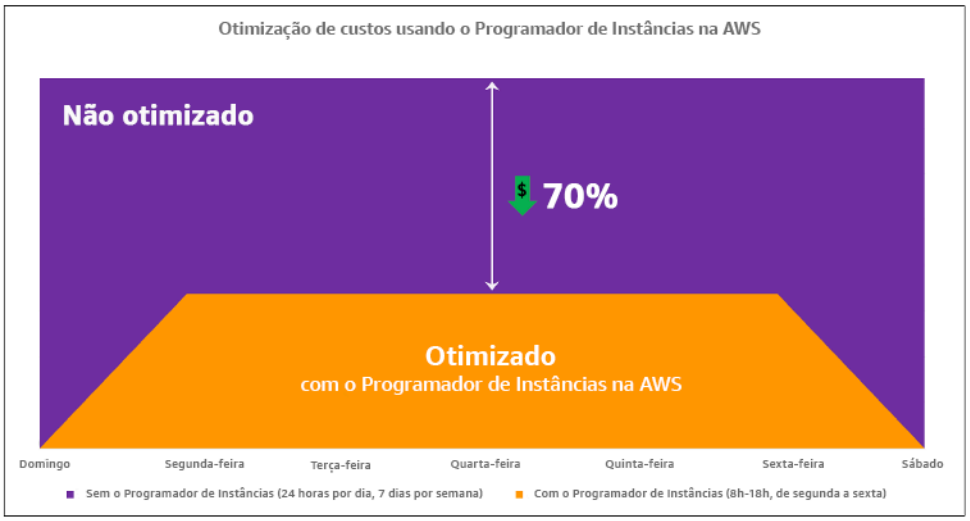
Os recursos são provisionados de maneira planejada. Isso pode ser feito com base na demanda, como por meio do ajuste de escala automático, ou com base no tempo, em que a demanda é previsível e os recursos são fornecidos em função do tempo. Esses métodos ocasionam a menor quantidade de superprovisionamento ou subprovisionamento.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Há várias maneiras como os clientes da AWS podem aumentar os recursos disponíveis para suas aplicações e fornecer recursos para atender à demanda. Uma dessas opções é usar o AWS Instance Scheduler, que automatiza o início e a interrupção de instâncias do Amazon Elastic Compute Cloud (Amazon EC2) e do Amazon Relational Database Service (Amazon RDS). A outra opção é usar o AWS Auto Scaling, que possibilita escalar automaticamente seus recursos de computação com base na demanda da aplicação ou do serviço. O fornecimento de recursos com base na demanda permitirá que você pague somente pelos recursos utilizados, reduza os custos lançando recursos quando eles forem necessários e os encerre quando não forem.

O Agendador de instâncias da AWS permite configurar a interrupção e o início de suas instâncias do Amazon EC2 e do Amazon RDS em horários definidos para que você possa atender à demanda pelos mesmos recursos em um padrão de tempo consistente, por exemplo, acesso diário dos usuários às instâncias do Amazon EC2 às 8h, que não são necessárias após as 18h. Essa solução ajuda a reduzir o custo operacional interrompendo recursos que não estão sendo usados e iniciá-los quando eles são necessários.



Otimização de custos com o Agendador de instâncias da AWS.

Você também pode configurar facilmente programações para suas instâncias do Amazon EC2 em suas contas e regiões com uma interface de usuário (IU) simples usando a Configuração rápida do AWS Systems Manager. É possível programar instâncias do Amazon EC2 e do Amazon RDS com o Agendador de instâncias da AWS e interromper e iniciar instâncias existentes. No entanto, não é possível parar e iniciar instâncias que fazem parte do seu grupo do Auto Scaling (ASG) ou que gerenciam serviços como o Amazon Redshift ou o Amazon OpenSearch Service. Os grupos do Auto Scaling têm seu próprio agendamento para as instâncias do grupo e essas instâncias são criadas.

O AWS Auto Scaling ajuda você a ajustar sua capacidade para manter uma performance estável e previsível pelo menor custo possível para atender às variações de demanda. Trata-se de um serviço totalmente gerenciado e gratuito para escalar a capacidade da aplicação e que se integra às instâncias do Amazon EC2 e às frotas spot, ao Amazon ECS, ao Amazon DynamoDB e ao Amazon Aurora. O Auto Scaling oferece descoberta automática de recursos para ajudar a encontrar recursos na sua workload que possam ser configurados, tem estratégias de ajuste de escala incorporadas para otimizar performance, custos ou um equilíbrio entre os dois, além de oferecer ajuste de escala preditivo para ajudar com picos que ocorrem regularmente.

Há várias opções de ajuste de escala disponíveis para escalar seu grupo do Auto Scaling:

- Manter níveis de instâncias atuais em todos os momentos
- Dimensionar manualmente
- Escala baseada em uma programação
- Escala com base em demanda
- Usar o ajuste de escala preditivo

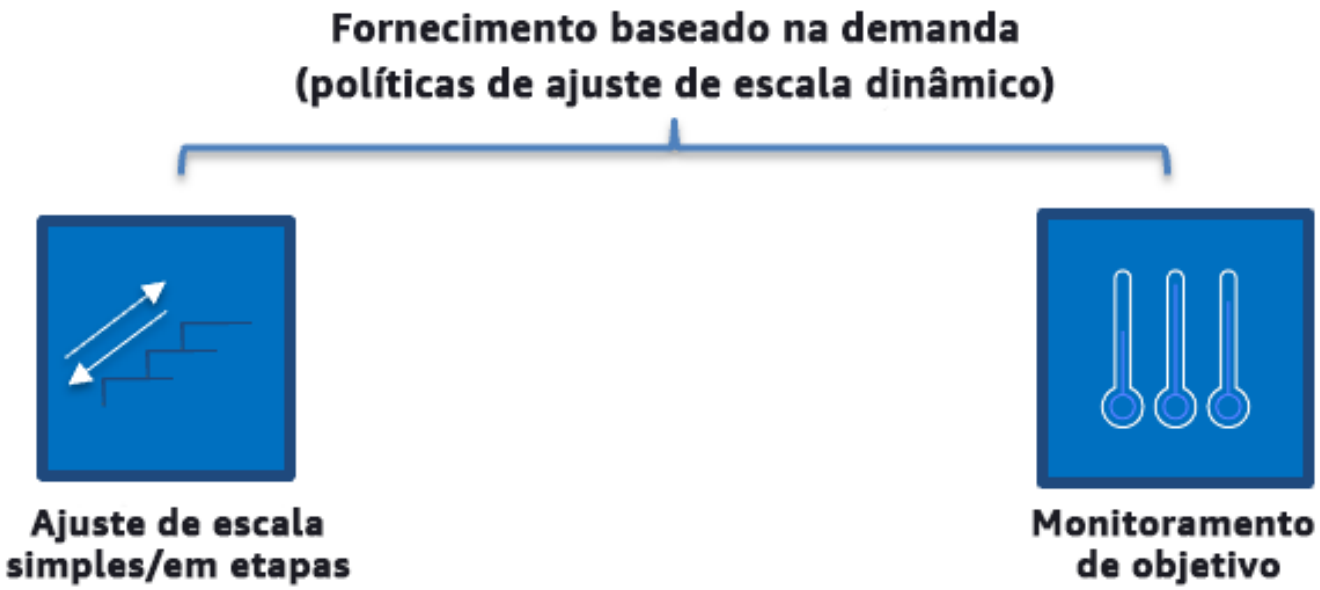
As políticas do Auto Scaling diferem e podem ser categorizadas como políticas de ajuste de escala dinâmicas e agendadas. As políticas dinâmicas são ajuste de escala manual ou dinâmico, ajuste de escala agendado ou preditivo. Você pode usar políticas para ajuste de escala dinâmico, agendado e preditivo. Você também pode usar métricas e alarmes do Amazon CloudWatch para acionar eventos de ajuste de escala para sua workload. Recomendamos usar modelos de execução para garantir que esteja acessando os recursos e melhorias mais recentes. Nem todos os recursos do Auto Scaling estão disponíveis quando você usa configurações de execução. Por exemplo, não é possível criar um grupo do Auto Scaling que execute instâncias spot e sob demanda ou que especifique vários tipos de instância. Você deve usar um modelo de execução para configurar esses recursos. Ao usar modelos de execução, recomendamos versionar cada um. Com o versionamento dos modelos de execução, você pode criar um subconjunto do conjunto completo de parâmetros. Em seguida, você pode reutilizá-lo para criar outras versões do mesmo modelo de execução.

É possível usar o AWS Auto Scaling ou incorporar ajuste de escala em seu código com as AWS APIs ou SDKs. Isso reduz os custos gerais da workload removendo o custo operacional de fazer alterações manualmente em seu ambiente, e quaisquer alterações podem ser realizadas muito mais rapidamente. Isso também atende à mobilização de recursos da workload de acordo com sua demanda a qualquer momento. Para seguir essa prática recomendada e fornecer recursos de forma dinâmica para sua organização, você precisa entender os ajustes de escala horizontal e vertical na Nuvem AWS, bem como a natureza das aplicações executadas em instâncias do Amazon EC2. É melhor para sua equipe de gerenciamento financeiro na nuvem trabalhar com equipes técnicas a fim de seguir essa prática recomendada.

O Elastic Load Balancing (ELB) ajuda você a escalar distribuindo a demanda entre vários recursos. Com o uso do ASG e do Elastic Load Balancing, você pode gerenciar as solicitações recebidas roteando o tráfego de forma ideal para que nenhuma instância fique sobrecarregada em um grupo do Auto Scaling. As solicitações seriam distribuídas entre todos os destinos de um grupo-alvo de forma contínua, sem considerar a capacidade nem a utilização.

As métricas típicas podem ser métricas padrão do Amazon EC2, como utilização de CPU, throughput de rede e latência de solicitação/resposta observada pelo Elastic Load Balancing. Quando possível, use uma métrica que seja indicativa da experiência do cliente. Normalmente é uma métrica personalizada que pode se originar do código da aplicação em sua workload. Para elaborar como atender à demanda dinamicamente neste documento, vamos agrupar o Auto Scaling em duas categorias, como modelos de fornecimento baseados na demanda e baseados no tempo, e nos aprofundarmos em cada uma delas.

Fornecimento baseado em demanda: utilize a elasticidade da nuvem para fornecer recursos para atender às mudanças na demanda e depender do estado de demanda quase em tempo real. Para fornecimento baseado em demanda, use as APIs ou os recursos de serviço para variar programaticamente a quantidade de recursos de nuvem em sua arquitetura. Isso permite que você ajuste a escala de componentes em sua arquitetura e aumente o número de recursos durante picos de demanda a fim de manter a performance e reduzir a capacidade quando a demanda diminui para reduzir os custos.

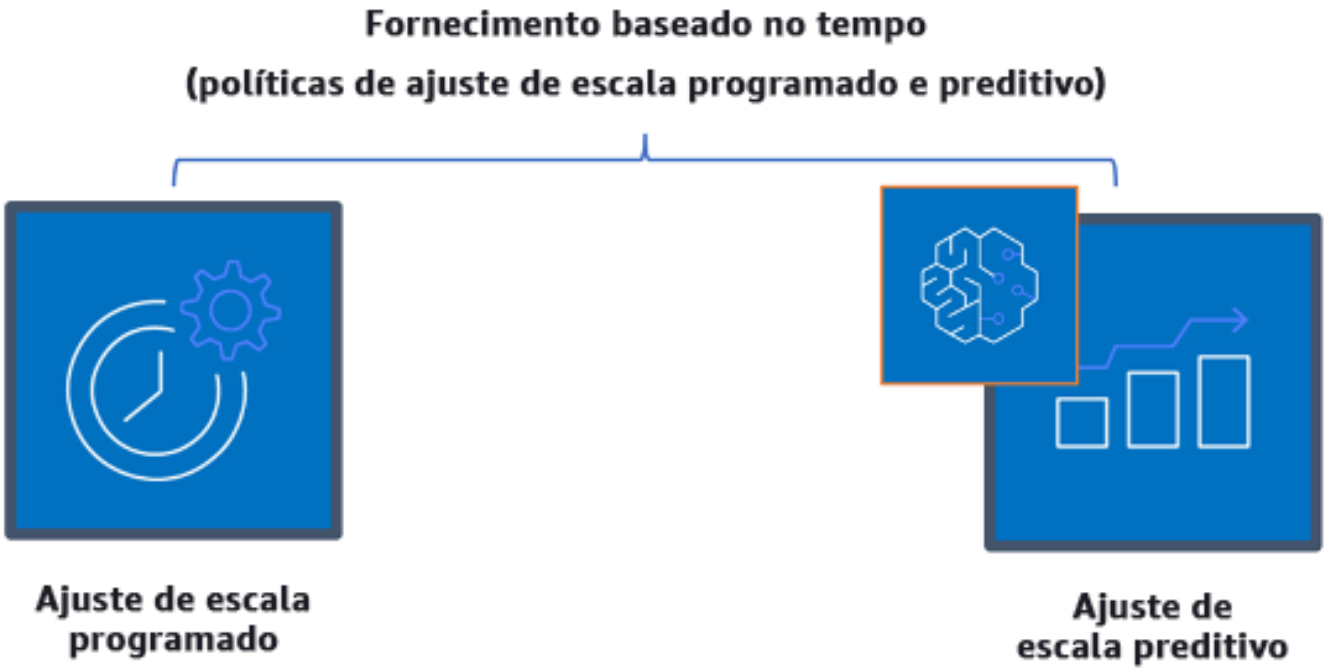


Políticas de ajuste de escala dinâmico com base na demanda

- Ajuste de escala simples/em etapas: monitora métricas e adiciona/remove instâncias de acordo com as etapas definidas manualmente pelos clientes.
- Monitoramento do objetivo: mecanismo de controle semelhante a um termostato que adiciona ou remove instâncias automaticamente para manter as métricas em um objetivo definido pelo cliente.

Ao arquitetar com uma abordagem baseada em demanda, tenha em mente dois pontos essenciais. Primeiro, entenda a rapidez com que você deve provisionar novos recursos. Segundo, entenda que o tamanho da margem entre fornecimento e demanda mudará. Você deve estar pronto para lidar com a taxa de alteração na demanda e também estar pronto para falhas de recursos.

Fornecimento com base em tempo: uma abordagem baseada em tempo alinha a capacidade de recursos à demanda previsível ou bem definida em relação ao tempo. Essa abordagem costuma não depender dos níveis de utilização dos recursos. Uma abordagem baseada em tempo garante que os recursos estejam disponíveis no momento específico em que são necessários e podem ser fornecidos sem nenhum atraso devido a procedimentos de inicialização e verificações do sistema ou de consistência. Usando uma abordagem baseada em tempo, você pode fornecer recursos adicionais ou aumentar a capacidade durante períodos ocupados.



Políticas de ajuste de escala baseado em tempo

Você pode usar o ajuste de escala automático agendado ou preditivo para implementar uma abordagem baseada em tempo. As workloads podem ser agendadas para aumentar ou reduzir a escala horizontalmente em horários definidos (por exemplo, o início do horário comercial), tornando os recursos disponíveis quando os usuários chegarem ou a demanda aumentar. O ajuste de escala preditivo usa padrões para aumentar a escala horizontalmente enquanto o ajuste de escala agendado usa horários predefinidos para isso. Você também pode usar uma estratégia de

seleção de tipo de instância baseada em atributo (ABS) em grupos do Auto Scaling, o que permite expressar seus requisitos de instância como um conjunto de atributos, como vCPU, memória e armazenamento. Isso também permite usar automaticamente os tipos de instância de geração mais recente quando eles são lançados e acessar uma variedade mais ampla de capacidade com instâncias spot do Amazon EC2. O Amazon EC2 Fleet e o Amazon EC2 Auto Scaling selecionam e executam instâncias que se ajustam aos atributos especificados, eliminando a necessidade de escolher manualmente os tipos de instância.

Você também pode aproveitar as APIs e os SDKs da AWS e o AWS CloudFormation para provisionar e desativar automaticamente ambientes inteiros conforme necessário. Essa abordagem é adequada para ambientes de desenvolvimento ou teste que são executados apenas em períodos ou horários comerciais definidos. Você pode usar APIs para ajustar a escala dos recursos dentro de um ambiente (ajuste de escala vertical). Por exemplo, você pode escalar uma workload de produção alterando o tamanho ou a classe da instância. Isso pode ser feito interrompendo e iniciando a instância e selecionando a classe ou o tamanho da instância diferente. Essa técnica também pode ser aplicada a outros recursos, como Volumes Elásticos do Amazon EBS, que podem ser modificados para aumentar o tamanho, ajustar a performance (IOPS) ou alterar o tipo de volume durante o uso.

Ao arquitetar com uma abordagem baseada em tempo, tenha em mente dois pontos essenciais. Primeiro, qual é a consistência do padrão de uso? Segundo, qual será o impacto se o padrão mudar? Você pode aumentar a precisão das previsões monitorando suas workloads e usando inteligência de negócios. Se você vir alterações significativas no padrão de uso, poderá ajustar os tempos para garantir que a cobertura seja fornecida.

Etapas de implementação

- Configure o ajuste de escala agendado: para alterações previsíveis na demanda, o ajuste de escala com base em tempo pode fornecer a quantidade correta de recursos em tempo hábil. Ele também será útil se a criação e a configuração de recursos não forem rápidas o suficiente para responder a alterações na demanda. Usando a análise de workload, configure o ajuste de escala agendado usando o AWS Auto Scaling. Para configurar o cronograma baseado em tempo, você pode usar o ajuste de escala preditivo ou agendado para aumentar o número de instâncias do Amazon EC2 em seus grupos do Auto Scaling com antecedência de acordo com as alterações de carga esperadas ou previsíveis.
- Configure o ajuste de escala preditivo: o ajuste de escala preditivo permite aumentar o número de instâncias do EC2 em seu grupo do Auto Scaling em antecipação aos padrões diários e semanais nos fluxos de tráfego. Se você tem picos de tráfego regulares e aplicações que demoram muito

para ser iniciadas, considere usar o ajuste de escala preditivo. O ajuste de escala preditivo pode ajudar você a escalar com maior rapidez inicializando a capacidade antes da carga projetada em comparação com o ajuste de escala dinâmico isolado, que é reativo por natureza. Por exemplo, se os usuários começarem a usar sua workload no início do horário comercial e não usá-la após o expediente, o ajuste de escala preditivo poderá adicionar capacidade antes do horário comercial, o que elimina o atraso do ajuste de escala dinâmico para reagir a mudanças no tráfego.

- Configure o ajuste de escala automático dinâmico: para configurar o ajuste de escala com base em métricas de workload ativas, use o Auto Scaling. Use a análise e configure o Auto Scaling para iniciar nos níveis de recursos corretos e garanta que a workload escale no tempo necessário. Você pode iniciar e escalar automaticamente uma frota de instâncias sob demanda e instâncias spot em um único grupo do Auto Scaling. Além de receber descontos pelo uso de instâncias spot, você pode usar instâncias reservadas ou um Savings Plan para receber taxas de desconto da definição de preço normal de instância sob demanda. Todos esses fatores combinados ajudam você a otimizar sua redução de custos para instâncias do Amazon EC2 e a obter a escala e a performance desejadas para a aplicação.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- Agendador de instâncias da AWS
- Escalar o tamanho do grupo do Auto Scaling
- Conceitos básicos do Amazon EC2 Auto Scaling
- Conceitos básicos do Amazon SQS
- Ajuste de escala agendado para o Amazon EC2 Auto Scaling
- Escala preditiva para o Amazon EC2 Auto Scaling

Vídeos relacionados:

- Políticas de ajuste de escala com monitoramento do objetivo para Auto Scaling
- Agendador de instâncias da AWS

Exemplos relacionados:

- Seleção de tipo de instância baseada em atributos para Auto Scaling para Amazon EC2 Fleet
- Otimizar o Amazon Elastic Container Service para custos usando ajuste de escala agendado
- Ajuste de escala preditivo com o Amazon EC2 Auto Scaling
- Como faço para usar o Agendador de instâncias com o AWS CloudFormation para agendar as instâncias do Amazon EC2?)

Otimização ao longo do tempo

Perguntas

- COST 10. Como avaliar os novos serviços?
- COST 11. Como avaliar o custo do esforço?

COST 10. Como avaliar os novos serviços?

À medida que a AWS lança novos serviços e recursos, uma das práticas recomendadas é avaliar suas decisões sobre a arquitetura existente a fim de garantir que elas ofereçam o melhor custo-benefício.

Práticas recomendadas

- COST10-BP01 Desenvolver um processo de revisão de workloads
- COST10-BP02 Revisar e analisar a workload regularmente

COST10-BP01 Desenvolver um processo de revisão de workloads

Desenvolva um processo que defina os critérios e o processo para a revisão de workloads. O esforço de análise deve refletir o benefício potencial. Por exemplo, workloads principais ou workloads com valor superior a 10% da fatura são revisadas trimestralmente ou a cada seis meses, enquanto workloads abaixo de 10% são revisadas anualmente.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Para manter a workload mais econômica, é necessário revisá-la regularmente para saber se há oportunidades de implementar novos serviços, recursos e componentes. Para obter custos gerais mais baixos, o processo deve ser proporcional à quantidade potencial de economia. Por exemplo, as workloads que representam 50% do seu gasto geral devem ser revisadas com mais

frequência e mais precisão do que as workloads que representam 5% do seu gasto geral. Leve em consideração quaisquer fatores externos ou volatilidade. Se a workload atender a uma área geográfica ou segmento de mercado específico e houver previsão de mudanças nessa área, revisões mais frequentes poderão resultar em economias de custos. Outro fator em questão é o esforço para implementar alterações. Se houver custos significativos em testes e validação de alterações, as revisões deverão ser menos frequentes.

Leve em consideração o custo de longo prazo de manutenção de componentes e recursos obsoletos e a incapacidade de implementar novos recursos neles. O custo atual de testes e validação pode exceder o benefício proposto. No entanto, ao longo do tempo, o custo de fazer a mudança pode aumentar significativamente à medida que a lacuna entre a workload e as tecnologias atuais aumenta, resultando em custos ainda maiores. Por exemplo, o custo da migração para uma nova linguagem de programação pode não ser econômico no momento. No entanto, em cinco anos, o custo de pessoas com qualificações nessa linguagem pode aumentar e, devido ao crescimento da workload, você estaria movendo um sistema ainda maior para a nova linguagem, exigindo ainda mais esforço do que anteriormente.

Divida sua workload em componentes, atribua o custo do componente (uma estimativa é suficiente) e liste os fatores (por exemplo, esforço e mercados externos) ao lado de cada componente. Use esses indicadores para determinar uma frequência de revisão para cada workload. Por exemplo, você pode ter servidores web como um alto custo, baixo esforço de alteração e fatores externos elevados, o que resulta em uma alta frequência de revisão. Um banco de dados central pode ser de custo médio, alto esforço de alteração e baixos fatores externos, resultando em uma média frequência de análise.

Defina um processo para avaliar novos serviços, padrões de design, tipos de recursos e configurações para otimizar o custo de sua workload conforme ficarem disponíveis. Semelhante aos processos de revisão do pilar Performance e revisão do pilar Confiabilidade, identifique, valide e priorize as atividades de otimização e melhoria e a remediação de problemas e incorpore-as à sua lista de pendências.

Etapas de implementação

- Defina a frequência de revisões: defina a frequência com que a workload e os respectivos componentes devem ser revisados. Aloque tempo e recursos para o aprimoramento contínuo e analise a frequência para melhorar a eficiência e a otimização da sua workload. Essa é uma combinação de fatores e pode diferir de workload para workload em sua organização e entre componentes na workload. Os fatores comuns incluem: a importância para a organização medida em termos de receita ou marca, o custo total da execução da workload (incluindo custos operacionais e de recursos), a complexidade da workload, a facilidade da implementação de uma

alteração, qualquer contrato de licenciamento de software e se uma alteração geraria aumentos significativos nos custos de licenciamento devido a licenciamento punitivo. Os componentes podem ser definidos de maneira funcional ou técnica, como bancos de dados e servidores Web ou recursos de computação e armazenamento. Equilibre os fatores de acordo e desenvolva um período para a workload e os respectivos componentes. Você pode decidir analisar a workload completa a cada 18 meses, analisar os servidores Web a cada seis meses, o banco de dados a cada doze meses, a computação e o armazenamento de curto prazo a cada seis meses e o armazenamento de longo prazo a cada doze meses.

- Defina o rigor da revisão: defina quanto esforço é gasto na revisão da workload ou dos respectivos componentes. Semelhante à frequência da análise, esse é um equilíbrio de vários fatores. Avalie e priorize oportunidades de melhorias para concentrar os esforços nos locais onde eles oferecem os maiores benefícios enquanto calcula quanto esforço é necessário para essas atividades. Se os resultados esperados não satisfizerem as metas e o esforço necessário custar mais, itere usando cursos de ação alternativos. Seus processos de análise devem incluir tempo e recursos dedicados para possibilitar melhorias incrementais contínuas. Por exemplo, você pode decidir gastar uma semana de análise no componente do banco de dados, uma semana de análise para recursos computacionais e quatro horas para análises de armazenamento.

Recursos

Documentos relacionados:

- Notícias do blog da AWS
- Tipos de computação em nuvem
- Novidades da AWS

Exemplos relacionados:

- Serviços proativos do AWS Support
- Revisões regulares da workload para workloads do SAP

COST10-BP02 Revisar e analisar a workload regularmente

As workloads existentes são revisadas regularmente com base em cada processo definido para descobrir se é possível adotar novos serviços, substituir serviços já em vigor ou refazer a arquitetura das workloads.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A AWS adiciona constantemente novos recursos para que você possa experimentar e novar mais rápido com a tecnologia mais recente. AWS Novidades detalha como a AWS está fazendo isso e fornece uma visão geral rápida dos serviços, recursos e anúncios de expansão regional da AWS à medida que são lançados. Você pode examinar detalhadamente os lançamentos que foram anunciados e usá-los para revisar e analisar suas workloads existentes. Para obter os benefícios de novos serviços e recursos da AWS, analise suas workloads e implemente novos serviços e recursos conforme necessário. Isso significa que você pode precisar substituir os serviços que você usa para a workload ou modernizar a workload para adotar novos serviços da AWS. Por exemplo, você pode revisar suas workloads e substituir o componente de mensagens pelo Amazon Simple Email Service. Fazer isso elimina os custos de operação e manutenção de uma frota de instâncias e, ao mesmo tempo, fornece toda a funcionalidade a um custo reduzido.

Para analisar sua workload e destacar possíveis oportunidades, considere não apenas novos serviços, mas também novas formas de criar soluções. Veja os vídeos Esta é a minha arquitetura na AWS para saber mais sobre os projetos de arquitetura de outros clientes, seus desafios e suas soluções. Confira a série All-In para descobrir aplicações reais de serviços da AWS e histórias de clientes. Você também pode assistir à série de vídeos De volta ao básico que explica, examina e detalha as práticas recomendadas básicas para padrões de arquitetura de nuvem. Outra fonte são os vídeos Como fazer isto, criados para ajudar pessoas com grandes ideias sobre como dar vida a seu produto mínimo viável (MVP) usando serviços da AWS. Desse modo, criadores do mundo inteiro que tiverem uma grande ideia poderão obter orientações arquiteturais de arquitetos de soluções da AWS experientes. Finalmente, você pode revisar os materiais de recursos de Conceitos básicos, os quais oferecem tutoriais passo a passo.

Antes de iniciar seu processo de avaliação, siga os requisitos de sua empresa com relação a workload, segurança e privacidade dos dados para usar requisitos específicos de serviço ou de região e performance e, ao mesmo tempo, siga o processo de avaliação que foi acordado.

Etapas de implementação

- Revise regularmente a workload: usando o processo definido, execute análises com a frequência especificada. Verifique se você despendeu a quantidade correta de esforço em cada componente. Esse processo seria semelhante ao processo de design inicial em que você selecionou serviços para otimização de custos. Analise os serviços e os benefícios que eles trariam, esse fator de tempo no custo de fazer a mudança, e não apenas os benefícios de longo prazo.

- Implemente novos serviços: se o resultado da análise for implementar alterações, primeiro execute uma linha de base da workload para saber o custo atual por saída. Implemente as alterações e, em seguida, execute uma análise para confirmar o novo custo por saída.

Recursos

Documentos relacionados:

- Notícias do blog da AWS
- Novidades da AWS
- Documentação do AWS
- Conceitos básicos da AWS
- Recursos gerais da AWS

Vídeos relacionados:

- AWS: Esta é a minha arquitetura
- AWS: De volta ao básico
- AWS: Série All-In
- Como fazer isto

COST 11. Como avaliar o custo do esforço?

Práticas recomendadas

- COST11-BP01 Realizar automações nas operações

COST11-BP01 Realizar automações nas operações

Avalie os custos operacionais na nuvem, concentrando-se em quantificar a economia de tempo e de esforço em tarefas administrativas e implantações, bem como em mitigar o risco de erros humanos, conformidade e outras operações por meio da automação. Avalie os custos associados e o tempo necessários para os esforços operacionais e implemente a automação de tarefas administrativas para minimizar o esforço manual sempre que possível.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

A automação de operações reduz a frequência das tarefas manuais, melhora a eficiência e beneficia os clientes por oferecer uma experiência consistente e confiável em implantação, administração ou operação de workloads. Você pode liberar recursos de infraestrutura de tarefas operacionais manuais e usá-los para tarefas e inovações de maior valor, o que melhora os resultados dos negócios. As empresas necessitam de um método comprovado e testado para gerenciar suas workloads na nuvem. Essa solução deve ser segura, rápida e econômica, com risco mínimo e máxima confiabilidade.

Primeiro, priorize suas atividades operacionais com base no esforço necessário examinando o custo geral das operações. Por exemplo, quanto tempo é preciso para implantar novos recursos na nuvem, realizar alterações de otimização nos recursos existentes ou implementar as configurações necessárias? Examine o custo total das ações humanas, incluindo o custo de operações e gerenciamento como fator. Priorize a automação das tarefas administrativas para reduzir o esforço humano.

A avaliação do esforço deve refletir o provável benefício. Por exemplo, examine o tempo gasto na execução de tarefas manuais em contraposição a tarefas automáticas. Priorize a automatização de atividades repetitivas, de alto valor, demoradas e complexas. As atividades que apresentam um alto valor ou alto risco de erro humano normalmente são o melhor lugar para começar a automatizar, pois, com frequência, o risco cria um custo operacional adicional não desejado (como horas extras de trabalho da equipe de operações).

Use ferramentas de automação, como o AWS Systems Manager ou o AWS Config para racionalizar processos de operações, conformidade, monitoramento, ciclo de vida e encerramento. Com serviços e ferramentas da AWS, além de produtos de terceiros, você pode personalizar as automações implementadas para atender às suas necessidades específicas. A tabela abaixo mostra alguns recursos e funções essenciais de operações que você pode obter com os serviços da AWS para automatizar a administração e a operação:

- AWS Audit Manager: audite continuamente o uso da AWS para simplificar a avaliação de risco e conformidade.
- AWS Backup: gerencie e automatize centralmente a proteção de dados para o Amazon.
- AWS Config: configure recursos computacionais, avalie, audite e analise configurações e o inventário de recursos.
- AWS CloudFormation: inicie recursos altamente disponíveis com a infraestrutura como código.
- AWS CloudTrail: gerenciamento, conformidade e controle de mudanças de TI.

- Agende eventos do Amazon EventBridge e acione o AWS Lambda para adotar medidas.
- AWS Lambda: automatize processos repetitivos acionando-os com eventos ou executando-os em um cronograma fixo com o AWS EventBridge.
- AWS Systems Manager: inicie e interrompa workloads, corrija sistemas operacionais, automatize a configuração e gerencie continuamente.
- AWS Step Functions: Agende trabalhos e automatize fluxos de trabalho.
- AWS Service Catalog: Consumo de modelos, infraestrutura como código com conformidade e controle.

Se você quiser adotar automações imediatamente com o uso de produtos e serviços da AWS e não contar com as habilidades necessárias em sua organização, entre em contato com AWS Managed Services (AMS), a AWS Professional Services ou parceiros da AWS para aumentar a adoção de automação e melhorar sua excelência operacional na nuvem.

A AWS Managed Services (AMS) é um serviço que opera a infraestrutura da AWS em nome de clientes e parceiros corporativos. Ele fornece um ambiente seguro e compatível no qual você pode implantar as workloads. O AMS usa modelos operacionais de nuvem empresarial com automação para permitir que você atenda aos requisitos da organização, migre para a nuvem mais rapidamente e reduza os custos de gerenciamento constantes.

A AWS Professional Services também pode ajudar você a alcançar os resultados de negócios desejados e a automatizar as operações com a AWS. Essa equipe ajuda os cliente a implantar operações de TI automatizadas, robustas e ágeis, bem como recursos de governança otimizados para a nuvem. Para ver exemplos detalhados de monitoramento e práticas recomendadas, consulte o whitepaper Pilar Excelência operacional.

Etapas de implementação

- Crie uma vez e implante várias: use infraestrutura como código, como CloudFormation, AWS SDK ou AWS CLI para implantar uma vez e usar várias vezes em ambientes semelhantes ou em cenários de recuperação de desastres. Marque com tags enquanto monitora o consumo conforme definido em outras práticas recomendadas. Use o AWS Launch Wizard para reduzir o tempo de implantação de muitas workloads corporativas populares. A AWS Launch Wizard orienta você no dimensionamento, na configuração e na implantação de workloads corporativas de acordo com as práticas recomendadas da AWS. Você também pode usar o Service Catalog, que ajuda a criar e gerenciar modelos aprovados de infraestrutura como código para uso na AWS para que qualquer pessoa possa descobrir recursos de nuvem de autoatendimento aprovados.

- Automatize a conformidade contínua: considere automatizar a avaliação e a correção das configurações registradas em relação a padrões predefinidos. Ao combinar o AWS Organizations com os recursos do AWS Config e AWS CloudFormation, você pode gerenciar e automatizar com eficiência a conformidade da configuração em grande escala para centenas de contas-membro. É possível analisar as mudanças nas configurações e nas relações entre os recursos da AWS e examinar em detalhe o histórico da configuração de um recurso.
- Automatização das tarefas de monitoramento: a AWS fornece várias ferramentas que você pode usar para monitorar serviços. É possível configurar essas ferramentas para automatizar as tarefas de monitoramento. Crie e implemente um plano de monitoramento que colete dados de todas as partes da workload para que você possa depurar com maior facilidade uma falha multiponto, caso ocorra. Por exemplo, é possível usar as ferramentas de monitoramento automatizadas para observar o Amazon EC2 e informar quando algo está errado nas verificações de status do sistema, nas verificações de status de instâncias e nos alarmes do Amazon CloudWatch.
- Automatize a manutenção e as operações: execute operações de rotina automaticamente sem intervenção humana. Usando serviços e ferramentas da AWS, você pode escolher quais automações da AWS deverão ser implementadas e personalizadas de acordo com seus requisitos específicos. Por exemplo, use o EC2 Image Builder para criar, testar e implantar imagens de máquinas virtuais e de contêineres para uso na AWS ou on-premises ou para aplicar patches em suas instâncias do EC2 com o AWS SSM. Se a ação desejada não puder ser realizada com serviços da AWS, ou se você precisar de ações mais complexas com recursos de filtragem, automatize suas operações usando o AWS Command Line Interface (AWS CLI) ou as ferramentas do AWS SDK. A AWS CLI oferece a capacidade de automatizar todo o processo de controle e gerenciamento de serviços da AWS com scripts sem usar o AWS Management Console. Selecione seus AWS SDKs preferidos para interagir com os serviços da AWS. Para outros exemplos de código, consulte o repositório de exemplos de código do AWS SDK.
- Crie um ciclo de vida contínuo com automações: é importante que você estabeleça e preserve políticas maduras de ciclo de vida, não apenas para regulamentações ou redundância, mas também para otimização de custos. É possível usar o AWS Backup para gerenciar e automatizar centralmente a proteção de datastores, como buckets, volumes, bancos de dados e sistemas de arquivos. Você também pode usar o Amazon Data Lifecycle Manager para automatizar a criação, a retenção e a exclusão de snapshots do EBS e de AMIs apoiadas pelo EBS.
- Exclua recursos desnecessários: é muito comum acumular recursos não utilizados no sandbox ou em Contas da AWS de desenvolvimento. Os desenvolvedores criam e experimentam vários serviços e recursos como parte do ciclo normal de desenvolvimento, mas não excluem esses recursos quando não são mais necessários. Recursos não utilizados podem gerar custos desnecessários e, às vezes, altos para a organização. A exclusão desses recursos pode reduzir os

custos de operação desses ambientes. Certifique-se de que seus dados não sejam necessários. Se não tiver certeza, faça backup. Você pode usar o AWS CloudFormation para limpar as pilhas implantadas, o que exclui automaticamente a maioria dos recursos definidos no modelo. Como alternativa, você pode criar uma automação para a exclusão de recursos da AWS usando ferramentas como o aws-nuke.

Recursos

Documentos relacionados:

- Modernizar as operações na Nuvem AWS
- Usar serviços da AWS para automação
- Infraestrutura e automação
- AWS Systems Manager Automation
- Monitoramento automatizado e manual
- Automações da AWS para administração e operações do SAP
- AWS Managed Services
- AWS Professional Services

Vídeos relacionados:

- Automatizar a conformidade contínua em grande escala na AWS
- Demonstração do AWS Backup: backup entre contas e regiões
- Correção de patches para suas instâncias do Amazon EC2

Exemplos relacionados:

- Reinventar as operações automatizadas (Parte I)
- Reinventar as operações automatizadas (Parte II)
- Automatizar a exclusão de recursos da AWS usando o aws-nuke
- Excluir volumes não utilizados do Amazon EBS com o AWS Config e o AWS SSM
- Automatizar a conformidade contínua em grande escala na AWS
- Automações de TI com o AWS Lambda

Sustentabilidade

O pilar Sustentabilidade abrange ações como compreender os impactos dos serviços usados, quantificá-los ao longo do ciclo de vida da workload e aplicar princípios e práticas recomendadas de design para reduzi-los ao criar workloads na nuvem. Recomendações sobre implementação estão disponíveis no whitepaper Pilar Sustentabilidade.

Áreas de práticas recomendadas

- Seleção da região
- Alinhamento com a demanda
- Software e arquitetura
- Dados
- Hardware e serviços
- Processo e cultura

Seleção da região

Pergunta

- SUS 1 Como selecionar regiões para sua workload?

SUS 1 Como selecionar regiões para sua workload?

A escolha da região para sua workload afeta significativamente seus KPIs, incluindo performance, custo e pegada de carbono. Para melhorar efetivamente esses KPIs, você deve escolher regiões para suas workloads com base em requisitos empresariais e metas de sustentabilidade.

Práticas recomendadas

- SUS01-BP01 Escolher a região com base nos requisitos empresariais e nas metas de sustentabilidade

SUS01-BP01 Escolher a região com base nos requisitos empresariais e nas metas de sustentabilidade

Escolha uma região para sua workload com base em seus requisitos empresariais e metas de sustentabilidade para otimizar seus KPIs, incluindo performance, custo e pegada de carbono.

Práticas comuns que devem ser evitadas:

- Selecionar a região da workload com base em sua localização.
- Consolidar todos os recursos da workload em uma única localização geográfica.

Benefícios de implementar esta prática recomendada: colocar uma workload próxima aos projetos de energia renovável da Amazon ou às regiões com baixa intensidade de carbono publicada pode ajudar a reduzir a pegada de carbono de uma workload na nuvem.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A Nuvem AWS é uma rede em constante expansão de regiões e pontos de presença (PoP) com uma infraestrutura de rede global que conecta uns aos outros. A escolha da região para sua workload afeta significativamente seus KPIs, incluindo performance, custo e pegada de carbono. Para melhorar efetivamente esses KPIs, você deve escolher regiões para sua workload com base em seus requisitos empresariais e metas de sustentabilidade.

Etapas de implementação

- Siga estas etapas para avaliar e selecionar possíveis regiões para sua workload com base em seus requisitos de negócios, incluindo conformidade, recursos disponíveis, custo e latência:
  - Confirme se essas regiões estão em conformidade com as regulamentações locais aplicáveis.
  - Use as Listas de serviços regionais da AWS para verificar se as regiões têm os serviços e recursos necessários para executar sua workload.
  - Calcule o custo da workload em cada região usando o AWS Pricing Calculator.
  - Teste a latência de rede entre as localizações de seus usuários finais e cada Região da AWS.
- Escolha regiões próximas aos projetos de energia renovável da Amazon e regiões onde a grade de intensidade de carbono publicada esteja abaixo de outros locais (ou regiões).
  - Identifique suas diretrizes de sustentabilidade relevantes para rastrear e comparar as emissões de carbono ano a ano com base no Protocolo de Gases de Efeito Estufa (métodos baseados no mercado e na localização).
  - Escolha a região com base no método que você utiliza para rastrear as emissões de carbono. Para obter mais detalhes sobre como escolher uma região em função das suas diretrizes de sustentabilidade, consulte Como selecionar uma região para sua workload com base em metas de sustentabilidade.

Recursos

Documentos relacionados:

- Noções básicas das suas estimativas de emissão de carbono
- Amazon ao redor do mundo
- Metodologia de energia renovável
- O que considerar ao selecionar uma região para suas workloads

Vídeos relacionados:

- AWS re:Invent 2023: Inovação sustentável na infraestrutura global da AWS
- AWS re:Invent 2023: Arquitetura sustentável: passado, presente e futuro
- AWS re:Invent 2022: Como entregar arquiteturas sustentáveis e de alta performance
- AWS re:Invent 2022: Arquitetar de forma sustentável e reduzir sua pegada de carbono da AWS
- AWS re:Invent 2022: Sustentabilidade na infraestrutura global da AWS

Alinhamento com a demanda

Pergunta

- SUS 2 Como alinhar recursos de nuvem à sua demanda?

SUS 2 Como alinhar recursos de nuvem à sua demanda?

A maneira como os usuários e as aplicações consomem suas workloads e outros recursos pode ajudar você a identificar melhorias para atingir metas de sustentabilidade. Escale a infraestrutura de forma que ela corresponda à demanda e use apenas os recursos mínimos necessários para oferecer suporte aos usuários. Alinhe os níveis de serviço às necessidades do cliente. Posicione os recursos a fim de limitar a rede necessária para que usuários e aplicações os consumam. Elimine ativos não utilizados. Forneça aos membros da sua equipe dispositivos compatíveis com suas necessidades e minimize o impacto na sustentabilidade.

Práticas recomendadas

- SUS02-BP01 Escalar a infraestrutura da workload dinamicamente
- SUS02-BP02 Alinhar os SLAs às metas de sustentabilidade

- SUS02-BP03 Interromper a criação e a manutenção de ativos não utilizados
- SUS02-BP04 Otimizar o posicionamento geográfico das workloads com base nos respectivos requisitos de rede
- SUS02-BP05 Otimizar os recursos dos membros da equipe para as atividades realizadas
- SUS02-BP06 Implementar armazenamento em buffer ou controle de utilização para nivelar a curva da demanda

SUS02-BP01 Escalar a infraestrutura da workload dinamicamente

Use a elasticidade da nuvem e escale sua infraestrutura de forma dinâmica para corresponder a oferta de recursos de nuvem à demanda e evitar capacidade superprovisionada em sua workload.

Práticas comuns que devem ser evitadas:

- Você não escala sua infraestrutura de acordo com a carga de usuários.
- Você escala sua infraestrutura manualmente o tempo todo.
- Manter a capacidade aumentada após um evento de ajuste de escala, em vez de reduzi-la novamente.

Benefícios de implementar esta prática recomendada: configurar e testar a elasticidade da workload ajuda a adequar eficientemente a oferta de recursos de nuvem à demanda e evitar o excesso de capacidade provisionada. É possível aproveitar a elasticidade na nuvem para escalar automaticamente a capacidade durante e depois de picos de demanda para garantir que esteja usando apenas o número exato de recursos necessários para atender aos requisitos do seu negócio.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A nuvem fornece a flexibilidade necessária para expandir ou reduzir seus recursos dinamicamente por meio de diversos mecanismos para atender a mudanças na demanda. O equilíbrio ideal entre a oferta e a demanda oferece o menor impacto ambiental para uma workload.

A demanda pode ser fixa ou variável, exigindo métricas e automação para garantir que o gerenciamento não se torne um gasto excessivo. As aplicações podem aumentar ou diminuir a escala verticalmente ao modificar o tamanho da instância e horizontalmente ao modificar o número de instâncias, ou ainda uma combinação de ambos.

Você pode usar diversas abordagens diferentes para corresponder a oferta de recursos com a demanda.

- Abordagem de rastreamento de métricas: monitore a métrica de ajuste de escala e aumente ou diminua automaticamente a capacidade conforme necessário.
- Ajuste de escala preditivo: aumente ou reduza a escala em antecipação às tendências diárias e semanais.
- Abordagem baseada em cronograma: defina seu próprio cronograma de ajuste de escala de acordo com as mudanças de carga previsíveis.
- Dimensionamento de serviços: escolha serviços (como sem servidor) cuja escala seja modificada de forma nativa por design ou forneçam o ajuste de escala automático como um recurso.

Identifique períodos de utilização baixa ou sem utilização e escale os recursos para eliminar a capacidade em excesso e melhorar a eficiência.

Etapas de implementação

- A elasticidade corresponde à oferta de recursos que você tem face à demanda por estes recursos. Instâncias, contêineres e funções fornecem mecanismos para elasticidade, seja em combinação com o ajuste de escala automático ou como um recurso do serviço. A AWS fornece uma variedade de mecanismos de ajuste de escala automático para garantir que as workloads possam reduzir a escala verticalmente de forma rápida e fácil durante períodos de baixa carga de usuários. Veja alguns exemplos de mecanismos de ajuste de escala automático:

Mecanismo de ajuste de escala automático	Onde usar
Amazon EC2 Auto Scaling	Use para verificar se você tem o número correto de instâncias do Amazon EC2 disponíveis para processar a carga de usuário para a sua aplicação.
Application Auto Scaling	Use para escalar automaticamente os recursos para serviços da AWS individua is além do Amazon EC2, como funções do Lambda ou serviços do Amazon Elastic Container Service (Amazon ECS).

Mecanismo de ajuste de escala automático	Onde usar
Kubernetes Cluster Autoscaler	Use para escalar automaticamente os clusters do Kubernetes na AWS.

- O ajuste de escala geralmente é discutido em relação a serviços de computação, como instâncias do Amazon EC2 ou funções do AWS Lambda. Considere a configuração de serviços não computacionais, como unidades de capacidade de leitura e gravação do Amazon DynamoDB ou fragmentos do Amazon Kinesis Data Streams para atender à demanda.
- Verifique se as métricas para aumentar ou reduzir a escala verticalmente são validadas em relação ao tipo da workload que está sendo implantada. Se você estiver implantando uma aplicação de transcodificação de vídeo, espera-se que a utilização da CPU seja de 100%, e essa não deve ser sua métrica principal. É possível usar uma métrica personalizada (como utilização de memória) para sua política de ajuste de escala, se necessário. Para escolher as métricas certas, considere a seguinte orientação para o Amazon EC2:
  - A métrica deve ser uma métrica de utilização válida e descrever o quanto uma instância está ocupada.
  - O valor da métrica deve aumentar e diminuir em proporção ao número das instâncias no grupo do Auto Scaling.
- Use o ajuste de escala dinâmico em vez do ajuste de escala manual para seu grupo do Auto Scaling. Também recomendamos usar políticas de ajuste de escala de rastreamento de metas em seu ajuste de escala dinâmico.
- Verifique se as implantações da workload podem lidar com eventos de aumento e redução horizontal da escala. Crie cenários de teste para eventos de redução horizontal da escala para verificar se a workload se comporta conforme o esperado e não afeta a experiência do usuário (como perda da sessão persistente). É possível usar o Histórico de atividades para verificar uma atividade de ajuste de escala para um grupo do Auto Scaling.
- Avalie sua workload em relação a padrões previsíveis e, ao antecipar alterações previstas e planejadas na demanda, ajuste a escala proativamente. Com o ajuste de escala preditivo, é possível eliminar a necessidade de superprovisionar a capacidade. Para obter mais informações, consulte Ajuste de escala preditivo com o Amazon EC2 Auto Scaling.

Recursos

Documentos relacionados:

- Conceitos básicos do Amazon EC2 Auto Scaling
- Ajuste de escala preditivo para o EC2 com Machine Learning
- Analisar o comportamento do usuário usando o Amazon OpenSearch Service, o Amazon Kinesis Data Firehose e o Kibana
- O que é o Amazon CloudWatch?
- Monitorar a workload de banco de dados com o Performance Insights no Amazon RDS
- Introdução de suporte nativo para ajuste de escala preditivo com o Amazon EC2 Auto Scaling
- Introdução ao Karpenter: um dimensionador automático de escala de clusters do Kubernetes de código aberto e alta performance
- Mergulho profundo no ajuste de escala automático de clusters do Amazon ECS

Vídeos relacionados:

- AWS re:Invent 2023: Escalar na AWS para seus primeiros 10 milhões de usuários
- AWS re:Invent 2023: Arquitetura sustentável: passado, presente e futuro
- AWS re:Invent 2022: Criar um ambiente de computação eficiente em termos de custo, energia e recursos
- AWS re:Invent 2022: Escalar contêineres de um para milhões de usuários
- AWS re:Invent 2023: Escalar a inferência de FM para centenas de modelos com o Amazon SageMaker
- AWS re:Invent 2023: Aproveitar o poder do Karpenter para escalar, otimizar e atualizar o Kubernetes

Exemplos relacionados:

- Ajuste de escala automático

SUS02-BP02 Alinhar os SLAs às metas de sustentabilidade

Analise e otimize os acordos de serviço (SLA) com base em suas metas de sustentabilidade para minimizar os recursos necessários a fim de oferecer compatibilidade com sua workload enquanto continua a atender às necessidades empresariais.

Práticas comuns que devem ser evitadas:

- SLAs de workload são desconhecidos ou ambíguos.
- Você define seu SLA apenas para disponibilidade e performance.
- Você usa o mesmo padrão de design (como arquitetura multi-AZ) para todas as suas workloads.

Benefícios de implementar esta prática recomendada: alinhar os SLAs às metas de sustentabilidade leva ao uso ideal dos recursos e, ao mesmo tempo, atende às necessidades de negócios.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Os SLAs definem o nível de serviço esperado de uma workload de nuvem, como tempo de resposta, disponibilidade e retenção de dados. Eles influenciam a arquitetura, o uso de recursos e o impacto ambiental de uma workload na nuvem. Em uma cadência regular, analise os SLAs e faça compensações que reduzam significativamente o uso de recursos em troca de reduções aceitáveis em níveis de serviço.

Etapas de implementação

- Entenda as metas de sustentabilidade: identifique metas de sustentabilidade em sua organização, como redução de carbono ou melhoria na utilização de recursos.
- Analise os SLAs: avalie seus SLAs para avaliar se eles atendem aos requisitos da sua empresa. Se estiver superando os SLAs, faça uma análise adicional.
- Entenda as vantagens e desvantagens: entenda as vantagens e desvantagens da complexidade da sua workload (como alto volume de usuários simultâneos), performance (como latência) e impacto na sustentabilidade (como os recursos necessários). Normalmente, priorizar dois dos fatores inviabiliza o terceiro.
- Ajuste os SLAs: ajuste os SLAs fazendo compensações que reduzam significativamente os impactos na sustentabilidade em troca de reduções aceitáveis em níveis de serviço.
  - Sustentabilidade e confiabilidade: workloads altamente disponíveis tendem a consumir mais recursos.
  - Sustentabilidade e performance: usar mais recursos para aumentar a performance pode ter um impacto ambiental maior.
  - Sustentabilidade e segurança: workloads excessivamente seguras podem ter um impacto ambiental maior.

- Defina SLAs de sustentabilidade, se possível: inclua SLAs de sustentabilidade para sua workload. Por exemplo, defina um nível mínimo de utilização como um SLA de sustentabilidade para as instâncias de computação.
- Use padrões de design eficientes: use padrões de design, como microsserviços na AWS, que priorizem funções essenciais aos negócios e permita níveis de serviço mais baixos (como objetivos de tempo de resposta ou de tempo de recuperação) para funções não essenciais.
- Estabeleça comunicação e responsabilidade: compartilhe os SLAs com todas as partes interessadas relevantes, incluindo sua equipe de desenvolvimento e seus clientes. Use relatórios para rastrear e monitorar os SLAs. Atribua responsabilidades para alcançar as metas de sustentabilidade dos SLAs.
- Use incentivos e recompensas: use incentivos e recompensas para alcançar ou superar os SLAs alinhados às metas de sustentabilidade.
- Revise e repita: revise e ajuste regularmente seus SLAs para garantir que estejam alinhados às metas de sustentabilidade e performance em evolução.

Recursos

Documentos relacionados:

- Entender padrões de resiliência e compromissos para arquitetar de forma eficiente na nuvem
- Importância do acordo de nível de serviço para provedores de SaaS

Vídeos relacionados:

- AWS re:Invent 2023: Capacidade, disponibilidade, eficiência de custos: escolha três
- AWS re:Invent 2023: Arquitetura sustentável: passado, presente e futuro
- AWS re:Invent 2023: padrões de integração avançados e compromissos para sistemas com acoplamento fraco
- AWS re:Invent 2022: Como entregar arquiteturas sustentáveis e de alta performance
- AWS re:Invent 2022: Criar um ambiente de computação eficiente em termos de custo, energia e recursos

SUS02-BP03 Interromper a criação e a manutenção de ativos não utilizados

Desative ativos em sua workload para reduzir o número de recursos necessários para atender à sua demanda e minimizar o desperdício.

Práticas comuns que devem ser evitadas:

- Você não analisa sua aplicação com relação a ativos redundantes ou não mais necessários.
- Você não remove ativos redundantes ou que não mais necessários.

Benefícios de implementar esta prática recomendada: a remoção de ativos não utilizados libera recursos e melhora a eficiência geral da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Os ativos ociosos consomem recursos de nuvem como espaço de armazenamento e potência computacional. Com a identificação e eliminação desses ativos, você pode liberar esses recursos e aumentar a eficiência da arquitetura de nuvem. Analise regularmente os ativos de aplicações (como relatórios pré-compilados, conjuntos de dados e imagens estáticas) e os padrões de acesso aos ativos para identificar redundâncias, subutilização e possíveis alvos de desativação. Remova esses ativos redundantes para diminuir o desperdício de recursos em sua workload.

Etapas de implementação

- Faça um inventário: faça um inventário abrangente para identificar todos os ativos em sua workload.
- Analise o uso: use ferramentas de monitoramento para identificar ativos estáticos que não são mais necessários.
- Remova ativos não utilizados: desenvolva um plano para remover ativos que não são mais necessários.
  - Antes de remover qualquer ativo, avalie o impacto da remoção sobre a arquitetura.
  - Consolide ativos gerados sobrepostos para remover o processamento redundante.
  - Atualize suas aplicações para que não produzam nem armazenem mais ativos que não são necessários.

- Comunique-se com terceiros: instrua terceiros a interromper a produção e o armazenamento de ativos gerenciados em seu nome que não sejam mais necessários. Peça que consolidem ativos redundantes.
- Use políticas de ciclo de vida: use políticas de ciclo de vida para excluir automaticamente ativos não utilizados.
  - Você pode usar o Amazon S3 Lifecycle para gerenciar seus objetos durante todo o ciclo de vida de cada um.
  - É possível usar o Amazon Data Lifecycle Manager para automatizar a criação, a retenção e a exclusão de snapshots do Amazon EBS e de AMIs apoiadas pelo Amazon EBS.
- Revise e otimize: revise regularmente sua workload para identificar e remover ativos ociosos.

Recursos

Documentos relacionados:

- Otimizar a sua infraestrutura da AWS para sustentabilidade, Parte II: Armazenamento
- Como faço para encerrar recursos ativos dos quais não preciso mais em minha Conta da AWS?

Vídeos relacionados:

- AWS re:Invent 2023: Arquitetura sustentável: passado, presente e futuro
- AWS re:Invent 2022: Preservar e maximizar o valor de ativos de mídia digital usando o Amazon S3
- AWS re:Invent 2023: Otimizar custos em seus ambientes com várias contas

SUS02-BP04 Otimizar o posicionamento geográfico das workloads com base nos respectivos requisitos de rede

Selecione locais e serviços de nuvem para sua workload que reduzam a distância que o tráfego de rede deve percorrer e diminua o total de recursos de rede necessários para comportar a workload.

Práticas comuns que devem ser evitadas:

- Selecione a região da workload com base em sua localização.
- Consolidar todos os recursos da workload em uma única localização geográfica.
- Todo o tráfego flui por meio dos data centers existentes.

Benefícios de implementar esta prática recomendada: implantar uma workload perto dos clientes proporciona a latência mais baixa enquanto reduz a movimentação de dados pela rede e reduz o impacto ambiental.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A infraestrutura da Nuvem AWS é construída em torno de opções de local, como regiões, zonas de disponibilidade, grupos de posicionamento e locais da borda, como AWS Outposts e zonas locais da AWS. Essas opções de local são responsáveis por manter a conectividade entre componentes da aplicação, serviços de nuvem, redes da borda e data centers on-premises.

Analise os padrões de acesso à rede em sua workload para identificar como usar essas opções de local de nuvem e reduzir a distância que o tráfego de rede precisa percorrer.

Etapas de implementação

- Analise os padrões de acesso à rede em sua workload para identificar como os usuários utilizam sua aplicação.
  - Use ferramentas de monitoramento, como o Amazon CloudWatch e o AWS CloudTrail, para coletar dados sobre atividades de rede.
  - Analise os dados para identificar o padrão de acesso à rede.
- Selecione as regiões para implantação da workload com base nos seguintes elementos fundamentais:
  - Sua meta de sustentabilidade: conforme explicado em Seleção de regiões.
  - Onde seus dados estão localizados: para aplicações com uso intenso de dados (como big data e machine learning), o código da aplicação deve ser executado o mais perto possível dos dados.
  - Onde seus usuários estão localizados: para aplicações voltadas ao usuário, escolha uma ou mais regiões perto dos clientes da workload.
  - Outras restrições: considere restrições como custo e conformidade, conforme explicado em O que considerar ao selecionar uma região para suas workloads.
- Use o armazenamento em cache local ou soluções de armazenamento em cache da AWS para dados usados com frequência a fim de aumentar a performance, reduzir a movimentação de dados e diminuir o impacto ambiental.

Serviço	Quando usar
Amazon CloudFront	Use para armazenar conteúdo estático em cache, como imagens, scripts e vídeos, além de conteúdo dinâmico como respostas de API ou aplicações Web.
Amazon ElastiCache	Use para armazenar conteúdo em cache para aplicações Web.
DynamoDB Accelerator	Use para adicionar aceleração na memória às suas tabelas do DynamoDB.

- Use serviços que podem ajudar você a executar código mais perto dos usuários da workload:

Serviço	Quando usar
Lambda@Edge	Use para operações com uso intenso de computação que são iniciadas quando objetos não estão no cache.
Amazon CloudFront Functions	Use para casos de uso simples, como solicitações HTTP(s) ou manipulações de resposta que podem ser iniciadas por funções de curta duração.
AWS IoT Greengrass	Use para executar computação local, mensagens e armazenamento de dados em cache para dispositivos conectados.

- Use o agrupamento de conexões para permitir a reutilização de conexões e reduzir os recursos necessários.
- Use datastores distribuídos que não dependem de conexões persistentes e atualizações síncronas para fins de consistência com o objetivo de atender a populações regionais.
- Substitua a capacidade de rede estática pré-provisionada por capacidade dinâmica compartilhada e divida o impacto sobre a sustentabilidade da capacidade de rede com outros assinantes.

Recursos

Documentos relacionados:

- Otimizar a sua infraestrutura da AWS para sustentabilidade, Parte III: Redes
- Documentação do Amazon ElastiCache
- O que é o Amazon CloudFront?
- Principais recursos do Amazon CloudFront
- Infraestrutura global da AWS
- Zonas locais da AWS e AWS Outposts: como escolher a tecnologia certa para sua workload de borda)
- Grupos de posicionamento
- Zonas locais da AWS
- AWS Outposts

Vídeos relacionados:

- Desmistificar a transferência de dados na AWS
- Escalar a performance da rede em instâncias do Amazon EC2 de última geração
- Vídeo explicativo de zonas locais da AWS
- Visão geral do AWS Outposts e como ele funciona
- AWS re:Invent 2023: Uma estratégia de migração para workloads periféricas e on-premises
- AWS re:Invent 2021: AWS Outposts: como trazer a experiência da AWS para ambientes on-premises
- AWS re:Invent 2020: AWS Wavelength:executar aplicações com latência ultrabaixa na borda 5G
- AWS re:Invent 2022: Zonas locais da AWS: como criar aplicações para uma borda distribuída
- AWS re:Invent 2021: Criar sites de baixa latência com o Amazon CloudFront
- AWS re:Invent 2022: Aprimorar a performance e a disponibilidade com o AWS Global Accelerator
- AWS re:Invent 2022: Criar sua rede de longa distância usando a AWS
- AWS re:Invent 2020: Gerenciamento de tráfego global com o Amazon Route 53

Exemplos relacionados:

- Workshops de redes da AWS
- Arquitetura para a sustentabilidade: reduza a movimentação de dados entre redes

SUS02-BP05 Otimizar os recursos dos membros da equipe para as atividades realizadas

Otimize os recursos fornecidos aos membros da equipe para minimizar o impacto sobre a sustentabilidade ambiental e, ao mesmo tempo, atender às suas necessidades.

Práticas comuns que devem ser evitadas:

- Você ignora o impacto dos dispositivos usados pelos membros da equipe sobre a eficiência geral de sua aplicação de nuvem.
- Você gerencia e atualiza manualmente os recursos usados pelos membros da equipe.

Benefícios de implementar esta prática recomendada: a otimização dos recursos dos membros da equipe melhora a eficiência geral das aplicações habilitadas para a nuvem.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Conheça os recursos que os membros da equipe usam para consumir seus serviços, o ciclo de vida esperado e o impacto financeiro e na sustentabilidade. Implemente estratégias para otimizar esses recursos. Por exemplo, realize operações complexas, como renderização e compilação, em infraestrutura escalável com alta utilização em vez de em sistemas de usuário único subutilizados com alto consumo de energia.

Etapas de implementação

- Use estações de trabalho energeticamente eficientes: forneça aos membros da equipe estações de trabalho e periféricos com baixo consumo de energia. Use recursos eficientes para gerenciamento de energia (como o modo de baixo consumo de energia) nesses dispositivos para reduzir o uso de energia
- Use virtualização: use áreas de trabalho virtuais e a transmissão de aplicações para limitar os requisitos de upgrade e dispositivos.
- Incentive a colaboração remota: incentive os membros da equipe a usar ferramentas de colaboração remota como o Amazon Chime ou o AWS Wickr para reduzir a necessidade de viagens e as emissões de carbono associadas.

- Use software energeticamente eficiente: forneça aos membros da equipe software de baixo consumo de energia removendo ou desativando recursos e processos desnecessários.
- Gerencie os ciclos de vida: avalie o impacto de processos e sistemas no ciclo de vida de seus dispositivos e escolha soluções que minimizem o requisito de substituição de dispositivos e, ao mesmo tempo, atendam aos requisitos comerciais. Mantenha e atualize regularmente estações de trabalho e o software para manter e melhorar a eficiência.
- Gerenciamento remoto de dispositivos: implemente o gerenciamento remoto de dispositivos para reduzir as viagens de negócios.
  - O AWS Systems Manager Fleet Manager é uma experiência de interface do usuário (IU) unificada que ajuda você a gerenciar remotamente os nós em execução no AWS ou on-premises.

Recursos

Documentos relacionados:

- O que é o Amazon WorkSpaces?
- Otimizador de custos para Amazon WorkSpaces
- Documentação do Amazon AppStream 2.0
- NICE DCV

Vídeos relacionados:

- Gerenciar os custos do Amazon WorkSpaces na AWS

SUS02-BP06 Implementar armazenamento em buffer ou controle de utilização para nivelar a curva da demanda

O armazenamento em buffer e o controle de utilização nivelam a curva da demanda e reduzem a capacidade provisionada necessária para sua workload.

Práticas comuns que devem ser evitadas:

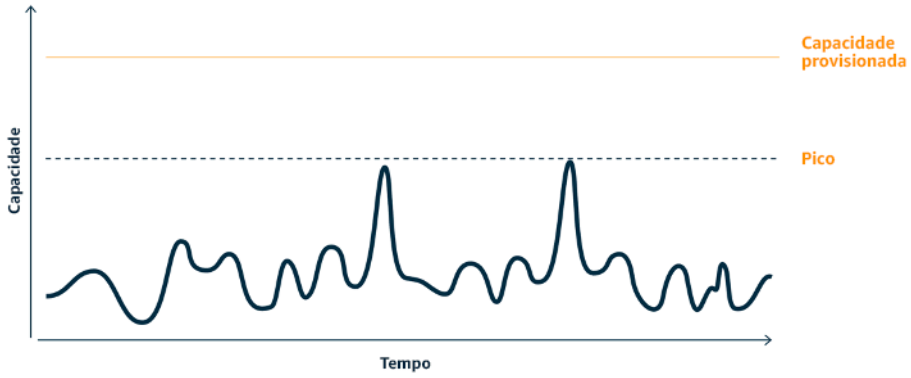
- Você processa imediatamente as solicitações de cliente embora isso não seja necessário.
- Você não analisa os requisitos das solicitações de cliente.

Benefícios de implementar esta prática recomendada: nivelar a curva de demanda reduz a capacidade provisionada necessária para a workload. Reduzir a capacidade provisionada significa diminuir o consumo de energia e o impacto ambiental.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

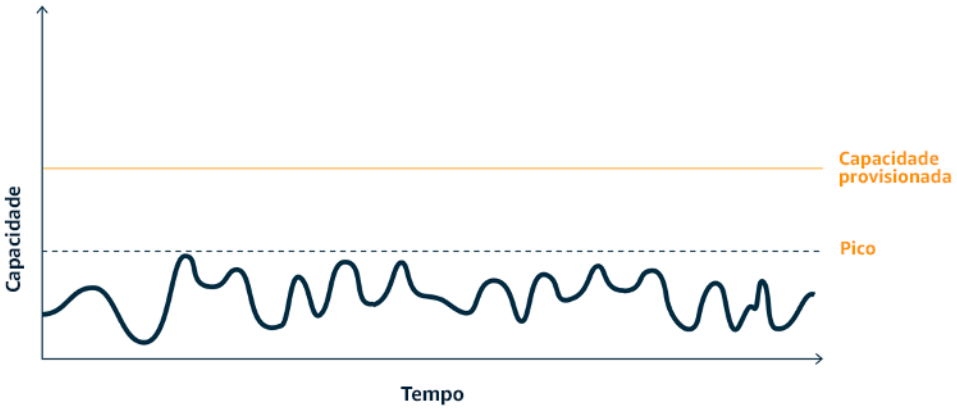
Orientação para implementação

Nivelar a curva da demanda pode ajudar você a reduzir a capacidade provisionada para uma workload e a diminuir o respectivo impacto ambiental. Considere a workload com a curva de demanda mostrada na figura a seguir. Essa workload tem dois picos e, para lidar com eles, é provisionada a capacidade de recurso mostrada pela linha laranja. Os recursos e energia usados para essa workload não são indicados pela área abaixo da curva da demanda, mas pela área abaixo da linha da capacidade provisionada, visto que é preciso ter capacidade provisionada para lidar com esses dois picos.



Curva de demanda com dois picos distintos que exigem alta capacidade provisionada.

Você pode usar o armazenamento em buffer ou o controle de utilização para modificar a curva da demanda e atenuar os picos, o que significa menor capacidade provisionada e menor consumo de energia. Implemente o controle de utilização quando seus clientes puderem realizar novas tentativas. Implemente o armazenamento em buffer para armazenar a solicitação e adiar o processamento até um momento posterior.



O efeito do controle de utilização na curva de demanda e na capacidade provisionada.

Etapas de implementação

- Analise as solicitações dos clientes para determinar como responder a elas. As perguntas a serem consideradas incluem:
  - Essa solicitação pode ser processada assincronamente?
  - O cliente tem capacidade de repetição?
- Se o cliente tiver capacidade de repetição, você poderá implementar o controle de utilização, que informa à origem que, se ela não puder atender à solicitação naquele momento, deverá tentar novamente mais tarde.
  - Você pode usar o Amazon API Gateway para implementar o controle de utilização.
- Para clientes que não podem realizar novas tentativas, é necessário implementar um buffer para nivelar a curva da demanda. O buffer adia o processamento de solicitações, permitindo que as aplicações executadas em diferentes taxas se comuniquem com eficácia. Uma abordagem baseada em buffer usa uma fila ou um fluxo para aceitar mensagens de produtores. As mensagens são lidas pelos consumidores e processadas, permitindo que as mensagens sejam executadas na taxa que atenda aos requisitos de negócios dos consumidores.
  - O Amazon Simple Queue Service (Amazon SQS) é um serviço gerenciado que fornece filas que permitem que um único consumidor leia mensagens individuais.
  - O Amazon Kinesis oferece um fluxo que permite que muitos consumidores leiam as mesmas mensagens.
- Analise a demanda geral, a taxa de alteração e o tempo de resposta necessário para dimensionar adequadamente o controle ou buffer necessário.

Recursos

Documentos relacionados:

- Conceitos básicos do Amazon SQS
- Integrar aplicações usando filas e mensagens
- Gerenciar e monitorar o controle de utilização de APIs em suas workloads
- Controlar a utilização de uma API REST em camadas e multilocatário em grande escala usando o API Gateway
- Integrar aplicações usando filas e mensagens

Vídeos relacionados:

- AWS re:Invent 2022: padrões de integração de aplicações para microsserviços
- AWS re:Invent 2023: economias inteligentes: estratégias de otimização de custos no Amazon EC2
- AWS re:Invent 2023: padrões de integração avançados e compromissos para sistemas com acoplamento fraco

Software e arquitetura

Pergunta

- SUS 3 Como aproveitar os padrões de software e arquitetura para apoiar as metas de sustentabilidade?

SUS 3 Como aproveitar os padrões de software e arquitetura para apoiar as metas de sustentabilidade?

Implemente padrões que suavizem os picos de carga e mantenham a alta utilização consistente de recursos implantados para minimizar os recursos consumidos. Os componentes podem ficar ociosos devido à falta de uso por conta das mudanças no comportamento do usuário ao longo do tempo. Revise os padrões e a arquitetura para consolidar os componentes subutilizados a fim de aumentar a utilização geral. Retire os componentes que não são mais necessários. Saiba qual é a performance dos componentes de sua workload e otimize os componentes que consomem a maioria dos recursos. Esteja ciente dos dispositivos que seus clientes usam para acessar seus serviços e implemente padrões a fim de minimizar a necessidade de upgrades de dispositivos.

Práticas recomendadas

- SUS03-BP01 Otimizar o software e a arquitetura para trabalhos assíncronos e agendados
- SUS03-BP02 Remover ou refatorar componentes da workload com pouco ou nenhum uso
- SUS03-BP03 Otimizar áreas de código que consomem mais tempo ou recursos
- SUS03-BP04 Otimizar o impacto sobre dispositivos e equipamentos
- SUS03-BP05 Usar padrões e arquiteturas de software que atendam melhor aos padrões de armazenamento e acesso a dados

SUS03-BP01 Otimizar o software e a arquitetura para trabalhos assíncronos e agendados

Use software eficiente e padrões de arquitetura, como orientado a filas, para manter uma alta e consistente utilização dos recursos implantados.

Práticas comuns que devem ser evitadas:

- Provisione em excesso os recursos em sua workload na nuvem para atender a picos imprevistos na demanda.
- Sua arquitetura não separa remetentes e destinatários de mensagens assíncronas por um componente de sistema de mensagens.

Benefícios de implementar esta prática recomendada:

- Padrões eficientes de software e arquitetura minimizam os recursos não utilizados em sua workload e melhoram a eficiência geral.
- Você pode dimensionar o processamento independentemente do recebimento de mensagens assíncronas.
- Por meio de um componente de mensagens, você relaxou os requisitos de disponibilidade que podem ser atendidos com menos recursos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Use padrões de arquitetura eficientes, como arquitetura orientada por eventos, que resultam na utilização uniforme dos componentes e minimizam o excesso de provisionamento em sua workload.

A utilização de padrões de arquitetura eficientes minimiza recursos ociosos por falta de uso devido a mudanças na demanda ao longo do tempo.

Entenda os requisitos de seus componentes de workload e adote padrões de arquitetura que aumentam a utilização geral dos recursos. Retire os componentes que não são mais necessários.

Etapas de implementação

- Analise a demanda da sua workload para determinar como responder a ela.
- Para solicitações ou trabalhos que não exigem respostas síncronas, use arquiteturas orientadas por filas e operadores de escalação automática para maximizar a utilização. Aqui estão alguns exemplos de quando você pode considerar a arquitetura orientada por filas:

Mecanismo de filas	Descrição
Filas de trabalhos do AWS Batch	Os trabalhos do AWS Batch são enviados para uma fila de trabalhos onde permanece m até que possam ser agendados para execução em um ambiente de computação.
Amazon Simple Queue Service e instâncias spot do Amazon EC2	Emparelhar o Amazon SQS e instâncias spot para criar uma arquitetura eficiente e tolerante a falhas.

- Para solicitações ou trabalhos que podem ser processados a qualquer momento, use mecanismos de agendamento para processar trabalhos em lote para maior eficiência. Aqui estão alguns exemplos de mecanismos de agendamento no AWS:

Mecanismos de agendamento	Descrição
Agendador do Amazon EventBridge	Um recurso do Amazon EventBridge que permite criar, executar e gerenciar tarefas agendadas em grande escala.
Agendamento do AWS Glue baseado em tempo	Defina um agendamento baseada em tempo para seus crawlers e trabalhos no AWS Glue.
Tarefas agendadas do Amazon Elastic Container Service (Amazon ECS)	O Amazon ECS oferece suporte à criação de tarefas agendadas. As tarefas programadas

Mecanismos de agendamento	Descrição
	usam as regras do Amazon EventBridge para executar tarefas em uma programação ou em uma resposta a um evento do EventBridge.
Agendador de instâncias	Configure agendamentos de início e término para suas instâncias do Amazon EC2 e do Amazon Relational Database Service.

- Se você usa mecanismos de pesquisa e webhooks em sua arquitetura, substitua-os por eventos. Use arquiteturas orientadas por eventos para criar workloads altamente eficientes.
- Aproveite a tecnologia sem servidor na AWS para eliminar a infraestrutura provisionada em excesso.
- Dimensione corretamente componentes individuais da sua arquitetura para evitar recursos ociosos aguardando entrada.
  - As Recomendações de dimensionamento correto em no AWS Cost Explorer ou o AWS Compute Optimizer podem ser usados para identificar oportunidades de dimensionamento correto.
  - Para obter mais detalhes, consulte Dimensionamento correto: provisionamento de instâncias para corresponder a workloads.

Recursos

Documentos relacionados:

- O que é o Amazon Simple Queue Service?
- O que é o Amazon MQ?
- Ajuste de escala baseado no Amazon SQS
- O que é AWS Step Functions?
- O que é AWS Lambda?
- Usar o AWS Lambda com o Amazon SQS
- O que é o Amazon EventBridge?
- Gerenciara fluxos de trabalho assíncronos com uma API REST

Vídeos relacionados:

- AWS re:Invent 2023: Navegar na jornada rumo à arquitetura orientada a eventos sem servidor
- AWS re:Invent 2023: Usar a tecnologia sem servidor para arquitetura orientada a eventos e design orientado por domínio
- AWS re:Invent 2023: Padrões avançados orientados a eventos com o Amazon EventBridge
- AWS re:Invent 2023: Arquitetura sustentável: passado, presente e futuro
- Padrões de mensagens assíncronas | Eventos da AWS

Exemplos relacionados:

- Arquitetura orientada a eventos com processadores AWS Graviton e instâncias spot do Amazon EC2

SUS03-BP02 Remover ou refatorar componentes da workload com pouco ou nenhum uso

Remova os componentes que não são mais utilizados nem necessários e refatore os componentes pouco usados para minimizar o desperdício em sua workload.

Práticas comuns que devem ser evitadas:

- Você não verifica regularmente o nível de utilização de componentes individuais da sua workload.
- Você não verifica as recomendações de ferramentas de dimensionamento correta da AWS, como o AWS Compute Optimizer.

Benefícios de implementar esta prática recomendada: a remoção de componentes não utilizados minimiza o desperdício e melhora a eficiência geral da sua workload na nuvem.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Avalie sua workload para identificar componentes ociosos ou não utilizados. Esse é um processo de melhoria interativo que pode ser iniciado por alterações na demanda ou pelo lançamento de um novo serviço de nuvem. Por exemplo, uma queda significativa no tempo de execução da função do AWS Lambda pode ser um indicador da necessidade de reduzir o tamanho da memória. Além disso, à medida que a AWS lança novos serviços e recursos, a arquitetura e os serviços ideais para sua workload podem mudar.

Monitore continuamente a atividade da workload e procure oportunidades para melhorar o nível de utilização de componentes individuais. Com a remoção de componentes ociosos e a execução de atividades de dimensionamento correto, você atende aos seus requisitos empresariais com menos recursos de nuvem.

Etapas de implementação

- Faça um inventário dos seus recursos da AWS. Na AWS, é possível ativar o Explorador de recursos da AWS para explorar e organizar seus recursos da AWS. Para obter mais detalhes, consulte AWS re:Invent 2022: Como gerenciar recursos e aplicações em grande escala AWS.
- Monitore e capture as métricas de utilização de componentes críticos de sua workload (como utilização de CPU, utilização de memória ou throughput de rede nas métricas do Amazon CloudWatch).
- Identifique componentes não utilizados ou subutilizados em sua arquitetura.
  - Para workloads estáveis, verifique as ferramentas de dimensionamento correto da AWS, como o AWS Compute Optimizer, em intervalos regulares para identificar componentes ociosos, não utilizados ou subutilizados.
  - Para workloads efêmeras, avalie as métricas de utilização para identificar componentes ociosos, não usados ou subutilizados.
- Retire componentes e ativos associados (como imagens do Amazon ECR) que não são mais necessários.
  - Limpeza automatizada de imagens não utilizadas no Amazon ECR
  - Excluir volumes não utilizados do Amazon Elastic Block Store (Amazon EBS) usando AWS Config e AWS Systems Manager
- Refatore ou consolide os componentes subutilizados com outros recursos para melhorar a eficiência da utilização. Por exemplo, é possível provisionar vários bancos de dados pequenos em uma única instância de banco de dados do Amazon RDS em vez de executar bancos de dados em instâncias individuais subutilizadas.
- Entenda os recursos provisionados pela sua workload para concluir uma unidade de trabalho.

Recursos

Documentos relacionados:

- AWS Trusted Advisor
- O que é o Amazon CloudWatch?

- Dimensionamento correto: provisionamento de instâncias para corresponder às workloads
- Como otimizar seus custos com as recomendações de redimensionamento

Vídeos relacionados:

- AWS re:Invent 2023: Capacidade, disponibilidade, eficiência de custos: escolha três

Exemplos relacionados:

- Otimizar padrões de hardware e observar KPIs de sustentabilidade

SUS03-BP03 Otimizar áreas de código que consomem mais tempo ou recursos

Otimize o código que é executado em diferentes componentes de sua arquitetura para minimizar o uso de recursos e, ao mesmo tempo, maximizar a performance.

Práticas comuns que devem ser evitadas:

- Você ignora a otimização de seu código para uso de recursos.
- Normalmente, você responde a problemas de performance aumentando os recursos.
- Seu processo de revisão e desenvolvimento de código não monitora alterações na performance.

Benefícios de implementar esta prática recomendada: o uso eficiente de código minimiza o uso de recursos e melhora a performance.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

É essencial examinar toda área funcional, incluindo o código referente a uma aplicação projetada para a nuvem, para otimizar o uso de recursos e a performance. Monitore continuamente a performance da workload em ambientes de compilação e na produção e identifique oportunidades para melhorar os trechos cujo uso de recursos é particularmente alto. Adote um processo de revisão regular para identificar erros ou antipadrões dentro do código que usa os recursos ineficazmente. Utilize algoritmos simples e eficientes que produzem os mesmos resultados para seu caso de uso.

Etapas de implementação

- Use uma linguagem de programação eficiente: use um sistema operacional e uma linguagem de programação eficientes para a workload. Para obter detalhes sobre linguagens de programação com eficiência energética (incluindo Rust), consulte Sustentabilidade com o Rust.
- Use um companheiro de codificação de IA: considere usar um companheiro de codificação de IA, como o Amazon CodeWhisperer, para escrever código com eficiência.
- Automatize as revisões de código: ao desenvolver suas workloads, adote um processo de revisão de código automatizado para melhorar a qualidade e identificar erros e práticas não recomendadas.
  - Automatize as revisões de código com o Amazon CodeGuru Reviewer
  - Detectar bugs de simultaneidade com o Amazon CodeGuru
  - Melhorar a qualidade do código para aplicações Python com o Amazon CodeGuru
- Use um criador de perfil de código: use um criador de perfil de código para identificar as áreas de código que gastam mais tempo ou usam mais recursos e as defina como alvos de otimização.
  - Reduzir a pegada de carbono de sua organização com o Amazon CodeGuru Profiler
  - Conceitos básicos do uso de memória em sua aplicação Java com o Amazon CodeGuru Profiler
  - Melhorar a experiência do cliente e reduzir os custos com o Amazon CodeGuru Profiler
- Monitore e otimize: use recursos de monitoramento contínuo para identificar componentes com altos requisitos de recursos ou configuração abaixo do ideal.
  - Substitua os algoritmos com uso intenso de computação por uma versão mais simples e mais eficiente que produza o mesmo resultado.
  - Remova códigos desnecessários, como classificações e formatações.
- Use refatoração ou transformação de código: explore a possibilidade da transformação de código do Amazon Q para manutenção e atualizações de aplicações.
  - Atualizar as versões de idioma com o Amazon Q Code Transformation
  - AWS re:Invent 2023: Automatizar as atualizações e a manutenção de aplicações usando o Amazon Q Code Transformation

Recursos

Documentos relacionados:

- O que é o Amazon CodeGuru Profiler?

- Instâncias FPGA
- Os AWS SDKs em Ferramentas para desenvolver na AWS

Vídeos relacionados:

- Melhorar a eficiência do código usando o Amazon CodeGuru Profiler
- AWS re:Invent 2023: Práticas recomendadas para o Amazon CodeWhisperer
- Automatizar revisões de código e recomendações de performance de aplicações com o Amazon CodeGuru

Exemplos relacionados:

- Otimizar código com o Amazon CodeGuru

SUS03-BP04 Otimizar o impacto sobre dispositivos e equipamentos

Conheça os dispositivos e equipamentos usados em sua arquitetura e use estratégias para reduzir o respectivo uso. Isso pode minimizar o impacto ambiental de modo geral de sua workload na nuvem.

Práticas comuns que devem ser evitadas:

- Você ignora o impacto ambiental dos dispositivos usados por seus clientes.
- Você gerencia e atualiza manualmente os recursos usados pelos clientes.

Benefícios de implementar esta prática recomendada: a implementação de padrões e recursos de software otimizados para o dispositivo do cliente pode reduzir o impacto ambiental geral da workload na nuvem.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Implementar padrões e recursos de software que são otimizados para os dispositivos do clientes pode reduzir o impacto ambiental de variadas maneiras:

- Implementar novos recursos que são compatíveis com versões anteriores pode reduzir o número de substituições de hardware.

- Otimizar uma aplicação para ser executada com eficiência nos dispositivos pode ajudar a reduzir o consumo de energia e a estender a duração da bateria (se eles forem alimentados por bateria).
- Otimizar uma aplicação para dispositivos também pode reduzir a transferência de dados ao longo da rede.

Conheça os dispositivos e equipamentos usados em sua arquitetura, o ciclo de vida esperado e o impacto da substituição desses componentes. Implemente padrões e recursos de software que possam ajudar a minimizar o consumo de energia do dispositivo, bem como a necessidade de os clientes substituírem o dispositivo e também atualizá-lo manualmente.

Etapas de implementação

- Conduza um inventário: faça um inventário dos dispositivos usados em sua arquitetura. Os dispositivos podem ser celular, tablet, dispositivos IoT, lâmpada inteligente ou até dispositivos inteligentes em uma fábrica.
- Use dispositivos energeticamente eficientes: considere usar dispositivos com baixo consumo de energia em sua arquitetura. Use as configurações de gerenciamento de energia nos dispositivos para que entrem no modo de baixo consumo quando não estiverem em uso.
- Execute aplicações eficientes: otimize a aplicação em execução nos dispositivos:
  - Use estratégias como execução de tarefas em segundo plano para reduzir o consumo de energia.
  - Considere a largura de banda da rede e a latência ao criar cargas úteis, e implemente recursos que ajudem suas aplicações a funcionar bem em links de baixa largura de banda e alta latência.
  - Converta cargas úteis e arquivos nos formatos otimizados exigidos pelos dispositivos. Por exemplo, é possível usar o Amazon Elastic Transcoder ou o AWS Elemental MediaConvert para converter arquivos de mídia digital grande e de alta qualidade em formatos que podem ser reproduzidos em dispositivos móveis, tablets, navegadores da web e televisões conectadas.
  - Realize atividades com computação intensa no lado do servidor (como renderização de imagens) ou use a transmissão de aplicações para melhorar a experiência do usuário em dispositivos mais antigos.
  - Faça a segmentação e a paginação dos dados de saída, especialmente para sessões interativas, a fim de gerenciar cargas úteis e limitar os requisitos de armazenamento local.
- Envolva fornecedores: trabalhe com fornecedores de dispositivos que usam materiais sustentáveis e fornecem transparência em suas cadeias de suprimentos e certificações ambientais.

- Use atualizações sem fios: use um mecanismo sem fio automatizado para implantar atualizações em um ou mais dispositivos.
  - É possível usar um pipeline de CI/CD para atualizar aplicações móveis.
  - O AWS IoT Device Management também pode ser usado para gerenciar remotamente dispositivos conectados em grande escala.
- Use parques de dispositivos gerenciados: para testar novos recursos e atualizações, use parques de dispositivos gerenciados com conjuntos representativos de hardware e itere o desenvolvimento para maximizar os dispositivos compatíveis. Para obter mais detalhes, consulte SUS06-BP04 Usar parques de dispositivos gerenciados para testes.
- Continue monitorando e melhorando: acompanhe o uso de energia dos dispositivos para identificar áreas de melhoria. Use novas tecnologias ou práticas recomendadas para melhorar os impactos ambientais desses dispositivos.

Recursos

Documentos relacionados:

- O que é AWS Device Farm?
- Documentação do Amazon AppStream 2.0
- NICE DCV
- Tutorial OTA para atualização de firmware em dispositivos que executam o FreeRTOS
- Otimizando seus dispositivos de IoT para sustentabilidade ambiental

Vídeos relacionados:

- AWS re:Invent 2023: Melhorar a qualidade das suas aplicações móveis e Web com o AWS Device Farm

SUS03-BP05 Usar padrões e arquiteturas de software que atendam melhor aos padrões de armazenamento e acesso a dados

Entenda como os dados são usados com sua workload, consumidos pelos usuários, transferidos e armazenados. Use os padrões e arquiteturas de software ideais para acesso e armazenamento de dados a fim de minimizar os recursos de computação, rede e armazenamento necessários para atender à workload.

Práticas comuns que devem ser evitadas:

- Você pressupõe que todas as workloads temja, padrões de acesso e armazenamento de dados semelhantes.
- Você usa apenas um nível de armazenamento, supondo que todas as workloads se encaixem nesse nível.
- Você pressupõe que os padrões de acesso aos dados permanecerão consistentes ao longo do tempo.
- Na eventualidade de uma alta expansão no acesso aos dados, sua arquitetura é capaz de comportá-la, mas isso faz com que os recursos fiquem ociosos na maior parte do tempo.

Benefícios de implementar esta prática recomendada: selecionar e otimizar sua arquitetura com base nos padrões de acesso e armazenamento de dados ajudará a diminuir a complexidade do desenvolvimento e aumentar a utilização geral. Compreender quando usar tabelas globais, provisionamento de dados e armazenamento em cache ajuda a reduzir a despesas operacionais indiretas e a escalar com base nas necessidades da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Use padrões de software e arquitetura que melhor se alinhem às características dos dados e aos padrões de acesso. Por exemplo, use uma arquitetura de dados moderna na AWS que permita que você use serviços personalizados e otimizados para seus casos de uso de análise exclusivos. Esses padrões de arquitetura possibilitam um processamento de dados eficiente e reduzem o uso de recursos.

Etapas de implementação

- Analise as características dos dados e os padrões de acesso para identificar a configuração correta para seus recursos de nuvem. Principais características a serem consideradas:
  - Tipos de dados: estruturados, semiestruturados e não estruturados
  - Crescimento de dados: limitado, ilimitado
  - Durabilidade dos dados: persistentes, efêmeros, transitórios
  - Padrões de acesso: leituras ou gravações, frequência, com picos ou consistente
- Use padrões de arquitetura que comportem melhor os padrões de armazenamento e acesso aos dados.

  - Padrões para permitir a persistência de dados
  - Vamos arquitetar! Arquiteturas de dados modernas
  - Bancos de dados na AWS: a ferramenta certa para o trabalho certo
- Use tecnologias que funcionam nativamente com dados compactados.
  - Formatos de arquivos compactados compatíveis com o Athena
  - Opções de formato para entradas e saídas de ETL no AWS Glue
  - Carregar arquivos de dados compactados do Amazon S3 com o Amazon Redshift
- Use serviços de análise específicos para processamento de dados em sua arquitetura. Para obter detalhes sobre serviços de análise da AWS com propósitos específicos, consulte AWS re:Invent 2022: construir arquiteturas de dados modernas na AWS.
- Use o mecanismo de banco de dados que melhor comporta seu padrão de consulta dominante. Gerencie seus índices de bancos de dados para garantir consultas eficientes. Para obter mais detalhes, consulte Bancos de dados da AWS e AWS re:Invent 2022: Modernizar aplicações com bancos de dados com propósitos específicos.
- Escolha protocolos de rede que reduzam a quantidade de capacidade de rede consumida em sua arquitetura.

Recursos

Documentos relacionados:

- COPY de formatos de dados colunares com o Amazon Redshift
- Converter formato de registros recebidos no Firehose
- Melhorar a performance de consultas no Amazon Athena fazendo a conversão para formatos colunares
- Monitorar a workload de banco de dados com o Performance Insights no Amazon Aurora
- Monitorar a workload de banco de dados com o Performance Insights no Amazon RDS
- Classe de armazenamento do Amazon S3 Intelligent-Tiering
- Criar uma loja de eventos CQRS com o Amazon DynamoDB

Vídeos relacionados:

- AWS re:Invent 2022: Construir arquiteturas de data mesh na AWS
- AWS re:Invent 2023: Mergulho profundo no Amazon Aurora e suas inovações

- AWS re:Invent 2023: Melhorar a eficiência do Amazon EBS e ser mais econômico
- AWS re:Invent 2023: Otimizar o preço e a performance do armazenamento com o Amazon S3
- AWS re:Invent 2023: Criar e otimizar data lakes no Amazon S3
- AWS re:Invent 2023: Padrões avançados orientados a eventos com o Amazon EventBridge

Exemplos relacionados:

- Workshop de bancos de dados com propósito específico na AWS
- Dia de imersão na arquitetura de dados moderna na AWS
- Criar um data mesh na AWS

Dados

Pergunta

- SUS 4 Como aproveitar as políticas e os padrões de gerenciamento de dados para apoiar as metas de sustentabilidade?

SUS 4 Como aproveitar as políticas e os padrões de gerenciamento de dados para apoiar as metas de sustentabilidade?

Implemente práticas de gerenciamento de dados para reduzir o armazenamento provisionado necessário para comportar a workload e os recursos exigidos para usá-la. Entenda seus dados e use as tecnologias e as configurações de armazenamento que promovam o valor empresarial dos dados de forma mais eficaz e a forma como eles são usados. Gerencie o ciclo de vida dos dados e opte por um armazenamento mais eficiente e com menor performance quando os requisitos diminuírem, excluindo os dados que não são mais necessários.

Práticas recomendadas

- SUS04-BP01 Implementar uma política de classificação de dados
- SUS04-BP02 Usar tecnologias compatíveis com seus padrões de acesso e de armazenamento de dados
- SUS04-BP03 Usar políticas para gerenciar o ciclo de vida de seus conjuntos de dados
- SUS04-BP04 Usar elasticidade e automação para expandir o armazenamento em bloco ou o sistema de arquivos

- SUS04-BP05 Remover dados desnecessários ou redundantes
- SUS04-BP06 Usar armazenamento ou sistemas de arquivos compartilhados para acessar dados comuns
- SUS04-BP07 Minimizar a movimentação de dados entre redes
- SUS04-BP08 Fazer backup de dados somente quando for difícil recriá-los

SUS04-BP01 Implementar uma política de classificação de dados

Classifique os dados para entender sua importância para os resultados comerciais e selecione o nível de armazenamento eficiente em termos de energia para armazenar os dados.

Práticas comuns que devem ser evitadas:

- Você não identifica ativos de dados com características semelhantes (como sensibilidade, importância empresarial ou requisitos regulatórios) que estão sendo processados ou armazenados.
- Você não implementou um catálogo de dados para criar um inventário de seus ativos de dados.

Benefícios de implementar esta prática recomendada: a implementação de uma política de classificação de dados permite determinar o nível de armazenamento de dados com maior eficiência energética.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A classificação de dados envolve a identificação dos tipos de dados que estão sendo processados e armazenados em um sistema de informação pertencente a uma organização ou operado por ela. Ela também envolve a decisão em relação à importância dos dados e ao impacto provável do comprometimento dos dados, perda ou uso incorreto.

Implemente a política de classificação de dados trabalhando de forma reversa a partir do uso contextual dos dados e criando um esquema de categorização que leve em conta a importância de determinado conjunto de dados para as operações de uma organização.

Etapas de implementação

- Faça o inventário dos dados: faça um inventário dos vários tipos de dados que existem para sua workload.

- Agrupe os dados: determine a importância, a confidencialidade, a integridade e a disponibilidade dos dados com base no risco para a organização. Use esses requisitos para agrupar dados em um dos níveis de classificação de dados adotados. Como exemplo, consulte Quatro etapas simples para classificar seus dados e proteger sua startup.
- Defina níveis e políticas de classificação de dados: para cada grupo de dados, defina o nível de classificação de dados (por exemplo, público ou confidencial) e as políticas de tratamento. Marque os dados adequadamente. Para obter mais detalhes sobre as categorias de classificação de dados, consulte o whitepaper Classificação de dados.
- Revise periodicamente: revise e audite periodicamente seu ambiente em busca de dados não marcados e não classificados. Utilize automação para identificar esses dados e classificá-los e marcá-los adequadamente. Como exemplo, consulte Catálogo de dados e crawlers na AWS Glue.
- Estabeleça um catálogo de dados: estabeleça um catálogo de dados que forneça recursos de auditoria e governança.
- Documentação: documente as políticas de classificação de dados e os procedimentos de tratamento para cada classe de dados.

Recursos

Documentos relacionados:

- Utilizar a Nuvem AWS para apoiar a classificação de dados
- Políticas de tags do AWS Organizations

Vídeos relacionados:

- AWS re:Invent 2022: Capacitar a agilidade com a governança de dados na AWS
- AWS re:Invent 2023: Proteção e resiliência de dados com armazenamento na AWS

SUS04-BP02 Usar tecnologias compatíveis com seus padrões de acesso e de armazenamento de dados

Use tecnologias de armazenamento mais adequadas à maneira como seus dados são acessados e armazenados a fim de reduzir os recursos provisionados e, ao mesmo tempo, comportar sua workload.

Práticas comuns que devem ser evitadas:

- Você pressupõe que todas as workloads temja, padrões de acesso e armazenamento de dados semelhantes.
- Você usa apenas um nível de armazenamento, supondo que todas as workloads se encaixem nesse nível.
- Você pressupõe que os padrões de acesso aos dados permanecerão consistentes ao longo do tempo.

Benefícios de implementar esta prática recomendada: selecionar e otimizar suas tecnologias de armazenamento com base em padrões de armazenamento e acesso aos dados ajudará a reduzir os recursos de nuvem necessários a fim de atender às suas necessidades empresariais e melhorar a eficiência geral da workload de nuvem.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Selecione a solução de armazenamento mais alinhada a seus padrões de acesso ou considere a possibilidade de alterar seus padrões de acesso para alinhamento com a solução de armazenamento a fim de maximizar a eficiência da performance.

Etapas de implementação

- Avalie dados e acesse características: avalie suas características de dados e padrão de acesso a fim de reunir as principais características de suas necessidades de armazenamento. Principais características a serem consideradas:
  - Tipos de dados: estruturados, semiestruturados e não estruturados
  - Crescimento de dados: limitado, ilimitado
  - Durabilidade dos dados: persistentes, efêmeros, transitórios
  - Padrões de acesso: leituras ou gravações, frequência, com picos ou consistente
- Escolha a tecnologia de armazenamento correta: migre os dados para a tecnologia de armazenamento apropriada que seja compatível com suas características de dados e padrão de acesso. Veja alguns exemplos de tecnologias de armazenamento da AWS e suas principais características:

Tipo	Tecnologia	Características principais
Armazenamento de objetos	Amazon S3	Um serviço de armazenamento de objetos com escalabilidade ilimitada, alta disponibilidade e várias opções de acessibilidade. A transferência e o acesso de objetos dentro e fora do Amazon S3 podem usar um serviço, como o Transfer Acceleration ou Access Points, para oferecer suporte à sua localização, necessidades de segurança e padrões de acesso.
Armazenamento de arquivamento	Amazon S3 Glacier	Classe de armazenamento do Amazon S3 desenvolvida para arquivamento de dados.
Sistema de arquivos compartilhado	Amazon Elastic File System (Amazon EFS)	Sistema de arquivos montável que pode ser acessado por vários tipos de soluções de computação. O Amazon EFS aumenta e reduz automaticamente o armazenamento e possui performance otimizada para oferecer baixas latências consistentes.

Tipo	Tecnologia	Características principais
Sistema de arquivos compartilhado	Amazon FSx	Baseia-se nas soluções de computação mais recentes da AWS para oferecer compatibilidade com quatro sistemas de arquivos usados com frequência: NetApp ONTAP, OpenZFS, Windows File Server e Lustre. A latência, o throughput e as IOPS do Amazon FSx variam de acordo com o sistema de arquivos e devem ser consideradas ao selecionar o sistema de arquivos certo para as necessidades da sua workload.
Armazenamento em bloco	Amazon Elastic Block Store (Amazon EBS)	Serviço de armazenamento em blocos fácil de usar, escalável e de alta performance projetado para o Amazon Elastic Compute Cloud (Amazon EC2). O Amazon EBS inclui armazenamento baseado em SSD para workloads transacionais de alto throughput e em HDD para workloads trabalho de alto throughput.

Tipo	Tecnologia	Características principais
Banco de dados relacional	Amazon Aurora, Amazon RDS, Amazon Redshift	Projetados para oferecer compatibilidade com transações ACID (atomicidade, consistência, isolamento, durabilidade) e manter a integridade referencial e uma forte consistência de dados. Muitas aplicações tradicionais, sistemas de planejamento de recursos empresariais (ERP), de gerenciamento de relacionamentos com o cliente (CRM) e de comércio eletrônico usam bancos de dados relacionais para armazenar seus dados.
Banco de dados de chave-valor	Amazon DynamoDB	Otimizados para padrões de acesso comuns, normalmente visando armazenar e recuperar grandes volumes de dados. Aplicações web de alto tráfego, sistemas de comércio eletrônico e aplicações de jogos são os casos de uso habituais para bancos de dados de chave-valor.

- Automatize a alocação de armazenamento: para sistemas de armazenamento que têm tamanho fixo, como o Amazon EBS ou o Amazon FSx, monitore o espaço de armazenamento disponível e automatize a alocação de armazenamento ao atingir um limite. É possível utilizar o Amazon CloudWatch para coletar e analisar diferentes métricas para o Amazon EBS e o Amazon FSx.
- Escolha a classe de armazenamento correta: escolha a classe de armazenamento apropriada para seus dados.

- As classes de armazenamento do Amazon S3 podem ser configuradas em nível de objeto. Um único bucket pode conter objetos armazenados em todas as classes de armazenamento.
- É possível usar políticas de ciclo de vida do Amazon S3 para realizar a transição automática de objetos entre classes de armazenamento ou remover dados sem nenhuma alteração na aplicação. Em geral, é necessário fazer uma compensação entre a eficiência dos recursos, a latência de acesso e a confiabilidade ao considerar esses mecanismos de armazenamento.

Recursos

Documentos relacionados:

- Tipos de volume do Amazon EBS
- Armazenamento de instâncias do Amazon EC2
- Amazon S3 Intelligent-Tiering
- Características de E/S do Amazon EBS
- Usar classes de armazenamento do Amazon S3
- O que é o Amazon S3 Glacier?

Vídeos relacionados:

- AWS re:Invent 2023: Melhorar a eficiência do Amazon EBS e ser mais econômico
- AWS re:Invent 2023: Otimizar o preço e a performance do armazenamento com o Amazon S3
- AWS re:Invent 2023: Criar e otimizar data lakes no Amazon S3
- AWS re:Invent 2022: Construir arquiteturas de dados modernos na AWS
- AWS re:Invent 2022: Modernizar aplicações com bancos de dados com propósito específico
- AWS re:Invent 2022: Construir arquiteturas de data mesh na AWS
- AWS re:Invent 2023: Mergulho profundo no Amazon Aurora e suas inovações
- AWS re:Invent 2023: Modelagem de dados com o Amazon DynamoDB

Exemplos relacionados:

- Exemplos do Amazon S3
- Workshop de bancos de dados com propósito específico na AWS
- Bancos de dados para desenvolvedores

- Dia de imersão na arquitetura de dados moderna na AWS
- Criar um data mesh na AWS

SUS04-BP03 Usar políticas para gerenciar o ciclo de vida de seus conjuntos de dados

Gerencie o ciclo de vida de todos os seus dados e aplique a exclusão automaticamente para minimizar o armazenamento total necessário para sua workload.

Práticas comuns que devem ser evitadas:

- Você exclui dados manualmente.
- Você não exclui nenhum de seus dados de workload.
- Você não move os dados para níveis de armazenamento mais eficientes em termos de energia com base em seus requisitos de retenção e acesso.

Benefícios de estabelecer essa prática recomendada: o uso de políticas de ciclo de vida de dados garante acesso e retenção eficientes de dados em uma workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Os conjuntos de dados geralmente têm diferentes requisitos de retenção e acesso durante seu ciclo de vida. Por exemplo, sua aplicação pode precisar de acesso frequente a alguns conjuntos de dados por um período limitado. Depois disso, esses conjuntos de dados são acessados com pouca frequência.

Para gerenciar com eficiência seus conjuntos de dados ao longo de seu ciclo de vida, configure políticas de ciclo de vida, as quais são regras que definem como lidar com conjuntos de dados.

Com as regras de configuração do ciclo de vida, é possível orientar o serviço de armazenamento específico a fazer a transição de um conjunto de dados para níveis de armazenamento mais eficientes em termos de energia, arquivá-lo ou excluí-lo.

Etapas de implementação

- Classifique os conjuntos de dados em sua workload.
- Defina procedimentos de manipulação para cada classe de dados.

- Defina políticas automatizadas de ciclo de vida para aplicar regras de ciclo de vida. Aqui estão alguns exemplos de como configurar políticas de ciclo de vida automatizadas para diferentes serviços de armazenamento do AWS:

Serviços de armazenamento	Como definir políticas de ciclo de vida automatizadas
Amazon S3	Você pode usar o Amazon S3 Lifecycle para gerenciar seus objetos durante todo o ciclo de vida de cada um. Se os padrões de acesso forem desconhecidos, variáveis ou imprevisíveis, você poderá usar o Amazon S3 Intelligent-Tiering para monitorar os padrões de acesso e mover automaticamente os objetos que não foram acessados para níveis de acesso de baixo custo. Também é possível usar as métricas da Lente de Armazenamento do Amazon S3 para identificar oportunidades de otimização e lacunas no gerenciamento do ciclo de vida.
Amazon Elastic Block Store	É possível usar o Amazon Data Lifecycle Manager para automatizar a criação, a retenção e a exclusão de snapshots do Amazon EBS e de AMIs apoiadas pelo Amazon EBS.
Amazon Elastic File System	O gerenciamento de ciclo de vida útil do Amazon EFS gerencia automaticamente o armazenamento de arquivos dos seus sistemas de arquivos.
Amazon Elastic Container Registry	As políticas de ciclo de vida do Amazon ECR permitem automatizar a limpeza de imagens de contêineres expirando as imagens com base na idade ou no número.

Serviços de armazenamento	Como definir políticas de ciclo de vida automatizadas
AWS Elemental MediaStore	Você pode usar uma política de ciclo de vida de objeto que determina por quanto tempo os objetos devem ser armazenados no contêiner MediaStore.

- Exclua volumes não utilizados, snapshots e dados que estão fora do período de retenção. Aproveite os recursos de serviços nativos, como o tempo de vida útil do Amazon DynamoDB ou a retenção de log do Amazon CloudWatch para exclusão.
- Agregue e compacte dados quando possível com base nas regras do ciclo de vida.

Recursos

Documentos relacionados:

- Otimizar regras de ciclo de vida do Amazon S3 com a analise de classes de armazenamento do Amazon S3
- Avaliar recursos com o Regras do AWS Config

Vídeos relacionados:

- AWS re:Invent 2021: Práticas recomendadas do ciclo de vida do Amazon S3 para otimizar seus gastos com armazenamento
- AWS re:Invent 2023: Otimizar o preço e a performance do armazenamento com o Amazon S3
- Simplificar o ciclo de vida de dados e otimizar os custos de armazenamento com o Amazon S3 Lifecycle
- Reduzir seus custos de armazenamento com a Lente de Armazenamento do Amazon S3

SUS04-BP04 Usar elasticidade e automação para expandir o armazenamento em bloco ou o sistema de arquivos

Use a elasticidade e automação para expandir o armazenamento em bloco ou o sistema de arquivos à medida que os dados aumentam para minimizar o total de armazenamento provisionado.

Práticas comuns que devem ser evitadas:

- Você adquire um grande armazenamento em bloco ou sistema de arquivos para necessidades futuras.
- Você provisiona em excesso as operações de entrada e saída por segundo (IOPS) de seu sistema de arquivos.
- Você não monitora a utilização de seus volumes de dados.

Benefícios de implementar esta prática recomendada: minimizar o provisionamento excessivo do sistema de armazenamento reduz os recursos ociosos e melhora a eficiência geral de sua workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Crie armazenamento em bloco ou sistemas de arquivos com alocação de tamanho, throughput e latência apropriados à workload. Use a elasticidade e automação para expandir o armazenamento em bloco ou o sistema de arquivos à medida que os dados aumentarem sem precisar provisionar em excesso esses serviços de armazenamento.

Etapas de implementação

- Para sistemas de armazenamento com tamanho fixo, como o Amazon EBS, certifique-se de monitorar a quantidade de armazenamento usada em comparação com o tamanho geral do armazenamento e, se possível, criar automação para aumentar o tamanho do armazenamento ao atingir um limite.
- Use volumes elásticos e serviços gerenciados de dados em bloco para automatizar a alocação de armazenamento adicional à medida que os seus dados persistentes aumentarem. Por exemplo, é possível usar Volumes Elásticos do Amazon EBS para alterar o tamanho ou o tipo de volume ou ajustar a performance de seus volumes do Amazon EBS.
- Escolha a classe de armazenamento, modo de performance e modo de throughput corretos para seu sistema de arquivos para atender à necessidade de seus negócios, sem a ultrapassar.
  - Performance do Amazon EFS
  - Performance do volume do Amazon EBS em instâncias do Linux
- Defina os níveis pretendidos de utilização para seus volumes de dados e redimensione os volumes fora dos intervalos esperados.
- Dimensione adequadamente volumes somente leitura para acomodar os dados.
- Migre os dados para depósitos de objetos a fim de evitar o provisionamento de capacidade em excesso que ocorre com os tamanhos de volumes fixos no armazenamento em bloco.

- Revise regularmente volumes elásticos e sistemas de arquivos para encerrar volumes ociosos, reduzir recursos com excesso de provisionamento e se ajustar ao tamanho de dados atual.

Recursos

Documentos relacionados:

- Estender um sistema de arquivos após redimensionar um volume do EBS
- Modificar um volume do EBS usando Volumes Elásticos do Amazon EBS
- Documentação do Amazon FSx
- O que é o Amazon Elastic File System?

Vídeos relacionados:

- Mergulho profundo nos Volumes Elásticos do Amazon EBS
- Amazon EBS e estratégias de otimização de snapshots para melhor performance e redução de custos
- Otimizar o Amazon EFS para custo e performance usando práticas recomendadas

SUS04-BP05 Remover dados desnecessários ou redundantes

Remova dados desnecessários ou redundantes para minimizar os recursos de armazenamento necessários para armazenar seus conjuntos de dados.

Práticas comuns que devem ser evitadas:

- Você duplica dados que podem ser facilmente obtidos ou recriados.
- Você faz backup de todos os dados sem considerar sua criticidade.
- Você apenas exclui dados irregularmente, em eventos operacionais ou não os exclui.
- Você armazena dados de forma redundante, independentemente da durabilidade do serviço de armazenamento.
- Você ativa o versionamento do Amazon S3 sem qualquer justificativa comercial.

Benefícios de implementar esta prática recomendada: a remoção de dados desnecessários reduz o tamanho de armazenamento necessário para sua workload e o impacto ambiental causado por ela.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Não armazene dados de que você não precisa. Automatize a exclusão de dados desnecessários. Use tecnologias que eliminam dados duplicados em níveis de arquivo e bloco. Aproveite a replicação de dados nativos e os recursos de redundância dos serviços.

Etapas de implementação

- Avalie se você pode evitar o armazenamento de dados usando conjuntos de dados existentes publicamente disponíveis no AWS Data Exchange e Open Data on AWS.
- Use mecanismos que possam duplicar dados no nível de bloco e objeto. Aqui estão alguns exemplos de como eliminar duplicações dados na AWS:

Serviços de armazenamento	Mecanismo de eliminação de duplicações
Amazon S3	Use o AWS Lake Formation FindMatches para encontrar registros correspondentes em um conjunto de dados (incluindo aqueles sem identificadores) usando a nova transformada de ML do FindMatches.
Amazon FSx	Use a eliminação de duplicação de dados no Amazon FSx para Windows.
Snapshots do Amazon Elastic Block Store	Snapshots são backups incrementais, o que significa que somente os blocos no dispositivo que tiverem mudado depois do snapshot mais recente serão salvos.

- Analise o acesso aos dados para identificar dados desnecessários. Automatize as políticas de ciclo de vida. Utilize recursos de serviços nativos, como o tempo de vida útil do Amazon DynamoDB, o ciclo de vida do Amazon S3 ou a retenção de logs do Amazon CloudWatch para exclusão.
- Use os recursos de virtualização de dados no AWS para manter os dados em sua origem e evitar a duplicação de dados.
  - Virtualização de dados nativos da nuvem na AWS
  - Otimizar o padrão de dados usando o compartilhamento de dados do Amazon Redshift

- Use tecnologia de backup capaz de fazer backups incrementais.
- Aproveite a durabilidade do Amazon S3 e a replicação do Amazon EBS, em vez de tecnologias autogerenciadas (como uma matriz redundante de discos independentes (RAID)), para atingir suas metas de durabilidade.
- Centralize o log e rastreie os dados, elimine a duplicação de entradas de log idênticas e estabeleça mecanismos para ajustar a prolixidade quando necessário.
- Preencha os caches com antecedência somente quando justificável.
- Estabeleça o monitoramento e a automação de cache para redimensionar o cache de forma adequada.
- Remova implantações e ativos desatualizados dios repositórios de objetos e caches de borda ao enviar novas versões da sua workload por push.

Recursos

Documentos relacionados:

- Alterar a retenção de dados de log no CloudWatch Logs
- Eliminação da duplicação de dados no Amazon FSx para Windows File Server
- Recursos do Amazon FSx para ONTAP, incluindo a eliminação da duplicação de dados
- Invalidação de arquivos no Amazon CloudFront
- Usar o AWS Backup para fazer backup e restaurar sistemas de arquivos do Amazon EFS
- O que é o Amazon CloudWatch Logs?
- Trabalhar com backups no Amazon RDS
- Integrar e eliminar duplicações de conjuntos de dados usando o AWS Lake Formation

Vídeos relacionados:

- Casos de uso de compartilhamento de dados do Amazon Redshift

Exemplos relacionados:

- Como analiso meus logs de acesso ao servidor do Amazon S3 usando o Amazon Athena?

SUS04-BP06 Usar armazenamento ou sistemas de arquivos compartilhados para acessar dados comuns

Adote armazenamento ou sistemas de arquivos compartilhados para evitar a duplicação de dados e viabilizar uma infraestrutura mais eficiente para a workload.

Práticas comuns que devem ser evitadas:

- Você provisiona armazenamento para cada cliente específico.
- Você não desanexa o volume de dados dos clientes inativos.
- Você não fornece acesso a armazenamento em plataformas e sistemas.

Benefícios de estabelecer essa prática recomendada: o uso de sistemas de arquivos ou armazenamento compartilhados permite compartilhar dados com um ou mais consumidores sem precisar copiá-los. Isso ajuda a reduzir os recursos de armazenamento necessários à workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Se você tiver vários usuários ou aplicações que acessam os mesmos conjuntos de dados, o uso da tecnologia de armazenamento compartilhado é essencial para viabilizar uma infraestrutura eficiente para a workload. A tecnologia de armazenamento compartilhado oferece um local central para armazenar e gerenciar conjuntos de dados e evitar a duplicação de dados. Ela também impõe a consistência dos dados em sistemas diferentes. Além disso, a tecnologia de armazenamento compartilhado permite o uso mais eficiente da potência computacional, visto que vários recursos podem acessar e processar os dados ao mesmo tempo em paralelo.

Busque dados dos serviços de armazenamento compartilhado somente quando necessário e desanexe os volumes não usados para liberar recursos.

Etapas de implementação

- Migre dados para o armazenamento compartilhado quando eles tiverem vários consumidores. Veja aqui alguns exemplos de tecnologia de armazenamento compartilhado na AWS:

Opções de armazenamento	Quando usar
Amazon EBS Multi-Attach	O Amazon EBS Multi-Attach permite que você anexe um único volume de SSD de IOPS

Opções de armazenamento	Quando usar
	provisionadas (io1 ou io2) a várias instâncias que estão na mesma zona de disponibilidade.
Amazon EFS	Consulte Quando escolher o Amazon EFS.
Amazon FSx	Consulte Escolher um sistema de arquivos Amazon FSx.
Amazon S3	As aplicações que não exigem uma estrutura de sistema de arquivos e são projetadas para funcionar com armazenamento de objetos podem usar o Amazon S3 como solução de armazenamento de objetos altamente escalável, durável e de baixo custo.

- Copie os dados para ou busque os dados de sistemas de arquivos compartilhados somente quando necessário. Como exemplo, você pode criar um sistema de arquivos Amazon FSx para Lustre apoiado pelo Amazon S3 e carregar somente o subconjunto de dados necessários para o processamento de trabalhos no Amazon FSx.
- Exclua dados conforme apropriado para os seus padrões de uso conforme descrito em SUS04-BP03 Usar políticas para gerenciar o ciclo de vida de seus conjuntos de dados.
- Desvincule volumes de clientes que não os estão usando ativamente.

Recursos

Documentos relacionados:

- Vincular o sistema de arquivos a um bucket do Amazon S3
- Usar o Amazon EFS para AWS Lambda em aplicações com tecnologia sem servidor
- O Amazon EFS Intelligent-Tiering otimiza os custos das workloads com padrões de acesso variáveis
- Como usar o Amazon FSx com seu repositório de dados on-premises

Vídeos relacionados:

- Otimização do custo de armazenamento com o Amazon EFS

- AWS re:Invent 2023: Novidades do armazenamento de arquivos na AWS
- AWS re:Invent 2023: Armazenamento de arquivos para criadores e cientistas de dados no Amazon Elastic File System

SUS04-BP07 Minimizar a movimentação de dados entre redes

Use armazenamento de objetos ou sistemas de arquivos compartilhados para acessar dados comuns e minimizar os recursos totais de rede exigidos para comportar a movimentação de dados da workload.

Práticas comuns que devem ser evitadas:

- Você armazena todos os dados na mesma Região da AWS independentemente de onde os usuários dos dados estão.
- Você não otimiza o tamanho e o formato dos dados antes de movimentá-los na rede.

Benefícios de implementar esta prática recomendada: otimizar a movimentação de dados na rede reduz os recursos totais de rede necessários à workload e reduz o respectivo impacto ambiental.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A movimentação de dados em sua organização exige recursos de computação, rede e armazenamento. Use técnicas para minimizar a movimentação de dados e melhorar a eficiência geral da workload.

Etapas de implementação

- Considere a proximidade dos dados ou dos usuários como um fator decisivo ao selecionar uma região para sua workload.
- Particione serviços consumidos regionalmente para que os dados específicos da região sejam armazenados na região em que eles são consumidos.
- Use formatos de arquivo eficientes (como Parquet ou ORC) e compacte os dados antes movimentá-los na rede.
- Não movimente dados não usados. Alguns exemplos que podem ajudar você a evitar a movimentação de dados não utilizados:
  - Reduza as respostas de API apenas aos dados relevantes.

  - Agregue os dados onde não houver necessidade de informações detalhadas (em nível de registro).
  - Consulte Laboratório do Well-Architected: Otimizar padrão de dados usando o compartilhamento de dados do Amazon Redshift.
  - Considere Compartilhamento de dados entre contas no AWS Lake Formation.
- Use serviços que possam ajudar você a executar código mais perto dos usuários da workload.

Serviço	Quando usar
Lambda@Edge	Use para operações com uso intenso de computação que são executadas quando objetos não estão no cache.
CloudFront Functions	Use para casos de uso simples como solicitações HTTP(s)/manipulações de resposta que podem ser iniciadas por funções de curta duração.
AWS IoT Greengrass	Execute computação local, mensagens e armazenamento de dados em cache para dispositivos conectados.

Recursos

Documentos relacionados:

- Otimizar a sua infraestrutura da AWS para sustentabilidade, Parte III: Redes
- Infraestrutura global da AWS
- Principais recursos do Amazon CloudFront incluindo a rede de borda global do CloudFront
- Compactação de solicitações HTTP no Amazon OpenSearch Service
- Intermediar a compactação de dados com o Amazon EMR
- Carregar arquivos de dados compactados do Amazon S3 no Amazon Redshift
- Envio de arquivos compactados com o Amazon CloudFront

Vídeos relacionados:

- Desmistificar a transferência de dados na AWS

Exemplos relacionados:

- Arquitetura para a sustentabilidade: reduza a movimentação de dados entre redes

SUS04-BP08 Fazer backup de dados somente quando for difícil recriá-los

Evite fazer backup de dados que não têm valor empresarial para minimizar os requisitos de recursos de armazenamento da workload.

Práticas comuns que devem ser evitadas:

- Você não tem uma estratégia de backup para seus dados.
- Você faz backup de dados que podem ser recriados com facilidade.

Benefícios de implementar esta prática recomendada: evitar o backup de dados não críticos reduz os recursos de armazenamento necessários para a workload e diminui seu impacto ambiental.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Evitar o backup de dados desnecessários pode ajudar a reduzir os custos e os recursos de armazenamento usados pela workload. Faça backup somente de dados com valor empresarial ou que sejam necessários para atender a requisitos de conformidade. Examine as políticas de backup e exclua armazenamentos temporários que não fornecem valor em um cenário de recuperação.

Etapas de implementação

- Implemente a política de classificação de dados conforme descrito em SUS04-BP01 Implementar uma política de classificação de dados.
- Use a importância da sua classificação de dados e projete uma estratégia de backup com base em seus objetivo de tempo de recuperação (RTO) e o objetivo de ponto de recuperação (RPO). Evite fazer backup de dados não essenciais.
  - Exclua dados que podem ser recriados com facilidade.
  - Exclua dados temporários dos seus backups.

  - Exclua cópias locais de dados, a menos que o tempo necessário para restaurar esses dados de um local comum exceda seus Acordos de Serviço (SLAs).
- Use uma solução automatizada ou um serviço gerenciado para fazer backup de dados essenciais aos negócios.
  - O AWS Backup é um serviço totalmente gerenciado que ajuda você a centralizar e automatizar a proteção de dados nos serviços da AWS, na nuvem e on-premises. Para obter orientação prática sobre como criar backups automatizados usando o AWS Backup, consulte Laboratórios do Well-Architected: Testar o backup e a restauração de dados.
  - Automatize backups e otimize os custos de backup para o Amazon EFS usando o AWS Backup.

Recursos

Práticas recomendadas relacionadas:

- REL09-BP01 Identificar e fazer backup de todos os dados que precisam de backup ou reproduzir os dados das fontes
- REL09-BP03 Fazer backup de dados automaticamente
- REL13-BP02 Usar estratégias de recuperação definidas para cumprir os objetivos de recuperação

Documentos relacionados:

- Usar o AWS Backup para fazer backup e restaurar sistemas de arquivos do Amazon EFS
- Snapshots do Amazon EBS
- Trabalhar com backups no Amazon Relational Database Service
- Parceiro da APN: parceiros que podem ajudar com o backup
- AWS Marketplace: produtos que podem ser usados para backup
- Fazer backup do Amazon EFS
- Fazer backup do Amazon FSx para Windows File Server
- Backup e restauração do Amazon ElastiCache (Redis OSS)

Vídeos relacionados:

- AWS re:Invent 2023: Estratégias de backup e recuperação de desastres para aumentar a resiliência

- AWS re:Invent 2023: Novidades no AWS Backup
- AWS re:Invent 2021: Backup, recuperação de desastres e proteção contra ransomware com a AWS

Exemplos relacionados:

- Laboratório do Well-Architected: Backup de dados

Hardware e serviços

Pergunta

- SUS 5 Como selecionar e usar hardware e serviços em nuvem na arquitetura para apoiar os objetivos de sustentabilidade?

SUS 5 Como selecionar e usar hardware e serviços em nuvem na arquitetura para apoiar os objetivos de sustentabilidade?

Procure oportunidades para reduzir os impactos na sustentabilidade da workload fazendo mudanças nas suas práticas de gerenciamento de hardware. Minimize a quantidade de hardware necessária para provisionar e implantar e escolha o hardware e os serviços mais eficientes para sua workload específica.

Práticas recomendadas

- SUS05-BP01 Usar a quantidade mínima de hardware para atender às suas necessidades
- SUS05-BP02 Usar tipos de instância com o mínimo de impacto
- SUS05-BP03 Usar serviços gerenciados
- SUS05-BP04 Otimizar o uso de aceleradores de computação baseados em hardware

SUS05-BP01 Usar a quantidade mínima de hardware para atender às suas necessidades

Use a quantidade mínima de hardware para sua workload para atender com eficiência às suas necessidades de negócios.

Práticas comuns que devem ser evitadas:

- Você não monitora a utilização de recursos.

- Você tem recursos com baixo nível de utilização em sua arquitetura.
- Você não analisa a utilização de hardware estático para determinar se é necessário redimensioná-lo.
- Você não define metas de utilização de hardware para sua estrutura de computação com base nos KPIs de negócios.

Benefícios de implementar esta prática recomendada: o dimensionamento correto de seus recursos de nuvem ajuda a reduzir o impacto ambiental de uma workload, economizar dinheiro e manter os benchmarks de performance.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Selecione do modo mais eficiente o número total de hardware necessário à workload para melhorar a eficiência geral. A Nuvem AWS fornece a flexibilidade de expandir ou reduzir o número de recursos dinamicamente por meio de diversos mecanismos, como o AWS Auto Scaling, para atender a mudanças na demanda. Ela também fornece APIs e SDKs que permitem que os recursos sejam modificados com o mínimo de esforço. Use esses recursos para fazer alterações frequentes nas implementações da workload. Além disso, use as orientações sobre dimensionamento correto das ferramentas da AWS para operar com eficiência o recursos de nuvem e atender às suas necessidades empresariais.

Etapas de implementação

- Escolha o tipo de instância: escolha o tipo de instância certo para melhor atender às suas necessidades. Para saber como escolher instâncias do Amazon Elastic Compute Cloud e usar mecanismos, como a seleção de instâncias com base em atributos, consulte:
  - Como faço para escolher o tipo de instância do Amazon EC2 apropriado para minha workload?
  - Seleção de tipo de instância baseada em atributos para Amazon EC2 Fleet.
  - Criar um grupo do Auto Scaling usando seleção de tipo de instância baseada em atributos.
- Escala: use pequenos incrementos para workloads variáveis.
- Use várias opções de compra de computação: equilibre a flexibilidade, a escalabilidade e a redução de custos da instância com várias opções de compra de computação.
  - As instâncias sob demanda do Amazon EC2 são mais adequadas para workloads novas, dinâmicas e com estado que não podem ser flexíveis em termos de tipo de instância, localização ou horário.

  - As instâncias spot do Amazon EC2 são uma ótima maneira de complementar as outras opções para aplicações que são flexíveis e tolerantes a falhas.
  - Aproveite os Savings Plans para computação para workloads de estado estável que permitem flexibilidade se suas necessidades (como AZ, região, famílias de instâncias ou tipos de instância) mudarem.
- Use a diversidade de instâncias e zonas de disponibilidade: maximize a disponibilidade das aplicações e aproveite o excesso de capacidade diversificando suas instâncias e zonas de disponibilidade.
- Dimensione as instâncias corretamente: use as recomendações de dimensionamento correto das ferramentas da AWS para fazer ajustes na workload. Para obter mais informações, consulte Como otimizar seu custo com as recomendações de redimensionamento correto e Dimensionamento correto: provisionamento de instâncias para corresponder às workloads.
  - Use as recomendações de dimensionamento correto no AWS Cost Explorer ou no AWS Compute Optimizer para identificar oportunidades de dimensionamento correto.
- Negocie acordos de serviço (SLAs): que permitam uma redução temporária na capacidade enquanto a automação implanta recursos de substituição.

Recursos

Documentos relacionados:

- Otimizar a sua infraestrutura da AWS para sustentabilidade, Parte I: Computação
- Seleção de tipo de instância baseada em atributos para Auto Scaling para Amazon EC2 Fleet
- Documentação do AWS Compute Optimizer
- Operação do Lambda: otimização da performance
- Documentação do Auto Scaling

Vídeos relacionados:

- AWS re:Invent 2023: Novidades no Amazon EC2
- AWS re:Invent 2023: Economias inteligentes: estratégias de otimização de custos no Amazon Elastic Compute Cloud
- AWS re:Invent 2022: Otimizar o Amazon Elastic Kubernetes Service para performance e custo na AWS

- AWS re:Invent 2023: Computação sustentável: reduzir custos e emissões de carbono com a AWS

SUS05-BP02 Usar tipos de instância com o mínimo de impacto

Monitore continuamente e use novos tipos de instância para aproveitar as melhorias de eficiência de energia.

Práticas comuns que devem ser evitadas:

- Você usa apenas uma família de instâncias.
- Você usa apenas instâncias x86.
- Você especifica um tipo de instância em sua configuração do Amazon EC2 Auto Scaling.
- Você usa instâncias da AWS de um modo para o qual elas não foram projetadas (por exemplo, você usa instâncias otimizadas para computação em uma workload com uso intenso de memória).
- Você não avalia os novos tipos de instância regularmente.
- Você não verifica as recomendações de ferramentas de dimensionamento correto da AWS, como o AWS Compute Optimizer.

Benefícios de implementar esta prática recomendada: ao usar instâncias com eficiência de energia e dimensionadas corretamente, é possível reduzir ainda mais o impacto ambiental e o custo da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Usar instâncias eficientes na workload de nuvem é essencial para reduzir o uso de recursos e os custos. Monitore continuamente o lançamento de novos tipos de instância e aproveite as melhorias de eficiência de energia, incluindo os tipos de instância projetados para comportar workloads específicas, como treinamento e inferência de machine learning e transcodificação de vídeo.

Etapas de implementação

- Conheça e explore tipos de instâncias: encontre tipos de instâncias que podem reduzir o impacto ambiental da sua workload.
  - Assine as Novidades da AWS para se manter em dia com as tecnologias e instâncias mais recentes da AWS.

  - Conheça os diversos tipos de instâncias da AWS.
  - Saiba mais sobre instâncias baseadas no AWS Graviton que oferecem a melhor performance por watt de uso de energia no Amazon EC2 assistindo ao re:Invent 2020: Mergulho profundo nas instâncias do Amazon EC2 baseadas no processador AWS Graviton2 e Mergulho profundo no AWS Graviton 3 e instâncias C7g do Amazon EC2.
- Use tipos de instâncias com o mínimo de impacto: planeje e migre sua workload para tipos de instância com impacto mínimo.
  - Defina um processo para avaliar novos recursos ou instâncias para a workload. Aproveite a agilidade da nuvem para testar rapidamente como novos tipos de instância podem melhorar a sustentabilidade ambiental de sua workload. Use métricas de proxy para mensurar quantos recursos são necessários para concluir uma unidade de trabalho.
  - Se possível, modifique sua workload para trabalhar com diferentes números de vCPUs e diferentes quantidades de memória para maximizar sua escolha de tipo de instância.
  - Pense em migrar a workload para instâncias baseadas em Graviton e melhorar a eficiência da performance da workload. Para obter mais informações sobre como mover workloads para o AWS Graviton, consulte Início rápido do AWS Graviton e Considerações ao fazer a transição de workloads para instâncias do Amazon Elastic Compute Cloud baseadas no AWS Graviton.
  - Considere selecionar a opção AWS Graviton em seu uso de serviços gerenciados da AWS.
  - Migre sua workload para regiões que ofereçam instâncias com o menor impacto na sustentabilidade e atendam aos seus requisitos de negócios.
  - Para workloads de machine learning, utilize hardware específico para sua workload, como AWS Trainium, AWS Inferentia e Amazon EC2 DL1. AWS Instâncias do Inferentia, como instâncias Inf2, oferecem performance até 50% melhor por watt em relação a instâncias comparáveis do Amazon EC2.
  - Use o Amazon SageMaker Inference Recommender para dimensionar corretamente o endpoint de inferência de ML.
  - Para workloads com picos (workloads com requisitos irregulares para capacidade adicional), use instâncias de performance expansível.
  - Para workloads sem estado e tolerantes a falhas, use instâncias spot do Amazon EC2 para aumentar a utilização geral da nuvem e reduzir o impacto na sustentabilidade de recursos não utilizados.
- Opere e otimize: opere e otimize a instância da sua workload.
  - Para workloads efêmeras, avalie as métricas da instância do Amazon CloudWatch, como `CPUUtilization`, para identificar se a instância está ociosa ou subutilizada.

- Para workloads estáveis, verifique as ferramentas de dimensionamento correto da AWS, como AWS Compute Optimizer, em intervalos regulares para identificar oportunidades de otimizar e dimensionar corretamente as instâncias. Para obter mais exemplos e recomendações, consulte os seguintes laboratórios:
  - Laboratório do Well-Architected: Recomendações de dimensionamento correto
  - Laboratório do Well-Architected: Dimensionamento correto com o Compute Optimizer
  - Laboratório do Well-Architected: Otimizar padrões de hardware e observar KPIs de sustentabilidade

Recursos

Documentos relacionados:

- Otimizar a sua infraestrutura da AWS para sustentabilidade, Parte I: Computação
- AWS Graviton
- DL1 do Amazon EC2
- Frotas de reserva de capacidade do Amazon EC2
- Frota spot do Amazon EC2
- Funções: configuração da função do Lambda
- Seleção de tipo de instância baseada em atributos para Amazon EC2 Fleet
- Criar aplicações sustentáveis, eficientes e com custo otimizado na AWS
- Como o Painel de Sustentabilidade da Contino ajuda os clientes a otimizar sua pegada de carbono

Vídeos relacionados:

- AWS re:Invent 2023: AWS Graviton: a melhor performance de preços para suas workloads da AWS
- AWS re:Invent 2023: Novos recursos de IA generativa do Amazon Elastic Compute Cloud no AWS Management Console
- AWS re:Invent 2023: Novidades do Amazon Elastic Compute Cloud
- AWS re:Invent 2023: Economias inteligentes: estratégias de otimização de custos no Amazon Elastic Compute Cloud
- AWS re:Invent 2021: Mergulho profundo no AWS Graviton 3 e instâncias C7g do Amazon EC2

- AWS re:Invent 2022: Criar um ambiente de computação eficiente em termos de custo, energia e recursos

Exemplos relacionados:

- Solução: orientações sobre como otimizar workloads de aprendizado profundo para sustentabilidade na AWS
- Migração de bancos de dados do Amazon Relational Database Service para o Graviton

SUS05-BP03 Usar serviços gerenciados

Use serviços gerenciados para operar com maior eficiência na nuvem.

Práticas comuns que devem ser evitadas:

- Você usa instâncias do Amazon EC2 com baixa utilização para executar suas aplicações.
- Sua equipe interna gerencia apenas a workload e não tem tempo para se concentrar em inovação ou simplificações.
- Você implanta e mantém tecnologias para tarefas que podem ser executadas com maior eficiência em serviços gerenciados.

Benefícios de implementar esta prática recomendada:

- Com o uso de serviços gerenciados, a responsabilidade é transferida para a AWS, que tem insights referentes a milhões de clientes que podem ajudar a promover inovações inéditas e melhorar a eficiência.
- O serviço gerenciado distribui o impacto ambiental do serviço entre vários usuários em virtude dos ambientes de gerenciamento de vários locatários.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Como os serviços gerenciados, a responsabilidade por manter a alta utilização e otimizar a sustentabilidade do hardware implantado é transferida para a AWS. Os serviços gerenciados também eliminam as despesas operacionais e administrativas da manutenção de um serviço, o que permite que sua equipe tenha mais tempo para se concentrar na inovação.

Avalie sua workload para identificar componentes que podem ser substituídos por serviços gerenciados da AWS. Por exemplo, o Amazon RDS, o Amazon Redshift e o Amazon ElastiCache fornecem um serviço de banco de dados gerenciado. O Amazon Athena, o Amazon EMR e o Amazon OpenSearch Service oferecem um serviço de análise gerenciado.

Etapas de implementação

1. Faça o inventário da workload: faça um inventário de serviços e componentes para sua workload.
2. Identifique candidatos: avalie e identifique componentes que podem ser substituídos por serviços gerenciados. Veja aqui alguns exemplos de quando considerar usar um serviço gerenciado:

Tarefa	O que usar na AWS
Hospedar um banco de dados	Use instâncias do Amazon Relational Database Service (Amazon RDS) gerenciadas em vez de manter instâncias do Amazon RDS no Amazon Elastic Compute Cloud (Amazon EC2).
Hospedar uma workload conteinerizada	Use o AWS Fargate em vez de implementar sua própria infraestrutura de contêineres.
Hospedar aplicações Web	Use o AWS Amplify Hosting como CI/CD totalmente gerenciado e serviço de hospedagem para sites estáticos e aplicações Web renderizadas no lado do servidor.

3. Crie um plano de migração: identifique dependências e crie um plano de migração. Atualize runbooks e playbook de forma apropriada.
   - O AWS Application Discovery Service coleta e apresenta automaticamente informações detalhadas sobre dependências e utilização de aplicações que ajudam a tomar decisões mais fundamentadas durante o planejamento da migração.
4. Faça testes: teste o serviço antes de migrar para o serviço gerenciado.
5. Substitua os serviços auto-hospedados: use seu plano de migração para substituir os serviços auto-hospedados por serviços gerenciados.
6. Monitore e ajuste: monitore continuamente o serviço após a conclusão da migração para fazer ajustes conforme necessário e otimizar o serviço.

Recursos

Documentos relacionados:

- Produtos da Nuvem AWS
- Calculadora de custo total de propriedade (TCO) da AWS
- Amazon DocumentDB
- Amazon Elastic Kubernetes Service (EKS)
- Amazon Managed Streaming for Apache Kafka (Amazon MSK)

Vídeos relacionados:

- AWS re:Invent 2021: Operações na nuvem em grande escala com o AWS Managed Services
- AWS re:Invent 2023: Práticas recomendadas para operar na AWS

SUS05-BP04 Otimizar o uso de aceleradores de computação baseados em hardware

Otimize o uso de instâncias com computação acelerada para reduzir as demandas de infraestrutura física de sua workload.

Práticas comuns que devem ser evitadas:

- Você não está monitorando o uso da GPU.
- Você está usando uma instância de finalidade geral para workload, enquanto uma instância criada especificamente pode oferecer maior performance, menor custo e melhor performance por watt.
- Você está usando aceleradores de computação baseados em hardware para tarefas em que são mais eficientes usando alternativas baseadas em CPU.

Benefícios de implementar esta prática recomendada: ao otimizar o uso de aceleradores baseados em hardware, é possível reduzir as demandas de infraestrutura física da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Se você precisar de alta capacidade de processamento, poderá se beneficiar do uso de instâncias com computação acelerada, que fornecem acesso a aceleradores de computação baseados em hardware, como unidades de processamento gráfico (GPUs) e matrizes de portas programáveis em

campo (FPGAs). Esses aceleradores de hardware executam certas funções, como processamento gráfico ou correspondência de padrões de dados, com mais eficiência do que alternativas baseadas em CPU. Muitas workloads aceleradas, como renderização, transcodificação e machine learning, são altamente variáveis em termos de uso de recursos. Execute esse hardware apenas pelo tempo necessário e desative-o com automação quando não precisar mais dele para minimizar o consumo de recursos.

Etapas de implementação

- Identifique quais instâncias com computação acelerada podem atender às suas necessidades.
- Para workloads de machine learning, utilize hardware específico para sua workload, como AWS Trainium, AWS Inferentia e Amazon EC2 DL1. AWS Instâncias do Inferentia, como instâncias Inf2, oferecem performance até 50% melhor por watt em relação a instâncias comparáveis do Amazon EC2.
- Colete métricas de uso para suas instâncias com computação acelerada. Por exemplo, você pode usar o agente do CloudWatch para coletar métricas como `utilization_gpu` e `utilization_memory` e para suas GPUs, conforme mostrado em Coletar métricas de GPU NVIDIA com o Amazon CloudWatch.
- Otimize o código, a operação de rede e as configurações dos aceleradores de hardware para garantir que o hardware subjacente seja totalmente utilizado.
  - Otimizar as configurações da GPU
  - Monitoramento e otimização da GPU na AMI de aprendizado profundo
  - Otimizar a E/S para ajuste de performance da GPU de treinamento de aprendizado profundo no Amazon SageMaker
- Use as mais recentes bibliotecas de alta performance e drivers de GPU.
- Use automação para liberar instâncias de GPU quando não estiverem em uso.

Recursos

Documentos relacionados:

- Computação acelerada
- Vamos arquitetar! Como arquitetar com chips e aceleradores personalizados
- Como faço para escolher o tipo de instância do Amazon EC2 apropriado para minha workload?
- Instâncias VT1 do Amazon EC2

- Escolher o melhor acelerador de IA e compilação de modelos para inferência de visão computacional com o Amazon SageMaker

Vídeos relacionados:

- AWS re:Invent 2021: Como selecionar instâncias de GPU do Amazon EC2 para aprendizado profundo
- AWS Online Tech Talks: Implantar inferência de aprendizado profundo eficiente em termos de custos
- AWS re:Invent 2023: IA de última geração com a AWS e a NVIDIA
- AWS re:Invent 2022 [NOVO LANÇAMENTO!]: Introdução as instâncias Inf2 do Amazon EC2 baseadas no AWS Inferentia2
- AWS re:Invent 2022: Acelere o aprendizado profundo e inove com mais rapidez com o AWS Trainium
- AWS re:Invent 2022: Aprendizado profundo na AWS com a NVIDIA: do treinamento à implantação

Processo e cultura

Pergunta

- SUS 6 Como os processos organizacionais apoiam as metas de sustentabilidade?

SUS 6 Como os processos organizacionais apoiam as metas de sustentabilidade?

Procure oportunidades para reduzir seu impacto na sustentabilidade fazendo mudanças nas suas práticas de desenvolvimento, teste e implantação.

Práticas recomendadas

- SUS06-BP01 Adotar métodos que podem apresentar melhorias na sustentabilidade rapidamente
- SUS06-BP02 Manter a workload atualizada
- SUS06-BP03 Aumentar a utilização de ambientes de compilação
- SUS06-BP04 Usar parques de dispositivos gerenciados para testes

SUS06-BP01 Adotar métodos que podem apresentar melhorias na sustentabilidade rapidamente

Adote métodos e processos para validar possíveis aprimoramentos, minimizar o custo dos testes e fornecer pequenas melhorias.

Práticas comuns que devem ser evitadas:

- A avaliação da sustentabilidade de sua aplicação é uma tarefa que é feita apenas uma vez no início de um projeto.
- Como o processo de lançamento para introduzir pequenas alterações em prol da eficiência dos recursos é muito trabalhoso, sua workload tornou-se ultrapassada.
- Você não tem mecanismos para melhorar a sustentabilidade de sua workload.

Benefícios de implementar esta prática recomendada: ao estabelecer um processo para introduzir e monitorar melhorias de sustentabilidade, você poderá adotar continuamente novos recursos e capacidades, remover problemas e melhorar a eficiência da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Teste e valide as possíveis melhorias de sustentabilidade antes de implantá-las na produção. Considere o custo do teste ao calcular o benefício futuro potencial de uma melhoria. Desenvolva métodos de teste de baixo custo para oferecer pequenas melhorias.

Etapas de implementação

- Entenda e comunique suas metas de sustentabilidade organizacional: entenda suas metas de sustentabilidade organizacional, como redução de carbono ou administração da água. Converta essas metas em requisitos de sustentabilidade para suas workloads na nuvem. Comunique esses requisitos às principais partes interessadas.
- Adicione requisitos de sustentabilidade à sua lista de pendências: adicione requisitos para melhoria da sustentabilidade à sua lista de pendências de desenvolvimento.
- Itere e melhore: use um processo de melhoria iterativo para identificar, avaliar, priorizar, testar e implantar essas melhorias.
- Teste usando o produto mínimo viável (MVP): desenvolva e teste possíveis melhorias usando os componentes representativos mínimos viáveis para reduzir o custo e o impacto ambiental dos testes.

- Racionalize o processo: melhore e otimize continuamente seus processos de desenvolvimento. A título de exemplo, automatize o processo de entrega de software usando pipelines de integração contínua e entrega contínua (CI/CD) a fim de testar e implantar possíveis melhorias para reduzir o nível de esforço e limitar os erros provocados por processos manuais.
- Treinamento e conscientização: realize programas de treinamento para os membros da sua equipe para educá-los em sustentabilidade e como suas atividades afetam suas metas de sustentabilidade organizacional.
- Avalie e ajuste: avalie continuamente o impacto das melhorias e faça ajustes conforme necessário.

Recursos

Documentos relacionados:

- A AWS viabiliza soluções de sustentabilidade
- Práticas escaláveis de desenvolvimento ágil baseadas no AWS CodeCommit

Vídeos relacionados:

- AWS re:Invent 2023: Arquitetura sustentável: passado, presente e futuro
- AWS re:Invent 2022: Como entregar arquiteturas sustentáveis e de alta performance
- AWS re:Invent 2022: Arquitetar de forma sustentável e reduzir sua pegada de carbono da AWS
- AWS re:Invent 2022: Sustentabilidade na infraestrutura global da AWS
- AWS re:Invent 2023: Novidades em observabilidade e operações na AWS

Exemplos relacionados:

- Laboratório do Well-Architected: Transformar relatórios de custo e uso em relatórios de eficiência

SUS06-BP02 Manter a workload atualizada

Mantenha sua workload atualizada para adotar recursos eficientes, eliminar problemas e melhorar a eficiência geral da workload.

Práticas comuns que devem ser evitadas:

- Você pressupõe que sua arquitetura atual é estática e não será atualizada ao longo do tempo.

- Você não tem nenhum sistema ou ritmo regular para avaliar se software ou pacotes atualizados são compatíveis com sua workload.

Benefícios de implementar esta prática recomendada: ao estabelecer um processo para manter a workload atualizada, você poderá adotar novos recursos e capacidades, resolver problemas e aumentar a eficiência da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Sistemas operacionais, runtimes, middleware, bibliotecas e aplicações atualizados podem melhorar a eficiência da workload e facilitar a adoção de tecnologias mais eficientes. Um software atualizado também pode incluir recursos para medir o impacto na sustentabilidade da workload com maior precisão, pois os fornecedores oferecem recursos para atender às suas próprias metas de sustentabilidade. Adote um ritmo regular para manter a workload atualizada com os recursos e versões mais recentes.

Etapas de implementação

- Defina um processo: use um processo e um cronograma para avaliar novos recursos ou instâncias para sua workload. Aproveite a agilidade da nuvem para testar rapidamente como novos recursos podem melhorar a workload com o objetivo de:
  - Reduzir impactos de sustentabilidade.
  - Obter eficiências de performance.
  - Remover barreiras a melhorias planejadas.
  - Aumentar sua capacidade de medir e gerenciar impactos na sustentabilidade.
- Faça o inventário da workload: faça o inventário de software e arquitetura da workload e identifique os componentes que precisam ser atualizados.
  - É possível usar o AWS Systems Manager Inventory para coletar metadados de sistema operacional (SO), aplicação e instância das instâncias do Amazon EC2 e entender rapidamente quais instâncias executam o software e as configurações exigidas pela política de software e quais instâncias precisam ser atualizadas.
- Avalie a nova atualização: entenda como atualizar os componentes da sua workload.

Componente de Workload	Como atualizar
Imagens de máquina	Use o EC2 Image Builder para gerenciar atualizações em imagens de máquina da Amazon (AMIs) para imagens de servidores Linux ou Windows.
Imagens de contêiner	Use o Amazon Elastic Container Registry (Amazon ECR) com seu pipeline atual para gerenciar as imagens do Amazon Elastic Container Service (Amazon ECS).
AWS Lambda	O AWS Lambda inclui recursos de gerenciamento de versão.

- Use automação: automatize as atualizações para reduzir o nível de esforço para implantar novos recursos e limitar erros causados por processos manuais.
  - É possível usar CI/CD para atualizar automaticamente AMIs, imagens de contêiner e outros artefatos relacionados à aplicação de nuvem.
  - Você pode usar ferramentas como o AWS Systems Manager Patch Manager para automatizar o processo de atualizações do sistema e agendar a atividade usando as Janelas de Manutenção do AWS Systems Manager.

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- Novidades da AWS
- Ferramentas de desenvolvedor AWS

Vídeos relacionados:

- AWS re:Invent 2022: Otimizar suas workloads da AWS com a orientação de práticas recomendadas
- All Things Patch: AWS Systems Manager

Exemplos relacionados:

- Laboratórios do Well-Architected: Gerenciamento de inventário e patches
- Laboratório: AWS Systems Manager

SUS06-BP03 Aumentar a utilização de ambientes de compilação

Aumente a utilização dos recursos para desenvolver, testar e compilar suas workloads.

Práticas comuns que devem ser evitadas:

- Você provisiona ou encerra manualmente seus ambientes de compilação.
- Você mantém seus ambientes de compilação em execução independentemente de atividades de teste, compilação ou lançamento (por exemplo, execução de um ambiente fora do horário de expediente dos membros de sua equipe de desenvolvimento).
- Você provisiona recursos em excesso para seus ambientes de compilação.

Benefícios de implementar esta prática recomendada: ao aumentar a utilização de ambientes de criação, você pode melhorar a eficiência geral da workload na nuvem e, ao mesmo tempo, alocar os recursos para que os criadores desenvolvam, testem e criem com eficiência.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Use a automação e a infraestrutura como código para ativar ambientes de compilação quando necessário e desativá-los quando não forem usados. Um padrão comum é programar períodos de disponibilidade que coincidam com as horas de trabalho dos membros da equipe de desenvolvimento. A configuração dos ambientes de teste deve ser bem semelhante à do ambiente de produção. Entretanto, procure oportunidades para usar tipos de instância com capacidade de expansão, instâncias spot do Amazon EC2, serviços de banco de dados com ajuste de escala automático, contêineres e tecnologias sem servidor para alinhar a capacidade de desenvolvimento e teste ao uso. Limite o volume de dados apenas para atender os requisitos de teste. Ao usar dados de produção no teste, explore possibilidades para compartilhar os dados da produção em vez de movimentá-los.

Etapas de implementação

- Use infraestrutura como código: use a infraestrutura como código para provisionar os ambientes de compilação.
- Use automação: use automação para gerenciar o ciclo de vida de seus ambientes de desenvolvimento e teste e maximizar a eficiência dos recursos de compilação.
- Maximize a utilização: use estratégias para maximizar a utilização de seus ambientes de desenvolvimento e teste.
  - Use ambientes representativos mínimos viáveis para desenvolver e testar possíveis melhorias.
  - Utilize tecnologias sem servidor, se possível.
  - Use instâncias sob demanda para complementar os dispositivos de desenvolvedor.
  - Use tipos de instância com capacidade de expansão, instâncias spot e outras tecnologias para alinhar a capacidade de compilação com o uso.
  - Adote serviços de nuvem nativos para acesso seguro ao shell de instância em vez de implantar frotas de hosts bastion.
  - Escale automaticamente seus recursos de compilação de acordo com seus trabalhos de compilação.

Recursos

Documentos relacionados:

- Gerenciador de sessões do AWS Systems Manager
- Instâncias de performance expansível do Amazon EC2
- O que é AWS CloudFormation?
- O que é o AWS CodeBuild?
- Agendador de instâncias na AWS

Vídeos relacionados:

- AWS re:Invent 2023: Integração e entrega contínuas para AWS

SUS06-BP04 Usar parques de dispositivos gerenciados para testes

Use parques de dispositivos gerenciados para testar com eficiência um novo recurso em um conjunto representativo de hardware.

Práticas comuns que devem ser evitadas:

- Você testa e implanta manualmente sua aplicação em dispositivos físicos individuais.
- Você não usa o serviço de testes de aplicação para testar e interagir com suas aplicações (por exemplo, Android, iOS e aplicações Web) em dispositivos físicos reais.

Benefícios de implementar esta prática recomendada: o uso de parques de dispositivos gerenciados para testar aplicações habilitadas para a nuvem oferece vários benefícios:

- Eles contam com recursos mais eficientes para testar a aplicação em uma ampla variedade de dispositivos.
- Eles eliminam a necessidade de infraestrutura interna para testes.
- Eles oferecem diversos tipos de dispositivo, incluindo hardware mais antigo e menos conhecido, eliminando a necessidade de atualizações de dispositivo desnecessárias.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Usar parques de dispositivos gerenciados pode ajudar a otimizar o processo de testes de novos recursos em um conjunto representativo de hardware. Os parques de dispositivos gerenciados oferecem diversos tipos de dispositivo, incluindo hardware mais antigo e menos conhecido, e evita o impacto sobre a sustentabilidade por parte do cliente devido a atualizações desnecessárias de dispositivo.

Etapas de implementação

- Defina seus requisitos de testes: defina seus requisitos e plano de testes (como tipo de teste, sistemas operacionais e programação dos testes).
  - O Amazon CloudWatch RUM pode ser usado para coletar e analisar dados do lado do cliente e definir seu plano de teste.
- Selecione um parque de dispositivos gerenciados: selecione um parque de dispositivos gerenciados capaz de suportar seus requisitos de teste. Por exemplo, é possível usar o AWS Device Farm para testar e entender o impacto das suas alterações em um conjunto representativo de hardware.
- Use automação: use a integração contínua/implantação contínua (CI/CD) para agendar e executar seus testes.

  - Integrar o AWS Device Farm com seu pipeline de CI/CD para executar testes de Selenium em vários navegadores
  - Criar e testar aplicações para iOS e iPadOS com o AWS DevOps e serviços móveis
- Revise e ajuste: revise continuamente os resultados dos testes e faça as melhorias necessárias.

Recursos

Documentos relacionados:

- Lista de dispositivos do AWS Device Farm
- Visualizar o painel do CloudWatch RUM

Vídeos relacionados:

- AWS re:Invent 2023: Melhorar a qualidade da sua aplicação móvel e Web usando o AWS Device Farm
- AWS re:Invent 2021: Otimizar aplicações com base em insights do usuário final com o Amazon CloudWatch RUM

Exemplos relacionados:

- Exemplo de aplicação do AWS Device Farm para Android
- Exemplo de aplicação do AWS Device Farm para iOS
- Testes do Appium Web para AWS Device Farm

Avisos

Os clientes são responsáveis por fazer a própria avaliação independente das informações contidas neste documento. Este documento: (a) é apenas para fins informativos, (b) representa as ofertas e práticas de produtos atuais da AWS, que estão sujeitas a alterações sem aviso prévio e (c) não criam nenhum compromisso ou garantia da AWS e de suas afiliadas, fornecedores ou licenciadores. Os produtos ou serviços da AWS são fornecidos “no estado em que se encontram”, sem garantias, representações ou condições de qualquer tipo, expressas ou implícitas. As responsabilidades e as obrigações da AWS com os seus clientes são controladas por contratos da AWS, e este documento não é parte, nem modifica, qualquer contrato entre a AWS e seus clientes.

Glossário da AWS

Para obter a terminologia mais recente da AWS, consulte o glossário da AWS na Referência do Glossário da AWS.