Uma abordagem em fases para engenharia de desempenho no Nuvem AWS

AWS Orientação prescritiva

AWS Orientação prescritiva: Uma abordagem em fases para engenharia de desempenho no Nuvem AWS

Table of Contents

Uma abordagem em fases para engenharia de desempenho no Nuvem AWS

Amazon Web Services (colaboradores)

Abril de 2024 (histórico do documento)

Este guia descreve as melhores práticas para planejar, criar e habilitar a engenharia de desempenho para cargas de trabalho de aplicativos executadas na Amazon Web Services ()AWS. Ele estabelece quatro pilares para a engenharia de desempenho e sugere abordagens diferentes para atender aos requisitos de desempenho dos aplicativos. Para cada pilar, este guia lista ferramentas e soluções para configurar os testes de desempenho e o ambiente de testes.

O que é engenharia de desempenho?

A engenharia de desempenho abrange as técnicas aplicadas durante o ciclo de vida de desenvolvimento de um sistema para garantir que os requisitos de desempenho não funcionais (como taxa de transferência, latência ou uso de memória) sejam atendidos.

Antes do início do teste de desempenho, você precisa configurar o ambiente de desempenho. Um ambiente de desempenho típico se baseia nos seguintes pilares:

- Geração de dados de teste
- Observabilidade do teste
- Automação de testes
- Relatórios de teste

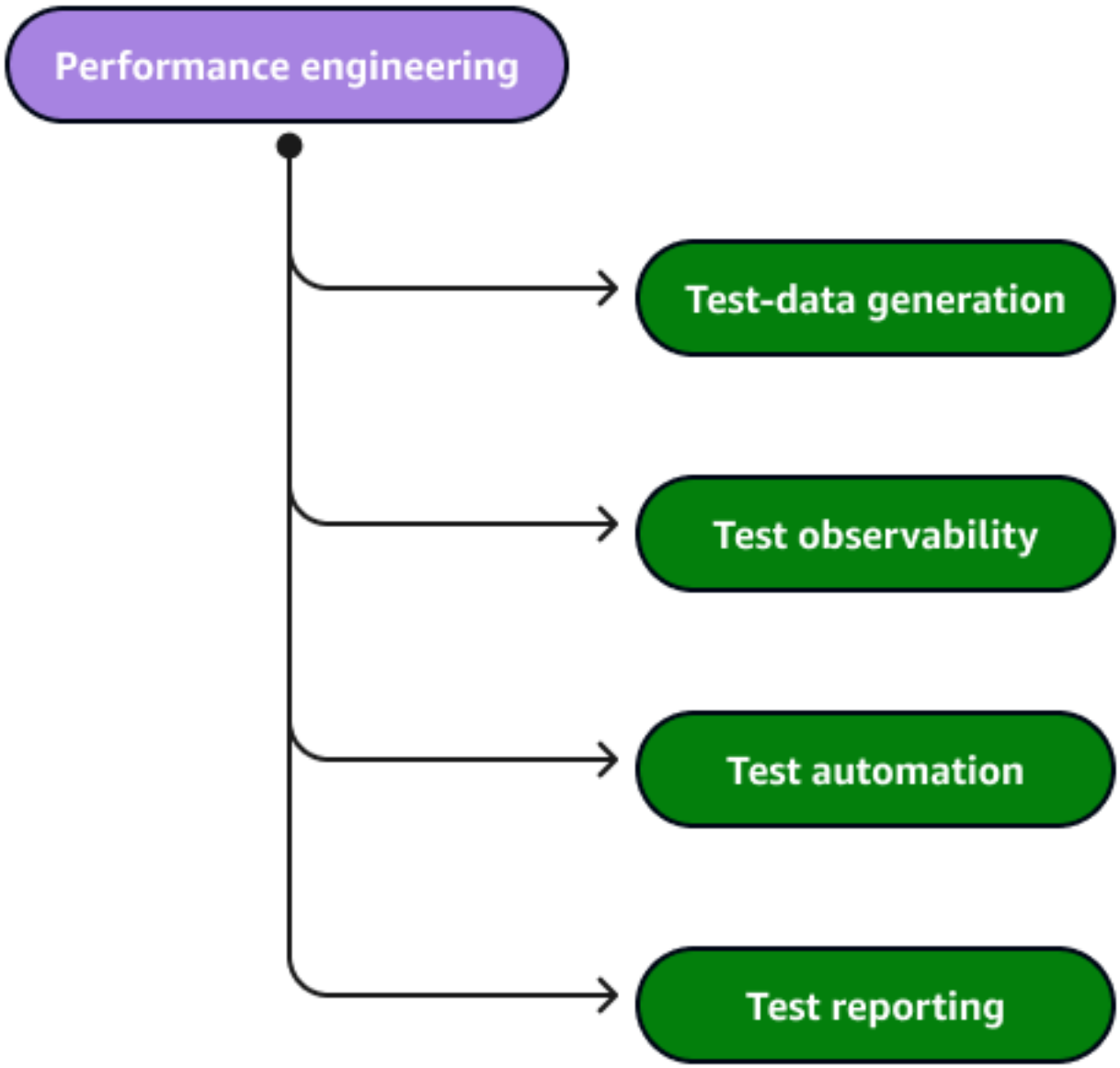
Por que usar a engenharia de desempenho?

A engenharia de desempenho é o processo de otimizar continuamente o desempenho do aplicativo desde o início da fase de projeto. Ele agrega grande valor aos negócios ao evitar o retrabalho e a refatoração do código em um estágio posterior do ciclo de desenvolvimento. Iniciar a engenharia de desempenho na fase de projeto resulta em um aplicativo com melhor desempenho porque o desempenho pode ser considerado no projeto. A engenharia de desempenho requer a participação ativa de arquitetos de sistemas DevOps, desenvolvedores e garantia de qualidade.

Os pilares da engenharia de desempenho

Para permitir uma mentalidade de engenharia de desempenho, é importante construir uma base sólida ao configurar a engenharia de desempenho para o aplicativo. A engenharia de desempenho exige a criação de quatro pilares principais:

- Geração de dados de teste — engenheiros de desempenho configuram ferramentas para gerar os dados de teste.
- Teste a observabilidade — Os engenheiros de desempenho configuram o ambiente de observabilidade para garantir que a execução do desempenho possa ser registrada e rastreada e que os recursos que manipulam as cargas sejam monitorados.
- Automação de testes — Os engenheiros de desempenho desenvolvem testes automatizados que simulam o tráfego do usuário e a carga do sistema usando ferramentas como Apache JMeter ou ghz.
- Relatórios de teste — Os dados são coletados sobre a configuração de cada teste executado junto com os resultados de desempenho. Os dados permitem correlacionar as alterações de configuração ao desempenho e fornecem informações valiosas.



A incorporação desses pilares incentivará a mentalidade de desempenho a partir das fases iniciais do design. Isso ajudará a evitar alterações no aplicativo ou no ambiente em fases posteriores de desenvolvimento e teste.

Geração de dados de teste

A geração de dados de teste envolve gerar e manter uma grande quantidade de dados para executar o caso de teste de desempenho. Esses dados gerados atuam como uma entrada para os casos de teste para que o aplicativo possa ser testado em um conjunto diversificado de dados.

Muitas vezes, gerar dados de teste é um processo complexo. No entanto, usar um conjunto de dados mal criado pode levar a um comportamento imprevisível do aplicativo no ambiente de produção. A geração de dados de teste para testes de desempenho difere das abordagens tradicionais de geração de dados de teste. Isso requer cenários reais, e a maioria dos clientes quer testar suas cargas de trabalho com dados semelhantes aos dados reais de produção. Os dados de teste gerados geralmente também precisam ser redefinidos ou atualizados para seu estado original após cada execução de teste, o que aumenta o tempo e o esforço.

A geração de dados de teste inclui as seguintes considerações principais:

- Precisão — A precisão dos dados é importante em todos os aspectos do teste. Dados imprecisos criam resultados imprecisos. Por exemplo, quando uma transação com cartão de crédito é gerada, ela não deve ser para uma data futura.
- Validade — Os dados devem ser válidos para o caso de uso. Por exemplo, ao testar transações com cartão de crédito, não é aconselhável gerar 10.000 transações por usuário por dia, pois isso se desvia significativamente do cenário de caso de uso válido.
- Automação — A automação da geração de dados de teste pode trazer benefícios de tempo e esforço. Isso também leva a uma automação de testes eficaz. A geração manual de dados de teste pode ter consequências em relação aos requisitos de qualidade e tempo.

  Existem diferentes mecanismos que podem ser adotados com base nos seguintes casos de uso:

  - Orientado por API — Nesse caso, o desenvolvedor fornece uma API de geração de dados de teste que o testador pode consumir para gerar dados. Usando ferramentas de teste JMeter, como, os testadores podem escalar a geração de dados usando uma API comercial. Por exemplo, se você tiver uma API para adicionar um usuário, poderá usar a mesma API para criar centenas de usuários com perfis diferentes. Da mesma forma, você pode excluir os usuários chamando a operação de exclusão da API. Para aplicativos complexos de fluxo de trabalho, o desenvolvedor pode fornecer uma API composta que pode gerar conjuntos de dados em diferentes componentes. Usando essa abordagem, os testadores podem escrever automação para gerar e excluir os conjuntos de dados com base em seus requisitos.

    No entanto, se o sistema for complexo ou o tempo de resposta da API por invocação for alto, pode levar muito tempo para configurar e eliminar os dados.
  - Orientado por instruções SQL — Uma abordagem alternativa é usar instruções SQL de back-end para gerar um grande volume de dados. O desenvolvedor pode fornecer instruções SQL baseadas em modelos para geração de dados de teste. Os testadores podem consumir as instruções para preencher os dados ou criar scripts de wrapper sobre essas instruções para

automatizar a geração de dados de teste. Usando essa abordagem, os testadores podem preencher e eliminar dados muito rapidamente se os dados precisarem ser redefinidos após a conclusão do teste. No entanto, essa abordagem requer acesso direto ao banco de dados do aplicativo, o que pode não ser possível em um ambiente seguro típico. Além disso, consultas inválidas podem resultar em preenchimento incorreto de dados, o que pode produzir resultados distorcidos. Os desenvolvedores também devem atualizar continuamente as instruções SQL no código do aplicativo para refletir as alterações feitas no aplicativo ao longo do tempo.

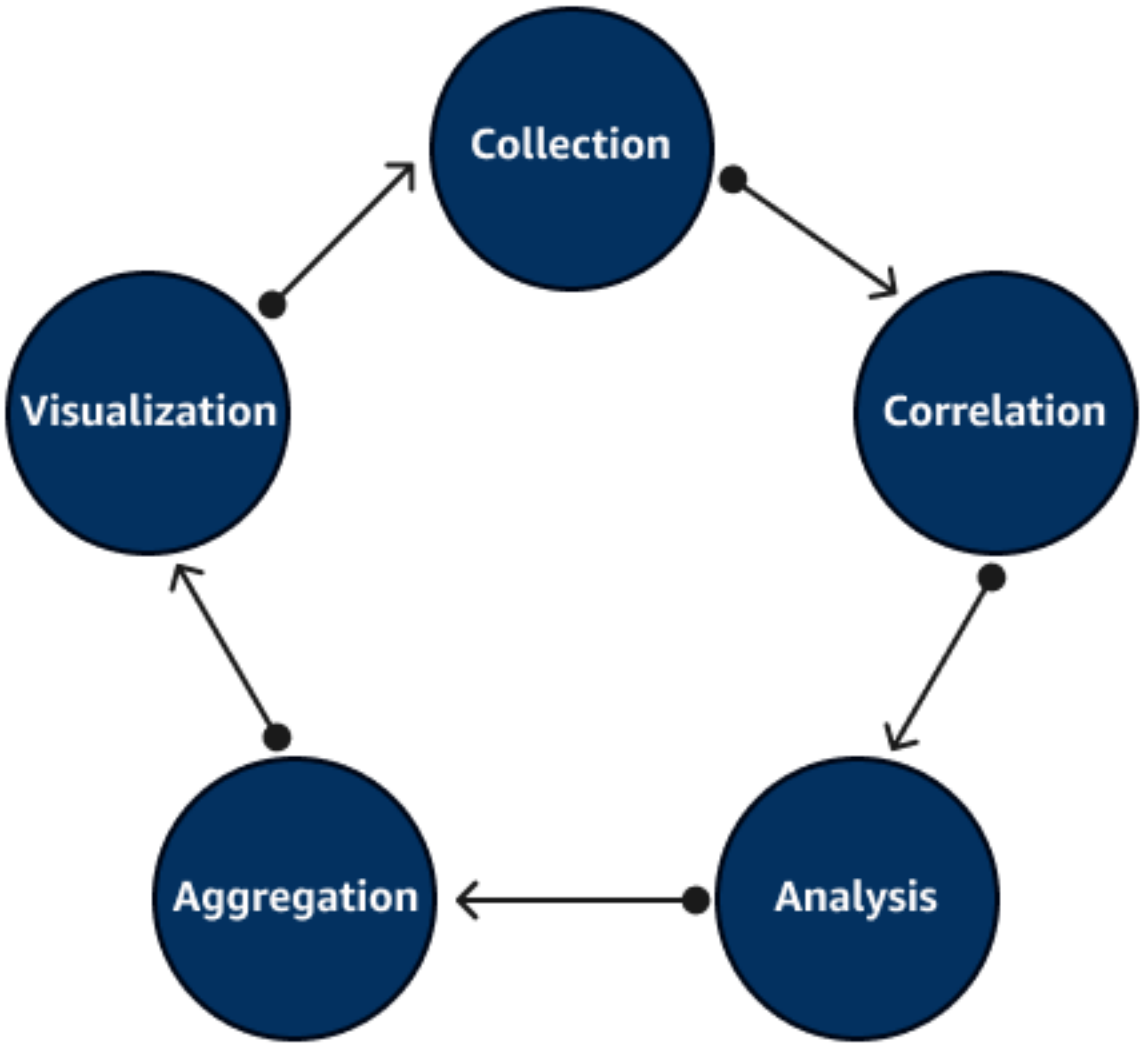
Ferramentas de geração de dados de teste

A AWS fornece ferramentas personalizadas nativas que você pode usar para geração de dados de teste:

- Gerador de dados do Amazon Kinesis — O Amazon Kinesis Data Generator (KDG) simplifica a tarefa de gerar dados e enviá-los para o Amazon Kinesis. A ferramenta fornece uma interface de usuário amigável que é executada diretamente no seu navegador. Para obter mais informações e uma implementação de referência, consulte a postagem do blog Teste sua solução de dados de streaming com o novo Amazon Kinesis Data Generator.
- AWS Glue Gerador de dados de AWS Glue teste — O gerador de dados de teste fornece uma estrutura configurável para geração de dados de teste usando AWS Glue PySpark trabalhos sem servidor. A descrição necessária dos dados de teste é totalmente configurável por meio de um arquivo de configuração YAML. Para obter mais informações e uma implementação de referência, consulte o GitHub repositório AWS Glue Test Data Generator.

Observabilidade do teste

A observabilidade do teste oferece suporte à coleta, correlação, agregação e análise da telemetria em sua rede, infraestrutura e aplicativos durante a execução do teste de desempenho. Você obtém informações completas sobre o comportamento, o desempenho e a integridade do seu sistema. Esses insights ajudam você a detectar, investigar e corrigir problemas com mais rapidez. Ao adicionar inteligência artificial e aprendizado de máquina, você pode reagir, prever e evitar problemas de forma proativa.



A observabilidade depende do registro, monitoramento e rastreamento. A responsabilidade de implementar essas atividades com sucesso abrange as equipes de aplicativos e infraestrutura.

No início da fase de design, as equipes de aplicativos devem entender o estado atual de sua pilha de observabilidade, incluindo registro, monitoramento e rastreamento. Eles podem então escolher ferramentas que se integrem com mais facilidade à pilha de observabilidade.

Da mesma forma, a equipe de infraestrutura é responsável por gerenciar e escalar a infraestrutura de observabilidade.

Considere os seguintes aspectos com relação à observabilidade do teste:

- Disponibilidade de registros e rastreamentos de aplicativos
- Correlação de registros e traços

- Disponibilidade de nós, contêineres e métricas de aplicativos
- Automação para configurar e atualizar a infraestrutura de observabilidade sob demanda
- Capacidade de visualizar a telemetria
- Dimensionamento da infraestrutura de observabilidade

Registro em log

O registro é o processo de manter dados sobre eventos que ocorrem em um sistema. O registro pode incluir problemas, erros ou informações sobre a operação atual. Os registros podem ser classificados em diferentes tipos, como os seguintes:

- Registro de eventos
- Registro do servidor
- Registro do sistema
- Registros de autorização e acesso
- Logs de auditoria

Um desenvolvedor pode pesquisar os registros em busca de códigos ou padrões de erro específicos, filtrá-los com base em campos específicos ou arquivá-los com segurança para análise futura. Os registros ajudam o desenvolvedor a realizar a análise da causa raiz dos problemas de desempenho e também a correlacionar os componentes do sistema.

A criação de uma solução de registro eficaz envolve uma estreita coordenação entre as equipes de aplicativos e infraestrutura. Os registros de aplicativos não são úteis, a menos que haja uma infraestrutura de registro escalável que ofereça suporte a casos de uso como análise, filtragem, armazenamento em buffer e correlação de registros. Casos de uso comuns, como gerar um ID de correlação, registrar o tempo de execução de métodos essenciais para os negócios e definir padrões de registro, podem ser simplificados.

Equipe de aplicação

Um desenvolvedor de aplicativos deve garantir que os registros gerados sigam as melhores práticas de registro. As melhores práticas incluem o seguinte:

- Gerando correlação IDs para rastrear solicitações exclusivas
- Registrando o tempo gasto pelos métodos essenciais para os negócios

- Registro em um nível de registro apropriado
- Compartilhando uma biblioteca de registro comum

Ao criar aplicativos que interagem com diferentes microsserviços, use esses princípios de design de registro para simplificar a filtragem e a extração de registros no back-end.

Gerando correlação IDs para rastrear solicitações exclusivas

Quando o aplicativo recebe a solicitação, ele pode verificar se um ID de correlação já está presente no cabeçalho. Se um ID não estiver presente, o aplicativo deverá gerar um ID. Por exemplo, um Application Load Balancer adiciona um cabeçalho chamado. `X-Amzn-Trace-Id` O aplicativo pode usar o cabeçalho para correlacionar a solicitação do balanceador de carga ao aplicativo. Da mesma forma, o aplicativo deve injetar `traceId` se estiver chamando microsserviços dependentes para que os registros gerados por diferentes componentes em um fluxo de solicitação sejam correlacionados.

Registrando o tempo gasto pelos métodos essenciais para os negócios

Quando o aplicativo recebe uma solicitação, ele interage com um componente diferente. O aplicativo deve registrar o tempo gasto com métodos essenciais para os negócios em um padrão definido. Isso pode facilitar a análise dos registros no back-end. Também pode ajudar você a gerar informações úteis a partir dos registros. Você pode usar abordagens como programação orientada a aspectos (AOP) para gerar esses registros para que você possa separar as questões de registro da sua lógica de negócios.

Registro em um nível de registro apropriado

O aplicativo deve gravar registros que tenham uma quantidade útil de informações. Use os níveis de registro para categorizar os eventos de acordo com sua gravidade. Por exemplo, uso `WARNING` e `ERROR` níveis para eventos importantes que precisam ser investigados. Use `INFO` e `DEBUG` para rastreamento detalhado e eventos de alto volume. Configure manipuladores de registros para capturar somente os níveis necessários na produção. Gerar muito registro no `INFO` nível não ajuda e aumenta a pressão na infraestrutura de back-end. `DEBUG`o registro pode ser útil, mas deve ser usado com cautela. O uso de `DEBUG` registros pode gerar um grande volume de dados, por isso não é recomendado em ambientes de teste de desempenho.

Compartilhando uma biblioteca de registro comum

As equipes de aplicativos devem usar uma biblioteca de registro comum, como, por exemplo AWS SDK para Java, com um padrão de registro comum predefinido que os desenvolvedores possam usar como dependências em seus projetos.

Equipe de infraestrutura

DevOps os engenheiros podem reduzir o esforço usando os seguintes princípios de design de registro ao filtrar e extrair registros no back-end. A equipe de infraestrutura deve configurar e oferecer suporte aos seguintes recursos.

Agente de registro

Um agente de registro (remetente de registros) é um programa que lê registros de um local e os envia para outro local. Os agentes de log são usados para ler arquivos de log armazenados em um computador e carregar eventos de log no back-end para centralização.

Os registros são dados não estruturados que precisam ser estruturados antes que você possa obter insights significativos a partir deles. Os agentes de log usam analisadores para ler instruções de log e extrair campos relevantes, como timestamp, nível de log e nome do serviço, e estruturam esses dados em um formato JSON. Ter um agente de registro leve na borda é útil porque leva a uma menor utilização de recursos. O agente de log pode enviar diretamente para o back-end ou usar um encaminhador de log intermediário que envia os dados para o back-end. Usar um encaminhador de registros descarrega o trabalho dos agentes de registro na origem.

Analisador de registros

Um analisador de registros converte os registros não estruturados em registros estruturados. Os analisadores do agente de log também enriquecem os registros adicionando metadados. A análise dos dados pode ser feita na fonte (final do aplicativo) ou centralmente. O esquema para armazenar os registros deve ser extensível para que você possa adicionar novos campos. Recomendamos o uso de formatos de log padrão, como JSON. No entanto, em alguns casos, os registros devem ser transformados em formatos JSON para uma melhor pesquisa. Escrever a expressão correta do analisador permite uma transformação eficiente.

Backend de registros

Um serviço de back-end de registros coleta, ingere e visualiza dados de log de várias fontes. O agente de log pode gravar diretamente no back-end ou usar um encaminhador de log intermediário. Durante o teste de desempenho, certifique-se de armazenar os registros para que possam ser pesquisados posteriormente. Armazene os registros no back-end separadamente para cada aplicativo. Por exemplo, use um índice dedicado para um aplicativo e use o padrão de índice para pesquisar registros espalhados por diferentes aplicativos relacionados. Recomendamos salvar pelo menos 7 dias de dados para pesquisar registros. No entanto, armazenar os dados por um período

maior pode resultar em custos de armazenamento desnecessários. Como um grande volume de registros é gerado durante o teste de desempenho, é importante que a infraestrutura de registro escale e dimensione corretamente o back-end de registro.

Visualização de registros

Para obter insights significativos e acionáveis dos registros do aplicativo, use ferramentas de visualização dedicadas para processar e transformar os dados brutos do registro em representações gráficas. Visualizações como tabelas, gráficos e painéis podem ajudar a descobrir tendências, padrões e anomalias que podem não ser facilmente aparentes ao analisar os registros brutos.

Os principais benefícios do uso de ferramentas de visualização incluem a capacidade de correlacionar dados em vários sistemas e aplicativos para identificar dependências e gargalos. Os painéis interativos oferecem suporte ao detalhamento dos dados em diferentes níveis de granularidade para solucionar problemas ou identificar tendências de uso. Plataformas especializadas de visualização de dados fornecem recursos como análise, alertas e compartilhamento de dados que podem aprimorar o monitoramento e a análise.

Ao usar o poder da visualização de dados nos registros de aplicativos, as equipes de desenvolvimento e operações podem obter visibilidade do desempenho do sistema e do aplicativo. Os insights derivados podem ser usados para diversas finalidades, incluindo otimizar a eficiência, melhorar a experiência do usuário, aprimorar a segurança e o planejamento da capacidade. O resultado final são painéis personalizados para várias partes interessadas, fornecendo at-a-glance visualizações que resumem os dados de log em informações úteis e perspicazes.

Automatizando a infraestrutura de registro

Como aplicativos diferentes têm requisitos diferentes, é importante automatizar a instalação e as operações da infraestrutura de registro. Use ferramentas de infraestrutura como código (IaC) para provisionar o back-end da infraestrutura de registro. Em seguida, você pode provisionar a infraestrutura de registro como um serviço compartilhado ou como uma implantação independente sob medida para um aplicativo específico.

Recomendamos que os desenvolvedores usem pipelines de entrega contínua (CD) para automatizar o seguinte:

- Implante a infraestrutura de registro sob demanda e desmonte-a quando não for necessária.
- Implante agentes de log em diferentes alvos.
- Implemente configurações de analisador e encaminhador de registros.

- Implemente painéis de aplicativos.

Ferramentas de registro

AWS fornece serviços nativos de registro, alarme e painel de controle. Os recursos a seguir são populares Serviços da AWS e são para registro em log:

- O Amazon OpenSearch Service ajuda as organizações a coletar, ingerir e visualizar dados de log de várias fontes. Para obter mais informações, consulte Registro centralizado com OpenSearch.
- O Amazon CloudWatch Agent e AWS o Fluent Bit são os agentes de log mais populares do mercado AWS. Para obter informações sobre como usar o CloudWatch agente com o Amazon CloudWatch Logs Insights, consulte a postagem do blog Simplificando os logs do servidor Apache com o Amazon CloudWatch Logs Insights. AWS Para a implementação de referência do Fluent Bit, consulte a postagem do blog Registro centralizado de contêineres com o Fluent Bit.

Monitoramento

O monitoramento é o processo de coletar métricas diferentes, como CPU e memória, e armazená-las em um banco de dados de séries temporais, como o Amazon Managed Service for Prometheus. O sistema de monitoramento pode ser baseado em push ou baseado em pull. Em sistemas baseados em push, a fonte envia métricas periodicamente para o banco de dados de séries temporais. Em sistemas baseados em pull, o raspador coleta métricas de várias fontes e as armazena no banco de dados de séries temporais. Os desenvolvedores podem analisar as métricas, filtrar as métricas e traçá-las ao longo do tempo para visualizar o desempenho. A implementação bem-sucedida do monitoramento pode ser dividida em duas grandes áreas: aplicação e infraestrutura.

Para desenvolvedores de aplicativos, as seguintes métricas são essenciais:

- Latência — O tempo gasto para receber uma resposta
- Taxa de transferência de solicitações — O número total de solicitações processadas por segundo
- Taxa de erro da solicitação — O número total de erros

Capture a utilização de recursos, a saturação e as contagens de erros de cada recurso (como o contêiner do aplicativo, o banco de dados) envolvido na transação comercial. Por exemplo, ao monitorar o uso da CPU, você pode monitorar a utilização média da CPU, a carga média e a carga máxima durante a execução do teste de desempenho. Quando um recurso atinge a

saturação durante o teste de estresse, mas pode não atingir a saturação durante uma execução de desempenho por um curto período de tempo.

Metrics

Os aplicativos podem usar atuadores diferentes, como atuadores de mola, para monitorar suas aplicações. Essas bibliotecas de nível de produção geralmente expõem um endpoint REST para monitorar informações sobre os aplicativos em execução. As bibliotecas podem monitorar a infraestrutura subjacente, as plataformas de aplicativos e outros recursos. Se alguma das métricas padrão não atender aos requisitos, o desenvolvedor deverá implementar métricas personalizadas. Métricas personalizadas podem ajudar a monitorar os principais indicadores de desempenho de negócios (KPIs) que não podem ser rastreados por meio de dados de implementações padrão. Por exemplo, talvez você queira monitorar uma operação comercial, como a latência de integração de API de terceiros ou o número total de transações concluídas.

Cardinalidade

Cardinalidade se refere ao número de séries temporais exclusivas de uma métrica. As métricas são rotuladas para fornecer informações adicionais. Por exemplo, um aplicativo baseado em REST que rastreia a contagem de solicitações de uma API específica indica uma cardinalidade de 1. Se você adicionar um rótulo de usuário para identificar a contagem de solicitações por usuário, a cardinalidade aumentará proporcionalmente ao número de usuários. Ao adicionar rótulos que criam cardinalidade, você pode dividir e dividir as métricas por vários grupos. É importante usar os rótulos certos para o caso de uso correto, pois a cardinalidade aumenta o número de séries de métricas no banco de dados de séries temporais de monitoramento de back-end.

Resolução

Em uma configuração de monitoramento típica, o aplicativo de monitoramento é configurado para extrair as métricas do aplicativo periodicamente. A periodicidade da raspagem define a granularidade dos dados de monitoramento. As métricas coletadas em intervalos mais curtos tendem a fornecer uma visão mais precisa do desempenho porque há mais pontos de dados disponíveis. No entanto, a carga no banco de dados de séries temporais aumenta à medida que mais entradas são armazenadas. Normalmente, uma granularidade de 60 segundos é a resolução padrão e 1 segundo é a alta resolução.

DevOps equipe

Os desenvolvedores de aplicativos geralmente pedem aos DevOps engenheiros que configurem um ambiente de monitoramento para visualizar as métricas da infraestrutura e dos aplicativos. O DevOps

engenheiro deve configurar um ambiente que seja escalável e ofereça suporte às ferramentas de visualização de dados usadas pelo desenvolvedor do aplicativo. Isso envolve coletar dados de monitoramento de diferentes fontes e enviá-los para um banco de dados central de séries temporais, como o Amazon Managed Service for Prometheus.

Backend de monitoramento

Um serviço de back-end de monitoramento oferece suporte à coleta, armazenamento, consulta e visualização de dados métricos. Normalmente, é um banco de dados de séries temporais, como o Amazon Managed Service for InfluxData Prometheus ou o InfluxDB. Usando um mecanismo de descoberta de serviços, o coletor de monitoramento pode coletar métricas de diferentes fontes e armazená-las. Durante o teste de desempenho, é importante armazenar os dados das métricas para que possam ser pesquisados posteriormente. Recomendamos salvar pelo menos 15 dias de dados para métricas. No entanto, armazenar as métricas por um período mais longo não agrega benefícios significativos e gera custos de armazenamento desnecessários. Como o teste de desempenho pode gerar um grande volume de métricas, é importante que a infraestrutura de métricas seja dimensionada e, ao mesmo tempo, forneça um desempenho rápido de consultas. O serviço de back-end de monitoramento fornece uma linguagem de consulta que pode ser usada para visualizar os dados de métricas.

Visualização

Forneça ferramentas de visualização que possam exibir os dados do aplicativo para fornecer insights significativos. O DevOps engenheiro e o desenvolvedor do aplicativo devem aprender a linguagem de consulta para o back-end de monitoramento e trabalhar em conjunto para gerar um modelo de painel que possa ser reutilizado. Nos painéis, inclua latência e erros, além de exibir a utilização e a saturação dos recursos na infraestrutura e nos recursos do aplicativo.

Automatizando a infraestrutura de monitoramento

Assim como o registro em log, é importante automatizar a instalação e a operação da infraestrutura de monitoramento para que você possa acomodar os diferentes requisitos de diferentes aplicativos. Use ferramentas de IaC para provisionar o back-end da infraestrutura de monitoramento. Em seguida, você pode provisionar a infraestrutura de monitoramento como um serviço compartilhado ou como uma implantação independente sob medida para um aplicativo específico.

Use pipelines de CD para automatizar o seguinte:

- Implante a infraestrutura de monitoramento sob demanda e desmonte-a quando não for necessária.

- Atualize a configuração de monitoramento para filtrar ou agregar métricas.
- Implemente painéis de aplicativos.

Ferramentas de monitoramento

O Amazon Managed Service for Prometheus é um serviço de monitoramento compatível com o Prometheus para infraestrutura de contêineres e métricas de aplicação para contêineres que você pode usar para monitorar com segurança ambientes de contêineres em grande escala. Para obter mais informações, consulte a postagem do blog Getting Started with Amazon Managed Service for Prometheus.

A Amazon CloudWatch fornece monitoramento completo em. AWS CloudWatch oferece suporte a soluções AWS nativas e de código aberto para que você possa entender o que está acontecendo em sua pilha de tecnologia a qualquer momento.

AWS As ferramentas nativas incluem o seguinte:

- CloudWatch Painéis da Amazon
- CloudWatch Container Insights
- CloudWatch métricas
- CloudWatch alarmes

A Amazon CloudWatch oferece recursos específicos que abordam casos de uso específicos, como monitoramento de contêineres por meio CloudWatch do Container Insights. Esses recursos são integrados para CloudWatch que você possa configurar registros, coleta de métricas e monitoramento.

Para seus aplicativos e microsserviços em contêineres, use o Container Insights para coletar, agregar e resumir métricas e registros. O Container Insights está disponível para as plataformas Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS) e Kubernetes na Amazon Elastic Compute Cloud (Amazon EC2). O Container Insights coleta dados como eventos de registro de desempenho no formato métrico incorporado. Essas entradas de eventos de registro de desempenho usam um esquema JSON estruturado que oferece suporte à ingestão e armazenamento de dados de alta cardinalidade em grande escala.

Para obter informações sobre a implementação do Container Insights com o Amazon EKS, consulte a postagem do blog Apresentando o Amazon CloudWatch Container Insights para o Amazon EKS Fargate usando o AWS Distro for. OpenTelemetry

Rastreamento

O rastreamento envolve o uso especializado de registrar informações sobre os processos de um programa. Os insights dos registros podem ajudar os engenheiros a depurar transações individuais e identificar gargalos. O rastreamento pode ser ativado automaticamente ou usando instrumentação manual.

Como um aplicativo se integra a serviços diferentes, é importante identificar o desempenho do aplicativo e dos serviços subjacentes. O traçado funciona com traços e vãos. Um rastreamento é o processo de solicitação completo, e cada rastreamento é composto por extensões. Um intervalo é um intervalo de tempo marcado e é a atividade dentro dos componentes ou serviços individuais de um sistema. Os rastreamentos fornecem uma visão geral do que acontece quando uma solicitação é feita a um aplicativo.

Equipe de aplicação

Os desenvolvedores de aplicativos instrumentam seus aplicativos enviando dados de rastreamento para solicitações de entrada e saída e outros eventos dentro do aplicativo, junto com metadados sobre cada solicitação. Para gerar rastreamentos, um aplicativo deve ser instrumentado para gerar rastreamentos. A instrumentação pode ser automática ou manual.

Instrumentação automática

Você pode coletar telemetria de um aplicativo usando instrumentação automática sem precisar modificar o código-fonte. Agentes de instrumentação automática podem gerar rastros de aplicativos de um aplicativo ou serviço. Normalmente, você usa alterações de configuração para adicionar o agente ou outro mecanismo.

A instrumentação da biblioteca envolve fazer alterações mínimas no código do aplicativo para adicionar instrumentação pré-construída. A instrumentação tem como alvo bibliotecas ou estruturas específicas, como o AWS SDK, clientes Apache HTTP ou clientes SQL.

Instrumentação manual

Nessa abordagem, os desenvolvedores de aplicativos adicionam código de instrumentação ao aplicativo em cada local em que desejam coletar informações de rastreamento. Por exemplo, use programação orientada a aspectos (AOP) para coletar dados AWS X-Ray de rastreamento. Os desenvolvedores podem usar SDKs para instrumentar seus aplicativos.

Amostragem

Os dados de rastreamento geralmente são gerados em grandes volumes. É importante ter um mecanismo para determinar se os dados de rastreamento devem ser exportados ou não. A amostragem é o processo de determinar quais dados devem ser exportados. Isso geralmente é feito para economizar custos. Ao personalizar regras de amostragem, você pode controlar a quantidade de dados registrados. Você também pode alterar o comportamento da amostragem sem alterar e reimplantar seu código. É importante controlar a taxa de amostragem para gerar a quantidade certa de traços.

Os desenvolvedores de aplicativos podem anotar os rastreamentos adicionando metadados como pares de valores-chave. As anotações enriquecem os traços e ajudam a refinar a filtragem no back-end.

DevOps equipe

DevOps Muitas vezes, os engenheiros são solicitados a configurar um ambiente de rastreamento para que o desenvolvedor do aplicativo visualize os rastreamentos da infraestrutura e dos aplicativos. A configuração do ambiente de rastreamento envolve coletar dados de rastreamento de diferentes fontes e enviá-los para um armazenamento central para visualização.

Backend de rastreamento

Um back-end de rastreamento é um serviço AWS X-Ray que coleta dados sobre solicitações atendidas pelo seu aplicativo. Ele fornece ferramentas que você pode usar para visualizar, filtrar e obter informações sobre esses dados para identificar problemas e oportunidades de otimização. Para qualquer solicitação rastreada para seu aplicativo, você pode ver informações detalhadas sobre a solicitação e a resposta e sobre outras chamadas que seu aplicativo faz para AWS recursos downstream, microsserviços, bancos de dados e web. APIs

Automatizando o rastreamento

Como aplicativos diferentes têm requisitos de rastreamento diferentes, é importante automatizar a configuração e a operação da infraestrutura de rastreamento. Use ferramentas de IaC para provisionar o back-end da infraestrutura de rastreamento.

Use pipelines de CD para automatizar o seguinte:

- Implante a infraestrutura de rastreamento sob demanda e destrua-a quando não for necessário.

- Implante a configuração de rastreamento em todos os aplicativos.

Ferramentas de rastreamento

AWS fornece os seguintes serviços para rastreamento e sua visualização associada:

- AWS X-Ray recebe rastreamentos de seu aplicativo, além de rastreamentos de AWS serviços que seu aplicativo usa que já estão integrados ao X-Ray. Existem vários SDKs agentes e ferramentas que podem ser usados para instrumentar seu aplicativo para rastreamento de raios-X. Para obter mais informações, consulte a documentação do AWS X-Ray.

  Os desenvolvedores também podem usar AWS X-Ray SDKs para enviar traços para o X-Ray. AWS X-Ray fornece SDKs Go,Java, Node.jsPython, .NET Ruby e. Cada X-Ray SDK fornece o seguinte:

  - Interceptadores a serem adicionados ao código para rastrear solicitações HTTP recebidas
  - Manipuladores de clientes para AWS instrumentar clientes SDK que seu aplicativo usa para chamar outros serviços AWS
  - Um cliente HTTP para instrumentar chamadas para outros serviços da web HTTP internos e externos

  O X-Ray SDKs também oferece suporte a chamadas de instrumentação para bancos de dados SQL, instrumentação automática de clientes AWS SDK e outros recursos. Em vez de enviar dados de rastreamento diretamente ao X-Ray, os SDK enviam documentos de segmentos JSON a um processo do daemon que escuta o tráfego UDP. O daemon do X-Ray armazena os segmentos em buffer em uma fila e os carrega em lote no X-Ray. Para obter mais informações sobre como instrumentar seu aplicativo usando um X-Ray SDK, consulte a documentação do X-Ray.
- O Amazon OpenSearch Service é um serviço AWS gerenciado para executar e escalar OpenSearch clusters, que pode ser usado para armazenar centralmente registros, métricas e rastreamentos. O plug-in de Observabilidade fornece uma experiência unificada para coletar e monitorar métricas, logs e rastreamentos de fontes de dados comuns. A coleta e o monitoramento de dados em um só lugar fornecem uma end-to-end capacidade de observação completa de toda a sua infraestrutura. Para obter informações sobre a implementação, consulte a documentação do OpenSearch serviço.
- AWS Distro for OpenTelemetry (ADOT) é uma AWS distribuição baseada no projeto Cloud Native Computing Foundation (CNCF). OpenTelemetry Atualmente, o ADOT inclui suporte de instrumentação automática para Java e Python. Além disso, o ADOT oferece suporte à

instrumentação automática de AWS Lambda funções e suas solicitações downstream usando Node.js e Python tempos de execuçãoJava, por meio do ADOT Managed Lambda Layers. Os desenvolvedores podem usar o coletor ADOT para enviar rastreamentos para diferentes back-ends, incluindo o AWS X-Ray Amazon Service. OpenSearch

Para ver um exemplo de referência de como instrumentar seu aplicativo usando o SDK ADOT, consulte a documentação. Para obter um exemplo de referência de como usar o SDK ADOT para enviar dados para o Amazon OpenSearch Service, consulte a documentação do OpenSearch serviço.

Para obter um exemplo de referência de como instrumentar seu aplicativo em execução no Amazon EKS, consulte a postagem do blog Coleta de métricas e rastreamentos usando complementos do Amazon EKS para o AWS Distro for. OpenTelemetry

Automação de testes

Testes automatizados com uma estrutura e ferramentas especializadas podem reduzir a intervenção humana e maximizar a qualidade. O teste de desempenho automatizado não é diferente dos testes de automação, como testes unitários e testes de integração.

Use DevOps pipelines nos diferentes estágios para testes de desempenho.

Os cinco estágios do pipeline de automação de testes são:

1. Configuração — Use as abordagens de dados de teste descritas na seção Geração de dados de teste para esse estágio. Gerar dados de teste realistas é fundamental para obter resultados de teste válidos. Você deve criar cuidadosamente diversos dados de teste que abranjam uma ampla variedade de casos de uso e correspondam de perto aos dados de produção ao vivo. Antes de executar testes de desempenho em grande escala, talvez seja necessário executar testes iniciais para validar os scripts de teste, os ambientes e as ferramentas de monitoramento.
2. Ferramenta de teste — Para realizar o teste de desempenho, selecione uma ferramenta de teste de carga apropriada, como JMeter ou ghz. Considere a melhor opção para suas necessidades comerciais em termos de simulação de cargas de usuários no mundo real.

3. Execução de teste — Com as ferramentas e os ambientes de teste estabelecidos, execute testes de end-to-end desempenho em uma variedade de cargas e durações esperadas do usuário. Durante todo o teste, monitore de perto a integridade do sistema que está sendo testado. Normalmente, esse é um estágio de longa duração. Monitore as taxas de erro para invalidação automática do teste e interrompa o teste se houver muitos erros.

   A ferramenta de teste de carga fornece informações sobre a utilização de recursos, tempos de resposta e possíveis gargalos.
4. Relatórios de teste — colete os resultados do teste junto com a configuração do aplicativo e do teste. Automatize a coleta da configuração do aplicativo, da configuração do teste e dos resultados, o que ajuda a registrar os dados relacionados ao teste de desempenho e armazená-los centralmente. Manter os dados de desempenho de forma centralizada ajuda a fornecer bons insights e auxilia na definição programática de critérios de sucesso para sua empresa.
5. Limpeza — Depois de concluir uma execução de teste de desempenho, redefina o ambiente de teste e os dados para se preparar para as execuções subsequentes. Primeiro, você reverte todas as alterações feitas nos dados de teste durante a execução. Você deve restaurar os bancos de dados e outros armazenamentos de dados ao estado original, revertendo todos os registros novos, atualizados ou excluídos gerados durante o teste.

Você pode reutilizar o pipeline para repetir o teste várias vezes até que os resultados reflitam o desempenho desejado. Você também pode usar o pipeline para validar se as alterações no código não prejudicam o desempenho. Você pode executar testes de validação de código fora do horário comercial e usar os dados de teste e observabilidade disponíveis para solucionar problemas.

As melhores práticas incluem o seguinte:

- Registre a hora de início e término e gere automaticamente URLs para registro. Isso ajuda você a filtrar os dados de observabilidade na janela de tempo apropriada. Sistemas de monitoramento e rastreamento.
- Injete identificadores de teste no cabeçalho ao invocar os testes. Os desenvolvedores de aplicativos podem enriquecer seus dados de registro, monitoramento e rastreamento usando o identificador como filtro no back-end.
- Limite o pipeline a apenas uma execução por vez. A execução de testes simultâneos gera ruído que pode causar confusão durante a solução de problemas. Também é importante executar o teste em um ambiente de desempenho dedicado.

Ferramentas de automação de testes

As ferramentas de teste desempenham um papel importante em qualquer automação de teste. As opções populares para ferramentas de teste de código aberto incluem o seguinte:

- Apache JMeter é o cavalo poderoso experiente. Com o passar dos anos, o Apache se JMeter tornou mais confiável e adicionou recursos. Com a interface gráfica, é possível criar testes complexos sem precisar conhecer uma linguagem de programação. Empresas como a BlazeMeter apoiam o Apache JMeter.
- O K6é uma ferramenta gratuita que oferece suporte, hospedagem da fonte de carga e uma interface Web integrada para organizar, executar e analisar testes de carga.
- O teste de carga do Vegeta segue um conceito diferente. Em vez de definir simultaneidade ou gerar carga em seu sistema, você define uma determinada taxa. A ferramenta então cria essa carga independente dos tempos de resposta do seu sistema.
- Hey e ab, a ferramenta de benchmarking do servidor Apache HTTP, são ferramentas básicas que você pode usar na linha de comando para executar a carga especificada em um único endpoint. Essa é a maneira mais rápida de gerar carga se você tem um servidor para executar as ferramentas. Até mesmo um laptop local funcionará, embora possa não ser poderoso o suficiente para produzir cargas elevadas.
- ghz é um utilitário de linha de comando e um pacote Go para testes de carga e benchmarking de serviços gRPC.

AWS fornece o teste de carga distribuída na AWS solução. A solução cria e simula milhares de usuários conectados gerando registros transacionais em um ritmo constante sem a necessidade de provisionar servidores. Para obter mais informações, consulte a Biblioteca de AWS soluções.

Você pode usar AWS CodePipeline para automatizar o pipeline de testes de desempenho. Para obter mais informações sobre como automatizar seus testes de API usando CodePipeline, consulte o AWS DevOps blog e a AWS documentação.

Relatórios de teste

Os relatórios de teste se referem à coleta, análise e apresentação de dados relacionados ao desempenho de sistemas, aplicativos, serviços ou processos. Envolve a medição de várias métricas e indicadores para avaliar a eficiência, a capacidade de resposta, a confiabilidade e a eficácia geral de um determinado sistema ou componente.

Os relatórios de testes de desempenho envolvem a escolha de métricas relevantes com base no contexto e nos objetivos da análise. As métricas de desempenho comuns incluem tempos de resposta, taxa de transferência, taxas de erro, utilização de recursos (CPU, memória, disco) e latência da rede.

Depois que os dados relacionados ao desempenho forem coletados, eles precisarão ser armazenados em um repositório central. Esses resultados de teste podem vir de diferentes ambientes, aplicativos e ferramentas de teste. Quando você tem várias cargas de trabalho em execução em ambientes diferentes, é difícil reunir dados relacionados ao desempenho e correlacionar esses pontos de dados para tirar conclusões informadas. Recomendamos definir um método padrão para coletar dados de métricas de desempenho usando um repositório central para armazenamento e visualização de dados.

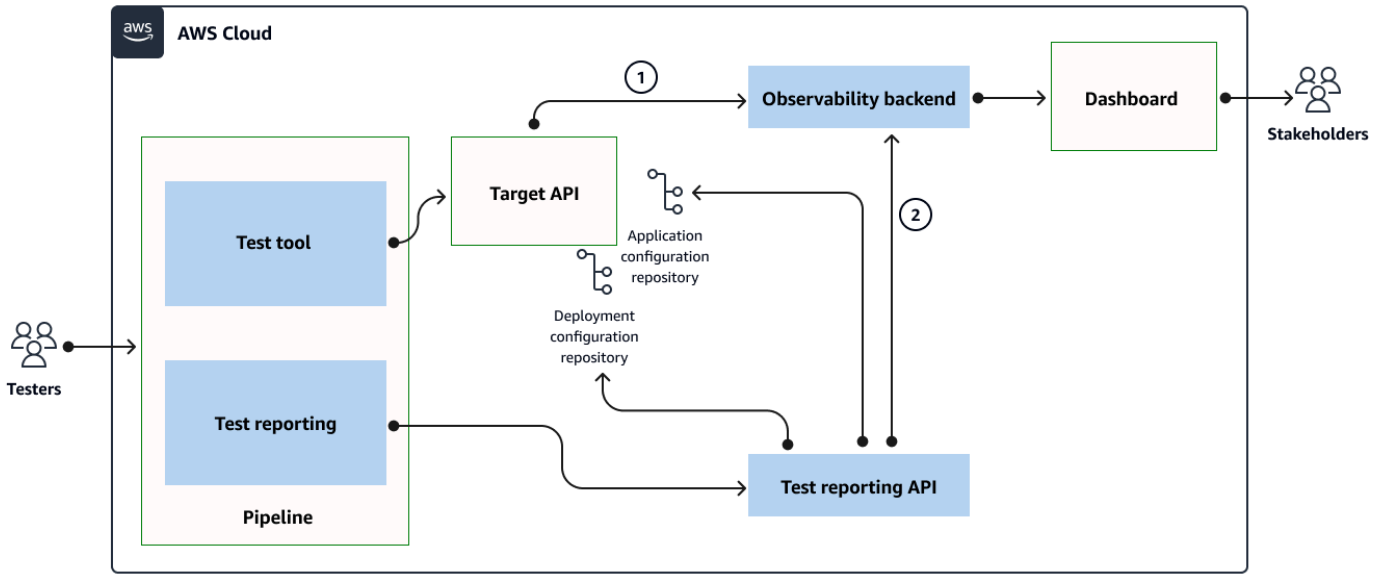
Gravação padronizada

Recomendamos padronizar a forma como as diferentes partes interessadas realizam os testes de desempenho e gravam os dados resultantes em um repositório central. Por exemplo, isso pode assumir a forma de uma API aceitando os resultados e armazenando-os em uma solução de armazenamento persistente. Em situações em que os dados precisam ser obtidos de fontes como GitOps o Amazon Managed Service for Prometheus, a API pode extrair diretamente esses detalhes das fontes especificadas com base em arquivos de esquema que descrevem como extrair os campos das especificações de implantação e das especificações do Kubernetes. Os arquivos de esquema podem usar `JSONPath` expressões ou a Prometheus Query Language (PromQL). Conforme mencionado anteriormente, as métricas coletadas devem ser relevantes para o contexto e as metas da análise de desempenho.

Os dados passados para a API podem incluir detalhes e tags relacionados ao aplicativo e ao ambiente para o qual o teste foi realizado. Isso ajuda na realização de análises nos dados do teste de desempenho.

Pilares da engenharia de desempenho em ação

A arquitetura de referência a seguir demonstra os pilares da engenharia de desempenho para testar uma API específica.



1. Os dados de registro, monitoramento e rastreamento são enviados da API de destino para o back-end.
2. Quando invocada, a API de relatórios de teste envia resultados e informações de configuração para o back-end.

O componente principal é a API ou o aplicativo de destino em teste. A API de destino é sincronizada com o repositório de configuração do aplicativo e o repositório de configuração de implantação de GitOps forma a obter as configurações mais recentes do aplicativo e da infraestrutura. Essa sincronização permite que os testes automatizados sejam executados no estado atual desejado do aplicativo e de sua infraestrutura de suporte, conforme definido nos repositórios Git.

O pipeline de automação de testes automatiza a geração dos dados do teste, a execução do teste e o relatório dos resultados do teste para a API de destino.

A API de destino gera insights de desempenho (métricas, registros e rastreamentos), usando as melhores práticas de observabilidade, e transmite dados de métricas para o back-end de observabilidade.

A API de relatórios de teste coleta todos os dados de relatórios relacionados ao teste (configuração e resultados do teste) e os armazena no back-end de observabilidade.

A agregação de informações de desempenho e dados de relatórios (configuração, resultados de testes) ajuda você a consultar dados relacionados ao desempenho para a API de destino. Por exemplo, você pode perguntar o seguinte:

- Quais são as dez transações mais lentas?
- Qual é o número médio P99, P90 de cada teste?
- Como as configurações das duas execuções de teste se comparam?

Correlacionar casos de teste com resultados, configurações e métricas ao longo de um período ajuda a identificar a melhor configuração e os resultados de desempenho.

Usando esses resultados de teste, você pode tomar decisões mais precisas e baseadas em dados para a API e ter confiança ao levar a API à produção.

Recursos

Serviços da AWS

- Amazon CloudWatch
- AWS CodePipeline
- AWS Distro para OpenTelemetry
- OpenSearch Serviço Amazon
- AWS X-Ray

Implementações

- amazon-kinesis-data-generator
- AWS Glue Gerador de dados de teste
- Teste de carga distribuída em AWS

Publicações no blog

- Registro centralizado de contêineres com FluentBit
- Teste sua solução de streaming de dados com o novo Amazon Kinesis Data Generator
- Apresentando o Amazon CloudWatch Container Insights para Amazon EKS Fargate usando AWS Distro para OpenTelemetry
- Rastreamento de aplicativos no Kubernetes com AWS X-Ray
- Coleta de métricas e traços usando complementos do Amazon EKS para AWS Distro for OpenTelemetry
- Introdução ao Amazon Managed Service para Prometheus

Workshop

- Introdução à AWS observabilidade

AWS Orientação prescritiva

- Aplicativos de teste de carga (guia)

Aplicativos de terceiros

- Apache JMeter
- K6
- Vegeta
- Olá e ab
- ghz

Colaboradores

Os colaboradores deste documento incluem:

- Varun Sharma, consultor principal sênior, AWS
- Akash Kumar, consultor principal sênior, AWS
- Archana Bhatnagar, gerente de consultório, AWS
- Pratik Sharma, Serviços Profissionais II, AWS

Histórico do documento

A tabela a seguir descreve alterações significativas feitas neste guia. Se desejar receber notificações sobre futuras atualizações, inscreva-se em um feed RSS.

Alteração	Descrição	Data
Publicação inicial	—	24 de abril de 2024

AWS Glossário de orientação prescritiva

A seguir estão os termos comumente usados em estratégias, guias e padrões fornecidos pela Orientação AWS Prescritiva. Para sugerir entradas, use o link Fornecer feedback no final do glossário.

Números

7 Rs

Sete estratégias comuns de migração para mover aplicações para a nuvem. Essas estratégias baseiam-se nos 5 Rs identificados pela Gartner em 2011 e consistem em:

- Refactor/re-architect — mova um aplicativo e modifique sua arquitetura aproveitando ao máximo os recursos nativos da nuvem para melhorar a agilidade, o desempenho e a escalabilidade. Isso normalmente envolve a portabilidade do sistema operacional e do banco de dados. Exemplo: migre seu banco de dados Oracle local para a Amazon PostgreSQL-Compatible Aurora Edition.
- Redefinir a plataforma (mover e redefinir [mover e redefinir (lift-and-reshape]):: mova uma aplicação para a nuvem e introduza algum nível de otimização a fim de aproveitar os recursos da nuvem. Exemplo: migrar seu banco de dados Oracle on-premises para o Amazon Relational Database Service (Amazon RDS) para Oracle na Nuvem AWS.
- Recomprar (drop and shop): mude para um produto diferente, normalmente migrando de uma licença tradicional para um modelo SaaS. Exemplo: Migre seu sistema de gerenciamento de relacionamento com o cliente (CRM) para o. Salesforce.com
- Redefinir a hospedagem (mover sem alterações [lift-and-shift]): mover uma aplicação para a nuvem sem fazer nenhuma alteração a fim de aproveitar os recursos da nuvem. Exemplo: migrar seu banco de dados Oracle on-premises para o Oracle em uma instância do EC2 na Nuvem AWS.
- Realocar (mover o hipervisor sem alterações [hypervisor-level lift-and-shift]): mover a infraestrutura para a nuvem sem comprar novo hardware, reescrever aplicações ou modificar suas operações existentes. Você migra servidores de uma plataforma on-premises para um serviço de nuvem para a mesma plataforma. Exemplo: Migrar um Microsoft Hyper-V aplicativo para o. AWS
- Reter (revisitar): mantenha as aplicações em seu ambiente de origem. Isso pode incluir aplicações que exigem grande refatoração, e você deseja adiar esse trabalho para um

momento posterior, e aplicações antigas que você deseja manter porque não há justificativa comercial para migrá-las.

- Retirar: desative ou remova aplicações que não são mais necessárias em seu ambiente de origem.

A

A2A () Agent-to-Agent

Um protocolo com estado para colaboração entre agentes, apoiando a delegação de tarefas e a transferência de estados.

ABAC

Consulte controle de acesso baseado em atributo.

serviços abstraídos

Veja serviços gerenciados.

ACID

Veja atomicidade, consistência, isolamento, durabilidade.

migração ativa-ativa

Um método de migração de banco de dados no qual os bancos de dados de origem e de destino são mantidos em sincronia (por meio de uma ferramenta de replicação bidirecional ou operações de gravação dupla), e ambos os bancos de dados lidam com transações de aplicações conectadas durante a migração. Esse método oferece suporte à migração em lotes pequenos e controlados, em vez de exigir uma substituição única. É mais flexível, mas exige mais trabalho do que a migração ativa-passiva.

migração ativa-passiva

Um método de migração de banco de dados em que os bancos de dados de origem e de destino são mantidos em sincronia, mas somente o banco de dados de origem manipula as transações das aplicações conectadas, enquanto os dados são replicados no banco de dados de destino. O banco de dados de destino não aceita nenhuma transação durante a migração.

Agente

Um sistema de IA que pode raciocinar, planejar e realizar ações de forma autônoma usando ferramentas para atingir metas.

Agente Ops

Práticas operacionais para criar, testar, implantar e executar agentes de IA na produção em grande escala.

AGGREGATE FUNCTION

Uma função SQL que opera em um grupo de linhas e calcula um único valor de retorno para o grupo. Exemplos de funções agregadas incluem `SUM` e `MAX`.

AI

Veja inteligência artificial.

AIOps

Veja operações de inteligência artificial.

anonimização

O processo de excluir permanentemente informações pessoais em um conjunto de dados. A anonimização pode ajudar a proteger a privacidade pessoal. Dados anônimos não são mais considerados dados pessoais.

antipadrões

Uma solução frequentemente usada para um problema recorrente em que a solução é contraproducente, ineficaz ou menos eficaz do que uma alternativa.

controle de aplicações

Uma abordagem de segurança que permite o uso somente de aplicações aprovadas para ajudar a proteger um sistema contra malware.

portfólio de aplicações

Uma coleção de informações detalhadas sobre cada aplicação usada por uma organização, incluindo o custo para criar e manter a aplicação e seu valor comercial. Essas informações são fundamentais para o processo de descoberta e análise de portfólio e ajudam a identificar e priorizar as aplicações a serem migradas, modernizadas e otimizadas.

inteligência artificial (IA)

O campo da ciência da computação que se dedica ao uso de tecnologias de computação para desempenhar funções cognitivas normalmente associadas aos humanos, como aprender, resolver problemas e reconhecer padrões. Para obter mais informações, consulte O que é inteligência artificial?

operações de inteligência artificial (AIOps)

O processo de usar técnicas de machine learning para resolver problemas operacionais, reduzir incidentes operacionais e intervenção humana e aumentar a qualidade do serviço. Para obter mais informações sobre como as AIOps são usadas na estratégia de migração para a AWS , consulte o guia de integração de operações.

criptografia assimétrica

Um algoritmo de criptografia que usa um par de chaves, uma chave pública para criptografia e uma chave privada para descriptografia. É possível compartilhar a chave pública porque ela não é usada na descriptografia, mas o acesso à chave privada deve ser altamente restrito.

atomicidade, consistência, isolamento, durabilidade (ACID)

Um conjunto de propriedades de software que garantem a validade dos dados e a confiabilidade operacional de um banco de dados, mesmo no caso de erros, falhas de energia ou outros problemas.

controle de acesso por atributo (ABAC)

A prática de criar permissões minuciosas com base nos atributos do usuário, como departamento, cargo e nome da equipe. Para obter mais informações, consulte ABAC AWS na documentação AWS Identity and Access Management (IAM).

fonte de dados autorizada

Um local onde você armazena a versão principal dos dados, que é considerada a fonte de informações mais confiável. Você pode copiar dados da fonte de dados autorizada para outros locais com o objetivo de processar ou modificar os dados, como anonimizá-los, redigi-los ou pseudonimizá-los.

Zona de disponibilidade

Um local distinto dentro de um Região da AWS que está isolado de falhas em outras zonas de disponibilidade e fornece conectividade de rede barata e de baixa latência a outras zonas de disponibilidade na mesma região.

AWS Estrutura de adoção da nuvem (AWS CAF)

Uma estrutura de diretrizes e melhores práticas AWS para ajudar as organizações a desenvolver um plano eficiente e eficaz para migrar com sucesso para a nuvem. AWS O CAF organiza a orientação em seis áreas de foco chamadas perspectivas: negócios, pessoas, governança, plataforma, segurança e operações. As perspectivas de negócios, pessoas e governança têm

como foco habilidades e processos de negócios; as perspectivas de plataforma, segurança e operações concentram-se em habilidades e processos técnicos. Por exemplo, a perspectiva das pessoas tem como alvo as partes interessadas que lidam com recursos humanos (RH), funções de pessoal e gerenciamento de pessoal. Nessa perspectiva, o AWS CAF fornece orientação para desenvolvimento, treinamento e comunicação de pessoas para ajudar a preparar a organização para a adoção bem-sucedida da nuvem. Para obter mais informações, consulte o site da AWS CAF e o whitepaper da AWS CAF.

AWS Estrutura de qualificação da carga de trabalho (AWS WQF)

Uma ferramenta que avalia as cargas de trabalho de migração do banco de dados, recomenda estratégias de migração e fornece estimativas de trabalho. AWS O WQF está incluído com AWS Schema Conversion Tool ()AWS SCT. Ela analisa esquemas de banco de dados e objetos de código, código de aplicações, dependências e características de performance, além de fornecer relatórios de avaliação.

B

bot malicioso

Um bot destinado a causar disrupção ou danos a indivíduos ou organizações.

BCP

Veja planejamento de continuidade de negócios

gráfico de comportamento

Uma visualização unificada e interativa do comportamento e das interações de recursos ao longo do tempo. É possível usar um gráfico de comportamento com o Amazon Detective para examinar tentativas de login malsucedidas, chamadas de API suspeitas e ações similares. Para obter mais informações, consulte Dados em um gráfico de comportamento na documentação do Detective.

sistema big-endian

Um sistema que armazena o byte mais significativo antes. Veja também endianness.

classificação binária

Um processo que prevê um resultado binário (uma de duas classes possíveis). Por exemplo, seu modelo de ML pode precisar prever problemas como “Este e-mail é ou não é spam?” ou “Este produto é um livro ou um carro?”

filtro de bloom

Uma estrutura de dados probabilística e eficiente em termos de memória que é usada para testar se um elemento é membro de um conjunto.

blue/green implantação

Uma estratégia de implantação em que você cria dois ambientes separados, mas idênticos. Você executa a versão atual da aplicação em um ambiente (azul) e a nova versão da aplicação no outro ambiente (verde). Essa estratégia ajuda você a reverter rapidamente com o mínimo de impacto.

bot

Uma aplicação de software que executa tarefas automatizadas na internet e simula a atividade ou interação humana. Alguns bots são úteis ou benéficos, como crawlers da web que indexam informações na internet. Outros bots, conhecidos como bots maliciosos, têm como objetivo causar disrupção ou danos a indivíduos ou organizações.

botnet

Redes de bots infectadas por malware e sob o controle de uma única parte, conhecidas como bot herder ou operador de bots. Os botnets são o mecanismo mais conhecido para escalar bots e seu impacto.

ramo

Uma área contida de um repositório de código. A primeira ramificação criada em um repositório é a ramificação principal. Você pode criar uma nova ramificação a partir de uma ramificação existente e, em seguida, desenvolver recursos ou corrigir bugs na nova ramificação. Uma ramificação que você cria para gerar um recurso é comumente chamada de ramificação de recurso. Quando o recurso estiver pronto para lançamento, você mesclará a ramificação do recurso de volta com a ramificação principal. Para obter mais informações, consulte Sobre filiais (GitHub documentação).

Acesso de emergência

Em circunstâncias excepcionais e por meio de um processo aprovado, um meio rápido para um usuário obter acesso a um Conta da AWS que ele normalmente não tem permissão para acessar. Para obter mais informações, consulte o indicador Implementar procedimentos de quebra de vidros na AWS Well-Architected orientação.

estratégia brownfield

A infraestrutura existente em seu ambiente. Ao adotar uma estratégia brownfield para uma arquitetura de sistema, você desenvolve a arquitetura de acordo com as restrições dos sistemas e da infraestrutura atuais. Se estiver expandindo a infraestrutura existente, poderá combinar as estratégias brownfield e greenfield.

cache do buffer

A área da memória em que os dados acessados com mais frequência são armazenados.

capacidade de negócios

O que uma empresa faz para gerar valor (por exemplo, vendas, atendimento ao cliente ou marketing). As arquiteturas de microsserviços e as decisões de desenvolvimento podem ser orientadas por recursos de negócios. Para obter mais informações, consulte a seção Organizados de acordo com as capacidades de negócios do whitepaper Executar microsserviços conteinerizados na AWS.

planejamento de continuidade de negócios (BCP)

Um plano que aborda o impacto potencial de um evento disruptivo, como uma migração em grande escala, nas operações e permite que uma empresa retome as operações rapidamente.

C

CAF

Veja AWS Cloud Adoption Framework.

implantação canário

O lançamento lento e incremental de uma versão para usuários finais. Quando estiver confiante, você implanta a nova versão e substitui a versão atual por completo.

CCoE

Veja Centro de Excelência da Nuvem.

CDC

Veja captura de dados de alteração.

captura de dados de alterações (CDC)

O processo de rastrear alterações em uma fonte de dados, como uma tabela de banco de dados, e registrar metadados sobre a alteração. É possível usar o CDC para várias finalidades, como auditar ou replicar alterações em um sistema de destino para manter a sincronização.

engenharia do caos

Introduzir intencionalmente falhas ou eventos disruptivos para testar a resiliência de um sistema. Você pode usar AWS Fault Injection Service (AWS FIS) para realizar experimentos que estressam suas AWS cargas de trabalho e avaliar sua resposta.

CI/CD

Veja integração e entrega contínuas.

classificação

Um processo de categorização que ajuda a gerar previsões. Os modelos de ML para problemas de classificação predizem um valor discreto. Os valores discretos são sempre diferentes uns dos outros. Por exemplo, um modelo pode precisar avaliar se há ou não um carro em uma imagem.

Desenvolvedor cidadão

Um usuário corporativo que cria aplicativos de IA usando plataformas sem code/low código sem habilidades técnicas especializadas.

criptografia no lado do cliente

Criptografia de dados localmente, antes que o alvo os AWS service (Serviço da AWS) receba.

Centro de Excelência da Nuvem (CCoE)

Uma equipe multidisciplinar que impulsiona os esforços de adoção da nuvem em toda a organização, incluindo o desenvolvimento de práticas recomendas de nuvem, a mobilização de recursos, o estabelecimento de cronogramas de migração e a liderança da organização em transformações em grande escala. Para obter mais informações, consulte as postagens do CCoE no blog de estratégia Nuvem AWS corporativa.

computação em nuvem

A tecnologia de nuvem normalmente usada para armazenamento de dados remoto e gerenciamento de dispositivos de IoT. A computação em nuvem é normalmente conectada à tecnologia de computação de borda.

modelo operacional em nuvem

Em uma organização de TI, o modelo operacional usado para criar, amadurecer e otimizar um ou mais ambientes de nuvem. Para obter mais informações, consulte Criar seu modelo operacional de nuvem.

estágios de adoção da nuvem

As quatro fases pelas quais as organizações normalmente passam ao migrar para a Nuvem AWS:

- Projeto: executar alguns projetos relacionados à nuvem para fins de prova de conceito e aprendizado
- Fundação: realizar investimentos fundamentais para escalar sua adoção da nuvem (por exemplo, criar uma zona de pouso, definir um CCoE, estabelecer um modelo de operações)
- Migração: migrar aplicações individuais
- Re-invention — Otimizando produtos e serviços e inovando na nuvem

Esses estágios foram definidos por Stephen Orban na postagem do blog The Journey Toward Cloud-First & the Stages of Adoption no blog Nuvem AWS Enterprise Strategy. Para obter informações sobre como eles se relacionam com a estratégia de AWS migração, consulte o guia de preparação para migração.

CMDB

Veja banco de dados de gerenciamento de configuração.

repositório de código

Um local onde o código-fonte e outros ativos, como documentação, amostras e scripts, são armazenados e atualizados por meio de processos de controle de versão. Os repositórios de nuvem comuns incluem o GitHub ou o Bitbucket Cloud. Cada versão do código é chamada de ramificação. Em uma estrutura de microsserviços, cada repositório é dedicado a uma única peça de funcionalidade. Um único CI/CD pipeline pode usar vários repositórios.

cache frio

Um cache de buffer que está vazio, não está bem preenchido ou contém dados obsoletos ou irrelevantes. Isso afeta a performance porque a instância do banco de dados deve ler da memória principal ou do disco, um processo que é mais lento do que a leitura do cache do buffer.

dados frios

Dados que raramente são acessados e geralmente são históricos. Ao consultar esse tipo de dados, consultas lentas geralmente são aceitáveis. Mover esses dados para níveis ou classes de armazenamento de baixo desempenho e menos caros pode reduzir os custos.

visão computacional (CV)

Um campo de IA que usa machine learning para analisar e extrair informações de formatos visuais, como vídeos e imagens digitais. Por exemplo, a Amazon SageMaker AI fornece algoritmos de processamento de imagem para CV.

desvio de configuração

Em uma workload, uma alteração de configuração em relação ao estado esperado. Isso pode fazer com que a workload se torne incompatível e, normalmente, é gradual e não intencional.

banco de dados de gerenciamento de configuração (CMDB)

Um repositório que armazena e gerencia informações sobre um banco de dados e seu ambiente de TI, incluindo componentes de hardware e software e suas configurações. Normalmente, os dados de um CMDB são usados no estágio de descoberta e análise do portfólio da migração.

pacote de conformidade

Uma coleção de AWS Config regras e ações de remediação que você pode montar para personalizar suas verificações de conformidade e segurança. Você pode implantar um pacote de conformidade como uma entidade única em uma Conta da AWS região ou em uma organização usando um modelo YAML. Para obter mais informações, consulte Pacotes de conformidade na documentação. AWS Config

integração contínua e entrega contínua (CI/CD)

O processo de automatizar os estágios de origem, criação, teste, preparação e produção do processo de lançamento do software. CI/CD é comumente descrito como um pipeline. CI/CD pode ajudá-lo a automatizar processos, melhorar a produtividade, melhorar a qualidade do código e entregar com mais rapidez. Para obter mais informações, consulte Benefícios da entrega contínua. CD também pode significar implantação contínua. Para obter mais informações, consulte Entrega contínua versus implantação contínua.

CV

Veja visão computacional.

D

dados em repouso

Dados estacionários em sua rede, por exemplo, dados que estão em um armazenamento.

classificação de dados

Um processo para identificar e categorizar os dados em sua rede com base em criticalidade e confidencialidade. É um componente crítico de qualquer estratégia de gerenciamento de riscos de segurança cibernética, pois ajuda a determinar os controles adequados de proteção e retenção para os dados. A classificação de dados é um componente do pilar de segurança na AWS Well-Architected Estrutura. Para obter mais informações, consulte Classificação de dados.

desvio de dados

Uma variação significativa entre os dados de produção e os dados usados para treinar um modelo de ML ou uma alteração significativa nos dados de entrada ao longo do tempo. O desvio de dados pode reduzir a qualidade geral, a precisão e a imparcialidade das previsões do modelo de ML.

dados em trânsito

Dados que estão se movendo ativamente pela sua rede, como entre os recursos da rede.

data mesh

Um framework de arquitetura que fornece propriedade de dados distribuída e descentralizada com gerenciamento e governança centralizados.

minimização de dados

O princípio de coletar e processar apenas os dados estritamente necessários. Praticar a minimização de dados no Nuvem AWS pode reduzir os riscos de privacidade, os custos e a pegada de carbono de sua análise.

perímetro de dados

Um conjunto de proteções preventivas em seu AWS ambiente que ajudam a garantir que somente identidades confiáveis acessem recursos confiáveis das redes esperadas. Para obter mais informações, consulte Construindo um perímetro de dados em. AWS

pré-processamento de dados

A transformação de dados brutos em um formato que seja facilmente analisado por seu modelo de ML. O pré-processamento de dados pode significar a remoção de determinadas colunas ou linhas e o tratamento de valores ausentes, inconsistentes ou duplicados.

proveniência dos dados

O processo de rastrear a origem e o histórico dos dados ao longo de seu ciclo de vida, por exemplo, como os dados foram gerados, transmitidos e armazenados.

titular dos dados

Um indivíduo cujos dados estão sendo coletados e processados.

data warehouse

Um sistema de gerenciamento de dados compatível com business intelligence, como analytics. Os data warehouses geralmente contêm grandes quantidades de dados históricos e geralmente são usados para consultas e análises.

linguagem de definição de dados (DDL)

Instruções ou comandos para criar ou modificar a estrutura de tabelas e objetos em um banco de dados.

linguagem de manipulação de dados (DML)

Instruções ou comandos para modificar (inserir, atualizar e excluir) informações em um banco de dados.

DDL

Veja linguagem de definição de banco de dados.

deep ensemble

A combinação de vários modelos de aprendizado profundo para gerar previsões. Os deep ensembles podem ser usados para produzir uma previsão mais precisa ou para estimar a incerteza nas previsões.

Aprendizado profundo

Um subcampo do ML que usa várias camadas de redes neurais artificiais para identificar o mapeamento entre os dados de entrada e as variáveis-alvo de interesse.

defesa completa

Uma abordagem de segurança da informação na qual uma série de mecanismos e controles de segurança são cuidadosamente distribuídos por toda a rede de computadores para proteger a confidencialidade, a integridade e a disponibilidade da rede e dos dados nela contidos. Ao adotar essa estratégia AWS, você adiciona vários controles em diferentes camadas da AWS Organizations estrutura para ajudar a proteger os recursos. Por exemplo, uma abordagem de defesa aprofundada pode combinar autenticação multifatorial, segmentação de rede e criptografia.

administrador delegado

Em AWS Organizations, um serviço compatível pode registrar uma conta de AWS membro para administrar as contas da organização e gerenciar as permissões desse serviço. Essa conta é chamada de administrador delegado para esse serviço Para obter mais informações e uma lista de serviços compatíveis, consulte Serviços que funcionam com o AWS Organizations na documentação do AWS Organizations .

implantação

O processo de criar uma aplicação, novos recursos ou correções de código disponíveis no ambiente de destino. A implantação envolve a implementação de mudanças em uma base de código e, em seguida, a criação e execução dessa base de código nos ambientes da aplicação

ambiente de desenvolvimento

Veja ambiente.

controle detectivo

Um controle de segurança projetado para detectar, registrar e alertar após a ocorrência de um evento. Esses controles são uma segunda linha de defesa, alertando você sobre eventos de segurança que contornaram os controles preventivos em vigor. Para obter mais informações, consulte Controles detectivos em Como implementar controles de segurança na AWS.

mapeamento do fluxo de valor de desenvolvimento (DVSM)

Um processo usado para identificar e priorizar restrições que afetam negativamente a velocidade e a qualidade em um ciclo de vida de desenvolvimento de software. O DVSM estende o processo de mapeamento do fluxo de valor originalmente projetado para práticas de manufatura enxuta. Ele se concentra nas etapas e equipes necessárias para criar e movimentar valor por meio do processo de desenvolvimento de software.

gêmeo digital

Uma representação virtual de um sistema real, como um prédio, fábrica, equipamento industrial ou linha de produção. Os gêmeos digitais oferecem suporte à manutenção preditiva, ao monitoramento remoto e à otimização da produção.

tabela de dimensões

Em um esquema em estrela, uma tabela menor que contém atributos de dados sobre dados quantitativos em uma tabela de fatos. Os atributos da tabela de dimensões geralmente são campos de texto ou números discretos que se comportam como texto. Esses atributos normalmente são usados para restringir consultas, filtrar e rotular conjuntos de resultados.

desastre

Um evento que impede que uma workload ou sistema cumpra seus objetivos de negócios em seu local principal de implantação. Esses eventos podem ser desastres naturais, falhas técnicas ou o resultado de ações humanas, como configuração incorreta não intencional ou ataque de malware.

Recuperação de desastres (RD)

A estratégia e o processo que você usa para minimizar o tempo de inatividade e a perda de dados causados por um disastre. Para obter mais informações, consulte Recuperação de desastres de cargas de trabalho em AWS: Recuperação na nuvem na AWS Well-Architected estrutura.

DML

Veja linguagem de manipulação de banco de dados.

design orientado por domínio

Uma abordagem ao desenvolvimento de um sistema de software complexo conectando seus componentes aos domínios em evolução, ou principais metas de negócios, atendidos por cada componente. Esse conceito foi introduzido por Eric Evans em seu livro Domain-Driven Design: Tackling Complexity in the Heart of Software (Boston: Addison-Wesley Professional, 2003). Para obter informações sobre como você pode usar o design orientado por domínio com o padrão strangler fig, consulte Modernizando os serviços web legados da Microsoft ASP.NET (ASMX) de forma incremental usando contêineres e o Amazon API Gateway.

DR

Veja recuperação de desastres.

Detecção da oscilação

Rastreamento de desvios de uma configuração de linha de base. Por exemplo, você pode usar AWS CloudFormation para detectar desvios nos recursos do sistema ou AWS Control Tower para detectar mudanças em seu landing zone que possam afetar a conformidade com os requisitos de governança.

DVSM

Veja mapeamento do fluxo de valor de desenvolvimento.

E

EDA

Veja análise exploratória de dados.

EDI

Veja intercâmbio eletrônico de dados.

computação de borda

A tecnologia que aumenta o poder computacional de dispositivos inteligentes nas bordas de uma rede de IoT. Quando comparada com a computação em nuvem, a computação de borda pode reduzir a latência da comunicação e melhorar o tempo de resposta.

intercâmbio eletrônico de dados (EDI)

A troca automatizada de documentos comerciais entre organizações. Para obter mais informações, consulte O que é EDI (Intercâmbio eletrônico de dados)?.

criptografia

Um processo de computação que transforma dados de texto simples, legíveis por humanos, em texto cifrado.

chave de criptografia

Uma sequência criptográfica de bits aleatórios que é gerada por um algoritmo de criptografia. As chaves podem variar em tamanho, e cada chave foi projetada para ser imprevisível e exclusiva.

endianismo

A ordem na qual os bytes são armazenados na memória do computador. Big-endian os sistemas armazenam primeiro o byte mais significativo. Little-endian os sistemas armazenam primeiro o byte menos significativo.

endpoint

Veja endpoint de serviço.

serviço de endpoint

Um serviço que pode ser hospedado em uma nuvem privada virtual (VPC) para ser compartilhado com outros usuários. Você pode criar um serviço de endpoint com AWS PrivateLink e conceder permissões a outros diretores Contas da AWS ou a AWS Identity and Access Management (IAM). Essas contas ou entidades principais podem se conectar ao serviço de endpoint de maneira privada criando endpoints da VPC de interface. Para obter mais informações, consulte Criar um serviço de endpoint na documentação do Amazon Virtual Private Cloud (Amazon VPC).

planejamento de recursos empresariais (ERP)

Um sistema que automatiza e gerencia os principais processos de negócios (como contabilidade, MES e gerenciamento de projetos) para uma empresa.

criptografia envelopada

O processo de criptografar uma chave de criptografia com outra chave de criptografia. Para obter mais informações, consulte Criptografia de envelope na documentação AWS Key Management Service (AWS KMS).

ambiente

Uma instância de uma aplicação em execução. Estes são tipos comuns de ambientes na computação em nuvem:

- ambiente de desenvolvimento: uma instância de uma aplicação em execução que está disponível somente para a equipe principal responsável pela manutenção da aplicação. Ambientes de desenvolvimento são usados para testar mudanças antes de promovê-las para ambientes superiores. Esse tipo de ambiente às vezes é chamado de ambiente de teste.
- ambientes inferiores: todos os ambientes de desenvolvimento para uma aplicação, como aqueles usados para compilações e testes iniciais.
- ambiente de produção: uma instância de uma aplicação em execução que os usuários finais podem acessar. Em um CI/CD pipeline, o ambiente de produção é o último ambiente de implantação.

- ambientes superiores: todos os ambientes que podem ser acessados por usuários que não sejam a equipe principal de desenvolvimento. Isso pode incluir um ambiente de produção, ambientes de pré-produção e ambientes para testes de aceitação do usuário.

epic

Em metodologias ágeis, categorias funcionais que ajudam a organizar e priorizar seu trabalho. Os epics fornecem uma descrição de alto nível dos requisitos e das tarefas de implementação. Por exemplo, os épicos de segurança AWS da CAF incluem gerenciamento de identidade e acesso, controles de detetive, segurança de infraestrutura, proteção de dados e resposta a incidentes. Para obter mais informações sobre epics na estratégia de migração da AWS , consulte o guia de implementação do programa.

ERP

Veja planejamento de recursos empresariais.

análise exploratória de dados (EDA)

O processo de analisar um conjunto de dados para entender suas principais características. Você coleta ou agrega dados e, em seguida, realiza investigações iniciais para encontrar padrões, detectar anomalias e verificar suposições. O EDA é realizado por meio do cálculo de estatísticas resumidas e da criação de visualizações de dados.

F

tabela de fatos

A tabela central em um esquema em estrela. Ela armazena dados quantitativos sobre as operações comerciais. Normalmente, uma tabela de fatos contém dois tipos de colunas: as que contêm medidas e as que contêm uma chave externa para uma tabela de dimensões.

Antecipar-se à falha

Uma filosofia que usa testes frequentes e incrementais para reduzir o ciclo de vida do desenvolvimento. É uma parte essencial de uma abordagem ágil.

delimitação de isolamento contra falhas

No Nuvem AWS, um limite, como uma zona de disponibilidade, Região da AWS um plano de controle ou um plano de dados, que limita o efeito de uma falha e ajuda a melhorar a resiliência das cargas de trabalho. Para obter mais informações, consulte AWS Fault Isolation Boundaries.

ramificação de recursos

Veja ramificação.

recursos

Os dados de entrada usados para fazer uma previsão. Por exemplo, em um contexto de manufatura, os recursos podem ser imagens capturadas periodicamente na linha de fabricação.

importância do recurso

O quanto um recurso é importante para as previsões de um modelo. Isso geralmente é expresso como uma pontuação numérica que pode ser calculada por meio de várias técnicas, como Shapley Additive Explanations (SHAP) e gradientes integrados. Para obter mais informações, consulte Interpretabilidade do modelo de aprendizado de máquina com AWS.

transformação de recursos

O processo de otimizar dados para o processo de ML, incluindo enriquecer dados com fontes adicionais, escalar valores ou extrair vários conjuntos de informações de um único campo de dados. Isso permite que o modelo de ML se beneficie dos dados. Por exemplo, se a data “2021-05-27 00:15:37” for dividida em “2021”, “maio”, “quinta” e “15”, isso poderá ajudar o algoritmo de aprendizado a aprender padrões diferenciados associados a diferentes componentes de dados.

prompt few shot

Fornecer a um LLM um pequeno número de exemplos que demonstram a tarefa e o resultado desejado antes de solicitar que ele execute uma tarefa semelhante. Essa técnica é uma aplicação do aprendizado contextual, em que os modelos aprendem com exemplos (fotos) incorporados aos prompts. Few-shot a solicitação pode ser eficaz para tarefas que exigem formatação, raciocínio ou conhecimento de domínio específicos. Veja também prompts zero-shot.

FGAC

Veja controle de acesso refinado.

Controle de acesso refinado (FGAC)

O uso de várias condições para permitir ou negar uma solicitação de acesso.

migração flash-cut

Um método de migração de banco de dados que usa replicação contínua de dados via captura de dados de alteração para migrar os dados no menor tempo possível, em vez de usar uma abordagem em fases. O objetivo é reduzir ao mínimo o tempo de inatividade.

FM

Veja modelo de base.

modelo de base (FM)

Uma grande rede neural de aprendizado profundo que treina em grandes conjuntos de dados generalizados e não rotulados. Os FMs são capazes de realizar uma ampla variedade de tarefas gerais, como entender a linguagem, gerar texto e imagens e conversar em linguagem natural. Para obter mais informações, consulte O que são modelos de base?.

Gateway FM

Um intermediário centralizado que controla e normaliza o acesso aos modelos de fundação. Também conhecido como gateway LLM.

G

IA generativa

Um subconjunto de modelos de IA que foram treinados em grandes quantidades de dados e que podem usar um simples prompt de texto para criar novos artefatos e conteúdo, como imagens, vídeos, texto e áudio. Para obter mais informações, consulte O que é IA generativa?.

bloqueio geográfico

Veja restrições geográficas.

restrições geográficas (bloqueio geográfico)

Na Amazon CloudFront, uma opção para impedir que usuários em países específicos acessem distribuições de conteúdo. É possível usar uma lista de permissões ou uma lista de bloqueios para especificar países aprovados e banidos. Para obter mais informações, consulte Restringir a distribuição geográfica do seu conteúdo na CloudFront documentação.

Fluxo de trabalho do GitFlow

Uma abordagem na qual ambientes inferiores e superiores usam ramificações diferentes em um repositório de código-fonte. O fluxo de trabalho do Gitflow é considerado legado, e o fluxo de trabalho trunk-based é a abordagem moderna e preferencial.

golden image

Um snapshot de um sistema ou software usado como modelo para implantar novas instâncias desse sistema ou software. Por exemplo, na manufatura, uma golden image pode ser usada para

provisionar software em vários dispositivos e ajudar a melhorar a velocidade, a escalabilidade e a produtividade nas operações de fabricação de dispositivos.

estratégia greenfield

A ausência de infraestrutura existente em um novo ambiente. Ao adotar uma estratégia greenfield para uma arquitetura de sistema, é possível selecionar todas as novas tecnologias sem a restrição da compatibilidade com a infraestrutura existente, também conhecida como brownfield. Se estiver expandindo a infraestrutura existente, poderá combinar as estratégias brownfield e greenfield.

barreira de proteção

Uma regra de alto nível que ajuda a gerenciar recursos, políticas e conformidade em todas as unidades organizacionais (UOs). Barreiras de proteção preventivas impõem políticas para garantir o alinhamento a padrões de conformidade. Elas são implementadas usando políticas de controle de serviço e limites de permissões do IAM. Barreiras de proteção detectivas detectam violações de políticas e problemas de conformidade e geram alertas para remediação. Eles são implementados usando AWS Config, AWS Security Hub CSPM, Amazon GuardDuty AWS Trusted Advisor, Amazon Inspector e verificações personalizadas AWS Lambda .

grades de proteção (IA)

Mecanismos de segurança que filtram, validam e restringem as entradas e saídas dos agentes para ajudar a garantir um comportamento de IA responsável e seguro.

H

HA

Veja alta disponibilidade.

migração heterogênea de bancos de dados

Migrar seu banco de dados de origem para um banco de dados de destino que usa um mecanismo de banco de dados diferente (por exemplo, Oracle para Amazon Aurora). A migração heterogênea geralmente faz parte de um esforço de redefinição da arquitetura, e converter o esquema pode ser uma tarefa complexa. O AWS fornece o AWS SCT para ajudar nas conversões de esquemas.

alta disponibilidade (HA)

A capacidade de uma workload operar continuamente, sem intervenção, em caso de desafios ou desastres. Os sistemas AH são projetados para realizar o failover automático, oferecer consistentemente desempenho de alta qualidade e lidar com diferentes cargas e falhas com impacto mínimo no desempenho.

modernização de historiador

Uma abordagem usada para modernizar e atualizar os sistemas de tecnologia operacional (OT) para melhor atender às necessidades do setor de manufatura. Um historiador é um tipo de banco de dados usado para coletar e armazenar dados de várias fontes em uma fábrica.

dados de hold-out

Uma parte dos dados históricos rotulados que são retidos de um conjunto de dados usado para treinar um modelo de machine learning. Você pode usar dados de hold-out para avaliar a performance do modelo comparando as predições do modelo com os dados de retenção.

humano no circuito (HiTL)

Um padrão de fluxo de trabalho em que a execução do agente é pausada para análise e aprovação humana em pontos críticos de decisão.

migração homogênea de bancos de dados

Migrar seu banco de dados de origem para um banco de dados de destino que compartilha o mesmo mecanismo de banco de dados (por exemplo, Microsoft SQL Server para Amazon RDS para SQL Server). A migração homogênea geralmente faz parte de um esforço de redefinição da hospedagem ou da plataforma. É possível usar utilitários de banco de dados nativos para migrar o esquema.

dados quentes

Dados acessados com frequência, como dados em tempo real ou dados translacionais recentes. Esses dados normalmente exigem uma camada ou classe de armazenamento de alto desempenho para fornecer respostas rápidas às consultas.

hotfix

Uma correção urgente para um problema crítico em um ambiente de produção. Devido à sua urgência, um hotfix geralmente é feito fora do fluxo de trabalho típico de uma DevOps versão.

período de hipercuidados

Imediatamente após a substituição, o período em que uma equipe de migração gerencia e monitora as aplicações migradas na nuvem para resolver quaisquer problemas. Normalmente, a duração desse período é de 1 a 4 dias. No final do período de hipercuidados, a equipe de migração normalmente transfere a responsabilidade pelas aplicações para a equipe de operações de nuvem.

eu

IaC

Veja infraestrutura como código.

Política baseada em identidade

Uma política anexada a um ou mais diretores do IAM que define suas permissões no Nuvem AWS ambiente.

aplicação ociosa

Uma aplicação que tem um uso médio de CPU e memória entre 5 e 20% em um período de 90 dias. Em um projeto de migração, é comum retirar essas aplicações ou retê-las on-premises.

IIoT

Veja Internet das Coisas Industrial.

infraestrutura imutável

Um modelo que implanta uma nova infraestrutura para workloads de produção em vez de atualizar, aplicar patches ou modificar a infraestrutura existente. Infraestruturas imutáveis são inerentemente mais consistentes, confiáveis e preditivas do que infraestruturas mutáveis. Para obter mais informações, consulte as melhores práticas de implantação usando infraestrutura imutável na AWS Well-Architected Estrutura.

VPC de entrada (admissão)

Em uma arquitetura de AWS várias contas, uma VPC que aceita, inspeciona e roteia conexões de rede de fora de um aplicativo. A Arquitetura de referência de segurança da AWS recomenda configurar sua conta de rede com VPCs de entrada, saída e inspeção para proteger a interface bidirecional entre a aplicação e a Internet em geral.

migração incremental

Uma estratégia de substituição na qual você migra a aplicação em pequenas partes, em vez de realizar uma única substituição completa. Por exemplo, é possível mover inicialmente apenas alguns microsserviços ou usuários para o novo sistema. Depois de verificar se tudo está funcionando corretamente, mova os microsserviços ou usuários adicionais de forma incremental até poder descomissionar seu sistema herdado. Essa estratégia reduz os riscos associados a migrações de grande porte.

Indústria 4.0

Um termo que foi introduzido por Klaus Schwab em 2016 para se referir à modernização dos processos de fabricação por meio de avanços na conectividade, dados em tempo real, automação, análise e. AI/ML

infraestrutura

Todos os recursos e ativos contidos no ambiente de uma aplicação.

Infraestrutura como código (IaC)

O processo de provisionamento e gerenciamento da infraestrutura de uma aplicação por meio de um conjunto de arquivos de configuração. A IaC foi projetada para ajudar você a centralizar o gerenciamento da infraestrutura, padronizar recursos e escalar rapidamente para que novos ambientes sejam reproduzíveis, confiáveis e consistentes.

Internet das Coisas Industrial (IIoT)

O uso de sensores e dispositivos conectados à Internet nos setores industriais, como manufatura, energia, automotivo, saúde, ciências biológicas e agricultura. Para obter mais informações, consulte Construir uma estratégia de transformação digital para a Internet das Coisas Industrial (IIoT).

VPC de inspeção

Em uma arquitetura de AWS várias contas, uma VPC centralizada que gerencia as inspeções do tráfego de rede entre VPCs (na mesma ou em diferentes Regiões da AWS), a Internet e as redes locais. A Arquitetura de referência de segurança da AWS recomenda configurar sua conta de rede com VPCs de entrada, saída e inspeção para proteger a interface bidirecional entre a aplicação e a Internet em geral.

Internet das coisas (IoT)

A rede de objetos físicos conectados com sensores ou processadores incorporados que se comunicam com outros dispositivos e sistemas pela Internet ou por uma rede de comunicação local. Para obter mais informações, consulte O que é IoT?

interpretabilidade

Uma característica de um modelo de machine learning que descreve o grau em que um ser humano pode entender como as previsões do modelo dependem de suas entradas. Para obter mais informações, consulte Interpretabilidade do modelo de aprendizado de máquina com AWS.

IoT

Veja Internet das Coisas.

Biblioteca de informações de TI (ITIL)

Um conjunto de práticas recomendadas para fornecer serviços de TI e alinhar esses serviços a requisitos de negócios. A ITIL fornece a base para o ITSM.

Gerenciamento de serviços de TI (ITSM)

Atividades associadas a design, implementação, gerenciamento e suporte de serviços de TI para uma organização. Para obter informações sobre a integração de operações em nuvem com ferramentas de ITSM, consulte o guia de integração de operações.

ITIL

Veja biblioteca de informações de TI.

ITSM

Veja gerenciamento de serviços de TI.

L

controle de acesso baseado em etiqueta (LBAC)

Uma implementação do controle de acesso obrigatório (MAC) em que os usuários e os dados em si recebem explicitamente um valor de etiqueta de segurança. A interseção entre a etiqueta de segurança do usuário e a etiqueta de segurança dos dados determina quais linhas e colunas podem ser vistas pelo usuário.

zona de pouso

Uma landing zone é um AWS ambiente bem arquitetado, com várias contas, escalável e seguro. Um ponto a partir do qual suas organizações podem iniciar e implantar rapidamente workloads e aplicações com confiança em seu ambiente de segurança e infraestrutura. Para obter mais informações sobre zonas de pouso, consulte Configurar um ambiente da AWS com várias contas seguro e escalável.

grande modelo de linguagem (LLM)

Um modelo de IA de aprendizado profundo pré-treinado em uma grande quantidade de dados. Um LLM pode realizar várias tarefas, como responder a perguntas, resumir documentos, traduzir texto para outros idiomas e completar frases. Para obter mais informações, consulte O que é grande modelo de linguagem (LLM)?.

migração de grande porte

Uma migração de 300 servidores ou mais.

LBAC

Veja controle de acesso baseado em rótulo.

privilégio mínimo

A prática recomendada de segurança de conceder as permissões mínimas necessárias para executar uma tarefa. Para obter mais informações, consulte Aplicar permissões de privilégios mínimos na documentação do IAM.

mover sem alterações (lift-and-shift)

Veja 7 Rs.

sistema little-endian

Um sistema que armazena o byte menos significativo antes. Veja também endianness.

LLM

Veja grande modelo de linguagem.

ambientes inferiores

Veja ambiente.

M

machine learning (ML)

Um tipo de inteligência artificial que usa algoritmos e técnicas para reconhecimento e aprendizado de padrões. O ML analisa e aprende com dados gravados, por exemplo, dados da Internet das Coisas (IoT), para gerar um modelo estatístico baseado em padrões. Para obter mais informações, consulte Machine learning.

ramificação principal

Veja ramificação.

Malware

Software projetado para comprometer a segurança ou a privacidade do computador. O malware pode interromper os sistemas do computador, vazar informações sensíveis ou obter acesso não autorizado. Exemplos de malware incluem vírus, worms, ransomware, cavalos de Troia, spyware e keyloggers.

Serviços gerenciados

Serviços da AWS para o qual AWS opera a camada de infraestrutura, o sistema operacional e as plataformas, e você acessa os endpoints para armazenar e recuperar dados. O Amazon Simple Storage Service (Amazon S3) e o Amazon DynamoDB são exemplos de serviços gerenciados. Eles também são conhecidos como serviços abstraídos.

sistema de execução de manufatura (MES)

Um sistema de software para rastrear, monitorar, documentar e controlar processos de produção que convertem matérias-primas em produtos acabados no chão de fábrica.

MAP

Veja Programa de Aceleração da Migração.

MCP

Consulte Protocolo de contexto do modelo.

Protocolo de contexto para modelos (MCP)

Um protocolo sem estado para comunicação entre agentes e ferramentas.

Servidor MCP

Um serviço que expõe uma ou mais ferramentas por meio do Model Context Protocol.

mecanismo

Um processo completo em que você cria uma ferramenta, impulsiona a adoção da ferramenta e, em seguida, inspeciona os resultados para fazer ajustes. Um mecanismo é um ciclo que se reforça e se aprimora à medida que opera. Para obter mais informações, consulte Criação de mecanismos na AWS Well-Architected estrutura.

conta de membro

Todos, Contas da AWS exceto a conta de gerenciamento, que fazem parte de uma organização em AWS Organizations. Uma conta só pode ser membro de uma organização de cada vez.

MES

Veja sistema de execução de manufatura.

Transporte de Telemetria de Enfileiramento de Mensagens (MQTT)

Um protocolo de comunicação leve, máquina a máquina (M2M), baseado no padrão, para dispositivos de IoT com recursos publish/subscribelimitados.

microsserviço

Um serviço pequeno e independente que se comunica por meio de APIs bem definidas e normalmente pertence a equipes pequenas e autônomas. Por exemplo, um sistema de seguradora pode incluir microsserviços que mapeiam as capacidades comerciais, como vendas ou marketing, ou subdomínios, como compras, reclamações ou análises. Os benefícios dos microsserviços incluem agilidade, escalabilidade flexível, fácil implantação, código reutilizável e resiliência. Para obter mais informações, consulte Integração de microsserviços usando serviços sem AWS servidor.

arquitetura de microsserviços

Uma abordagem à criação de aplicações com componentes independentes que executam cada processo de aplicação como um microsserviço. Esses microsserviços se comunicam por meio de uma interface bem definida usando APIs leves. Cada microsserviço nessa arquitetura pode ser atualizado, implantado e escalado para atender à demanda por funções específicas de uma aplicação. Para obter mais informações, consulte Implementação de microsserviços em. AWS

Programa de Aceleração da Migração (MAP)

Um AWS programa que fornece suporte de consultoria, treinamento e serviços para ajudar as organizações a criar uma base operacional sólida para migrar para a nuvem e ajudar a

compensar o custo inicial das migrações. O MAP inclui uma metodologia de migração para executar migrações legadas de forma metódica e um conjunto de ferramentas para automatizar e acelerar cenários comuns de migração.

migração em escala

O processo de mover a maior parte do portfólio de aplicações para a nuvem em ondas, com mais aplicações sendo movidas em um ritmo mais rápido a cada onda. Essa fase usa as práticas recomendadas e lições aprendidas nas fases anteriores para implementar uma fábrica de migração de equipes, ferramentas e processos para agilizar a migração de workloads por meio de automação e entrega ágeis. Esta é a terceira fase da estratégia de migração para a AWS.

fábrica de migração

Cross-functional equipes que simplificam a migração de cargas de trabalho por meio de abordagens automatizadas e ágeis. As equipes da fábrica de migração geralmente incluem operações, analistas e proprietários de negócios, engenheiros de migração, desenvolvedores e DevOps profissionais que trabalham em sprints. Entre 20 e 50% de um portfólio de aplicações corporativas consiste em padrões repetidos que podem ser otimizados por meio de uma abordagem de fábrica. Para obter mais informações, consulte discussão sobre fábricas de migração e o guia do Cloud Migration Factory neste conjunto de conteúdo.

metadados de migração

As informações sobre a aplicação e o servidor necessárias para concluir a migração. Cada padrão de migração exige um conjunto de metadados de migração diferente. Exemplos de metadados de migração incluem a sub-rede, o grupo de segurança e AWS a conta de destino.

padrão de migração

Uma tarefa de migração repetível que detalha a estratégia de migração, o destino da migração e a aplicação ou o serviço de migração usado. Exemplo: rehospede a migração para o Amazon EC2 AWS com o Application Migration Service.

Avaliação de Portfólio para Migração (MPA)

Uma ferramenta on-line que fornece informações para validar o caso de negócios para migrar para a Nuvem AWS. O MPA fornece avaliação detalhada do portfólio (dimensionamento correto do servidor, preços, comparações de TCO, análise de custos de migração), bem como planejamento de migração (análise e coleta de dados de aplicações, agrupamento de aplicações, priorização de migração e planejamento de ondas). A ferramenta MPA (requer login) está disponível gratuitamente para todos os AWS consultores e consultores parceiros da APN.

Avaliação de Preparação para Migração (MRA)

O processo de obter insights sobre o status de prontidão de uma organização para a nuvem, identificar pontos fortes e fracos e criar um plano de ação para fechar as lacunas identificadas, usando o CAF. AWS Para mais informações, consulte o guia de preparação para migração. A MRA é a primeira fase da estratégia de migração para a AWS.

estratégia de migração

A abordagem usada para migrar uma workload para a Nuvem AWS. Para obter mais informações, veja a entrada 7 Rs neste glossário e consulte Mobilize sua organização para acelerar migrações em grande escala.

ML

Veja machine learning.

modernização

Transformar uma aplicação desatualizada (herdada ou monolítica) e sua infraestrutura em um sistema ágil, elástico e altamente disponível na nuvem para reduzir custos, ganhar eficiência e aproveitar as inovações. Para obter mais informações, consulte Strategy for modernizing applications in the Nuvem AWS.

avaliação de preparação para modernização

Uma avaliação que ajuda a determinar a preparação para modernização das aplicações de uma organização. Ela identifica benefícios, riscos e dependências e determina o quão bem a organização pode acomodar o estado futuro dessas aplicações. O resultado da avaliação é um esquema da arquitetura de destino, um roteiro que detalha as fases de desenvolvimento e os marcos do processo de modernização e um plano de ação para abordar as lacunas identificadas. Para obter mais informações, consulte Evaluating modernization readiness for applications in the Nuvem AWS.

aplicações monolíticas (monólitos)

Aplicações que são executadas como um único serviço com processos fortemente acoplados. As aplicações monolíticas apresentam várias desvantagens. Se um recurso da aplicação apresentar um aumento na demanda, toda a arquitetura deverá ser escalada. Adicionar ou melhorar os recursos de uma aplicação monolítica também se torna mais complexo quando a base de código cresce. Para resolver esses problemas, é possível criar uma arquitetura de microsserviços. Para obter mais informações, consulte Decompor monólitos em microsserviços.

MPA

Veja Avaliação do Portfólio para Migração.

MQTT

Veja Transporte de Telemetria de Enfileiramento de Mensagens.

classificação multiclasse

Um processo que ajuda a gerar previsões para várias classes (prevendo um ou mais de dois resultados). Por exemplo, um modelo de ML pode perguntar “Este produto é um livro, um carro ou um telefone?” ou “Qual categoria de produtos é mais interessante para este cliente?”

infraestrutura mutável

Um modelo que atualiza e modifica a infraestrutura existente para workloads de produção. Para melhorar a consistência, confiabilidade e previsibilidade, a AWS Well-Architected Estrutura recomenda o uso de infraestrutura imutável como uma prática recomendada.

O

OAC

Veja controle de acesso de origem.

OAI

Veja identidade de acesso de origem.

OCM

Veja gerenciamento de alterações organizacionais.

migração offline

Um método de migração no qual a workload de origem é desativada durante o processo de migração. Esse método envolve tempo de inatividade prolongado e geralmente é usado para workloads pequenas e não críticas.

OI

Veja integração de operações.

Ola

Veja acordo de nível operacional.

migração online

Um método de migração no qual a workload de origem é copiada para o sistema de destino sem ser colocada offline. As aplicações conectadas à workload podem continuar funcionando durante a migração. Esse método envolve um tempo de inatividade nulo ou mínimo e normalmente é usado para workloads essenciais para a produção.

OPC-UA

Veja Open Process Communications - Unified Architecture.

Comunicação de processo aberto - Arquitetura unificada (OPC-UA)

Um protocolo de comunicação máquina a máquina (M2M) para automação industrial. OPC-UA fornece um padrão de interoperabilidade com esquemas de criptografia, autenticação e autorização de dados.

acordo de nível operacional (OLA)

Um acordo que esclarece o que os grupos funcionais de TI prometem oferecer uns aos outros para apoiar um acordo de serviço (SLA).

análise de prontidão operacional (ORR)

Uma lista de verificação de perguntas e práticas recomendadas associadas que ajudam você a entender, avaliar, prevenir ou reduzir o escopo de incidentes e possíveis falhas. Para obter mais informações, consulte Operational Readiness Reviews (ORR) na AWS Well-Architected Estrutura.

tecnologia operacional (TO)

Sistemas de hardware e software que trabalham com o ambiente físico para controlar operações, equipamentos e infraestrutura industriais. Na manufatura, a integração dos sistemas de tecnologia da informação (TI) e tecnologia operacional (TO) é o foco principal das transformações da Indústria 4.0.

integração de operações (OI)

O processo de modernização das operações na nuvem, que envolve planejamento de preparação, automação e integração. Para obter mais informações, consulte o guia de integração de operações.

trilha organizacional

Uma trilha criada por ela AWS CloudTrail registra todos os eventos de todos Contas da AWS em uma organização em AWS Organizations. Essa trilha é criada em cada Conta da AWS que faz parte da organização e monitora a atividade em cada conta. Para obter mais informações, consulte Criação de uma trilha para uma organização na CloudTrail documentação.

gerenciamento de alterações organizacionais (OCM)

Uma estrutura para gerenciar grandes transformações de negócios disruptivas de uma perspectiva de pessoas, cultura e liderança. O OCM ajuda as organizações a se prepararem e fazerem a transição para novos sistemas e estratégias, acelerando a adoção de alterações, abordando questões de transição e promovendo mudanças culturais e organizacionais. Na estratégia de AWS migração, essa estrutura é chamada de aceleração de pessoas, devido à velocidade de mudança necessária nos projetos de adoção da nuvem. Para obter mais informações, consulte o guia do OCM.

controle de acesso de origem (OAC)

Em CloudFront, uma opção aprimorada para restringir o acesso para proteger seu conteúdo do Amazon Simple Storage Service (Amazon S3). O OAC oferece suporte a todos os buckets do S3 Regiões da AWS, à criptografia do lado do servidor com AWS KMS (SSE-KMS) e à dinâmica `PUT` e `DELETE` às solicitações ao bucket do S3.

Identidade do acesso de origem (OAI)

Em CloudFront, uma opção para restringir o acesso para proteger seu conteúdo do Amazon S3. Quando você usa o OAI, CloudFront cria um principal com o qual o Amazon S3 pode se autenticar. Os diretores autenticados podem acessar o conteúdo em um bucket do S3 somente por meio de uma distribuição específica. CloudFront Veja também OAC, que fornece um controle de acesso mais granular e aprimorado.

ORR

Veja análise de prontidão operacional.

OT

Veja tecnologia operacional.

VPC de saída (egresso)

Em uma arquitetura de AWS várias contas, uma VPC que gerencia conexões de rede que são iniciadas de dentro de um aplicativo. A Arquitetura de referência de segurança da AWS

recomenda configurar sua conta de rede com VPCs de entrada, saída e inspeção para proteger a interface bidirecional entre a aplicação e a Internet em geral.

P

limite de permissões

Uma política de gerenciamento do IAM anexada a entidades principais do IAM para definir as permissões máximas que o usuário ou perfil podem ter. Para obter mais informações, consulte Limites de permissões na documentação do IAM.

Informações de identificação pessoal (PII)

Informações que, quando visualizadas diretamente ou combinadas com outros dados relacionados, podem ser usadas para inferir razoavelmente a identidade de um indivíduo. Exemplos de PII incluem nomes, endereços e informações de contato.

PII

Veja informações de identificação pessoal.

manual

Um conjunto de etapas predefinidas que capturam o trabalho associado às migrações, como a entrega das principais funções operacionais na nuvem. Um manual pode assumir a forma de scripts, runbooks automatizados ou um resumo dos processos ou etapas necessários para operar seu ambiente modernizado.

PLC

Veja controlador lógico programável.

PLM

Veja gerenciamento do ciclo de vida do produto.

política

Um objeto que pode definir permissões (veja política baseada em identidade), especificar condições de acesso (veja políta baseada em recurso) ou definir as permissões máximas para todas as contas em uma organização no AWS Organizations (veja política de controle de serviços).

persistência poliglota

Escolher de forma independente a tecnologia de armazenamento de dados de um microsserviço com base em padrões de acesso a dados e outros requisitos. Se seus microsserviços tiverem a mesma tecnologia de armazenamento de dados, eles poderão enfrentar desafios de implementação ou apresentar baixa performance. Os microsserviços serão implementados com mais facilidade e alcançarão performance e escalabilidade melhores se usarem o armazenamento de dados mais bem adaptado às suas necessidades.

avaliação do portfólio

Um processo de descobrir, analisar e priorizar o portfólio de aplicações para planejar a migração. Para obter mais informações, consulte Avaliar a preparação para a migração.

predicado

Uma condição de consulta que retorna `true` ou `false`, normalmente localizada em uma cláusula `WHERE`.

pushdown de predicados

Uma técnica de otimização de consultas de banco de dados que filtra os dados na consulta antes da transferência. Isso reduz a quantidade de dados que devem ser recuperados e processados do banco de dados relacional e melhora a performance das consultas.

controle preventivo

Um controle de segurança projetado para evitar que um evento ocorra. Esses controles são a primeira linha de defesa para ajudar a evitar acesso não autorizado ou alterações indesejadas em sua rede. Para obter mais informações, consulte Controles preventivos em Como implementar controles de segurança na AWS.

principal (entidade principal)

Uma entidade AWS que pode realizar ações e acessar recursos. Essa entidade geralmente é um usuário raiz para um Conta da AWS, uma função do IAM ou um usuário. Para obter mais informações, consulte Entidade principal em Termos e conceitos de perfis na documentação do IAM.

Privacidade por design

Uma abordagem em engenharia de sistemas que leva em consideração a privacidade em todo o processo de desenvolvimento.

zonas hospedadas privadas

Um contêiner que armazena informações sobre como você quer que o Amazon Route 53 responda a consultas ao DNS para um domínio e seus subdomínios dentro de uma ou mais VPCs. Para obter mais informações, consulte Como trabalhar com zonas hospedadas privadas na documentação do Route 53.

controle proativo

Um controle de segurança desenvolvido para evitar a implantação de recursos não conformes. Esses controles verificam os recursos antes de serem provisionados. Se o recurso não estiver em conformidade com o controle, ele não será provisionado. Para obter mais informações, consulte o guia de referência de controles na AWS Control Tower documentação e consulte Controles proativos em Implementação de controles de segurança em AWS.

gerenciamento do ciclo de vida do produto (PLM)

O gerenciamento de dados e processos de um produto em todo o seu ciclo de vida, desde a concepção, o desenvolvimento e o lançamento, passando pelo crescimento e maturidade, até o declínio e a remoção.

ambiente de produção

Veja ambiente.

controlador lógico programável (PLC)

Na manufatura, um computador altamente confiável e adaptável que monitora as máquinas e automatiza os processos de fabricação.

encadeamento de prompts

Uso da saída de um prompt do LLM como entrada para o próximo prompt para gerar respostas melhores. Essa técnica é usada para dividir uma tarefa complexa em subtarefas, ou para refinar ou expandir iterativamente uma resposta preliminar. Isso ajuda a melhorar a precisão e a relevância das respostas de um modelo e permite resultados mais granulares e personalizados.

pseudonimização

O processo de substituir identificadores pessoais em um conjunto de dados por valores de espaço reservado. A pseudonimização pode ajudar a proteger a privacidade pessoal. Os dados pseudonimizados ainda são considerados dados pessoais.

publish/subscribe (pub/sub)

Um padrão que permite comunicações assíncronas entre microsserviços para melhorar a escalabilidade e a capacidade de resposta. Por exemplo, em um MES baseado em microsserviços, um microsserviço pode publicar mensagens de eventos em um canal em que outros microsserviços possam assinar. O sistema pode adicionar novos microsserviços sem alterar o serviço de publicação.

Q

plano de consulta

Uma série de etapas, como instruções, usadas para acessar os dados em um sistema de banco de dados relacional SQL.

regressão de planos de consultas

Quando um otimizador de serviço de banco de dados escolhe um plano menos adequado do que escolhia antes de uma determinada alteração no ambiente de banco de dados ocorrer. Isso pode ser causado por alterações em estatísticas, restrições, configurações do ambiente, associações de parâmetros de consulta e atualizações do mecanismo de banco de dados.

R

Matriz RACI

Veja responsável, aprovador, consultado, informado (RACI).

RAG

Veja geração aumentada via recuperação.

ransomware

Um software mal-intencionado desenvolvido para bloquear o acesso a um sistema ou dados de computador até que um pagamento seja feito.

Matriz RASCI

Veja responsável, aprovador, consultado, informado (RACI).

RCAC

Veja controle de acesso por linha e coluna.

réplica de leitura

Uma cópia de um banco de dados usada somente para leitura. É possível encaminhar consultas para a réplica de leitura e reduzir a carga no banco de dados principal.

Redefinir arquitetura

Veja 7 Rs.

objetivo de ponto de recuperação (RPO).

O máximo período de tempo 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.

objetivo de tempo de recuperação (RTO)

O máximo atraso aceitável entre a interrupção e a restauração do serviço.

refatorar

Veja 7 Rs.

Região

Uma coleção de AWS recursos em uma área geográfica. Cada um Região da AWS é isolado e independente dos outros para fornecer tolerância a falhas, estabilidade e resiliência. Para obter informações, consulte Specify which Regiões da AWS your account can use.

regressão

Uma técnica de ML que prevê um valor numérico. Por exemplo, para resolver o problema de “Por qual preço esta casa será vendida?” um modelo de ML pode usar um modelo de regressão linear para prever o preço de venda de uma casa com base em fatos conhecidos sobre a casa (por exemplo, a metragem quadrada).

redefinir a hospedagem

Veja 7 Rs.

versão

Em um processo de implantação, o ato de promover mudanças em um ambiente de produção.

realocar

Veja 7 Rs.

redefinir a plataforma

Veja 7 Rs.

recomprar

Veja 7 Rs.

resiliência

A capacidade de uma aplicação de resistir ou se recuperar de interrupções. Alta disponibilidade e recuperação de desastres são considerações comuns ao planejar a resiliência na Nuvem AWS. Para obter mais informações, consulte Nuvem AWS Resilience.

política baseada em recurso

Uma política associada a um recurso, como um bucket do Amazon S3, um endpoint ou uma chave de criptografia. Esse tipo de política especifica quais entidades principais têm acesso permitido, ações válidas e quaisquer outras condições que devem ser atendidas.

matriz responsável, accountable, consultada, informada (RACI)

Uma matriz que define as funções e responsabilidades de todas as partes envolvidas nas atividades de migração e nas operações de nuvem. O nome da matriz é derivado dos tipos de responsabilidade definidos na matriz: responsável (R), responsabilizável (A), consultado (C) e informado (I). O tipo de suporte (S) é opcional. Se você incluir suporte, a matriz será chamada de matriz RASCI e, se excluir, será chamada de matriz RACI.

controle responsivo

Um controle de segurança desenvolvido para conduzir a remediação de eventos adversos ou desvios em relação à linha de base de segurança. Para obter mais informações, consulte Controles responsivos em Como implementar controles de segurança na AWS.

reter

Veja 7 Rs.

Retirada

Veja 7 Rs.

Geração Aumentada de Recuperação (RAG)

Uma tecnologia de IA generativa em que um LLM faz referência a uma fonte de dados autorizada que está fora de suas fontes de dados de treinamento antes de gerar uma resposta. Por exemplo, um modelo RAG pode realizar uma pesquisa semântica na base de conhecimento ou nos dados personalizados de uma organização. Para obter mais informações, consulte O que é RAG (geração aumentada via recuperação)?.

alternância

O processo de atualizar periodicamente um segredo para dificultar o acesso de um invasor às credenciais.

controle de acesso por linha e coluna (RCAC)

O uso de expressões SQL básicas e flexíveis que tenham regras de acesso definidas. O RCAC consiste em permissões de linha e máscaras de coluna.

RPO

Veja objetivo de ponto de recuperação.

RTO

Veja objetivo de tempo de recuperação.

runbook

Um conjunto de procedimentos manuais ou automatizados necessários para realizar uma tarefa específica. Eles são normalmente criados para agilizar operações ou procedimentos repetitivos com altas taxas de erro.

S

SAML 2.0

Um padrão aberto que muitos provedores de identidade (IdPs) usam. Esse recurso permite o login único federado (SSO), para que os usuários possam fazer login Console de gerenciamento da AWS ou chamar as operações da AWS API sem que você precise criar um usuário no IAM para todos em sua organização. Para obter mais informações sobre a federação baseada em SAML 2.0, consulte Sobre a federação baseada em SAML 2.0 na documentação do IAM.

SCADA

Veja controle de supervisão e aquisição de dados.

SCP

Veja política de controle de serviço.

secret

Em AWS Secrets Manager, informações confidenciais ou restritas, como uma senha ou credenciais de usuário, que você armazena de forma criptografada. Consiste no valor secreto e em seus metadados. O valor secreto pode ser binário, uma única string ou várias strings. Para obter mais informações, consulte What's in a Secrets Manager secret? na documentação do Secrets Manager.

segurança desde a concepção

Uma abordagem em engenharia de sistemas que leva em consideração a segurança em todo o processo de desenvolvimento.

controle de segurança

Uma barreira de proteção técnica ou administrativa que impede, detecta ou reduz a capacidade de uma ameaça explorar uma vulnerabilidade de segurança. Existem quatro tipos primários de controles de segurança: preventivos, detectivos, responsivos e proativos.

hardening da segurança

O processo de reduzir a superfície de ataque para torná-la mais resistente a ataques. Isso pode incluir ações como remover recursos que não são mais necessários, implementar a prática recomendada de segurança de conceder privilégios mínimos ou desativar recursos desnecessários em arquivos de configuração.

sistema de gerenciamento de eventos e informações de segurança (SIEM)

Ferramentas e serviços que combinam sistemas de gerenciamento de informações de segurança (SIM) e gerenciamento de eventos de segurança (SEM). Um sistema SIEM coleta, monitora e analisa dados de servidores, redes, dispositivos e outras fontes para detectar ameaças e violações de segurança e gerar alertas.

automação de resposta de segurança

Uma ação predefinida e programada projetada para responder ou remediar automaticamente um evento de segurança. Essas automações servem como controles de segurança responsivos ou detectivos que ajudam você a implementar as melhores práticas AWS de segurança. Exemplos de ações de resposta automatizada incluem a modificação de um grupo de segurança da VPC, a aplicação de patches em uma instância do Amazon EC2 ou a alternância de credenciais.

Criptografia do lado do servidor

Criptografia dos dados em seu destino, por AWS service (Serviço da AWS) quem os recebe.

política de controle de serviços (SCP)

Uma política que fornece controle centralizado sobre as permissões de todas as contas em uma organização no AWS Organizations. As SCPs definem barreiras de proteção ou estabelecem limites para as ações que um administrador pode delegar a usuários ou perfis. É possível usar SCPs como listas de permissão ou de negação para especificar quais serviços ou ações são permitidos ou proibidos. Para obter mais informações, consulte Políticas de controle de serviço na AWS Organizations documentação.

service endpoint (endpoint de serviço)

O URL do ponto de entrada para um AWS service (Serviço da AWS). Você pode usar o endpoint para se conectar programaticamente ao serviço de destino. Para obter mais informações, consulte Endpoints do AWS service (Serviço da AWS) na Referência geral da AWS.

acordo de serviço (SLA)

Um acordo que esclarece o que uma equipe de TI promete fornecer aos clientes, como tempo de atividade e performance do serviço.

indicador de nível de serviço (SLI)

Uma avaliação de um aspecto de performance de um serviço, como taxa de erro, disponibilidade ou throughput.

objetivo de nível de serviço (SLO)

Uma métrica alvo que representa a integridade de um serviço, conforme avaliado por um indicador de nível de serviço.

modelo de responsabilidade compartilhada

Um modelo que descreve a responsabilidade com a qual você compartilha AWS pela segurança e conformidade na nuvem. AWS é responsável pela segurança da nuvem, enquanto você é responsável pela segurança na nuvem. Para obter mais informações, consulte o Modelo de responsabilidade compartilhada.

Inteligência artificial sombria

Aplicativos de IA não autorizados criados ou usados fora dos canais controlados dentro de uma organização.

SIEM

Veja sistema de gerenciamento de eventos e informações de segurança.

ponto único de falha (SPOF)

Uma falha em um único componente crítico de uma aplicação que pode interromper o sistema.

SLA

Veja cacordo de serviço.

SLI

Veja indicador de nível de serviço.

SLO

Veja objetivo de nível de serviço.

modelo dividir e semear

Um padrão para escalar e acelerar projetos de modernização. À medida que novos recursos e lançamentos de produtos são definidos, a equipe principal se divide para criar novas equipes de produtos. Isso ajuda a escalar os recursos e os serviços da sua organização, melhora a produtividade do desenvolvedor e possibilita inovações rápidas. Para obter mais informações, consulte Phased approach to modernizing applications in the Nuvem AWS.

SPOF

Veja ponto único de falha.

esquema em estrela

Uma estrutura organizacional de banco de dados que usa uma grande tabela de fatos para armazenar dados transacionais ou medidos e usa uma ou mais tabelas dimensionais menores para armazenar atributos de dados. Essa estrutura foi projetada para ser usada em um data warehouse ou para fins de inteligência comercial.

padrão strangler fig

Uma abordagem à modernização de sistemas monolíticos que consiste em reescrever e substituir incrementalmente a funcionalidade do sistema até que o sistema herdado possa ser desativado. Esse padrão usa a analogia de uma videira que cresce e se torna uma árvore estabelecida e, eventualmente, supera e substitui sua hospedeira. O padrão foi apresentado por Martin Fowler

como forma de gerenciar riscos ao reescrever sistemas monolíticos. Para ver um exemplo de como aplicar esse padrão, consulte Modernizando os serviços web legados da Microsoft ASP.NET (ASMX) de forma incremental usando contêineres e o Amazon API Gateway.

sub-rede

Um intervalo de endereços IP na VPC. Cada sub-rede fica alocada em uma única zona de disponibilidade.

controle supervisório e aquisição de dados (SCADA)

Na manufatura, um sistema que usa hardware e software para monitorar ativos físicos e operações de produção.

symmetric encryption (criptografia simétrica)

Um algoritmo de criptografia que usa a mesma chave para criptografar e descriptografar dados.

testes sintéticos

Testar um sistema de forma que simule as interações do usuário para detectar possíveis problemas ou monitorar a performance. Você pode usar o Amazon CloudWatch Synthetics para criar esses testes.

prompt do sistema

Uma técnica para fornecer contexto, instruções ou orientações a um LLM a fim de direcionar seu comportamento. Os prompts do sistema ajudam a definir o contexto e a estabelecer regras para interações com os usuários.

T

tags

Key-value pares que atuam como metadados para organizar seus AWS recursos. As tags podem ajudar você a gerenciar, identificar, organizar, pesquisar e filtrar recursos da . Para obter mais informações, consulte Marcar seus recursos do AWS.

variável-alvo

O valor que você está tentando prever no ML supervisionado. Ela também é conhecida como variável de resultado. Por exemplo, em uma configuração de fabricação, a variável-alvo pode ser um defeito do produto.

lista de tarefas

Uma ferramenta usada para monitorar o progresso por meio de um runbook. Uma lista de tarefas contém uma visão geral do runbook e uma lista de tarefas gerais a serem concluídas. Para cada tarefa geral, ela inclui o tempo estimado necessário, o proprietário e o progresso.

ambiente de teste

Veja ambiente.

treinamento

O processo de fornecer dados para que seu modelo de ML aprenda. Os dados de treinamento devem conter a resposta correta. O algoritmo de aprendizado descobre padrões nos dados de treinamento que mapeiam os atributos dos dados de entrada no destino (a resposta que você deseja prever). Ele gera um modelo de ML que captura esses padrões. Você pode usar o modelo de ML para obter previsões de novos dados cujo destino você não conhece.

ferramenta

Uma função ou API que um agente pode invocar para realizar operações em sistemas externos.

gateway de trânsito

Um hub de trânsito de rede que pode ser usado para interconectar as VPCs e as redes on-premises. Para obter mais informações, consulte O que é um gateway de trânsito na AWS Transit Gateway documentação.

fluxo de trabalho baseado em troncos

Uma abordagem na qual os desenvolvedores criam e testam recursos localmente em uma ramificação de recursos e, em seguida, mesclam essas alterações na ramificação principal. A ramificação principal é então criada para os ambientes de desenvolvimento, pré-produção e produção, sequencialmente.

Acesso confiável

Conceder permissões a um serviço que você especifica para realizar tarefas em sua organização AWS Organizations e em suas contas em seu nome. O serviço confiável cria um perfil vinculado ao serviço em cada conta, quando esse perfil é necessário, para realizar tarefas de gerenciamento para você. Para obter mais informações, consulte Usando AWS Organizations com outros AWS serviços na AWS Organizations documentação.

tuning (ajustar)

Alterar aspectos do processo de treinamento para melhorar a precisão do modelo de ML. Por exemplo, você pode treinar o modelo de ML gerando um conjunto de rótulos, adicionando rótulos e repetindo essas etapas várias vezes em configurações diferentes para otimizar o modelo.

equipe de duas pizzas

Uma pequena DevOps equipe que você pode alimentar com duas pizzas. Uma equipe de duas pizzas garante a melhor oportunidade possível de colaboração no desenvolvimento de software.

U

incerteza

Um conceito que se refere a informações imprecisas, incompletas ou desconhecidas que podem minar a confiabilidade dos modelos preditivos de ML. Há dois tipos de incertezas: a incerteza epistêmica é causada por dados limitados e incompletos, enquanto a incerteza aleatória é causada pelo ruído e pela aleatoriedade inerentes aos dados.

tarefas indiferenciadas

Também conhecido como trabalho pesado, trabalho necessário para criar e operar um aplicativo, mas que não fornece valor direto ao usuário final nem oferece vantagem competitiva. Exemplos de tarefas indiferenciadas incluem aquisição, manutenção e planejamento de capacidade.

ambientes superiores

Veja ambiente.

V

aspiração

Uma operação de manutenção de banco de dados que envolve limpeza após atualizações incrementais para recuperar armazenamento e melhorar a performance.

controle de versões

Processos e ferramentas que rastreiam mudanças, como alterações no código-fonte em um repositório.

emparelhamento de VPC

Uma conexão entre duas VPCs que permite rotear tráfego usando endereços IP privados. Para ter mais informações, consulte O que é emparelhamento de VPC? na documentação da Amazon VPC.

Vulnerabilidade

Uma falha de software ou hardware que compromete a segurança do sistema.

W

cache quente

Um cache de buffer que contém dados atuais e relevantes que são acessados com frequência. A instância do banco de dados pode ler do cache do buffer, o que é mais rápido do que ler da memória principal ou do disco.

dados mornos

Dados acessados raramente. Ao consultar esse tipo de dados, consultas moderadamente lentas geralmente são aceitáveis.

função de janela

Uma função SQL que executa um cálculo em um grupo de linhas que se relacionam de alguma forma com o registro atual. As funções de janela são úteis para processar tarefas, como calcular uma média móvel ou acessar o valor das linhas com base na posição relativa da linha atual.

workload

Uma coleção de códigos e recursos que geram valor empresarial, como uma aplicação voltada para o cliente ou um processo de backend.

workstreams

Grupos funcionais em um projeto de migração que são responsáveis por um conjunto específico de tarefas. Cada workstream é independente, mas oferece suporte aos outros workstreams do projeto. Por exemplo, o workstream de portfólio é responsável por priorizar aplicações, planejar ondas e coletar metadados de migração. O workstream de portfólio entrega esses ativos ao workstream de migração, que então migra os servidores e as aplicações.

WORM

Veja gravação única e várias leituras.

WQF

Veja AWS Workload Qualification Framework.

gravação única e várias leituras (WORM)

Um modelo de armazenamento que grava dados uma única vez e evita que os dados sejam excluídos ou modificados. Os usuários autorizados podem ler os dados quantas vezes forem necessárias, mas não podem alterá-los. Essa infraestrutura de armazenamento de dados é considerada imutável.

Z

exploração de dia zero

Um ataque, normalmente malware, que tira proveito de uma vulnerabilidade zero-day.

vulnerabilidade de dia zero

Uma falha ou vulnerabilidade não mitigada em um sistema de produção. Os agentes de ameaças podem usar esse tipo de vulnerabilidade para atacar o sistema. Os desenvolvedores frequentemente ficam cientes da vulnerabilidade como resultado do ataque.

prompt zero shot

Fornecer a um LLM instruções para realizar uma tarefa, mas sem exemplos (shots) que possam ajudar a orientá-lo. O LLM deve usar seu conhecimento pré-treinado para lidar com a tarefa. A eficácia dos prompts zero-shot depende da complexidade da tarefa e da qualidade do prompt. Veja também prompts few-shot.

aplicação zumbi

Uma aplicação que tem um uso médio de CPU e memória inferior a 5%. Em um projeto de migração, é comum retirar essas aplicações.

As traduções são geradas por tradução automática. Em caso de conflito entre o conteúdo da tradução e da versão original em inglês, a versão em inglês prevalecerá.