AWS Well-Architected Framework

Pilar de eficiencia del rendimiento

Pilar de eficiencia del rendimiento: AWS Well-Architected Framework

Table of Contents

Resumen e introducción 1
Resumen 1
Introducción 1
Eficiencia del rendimiento 3
Principios de diseño 3
Definición 4
Selección de la arquitectura 5
PERF01-BP01 Descubrir y comprender los servicios y las características disponibles en la nube 5
Guía para la implementación 6
Recursos 7
PERF01-BP02 Seguir las recomendaciones de su proveedor de servicios en la nube o de un socio adecuado para conocer los modelos arquitectónicos y las prácticas recomendadas 8
Guía para la implementación 6
Recursos 7
PERF01-BP03 Tener en cuenta los costes en sus decisiones arquitectónicas 10
Guía para la implementación 6
Recursos 7
PERF01-BP04 Analizar cómo sus decisiones afectan a los clientes y a la eficiencia de la arquitectura 12
Guía para la implementación 6
Recursos 7
PERF01-BP05 Usar políticas y arquitecturas de referencia 14
Guía para la implementación 6
Recursos 7
PERF01-BP06 Realizar pruebas comparativas para tomar decisiones arquitectónicas 16
Guía para la implementación 6
Recursos 7
PERF01-BP07 Aplicar un enfoque basado en los datos en sus decisiones arquitectónicas 19
Guía para la implementación 6
Recursos 7
Computación y hardware 22
PERF02-BP01 Seleccionar las mejores opciones computacionales para su carga de trabajo 22
Guía para la implementación 6

Pasos para la implementación 6
Recursos 7
PERF02-BP02 Comprender las opciones de configuración y las características de computación disponibles 26
Guía para la implementación 6
Pasos para la implementación 6
Recursos 7
PERF02-BP03 Recopilar métricas relacionadas con la computación 30
Guía para la implementación 6
Pasos para la implementación 6
Recursos 7
PERF02-BP04 Configurar y dimensionar correctamente los recursos de computación 33
Guía para la implementación 6
Recursos 7
PERF02-BP05 Escalar los recursos computacionales de forma dinámica 35
Guía para la implementación 6
Recursos 7
PERF02-BP06 Utilización de aceleradores computacionales optimizados basados en hardware 39
Guía para la implementación 6
Recursos 7
Administración de datos 42
PERF03-BP01 Utilización de un almacén de datos personalizado que se adapte mejor a los requisitos de acceso y almacenamiento de datos 42
Guía para la implementación 6
Recursos 7
PERF03-BP02 Evaluar las opciones de configuración disponibles 54
Guía para la implementación 6
Recursos 7
PERF03-BP03 Recopilar y registrar las métricas de rendimiento del almacén de datos 59
Guía para la implementación 6
Pasos para la implementación 6
Recursos 7
PERF03-BP04 Implementar estrategias para mejorar el rendimiento de las consultas en el almacén de datos 62
Guía para la implementación 6

Recursos 7
PERF03-BP05 Implementar patrones de acceso a datos que utilicen el almacenamiento en caché 64
Guía para la implementación 6
Recursos 7
Redes y entrega de contenido 69
PERF04-BP01 Comprender cómo afectan las redes al rendimiento 69
Guía para la implementación 6
Recursos 7
PERF04-BP02 Evaluar las características de las redes disponibles 73
Guía para la implementación 6
Recursos 7
PERF04-BP03 Elegir la conectividad o VPN dedicadas adecuadas para la carga de trabajo 80
Guía para la implementación 6
Recursos 7
PERF04-BP04 Utilizar el equilibrio de carga para distribuir el tráfico entre varios recursos 83
Guía para la implementación 6
Recursos 7
PERF04-BP05 Elegir los protocolos de red para mejorar el rendimiento 87
Guía para la implementación 6
Recursos 7
PERF04-BP06 Elegir la ubicación de la carga de trabajo en función de los requisitos de la red 91
Guía para la implementación 6
Recursos 7
PERF04-BP07 Optimizar la configuración de red según las métricas 96
Guía para la implementación 6
Recursos 7
Proceso y cultura 102
PERF05-BP01 Establecer indicadores clave de rendimiento (KPI) para medir el estado y el rendimiento de la carga de trabajo 104
Guía para la implementación 6
Pasos para la implementación 6
Recursos 7
PERF05-BP02 Utilizar soluciones de supervisión para saber en qué áreas es más crítico el rendimiento 107

Guía para la implementación 6
Recursos 7
PERF05-BP03 Definir un proceso para mejorar el rendimiento de la carga de trabajo 110
Guía para la implementación 6
Recursos 7
PERF05-BP04 Realizar pruebas de la carga de trabajo 112
Guía para la implementación 6
Recursos 7
PERF05-BP05 Utilizar la automatización para solucionar de forma proactiva los problemas relacionados con el rendimiento 114
Guía para la implementación 6
Recursos 7
PERF05-BP06 Mantener la carga de trabajo y los servicios actualizados 117
Guía para la implementación 6
Pasos para la implementación 6
Recursos 7
PERF05-BP07 Revisar las métricas a intervalos regulares 119
Guía para la implementación 6
Recursos 7
Conclusión 122
Colaboradores 123
Otra documentación 124
Revisiones del documento 125
AWS Glossary 127

Pilar de eficiencia del rendimiento: AWS Well-Architected Framework.

Fecha de publicación: 27 de junio de 2024 (Revisiones del documento)

Resumen

Este documento técnico se centra en el pilar de la eficiencia del rendimiento de AWS Well-Architected Framework. El objetivo de este documento es proporcionar una guía que ayude a los clientes a utilizar los recursos de la nube de manera eficiente para satisfacer sus requisitos empresariales y a mantener esa eficiencia a medida que la demanda cambia y las tecnologías evolucionan.

Introducción

El marco AWS Well-Architected Framework le ayuda a comprender las ventajas y desventajas de las decisiones que toma al crear cargas de trabajo en AWS. Mediante el uso del marco, podrá conocer las prácticas recomendadas de arquitectura para diseñar y operar cargas de trabajo en la nube que sean fiables, seguras, eficaces, rentables y sostenibles. El marco ofrece una forma de medir sus arquitecturas de forma constante en función de las prácticas recomendadas de arquitectura y de identificar áreas de mejora. Creemos que contar con cargas de trabajo de buena arquitectura aumenta en gran medida la probabilidad de éxito empresarial.

El marco se basa en seis pilares:

- Excelencia operativa
- Seguridad
- Fiabilidad
- Eficiencia del rendimiento
- Optimización de costes
- Sostenibilidad

Este documento se centra en la aplicación de los principios del pilar de eficiencia del rendimiento a sus cargas de trabajo. Alcanzar un rendimiento alto y duradero puede ser un desafío en los entornos

locales tradicionales. El uso de los principios de este documento le ayudará a crear arquitecturas en AWS que ofrezcan un rendimiento sostenido a lo largo del tiempo. La guía y las prácticas recomendadas de este documento se dividen en cinco áreas de interés clave que sirven como principios rectores para crear soluciones en la nube eficientes en términos de rendimiento en AWS. Estas áreas de interés son:

- Selección de la arquitectura
- Computación y hardware
- Administración de datos
- Redes y entrega de contenido
- Proceso y cultura

Este documento está destinado a aquellos que ocupan puestos en tecnología, como los directores de tecnología (CTO), arquitectos, desarrolladores y miembros del equipo de operaciones. Después de leer este documento, comprenderá mejor qué prácticas recomendadas y estrategias de AWS se deben utilizar cuando diseñe una arquitectura eficiente en la nube.

Eficiencia del rendimiento

El pilar de la eficiencia del rendimiento se centra en el uso eficiente de los recursos de computación para satisfacer los requisitos y en cómo mantener la eficiencia a medida que la demanda cambia y las tecnologías evolucionan.

Temas

- Principios de diseño
- Definición

Principios de diseño

Los siguientes principios de diseño pueden ayudarle a conseguir y mantener cargas de trabajo eficientes en la nube.

- Democratizar las tecnologías avanzadas: facilite a su equipo la implementación de tecnologías avanzadas mediante la delegación de tareas complejas a su proveedor de servicios en la nube. En lugar de pedir a su equipo de TI que aprenda a alojar y ejecutar una tecnología nueva, considere la posibilidad de consumir la tecnología como un servicio. Por ejemplo, las bases de datos NoSQL, la transcodificación de medios y el machine learning son tecnologías que requieren conocimientos especializados. En la nube, estas tecnologías se convierten en servicios que su equipo puede consumir, lo que permite que su equipo se centre en el desarrollo de productos, y no en aprovisionar o administrar recursos.
- Adoptar un enfoque global en cuestión de minutos: el despliegue de su carga de trabajo en varias regiones de AWS del mundo le permite ofrecer una menor latencia y una mejor experiencia a sus clientes con un coste mínimo.
- Utilizar arquitecturas sin servidor: las arquitecturas sin servidor eliminan la necesidad de ejecutar y mantener servidores físicos para las actividades de computación tradicionales. Por ejemplo, los servicios de almacenamiento sin servidor pueden servir como sitios web estáticos, con lo que se elimina la necesidad de servidores web. Además, los servicios basados en eventos pueden alojar código. Esto elimina la carga operativa de administrar servidores físicos y puede reducir los costes de transacciones porque los servicios administrados operan a escala de la nube.
- Experimentar con más frecuencia: Los recursos virtuales y automatizables permiten realizar pruebas comparativas con rapidez mediante diferentes tipos de instancias, almacenamiento y configuraciones.

- Considerar la simpatía mecánica: utilice el enfoque tecnológico que mejor se adapte a sus objetivos. Por ejemplo, piense en los patrones de acceso a datos al elegir la base de datos o el almacenamiento de su carga de trabajo.

Definición

Céntrese en las siguientes áreas para lograr la eficiencia del rendimiento en la nube:

- Selección de la arquitectura
- Computación y hardware
- Administración de datos
- Redes y entrega de contenido
- Proceso y cultura

Adopte un enfoque basado en datos para crear una arquitectura de alto rendimiento. Recopile datos sobre todos los aspectos de la arquitectura, desde el diseño de alto nivel hasta la selección y configuración de los tipos de recursos.

Revisar periódicamente sus opciones le permitirá estar seguro de que aprovecha la continua evolución de la nube de AWS. Mediante la supervisión se asegura de conocer cualquier desviación del rendimiento esperado. Haga compensaciones en su arquitectura para mejorar el rendimiento, tales como el uso de la compresión o el almacenamiento en caché, o bien la mitigación de los requisitos de consistencia.

Selección de la arquitectura

La solución óptima para una carga de trabajo concreta varía y las soluciones suelen combinar varios enfoques. Las cargas de trabajo Well-Architected utilizan varias soluciones y admiten diferentes características para mejorar el rendimiento.

Los recursos de AWS están disponibles en muchos tipos y configuraciones, lo que facilita encontrar un enfoque que se ajuste a sus necesidades. También puede encontrar opciones que no se logran fácilmente con una infraestructura en las instalaciones. Por ejemplo, un servicio administrado como Amazon DynamoDB ofrece una base de datos NoSQL completamente administrada con una latencia de milisegundos de un solo dígito a cualquier escala.

Esta área de enfoque comparte guías y prácticas recomendadas sobre cómo seleccionar patrones de arquitectura y recursos en la nube eficientes y de alto rendimiento.

Prácticas recomendadas

- PERF01-BP01 Descubrir y comprender los servicios y las características disponibles en la nube
- PERF01-BP02 Seguir las recomendaciones de su proveedor de servicios en la nube o de un socio adecuado para conocer los modelos arquitectónicos y las prácticas recomendadas
- PERF01-BP03 Tener en cuenta los costes en sus decisiones arquitectónicas
- PERF01-BP04 Analizar cómo sus decisiones afectan a los clientes y a la eficiencia de la arquitectura
- PERF01-BP05 Usar políticas y arquitecturas de referencia
- PERF01-BP06 Realizar pruebas comparativas para tomar decisiones arquitectónicas
- PERF01-BP07 Aplicar un enfoque basado en los datos en sus decisiones arquitectónicas

PERF01-BP01 Descubrir y comprender los servicios y las características disponibles en la nube

Investigue continuamente los servicios y configuraciones disponibles que pueden ayudarle a tomar mejores decisiones arquitectónicas y a mejorar la eficiencia del rendimiento de la arquitectura de su carga de trabajo.

Patrones comunes de uso no recomendados:

- Utiliza la nube como un centro de datos coubicado.
- Después de migrar a la nube, no moderniza la aplicación.
- Utiliza un único tipo de almacenamiento para todo lo que necesita conservar.
- Utiliza los tipos de instancia que más se ajustan a sus estándares actuales, pero son más grandes cuando es necesario.
- Implementa y administra tecnologías que están disponibles como servicios administrados.

Beneficios de establecer esta práctica recomendada: al explorar nuevos servicios y configuraciones, es posible que pueda mejorar considerablemente el rendimiento, reducir los costes y optimizar el esfuerzo necesario para mantener la carga de trabajo. También podrá reducir el tiempo de amortización de los productos habilitados para la nube.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Alto

Guía para la implementación

AWS lanza nuevos servicios y características de forma continua que pueden mejorar el rendimiento y reducir el coste de las cargas de trabajo en la nube. Para mantener un rendimiento eficaz en la nube, es crucial estar al tanto de estos nuevos servicios y características. Modernizar la arquitectura de la carga de trabajo también le ayudará a acelerar la productividad, a impulsar la innovación y a descubrir más oportunidades de crecimiento.

Pasos para la implementación

- Haga un inventario del software y la arquitectura de su carga de trabajo para los servicios relacionados. Decida la categoría de productos sobre la que desea obtener más información.
- Explore las ofertas de AWS para identificar y conocer los servicios y las opciones de configuración pertinentes que pueden ayudarle a mejorar el rendimiento y a reducir los costes y la complejidad operativa.
  - Nube de Amazon Web Services
  - AWS Academy
  - Novedades en AWS
  - Blog de AWS
  - AWS Skill Builder
  - Eventos y Webinars de AWS

  - Formación de AWS and Certifications
  - Canal de YouTube de AWS
  - Talleres de AWS
  - Comunidades de AWS
- Utilice entornos aislados (que no sean de producción) para aprender y experimentar con los nuevos servicios sin incurrir en costes extraordinarios.
- Obtenga información continua sobre los nuevos servicios y características de la nube.

Recursos

Documentos relacionados:

- Descripción general de Amazon Web Services
- Características de Amazon EC2
- Aprenda paso a paso con un plan de aprendizaje para socios de AWS
- AWS Training and Certification
- My learning path to become an AWS solutions architect
- Centro de arquitectura de AWS
- AWS Partner Network
- La Biblioteca de soluciones de AWS
- Centro de conocimiento de AWS
- Build modern applications on AWS

Vídeos relacionados:

- AWS re:Invent 2023 - What's new with Amazon EC2
- AWS re:Invent 2022 - Reduce your operational and infrastructure costs with Amazon ECS
- AWS re:Invent 2023 - Build with the efficiency, agility & innovation of the cloud with AWS
- AWS re:Invent 2022 - Deploy ML models for inference at high performance and low cost
- This is my Architecture

Ejemplos relacionados:

- Ejemplos de AWS
- Ejemplos de SDK de AWS

PERF01-BP02 Seguir las recomendaciones de su proveedor de servicios en la nube o de un socio adecuado para conocer los modelos arquitectónicos y las prácticas recomendadas

Utilice los recursos corporativos de la nube, como la documentación, los arquitectos de soluciones, los servicios profesionales o los socios adecuados, para que le sirvan de guía en sus decisiones arquitectónicas. Estos recursos le ayudarán a revisar y mejorar su arquitectura para obtener un rendimiento óptimo.

Patrones comunes de uso no recomendados:

- Utiliza AWS como un proveedor de servicios en la nube al uso.
- Utiliza los servicios de AWS de una manera para la que no fueron diseñados.
- Sigue todas las directrices sin tener en cuenta su contexto empresarial.

Beneficios de establecer esta práctica recomendada: seguir las directrices de un proveedor de servicios en la nube o de un socio adecuado puede ayudarle a tomar las decisiones arquitectónicas correctas para su carga de trabajo y a ganar confianza en sus decisiones.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Medio

Guía para la implementación

AWS ofrece un gran número de directrices, documentación y recursos que pueden ayudarle a crear y gestionar cargas de trabajo en la nube de forma eficiente. La documentación de AWS contiene ejemplos de código, tutoriales y explicaciones detalladas de los servicios. Además de la documentación, AWS ofrece programas de formación y certificación, arquitectos de soluciones y servicios profesionales que pueden ayudar a los clientes a explorar diferentes aspectos de los servicios en la nube y a implementar una arquitectura de nube eficiente en AWS.

Aproveche estos recursos para obtener valiosos conocimientos y prácticas recomendadas, ahorrar tiempo y lograr mejores resultados en la Nube de AWS.

Pasos para la implementación

- Revise la documentación y las directrices de AWS y siga las prácticas recomendadas. Estos recursos pueden ayudarle a elegir y configurar los servicios de manera eficaz y a lograr un mejor rendimiento.
  - Documentación de AWS (como guías de usuario y documentos técnicos)
  - Blog de AWS
  - Formación de AWS and Certifications
  - Canal de YouTube de AWS
- Únase a los eventos de los socios de AWS (como los AWS Global Summits, AWS re:invent, grupos de usuarios y talleres) para aprender de la mano de expertos de AWS las prácticas recomendadas acerca de cómo usar los servicios de AWS.
  - Aprenda paso a paso con un plan de aprendizaje para socios de AWS
  - Eventos y Webinars de AWS
  - Talleres de AWS
  - Comunidades de AWS
- Póngase en contacto con AWS cuando necesite más ayuda o información sobre un producto. Los arquitectos de soluciones de AWS y los servicios profesionales de AWS proporcionan orientación para la implementación de soluciones. Los socios de AWS ponen a su disposición la experiencia de AWS para ayudarle a mejorar la agilidad y la innovación para su empresa.
- Utilice AWS Support si necesita soporte técnico para usar un servicio de forma eficaz. Nuestros planes de soporte están diseñados para brindarle la combinación perfecta de herramientas y ofrecerle acceso a conocimientos especializados para que pueda tener éxito con AWS mientras optimiza el rendimiento, administra los riesgos y mantiene los costes bajo control.

Recursos

Documentos relacionados:

- Centro de arquitectura de AWS
- AWS Partner Network
- La Biblioteca de soluciones de AWS
- Centro de conocimiento de AWS
- AWS Enterprise Support

Vídeos relacionados:

- This is my Architecture
- AWS re:Invent 2023 - Advanced event-driven patterns with Amazon EventBridge
- AWS re:Invent 2023 - Implementing distributed design patterns on AWS
- AWS re:Invent 2023 - Application architecture as code

Ejemplos relacionados:

- Ejemplos de AWS
- Ejemplos de SDK de AWS
- Arquitectura de referencia de análisis de AWS

PERF01-BP03 Tener en cuenta los costes en sus decisiones arquitectónicas

Tenga en cuenta los costes en sus decisiones arquitectónicas para mejorar la utilización de los recursos y la eficiencia del rendimiento de su carga de trabajo en la nube. Si conoce las implicaciones financieras de su carga de trabajo en la nube, es más probable que aproveche los recursos de forma eficiente y reduzca las prácticas innecesarias.

Patrones comunes de uso no recomendados:

- Solo utiliza una familia de instancias.
- No contempla la posibilidad de utilizar soluciones con licencia en lugar de soluciones de código abierto.
- No tienen políticas definidas sobre el ciclo de vida del almacenamiento.
- No revisa los nuevos servicios y características de la Nube de AWS.
- Solo utiliza el almacenamiento de bloques.

Beneficios de establecer esta práctica recomendada: si tiene en cuenta los costes a la hora de tomar decisiones, tendrá la oportunidad de utilizar recursos más eficientes y explorar otras inversiones.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Medio

Guía para la implementación

Si optimiza las cargas de trabajo con arreglo a los costes, puede mejorar la utilización de los recursos y evitar pérdidas en una carga de trabajo en la nube. Por lo general, al contemplar los costes en las decisiones de arquitectura, los componentes de la carga de trabajo se dimensionan correctamente y se favorece la elasticidad, lo que se traduce en una mejora de la eficiencia del rendimiento de las cargas de trabajo en la nube.

Pasos para la implementación

- Establezca objetivos de costes, como los límites presupuestarios de la carga de trabajo en la nube.
- Identifique los componentes clave (como las instancias y el almacenamiento) que influyen en los costes de su carga de trabajo. Puede usar el AWS Pricing Calculator y AWS Cost Explorer para identificar los principales factores que influyen en los costes de su carga de trabajo.
- Comprenda los modelos de precios en la nube, como instancias bajo demanda, reservadas, Savings Plans e instancias de spot.
- Utilice las prácticas recomendadas de optimización de costes de Well-Architected para optimizar los costes de estos componentes clave.
- Supervise y analice los costes de forma continua para identificar oportunidades que le permitan optimizar los gastos de su carga de trabajo.
  - Utilice AWS Budgets para recibir alertas sobre costes inaceptables.
  - Utilice AWS Compute Optimizer o bien AWS Trusted Advisor para obtener recomendaciones sobre la optimización de costes.
  - Utilice la detección de anomalías en los costes de AWS para detectar automáticamente las anomalías en los costes y analizar la causa raíz.

Recursos

Documentos relacionados:

- What is AWS Billing and Cost Management?
- Optimización de costes con AWS
- Choosing an AWS cost management strategy
- A Beginner's Guide to AWS Cost Management
- A Detailed Overview of the Cost Intelligence Dashboard

- Centro de arquitectura de AWS
- La Biblioteca de soluciones de AWS
- Centro de conocimiento de AWS

Vídeos relacionados:

- This is my Architecture
- AWS re:Invent 2023 - What's new with AWS cost optimization
- AWS re:Invent 2023 - Optimize cost and performance and track progress toward mitigation
- AWS re:Invent 2023 - AWS storage cost-optimization best practices
- AWS re:Invent 2023 - Optimize costs in your multi-account environments

Ejemplos relacionados:

- Código de demostración de AWS Compute Optimizer
- Taller de optimización de costes
- Cloud Financial Management Technical Implementation Playbooks
- Optimización de startups: ajuste del rendimiento de las aplicaciones para lograr la máxima eficiencia
- Serverless Optimization Workshop (Performance and Cost)
- Scaling cost effective architectures

PERF01-BP04 Analizar cómo sus decisiones afectan a los clientes y a la eficiencia de la arquitectura

Cuando evalúe las mejoras relacionadas con el rendimiento, debe determinar qué decisiones afectarán a sus clientes y a la eficiencia de la carga de trabajo. Por ejemplo, si el uso de un almacén de datos clave-valor mejora el rendimiento del sistema, es importante analizar cómo la naturaleza eventualmente consistente de este cambio afectaría a los clientes.

Patrones comunes de uso no recomendados:

- Da por hecho que habría que implementar todas las ventajas relacionadas con el rendimiento, aunque esta implementación tenga repercusiones.

- Solo evalúa los cambios en las cargas de trabajo cuando un problema de rendimiento ha alcanzado un punto crítico.

Beneficios de establecer esta práctica recomendada: Al evaluar las mejoras potenciales relacionadas con el rendimiento, debe decidir si las compensaciones que exigen los cambios son aceptables de acuerdo con los requisitos de la carga de trabajo. En algunos casos, es posible que tenga que implementar controles adicionales para contrarrestar estas repercusiones.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Alto

Guía para la implementación

Identifique las áreas críticas de la arquitectura en términos de cómo afectan al rendimiento y a los clientes. Determine cómo puede hacer mejoras, qué repercusiones tienen esas mejoras y cómo afectan al sistema y a la experiencia del usuario. Por ejemplo, la implementación de datos en caché puede mejorar drásticamente el rendimiento, pero requiere una estrategia clara sobre cómo y cuándo actualizar o invalidar los datos en caché para evitar un comportamiento incorrecto del sistema.

Pasos para la implementación

- Estudie los requisitos de la carga de trabajo y los SLA.
- Defina claramente los factores de la evaluación. Estos factores pueden estar relacionados con los costes, la fiabilidad, la seguridad y el rendimiento de su carga de trabajo.
- Seleccione una arquitectura y unos servicios que puedan satisfacer sus necesidades.
- Realice experimentos y pruebas de conceptos (POC) para analizar las repercusiones y el impacto que pueden tener en los clientes y en la eficiencia de la arquitectura. Por lo general, las cargas de trabajo seguras, de alto rendimiento y de alta disponibilidad consumen más recursos de la nube, aunque proporcionan una mejor experiencia al cliente. Comprenda las ventajas y desventajas de la complejidad, el rendimiento y el coste de su carga de trabajo. Por lo general, priorizar dos de los factores se produce a expensas del tercero.

Recursos

Documentos relacionados:

- Amazon Builders' Library
- Amazon QuickSight KPIs

- Amazon CloudWatch RUM
- Documentación de X-Ray
- Understand resiliency patterns and trade-offs to architect efficiently in the cloud

Vídeos relacionados:

- Optimize applications through Amazon CloudWatch RUM
- AWS re:Invent 2023 - Capacity, availability, cost efficiency: Pick three
- AWS re:Invent 2023 - Advanced integration patterns & trade-offs for loosely coupled systems

Ejemplos relacionados:

- Medición del tiempo de carga de la página con Amazon CloudWatch Synthetics
- Cliente web de Amazon CloudWatch RUM

PERF01-BP05 Usar políticas y arquitecturas de referencia

Cuando elija los servicios y las configuraciones, utilice políticas internas y arquitecturas de referencia existentes para ser más eficiente al diseñar e implementar su carga de trabajo.

Patrones comunes de uso no recomendados:

- Permite usar una gran variedad de tecnologías, lo que puede incidir en los gastos generales de administración de la empresa.

Beneficios de establecer esta práctica recomendada: establecer una política para la elección de la arquitectura, la tecnología y el proveedor permite tomar decisiones de forma rápida.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Medio

Guía para la implementación

Contar con políticas internas para seleccionar los recursos y la arquitectura proporciona estándares y pautas que pueden seguirse al tomar decisiones arquitectónicas. Estas directrices agilizan el proceso de toma de decisiones a la hora de elegir el servicio de nube correcto y pueden ayudar a mejorar la eficiencia del rendimiento. Despliegue la carga de trabajo utilizando políticas o arquitecturas de

referencia. Integre los servicios en su despliegue en la nube y, a continuación, utilice las pruebas de rendimiento para asegurarse de que puede seguir cumpliendo los requisitos establecidos.

Pasos para la implementación

- Conozca al detalle los requisitos de su carga de trabajo en la nube.
- Consulte políticas internas y externas para identificar las más relevantes.
- Utilice las arquitecturas de referencia adecuadas que le ofrece AWS o las prácticas recomendadas por el sector.
- Cree un conjunto coherente de políticas, estándares, arquitecturas de referencia y pautas prescriptivas para situaciones comunes. De este modo, sus equipos podrán avanzar más rápido. Adapte los activos a su sector, si procede.
- Coteje estas políticas y arquitecturas de referencia con su carga de trabajo en entornos aislados.
- Manténgase al tanto de los estándares sectoriales y las actualizaciones de AWS para asegurarse de que las políticas y las arquitecturas de referencia le ayudan a optimizar su carga de trabajo en la nube.

Recursos

Documentos relacionados:

- Centro de arquitectura de AWS
- AWS Partner Network
- La Biblioteca de soluciones de AWS
- Centro de conocimiento de AWS
- AWS Architecture Blog

Vídeos relacionados:

- This is my Architecture
- AWS re:Invent 2022 - Accelerate value for your business with SAP & AWS reference architecture

Ejemplos relacionados:

- Ejemplos de AWS

- Ejemplos de SDK de AWS

PERF01-BP06 Realizar pruebas comparativas para tomar decisiones arquitectónicas

Mida el rendimiento de una carga de trabajo existente para entender cómo rinde en la nube y fundamentar sus decisiones arquitectónicas en esos datos.

Antipatrones usuales:

- Utiliza pruebas comparativas de uso común que no son indicativas de las características concretas de su carga de trabajo.
- La única referencia que tiene en cuenta son los comentarios y las percepciones de los clientes.

Ventajas de aplicar esta práctica recomendada: realizar pruebas comparativas en la implementación actual le permite medir las mejoras de rendimiento.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: medio

Guía para la implementación

Utilice la evaluación comparativa con pruebas sintéticas para evaluar el rendimiento de los componentes de su carga de trabajo. Las pruebas comparativas suelen ser más rápidas de configurar que las pruebas de carga y se utilizan para evaluar la tecnología de un componente concreto. Estas pruebas comparativas suelen usarse al comienzo de un nuevo proyecto, cuando aún no se tiene una solución completa para realizar una prueba de carga.

Para comparar las cargas de trabajo, puede diseñar sus propias pruebas comparativas personalizadas o usar una prueba estándar del sector, como TPC-DS. Las pruebas comparativas sectoriales son útiles cuando se comparan entornos. Los puntos de referencia personalizados son útiles para encontrar tipos específicos de operaciones que espera realizar en su arquitectura.

Con las pruebas comparativas, es importante realizar los preparativos necesarios en el entorno de prueba para asegurarse de que los resultados obtenidos son válidos. Ejecute la misma comparativa muchas veces para asegurarse de que detecta cualquier variación que haya podido surgir con el tiempo.

Como las pruebas comparativas por lo general se ejecutan más rápido que las pruebas de carga, pueden usarse antes en la canalización de despliegue para y proporcionan información de una forma más rápida sobre las desviaciones del rendimiento. Al evaluar un cambio importante en un componente o servicio, puede resultar más rápido usar una prueba comparativa para determinar si el esfuerzo que conlleva el cambio es justificable. Es importante usar pruebas de carga junto con las pruebas comparativas, ya que las pruebas de carga le informan del rendimiento de la carga de trabajo en producción.

Pasos para la implementación

- Planificar y definir:
  - Defina los objetivos, la base de referencia, los escenarios de prueba, las métricas (como la utilización de la CPU, la latencia o el rendimiento) y los KPI para el punto de referencia.
  - Céntrese en los requisitos de los usuarios en lo que respecta a la experiencia de usuario y factores como el tiempo de respuesta y la accesibilidad.
  - Identifique una herramienta de pruebas comparativas que sea adecuada para su carga de trabajo. Puede usar los servicios de AWS (como Amazon CloudWatch) o una herramienta de terceros que sea compatible con la carga de trabajo.
- Configurar e instrumentar:
  - Configure el entorno y los recursos.
  - Implemente la supervisión y el registro para recopilar los resultados de las pruebas.
- Comparar y supervisar:
  - Realice las pruebas comparativas y supervise las métricas durante la prueba.
- Analizar y documentar:
  - Documente el proceso de evaluación comparativa y los resultados.
  - Analice los resultados para identificar los cuellos de botella, las tendencias y las áreas de mejora.
  - Utilice los resultados de las pruebas para tomar decisiones arquitectónicas y ajustar la carga de trabajo. Para ello, puede ser necesario cambiar los servicios o adoptar nuevas características.
- Optimizar y repetir:
  - Ajuste las configuraciones y asignaciones de los recursos en función de los puntos de referencia.
  - Vuelva a probar la carga de trabajo después del ajuste para validar las mejoras.
  - Documente la información obtenida y repita el proceso para identificar otras áreas de mejora.

Recursos

Documentos relacionados:

- Centro de arquitectura de AWS
- AWS Partner Network
- Biblioteca de soluciones de AWS
- Centro de conocimientos de AWS
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Genomics workflows, Part 5: automated benchmarking
- Benchmark and optimize endpoint deployment in Amazon SageMaker JumpStart

Vídeos relacionados:

- AWS re:Invent 2023 - Benchmarking AWS Lambda cold starts
- Benchmarking stateful services in the cloud
- This is my Architecture
- Optimize applications through Amazon CloudWatch RUM
- Demostración de Amazon CloudWatch Synthetics

Ejemplos relacionados:

- AWS Samples
- AWS SDK Examples
- Pruebas de carga distribuidas
- Measure page load time with Amazon CloudWatch Synthetics
- Amazon CloudWatch RUM Web Client

PERF01-BP07 Aplicar un enfoque basado en los datos en sus decisiones arquitectónicas

Defina un enfoque claro basado en los datos para utilizarlo cuando tome decisiones arquitectónicas y asegurarse de que se utilizan los servicios y las configuraciones de nube correctos para satisfacer las necesidades específicas de su empresa.

Patrones comunes de uso no recomendados:

- Presupone que la arquitectura actual es estática y no debe actualizarse con el tiempo.
- Las decisiones arquitectónicas que toma se basan en conjeturas y suposiciones.
- Se introducen cambios en la arquitectura a lo largo del tiempo sin justificación.

Beneficios de establecer esta práctica recomendada: al contar con un enfoque bien definido y aplicarlo a la hora de optar por las opciones arquitectónicas, se utilizan los datos para influir en el diseño de la carga de trabajo y tomar decisiones fundamentadas a lo largo del tiempo.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Medio

Guía para la implementación

Para seleccionar los recursos y los servicios de su arquitectura, aproveche la experiencia y los conocimientos sobre la nube del personal interno o utilice recursos externos, como los casos de uso publicados o los documentos técnicos. Debe contar con un proceso bien definido que contribuya a probar y comparar los servicios que podrían utilizarse en su carga de trabajo.

La lista de tareas pendientes para las cargas de trabajo críticas no solo debe incluir casos de usuario que brinden una funcionalidad relevante para la empresa y los usuarios, sino también casos técnicos que conformen un plan arquitectónico para la carga de trabajo. Este plan se nutre de nuevos avances en tecnología y nuevos servicios, que se incorporan con arreglo a los datos y de forma justificada. Esto garantiza que la arquitectura siempre está preparada para el futuro y no se queda anquilosada.

Pasos para la implementación

- Hable con las principales partes interesadas para definir los requisitos de la carga de trabajo, incluidas las consideraciones de rendimiento, disponibilidad y costes. Tenga en cuenta factores como la cantidad de usuarios y el modo de uso de la carga de trabajo.

- Cree un plan arquitectónico o una lista de tareas pendientes relacionadas con la tecnología que tengan la misma prioridad que las tareas pendientes relacionadas con la funcionalidad.
- Evalúe los diferentes servicios en la nube (para obtener más información, consulte PERF01-BP01 Descubrir y comprender los servicios y las características disponibles en la nube).
- Analice diferentes patrones arquitectónicos, como los microservicios o la computación sin servidor, que se ajusten a sus requisitos de rendimiento (para obtener más información, consulte PERF01-BP02 Seguir las recomendaciones de su proveedor de servicios en la nube o de un socio adecuado para conocer los modelos arquitectónicos y las prácticas recomendadas).

- Consulte otros equipos, diagramas de arquitectura y recursos, como los arquitectos de soluciones de AWS, Centro de arquitectura de AWSy AWS Partner Network, para ayudarle a elegir la arquitectura adecuada para su carga de trabajo.

- Defina métricas, como el rendimiento y el tiempo de respuesta, que puedan ayudarle a evaluar el desempeño de su carga de trabajo.
- Pruebe y utilice las métricas definidas para validar el rendimiento de la arquitectura seleccionada.
- Mantenga un control continuo y realice los ajustes necesarios para garantizar el rendimiento óptimo de su arquitectura.
- Documente la arquitectura seleccionada y las decisiones adoptadas de forma que sirvan de referencia para futuras actualizaciones y formaciones.
- Revise y actualice continuamente el enfoque de selección de arquitectura con arreglo a los nuevos conocimientos, las nuevas tecnologías y las métricas que indiquen un cambio necesario o un problema en el enfoque actual.

Recursos

Documentos relacionados:

- La Biblioteca de soluciones de AWS
- Centro de conocimiento de AWS
- Architectural Patterns to Build End-to-End Data Driven Applications on AWS

Vídeos relacionados:

- This is my Architecture
- AWS re:Invent 2021 - Data-driven enterprise: Going from vision to value
- AWS re:Invent 2022 - Delivering sustainable, high-performing architectures
- AWS re:Invent 2023 - Optimize cost and performance and track progress toward mitigation
- AWS re:Invent 2022 - AWS optimization: Actionable steps for immediate results

Ejemplos relacionados:

- Ejemplos de AWS
- Ejemplos de SDK de AWS

Computación y hardware

La elección óptima de computación para una carga de trabajo concreta puede variar en función del diseño de la aplicación, los patrones de uso y los ajustes de configuración. Las arquitecturas pueden usar diferentes opciones de computación para varios componentes y admiten diferentes características para mejorar el rendimiento. Seleccionar la opción de computación incorrecta para una arquitectura puede disminuir la eficiencia del rendimiento.

Esta área de interés comparte guías y prácticas recomendadas sobre cómo identificar y optimizar las opciones de computación para lograr la eficiencia del rendimiento en la nube.

Prácticas recomendadas

- PERF02-BP01 Seleccionar las mejores opciones computacionales para su carga de trabajo
- PERF02-BP02 Comprender las opciones de configuración y las características de computación disponibles
- PERF02-BP03 Recopilar métricas relacionadas con la computación
- PERF02-BP04 Configurar y dimensionar correctamente los recursos de computación
- PERF02-BP05 Escalar los recursos computacionales de forma dinámica
- PERF02-BP06 Utilización de aceleradores computacionales optimizados basados en hardware

PERF02-BP01 Seleccionar las mejores opciones computacionales para su carga de trabajo

Si selecciona la opción computacional más adecuada para su carga de trabajo, podrá mejorar el rendimiento, reducir los costes de infraestructura innecesarios y aligerar los esfuerzos operativos necesarios para mantener esa carga de trabajo.

Antipatrones usuales:

- Se utiliza la misma opción computacional que en el entorno local.
- No se tiene información suficiente sobre las opciones de computación, las características y las soluciones de la nube, y cómo estas podrían mejorar el rendimiento informático.
- Se ha sobreaprovisionado una opción de computación existente para cumplir los requisitos de escalamiento o rendimiento cuando una opción de computación alternativa se ajustaría con mayor precisión a las características de la carga de trabajo.

Ventajas de aplicar esta práctica recomendada: al identificar los requisitos de computación y evaluarlos con arreglo a las opciones disponibles, puede hacer que su carga de trabajo sea más eficiente en términos de recursos.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: alto

Guía para la implementación

Para optimizar las cargas de trabajo en la nube y lograr un rendimiento eficiente, es importante seleccionar las opciones de computación más adecuadas para su caso de uso y los requisitos de rendimiento. AWS ofrece una variedad de opciones de computación que se adaptan a diferentes cargas de trabajo en la nube. Por ejemplo, puede utilizar Amazon EC2 para lanzar y administrar servidores virtuales, AWS Lambda para ejecutar código sin tener que aprovisionar ni administrar servidores, Amazon ECS o Amazon EKS para ejecutar y administrar contenedores, oAWS Batch para procesar grandes volúmenes de datos en paralelo. En función de sus necesidades de computación y escalamiento, debe elegir y configurar la solución computacional que sea óptima para su caso. También puede considerar la posibilidad de usar diferentes tipos de soluciones computacionales en una misma carga de trabajo, ya que cada una de ellas tiene sus propias ventajas e inconvenientes.

Los siguientes pasos le ayudarán a seleccionar las opciones computacionales adecuadas que se adaptan a las características de su carga de trabajo y a los requisitos de rendimiento.

Pasos para la implementación

- Sepa cuáles son los requisitos computacionales de su carga de trabajo. Algunos de los principales requisitos son las necesidades de procesamiento, los patrones de tráfico, los patrones de acceso a los datos, las necesidades de escalamiento y los requisitos de latencia.
- Descubra las diferentes opciones de computación disponibles para su carga de trabajo en AWS (tal y como se indica en PERF01-BP01 Descubrir y comprender los servicios y las características disponibles en la nube. Estas son algunas de las opciones de computación clave de AWS, sus características y casos de uso comunes:

AWS service	Key characteristics	Common use cases
Amazon Elastic Compute Cloud (Amazon EC2)	Has dedicated option for hardware, license requirements, large selection of different	Lift and shift migrations, monolithic application, hybrid

AWS service	Key characteristics	Common use cases
	instance families, processor types and compute accelerators	environments, enterprise applications
Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS)	Easy deployment, consistent environments, scalable	Microservices, hybrid environments
AWS Lambda	Computación sin servidor service that runs code in response to events and automatically manages the underlying compute resources.	Microservices, event-driven applications
AWS Batch	Efficiently and dynamically provisions and scales Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS), and AWS Fargate compute resources, with an option to use On-Demand or Spot Instances based on your job requirements	HPC, train ML models
Amazon Lightsail	Preconfigured Linux and Windows application for running small workloads	Simple web applications, custom website

- Calcule el coste (por ejemplo, el coste por hora o la transferencia de datos) y los gastos generales de administración (como la aplicación de parches y el escalamiento) asociados a cada opción de computación.

- Realice experimentos y pruebas comparativas en un entorno que no sea de producción para identificar qué opción de computación puede satisfacer mejor los requisitos de su carga de trabajo.
- Una vez que haya probado e identificado su nueva solución de computación, planifique la migración y valide sus métricas de rendimiento.
- Utilice las herramientas de supervisión de AWS, como Amazon CloudWatch, y los servicios de optimización, como AWS Compute Optimizer, para optimizar continuamente los recursos de computación en función de los patrones de uso reales.

Recursos

Documentos relacionados:

- Cloud Compute with AWS (Computación en la nube con AWS)
- Amazon EC2 Instance Types
- Amazon EKS Containers: Amazon EKS Worker Nodes
- Amazon ECS Containers: Amazon ECS Container Instances
- Functions: Lambda Function Configuration
- Prescriptive Guidance for Containers
- Prescriptive Guidance for Serverless

Vídeos relacionados:

- AWS re:Invent 2023 - AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 - New Amazon Elastic Compute Cloud generative AI capabilities in AMS
- AWS re:Invent 2023 - What's new with Amazon Elastic Compute Cloud
- AWS re:Invent 2023 - Smart savings: Amazon Elastic Compute Cloud cost-optimization strategies
- AWS re:Invent 2021 - Powering next-gen Amazon Elastic Compute Cloud: Deep dive on the Nitro System
- AWS re:Invent 2019 - Optimize performance and cost for your AWS compute
- AWS re:Invent 2019 - Amazon Elastic Compute Cloud foundations
- AWS re:Invent 2022 - Deploy ML models for inference at high performance and low cost
- AWS re:Invent 2019 - Optimize performance and cost for your AWS compute
- Amazon EC2 foundations

- Deploy ML models for inference at high performance and low cost

Ejemplos relacionados:

- Migrating the Web application to containers
- Run a Serverless Hello World
- Amazon EKS Workshop
- Amazon EC2 Workshop
- Efficient and Resilient Workloads with Amazon Elastic Compute Cloud Auto Scaling
- Migrating to AWS Graviton with Container Services

PERF02-BP02 Comprender las opciones de configuración y las características de computación disponibles

Conozca las opciones de configuración y las características disponibles para su servicio de computación, lo que le ayudará a aprovisionar la cantidad de recursos adecuada y a conseguir un rendimiento más eficiente.

Patrones comunes de uso no recomendados:

- No evalúan las opciones de computación ni las familias de instancias disponibles con arreglo a las características de la carga de trabajo.
- Produce un aprovisionamiento excesivo de recursos informáticos para satisfacer los picos de demanda.

Beneficios de establecer esta práctica recomendada: familiarícese con las configuraciones y las características computacionales de AWS para utilizar una solución computacional optimizada que se ajuste a las características y necesidades de su carga de trabajo.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Medio

Guía para la implementación

Cada solución computacional tiene disponibles configuraciones y características únicas que admiten diferentes características y requisitos de la carga de trabajo. Descubra cómo estas opciones complementan su carga de trabajo y determine qué opciones de configuración son mejores para su

caso. Algunas de estas opciones pueden ser, por ejemplo, la familia de instancias, el tamaño, las características (GPU, E/S, etc.), la capacidad de ampliación, los tiempos de espera, los tamaños de funciones, las instancias de contenedor y la simultaneidad. Si su carga de trabajo ha estado utilizando la misma opción de computación durante más de cuatro semanas y prevé que las características seguirán siendo las mismas en el futuro, puede utilizar AWS Compute Optimizer para averiguar si la opción de computación actual es adecuada para las cargas de trabajo desde el punto de vista de la CPU y la memoria.

Pasos para la implementación

1. Sepa cuáles son los requisitos de la carga de trabajo (como los requisitos de CPU, la memoria y la latencia).
2. Consulte la documentación y las prácticas recomendadas de AWS para obtener información sobre las opciones de configuración recomendadas que pueden ayudar a mejorar el rendimiento computacional. Estas son algunas de las principales opciones de configuración que debe tener en cuenta:

Opción de configuración	Ejemplos
Tipo de instancia	• Las instancias optimizadas para la computación son ideales para las cargas de trabajo que requieren una relación entre vCPU y memoria más alta. • Las instancias optimizadas para la memoria ofrecen grandes cantidades de memoria para admitir cargas de trabajo que hacen un uso intensivo de la memoria. • Las instancias optimizadas para el almacenamiento están diseñadas para cargas de trabajo que requieren un alto acceso secuencial de lectura y escritura (IOPS) al almacenamiento local.
Modelo de precios	• Las instancias bajo demanda le permiten utilizar la capacidad de computación por horas o por segundos sin compromis o a largo plazo. Estas instancias son

Opción de configuración	Ejemplos
	adecuadas para ampliar la capacidad por encima de las necesidades de rendimiento estándar. • Savings Plans ofrecen un ahorro significativo en comparación con las instancias bajo demanda a cambio del compromiso de utilizar una cantidad específica de capacidad de computación durante un período de uno o tres años. • Las instancias de spot le permiten aprovechar la capacidad de las instancias que no se utilizan en cargas de trabajo sin estado y tolerantes a errores con descuento.
Auto Scaling	Utilice la configuración de Auto Scaling para ajustar los recursos computacionales con los patrones de tráfico.
Tamaño	• Utilice Compute Optimizer para obtener recomendaciones con tecnología de machine learning sobre qué configuración de computación se ajusta mejor a sus características de computación. • Utilice el ajuste de potencia de AWS Lambda para seleccionar la mejor configuración para su función Lambda.

Opción de configuración	Ejemplos
Aceleradores de cómputo basados en hardware	• Las instancias de computación acelerada ejecutan funciones, como procesamiento de gráficos o búsqueda de patrones de datos, de manera más eficiente que las alternativas basadas en CPU. • Para las cargas de trabajo de machine learning, utilice hardware personalizado específico para su carga de trabajo, como AWS Trainium, AWS Inferentia y Amazon EC2 DL1

Recursos

Documentos relacionados:

- Cloud Compute with AWS
- Amazon EC2 Instance Types
- Control de los estados del procesador de la instancia Amazon EC2
- Amazon EKS Containers: Amazon EKS Worker Nodes
- Amazon ECS Containers: Amazon ECS Container Instances
- Funciones: configuración de funciones de Lambda

Vídeos relacionados:

- AWS re:Invent 2023 – AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 – New Amazon EC2 generative AI capabilities in AWS Management Console
- AWS re:Invent 2023 – What's new with Amazon EC2
- AWS re:Invent 2023 – Smart savings: Amazon EC2 cost-optimization strategies
- AWS re:Invent 2021 – Powering next-gen Amazon EC2: Deep dive on the Nitro System
- AWS re:Invent 2019 – Amazon EC2 foundations
- AWS re:Invent 2022 – https://www.youtube.com/watch?v=5B4-s_ivn1o

Ejemplos relacionados:

- Código de demostración de Compute Optimizer
- Taller sobre instancias de spot de Amazon EC2
- Efficient and Resilient Workloads with Amazon EC2 AWS Auto Scaling
- Taller para desarrolladores de Graviton
- AWS for Microsoft Workloads Immersion Day
- AWS for Linux Workloads Immersion Day
- Código de demostración de AWS Compute Optimizer
- Taller de Amazon EKS

PERF02-BP03 Recopilar métricas relacionadas con la computación

Registre y supervise las métricas relacionadas con los recursos de computación para comprender mejor el rendimiento de los recursos informáticos y mejorar su rendimiento y su utilización.

Patrones comunes de uso no recomendados:

- Solo se utiliza la búsqueda manual de métricas en los archivos de registro.
- Solo utiliza las métricas predeterminadas registradas en el software de supervisión seleccionado.
- Solo se revisan las métricas cuando hay un problema.

Beneficios de establecer esta práctica recomendada: recopilar métricas relacionadas con el rendimiento le ayudará a ajustar el rendimiento de las aplicaciones a los requisitos empresariales para garantizar que cumple con las necesidades de su carga de trabajo. También puede ayudarle a mejorar continuamente el rendimiento y la utilización de los recursos en su carga de trabajo.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Alto

Guía para la implementación

Las cargas de trabajo en la nube pueden generar grandes volúmenes de datos, como métricas, registros y eventos. En Nube de AWS, la recopilación de métricas es un paso crucial para mejorar la seguridad, la rentabilidad, el rendimiento y la sostenibilidad. AWS ofrece una amplia variedad de métricas relacionadas con el rendimiento a través de servicios de supervisión como Amazon CloudWatch, que le proporcionan una información valiosa. Las métricas como la utilización de

la CPU, la utilización de la memoria, las operaciones de E/S del disco y la entrada y salida de la red pueden proporcionar información sobre los niveles de utilización o los cuellos de botella del rendimiento. Utilice estas métricas como parte de un enfoque basado en datos para ajustar y optimizar activamente los recursos de su carga de trabajo. En un supuesto ideal, debería recopilar todas las métricas relacionadas con sus recursos de computación en una única plataforma que tuviera políticas de retención implementadas para satisfacer los objetivos operativos y financieros.

Pasos para la implementación

1. Identifique qué métricas relacionadas con el rendimiento son relevantes para su carga de trabajo. Debe recopilar métricas sobre la utilización de los recursos y la forma en que funciona su carga de trabajo en la nube (por ejemplo, el tiempo de respuesta y el rendimiento).
    a. Métricas predeterminadas de Amazon EC2
    b. Métricas predeterminadas de Amazon ECS
    c. Métricas predeterminadas de Amazon EKS
    d. Métricas predeterminadas de Lambda
    e. Métricas de memoria y disco de Amazon EC2
2. Elija y configure la solución de registro y supervisión adecuada para su carga de trabajo.
    a. Observabilidad nativa de AWS
    b. AWS Distro para OpenTelemetry
    c. Amazon Managed Service for Prometheus
3. Defina el filtro y la agregación que se necesitan para las métricas en función de los requisitos de su carga de trabajo.
    a. Cuantifique métricas de aplicación personalizadas con Amazon CloudWatch Logs y filtros de métrica
    b. Recopile métricas personalizadas con el etiquetado estratégico de Amazon CloudWatch
4. Configure políticas de retención de datos para que las métricas se ajusten a los objetivos operativos y de seguridad.
    a. Retención de datos predeterminada para métricas de CloudWatch
    b. Retención de datos predeterminada para CloudWatch Logs
5. Si es necesario, cree alarmas y notificaciones para sus métricas, lo que le ayudará a responder de manera proactiva a los problemas relacionados con el rendimiento.
    a. Cree alarmas para métricas personalizadas con la detección de anomalías de Amazon CloudWatch

   b. Cree métricas y alarmas para páginas web específicas con Amazon CloudWatch RUM

6. Utilice la automatización para desplegar los agentes de agregación de métricas y registros.
   a. Automatización de AWS Systems Manager
   b. Colector de OpenTelemetry

Recursos

Documentos relacionados:

- Monitoreo y observabilidad
- Best practices: implementing observability with AWS
- Documentación de Amazon CloudWatch
- Recopilación de métricas y registros de instancias Amazon EC2 y en los servidores en las instalaciones con el agente de CloudWatch
- Accessing Amazon CloudWatch Logs for AWS Lambda
- Using CloudWatch Logs with container instances
- Publique métricas personalizadas
- AWS Answers: Centralized Logging
- Servicios de AWS que publican métricas de CloudWatch
- Monitoring Amazon EKS on AWS Fargate

Vídeos relacionados:

- AWS re:Invent 2023 – [LAUNCH] Application monitoring for modern workloads
- AWS re:Invent 2023 – Implementing application observability
- AWS re:Invent 2023 – Building an effective observability strategy
- AWS re:Invent 2023 – Seamless observability with AWS Distro for OpenTelemetry
- Application Performance Management on AWS

Ejemplos relacionados:

- AWS for Linux Workloads Immersion Day- Amazon CloudWatch
- Monitoring Amazon ECS clusters and containers

- Monitorizar con paneles de Amazon CloudWatch
- Taller de Amazon EKS

PERF02-BP04 Configurar y dimensionar correctamente los recursos de computación

Configure y dimensione correctamente los recursos de computación para que se ajusten a los requisitos de rendimiento de su carga de trabajo y evitar la infrautilización o el uso excesivo de recursos.

Patrones comunes de uso no recomendados:

- Ignora los requisitos de rendimiento de la carga de trabajo, lo que genera una falta o un exceso de aprovisionamiento de recursos computacionales.
- Solo elige la instancia más grande o más pequeña disponible para todas las cargas de trabajo.
- Solo usa una familia de instancias para facilitar la administración.
- No tiene en cuenta las recomendaciones de AWS Cost Explorer o Compute Optimizer para ajustar el tamaño.
- No somete a nuevas evaluaciones a la carga de trabajo para determinar la idoneidad de nuevos tipos de instancias.
- Solo certifica una pequeña cantidad de configuraciones de instancias para su organización.

Beneficios de establecer esta práctica recomendada: el dimensionamiento correcto de los recursos computacionales garantiza un funcionamiento óptimo en la nube al evitar que se produzca un exceso o falta de aprovisionamiento de recursos. El dimensionamiento adecuado de los recursos computacionales generalmente se traduce en un mayor rendimiento y una mejor experiencia del cliente, al tiempo que se reducen los costes.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Medio

Guía para la implementación

Un dimensionamiento correcto permite a las organizaciones gestionar la infraestructura en la nube de manera eficiente y rentable, al tiempo que abordan sus necesidades empresariales. Un aprovisionamiento excesivo de recursos en la nube puede generar costes adicionales, mientras que

un aprovisionamiento insuficiente puede provocar un rendimiento deficiente y una experiencia de cliente negativa. AWS proporciona herramientas como AWS Compute Optimizer y AWS Trusted Advisor que utilizan datos históricos para ofrecer recomendaciones que permiten dimensionar correctamente los recursos informáticos.

Pasos para la implementación

- Elija el tipo de instancia que mejor se adapte a sus necesidades:
  - How do I choose the appropriate Amazon EC2 instance type for my workload? (¿Cómo elijo el tipo de instancia de EC2 apropiado para mi carga de trabajo?)
  - Selección de tipo de instancia basada en atributos para la flota de Amazon EC2
  - Crear un grupo de Auto Scaling con la selección de un tipo de instancia basada en atributos
  - Optimizar los costes computacionales de Kubernetes con la consolidación de Karpenter
- Analice las distintas características de rendimiento de su carga de trabajo y la relación que tienen con el uso de memoria, redes y CPU. Use estos datos para elegir recursos que encajen bien con el perfil de la carga de trabajo y los objetivos de rendimiento.
- Controle el uso de los recursos con las herramientas de supervisión de AWS, comoAmazon CloudWatch.
- Seleccione la configuración correcta para cada recurso informático.
  - Para las cargas de trabajo efímeras, evalúe las métricas de Amazon CloudWatch de instancias como `CPUUtilization` para identificar si la instancia está infrautilizada o sobreutilizada.
  - En las cargas de trabajo estables, consulte regularmente las herramientas de dimensionamiento de AWS, como AWS Compute OptimizerAWS Trusted Advisor, para identificar oportunidades de optimizar y dimensionar las instancias de forma correcta.
- Pruebe los cambios de configuración en un entorno que no sea de producción antes de implementarlos en un entorno activo.
- Revalúe continuamente las nuevas ofertas de computación y compárelas con las necesidades de la carga de trabajo.

Recursos

Documentos relacionados:

- Cloud Compute with AWS

- Tipos de instancias de Amazon EC2
- Amazon ECS Containers: Amazon ECS Container Instances
- Amazon EKS Containers: Amazon EKS Worker Nodes
- Funciones: configuración de funciones de Lambda
- Control de los estados del procesador de la instancia Amazon EC2

Vídeos relacionados:

- Amazon EC2 foundations
- AWS re:Invent 2023 – AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 – New Amazon EC2 generative AI capabilities in AWS Management Console
- AWS re:Invent 2023 – What's new with Amazon EC2
- AWS re:Invent 2023 – Smart savings: Amazon EC2 cost-optimization strategies
- AWS re:Invent 2021 – Powering next-gen Amazon EC2: Deep dive on the Nitro System
- AWS re:Invent 2019 – Amazon EC2 foundations

Ejemplos relacionados:

- Código de demostración de AWS Compute Optimizer
- Taller de Amazon EKS
- Recomendaciones de tamaño adecuado

PERF02-BP05 Escalar los recursos computacionales de forma dinámica

Utilice la elasticidad de la nube para aumentar o reducir sus recursos computacionales de forma dinámica de forma que se ajusten a sus necesidades, lo que evitará un aprovisionamiento de capacidad excesivo o insuficiente para su carga de trabajo.

Patrones comunes de uso no recomendados:

- Reacciona a las alarmas aumentando manualmente la capacidad.
- Utiliza las mismas directrices de dimensionamiento (por lo general, una infraestructura estática) que en el entorno local.

- Deja la capacidad aumentada después de un evento de ajuste de escala en lugar de volver a desescalar verticalmente.

Beneficios de establecer esta práctica recomendada: configurar y probar la elasticidad de los recursos informáticos puede ayudarlo a ahorrar dinero, mantener los puntos de referencia de rendimiento y mejorar la fiabilidad a medida que cambia el tráfico.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Alto

Guía para la implementación

AWS le ofrece la flexibilidad necesaria para aumentar o reducir los recursos de forma dinámica a través de una gran variedad de mecanismos de escalamiento que se ajustan a los cambios de demanda. Junto con las métricas relacionadas con la computación, el escalamiento dinámico permite que las cargas de trabajo respondan automáticamente a los cambios y utilicen el conjunto óptimo de recursos informáticos para lograr su objetivo.

Puede usar distintos enfoques para hacer que el suministro de recursos coincida con la demanda.

- Enfoque de seguimiento de objetivos: supervise la métrica de escalamiento y aumente o reduzca de forma automática la capacidad en función de sus necesidades.
- Escalamiento predictivo: escale de antemano según las tendencias diarias y semanales previstas.
- Enfoque basado en programación: establezca su propia programación de escalamiento según los cambios de carga predecibles.
- Escalamiento de servicios: elija servicios (como los servicios sin servidor) diseñados para escalar automáticamente.

Debe asegurarse de que los despliegues de la carga de trabajo puedan manejar eventos de escalamiento y desescalamiento verticales.

Pasos para la implementación

- Las instancias de computación, los contenedores y las funciones proporcionan mecanismos que favorecen la elasticidad, ya sea en combinación con funciones de escalamiento automático o como características del servicio. Estos son algunos ejemplos de mecanismos de escalamiento automático:

Mecanismo de escalamiento automático	Dónde se usa
Amazon EC2 Auto Scaling	Para asegurarse de que tiene el número correcto de instancias Amazon EC2 disponibles para gestionar la carga de usuarios de su aplicación.
Application Auto Scaling	Para escalar automáticamente los recursos de servicios de AWS individuales más allá de Amazon EC2, como funciones AWS Lambda o servicios Amazon Elastic Container Service (Amazon ECS) .
Kubernetes Cluster Autoscaler/Karpenter	Para escalar automáticamente clústeres de Kubernetes.

- Normalmente, se habla del escalamiento en relación con los servicios de computación, como las instancias de Amazon EC2 o las funciones de AWS Lambda. No olvide que también debe tener en cuenta la configuración de otros servicios no computacionales como AWS Glue para satisfacer la demanda.
- Asegúrese de que las métricas de escalamiento se ajustan a las características de la carga de trabajo que se está desplegando. Si está desplegando una aplicación de transcodificación de vídeo, se espera una utilización del 100 % de la CPU y no debería ser su métrica principal. En su lugar, utilice la profundidad de la cola de trabajos de transcodificación. Puede usar una métrica personalizada para su política de escalamiento, si es necesario. Para elegir las métricas adecuadas, tenga en cuenta las siguientes directrices para Amazon EC2:
  - La métrica debe ser una métrica de utilización válida y describir el grado de ocupación de una instancia.
  - El valor de la métrica debe aumentar o disminuir proporcionalmente al número de instancias del grupo de Auto Scaling.
- Asegúrese de utilizar el escalado dinámico en vez del escalado manual para su grupo de Auto Scaling. También le recomendamos que utilice políticas de escalado de seguimiento de destino en su escalado dinámico.

- Compruebe que los despliegues de la carga de trabajo puedan gestionar ambos eventos de escalamiento (escalamiento y desescalamiento verticales). Como ejemplo, puede usar el historia de actividades para verificar una actividad de escalamiento para un grupo de Auto Scaling.
- Evalúe los patrones predecibles de su carga de trabajo y escale de forma proactiva al anticiparse a los cambios previstos y planeados en la demanda. Con el escalamiento predictivo, puede eliminar la necesidad de aprovisionar capacidad en exceso. Para obtener más detalles, consulte Predictive scaling with Amazon EC2 Auto Scaling (Escalamiento predictivo con Amazon EC2 Auto Scaling).

Recursos

Documentos relacionados:

- Cloud Compute with AWS
- Tipos de instancias de Amazon EC2
- Amazon ECS Containers: Amazon ECS Container Instances
- Amazon EKS Containers: Amazon EKS Worker Nodes
- Funciones: configuración de funciones de Lambda
- Control de los estados del procesador de la instancia Amazon EC2
- Deep Dive on Amazon ECS Cluster Auto Scaling
- Introducing Karpenter – An Open-Source High-Performance Kubernetes Cluster Autoscaler

Vídeos relacionados:

- AWS re:Invent 2023 – AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 – New Amazon EC2 generative AI capabilities in AWS Management Console
- AWS re:Invent 2023 – What's new with Amazon EC2
- AWS re:Invent 2023 – Smart savings: Amazon EC2 cost-optimization strategies
- AWS re:Invent 2021 – Powering next-gen Amazon EC2: Deep dive on the Nitro System
- AWS re:Invent 2019 – Amazon EC2 foundations

Ejemplos relacionados:

- Amazon EC2 Auto Scaling Group Examples
- Amazon EKS Workshop

- Scale your Amazon EKS workloads by running on IPv6

PERF02-BP06 Utilización de aceleradores computacionales optimizados basados en hardware

Use aceleradores de hardware para realizar ciertas funciones de manera más eficiente que con las alternativas basadas en CPU.

Antipatrones usuales:

- En su carga de trabajo, no ha comparado una instancia de uso general con una instancia personalizada que pueda ofrecer mayor rendimiento y costes más reducidos.
- Utiliza aceleradores computacionales basados en hardware para tareas en las que podría ser más eficiente utilizar alternativas basadas en CPU.
- No supervisa el uso de GPU.

Ventajas de aplicar esta práctica recomendada: al utilizar aceleradores basados en hardware, como unidades de procesamiento de gráficos (GPU) y matrices de puertas programables en campo (FPGA), puede ejecutar determinadas funciones de procesamiento de manera más eficiente.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: medio

Guía para la implementación

Las instancias de computación acelerada proporcionan acceso a aceleradores de computación basados en hardware, como las GPU y las FPGA. Estos aceleradores de hardware realizan ciertas funciones, como el procesamiento gráfico o la concordancia de patrones de datos, de forma más eficiente que las alternativas basadas en CPU. Muchas cargas de trabajo aceleradas, como el renderizado, la transcodificación y el machine learning, son muy variables en cuanto al uso de recursos. Ejecute este hardware únicamente durante el tiempo necesario y retírelo de forma automatizada cuando no sea necesario para mejorar la eficiencia general del rendimiento.

Pasos para la implementación

- Identifique qué instancias de computación acelerada pueden satisfacer sus necesidades.
- En las cargas de trabajo de machine learning, utilice un hardware personalizado específico para la carga de trabajo, como AWS Trainium, AWS Inferentia y Amazon EC2 DL1. Las instancias de

AWS Inferentia, como las instancias Inf2, ofrecen hasta un 50 % más de rendimiento por vatio que las instancias de Amazon EC2 equivalentes.

- Recopile las métricas de uso de sus instancias de computación acelerada. Por ejemplo, puede utilizar el agente de CloudWatch para recopilar las métricas, como `utilization_gpu` y `utilization_memory`, de sus GPU, tal y como se muestra en Collect NVIDIA GPU metrics with Amazon CloudWatch.
- Optimice el código, el funcionamiento de la red y la configuración de los aceleradores de hardware para asegurarse de que se aprovecha al máximo el hardware subyacente.
  - Optimizar la configuración de GPU
  - GPU Monitoring and Optimization in the Deep Learning AMI (Supervisión y optimización de la GPU en la AMI de aprendizaje profundo)
  - Optimizing I/O for GPU performance tuning of deep learning training in Amazon SageMaker (Optimización de la E/S para el ajuste del rendimiento de la GPU en el entrenamiento del aprendizaje profundo en Amazon SageMaker)
- Utilice las bibliotecas de alto rendimiento y los controladores de GPU más recientes.
- Use la automatización para liberar instancias de GPU cuando no se estén usando.

Recursos

Documentos relacionados:

- Uso de GPU en Amazon Elastic Container Service
- GPU instances
- Instances with AWS Trainium
- Instances with AWS Inferentia
- Let's Architect! Architecting with custom chips and accelerators

- Computación acelerada
- Instancias VT1 de Amazon EC2
- How do I choose the appropriate Amazon EC2 instance type for my workload?
- Choose the best AI accelerator and model compilation for computer vision inference with Amazon SageMaker

Vídeos relacionados:

- AWS re:Invent 2021 - How to select Amazon Elastic Compute Cloud GPU instances for deep learning
- AWS re:Invent 2022 - [NEW LAUNCH!] Introducing AWS Inferentia2-based Amazon EC2 Inf2 instances
- AWS re:Invent 2022 - Accelerate deep learning and innovate faster with AWS Trainium
- AWS re:Invent 2022 - Deep learning on AWS with NVIDIA: From training to deployment

Ejemplos relacionados:

- Amazon SageMaker and NVIDIA GPU Cloud (NGC)
- Use SageMaker with Trainium and Inferentia for optimized deep learning training and inferencing workloads
- Optimizing NLP models with Amazon Elastic Compute Cloud Inf1 instances in Amazon SageMaker

Administración de datos

La solución de administración de datos óptima para un sistema concreto varía según el tipo de datos (bloque, archivo u objeto), patrones de acceso (aleatorio o secuencial), rendimiento requerido, frecuencia de acceso (en línea, fuera de línea, archivo), frecuencia de actualización (WORM, dinámica) y restricciones de disponibilidad y durabilidad. Las cargas de trabajo Well-Architected utilizan almacenes de datos diseñados de manera específica que admiten diferentes características para mejorar el rendimiento.

Esta área de enfoque comparte la guía y las prácticas recomendadas para optimizar el almacenamiento de datos, los patrones de movimiento y acceso y la eficiencia del rendimiento de los almacenes de datos.

Prácticas recomendadas

- PERF03-BP01 Utilización de un almacén de datos personalizado que se adapte mejor a los requisitos de acceso y almacenamiento de datos
- PERF03-BP02 Evaluar las opciones de configuración disponibles
- PERF03-BP03 Recopilar y registrar las métricas de rendimiento del almacén de datos
- PERF03-BP04 Implementar estrategias para mejorar el rendimiento de las consultas en el almacén de datos
- PERF03-BP05 Implementar patrones de acceso a datos que utilicen el almacenamiento en caché

PERF03-BP01 Utilización de un almacén de datos personalizado que se adapte mejor a los requisitos de acceso y almacenamiento de datos

Debe saber cuáles son las características de los datos (por ejemplo, si se pueden compartir, su tamaño, los patrones de acceso, la latencia, el rendimiento y su persistencia) para seleccionar los almacenes de datos personalizados acordes a su carga de trabajo (almacenamiento o base de datos).

Antipatrones usuales:

- Utiliza exclusivamente un almacén de datos porque la experiencia y los conocimientos internos se limitan a un tipo concreto de solución de base de datos.

- Presupone que todas las cargas de trabajo tienen unos requisitos similares en relación con el almacenamiento de datos y el acceso a la información.
- No ha implementado un catálogo de datos para inventariar sus activos de datos.

Beneficios de establecer esta práctica recomendada: conocer las características y los requisitos de los datos le permite determinar cuál es la tecnología de almacenamiento más eficiente y funcional adecuada para las necesidades de su carga de trabajo.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: alto

Guía para la implementación

Al seleccionar e implementar el almacenamiento de datos, asegúrese de que las características de consulta, escalamiento y almacenamiento se ajusten a los requisitos de datos de la carga de trabajo. AWS ofrece un gran número de tecnologías de almacenamiento y bases de datos, como el almacenamiento en bloques, el almacenamiento de objetos, el almacenamiento en streaming, los sistemas de archivos, las bases de datos relacionales, las bases de datos de clave-valor, las bases de datos de documentos, las bases de datos en memoria, las bases de datos de grafos, las bases de datos de series temporales y las bases de datos de libro mayor. Cada solución de administración de datos tiene opciones y configuraciones a su disposición que se ajustan a los casos de uso y a los modelos de datos. Si conoce las características y los requisitos de los datos, puede dejar atrás la tecnología de almacenamiento monolítica y los enfoques restrictivos de «una misma cosa vale para todo», y centrarse en gestionar correctamente los datos.

Pasos para la implementación

- Realice un inventario de los distintos tipos de datos que existen en su carga de trabajo.
- Estudie y documente las características y los requisitos de los datos, como:
  - Tipo de datos (no estructurados, semiestructurados o relacionales)
  - Volumen y crecimiento de los datos
  - Durabilidad de los datos: persistentes, efímeros o transitorios
  - Requisitos de ACID (atomicidad, consistencia, aislamiento, durabilidad)
  - Patrones de acceso a los datos (lectura o escritura intensivas)
  - Latencia
  - Rendimiento
  - IOPS (operaciones de entrada/salida por segundo)

  - Período de retención de los datos
- Obtenga información sobre los diferentes almacenes de datos (servicios de base de datos y almacenamiento) disponibles para su carga de trabajo en AWS que se ajustan a las características de los datos (tal y como se describe en PERF01-BP01 Descubrir y comprender los servicios y las características disponibles en la nube). Estos son algunos ejemplos de tecnologías de almacenamiento de AWS y sus principales características:

Tipo	Servicios de AWS	Características clave
Object storage	Amazon S3	Unlimited scalability, high availability, and multiple options for accessibility. Transferring and accessing objects in and out of Amazon S3 can use a service, such as Aceleración de transferencia or Puntos de acceso, to support your location, security needs, and access patterns.
Archiving storage	Amazon S3 Glacier	Built for data archiving.
Streaming storage	Amazon Kinesis Amazon Managed Streaming for Apache Kafka (Amazon MSK)	Efficient ingestion and storage of streaming data.
Shared file system	Amazon Elastic File System (Amazon EFS)	Sistema de archivos montable al que pueden acceder varios tipos de soluciones de computación.
Shared file system	Amazon FSx	Built on the latest AWS compute solutions to support four commonly used file systems: NetApp ONTAP, OpenZFS, Windows File

Tipo	Servicios de AWS	Características clave
		Server, and Lustre. Amazon FSx su latencia, rendimiento y E/S por segundo vary per file system and should be considered when selecting the right file system for your workload needs.
Block storage	Amazon Elastic Block Store (Amazon EBS)	Scalable, high-performance block-storage service designed for Amazon Elastic Compute Cloud (Amazon EC2). Amazon EBS includes SSD-backed storage for transactional, IOPS-intensive workloads and HDD-backed storage for throughput-intensive workloads.
Relational database	Amazon Aurora, Amazon RDS, Amazon Redshift.	Designed to support ACID (atomicity, consistency, isolation, durability) transactions, and maintain referential integrity and strong data consistency. Many traditional applications, enterprise resource planning (ERP), customer relationship management (CRM), and ecommerce use relational databases to store their data.

Tipo	Servicios de AWS	Características clave
Key-value database	Amazon DynamoDB	Optimized for common access patterns, typically to store and retrieve large volumes of data. High-traffic web apps, ecommerce systems, and gaming applications are typical use-cases for key-value databases.
Document database	Amazon DocumentDB	Designed to store semi-structured data as JSON-like documents. These databases help developers build and update applications such as content management, catalogs, and user profiles quickly.
In-memory database	Amazon ElastiCache , Amazon MemoryDB para Redis	Used for applications that require real-time access to data, lowest latency and highest throughput. You may use in-memory databases for application caching, session management, gaming leaderboards, low latency ML feature store, microservices messaging system, and a high-throughput streaming mechanism

Tipo	Servicios de AWS	Características clave
Graph database	Amazon Neptune	Used for applications that must navigate and query millions of relationships between highly connected graph datasets with milliseco nd latency at large scale. Many companies use graph databases for fraud detection , social networking, and recommendation engines.
Time Series database	Amazon Timestream	Used to efficiently collect, synthesize, and derive insights from data that changes over time. IoT applications, DevOps, and industrial telemetry can utilize time-series databases.
Wide column	Amazon Keyspaces (para Apache Cassandra)	Uses tables, rows, and columns, but unlike a relational database, the names and format of the columns can vary from row to row in the same table. You typically see a wide column store in high scale industrial apps for equipment maintenance, fleet management, and route optimization.

Tipo	Servicios de AWS	Características clave
Ledger	Amazon Quantum Ledger Database (Amazon QLDB)	Provides a centralized and trusted authority to maintain a scalable, immutable, and cryptographically verifiabl e record of transactions for every application. We see ledger databases used for systems of record, supply chain, registrations, and even banking transactions.

- Si está creando una plataforma de datos, aproveche la arquitectura de datos moderna de AWS para integrar un lago de datos, un almacenamiento de datos y almacenes de datos personalizados.
- Las principales preguntas que debe hacerse al elegir un almacén de datos para su carga de trabajo son las siguientes:

Question	Things to consider
How is the data structured?	• Si los datos no están estructurados, considere la posibilidad de usar un almacén de objetos, como Amazon S3, o una base de datos NoSQL, como Amazon DocumentDB. • Para los datos de clave-valor, podría usar DynamoDB, Amazon ElastiCache for Redis o Amazon MemoryDB for Redis.
What level of referential integrity is required?	• En el caso de las restricciones de clave externa, las bases de datos relacionales, como Amazon RDS y Aurora, pueden proporcionar este nivel de integridad. • Normalmente, en un modelo de datos NoSQL, los datos se desnormalizarían

Question	Things to consider
	en un documento o una colección de documentos en lugar de combinarse en diferentes documentos o tablas, lo que permitiría recuperarlos en una única solicitud.
Is ACID (atomicity, consistency, isolation, durability) compliance required?	• Si las propiedades ACID asociadas a las bases de datos relacionales son necesaria s, considere la posibilidad de usar una base de datos relacional, como Amazon RDS y Aurora. • Si se requiere una consistencia sólida para la base de datos NoSQL, puede usar lecturas con coherencia fuerte a través de DynamoDB.
How will the storage requirements change over time? How does this impact scalability?	• Las bases de datos sin servidor, como DynamoDB y Amazon Quantum Ledger Database (Amazon QLDB), se escalarán dinámicamente. • Las bases de datos relacionales tienen límites máximos de almacenamiento aprovisionado y, a menudo, cuando alcanzan estos límites, es necesario hacer particiones horizontales a través de diversos mecanismos, como el particion amiento.

Question	Things to consider
What is the proportion of read queries in relation to write queries? Would caching be likely to improve performance?	• Las cargas de trabajo que requieren muchas lecturas pueden beneficiarse de una capa de almacenamiento en caché, como ElastiCache o DAX, si la base de datos es DynamoDB. • Las lecturas también pueden descargarse en réplicas de lectura con bases de datos relacionales, como Amazon RDS.
Does storage and modification (OLTP - Online Transaction Processing) or retrieval and reporting (OLAP - Online Analytical Processin g) have a higher priority?	• Para el procesamiento transaccional de lecturas de alto rendimiento sin realizar cambios, considere la posibilidad de usar una base de datos NoSQL, como. DynamoDB. • En el caso de los patrones de lectura complejos y de alto rendimiento (como una combinación) que tienen coherencia, use Amazon RDS. • Para las consultas analíticas, podría utilizar una base de datos en columnas, como Amazon Redshift, o exportar los datos a Amazon S3 y realizar análisis con Athena o Amazon QuickSight.

Question	Things to consider
What level of durability does the data require?	• Aurora replica los datos automáticamente en tres zonas de disponibilidad de una región, lo que significa que los datos tendrán una gran durabilidad y menos posibilidades de sufrir pérdidas. • DynamoDB se replica automáticamente en varias zonas de disponibilidad, lo que proporciona una elevada disponibilidad y durabilidad de los datos. • Amazon S3 proporciona un nivel de durabilidad de once nueves. Muchos servicios de bases de datos, como Amazon RDS y DynamoDB, permiten exportar datos a Amazon S3 para retenerlos y archivarlos durante largos períodos de tiempo.
Is there a desire to move away from commercial database engines or licensing costs?	• Considere la posibilidad de utilizar motores de código abierto como PostgreSQL y MySQL en Amazon RDS o Aurora. • Utilice AWS Database Migration Service y AWS Schema Conversion Tool para migrar motores de bases de datos comerciales a motores de código abierto.
What is the operational expectation for the database? Is moving to managed services a primary concern?	• Si usa Amazon RDS en lugar de Amazon EC2 y utiliza DynamoDB o Amazon DocumentDB en lugar de alojar una base de datos NoSQL en sus propios sistemas, puede reducir los costes operativos.

Question	Things to consider
How is the database currently accessed? Is it only application access, or are there business intelligence (BI) users and other connected off-the-shelf applications?	• Si tiene dependencias en herramientas externas, es posible que necesite mantener la compatibilidad con las bases de datos que se utilizan allí. Amazon RDS es totalmente compatible con las diferentes versiones de motores que admite, como Microsoft SQL Server, Oracle, MySQL y PostgreSQL.

- Realice experimentos y pruebas comparativas en un entorno que no sea de producción para identificar qué almacén de datos se ajusta a los requisitos de su carga de trabajo.

Recursos

Documentos relacionados:

- Amazon EBS Volume Types
- Amazon EC2 Storage
- Amazon EFS: Amazon EFS Performance
- Amazon FSx for Lustre Performance
- Amazon FSx for Windows File Server Performance
- Amazon S3 Glacier: S3 Glacier Documentation
- Amazon S3: Request Rate and Performance Considerations
- Almacenamiento en la nube en AWS
- Amazon EBS I/O Characteristics
- Bases de datos en la nube de AWS
- AWS Database Caching
- DynamoDB Accelerator
- Amazon Aurora best practices
- Amazon Redshift performance
- Amazon Athena top 10 performance tips
- Amazon Redshift Spectrum best practices

- Amazon DynamoDB best practices
- Choose between Amazon EC2 and Amazon RDS
- Best Practices for Implementing Amazon ElastiCache

Vídeos relacionados:

- AWS re:Invent 2023: Improve Amazon Elastic Block Store efficiency and be more cost-efficient
- AWS re:Invent 2023: Optimizing storage price and performance with Amazon Simple Storage Service
- AWS re:Invent 2023: Building and optimizing a data lake on Amazon Simple Storage Service
- AWS re:Invent 2022: Building modern data architectures on AWS
- AWS re:Invent 2022: Building data mesh architectures on AWS
- AWS re:Invent 2023: Deep dive into Amazon Aurora and its innovations
- AWS re:Invent 2023: Advanced data modeling with Amazon DynamoDB
- AWS re:Invent 2022: Modernize apps with purpose-built databases
- Amazon DynamoDB deep dive: Advanced design patterns

Ejemplos relacionados:

- AWS Purpose Built Databases Workshop
- Databases for Developers
- AWS Modern Data Architecture Immersion Day
- Build a Data Mesh on AWS
- Amazon S3 Examples
- Optimize Data Pattern using Amazon Redshift Data Sharing
- Database Migrations
- MS SQL Server - AWS Database Migration Service (AWS DMS) Replication Demo
- Database Modernization Hands On Workshop
- Amazon Neptune Samples

PERF03-BP02 Evaluar las opciones de configuración disponibles

Estudie y evalúe las diversas características y opciones de configuración disponibles para sus almacenes de datos a fin de optimizar el espacio de almacenamiento y el rendimiento de su carga de trabajo.

Antipatrones usuales:

- Utiliza el mismo tipo de almacenamiento (por ejemplo, Amazon EBS) para todas sus cargas de trabajo.
- Utiliza IOPS aprovisionadas en todas las cargas de trabajo sin realizar pruebas en el mundo real con todos los niveles de almacenamiento.
- No conoce las opciones de configuración de la solución de administración de datos que ha elegido.
- La única opción que contempla es aumentar el tamaño de las instancias, sin valorar otras opciones de configuración disponibles.
- No realiza pruebas en las características de escalamiento de su almacén de datos.

Ventajas de aplicar esta práctica recomendada: al explorar y probar las configuraciones del almacén de datos, es posible que pueda rebajar el coste de la infraestructura, mejorar el rendimiento y reducir el esfuerzo necesario para mantener sus cargas de trabajo.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: medio

Guía para la implementación

En una carga de trabajo, puede haber uno o varios almacenes de datos que se utilicen en función de los requisitos de almacenamiento y acceso. Para optimizar los costes y la eficiencia del rendimiento, debe evaluar los patrones de acceso a los datos y determinar cuáles son las configuraciones de almacenamiento de datos adecuadas. Cuando explore las opciones de almacenamiento de datos, tenga en cuenta diversos aspectos, como las opciones de almacenamiento, la memoria, los recursos de computación, la réplica de lectura, los requisitos de coherencia, la agrupación de conexiones y las opciones de almacenamiento en caché. Pruebe estas diferentes opciones de configuración para mejorar las métricas de eficiencia del rendimiento.

Pasos para la implementación

- Estudie las configuraciones actuales (como el tipo de instancia, el tamaño de almacenamiento o la versión del motor de base de datos) de su almacén de datos.

- Consulte la documentación y las prácticas recomendadas de AWS para obtener información sobre las opciones de configuración recomendadas que pueden ayudarle a mejorar el rendimiento de su almacén de datos. Las principales opciones de almacenamiento de datos que debe tener en cuenta son las siguientes:

Configuration option	Examples
Offloading reads (like read replicas and caching)	• En el caso de las tablas de DynamoDB, puede descargar las lecturas utilizando DAX para el almacenamiento en caché. • Puede crear un clúster de Amazon ElastiCache for Redis y configurar la aplicación para que lea primero la memoria caché y, si el elemento solicitado no está presente, recurra a la base de datos. • Las bases de datos relacionales, como Amazon RDS y Aurora, y las bases de datos NoSQL aprovisionadas, como Neptune y Amazon DocumentDB, permiten añadir réplicas de lectura para descargar las partes de lectura de la carga de trabajo. • Las bases de datos sin servidor, como DynamoDB, se escalarán automátic amente. Asegúrese de que tiene suficient es unidades de capacidad de lectura (RCU) aprovisionadas para gestionar la carga de trabajo.

Configuration option	Examples
Scaling writes (like partition key sharding or introducing a queue)	• En el caso de las bases de datos relacionales, puede aumentar el tamaño de la instancia para acomodar una mayor carga de trabajo o aumentar las IOPS aprovisionadas para mejorar el rendimiento del almacenamiento subyacente. • También puede introducir una cola delante de la base de datos en lugar de escribir directamente en la base de datos. Este patrón permite desacoplar la ingesta de la base de datos y controlar el caudal para que la base de datos no se vea desbordada. • Si agrupa las solicitudes de escritura en lugar de crear muchas transacciones de corta duración, puede mejorar el rendimiento de las bases de datos relacionales con un gran volumen de operaciones de escritura. • Las bases de datos sin servidor como DynamoDB pueden escalar el rendimiento de escritura automáticamente o ajustando las unidades de capacidad de escritura (WCU) aprovisionadas en función del modo de capacidad. • Puede tener problemas con las particiones activas si alcanza los límites de rendimiento de una clave de partición determinada. Esto puede mitigarse eligiendo una clave de partición distribuida de manera más uniforme o particionando la escritura en función de la clave de partición.

Configuration option	Examples
Policies to manage the lifecycle of your datasets	• Puede usar Amazon S3 Lifecycle para administrar los objetos a lo largo de su ciclo de vida. Si los patrones de acceso no se conocen, experimentan cambios o son impredecibles, puede utilizar Amazon S3 Intelligent-Tiering, que supervisa los patrones de acceso y mueve automátic amente los objetos a los que no se ha accedido a niveles de acceso más baratos. Puede utilizar las métricas de Amazon S3 Storage Lens para identificar las oportunid ades de optimización y las lagunas en la administración del ciclo de vida. • La administración del ciclo de vida de Amazon EFS gestiona automáticamente el almacenamiento en los sistemas de archivos.
Connection management and pooling	• Amazon RDS Proxy puede utilizarse con Amazon RDS y Aurora para administrar conexiones a la base de datos. • Las bases de datos sin servidor como DynamoDB no tienen conexiones asociadas , pero tienen en cuenta la capacidad aprovisionada y las políticas de escalamie nto automático para hacer frente a los picos de carga.

- Realice experimentos y pruebas comparativas en un entorno que no sea de producción para identificar qué opción de computación se ajusta a los requisitos de la carga de trabajo.
- Una vez hecho esto, planifique la migración y valide las métricas de rendimiento.
- Use las herramientas de supervisión de AWS (como Amazon CloudWatch) y de optimización (como Amazon S3 Storage Lens) para optimizar continuamente el almacén de datos con patrones de uso reales.

Recursos

Documentos relacionados:

- Almacenamiento en la nube en AWS
- Amazon EBS Volume Types
- Amazon EC2 Storage
- Amazon EFS: Amazon EFS Performance
- Amazon FSx for Lustre Performance
- Amazon FSx for Windows File Server Performance
- Amazon S3 Glacier: S3 Glacier Documentation
- Amazon S3: Request Rate and Performance Considerations
- Amazon EBS I/O Characteristics
- Bases de datos en la nube de AWS
- AWS Database Caching
- DynamoDB Accelerator
- Amazon Aurora best practices
- Amazon Redshift performance
- Amazon Athena top 10 performance tips
- Amazon Redshift Spectrum best practices
- Amazon DynamoDB best practices

Vídeos relacionados:

- AWS re:Invent 2023: Improve Amazon Elastic Block Store efficiency and be more cost-efficient
- AWS re:Invent 2023: Optimize storage price and performance with Amazon Simple Storage Service
- AWS re:Invent 2023: Building and optimizing a data lake on Amazon Simple Storage Service
- AWS re:Invent 2023: What's new with AWS file storage
- AWS re:Invent 2023: Dive deep into Amazon DynamoDB

Ejemplos relacionados:

- AWS Purpose Built Databases Workshop
- Databases for Developers
- AWS Modern Data Architecture Immersion Day
- Amazon EBS Autoscale
- Amazon S3 Examples
- Amazon DynamoDB Examples
- AWS Database migration samples
- Database Modernization Workshop
- Working with parameters on your Amazon RDS for Postgress DB

PERF03-BP03 Recopilar y registrar las métricas de rendimiento del almacén de datos

Supervise y registre las métricas de rendimiento relevantes del almacén de datos para saber cómo funcionan las soluciones de administración de datos. Estas métricas pueden ayudarle a optimizar el almacén de datos, a garantizar que se cumplen los requisitos de la carga de trabajo y a proporcionar una visión general clara del rendimiento de la carga de trabajo.

Patrones comunes de uso no recomendados:

- Solo se utiliza la búsqueda manual de métricas en los archivos de registro.
- Solo publica métricas en las herramientas internas que su equipo utiliza y no tiene una imagen completa de su carga de trabajo.
- Solo se utilizan las métricas predeterminadas registradas por el software de supervisión seleccionado.
- Solo se revisan las métricas cuando hay un problema.
- Solo se supervisan las métricas en el nivel del sistema y no se captura las métricas de acceso o de uso de datos.

Beneficios de establecer esta práctica recomendada: instaurar una base de referencia de rendimiento le ayuda a comprender el comportamiento habitual y los requisitos de las cargas de trabajo. Los patrones anómalos pueden identificarse y depurarse más rápidamente, lo que mejora el rendimiento y la fiabilidad del almacén de datos.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Alto

Guía para la implementación

Para supervisar el rendimiento de sus almacenes de trabajo, debe registrar diversas métricas de rendimiento a lo largo del tiempo. De este modo, podrá detectar anomalías y medir el rendimiento con respecto a las métricas de la empresa para asegurarse de que se están satisfaciendo las necesidades de su carga de trabajo.

Las métricas deben incluir tanto el sistema subyacente que da servicio al almacén de datos como las métricas de la base de datos. Las métricas del sistema subyacente podrían ser la utilización de la CPU, la memoria, el almacenamiento en disco disponible, las operaciones de E/S del disco, la proporción de aciertos de la caché y las métricas de entrada y salida de la red, mientras que las métricas del almacén de datos podrían ser las transacciones por segundo, las consultas principales, las tasas medias de consultas, los tiempos de respuesta, el uso de índices, los bloqueos de tablas, los tiempos de espera de las consultas y el número de conexiones abiertas. Estos datos son cruciales para entender cómo funciona la carga de trabajo y cómo se utiliza la solución de administración de datos. Utilice estas métricas como parte de un enfoque basado en datos para ajustar y optimizar los recursos de la carga de trabajo.

Use herramientas, bibliotecas y sistemas que registren las medidas de rendimiento relacionadas con el rendimiento de la base de datos.

Pasos para la implementación

1. Identifique las métricas de rendimiento clave del almacén de datos que desee supervisar.
    a. Métricas y dimensiones de Amazon S3
    b. Supervisión de las métricas de una instancia de Amazon RDS
    c. Supervisión de la carga de bases de datos con Información sobre rendimiento en Amazon RDS
    d. Descripción general de la supervisión mejorada
    e. Métricas y dimensiones de DynamoDB
    f. Supervisión de DynamoDB Accelerator
    g. Supervisión de Amazon MemoryDB for Redis con Amazon CloudWatch
    h. ¿Qué métricas debo supervisar?
    i. Supervisión del rendimiento del clúster de Amazon Redshift
    j. Métricas y dimensiones de Timestream

   k. Métricas de Amazon CloudWatch para Amazon Aurora
   l. Registro y supervisión en Amazon Keyspaces (for Apache Cassandra)
   m. Supervisión de recursos de Amazon Neptune
2. Use una solución de registro y supervisión aprobada para recopilar estas métricas. Amazon CloudWatch puede recopilar métricas en todos los recursos de su arquitectura. También puede recopilar y publicar métricas del cliente para negocios de superficie o métricas derivadas. Utilice CloudWatch o soluciones de terceros para establecer alarmas que avisen cuando se superen los umbrales.
3. Compruebe si la supervisión del almacén de datos puede beneficiarse de una solución de machine learning que detecte anomalías de rendimiento.
   a. Amazon DevOps Guru para Amazon RDS brinda visibilidad sobre los problemas de rendimiento y recomienda acciones correctivas.
4. Configure la retención de datos de la solución de supervisión y registro para que se ajuste a sus objetivos operativos y de seguridad.
   a. Retención de datos predeterminada para métricas de CloudWatch
   b. Retención de datos predeterminada para CloudWatch Logs

Recursos

Documentos relacionados:

- AWS Database Caching
- Amazon Athena top 10 performance tips
- Prácticas recomendadas para Amazon Aurora
- DynamoDB Accelerator
- Amazon DynamoDB best practices
- Amazon Redshift Spectrum best practices
- Desempeño de Amazon Redshift
- Cloud Databases with AWS
- Información sobre rendimiento de Amazon RDS

Vídeos relacionados:

- AWS re:Invent 2022 - Performance monitoring with Amazon RDS and Aurora, featuring Autodesk
- Database Performance Monitoring and Tuning with Amazon DevOps Guru for Amazon RDS
- AWS re:Invent 2023 - What's new with AWS file storage
- AWS re:Invent 2023 - Dive deep into Amazon DynamoDB
- AWS re:Invent 2023 - Building and optimizing a data lake on Amazon S3
- AWS re:Invent 2023 - What's new with AWS file storage
- AWS re:Invent 2023 - Dive deep into Amazon DynamoDB
- Best Practices for Monitoring Redis Workloads on Amazon ElastiCache

Ejemplos relacionados:

- AWS Dataset Ingestion Metrics Collection Framework
- Amazon RDS Monitoring Workshop
- AWS Purpose Built Databases Workshop

PERF03-BP04 Implementar estrategias para mejorar el rendimiento de las consultas en el almacén de datos

Implemente estrategias que permitan optimizar los datos y mejorar las consultas para aumentar la escalabilidad y conseguir un rendimiento eficiente para su carga de trabajo.

Patrones comunes de uso no recomendados:

- No divide en particiones los datos en su almacén de datos.
- Almacena los datos en un solo formato en su almacén de datos.
- No utiliza índices en su almacén de datos.

Beneficios de establecer esta práctica recomendada: al optimizar el rendimiento de los datos y las consultas, se consigue una mayor eficiencia, una reducción de los costes y una mejor experiencia de usuario.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Medio

Guía para la implementación

La optimización de los datos y el ajuste de las consultas son aspectos fundamentales en la eficiencia del rendimiento de un almacén de datos, ya que afectan al rendimiento y a la capacidad de respuesta de toda la carga de trabajo en la nube. Las consultas que no están optimizadas pueden aumentar el uso de recursos y generar cuellos de botella, lo que reduce la eficiencia general de los almacenes de datos.

La optimización de datos incluye diversas técnicas que garantizan la eficiencia del almacenamiento de datos y su acceso. Esto también ayuda a mejorar el rendimiento de las consultas en un almacén de datos. Algunas de las estrategias clave son la partición, la compresión y la desnormalización de los datos, lo que ayuda a optimizarlos tanto a la hora de almacenarlos como de acceder a ellos.

Pasos para la implementación

- Estudie y analice las consultas de datos críticos que se realizan en el almacén de datos.
- Identifique las consultas de ejecución lenta del almacén de datos y utilice planes de consulta para conocer su estado actual.
  - Análisis del plan de consulta en Amazon Redshift
  - Uso de EXPLAIN y EXPLAIN ANALYZE en Athena
- Implemente estrategias para mejorar el rendimiento de las consultas. Algunas de las estrategias clave son:
  - Uso de un formato de archivo en columnas (como Parquet u ORC).
  - Comprimir los datos en el almacén de datos para reducir el espacio de almacenamiento y la operación de E/S.
  - Crear particiones de datos para dividir la información en partes más pequeñas y reducir el tiempo de análisis de los datos.
    - Partición de datos en Athena
    - Particiones y distribución de datos
  - Indexar los datos de las columnas más frecuentes de la consulta.
  - Utilizar vistas materializadas para consultas frecuentes.
    - Comprensión de las vistas materializadas
    - Creación de vistas materializadas en Amazon Redshift
  - Elegir la operación de unión correcta para la consulta. Cuando una dos tablas, especifique la tabla mayor en el lado izquierdo de la unión y la tabla menor en el lado derecho de la unión.

  - Usar una solución de almacenamiento en caché distribuida para mejorar la latencia y reducir la cantidad de operaciones de E/S de la base de datos.
  - Realizar un mantenimiento regular, como la ejecución de estadísticas.
- Experimente y pruebe estrategias en un entorno que no sea de producción.

Recursos

Documentos relacionados:

- Prácticas recomendadas para Amazon Aurora
- Desempeño de Amazon Redshift
- Amazon Athena top 10 performance tips
- AWS Database Caching
- Best Practices for Implementing Amazon ElastiCache
- Partición de datos en Athena

Vídeos relacionados:

- AWS re:Invent 2023 - AWS storage cost-optimization best practices
- AWS re:Invent 2022 - Performance monitoring with Amazon RDS and Aurora, featuring Autodesk
- Optimize Amazon Athena Queries with New Query Analysis Tools

Ejemplos relacionados:

- Amazon S3 Select - Querying data without servers or databases
- AWS Purpose Built Databases Workshop

PERF03-BP05 Implementar patrones de acceso a datos que utilicen el almacenamiento en caché

Implemente patrones de acceso que puedan beneficiarse del almacenamiento en caché de los datos para lograr una recuperación rápida de los datos a los que se accede con frecuencia.

Patrones comunes de uso no recomendados:

- Almacena en caché datos que cambian con frecuencia.
- Confía en los datos en caché como si estuvieran almacenados de forma duradera y siempre disponibles.
- No tiene en cuenta la coherencia de los datos en caché.
- No supervisa la eficiencia de su implementación de almacenamiento en caché.

Beneficios de establecer esta práctica recomendada: El almacenamiento de datos en una memoria caché puede mejorar la latencia de lectura, el rendimiento de lectura, la experiencia del usuario y la eficiencia general, además de reducir los costes.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: medio

Guía para la implementación

Una memoria caché es un componente de software o hardware destinado a almacenar datos para que las futuras solicitudes de los mismos se puedan atender de manera más rápida o eficiente. Los datos almacenados en una memoria caché pueden reconstruirse si se pierden repitiendo un cálculo anterior o recuperándolos de otro almacén de datos.

El almacenamiento en caché de los datos puede ser una de las estrategias más eficaces para mejorar el rendimiento general de la aplicación y reducir la carga sobre los orígenes de datos principales subyacentes. Los datos pueden almacenarse en caché en varios niveles de la aplicación, como dentro de la aplicación que realiza llamadas remotas, lo que se conoce como almacenamiento en caché del lado del cliente, o mediante un servicio secundario rápido para almacenar los datos, lo que se conoce como almacenamiento remoto en caché.

Almacenamiento en caché del lado del cliente

Con el almacenamiento en caché del lado del cliente, cada cliente (una aplicación o servicio que consulta el almacén de datos del backend) puede almacenar los resultados de sus consultas únicas de forma local durante un período de tiempo determinado. Esto puede reducir el número de solicitudes a través de la red a un almacén de datos al comprobar primero la memoria caché del cliente local. Si no hay resultados presentes, la aplicación puede consultar el almacén de datos y almacenar esos resultados localmente. Este patrón permite a cada cliente almacenar los datos en la ubicación más cercana posible (el propio cliente), lo que tiene como resultado la latencia más baja posible. Los clientes también pueden seguir atendiendo algunas consultas cuando el almacén de datos del backend no esté disponible, lo que aumenta la disponibilidad de todo el sistema.

Una desventaja de este enfoque es que, cuando hay varios clientes implicados, pueden almacenar los mismos datos en caché localmente, lo que se traduce en un uso duplicado del almacenamiento y en una incoherencia de los datos entre esos clientes. Un cliente puede almacenar en caché los resultados de una consulta y, un minuto después, otro cliente puede ejecutar la misma consulta y obtener un resultado diferente.

Almacenamiento remoto en caché

Para resolver el problema de la duplicación de datos entre clientes, se puede utilizar un servicio externo rápido, o memoria caché remota, para almacenar los datos consultados. En lugar de comprobar un almacén de datos local, cada cliente comprobará la memoria caché remota antes de consultar el almacén de datos del backend. Esta estrategia facilita respuestas más coherentes entre los clientes, una mayor eficiencia en los datos almacenados y un mayor volumen de datos en caché, ya que el espacio de almacenamiento se escala independientemente de los clientes.

La desventaja de una memoria caché remota es que es posible que todo el sistema tenga una latencia mayor, ya que se requiere un salto de red adicional para comprobar la memoria caché remota. A fin de mejorar la latencia, es posible utilizar el almacenamiento en caché del lado del cliente junto con el almacenamiento en caché remoto para el almacenamiento en caché de varios niveles.

Pasos para la implementación

1. Identifique las bases de datos, las API y los servicios de red que podrían beneficiarse del almacenamiento en caché. Los servicios que tienen cargas de trabajo de lectura pesadas, tienen una alta relación de lectura y escritura o son caros de escalar son candidatos para el almacenamiento en caché.
   - Almacenamiento en caché de base de datos
   - Habilitación del almacenamiento en caché de la API para mejorar la capacidad de respuesta
2. Identifique el tipo de estrategia de almacenamiento en caché adecuada que mejor se adapte a su patrón de acceso.
   - Estrategias de almacenamiento en caché
   - Soluciones de almacenamiento en caché de AWS
3. Siga las prácticas recomendadas del almacenamiento en caché para su almacén de datos.
4. Configure una estrategia de invalidación de caché, como un tiempo de vida (TTL), para todos los datos que equilibre la actualización de los datos y reduzca la presión sobre el almacén de datos de backend.

5. Habilite características como reintentos de conexión automáticos, retroceso exponencial, tiempos de espera del lado del cliente y agrupación de conexiones en el cliente, si están disponibles, ya que pueden mejorar el rendimiento y la fiabilidad.
   - Prácticas recomendadas: clientes de Redis y Amazon ElastiCache for Redis
6. Supervise la tasa de aciertos de la caché con un objetivo del 80 % o superior. Los valores más bajos pueden indicar un tamaño de caché insuficiente o un patrón de acceso que no se beneficia del almacenamiento en caché.
   - ¿Qué métricas debo supervisar?
   - Best practices for monitoring Redis workloads on Amazon ElastiCache
   - Monitoring best practices with Amazon ElastiCache for Redis using Amazon CloudWatch
7. Implemente la replicación de datos para descargar las lecturas en varias instancias y mejorar el rendimiento y la disponibilidad de la lectura de datos.

Recursos

Documentos relacionados:

- Uso del enfoque Well-Architected de Amazon ElastiCache
- Monitoring best practices with Amazon ElastiCache for Redis using Amazon CloudWatch
- ¿Qué métricas debo supervisar?
- Documento técnico Performance at Scale with Amazon ElastiCache
- Desafíos y estrategias del almacenamiento en caché

Vídeos relacionados:

- Amazon ElastiCache Learning Path
- Design for success with Amazon ElastiCache best practices
- AWS re:Invent 2020 - Design for success with Amazon ElastiCache best practices
- AWS re:Invent 2023 - [LAUNCH] Introducing Amazon ElastiCache Serverless
- AWS re:Invent 2022 - 5 great ways to reimagine your data layer with Redis
- AWS re:Invent 2021 - Deep dive on Amazon ElastiCache for Redis

Ejemplos relacionados:

- Boosting MySQL database performance with Amazon ElastiCache for Redis

Redes y entrega de contenido

La solución de redes óptima para una carga de trabajo varía según los requisitos de latencia, rendimiento, fluctuaciones y ancho de banda. Las limitaciones físicas, como los recursos de usuario o locales, determinan las opciones de ubicación. Estas limitaciones pueden compensarse con las ubicaciones periféricas o la ubicación de los recursos.

En AWS, las redes se virtualizan y están disponibles en diversos tipos y configuraciones. Esto facilita la adaptación de las redes a sus necesidades. AWS ofrece características de producto, como por ejemplo redes mejoradas, instancias optimizadas para redes de Amazon EC2, aceleración de la transferencia de Amazon S3 y Amazon CloudFront dinámico, con el fin de optimizar el tráfico de red. AWS también ofrece características de red, como enrutamiento de latencia de Amazon Route 53, puntos de conexión de Amazon VPC, AWS Direct Connect y AWS Global Accelerator, para reducir la distancia o las fluctuaciones de red.

Esta área de enfoque comparte la guía y las prácticas recomendadas para diseñar, configurar y operar soluciones de redes y entrega de contenido eficientes en la nube.

Prácticas recomendadas

- PERF04-BP01 Comprender cómo afectan las redes al rendimiento
- PERF04-BP02 Evaluar las características de las redes disponibles
- PERF04-BP03 Elegir la conectividad o VPN dedicadas adecuadas para la carga de trabajo
- PERF04-BP04 Utilizar el equilibrio de carga para distribuir el tráfico entre varios recursos
- PERF04-BP05 Elegir los protocolos de red para mejorar el rendimiento
- PERF04-BP06 Elegir la ubicación de la carga de trabajo en función de los requisitos de la red
- PERF04-BP07 Optimizar la configuración de red según las métricas

PERF04-BP01 Comprender cómo afectan las redes al rendimiento

Analice y comprenda cómo las decisiones relacionadas con la red afectan a su carga de trabajo para ofrecer un rendimiento eficiente y una mejor experiencia de usuario.

Patrones comunes de uso no recomendados:

- Todo el tráfico fluye a través de sus centros de datos existentes.

- Enruta todo el tráfico a través de firewalls centrales en lugar de utilizar herramientas de seguridad de red nativas en la nube.
- Aprovisiona conexiones de AWS Direct Connect sin comprender los requisitos de uso reales.
- No tiene en cuenta las características de la carga de trabajo ni la sobrecarga de cifrado al definir sus soluciones de redes.
- Utiliza conceptos y estrategias locales para las soluciones de redes en la nube.

Beneficios de establecer esta práctica recomendada: comprender el impacto de las redes en el rendimiento de la carga de trabajo le ayuda a identificar posibles cuellos de botella, mejorar la experiencia del usuario, aumentar la fiabilidad y reducir el mantenimiento operativo a medida que cambia la carga de trabajo.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Alto

Guía para la implementación

La red es responsable de la conectividad entre los componentes de las aplicaciones, los servicios en la nube, las redes periféricas y los datos locales, por lo que puede tener un gran impacto en el rendimiento de las cargas de trabajo. Además del rendimiento de la carga de trabajo, la experiencia del usuario también puede verse afectada por la latencia de la red, el ancho de banda, los protocolos, la ubicación, la congestión de la red, las fluctuaciones, el rendimiento y las reglas de enrutamiento.

Disponga de una lista documentada de los requisitos de redes de la carga de trabajo, incluida la latencia, el tamaño de los paquetes, las reglas de enrutamiento, los protocolos y los patrones de tráfico que admiten. Examine las soluciones de red disponibles e identifique qué servicio se ajusta a las características de red de su carga de trabajo. Las redes basadas en la nube se pueden reconstruir rápidamente, de modo que hacer evolucionar su arquitectura de red con el tiempo resulta necesario para mantener la eficiencia del rendimiento.

Pasos para la aplicación:

1. Defina y documente los requisitos de rendimiento de la red e incluya métricas como la latencia de red, el ancho de banda, los protocolos, las ubicaciones, los patrones de tráfico (picos y frecuencia), el rendimiento, el cifrado, la inspección y las reglas de enrutamiento.
2. Obtenga información sobre los servicios de redes de AWS clave, como VPC, AWS Direct Connect, Elastic Load Balancing (ELB)y Amazon Route 53.

3. Recoja las siguientes características clave de la red:

Características	Herramientas y métricas
Características fundamentales de las redes	• Registros de flujo de VPC • Registros de flujo de AWS Transit Gateway • Métricas de AWS Transit Gateway • Métricas de AWS PrivateLink
Características de las redes de aplicaciones	• Elastic Fabric Adapter • Métricas de AWS App Mesh • Métricas de Amazon API Gateway
Características de las redes de periferia	• Métricas de Amazon CloudFront • Métricas de Amazon Route 53 • Métricas de AWS Global Accelerator
Características de las redes híbridas	• Métricas de AWS Direct Connect • Métricas de AWS Site-to-Site VPN • Métricas de AWS Client VPN • Métricas de Nube de AWS
Características de las redes de seguridad	• Métricas de AWS Shield, AWS WAF y AWS Network Firewall
Características de rastreo	• AWS X-Ray • VPC Reachability Analyzer • Network Access Analyzer • Amazon Inspector • Amazon CloudWatch RUM

4. Compare y pruebe el rendimiento de la red:

   a. Compare el rendimiento de la red ya que algunos factores pueden afectar al rendimiento de red de Amazon EC2 cuando las instancias están en la misma VPC. Mida el ancho de banda de la red entre las instancias Linux de Amazon EC2 en la misma VPC.

b. Realice pruebas de carga para experimentar con soluciones y opciones de redes.

Recursos

Documentos relacionados:

- Application Load Balancer
- EC2 Enhanced Networking on Linux (Redes mejoradas EC2 en Linux)
- EC2 Enhanced Networking on Windows (Redes mejoradas de EC2 en Windows)
- EC2 Placement Groups (Grupos de ubicación de EC2)
- Enabling Enhanced Networking with the Elastic Network Adapter (ENA) on Linux Instances (Habilitar redes mejoradas con Elastic Network Adapter [ENA] en las instancias de Linux)
- Network Load Balancer
- Productos de redes con AWS
- Transit Gateway
- Transición al direccionamiento basado en la latencia en Amazon Route 53
- Puntos de enlace de VPC

Vídeos relacionados:

- AWS re:Invent 2023 - AWS networking foundations
- AWS re:Invent 2023 - What can networking do for your application?
- AWS re:Invent 2023 - Advanced VPC designs and new capabilities
- AWS re:Invent 2023 - A developer’s guide to cloud networking
- AWS re:Invent 2019 - Connectivity to AWS and hybrid AWS network architectures
- AWS re:Invent 2019 - Optimizing Network Performance for Amazon EC2 Instances
- AWS Summit Online - Improve Global Network Performance for Applications
- AWS re:Invent 2020 - Networking best practices and tips with the Well-Architected Framework
- AWS re:Invent 2020 - AWS networking best practices in large-scale migrations

Ejemplos relacionados:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway y soluciones de seguridad escalables)
- AWS Networking Workshops
- Hands-on Network Firewall Workshop
- Observing and Diagnosing your Network on AWS
- Finding and addressing Network Misconfigurations on AWS

PERF04-BP02 Evaluar las características de las redes disponibles

Evalúe las características de la red en la nube que pueden aumentar el rendimiento. Medir el impacto de estas características a través de pruebas, métricas y análisis. Por ejemplo, aproveche las características a nivel de red que están disponibles para reducir la latencia, la distancia de la red o las fluctuaciones.

Patrones comunes de uso no recomendados:

- Se mantiene dentro de una región porque es allí donde se encuentra físicamente su sede.
- Utiliza firewalls en lugar de grupos de seguridad para filtrar el tráfico.
- Se infringe la TLS para inspeccionar el tráfico en lugar de confiar en grupos de seguridad, políticas de puntos de conexión y otras funciones nativas en la nube.
- Solo utiliza la segmentación basada en subredes en lugar de grupos de seguridad.

Beneficios de establecer esta práctica recomendada: Evaluar todas las características y opciones del servicio puede aumentar el rendimiento de su carga de trabajo, disminuir el esfuerzo necesario para mantener su carga de trabajo y aumentar su posición de seguridad general. Puede utilizar la estructura global de AWS para ofrecer una experiencia de red óptima a sus clientes.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Alto

Guía para la implementación

AWS ofrece servicios como AWS Global Accelerator y Amazon CloudFront que pueden ayudar a mejorar el rendimiento de la red, mientras que la mayoría de los servicios de AWS tienen características de producto (como la característica Amazon S3 Transfer Acceleration) para optimizar el tráfico de la red.

Revise qué opciones de configuración relacionadas con la red tiene a su disposición y cómo podrían afectar a su carga de trabajo. La optimización del rendimiento depende de comprender cómo interactúan estas opciones con su arquitectura y el impacto que tendrán tanto en el rendimiento medido como en la experiencia del usuario.

Pasos para la implementación

- Cree una lista de componentes de la carga de trabajo.
  - Piense en la posibilidad de usar Nube de AWS WAN para diseñar, administrar y supervisar la red de su organización al crear una red global unificada.
  - Supervise sus redes globales y principales con métricas de Amazon CloudWatch Logs. Utilice Amazon CloudWatch RUM, que proporciona información para ayudar a identificar, comprender y mejorar la experiencia digital de los usuarios.
  - Vea la latencia de red agregada entre las Regiones de AWS y las zonas de disponibilidad y dentro de cada zona de disponibilidad mediante AWS Network Manager para obtener información sobre cómo se relaciona el rendimiento de su aplicación con el rendimiento de la red de AWS subyacente.
  - Utilice una herramienta de base de datos de administración de la configuración (CMDB) existente o un servicio como AWS Config para crear un inventario de su carga de trabajo y cómo está configurada.
- Si se trata de una carga de trabajo existente, identifique y documente el punto de referencia para sus métricas de rendimiento, centrándose en los cuellos de botella y las áreas a mejorar. Las métricas de red relacionadas con el rendimiento variarán según la carga de trabajo en función de los requisitos empresariales y las características de la carga de trabajo. Para empezar, podría ser importante revisar estas métricas para su carga de trabajo: ancho de banda, latencia, pérdida de paquetes, fluctuación y retransmisiones.
- Si se trata de una nueva carga de trabajo, realice pruebas de carga para identificar cuellos de botella en el rendimiento.
- Para los cuellos de botella de rendimiento que identifique, revise las opciones de configuración de sus soluciones para identificar las oportunidades de mejora del rendimiento. Eche un vistazo a las siguientes opciones y características de red clave:

Oportunidad de mejora	Solución
Rutas de red	Utilice Network Access Analyzer para identificar rutas.
Protocolos de red	Consulte PERF04-BP05 Elegir los protocolos de red para mejorar el rendimiento
Topología de la red	Evalúe sus compensaciones operativas y de rendimiento entre Interconexión de VPC y AWS Transit Gateway al conectar varias cuentas. AWS Transit Gateway simplifica la forma de interconectar todas sus VPC, que pueden abarcar miles de Cuentas de AWS y sus redes locales. Comparta su AWS Transit Gateway entre varias cuentas utilizando AWS Resource Access Manager. Consulte PERF04-BP03 Elegir la conectividad o VPN dedicadas adecuadas para la carga de trabajo

Oportunidad de mejora	Solución
Servicios de red	AWS Global Accelerator es un servicio de redes que mejora el rendimiento del tráfico de los usuarios hasta un 60 % al utilizar la infraestructura de red global de AWS. Amazon CloudFront puede mejorar el rendimiento de la carga de trabajo, la entrega de contenido y la latencia a nivel mundial. Utilice Lambda@edge para ejecutar funciones que personalicen el contenido que CloudFront ofrece más cerca de los usuarios, reduzcan la latencia y mejoren el rendimiento. Amazon Route 53 ofrece opciones de enrutamiento basado en la latencia, enrutamiento de geolocalización, enrutamiento de geoproximidad y enrutamiento basado en IP para ayudar a mejorar el rendimiento de su carga de trabajo para una audiencia a nivel mundial. Identifique qué opción de enrutamiento optimizaría el rendimiento de su carga de trabajo revisando el tráfico de la misma y la ubicación de los usuarios cuando la carga de trabajo se distribuya globalmente.

Oportunidad de mejora	Solución
Características de los recursos de almacenamiento	Amazon S3 Transfer Acceleration es una característica que permite que los usuarios externos se beneficien de las optimizaciones de redes de CloudFront para cargar datos en Amazon S3. Esto mejora la capacidad de transferir grandes cantidades de datos desde ubicaciones remotas que no tienen conectividad dedicada a la Nube de AWS. Puntos de acceso multirregión de Amazon S3 replica el contenido en varias regiones y simplifica la carga de trabajo proporcionando un punto de acceso. Cuando se utiliza un punto de acceso multirregión, se pueden solicitar o escribir datos en Amazon S3 con el servicio que identifica el bucket de menor latencia.

Oportunidad de mejora	Solución
Características de recursos de computación	Las interfaces de redes elásticas (ENA) utilizadas por las instancias de Amazon EC2, los contenedores y las funciones de Lambda están limitadas por el flujo. Revise sus grupos de colocación para optimizar su rendimiento de red de EC2. Para evitar un cuello de botella por cada flujo, diseñe su aplicación para que utilice varios flujos. Para supervisar y obtener visibilidad de las métricas de red relacionadas con la computación, utilice métricas de CloudWatch y ethtool. La `ethtool` se incluye en el controlador ENA y expone métricas adicionales relacionadas con la red que pueden publicarse como una métrica personalizada en CloudWatch. Los Elastic Network Adapters (ENA) proporcionan una mayor optimización al ofrecer un mejor rendimiento para sus instancias dentro de un grupo con ubicación en clúster. Elastic Fabric Adapter (EFA) es una interfaz de red para instancias de Amazon EC2 que permite ejecutar cargas de trabajo que requieren altos niveles de comunicación entre nodos a escala en AWS. Las instancias optimizadas para Amazon EBS utilizan una pila de configuración optimizada y ofrecen capacidad dedicada adicional para aumentar la E/S de Amazon EBS.

Recursos

Documentos relacionados:

- Application Load Balancer
- EC2 Enhanced Networking on Linux (Redes mejoradas EC2 en Linux)
- EC2 Enhanced Networking on Windows (Redes mejoradas de EC2 en Windows)
- EC2 Placement Groups (Grupos de ubicación de EC2)
- Enabling Enhanced Networking with the Elastic Network Adapter (ENA) on Linux Instances (Habilitar redes mejoradas con Elastic Network Adapter [ENA] en las instancias de Linux)
- Network Load Balancer
- Productos de redes con AWS
- Transición al enrutamiento basado en la latencia en Amazon Route 53
- Puntos de conexión de VPC
- Registros de flujo de VPC

Vídeos relacionados:

- AWS re:Invent 2023 – Ready for what's next? Designing networks for growth and flexibility
- AWS re:Invent 2023 – Advanced VPC designs and new capabilities
- AWS re:Invent 2023 – A developer's guide to cloud networking
- AWS re:Invent 2022 – Dive deep on AWS networking infrastructure
- AWS re:Invent 2019 – Connectivity to AWS and hybrid AWS network architectures
- AWS re:Invent 2018 – Optimizing Network Performance for Amazon EC2 Instances
- AWS Global Accelerator

Ejemplos relacionados:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway y soluciones de seguridad escalables)
- AWS Networking Workshops
- Observación y diagnóstico de su red
- Finding and addressing network misconfigurations on AWS

PERF04-BP03 Elegir la conectividad o VPN dedicadas adecuadas para la carga de trabajo

Cuando se requiera conectividad híbrida para conectar los recursos locales y de la nube, aprovisione el ancho de banda adecuado para satisfacer sus requisitos de rendimiento. Calcule los requisitos de ancho de banda y de latencia para la carga de trabajo híbrida. Estas cifras determinarán los requisitos de tamaño.

Patrones comunes de uso no recomendados:

- Solo evalúa las soluciones de VPN para los requisitos de cifrado de su red.
- No evalúa las opciones de conectividad redundante o de respaldo.
- No identifica todos los requisitos de la carga de trabajo (necesidades de cifrado, protocolo, ancho de banda y tráfico).

Beneficios de establecer esta práctica recomendada: La selección y configuración de las soluciones de conectividad adecuadas aumentará la fiabilidad de su carga de trabajo y maximizará el rendimiento. Si identifica los requisitos de la carga de trabajo, planifica con antelación y evalúa las soluciones híbridas, puede minimizar los costosos cambios en la red física y los gastos operativos, a la vez que acelera el tiempo de rentabilización.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Alto

Guía para la implementación

Desarrolle una arquitectura de red híbrida basada en sus requisitos de ancho de banda. AWS Direct Connect le permite conectar su red local de forma privada con AWS. Es conveniente cuando se necesita un gran ancho de banda y baja latencia con un rendimiento uniforme. Una conexión VPN establece una conexión segura a través de Internet. Se usa cuando solo se requiere una conexión temporal, cuando el coste es un factor o como alternativa mientras se espera que se establezca una conectividad de red física resiliente durante el uso de AWS Direct Connect.

Si sus requisitos de ancho de banda son elevados, podría considerar la posibilidad de utilizar varios servicios de AWS Direct Connect o VPN. Es posible equilibrar la carga del tráfico entre los servicios, aunque no recomendamos equilibrar la carga entre AWS Direct Connect y una VPN debido a las diferencias de latencia y ancho de banda.

Pasos para la implementación

1. Calcule los requisitos de ancho de banda y de latencia de sus aplicaciones actuales.
   a. En el caso de cargas de trabajo existentes que se trasladan a AWS, utilice los datos de sus sistemas internos de supervisión de red.
   b. En el caso de cargas de trabajo nuevas o existentes para las que no disponga de datos de supervisión, consulte con los propietarios del producto para determinar las métricas de rendimiento adecuadas y ofrecer una buena experiencia de usuario.
2. Seleccione una conexión dedicada o VPN como opción de conectividad. En función de todos los requisitos de la carga de trabajo (necesidades de cifrado, ancho de banda y tráfico), puede elegir AWS Direct Connect o AWS VPN (o ambas). El siguiente diagrama puede ayudarle a elegir el tipo de conexión adecuado.
   a. AWS Direct Connect ofrece conectividad dedicada al entorno de AWS, desde 50 Mbps hasta 100 Gbps, mediante conexiones dedicadas o conexiones alojadas. Esto le ofrece un ancho de banda aprovisionado y una latencia administrada y controlada, a fin de que su carga de trabajo pueda conectarse de manera eficiente a otros entornos. Mediante el uso de socios de AWS Direct Connect, puede disponer de conectividad de extremo a extremo desde varios entornos, lo que proporciona una red ampliada con un rendimiento coherente. AWS ofrece un ancho de banda de conexión directa escalable mediante 100 Gbps nativos, un grupo de agregación de enlaces (LAG) o varias rutas de igual coste (ECMP) con BGP.
   b. La AWS Site-to-Site VPN proporciona un servicio de VPN administrado compatible con la seguridad del protocolo de Internet (IPsec). Cuando se crea una conexión VPN, cada conexión VPN incluye dos túneles para ofrecer una alta disponibilidad.
3. Siga la documentación de AWS para elegir la opción de conectividad adecuada:
   a. Si decide usar AWS Direct Connect, seleccione el ancho de banda adecuado para su conectividad.
   b. Si utiliza una AWS Site-to-Site VPN a través de numerosas ubicaciones para conectarse a una Región de AWS, use una conexión de Site-to-Site VPN acelerada para tener la oportunidad de mejorar el rendimiento de la red.
   c. Si el diseño de su red consiste en una conexión VPN IPsec a través de AWS Direct Connect, considere usar una VPN con IP privada para mejorar la seguridad y lograr la segmentación. La VPN con IP privada de AWS Site-to-Site se despliega sobre la interfaz virtual de tránsito (VIF).
   d. AWS Direct Connect SiteLink permite crear conexiones redundantes y de baja latencia entre sus centros de datos de todo el mundo mediante el envío de datos a través de la ruta más corta entre las ubicaciones de AWS Direct Connect, sin pasar por las Regiones de AWS.

4. Valide la configuración de la conectividad antes del despliegue en producción. Lleve a cabo pruebas de seguridad y rendimiento para asegurarse de que cumple los requisitos de ancho de banda, fiabilidad, latencia y cumplimiento.
5. Supervise periódicamente el rendimiento y el uso de la conectividad y optimícelo si es necesario.

Diagrama de flujo de rendimiento determinístico

Recursos

Documentos relacionados:

- Productos de redes con AWS
- AWS Transit Gateway
- Puntos de enlace de VPC
- Building a Scalable and Secure Multi-VPC AWS Network Infrastructure
- Cliente VPN

Vídeos relacionados:

- AWS re:Invent 2023 – Building hybrid network connectivity with AWS
- AWS re:Invent 2023 – Secure remote connectivity to AWS
- AWS re:Invent 2022 – Optimizing performance with Amazon CloudFront
- AWS re:Invent 2019 – Connectivity to AWS and hybrid AWS network architectures
- AWS re:Invent 2020 – AWS Transit Gateway Connect

Ejemplos relacionados:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway y soluciones de seguridad escalables)
- AWS Networking Workshops (Talleres de red de AWS)

PERF04-BP04 Utilizar el equilibrio de carga para distribuir el tráfico entre varios recursos

Distribuya el tráfico entre varios recursos o servicios para que su carga de trabajo aproveche la elasticidad que ofrece la nube. También puede utilizar el equilibrio de carga para descargar la terminación del cifrado con el objetivo de mejorar el rendimiento, la fiabilidad y administrar y enrutar el tráfico de manera eficaz.

Antipatrones usuales:

- No se tienen en cuenta los requisitos de la carga de trabajo al elegir el tipo de equilibrador de carga.
- No se aprovechan las características del equilibrador de carga para optimizar el rendimiento.
- La carga de trabajo se expone directamente a Internet sin un equilibrador de carga.
- Enruta todo el tráfico de Internet a través de los equilibradores de carga existentes.
- Utiliza el equilibrio de carga TCP genérico y hace que cada nodo de computación gestione el cifrado SSL.

Beneficios de establecer esta práctica recomendada: un equilibrador de carga gestiona la carga variable del tráfico de la aplicación en una única zona de disponibilidad o en varias zonas de disponibilidad y facilita una alta disponibilidad, un escalamiento automático y una mejor utilización de la carga de trabajo.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: alto

Guía para la implementación

Los equilibradores de carga actúan como punto de entrada de la carga de trabajo y, a partir de ahí, distribuyen el tráfico a los destinos de backend, como instancias de computación o contenedores, para mejorar la utilización.

La elección del tipo de equilibrador de carga adecuado es el primer paso para optimizar su arquitectura. Comience por enumerar las características de su carga de trabajo, como el protocolo (por ejemplo, TCP, HTTP, TLS o WebSockets), el tipo de destino (como instancias, contenedores o sin servidor), los requisitos de la aplicación (como conexiones de larga duración, autenticación de usuarios o permanencia) y la ubicación (como región, Local Zone, Outpost o aislamiento zonal).

AWS le ofrece varios modelos para que sus aplicaciones utilicen el equilibrio de carga. Application Load Balancer es el más adecuado para el equilibrio de carga del tráfico de HTTP y HTTPS y entrega un enrutamiento de solicitudes avanzado centrado en la entrega de arquitecturas de aplicaciones modernas, incluidos los microservicios y los contenedores.

Network Load Balancer es el más adecuado para el equilibrio de carga del tráfico de TCP en donde se necesite un rendimiento extremo. Es capaz de gestionar millones de solicitudes por segundo manteniendo latencias ultrabajas, y está optimizado para manejar patrones de tráfico repentinos y volátiles.

Elastic Load Balancing proporciona administración de certificados y descifrado SSL/TLS integrados, lo que le permite la flexibilidad de administrar de forma centralizada la configuración SSL del equilibrador de carga y descargar el trabajo intensivo de la CPU de su carga de trabajo.

Una vez elegido el equilibrador de carga adecuado, puede empezar a utilizar sus características para reducir el esfuerzo que debe realizar su backend para atender al tráfico.

Por ejemplo, al utilizar tanto Application Load Balancer (ALB) como Network Load Balancer (NLB), puede realizar la descarga de cifrado SSL/TLS, lo que da la oportunidad de evitar que sus destinos completen el establecimiento de comunicación TLS, que consume mucha CPU, y también para mejorar la administración de certificados.

Cuando configura la descarga SSL/TLS en el equilibrador de carga, este se ocupa del cifrado del tráfico desde y hacia los clientes, al tiempo que entrega el tráfico sin cifrar a sus backends, lo que libera recursos de backend y mejora el tiempo de respuesta para los clientes.

Application Load Balancer también puede atender el tráfico HTTP/2 sin necesidad de soporte en sus destinos. Esta simple decisión puede mejorar el tiempo de respuesta de su aplicación, ya que HTTP/2 utiliza las conexiones TCP de forma más eficiente.

Los requisitos de latencia de la carga de trabajo deben tenerse en cuenta a la hora de definir la arquitectura. Por ejemplo, si tiene una aplicación sensible a la latencia, puede decidir utilizar Network Load Balancer, que ofrece latencias extremadamente bajas. Como alternativa, puede decidir acercar su carga de trabajo a sus clientes con Application Load Balancer en zonas locales de AWS o incluso AWS Outposts.

Otra consideración para las cargas de trabajo sensibles a la latencia es el equilibrio de carga entre zonas. Con el equilibrio de carga entre zonas, cada nodo del equilibrador de carga distribuye el tráfico entre los destinos registrados en todas las zonas de disponibilidad permitidas.

Utilice Auto Scaling integrado con su equilibrador de carga. Uno de los aspectos clave de un sistema con un rendimiento eficiente tiene que ver con el redimensionamiento correcto de sus recursos de backend. Para ello, puede utilizar las integraciones del equilibrador de carga para los recursos de destino de backend. Mediante la integración del equilibrador de carga con los grupos de Auto Scaling, los destinos se añadirán o eliminarán del equilibrador de carga según sea necesario y en respuesta al tráfico entrante. Los equilibradores de carga también pueden integrarse con Amazon ECS y Amazon EKS para cargas de trabajo en contenedores.

- Amazon ECS: equilibro de la carga de servicios
- Equilibrio de carga de aplicaciones en Amazon EKS
- Equilibrio de carga de red en Amazon EKS

Pasos para la implementación

- Defina sus requisitos de equilibrio de carga, incluidos el volumen de tráfico, la disponibilidad y la escalabilidad de las aplicaciones.
- Elija el tipo de equilibrador de carga adecuado para su aplicación.
  - Utilice Application Load Balancer para cargas de trabajo HTTP/HTTPS.
  - Utilice Network Load Balancer para cargas de trabajo distintas de HTTP que se ejecuten en TCP o UDP.
  - Utilice una combinación de ambos (ALB como destino de NLB) si desea aprovechar las características de ambos productos. Por ejemplo, puede hacerlo si desea utilizar las IP estáticas de NLB junto con el enrutamiento basado en encabezado HTTP de ALB, o si desea exponer su carga de trabajo HTTP a una AWS PrivateLink.
  - Para obtener una comparación completa de los equilibradores de carga, consulte la comparación de productos de ELB.
- Utilice la descarga SSL/TLS si es posible.
  - Configure los agentes de escucha HTTPS/TLS con Application Load Balancer y Network Load Balancer integrados con AWS Certificate Manager.
  - Tenga en cuenta que algunas cargas de trabajo pueden requerir cifrado de extremo a extremo por motivos de conformidad. En este caso, es un requisito permitir el cifrado en los destinos.
  - Para conocer las prácticas recomendadas de seguridad, consulte SEC09-BP02 Aplicar el cifrado en tránsito.
- Seleccione el algoritmo de enrutamiento adecuado (solo ALB).

  - El algoritmo de enrutamiento puede marcar la diferencia en el grado de utilización de sus destinos de backend y, por lo tanto, en su repercusión en el rendimiento. Por ejemplo, ALB proporciona dos opciones para los algoritmos de enrutamiento:
  - Solicitudes menos pendientes: utilice esta opción para lograr una mejor distribución de la carga a sus destinos de backend para los casos en que las solicitudes de la aplicación varíen en complejidad o los destinos varíen en capacidad de procesamiento.
  - Distribución: utilice esta otra opción cuando las solicitudes y los destinos sean similares, o si necesita distribuir las solicitudes equitativamente entre los destinos.
- Considere el aislamiento entre zonas o zonal.
  - Desactive el aislamiento entre zonas (aislamiento zonal) para mejorar la latencia y los dominios de error zonal. Está desactivado de forma predeterminada en NLB. En ALB, puede desactivarlo por grupo de destino.
  - Active el aislamiento entre zonas para aumentar la disponibilidad y flexibilidad. Está activado de forma predeterminada en ALB. En NLB, puede activarlo por grupo de destino.
- Active la conexión persistente HTTP para sus cargas de trabajo HTTP (solo ALB). Con esta característica, el equilibrador de carga puede reutilizar las conexiones de backend hasta que expire el tiempo de espera activo, lo que mejora el tiempo de solicitud y respuesta HTTP, además de reducir la utilización de recursos en los destinos de backend. Para obtener información detallada sobre cómo hacer esto para Apache y Nginx, consulte What are the optimal settings for using Apache or NGINX as a backend server for ELB?
- Active la supervisión de su equilibrador de carga.
  - Active los registros de acceso para su Application Load Balancer y Network Load Balancer.
  - Los principales campos a tener en cuenta para ALB son `request_processing_time`, `request_processing_time` y `response_processing_time`.
  - Los principales campos a tener en cuenta para NLB son `connection_time` y `tls_handshake_time`.
  - Esté preparado para consultar los registros cuando los necesite. Puede utilizar Amazon Athena para consultar tanto los registros de ALB como los registros de NLB.
  - Cree alarmas para las métricas relacionadas con el rendimiento, como `TargetResponseTime` para ALB.

Recursos

Documentos relacionados:

- Comparación de productos ELB
- Infraestructura global de AWS
- Improving Performance and Reducing Cost Using Availability Zone Affinity
- Step by step for Log Analysis with Amazon Athena
- Querying Application Load Balancer logs
- Monitor your Application Load Balancers
- Monitor your Network Load Balancer
- Use Elastic Load Balancing to distribute traffic across the instances in your Auto Scaling group

Vídeos relacionados:

- AWS re:Invent 2023: What can networking do for your application?
- AWS re:Inforce 20: How to use Elastic Load Balancing to enhance your security posture at scale
- AWS re:Invent 2018: Elastic Load Balancing: Deep Dive and Best Practices
- AWS re:Invent 2021 - How to choose the right load balancer for your AWS workloads
- AWS re:Invent 2019: Get the most from Elastic Load Balancing for different workloads

Ejemplos relacionados:

- Gateway Load Balancer
- CDK and AWS CloudFormation samples for Log Analysis with Amazon Athena

PERF04-BP05 Elegir los protocolos de red para mejorar el rendimiento

Tome decisiones sobre los protocolos de comunicación entre sistemas y redes en función del impacto en el rendimiento de la carga de trabajo.

Existe una relación entre la latencia y el ancho de banda para lograr el rendimiento. Si la transferencia de archivos utiliza el protocolo de control de transmisión (TCP), las latencias más

altas probablemente reducirán el rendimiento general. Existen enfoques para solucionar esto con el ajuste de TCP y protocolos de transferencia optimizados, pero una solución es utilizar el protocolo de datagramas de usuario (UDP).

Patrones comunes de uso no recomendados:

- Utiliza TCP para todas las cargas de trabajo, independientemente de los requisitos de rendimiento.

Beneficios de establecer esta práctica recomendada: Verificar que se utiliza un protocolo adecuado para la comunicación entre los usuarios y los componentes de la carga de trabajo ayuda a mejorar la experiencia general del usuario para sus aplicaciones. Por ejemplo, UDP sin conexión permite una alta velocidad, pero no ofrece retransmisión ni alta fiabilidad. TCP es un protocolo con todas las características, pero requiere una mayor sobrecarga para procesar los paquetes.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Medio

Guía para la implementación

Si tiene la capacidad de elegir diferentes protocolos para su aplicación y tiene experiencia en esta área, optimice la aplicación y la experiencia del usuario final utilizando un protocolo diferente. Tenga en cuenta que este enfoque presenta una dificultad significativa y solo debe intentarse si primero ha optimizado su aplicación de otras maneras.

Una consideración primordial para mejorar el rendimiento de la carga de trabajo es comprender los requisitos de latencia y rendimiento, y luego elegir protocolos de red que optimicen el rendimiento.

Cuándo considerar el uso de TCP

TCP proporciona una entrega de datos fiable, y se puede utilizar para la comunicación entre los componentes de la carga de trabajo cuando la fiabilidad y la entrega garantizada de datos es importante. Muchas aplicaciones basadas en web dependen de protocolos basados en TCP, como HTTP y HTTPS, con el fin de abrir sockets TCP para la comunicación entre componentes de la aplicación. La transferencia de datos de correo electrónico y archivos son aplicaciones habituales que también utilizan TCP, ya que es un mecanismo de transferencia sencillo y fiable entre los componentes de la aplicación. El uso de TLS con TCP puede añadir cierta sobrecarga a la comunicación, lo que puede provocar un aumento de la latencia y una reducción del rendimiento, pero tiene la ventaja de seguridad. La sobrecarga proviene principalmente de la sobrecarga añadida del proceso de establecimiento de comunicación, que puede tardar varias idas y vueltas

en completarse. Una vez completado el proceso, la sobrecarga de cifrado y descifrado de datos es relativamente pequeña.

Cuándo considerar el uso de UDP

UDP es un protocolo sin conexión y, por tanto, adecuado para aplicaciones que necesitan una transmisión rápida y eficiente, como datos de registro, supervisión y VoIP. Además, considere el uso de UDP si tiene componentes de carga de trabajo que responden a pequeñas consultas de un gran número de clientes, a fin de garantizar un rendimiento óptimo de la carga de trabajo. La seguridad de la capa de transporte de datagramas (DTLS) es el equivalente UDP de la seguridad de la capa de transporte (TLS). Cuando se utiliza DTLS con UDP, la sobrecarga proviene del cifrado y descifrado de los datos, ya que el proceso de establecimiento de comunicación se simplifica. DTLS también añade una pequeña cantidad de sobrecarga a los paquetes UDP, ya que incluye campos adicionales para indicar los parámetros de seguridad y detectar manipulaciones.

Cuándo considerar el uso de SRD

Scalable reliable datagram (SRD) es un protocolo de transporte de red optimizado para cargas de trabajo de alto rendimiento debido a su capacidad para equilibrar la carga de tráfico a través de numerosas rutas y recuperarse rápidamente de las caídas de paquetes o errores de enlace. Por lo tanto, es mejor utilizar SRD para cargas de trabajo de computación de alto rendimiento (HPC) que exigen un alto rendimiento y una comunicación de baja latencia entre nodos de computación. Esto incluye tareas de procesamiento paralelo como simulación, modelado y análisis de datos que impliquen una gran cantidad de transferencia de datos entre nodos.

Pasos para la implementación

1. Utilice la AWS Global Accelerator y AWS Transfer Family para mejorar el rendimiento de sus aplicaciones de transferencia de archivos en línea. El servicio AWS Global Accelerator le ayuda a conseguir una latencia menor entre sus dispositivos cliente y su carga de trabajo en AWS. Con AWS Transfer Family, puede utilizar protocolos basados en TCP como el protocolo de transferencia de archivos de shell seguro (SFTP) y el protocolo de transferencia de archivos sobre SSL (FTPS) para escalar y administrar de forma segura las transferencias de archivos a los servicios de almacenamiento de AWS.
2. Utilice la latencia de la red para determinar si TCP es adecuado para la comunicación entre los componentes de la carga de trabajo. Si la latencia de la red entre la aplicación cliente y el servidor es alta, la comunicación TCP de tres vías puede tardar un tiempo, lo que afectará a la capacidad de respuesta de la aplicación. Para medir la latencia de la red pueden utilizarse métricas, como el

tiempo hasta el primer byte (TTFB) y el tiempo de ida y vuelta (RTT). Si su carga de trabajo ofrece contenido dinámico a los usuarios, considere la posibilidad de utilizar Amazon CloudFront, que establece una conexión persistente con cada origen para el contenido dinámico para eliminar el tiempo de configuración de la conexión que, de otro modo, ralentizaría cada solicitud del cliente.

3. El uso de TLS con TCP o UDP puede aumentar la latencia y reducir el rendimiento de la carga de trabajo debido al impacto del cifrado y el descifrado. Para este tipo de cargas de trabajo, considere la posibilidad de descargar SSL/TLS en Elastic Load Balancing para mejorar el rendimiento de la carga de trabajo al permitir que el equilibrador de carga gestione el proceso de cifrado y descifrado SSL/TLS, en lugar de que lo hagan las instancias de backend. Esto puede ayudar a reducir la utilización de la CPU en las instancias backend, lo que puede mejorar el rendimiento y aumentar la capacidad.
4. Utilice la Network Load Balancer (NLB) para desplegar servicios que dependan del protocolo UDP, como autenticación y autorización, registro, DNS, IoT y streaming multimedia, para mejorar el rendimiento y la fiabilidad de su carga de trabajo. El NLB distribuye el tráfico UDP entrante entre varios destinos, lo que le permite escalar su carga de trabajo horizontalmente, aumentar la capacidad y reducir la sobrecarga de un único destino.
5. Para sus cargas de trabajo de computación de alto rendimiento (HPC), considere la posibilidad de utilizar la funcionalidad Elastic Network Adapter (ENA) que utiliza el protocolo SRD para mejorar el rendimiento de la red al proporcionar un mayor ancho de banda de flujo único (25 Gbps) y una menor latencia de cola (percentil 99,0) para el tráfico de red entre instancias de EC2.
6. Utilice la Application Load Balancer (ALB) para enrutar y equilibrar la carga del tráfico gRPC (llamadas a procedimientos remotos) entre componentes de carga de trabajo o entre clientes y servicios gRPC. gRPC utiliza el protocolo HTTP/2 basado en TCP para el transporte y proporciona ventajas de rendimiento como una huella de red más ligera, compresión, serialización binaria eficiente, compatibilidad con numerosos idiomas y streaming bidireccional.

Recursos

Documentos relacionados:

- How to route UDP traffic into Kubernetes
- Application Load Balancer
- EC2 Enhanced Networking on Linux (Redes mejoradas EC2 en Linux)
- EC2 Enhanced Networking on Windows (Redes mejoradas de EC2 en Windows)
- EC2 Placement Groups (Grupos de ubicación de EC2)

- Enabling Enhanced Networking with the Elastic Network Adapter (ENA) on Linux Instances (Habilitar redes mejoradas con Elastic Network Adapter [ENA] en las instancias de Linux)
- Network Load Balancer
- Productos de redes con AWS
- Transición al enrutamiento basado en la latencia en Amazon Route 53
- Puntos de enlace de VPC

Vídeos relacionados:

- AWS re:Invent 2022 – Scaling network performance on next-gen Amazon Elastic Compute Cloud instances
- AWS re:Invent 2022 – Application networking foundations

Ejemplos relacionados:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway y soluciones de seguridad escalables)
- AWS Networking Workshops (Talleres de red de AWS)

PERF04-BP06 Elegir la ubicación de la carga de trabajo en función de los requisitos de la red

Evalúe las opciones de colocación de recursos para reducir la latencia de la red y mejorar el rendimiento, lo que proporcionará una experiencia de usuario óptima al reducir los tiempos de carga de las páginas y de transferencia de datos.

Antipatrones usuales:

- Consolida todos los recursos de la carga de trabajo en una ubicación geográfica.
- Ha elegido la región más cercana a su ubicación, pero no al usuario final de la carga de trabajo.

Beneficios de establecer esta práctica recomendada: La experiencia del usuario se ve muy afectada por la latencia entre el usuario y la aplicación. Al utilizar las Regiones de AWS adecuadas y la red global privada de AWS, puede reducir la latencia y ofrecer una mejor experiencia a los usuarios remotos.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: medio

Guía para la implementación

Los recursos, como las instancias de Amazon EC2, se colocan en zonas de disponibilidad dentro de Regiones de AWS, zonas locales de AWS, AWS Outposts o zonas de AWS Wavelength. La selección de esta ubicación influye en la latencia y el rendimiento de la red desde una ubicación de usuario. Los servicios de periferia, como Amazon CloudFront y AWS Global Accelerator también se pueden utilizar para mejorar el rendimiento de la red al almacenar contenido en caché en ubicaciones periféricas o proporcionar a los usuarios una ruta óptima a la carga de trabajo a través de la red global de AWS.

Amazon EC2 ofrece grupos de colocación para la creación de redes. Un grupo de registro es una agrupación lógica de instancias para reducir la latencia. El uso de grupos de colocación con tipos de instancias compatibles y un Elastic Network Adapter (ENA) permite que las cargas de trabajo participen en una red de 25 Gbps de baja latencia y fluctuación reducida. Se recomiendan grupos de colocación para cargas de trabajo que aprovechan la baja latencia de red, el alto rendimiento de red o ambos.

Los servicios sensibles a la latencia se prestan en ubicaciones periféricas mediante una red global de AWS, comoAmazon CloudFront. Estas ubicaciones periféricas normalmente prestan servicios como red de entrega de contenido (CDN) y sistema de nombres de dominio (DNS). Al tener estos servicios en la periferia, las cargas de trabajo pueden responder con baja latencia a las solicitudes de contenido o de resolución de DNS. Estos servicios pueden ofrecer servicios geográficos como la geolocalización del contenido (que proporciona contenido diferente según la ubicación de los usuarios finales) o el enrutamiento basado en la latencia para dirigir a los usuarios finales hacia la región más cercana (latencia mínima).

Utilice los servicios periféricos para reducir la latencia y permitir el almacenamiento en caché del contenido. Configure correctamente el control de caché para DNS y HTTP/HTTPS a fin de obtener el mayor beneficio de estos enfoques.

Pasos para la implementación

- Recoja información sobre el tráfico IP que entra y sale de las interfaces de red.
  - Registro del tráfico de IP con registros de flujo de la VPC
  - Cómo se conserva la dirección IP del cliente en AWS Global Accelerator
- Analice los patrones de acceso de la red en su carga de trabajo para identificar cómo utilizan los usuarios su aplicación.

  - Utilice herramientas de supervisión, como Amazon CloudWatch y AWS CloudTrail, para recopilar datos sobre las actividades de red.
  - Analice los datos para identificar el patrón de acceso de la red.
- Seleccione regiones para el despliegue de la carga de trabajo en función de los siguientes elementos clave:
  - Dónde se encuentran sus datos: en el caso de las aplicaciones con gran cantidad de datos (como big data y machine learning), el código de la aplicación debe ejecutarse lo más cerca posible de los datos.
  - Dónde se encuentran sus usuarios: en el caso de las aplicaciones orientadas al usuario, elija una región (o varias regiones) cerca de los usuarios de su carga de trabajo.
  - Otras restricciones: tenga en cuenta restricciones como el coste y el cumplimiento, tal y como se explica en What to Consider when Selecting a Region for your Workloads.
- Utilice zonas locales de AWS para ejecutar cargas de trabajo como la renderización de vídeo. Las zonas locales le permiten beneficiarse de tener recursos de computación y almacenamiento más cerca de los usuarios finales.
- Utilice AWS Outposts para cargas de trabajo que deban seguir siendo locales y en las que desee que esa carga de trabajo se ejecute sin problemas con el resto de sus demás cargas de trabajo en AWS.
- Aplicaciones como la transmisión de vídeo en directo de alta resolución, audio de alta fidelidad y realidad aumentada/realidad virtual (RA/RV) requieren una latencia ultrabaja para dispositivos 5G. Para este tipo de aplicaciones, considere AWS Wavelength. AWS Wavelength integra los servicios de computación y almacenamiento de AWS en las redes 5G, lo que proporciona una infraestructura de computación periférica móvil para desarrollar, desplegar y escalar aplicaciones de latencia ultrabaja.
- Utilice almacenamiento en caché local o soluciones de almacenamiento en caché de AWS para los recursos de uso frecuente con el fin de mejorar el rendimiento, reducir el movimiento de datos y disminuir el impacto medioambiental.

Service	When to use
Amazon CloudFront	Se usa para almacenar en caché el contenido estático como imágenes, scripts y vídeos, así como el contenido dinámico como respuestas de API y aplicaciones web.

Service	When to use
Amazon ElastiCache	Se usa para almacenar en caché el contenido de las aplicaciones web.
DynamoDB Accelerator	Se usa para añadir aceleración en memoria a sus tablas de DynamoDB.

- Utilice servicios que puedan ayudarle a ejecutar el código más cerca de los usuarios de su carga de trabajo, como estas:

Service	When to use
Lambda@edge	Se usa para las operaciones que utilizan muchos recursos de computación que se inician cuando los objetos no están en la memoria caché.
Amazon CloudFront Functions	Se usan para casos de uso sencillos como las manipulaciones de solicitudes o respuesta s HTTP(s) que pueden iniciarse mediante funciones de corta duración.
AWS IoT Greengrass	Se usa para ejecutar la computación local, la mensajería y el almacenamiento en caché de datos para los dispositivos conectados.

- Algunas aplicaciones requieren puntos de entrada fijos o un mayor rendimiento mediante el aumento del rendimiento y la reducción de la fluctuación y de la latencia del primer byte. Estas aplicaciones pueden beneficiarse de los servicios de red que proporcionan direcciones IP estáticas de difusión por proximidad y terminación TCP en ubicaciones periféricas. AWS Global Accelerator puede mejorar el rendimiento de las aplicaciones hasta en un 60 % y proporcionar una rápida conmutación por error para arquitecturas multirregión. AWS Global Accelerator le proporciona direcciones IP estáticas de difusión por proximidad que sirven como punto de entrada fijo para las aplicaciones alojadas en una o más Regiones de AWS. Estas direcciones IP permiten que el tráfico entre en la red global de AWS lo más cerca posible de sus usuarios. AWS Global Accelerator reduce el tiempo de configuración de la conexión inicial al establecer una conexión TCP entre el cliente y la ubicación periférica de AWS más cercana al cliente. Revise el uso

de AWS Global Accelerator para mejorar el rendimiento de sus cargas de trabajo TCP/UDP y proporcionar una rápida conmutación por error para arquitecturas multirregión.

Recursos

Prácticas recomendadas relacionadas:

- COST07-BP02 Implementar regiones según los costes
- COST08-BP03 Implementar servicios para reducir los costes de transferencia de datos
- REL10-BP01 Desplegar la carga de trabajo en varias ubicaciones
- REL10-BP02 Seleccionar las ubicaciones adecuadas para el despliegue en varias ubicaciones
- SUS01-BP01 Elegir la región basándose tanto en los requisitos empresariales como en los objetivos de sostenibilidad
- SUS02-BP04 Optimizar la ubicación geográfica de las cargas de trabajo en función de sus requisitos de red
- SUS04-BP07: Minimización del movimiento de datos entre redes

Documentos relacionados:

- Infraestructura global de AWS
- AWS Local Zones and AWS Outposts, choosing the right technology for your edge workload (AWS Local Zones y AWS Outposts: elegir la tecnología adecuada para su carga de trabajo de periferia)
- Grupos de ubicación
- Zonas locales de AWS
- AWS Outposts
- AWS Wavelength
- Amazon CloudFront
- AWS Global Accelerator
- AWS Direct Connect
- AWS Site-to-Site VPN
- Amazon Route 53

Vídeos relacionados:

- AWS Local Zones Explainer Video (Vídeo explicativo de AWS Local Zones)
- AWS Outposts: Overview and How it Works
- AWS re:Invent 2023 - A migration strategy for edge and on-premises workloads
- AWS re:Invent 2021 - AWS Outposts: Bringing the AWS experience on premises (AWS re:Invent 2021: Llevar la experiencia de AWS al entorno local)
- AWS re:Invent 2020: AWS Wavelength: Run apps with ultra-low latency at 5G edge
- AWS re:Invent 2022 - AWS Local Zones: Building applications for a distributed edge (AWS re:Invent 2022: AWS Local Zones: creación de aplicaciones para una periferia distribuida)
- AWS re:Invent 2021 - Building low-latency websites with Amazon CloudFront (AWS re:Invent 2021: Creación de sitios web de baja latencia con Amazon CloudFront)
- AWS re:Invent 2022 - Improve performance and availability with AWS Global Accelerator (AWS re:Invent 2022: Mejorar el rendimiento y la disponibilidad con AWS Global Accelerator)
- AWS re:Invent 2022 - Build your global wide area network using AWS (AWS re:Invent 2022: Construya su red mundial de área extensa con AWS)
- AWS re:Invent 2020: Global traffic management with Amazon Route 53 (AWS re:Invent 2020: Administración de tráfico global con Amazon Route 53)

Ejemplos relacionados:

- AWS Global Accelerator Custom Routing Workshop
- Handling Rewrites and Redirects using Edge Functions (Gestión de reescrituras y redireccionamientos mediante funciones periféricas)

PERF04-BP07 Optimizar la configuración de red según las métricas

Utilice los datos recogidos y analizados para tomar decisiones informadas sobre la optimización de la configuración de su red.

Patrones comunes de uso no recomendados:

- Supone que todos los problemas de rendimiento están relacionados con las aplicaciones.
- Solo hace pruebas del rendimiento de la red desde una ubicación cercana al punto de implementación de la carga de trabajo.

- Se utilizan configuraciones predeterminadas para todos los servicios de red.
- Se sobreaprovisionan los recursos de red para proporcionar capacidad suficiente.

Beneficios de establecer esta práctica recomendada: la recopilación de las métricas necesarias de su red de AWS y la implementación de herramientas de supervisión de red le permiten comprender el rendimiento de la red y optimizar las configuraciones de la red.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Bajo

Guía para la implementación

La supervisión del tráfico hacia y desde VPC, subredes o interfaces de red es crucial para comprender cómo utilizar los recursos de red de AWS y optimizar las configuraciones de la red. Mediante las siguientes herramientas de red de AWS, puede inspeccionar más a fondo la información sobre el uso del tráfico, el acceso a la red y los registros.

Pasos para la implementación

- Identifique las métricas clave de rendimiento, como la latencia o la pérdida de paquetes, que desee recopilar. AWS proporciona varias herramientas que pueden ayudarle a recopilar estas métricas. Mediante las siguientes herramientas, puede inspeccionar más a fondo la información sobre el uso del tráfico, el acceso a la red y los registros:

Herramienta de AWS	Dónde se usa
Amazon VPC IP Address Manager.	Utilice IPAM para planificar, seguir y supervisar las direcciones IP para sus cargas de trabajo de AWS y locales. Esta es una práctica recomendada para optimizar el uso y la asignación de direcciones IP.
Registros de flujo de VPC	Utilice los registros de flujo de la VPC para capturar información detallada sobre el tráfico hacia y desde las interfaces de red en sus VPC. Con los registros de flujo de la VPC, puede diagnosticar reglas de grupos de seguridad excesivamente restrictivas o

Herramienta de AWS	Dónde se usa
	permisivas y determinar la dirección del tráfico hacia y desde las interfaces de red.
Registros de flujo de AWS Transit Gateway	Utilice los registros de AWS Transit Gateway flujo para recoger información sobre el tráfico IP que entra y sale de sus puertas de enlace de tránsito.
Registro de consultas de DNS	Registre información sobre las consultas de DNS públicas o privadas que recibe Route 53. Con los registros de DNS, puede optimizar las configuraciones de DNS al conocer el dominio o subdominio que se solicitó o las ubicacion es periféricas de Route 53 que respondieron a las consultas de DNS.
Reachability Analyzer	Reachability Analyzer le ayuda a analizar y depurar la accesibilidad de la red. Reachabil ity Analyzer es una herramienta de análisis de configuración que le permite realizar pruebas de conectividad entre un recurso de origen y un recurso de destino en sus VPC. Esta herramienta le ayuda a verificar que la configuración de su red coincida con la conectividad prevista.
Network Access Analyzer	Network Access Analyzer le ayuda a comprender el acceso a la red a sus recursos. Puede utilizar el Network Access Analyzer para especificar los requisitos de acceso a la red e identificar posibles rutas de red que no cumplan los requisitos especificados. Al optimizar su configuración de red correspon diente, puede comprender y verificar el estado de su red y demostrar si su red en AWS cumple con sus requisitos de conformidad.

Herramienta de AWS	Dónde se usa
Amazon CloudWatch	Utilice la configuración de Amazon CloudWatch y habilite las métricas adecuadas para las opciones de red. Asegúrese de elegir la métrica de red adecuada para su carga de trabajo. Por ejemplo, puede activar métricas para el uso de direcciones de red VPC, la puerta de enlace NAT de VPC, AWS Transit Gateway, túneles de VPN, AWS Network Firewall, Elastic Load Balancing y AWS Direct Connect. La supervisión continua de las métricas es una práctica recomendada para observar y comprender el estado y el uso de su red, lo que le ayuda a optimizar la configuración de la red basándose en sus observaciones.
AWS Network Manager	Con AWS Network Manager, puede supervisar el rendimiento histórico y en tiempo real de la red global de AWS con fines operativos y de planificación. Network Manager proporciona una latencia de red agregada entre las Regiones de AWS y las zonas de disponibilidad y dentro de cada zona de disponibilidad, lo que le permite comprender mejor la relación entre el rendimiento de las aplicaciones y el rendimiento de la red de AWS subyacente.
Amazon CloudWatch RUM	Use Amazon CloudWatch RUM para recopilar las métricas que le proporcionan información que le ayuda a identificar, comprender y mejorar la experiencia del usuario.

- Identifique los principales interlocutores y patrones de tráfico de las aplicaciones mediante VPC y AWS Transit Gateway Flow Logs.

- Evalúe y optimice su arquitectura de red actual, incluidas las VPC, las subredes y el enrutamiento. Por ejemplo, puede evaluar cómo diferentes emparejamientos de VPC o AWS Transit Gateway pueden ayudarle a mejorar las redes de su arquitectura.
- Evalúe las rutas de enrutamiento de su red para verificar que siempre se utilice la ruta más corta entre los destinos. Network Access Analyzer puede ayudarle a hacerlo.

Recursos

Documentos relacionados:

- Habilite los registros de consultas de DNS públicos.
- What is IPAM?
- What is Reachability Analyzer?
- What is Network Access Analyzer?
- Métricas de CloudWatch para sus VPC
- Optimize performance and reduce costs for network analytics with VPC Flow Logs in Apache Parquet format (Optimice el rendimiento y reduzca los costes de los análisis de red con los registros de flujo de VPC en formato Apache Parquet)
- Monitoring your global and core networks with Amazon CloudWatch metrics
- Continuously monitor network traffic and resources (Supervisar de forma continua el tráfico y los recursos de red)

Vídeos relacionados:

- AWS re:Invent 2023 – A developer's guide to cloud networking
- AWS re:Invent 2023 – Ready for what’s next? Designing networks for growth and flexibility
- AWS re:Invent 2023 – Advanced VPC designs and new capabilities
- AWS re:Invent 2022 – Dive deep on AWS networking infrastructure
- AWS re:Invent 2020 – Networking best practices and tips with the AWS Well-Architected Framework
- AWS re:Invent 2020 – Monitoring and troubleshooting network traffic

Ejemplos relacionados:

- AWS Networking Workshops (Talleres de red de AWS)
- AWS Network Monitoring
- Observing and diagnosing your network on AWS
- Finding and addressing network misconfigurations on AWS

Proceso y cultura

Al diseñar cargas de trabajo, hay principios y prácticas que puede adoptar para ayudarle a ejecutar mejor cargas de trabajo en la nube eficientes y de alto rendimiento. Esta área de enfoque ofrece las prácticas recomendadas para ayudarle adoptar una cultura que fomente la eficiencia del rendimiento de las cargas de trabajo en la nube.

Tenga en cuenta estos principios clave para crear esta cultura:

- Infraestructura como código: defina su infraestructura como código mediante enfoques como las plantillas de AWS CloudFormation. El uso de plantillas le permite colocar su infraestructura en un control fuente junto con su código de aplicación y configuraciones. Esto le permite aplicar las mismas prácticas que utiliza para desarrollar software en su infraestructura con la finalidad de que pueda iterar rápidamente.
- Canalización de despliegue: utilice una canalización de integración continua o de despliegue continuo (CI/CD), como por ejemplo, el repositorio del código fuente, los sistemas de diseño, el despliegue y la automatización de pruebas, para desplegar su infraestructura. Esto le permite desplegar de manera repetible, coherente y por un bajo coste mientras itera.
- Métricas bien definidas: configure y supervise las métricas para recoger indicadores clave de rendimiento (KPI). Recomendamos que utilice tanto métricas técnicas, como comerciales. Para aplicaciones móviles o sitios web, las métricas clave registran el tiempo para el primer byte o la renderización. Otras métricas que generalmente se aplican incluyen el recuento de subprocesos, la tasa de recolección de basura y los estados de espera. Las métricas comerciales, como el costo acumulado agregado por solicitud, puede alertarle sobre formas de reducir costos. Considere con cuidado cómo planifica interpretar las métricas. Por ejemplo, podría elegir el percentil máximo o el 99.º, en vez del promedio.
- Prueba de rendimiento automática: como parte de su proceso de despliegue, inicie automáticamente las pruebas de rendimiento después de que las pruebas de ejecución más rápida se hayan superado con éxito. La automatización debería crear un nuevo entorno, establecer condiciones iniciales como datos de prueba y luego ejecutar una serie de puntos de referencia y pruebas de carga. Los resultados de estas pruebas deberían estar vinculados al diseño, para que pueda seguir los cambios del rendimiento en el tiempo. Para las pruebas de larga ejecución, puede hacer que esta parte de la canalización sea asíncrona al resto del diseño. Alternativamente, podría ejecutar las pruebas de rendimiento durante la noche con instancias de spot de Amazon EC2.

- Generación de cargas: debe crear una serie de scripts de prueba que repliquen trayectos de usuario sintéticos o pregrabados. Estos scripts deben ser idempotentes y no acoplados, y podría necesitar incluir scripts de precalentamiento para obtener resultados válidos. En la medida de lo posible, sus scripts de prueba deben replicar el comportamiento de uso en la producción. Puede utilizar soluciones de software o de software como servicio (SaaS) para generar la carga. Piense en la posibilidad de usar soluciones de AWS Marketplace e instancias de spot, que pueden ser formas rentables de generar la carga.
- Visibilidad de rendimiento: las métricas clave deben ser visibles para su equipo, especialmente las métricas para cada versión de diseño. Esto le permite ver cualquier tendencia significativa, sea positiva o negativa, con el paso del tiempo. También debería exponer métricas en la cantidad de errores o excepciones para garantizar que está poniendo a prueba un sistema de trabajo.
- Visualización: utilice técnicas de visualización que dejen claro dónde se presentan problemas de rendimiento, puntos críticos, estados de espera o un uso bajo. Superponga las métricas de rendimiento sobre los diagramas de arquitectura: los gráficos de llamadas o el código pueden ayudar a identificar problemas con mayor rapidez.
- Proceso de revisión periódico: el mal funcionamiento de las arquitecturas suele ser el resultado de un proceso de revisión del rendimiento inexistente o deficiente. Si su arquitectura tiene un bajo rendimiento, la implementación de un proceso de revisión del rendimiento le permitirá impulsar la mejora iterativa.
- Optimización continua: adopte una cultura que optimice continuamente la eficiencia del rendimiento de su carga de trabajo en la nube.

Prácticas recomendadas

- PERF05-BP01 Establecer indicadores clave de rendimiento (KPI) para medir el estado y el rendimiento de la carga de trabajo
- PERF05-BP02 Utilizar soluciones de supervisión para saber en qué áreas es más crítico el rendimiento
- PERF05-BP03 Definir un proceso para mejorar el rendimiento de la carga de trabajo
- PERF05-BP04 Realizar pruebas de la carga de trabajo
- PERF05-BP05 Utilizar la automatización para solucionar de forma proactiva los problemas relacionados con el rendimiento
- PERF05-BP06 Mantener la carga de trabajo y los servicios actualizados
- PERF05-BP07 Revisar las métricas a intervalos regulares

PERF05-BP01 Establecer indicadores clave de rendimiento (KPI) para medir el estado y el rendimiento de la carga de trabajo

Identifique los KPI que miden de forma cuantitativa y cualitativa el rendimiento de la carga de trabajo. Los KPI ayudan a medir el estado y el rendimiento de una carga de trabajo en relación con un objetivo empresarial.

Antipatrones usuales:

- Supervisa únicamente las métricas del nivel del sistema para obtener información sobre su carga de trabajo sin comprender el impacto empresarial de dichas métricas.
- Presupone que los KPI ya se están publicando y compartiendo como datos de métricas estándar.
- No tiene definido un KPI cuantitativo y medible.
- Los KPI no se corresponden con los objetivos o estrategias empresariales.

Ventajas de aplicar esta práctica recomendada: identificar los KPI específicos que representan el estado y el rendimiento de la carga de trabajo ayuda a alinear a los equipos con sus prioridades y a definir resultados empresariales correctos. Al compartir estas métricas con todos los departamentos, se obtiene información y se fomenta un enfoque coherente en relación con los umbrales, las expectativas y las repercusiones empresariales.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: alto

Guía para la implementación

Los KPI ayudan a las empresas y a los equipos de ingeniería a organizarse en función de la medición de los objetivos y estrategias, y del modo en que estos factores se combinan para producir resultados empresariales. Por ejemplo, en una carga de trabajo de un sitio web, el tiempo de carga de la página podría usarse como indicativo del rendimiento general. Esta métrica sería uno de los múltiples puntos de datos que miden la experiencia del usuario. Además de identificar los umbrales de los tiempos de carga de la página, debería documentar el resultado previsto o el riesgo empresarial si no se cumple el ideal de rendimiento. Si una página tarda en cargarse, los usuarios finales se ven directamente afectados, se reduce su valoración de la experiencia y se pueden perder clientes. Cuando defina los umbrales de KPI, combine tanto las referencias del sector como las expectativas de los usuarios finales. Por ejemplo, si la referencia sectorial actual es que una página web se cargue en dos segundos, pero los usuarios esperan que tarde solamente un segundo, debería tener en cuenta estos dos puntos de datos al establecer el KPI.

El equipo debe evaluar los KPI de su carga de trabajo utilizando datos detallados en tiempo real y datos históricos como referencia, y crear paneles en los que se realicen cálculos de métricas sobre los datos de los KPI para obtener información sobre las operaciones y la utilización. Los KPI deben documentarse e incluir umbrales que respalden los objetivos y las estrategias de la empresa, además de asignarse a las métricas que se estén supervisando. Los KPI deberían revisarse siempre que cambien los objetivos empresariales, las estrategias o los requisitos del usuario final.

Pasos para la implementación

- Identificar a las partes interesadas: identifique a las principales partes interesadas del negocio, incluidos los equipos de desarrollo y operaciones, y documéntelas.
- Definir los objetivos: colabore con estas partes interesadas para definir y documentar los objetivos de la carga de trabajo. Tenga en cuenta los aspectos esenciales de desempeño de las cargas de trabajo, como, por ejemplo, el rendimiento, el tiempo de respuesta y el coste, así como los objetivos empresariales, como, por ejemplo, la satisfacción del usuario.
- Revisar las prácticas recomendadas del sector: revise las prácticas recomendadas del sector para identificar los KPI pertinentes que se ajustan a los objetivos de la carga de trabajo.
- Identificar las métricas: identifique las métricas que se ajusten a los objetivos de la carga de trabajo y que puedan ayudarle a medir el rendimiento y los objetivos empresariales. Establezca los KPI en función de estas métricas. Las métricas de ejemplo son medidas como el tiempo promedio de respuesta o el número de usuarios simultáneos.
- Definir y documentar los KPI: utilice las prácticas recomendadas del sector y los objetivos de la carga de trabajo para establecer los objetivos del KPI de la carga de trabajo. Utilice esta información para establecer los umbrales de gravedad o el nivel de alarma de los KPI. Identifique y documente el riesgo y el impacto del incumplimiento de los KPI.
- Implementar la supervisión: utilice herramientas de supervisión, como, por ejemplo, Amazon CloudWatch o AWS Config para recopilar métricas y medir los KPI.
- Comunicar visualmente los KPI: utilice herramientas de panel, como, por ejemplo, Amazon QuickSight para visualizar los KPI y comunicárselos a las partes interesadas.
- Analizar y optimizar: revise y analice periódicamente los KPI para identificar las áreas de la carga de trabajo que deben mejorarse. Colabore con las partes interesadas para implementar estas mejoras.
- Revisar y perfeccionar: revise periódicamente las métricas y los KPI para evaluar su eficacia, especialmente cuando cambien los objetivos empresariales o el rendimiento de la carga de trabajo.

Recursos

Documentos relacionados:

- CloudWatch documentation
- AWS Partner con competencias en supervisión, registro y rendimiento
- AWS observability tools
- The Importance of Key Performance Indicators (KPIs) for Large-Scale Cloud Migrations
- How to track your cost optimization KPIs with the KPI Dashboard
- X-Ray Documentation
- Using Amazon CloudWatch dashboards
- Amazon QuickSight KPIs

Vídeos relacionados:

- AWS re:Invent 2023 - Optimize cost and performance and track progress toward mitigation
- AWS re:Invent 2023 - Manage resource lifecycle events at scale with AWS Health
- AWS re:Invent 2023 - Performance & efficiency at Pinterest: Optimizing the latest instances
- AWS re:Invent 2022 - AWS optimization: Actionable steps for immediate results
- AWS re:Invent 2023 - Building an effective observability strategy
- AWS Summit SF 2022 - Full-stack observability and application monitoring with AWS
- AWS re:Invent 2023 - Scaling on AWS for the first 10 million users
- AWS re:Invent 2022 - How Amazon uses better metrics for improved website performance
- Creating an Effective Metrics Strategy for Your Business | AWS Events

Ejemplos relacionados:

- Creación de un panel con Amazon QuickSight

PERF05-BP02 Utilizar soluciones de supervisión para saber en qué áreas es más crítico el rendimiento

Comprenda y detecte las áreas en las que un aumento del rendimiento de la carga de trabajo tendrá un impacto positivo en la eficiencia o en la experiencia del cliente. Por ejemplo, un sitio web que tenga una gran interacción del cliente se beneficiaría de utilizar servicios en la periferia para acercar la entrega de contenido a los clientes.

Antipatrones usuales:

- Supone que las métricas de computación estándares como el uso de CPU o la presión sobre la memoria son suficientes para detectar problemas de rendimiento.
- Solo se utilizan las métricas predeterminadas registradas por el software de supervisión seleccionado.
- Solo se revisan las métricas cuando hay un problema.

Ventajas de aplicar esta práctica recomendada: conocer las áreas críticas del rendimiento ayuda a los propietarios de las cargas de trabajo a supervisar los KPI y a priorizar las mejoras de mayor impacto.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: alto

Guía para la implementación

Configure el seguimiento de extremo a extremo para identificar los patrones de tráfico, la latencia y las áreas esenciales de rendimiento. Supervise los patrones de acceso a los datos para detectar consultas lentas o datos deficientemente fragmentados y particionados. Identifique las áreas restringidas de la carga de trabajo mediante pruebas de carga o supervisión.

aumentar la eficiencia del rendimiento mediante la comprensión de su arquitectura, patrones de tráfico y patrones de acceso a los datos e identificar sus tiempos de latencia y procesamiento. Identifique los posibles cuellos de botella que puedan afectar a la experiencia del cliente a medida que aumenta la carga de trabajo. Al identificar esas áreas, fíjese en qué solución podría desplegar para acabar con los problemas de rendimiento.

Pasos para la implementación

- Configure la supervisión de extremo a extremo para capturar todos los componentes y métricas de la carga de trabajo. A continuación, se muestran algunos ejemplos de soluciones de supervisión de AWS.

Service	Where to use
Amazon CloudWatch Real-User Monitoring (RUM)	To capture application performance metrics from real user client-side and frontend sessions.
AWS X-Ray	To trace traffic through the application layers and identify latency between components and dependencies. Use X-Ray service maps to see relationships and latency between workload components.
Información sobre el rendimiento de Amazon Relational Database Service	To view database performance metrics and identify performance improvements.
Amazon RDS Enhanced Monitoring	To view database OS performance metrics.
Amazon DevOps Guru	To detect abnormal operating patterns so you can identify operational issues before they impact your customers.

- Lleve a cabo pruebas para generar métricas, identificar patrones de tráfico, cuellos de botella y áreas críticas de rendimiento. Estos son algunos ejemplos de cómo se realizan las pruebas:
  - Configure valores controlados sintéticos de CloudWatch para imitar las actividades de los usuarios en el navegador mediante programación utilizando expresiones de velocidad o tareas cron de Linux y generar métricas coherentes a lo largo del tiempo.
  - Utilice la solución Pruebas de carga distribuidas en AWS para generar picos de tráfico o probar la carga de trabajo al ritmo de crecimiento esperado.
- Evalúe las métricas y la telemetría para identificar sus áreas fundamentales de rendimiento. Revise estas áreas con su equipo con el fin de analizar la supervisión y las soluciones para evitar los cuellos de botella.

- Experimente con las mejoras de rendimiento y mida los cambios con datos. Por ejemplo, puede usar CloudWatch Evidently para probar las nuevas mejoras y el impacto del rendimiento en su carga de trabajo.

Recursos

Documentos relacionados:

- What's new in AWS Observability at re:Invent 2023
- Amazon Builders' Library
- X-Ray Documentation
- Amazon CloudWatch RUM
- Amazon DevOps Guru

Vídeos relacionados:

- AWS re:Invent 2023 - [LAUNCH] Application monitoring for modern workloads
- AWS re:Invent 2023 - Implementing application observability
- AWS re:Invent 2023 - Building an effective observability strategy
- AWS Summit SF 2022 - Full-stack observability and application monitoring with AWS
- AWS re:Invent 2022 - AWS optimization: Actionable steps for immediate results
- AWS re:Invent 2022 - The Amazon Builders’ Library: 25 years of Amazon operational excellence
- AWS re:Invent 2022 - How Amazon uses better metrics for improved website performance
- Visual Monitoring of Applications with Amazon CloudWatch Synthetics

Ejemplos relacionados:

- Measure page load time with Amazon CloudWatch Synthetics
- Amazon CloudWatch RUM Web Client
- X-Ray SDK for Python
- Pruebas de carga distribuidas en AWS

PERF05-BP03 Definir un proceso para mejorar el rendimiento de la carga de trabajo

Definir un proceso para evaluar nuevos servicios, patrones de diseño, tipos de recursos y configuraciones a medida que estén disponibles. Por ejemplo, ejecute las pruebas de rendimiento existentes en las nuevas ofertas de instancias a fin de determinar su capacidad para mejorar su carga de trabajo.

Antipatrones usuales:

- Presupone que la arquitectura actual es estática y no se va a actualizar con el tiempo.
- Incorpora cambios en la arquitectura a lo largo del tiempo sin justificación de métricas.

Ventajas de aplicar esta práctica recomendada: al definir un proceso para realizar cambios en la arquitectura, puede utilizar los datos recopilados para influir en el diseño de la carga de trabajo a lo largo del tiempo.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: medio

Guía para la implementación

El rendimiento de su carga de trabajo tiene algunas limitaciones clave. Documéntelos para que sepa qué tipos de innovación pueden mejorar el rendimiento de su carga de trabajo. Utilice esta información cuando conozca nuevos servicios o tecnologías a medida que estén disponibles para identificar formas de mitigar las limitaciones o cuellos de botella.

Identifique las principales limitaciones en el rendimiento de su carga de trabajo Documente las limitaciones de rendimiento de la carga de trabajo para que sepa qué tipos de innovación pueden mejorar el rendimiento de la carga de trabajo.

Pasos para la implementación

- Identificar los KPI: identifique los KPI de rendimiento de la carga de trabajo, tal y como se describe en PERF05-BP01 Establecer indicadores clave de rendimiento (KPI) para medir el estado y el rendimiento de la carga de trabajo para establecer los puntos de referencia de dicha carga.
- Implementar la supervisión: utilice AWS observability tools para recopilar métricas de rendimiento y medir los KPI.

- Llevar a cabo análisis: realice un análisis exhaustivo para identificar las áreas de la carga de trabajo (como la configuración y el código de la aplicación) que tienen un rendimiento inferior, tal y como se describe en PERF05-BP02 Utilizar soluciones de supervisión para saber en qué áreas es más crítico el rendimiento. Utilice sus herramientas de análisis y rendimiento para identificar las estrategias de mejora del rendimiento.
- Validar las mejoras: utilice entornos de pruebas o de preproducción para validar la eficacia de las estrategias de mejora.
- Implementar los cambios: implemente los cambios en la producción y supervise continuamente el rendimiento de la carga de trabajo. Documente las mejoras y comunique los cambios a las partes interesadas.
- Revisar y perfeccionar: revise periódicamente su proceso de mejora del rendimiento para identificar las áreas que deben mejorarse.

Recursos

Documentos relacionados:

- Blog de AWS
- Novedades de AWS
- AWS Skill Builder

Vídeos relacionados:

- AWS re:Invent 2022 - Delivering sustainable, high-performing architectures
- AWS re:Invent 2023 - Optimize cost and performance and track progress toward mitigation
- AWS re:Invent 2022 - AWS optimization: Actionable steps for immediate results
- AWS re:Invent 2022 - Optimize your AWS workloads with best-practice guidance

Ejemplos relacionados:

- Github de AWS

PERF05-BP04 Realizar pruebas de la carga de trabajo

Realice una prueba de carga en su carga de trabajo para comprobar que puede gestionar la carga de producción e identificar cualquier cuello de botella en el rendimiento.

Antipatrones usuales:

- Realiza pruebas de carga de partes individuales de su carga de trabajo, pero no de la carga completa.
- Realiza pruebas de carga en una infraestructura que no es la misma que su entorno de producción.
- Solo realiza pruebas de carga hasta su carga prevista y no más allá, para ayudar a prever dónde puede tener problemas en el futuro.
- Realiza pruebas de carga sin consultar la política de pruebas de Amazon EC2 ni presentar un formulario de envío de eventos simulados. Esto hace que la prueba no se ejecute, ya que parece un evento de denegación de servicio.

Ventajas de aplicar esta práctica recomendada: al medir el rendimiento en una prueba de carga, podrá ver qué áreas se van a ver afectadas cuando aumente la carga. De este modo, podrá anticipar los cambios necesarios antes de que afecten a la carga de trabajo.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: bajo

Guía para la implementación

Las pruebas de carga en la nube es un proceso que permite medir el rendimiento de la carga de trabajo en la nube bajo condiciones realistas, con la carga de usuarios esperada. Este proceso implica el aprovisionamiento de un entorno de nube similar al de producción, el uso de herramientas de pruebas de carga para generar la carga y el análisis de métricas para evaluar la capacidad de la carga de trabajo a la hora de gestionar una carga realista. Las pruebas de carga deben ejecutarse con versiones sintéticas o saneadas de los datos de producción (debe eliminarse la información confidencial o de identificación). Realice automáticamente pruebas de carga en la canalización de entrega y compare los resultados con los KPI y los umbrales predefinidos. Este proceso le ayudará a seguir alcanzando el rendimiento requerido.

Pasos para la implementación

- Definir los objetivos de prueba: identifique los aspectos de rendimiento de la carga de trabajo que desea evaluar, como, por ejemplo, el rendimiento y el tiempo de respuesta.
- Seleccionar una herramienta de prueba: elija y configure la herramienta de prueba de carga que se ajuste a su carga de trabajo.
- Configurar el entorno: configure el entorno de prueba en función del entorno de producción. Puede usar los servicios de AWS para ejecutar entornos a escala de producción y poner a prueba su arquitectura.
- Implementar la supervisión: utilice herramientas de supervisión, como, por ejemplo, Amazon CloudWatch, para recopilar métricas de todos los recursos de la arquitectura. También puede recopilar y publicar métricas personalizadas.
- Definir escenarios: defina los escenarios y los parámetros de las pruebas de carga (como la duración de la prueba y el número de usuarios).
- Realizar pruebas de carga: cree escenarios de prueba a escala. Utilice la Nube de AWS para probar la carga de trabajo y detectar las áreas en las que el escalamiento no se realiza correctamente o no se produce de forma lineal. Por ejemplo, utilice Spot Instances para generar cargas a bajo costo y descubrir obstáculos antes que se experimenten en la producción
- Analizar los resultados de las pruebas: analice los resultados para identificar los cuellos de botella del rendimiento y las áreas en las que se pueden mejorar.
- Documentar y compartir los resultados: documente e informe sobre los resultados y recomendaciones. Comparta esta información con las partes interesadas para que puedan tomar decisiones fundamentadas con respecto a las estrategias de optimización del rendimiento.
- Iterar continuamente: las pruebas de carga deben realizarse con regularidad, especialmente después de un cambio en el sistema realizado por una actualización.

Recursos

Documentos relacionados:

- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Pruebas de carga distribuidas en AWS

Vídeos relacionados:

- AWS Summit ANZ 2023: Accelerate with confidence through AWS Distributed Load Testing
- AWS re:Invent 2022 - Scaling on AWS for your first 10 million users
- Solving with AWS Solutions: Distributed Load Testing

- AWS re:Invent 2021 - Optimize applications through end user insights with Amazon CloudWatch RUM
- Demostración de Amazon CloudWatch Synthetics

Ejemplos relacionados:

- Pruebas de carga distribuidas en AWS

PERF05-BP05 Utilizar la automatización para solucionar de forma proactiva los problemas relacionados con el rendimiento

Utilice los indicadores clave de rendimiento (KPI), junto con los sistemas de supervisión y alerta, para abordar de manera proactiva los problemas relacionados con el rendimiento.

Antipatrones usuales:

- Únicamente permite que el personal de operaciones pueda llevar a cabo cambios operativos en la carga de trabajo.
- Permite que todas las alarmas se filtren al equipo de operaciones sin medidas de corrección proactivas.

Ventajas de aplicar esta práctica recomendada: al solucionar de forma proactiva las acciones de alarma, al personal de soporte podrá concentrarse en aquellos elementos que no pueden abordarse de forma automática. De este modo, el personal de operaciones podrá gestionar todas las alarmas sin sentirse abrumado y concentrarse exclusivamente en las alarmas críticas.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: bajo

Guía para la implementación

Usa alarmas para activar acciones automatizadas y corregir los problemas siempre que sea posible. Escala la alarma a aquellos capaces de responder cuando no se pueda recurrir a la respuesta automatizada. Por ejemplo, podría tener un sistema capaz de predecir los valores esperados de los indicadores clave de rendimiento (KPI) y emitir alarmas cuando se sobrepasen ciertos umbrales, o una herramienta que pudiera detener o revertir automáticamente los despliegues si los KPI están fuera de los valores esperados.

Implementar procesos que proporcionen visibilidad del rendimiento a medida que ejecuta la carga de trabajo. Cree paneles de supervisión y establezca normas de referencia sobre las expectativas del rendimiento para determinar si la carga de trabajo funciona de manera óptima.

Pasos para la implementación

- Identificar el flujo de trabajo de corrección: identifique y estudie si el problema de rendimiento puede solucionarse automáticamente. Use soluciones de supervisión de AWS, como Amazon CloudWatch o AWS X-Ray, que le ayuden a comprender mejor la causa principal del problema.
- Definir el proceso de automatización: cree un proceso de corrección paso a paso que pueda usarse para solucionar el problema automáticamente.
- Configurar el evento de inicio: configure el evento para iniciar automáticamente el proceso de corrección. Por ejemplo, puede definir un activador que reinicie automáticamente una instancia cuando se alcance un determinado umbral de uso de la CPU.
- Automatizar la corrección: utilice los servicios y las tecnologías de AWS para automatizar el proceso de corrección. Por ejemplo, la Automatización de AWS Systems Manager proporciona un mecanismo seguro y escalable para automatizar el proceso de corrección. Asegúrese de usar la lógica de autorrecuperación para revertir los cambios si el problema no se soluciona correctamente.
- Probar el flujo de trabajo: pruebe el proceso de corrección automatizado en un entorno de preproducción.
- Implementar el flujo de trabajo: implemente la corrección automatizada en el entorno de producción.
- Desarrollar una guía de estrategias: desarrolle y documente una guía de estrategias que describa los pasos del plan de corrección, incluidos los eventos de inicio, la lógica de corrección y las medidas adoptadas. Asegúrese de que las partes interesadas reciban formación para que puedan responder de manera eficaz a los eventos de corrección automatizada.

- Revisar y perfeccionar: evalúe periódicamente la eficacia del flujo de trabajo de corrección automatizada. Ajuste los eventos de inicio y la lógica de corrección si es necesario.

Recursos

Documentos relacionados:

- CloudWatch Documentation
- Socios de AWS Partner Network con competencias en supervisión, registro y rendimiento
- X-Ray Documentation
- Using Alarms and Alarm Actions in CloudWatch
- Build a Cloud Automation Practice for Operational Excellence: Best Practices from AWS Managed Services
- Automate your Amazon Redshift performance tuning with automatic table optimization

Vídeos relacionados:

- AWS re:Invent 2023 - Strategies for automated scaling, remediation, and smart self-healing
- AWS re:Invent 2023 - [LAUNCH] Application monitoring for modern workloads
- AWS re:Invent 2023 - Implementing application observability
- AWS re:Invent 2021 - Intelligently automating cloud operations
- AWS re:Invent 2022 - Setting up controls at scale in your AWS environment
- AWS re:Invent 2022 - Automating patch management and compliance using AWS
- AWS re:Invent 2022 - How Amazon uses better metrics for improved website performance
- AWS re:Invent 2023 - Take a load off: Diagnose & resolve performance issues with Amazon RDS
- AWS re:Invent 2021 -{New Launch} Automatically detect and resolve issues with Amazon DevOps Guru
- AWS re:Invent 2023 - Centralize your operations

Ejemplos relacionados:

- CloudWatch Logs Customize Alarms

PERF05-BP06 Mantener la carga de trabajo y los servicios actualizados

Manténgase al tanto de los nuevos servicios y características de la nube para adoptar características eficientes, resolver problemas y mejorar la eficiencia general del rendimiento de la carga de trabajo.

Antipatrones usuales:

- Asume que su arquitectura actual es estática y no se actualizará con el tiempo.
- No dispone de sistemas ni de una cadencia regular para evaluar si los programas y paquetes actualizados son compatibles con su carga de trabajo.

Ventaja de aplicar esta práctica recomendada: al establecer un proceso que le permita estar al tanto de los nuevos servicios y ofertas, puede adoptar nuevas características y funcionalidades, resolver problemas y mejorar el rendimiento de la carga de trabajo.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: bajo

Guía para la implementación

Evalúe mecanismos para mejorar el rendimiento a medida que disponga de nuevos servicios, patrones de diseño y características de productos. Determine cuáles de ellas podrían mejorar el rendimiento o aumentar la eficiencia de la carga de trabajo mediante una evaluación, un debate interno o un análisis externo. Defina un proceso para evaluar las actualizaciones, las nuevas características y servicios pertinentes para su carga de trabajo. Por ejemplo, cree una prueba de concepto que utilice nuevas tecnologías o consulte a un grupo interno. Cuando pruebe nuevas ideas o servicios, realice pruebas de rendimiento para medir el impacto que tienen en el rendimiento de la carga de trabajo.

Pasos para la implementación

- Realizar un inventario de la carga de trabajo: realice un inventario del software y la arquitectura de su carga de trabajo e identifique los componentes que deben actualizarse.
- Identificar los orígenes de actualización: identifique las noticias y los orígenes de actualización relacionados con los componentes de su carga de trabajo. Por ejemplo, puede suscribirse al blog Novedades de AWS para examinar los productos que se ajustan al componente de la carga de trabajo. Puede suscribirse a la fuente RSS o administrar las suscripciones de correo electrónico.

- Definir un calendario de actualización: establezca un calendario para evaluar nuevos servicios y características en la carga de trabajo.
  - Puede usar AWS Systems Manager Inventory para recopilar metadatos de las instancias, las aplicaciones y el sistema operativo (SO) en las instancias de Amazon EC2 y saber rápidamente qué instancias están ejecutando el software y las configuraciones requeridas por la política de software y qué instancias deben actualizarse.
- Evaluar la nueva actualización: entienda cómo actualizar los componentes de la carga de trabajo. Aproveche la agilidad de la nube para probar rápidamente cómo las nuevas características pueden mejorar la eficiencia del rendimiento de la carga de trabajo.
- Usar la automatización: utilice la automatización del proceso de actualización para reducir el nivel de esfuerzo para desplegar nuevas características y limitar los errores que provocan los procesos manuales.
  - Puede utilizar CI/CD para actualizar automáticamente las AMI, las imágenes de contenedor y otros artefactos relacionados con su aplicación en la nube.
  - Puede utilizar herramientas como AWS Systems Manager Patch Manager para automatizar el proceso de las actualizaciones del sistema y programar la actividad con Ventanas de mantenimiento de AWS Systems Manager.
- Documentar el proceso: documente su proceso para evaluar las actualizaciones y los nuevos servicios. Proporcione a los propietarios el tiempo y el espacio necesarios para investigar, probar, experimentar y validar las actualizaciones y los nuevos servicios. Consulte los requisitos empresariales documentados y los KPI para ayudar a priorizar qué actualización tendrá un impacto empresarial positivo.

Recursos

Documentos relacionados:

- Blog de AWS
- Novedades de AWS
- Implementing up-to-date images with automated EC2 Image Builder pipelines

Vídeos relacionados:

- AWS re:Inforce 2022 - Automating patch management and compliance using AWS
- All Things Patch: AWS Systems Manager | AWS Events

Ejemplos relacionados:

- Inventory and Patch Management
- Taller sobre observabilidad

PERF05-BP07 Revisar las métricas a intervalos regulares

Revise qué métricas se están recopilando durante el mantenimiento rutinario o en respuesta a eventos o incidentes. Utilice estas revisiones para determinar qué métricas son esenciales para abordar los problemas y qué otras métricas, en caso de que se estén supervisando, podrían ayudar a identificar, abordar o prevenir problemas.

Antipatrones usuales:

- Permite que las métricas se mantengan en un estado de alarma durante un periodo de tiempo prolongado.
- Crea alarmas que no puede accionar un sistema de automatización.

Ventajas de aplicar esta práctica recomendada: revise continuamente las métricas que se recopilan para comprobar que identifican, abordan o previenen los problemas de manera adecuada. Las métricas también pueden quedarse obsoletas si deja que permanezcan en un estado de alarma durante mucho tiempo.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: medio

Guía para la implementación

Mejore continuamente la recopilación y la supervisión de métricas. Como parte de la respuesta a incidentes o sucesos, evalúe qué parámetros fueron útiles para abordar el problema y qué parámetros podrían haber ayudado a los que no se están controlando actualmente. Utilice este método para mejorar la calidad de las métricas que recopila, de modo que pueda prevenir o resolver incidentes en el futuro con mayor rapidez.

Como parte de la respuesta a incidentes o sucesos, evalúe qué parámetros fueron útiles para abordar el problema y qué parámetros podrían haber ayudado a los que no se están controlando actualmente. Utilícelo para mejorar la calidad de la métrica que recopila, de modo que pueda prevenir o resolver más rápidamente futuros incidentes.

Pasos para la implementación

- Definir las métricas: defina las métricas esenciales de rendimiento para supervisar que estén en consonancia con el objetivo de carga de trabajo, incluidas métricas como el tiempo de respuesta y la utilización de los recursos.
- Establecer bases de referencia: establezca una base de referencia y el valor que desee para cada métrica. La base de referencia debe proporcionar puntos de referencia para identificar desviaciones o anomalías.
- Establecer una cadencia: establezca una cadencia (como, por ejemplo, semanal o mensual) para revisar las métricas esenciales.
- Identificar los problemas de rendimiento: durante cada revisión, evalúe las tendencias y las desviaciones con respecto a los valores de la base de referencia. Busque cualquier cuello de botella o anomalía en el rendimiento. Lleve a cabo un análisis exhaustivo de la causa raíz de los problemas identificados para conocer qué los provoca.
- Identificar las medidas correctivas: utilice los análisis para identificar las medidas correctivas. Entre dichas medidas se pueden incluir el ajuste de parámetros, la corrección de errores y el escalamiento de los recursos.
- Documentar los resultados: documente los resultados, incluidos los problemas identificados, las causas principales y las medidas correctivas.
- Iterar y mejorar: evalúe y mejore continuamente el proceso de revisión de las métricas. Aplique lo que ha aprendido de la revisión anterior para mejorar el proceso con el tiempo.

Recursos

Documentos relacionados:

- CloudWatch Documentation
- Collect metrics and logs from Amazon EC2 Instances and on-premises servers with the CloudWatch Agent
- Consulte sus métricas con CloudWatch Metrics Insights
- Socios de AWS Partner Network con competencias en supervisión, registro y rendimiento
- X-Ray Documentation

Vídeos relacionados:

- AWS re:Invent 2022 - Setting up controls at scale in your AWS environment
- AWS re:Invent 2022 - How Amazon uses better metrics for improved website performance
- AWS re:Invent 2023 - Building an effective observability strategy
- AWS Summit SF 2022 - Full-stack observability and application monitoring with AWS
- AWS re:Invent 2023 - Take a load off: Diagnose & resolve performance issues with Amazon RDS

Ejemplos relacionados:

- Creación de un panel con Amazon QuickSight
- CloudWatch Dashboards

Conclusión

Lograr y mantener una eficiencia del rendimiento requiere un enfoque impulsado por los datos. Considere activamente patrones de acceso y compensaciones que le permitirán optimizar a fin de maximizar el rendimiento. El uso de un proceso de revisión basado en puntos de referencia y pruebas de carga permite seleccionar las configuraciones y tipos de recursos apropiados. Tratar su infraestructura como código permite evolucionar su arquitectura de forma rápida y segura, mientras usa los datos para tomar decisiones basadas en hechos sobre su arquitectura. Llevar a cabo una combinación de supervisión activa y pasiva le garantiza que el rendimiento de su arquitectura no se degrade con el tiempo.

AWS se esfuerza por ayudarle a diseñar arquitecturas que se ejecuten de forma eficiente y que ofrezcan valor empresarial. Utilice las herramientas y técnicas que se exponen en este documento para garantizar el éxito.

Colaboradores

Las siguientes personas y organizaciones han contribuido a este documento:

- Sam Mokhtari, Senior Efficiency Lead Solutions Architect, Amazon Web Services
- Josh Hart, Solutions Architect, Amazon Web Services
- Richard Trabing, Solutions Architect, Amazon Web Services
- Brett Looney, Principal Solutions Architect, Amazon Web Services
- Nina Vogl, Principal Solutions Architect, Amazon Web Services
- Eric Pullen, Solutions Architect, Amazon Web Services
- Julien Lépine, Specialist SA Manager, Amazon Web Services
- Ronnen Slasky, Solutions Architect, Amazon Web Services

Otra documentación

Para obtener más ayuda, consulte estas fuentes:

- AWS Well-Architected Framework
- Centro de arquitectura de AWS

Revisiones del documento

Para recibir notificaciones sobre las actualizaciones de este documento técnico, suscríbase al canal RSS.

Cambio	Descripción	Fecha
Actualizaciones de la guía sobre las prácticas recomendadas	Actualizaciones menores en todas las prácticas recomendadas.	June 27, 2024
Actualización y reestructuración importantes	El pilar se ha reestructurado para que incluya cinco áreas de prácticas recomendadas (en vez de ocho). El contenido se ha consolidado en las cinco áreas y se ha actualizado. Las nuevas áreas de prácticas recomendadas son Selección de la arquitectura, Computación y hardware, Administración de datos, Redes y entrega de contenidoy Proceso y cultura.	October 3, 2023
Actualización menor	Eliminación del lenguaje no inclusivo.	April 13, 2023
Actualizaciones del nuevo marco	Prácticas recomendadas actualizadas con guía prescriptiva y prácticas recomendadas añadidas.	April 10, 2023
Documento técnico actualizado	Prácticas recomendadas actualizadas con nueva guía de implementación.	December 15, 2022

Documento técnico actualizado	Se han ampliado las prácticas recomendadas y se han añadido planes de mejora.	October 20, 2022
Actualización menor	Se ha eliminado el lenguaje no inclusivo.	April 22, 2022
Actualización menor	Se ha añadido Pilar de sostenibilidad a la introducción.	December 2, 2021
Actualizaciones menores	Se han actualizado los enlaces.	March 10, 2021
Actualizaciones menores	Se ha cambiado el tiempo de espera de AWS Lambda a 900 segundos y se ha corregido el nombre de Amazon Keyspaces (for Apache Cassandra).	October 5, 2020
Actualización menor	Se ha corregido el enlace que no funciona.	July 15, 2020
Actualizaciones del nuevo marco	Revisión importante y actualización de contenidos	July 8, 2020
Documento técnico actualizado	Actualización menor por problemas gramaticales	July 1, 2018
Documento técnico actualizado	El documento técnico se ha actualizado para reflejar los cambios en AWS	November 1, 2017
Publicación inicial	Publicación de Pilar de eficiencia del rendimiento: AWS Well-Architected Framework.	November 1, 2016

AWS Glossary

For the latest AWS terminology, see the AWS glossary in the Glosario de AWS Reference.