AWS Well-Architected Framework

Performance Efficiency Pillar

Performance Efficiency Pillar: AWS Well-Architected Framework

Table of Contents

Pilar de eficiencia del rendimiento: Marco de AWS Well-Architected

Fecha de publicación: 6 de noviembre de 2024 (Revisiones del documento)

Este documento técnico se centra en el pilar de la eficiencia del rendimiento del Marco de AWS Well-Architected. Proporciona una guía para ayudar a los clientes a aplicar las prácticas recomendadas a la hora de diseñar, entregar y mantener los entornos de AWS.

Introducción

El Marco de AWS Well-Architected 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 Well-Architected aumenta en gran medida la probabilidad de éxito empresarial.

El marco se basa en seis pilares:

- Excelencia operativa
- Seguridad
- Fiabilidad
- Eficacia 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 en las instalaciones tradicionales. El uso de los principios de este documento le permitirá 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 eficiencia del rendimiento incluye la capacidad para utilizar los recursos de la nube de forma eficaz a fin de que satisfagan los requisitos de rendimiento y para mantener dicha eficacia a medida que la demanda cambia y las tecnologías evolucionan.

Temas

Principios de diseño

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

- Democratización de 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.
- Adopción de un enfoque global en cuestión de minutos: la implementación 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 costo mínimo.
- Uso de 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 costos de transacciones porque los servicios administrados operan a escala de la nube.
- Experimentación más frecuente: los recursos virtuales y automatizables permiten hacer pruebas comparativas con rapidez mediante diferentes tipos de instancias, almacenamiento y configuraciones.

- Consideración de 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 general hasta la selección y configuración de los tipos de recursos.

Revisar periódicamente sus opciones le permitirá asegurarse 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 de 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 Descubrimiento y comprensión de los servicios y las características disponibles en la nube
- PERF01-BP02 Uso de las recomendaciones del proveedor de servicios en la nube o de un socio adecuado para conocer los modelos de arquitectura y las prácticas recomendadas
- PERF01-BP03 Contemplación de los costos en las decisiones sobre arquitectura
- PERF01-BP04 Evaluación del efecto de las decisiones en los clientes y en la eficiencia de la arquitectura
- PERF01-BP05 Uso de políticas y arquitecturas de referencia
- PERF01-BP06 Uso de pruebas comparativas para tomar decisiones arquitectónicas
- PERF01-BP07 Uso de un enfoque basado en los datos en sus decisiones arquitectónicas

PERF01-BP01 Descubrimiento y comprensión de 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 costos 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 costo 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 ayudarlo a mejorar el rendimiento y a reducir los costos y la complejidad operativa.
  - Amazon Web Services Cloud
  - AWS Academy
  - Novedades de AWS
  - Blog de AWS
  - Skill Builder de AWS
  - Eventos y seminarios web de AWS

  - Formación de AWS and Certifications
  - Canal de YouTube de AWS
  - AWS Workshops
  - AWS Communities
- Use Amazon Q para obtener información y consejos pertinentes sobre los servicios.
- Utilice entornos de pruebas (que no sean de producción) para aprender y experimentar con los nuevos servicios sin incurrir en costos extraordinarios.
- Obtenga información continua sobre los nuevos servicios y características de la nube.

Recursos

Documentos relacionados:

- Overview of Amazon Web Services
- Características de Amazon EC2
- Aprenda paso a paso con un plan de aprendizaje para socios de AWS
- Capacitación y certificación de AWS
- My learning path to become an AWS solutions architect
- Centro de arquitectura de AWS
- AWS Partner Network
- Biblioteca de soluciones de AWS
- Centro de conocimiento de AWS
- Cree aplicaciones modernas en AWS

Videos 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 del AWS
- Ejemplos del AWS SDK

PERF01-BP02 Uso de las recomendaciones del proveedor de servicios en la nube o de un socio adecuado para conocer los modelos de arquitectura 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 se diseñaron.
- 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 sobre arquitectura 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 administrar 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 en la 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 seminarios web de AWS
  - AWS Workshops
  - AWS Communities
- Contacte con AWS cuando necesite más ayuda o información sobre un producto. AWS Los Solutions Architects y AWS Professional Services proporcionan orientación para la implementación de soluciones. AWS Los socios ponen a su disposición el conocimiento experto de AWS para ayudarle a mejorar la agilidad y la innovación para su empresa.
- Use Soporte si necesita asistencia técnica para usar un servicio de forma eficaz. Nuestros planes de asistencia están diseñados para ofrecerle la combinación perfecta de herramientas junto con el acceso a conocimientos especializados para que pueda tener éxito con AWS mientras optimiza el rendimiento, administra los riesgos y mantiene los costos bajo control.

Recursos

Documentos relacionados:

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

Videos 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 del AWS
- Ejemplos del AWS SDK
- AWS Analytics Reference Architecture

PERF01-BP03 Contemplación de los costos en las decisiones sobre arquitectura

Tenga en cuenta los costos en sus decisiones arquitectónicas para mejorar el uso 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 tiene 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 costos 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 costos, puede mejorar el uso de los recursos y evitar pérdidas en una carga de trabajo en la nube. Por lo general, al contemplar los costos 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 costos, 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 costos de su carga de trabajo. Puede usar Calculadora de precios de AWS y AWS Cost Explorer para identificar los principales factores que influyen en los costos de su carga de trabajo.
- Consulte los modelos de precios en la nube, como instancias bajo demanda, instancias reservadas, Savings Plans e instancias de spot.
- Utilice las prácticas recomendadas de optimización de costos de Well-Architected para optimizar estos componentes clave en términos de costos.
- Supervise y analice los costos de forma continua para identificar oportunidades que le permitan optimizar los gastos de su carga de trabajo.
  - Use AWS Budgets para recibir alertas sobre costos inaceptables.
  - Use AWS Compute Optimizer o AWS Trusted Advisor para obtener recomendaciones sobre la optimización de costos.
  - Use la Detección de anomalías en los costos de AWS para detectar automáticamente las anomalías en los costos y analizar la causa raíz.

Recursos

Documentos relacionados:

- What is AWS Billing and Cost Management?
- Optimización de costos 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
- Biblioteca de soluciones de AWS
- Centro de conocimiento de AWS

Videos 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
- Cost Optimization Workshop
- Cloud Financial Management Technical Implementation Playbooks
- Startup optimization: Tuning application performance for maximum efficiency
- Serverless Optimization Workshop (Performance and Cost)
- Scaling cost effective architectures

PERF01-BP04 Evaluación del efecto de las decisiones en los clientes y en 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

- Comprenda los requisitos de la carga de trabajo y los SLA.
- Defina claramente los factores de la evaluación. Estos factores pueden estar relacionados con los costos, la fiabilidad, la seguridad y el rendimiento de su carga de trabajo.
- Seleccione una arquitectura y unos servicios que puedan satisfacer sus necesidades.
- Lleve a cabo 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 compensaciones de la complejidad, el rendimiento y el costo 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
- KPI de QuickSight

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

Videos 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:

- Measure page load time with Amazon CloudWatch Synthetics
- Cliente web de Amazon CloudWatch RUM

PERF01-BP05 Uso de 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 sobre arquitectura. 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. Implemente la carga de trabajo a través de políticas o

arquitecturas de referencia. Integre los servicios en su implementación 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 de pruebas.
- 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
- Biblioteca de soluciones de AWS
- Centro de conocimiento de AWS
- AWS Architecture Blog

Videos relacionados:

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

Ejemplos relacionados:

- Ejemplos del AWS
- Ejemplos del AWS SDK

PERF01-BP06 Uso de 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 sobre arquitectura en esos datos.

Patrones comunes de uso no recomendados:

- 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.

Beneficios de establecer esta práctica recomendada: el estudio comparativo de su implementación actual le permite medir las mejoras del 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 hacer una prueba de carga.

Puede crear sus propias pruebas comparativas personalizadas, o bien usar un estándar del sector, como TPC-DS, para comparar sus cargas de trabajo. 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 llevar a cabo en su arquitectura.

Con las pruebas comparativas, es importante llevar a cabo 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 implementación 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

- Planificación y definición:
  - 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 su carga de trabajo.
- Configuración e instrumentación:
  - Configure el entorno y los recursos.
  - Implemente la supervisión y el registro para recopilar los resultados de las pruebas.
- Comparación y supervisión:
  - Haga las pruebas comparativas y supervise las métricas durante la prueba.
- Análisis y documentación:
  - 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 conocimiento de AWS
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Genomics workflows, Part 5: automated benchmarking
- Benchmark and optimize endpoint deployment in Amazon SageMaker AI JumpStart

Videos 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
- Demo of Amazon CloudWatch Synthetics

Ejemplos relacionados:

- Ejemplos del AWS
- Ejemplos del AWS SDK
- Pruebas de carga distribuidas
- Measure page load time with Amazon CloudWatch Synthetics
- Cliente web de Amazon CloudWatch RUM

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

Defina un enfoque claro basado en los datos para utilizarlo cuando tome decisiones sobre arquitectura y asegurarse de que se utilizan los servicios y las configuraciones en la 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 una 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 costos. Tenga en cuenta factores como la cantidad de usuarios y el modo de uso de la carga de trabajo.

- Cree un plan de arquitectura o una lista de tareas pendientes relacionadas con la tecnología que tengan la misma prioridad que las tareas pendientes relacionadas con la funcionalidad.
- Evalúe y valore los diferentes servicios en la nube (para obtener más información, consulte PERF01-BP01 Descubrimiento y comprensión de 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 Uso de las recomendaciones del proveedor de servicios en la nube o de un socio adecuado para conocer los modelos de arquitectura y las prácticas recomendadas).

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

- Defina métricas, como el rendimiento y el tiempo de respuesta, que puedan ser de ayuda a la hora de evaluar el rendimiento de su carga de trabajo.
- Pruebe y utilice las métricas definidas para validar el rendimiento de la arquitectura seleccionada.
- Mantenga un control continuo y haga 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:

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

Videos 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 del AWS
- Ejemplos del AWS SDK

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. No seleccionar la opción de computación correcta 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 Selección de las mejores opciones computacionales para su carga de trabajo
- PERF02-BP02 Comprensión de las opciones de configuración y las características de computación disponibles
- PERF02-BP03 Recopilación de métricas relacionadas con la computación
- PERF02-BP04 Configuración y dimensionamiento correcto de los recursos de computación
- PERF02-BP05 Escalado de los recursos de computación de forma dinámica
- PERF02-BP06 Uso de aceleradores de computación optimizados basados en hardware

PERF02-BP01 Selección de 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 costos de infraestructura innecesarios y aligerar los esfuerzos operativos necesarios para mantener esa carga de trabajo.

Patrones comunes de uso no recomendados:

- Utiliza la misma opción computacional que en el entorno en las instalaciones.
- No tiene información suficiente sobre las opciones de computación, las características y las soluciones de la nube, y cómo estas podrían mejorar el rendimiento de computación.
- Ha aprovisionado en exceso una opción de computación existente para cumplir los requisitos de escalado o rendimiento cuando una opción de computación alternativa se ajustaría con mayor precisión a las características de la carga de trabajo.

Beneficios de establecer una 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 usar Amazon EC2 para lanzar y administrar servidores virtuales, AWS Lambda para poner en marcha código sin tener que aprovisionar o administrar servidores, Amazon ECS o Amazon EKS para poner en marcha y administrar contenedores, o AWS Batch para procesar grandes volúmenes de datos en paralelo. En función de sus necesidades de computación y escalado, 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 permitirán 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

- Comprenda 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 escalado y los requisitos de latencia.
- Obtenga información sobre los diferentes servicios de computación de AWS para su carga de trabajo. Para obtener más información, consulte PERF01-BP01 Descubrimiento y comprensión de los servicios y las características disponibles en la nube. Estas son algunas de las principales opciones de computación de AWS, sus características y casos de uso comunes:

Servicio de AWS	Características clave	Casos de uso comunes
Amazon Elastic Compute Cloud (Amazon EC2)	Cuenta con una opción dedicada para hardware, requisitos de licencia, una	Migraciones mediante lift-and-shift, aplicación monolític

Servicio de AWS	Características clave	Casos de uso comunes
	amplia selección de distintas familias de instancias, tipos de procesadores y acelerado res de computación.	a, entornos híbridos, aplicacio nes empresariales
Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS)	Implementación sencilla, entornos coherentes, escalable	Microservicios, entornos híbridos
AWS Lambda	Servicio de computación sin servidor que pone en marcha código como respuesta a eventos y administra automáticamente los recursos de computación subyacentes.	Microservicios, aplicaciones basadas en eventos
AWS Batch	Aprovisiona y escala de manera eficiente y dinámica Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernete s Service (Amazon EKS) y recursos de computación de AWS Fargate, con la opción de usar instancias de spot o bajo demanda en función de los requisitos de su trabajo	HPC, entrenamiento de modelos de ML
Amazon Lightsail	Aplicación de Linux y Windows preconfigurada para poner en marcha cargas de trabajo pequeñas	Aplicaciones web simples, sitio web personalizado

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

- Computación en la nube con AWS
- Tipos de instancias de Amazon EC2
- Contenedores de Amazon EKS: nodos de trabajo de Amazon EKS
- Contenedores de Amazon ECS: instancias de contenedor de Amazon ECS
- Funciones: configuración de funciones de Lambda
- Prescriptive Guidance for Containers
- Prescriptive Guidance for Serverless

Videos 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
- Implementación de modelos de ML para realizar inferencias con un alto rendimiento y un bajo costo

Ejemplos relacionados:

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

PERF02-BP02 Comprensión de 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 permitirá aprovisionar la cantidad de recursos adecuada y 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.
- Aprovisiona un exceso de recursos de computación 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 de computación 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 comprobar si su opción computacional actual es apropiada para las cargas de trabajo en cuanto a CPU y memoria.

Pasos para la implementación

- Sepa cuáles son los requisitos de la carga de trabajo (como los requisitos de CPU, la memoria y la latencia).
- 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

Opción de configuración	Ejemplos
	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 adecuadas para ampliar la capacidad por encima de las necesidades de rendimiento estándar. • Los 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	Use la configuración de escalado automático para ajustar los recursos de computación a los patrones de tráfico.
Ajuste del tamaño	• Use Compute Optimizer para obtener recomendaciones con tecnología de machine learning sobre qué configuración de computación se ajusta mejor a sus características de computación. • Use AWS Lambda Power Tuning para seleccionar la mejor configuración para su función de Lambda.

Opción de configuración	Ejemplos
Aceleradores de computación basados en hardware	• Las instancias de computación acelerada ponen en marcha funciones de diversos tipos, por ejemplo, de procesamiento de gráficos o de 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:

- Computación en la nube con AWS
- Tipos de instancias de Amazon EC2
- Control de los estados del procesador de la instancia de Amazon EC2 Linux
- Contenedores de Amazon EKS: nodos de trabajo de Amazon EKS
- Contenedores de Amazon ECS: instancias de contenedor de Amazon ECS
- Funciones: configuración de funciones de Lambda

Videos 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 – Optimizing Amazon EKS for performance and cost on AWS

Ejemplos relacionados:

- Código de demostración de Compute Optimizer
- Amazon EC2 spot instances workshop
- Efficient and Resilient Workloads with Amazon EC2 AWS Auto Scaling
- Graviton developer workshop
- 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 Recopilación de métricas relacionadas con la computación

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

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 permitirá ajustar el rendimiento de las aplicaciones a los requisitos empresariales para garantizar que cumple con las necesidades de su carga de trabajo. También puede ser de ayuda para mejorar continuamente el rendimiento y el uso 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 para proporcionarle 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 uso 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

- Identifique qué métricas relacionadas con el rendimiento son relevantes para su carga de trabajo. Debe recopilar métricas sobre el uso 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).
  - Amazon EC2 default metrics
  - Amazon ECS default metrics
  - Amazon EKS default metrics
  - Lambda default metrics
  - Amazon EC2 memory and disk metrics
- Elija y configure la solución de registro y supervisión adecuada para su carga de trabajo.
  - AWS native Observability
  - AWS Distro para OpenTelemetry
  - Servicio administrado por Amazon para Prometheus
- Defina el filtro y la agregación que se necesitan para las métricas en función de los requisitos de su carga de trabajo.
  - Quantify custom application metrics with Amazon CloudWatch Logs and metric filters
  - Collect custom metrics with Amazon CloudWatch strategic tagging
- Configure políticas de retención de datos para que las métricas se ajusten a los objetivos operativos y de seguridad.
  - Retención de datos predeterminada para las métricas de CloudWatch
  - Retención de datos predeterminada para Registros de CloudWatch
- 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.
  - Create alarms for custom metrics using Amazon CloudWatch anomaly detection

    - Create metrics and alarms for specific web pages with Amazon CloudWatch RUM
- Utilice la automatización para implementar los agentes de agregación de métricas y registros.
    - AWS Systems Manager automation
    - OpenTelemetry Collector

Recursos

Documentos relacionados:

- Monitoreo y observabilidad
- Best practices: implementing observability with AWS
- Documentación de Amazon CloudWatch
- Recopilación de métricas y registros de instancias de Amazon EC2 y en los servidores en las instalaciones con el agente de CloudWatch
- Uso de Registros de Amazon CloudWatch con AWS Lambda
- Uso de Registros de CloudWatch con instancias de contenedor
- Publish custom metrics
- AWS Answers: Registro centralizado
- Servicios de AWS que publican métricas de CloudWatch
- Monitoring Amazon EKS on AWS Fargate

Videos 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

- Monitoring with Amazon CloudWatch dashboards
- Taller de Amazon EKS

PERF02-BP04 Configuración y dimensionamiento correcto de 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 de computación.
- 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 de computación 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 costos.

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 los recursos en la nube puede generar costos adicionales, mientras que un aprovisionamiento insuficiente puede provocar un rendimiento deficiente y una experiencia

negativa para el cliente. AWS proporciona herramientas como AWS Compute Optimizer y AWS Trusted Advisor, que utilizan datos históricos para ofrecer recomendaciones sobre el tamaño adecuado de sus recursos de computación.

Pasos para la implementación

- Elija el tipo de instancia que mejor se adapte a sus necesidades:
  - ¿Cómo elijo el tipo de instancia de Amazon EC2 apropiado para mi carga de trabajo?
  - Selección de tipo de instancia basada en atributos para la Flota de Amazon EC2
  - Create an Auto Scaling group using attribute-based instance type selection
  - Optimizing your Kubernetes compute costs with Karpenter consolidation
- 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, como Amazon CloudWatch.
- Seleccione la configuración correcta para cada recurso de computación.
  - En el caso de cargas de trabajo efímeras, evalúe las métricas de Amazon CloudWatch de la instancia, 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 Optimizer y AWS Trusted Advisor, para identificar oportunidades de optimizar y dimensionar correctamente el recurso de computación.
- 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:

- Computación en la nube con AWS
- Tipos de instancias de Amazon EC2
- Contenedores de Amazon ECS: instancias de contenedor de Amazon ECS

- Contenedores de Amazon EKS: nodos de trabajo de Amazon EKS
- Funciones: configuración de funciones de Lambda
- Control de los estados del procesador de la instancia de Amazon EC2

Videos 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
- Right-sizing recommendations

PERF02-BP05 Escalado de los recursos de computación 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 mediante el aumento manual de la capacidad.
- Utiliza las mismas directrices de dimensionamiento (por lo general, una infraestructura estática) que en el entorno en las instalaciones.
- Dejar 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 de computación puede ser útil para 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 escalado que se ajustan a los cambios de demanda. Junto con las métricas relacionadas con la computación, el escalado dinámico permite que las cargas de trabajo respondan automáticamente a los cambios y utilicen el conjunto óptimo de recursos de computación para lograr su objetivo.

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

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

Debe asegurarse de que las implementaciones de la carga de trabajo puedan manejar eventos de escalado vertical y reducción vertical.

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 escalado automático o como características del servicio. Estos son algunos ejemplos de mecanismos de escalado automático:

Mecanismo de escalado automático	Dónde se usa
Amazon EC2 Auto Scaling	Para garantizar que cuenta con la cantidad correcta de instancias de Amazon EC2

Mecanismo de escalado automático	Dónde se usa
	disponibles para controlar la carga de usuarios de su aplicación.
Aplicación de escalado automático	Para escalar automáticamente los recursos de servicios de AWS específicos más allá de Amazon EC2, como las funciones de AWS Lambda o los servicios de Amazon Elastic Container Service (Amazon ECS).
Kubernetes Cluster Autoscaler/Karpenter	Para escalar automáticamente clústeres de Kubernetes.

- Normalmente, se habla del escalado 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 adaptarse a la demanda.
- Asegúrese de que las métricas de escalado se ajusten a las características de la carga de trabajo que se implementa. Si está implementando una aplicación de transcodificación de vídeo, se espera un uso 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. Si es necesario, puede utilizar una métrica personalizada para su política de escalado. 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 escalado automático.
- Asegúrese de usar el escalado dinámico en lugar del escalado manual para su grupo de escalado automático. También le recomendamos que utilice políticas de escalado de seguimiento objetivo en su escalado dinámico.
- Compruebe que las implementaciones de la carga de trabajo puedan gestionar ambos eventos de escalado (escalado vertical y reducción vertical). Por ejemplo, puede usar el historial de actividad para verificar la actividad de escalado de un grupo de escalado automático.
- Evalúe los patrones predecibles de su carga de trabajo y escale de forma proactiva para anticiparse a los cambios previstos y planeados en la demanda. Con el escalado predictivo, puede

eliminar la necesidad de aprovisionar capacidad en exceso. Para más información, consulte Predictive Scaling with Amazon EC2 Auto Scaling.

Recursos

Documentos relacionados:

- Computación en la nube con AWS
- Tipos de instancias de Amazon EC2
- Contenedores de Amazon ECS: instancias de contenedor de Amazon ECS
- Contenedores de Amazon EKS: nodos de trabajo de Amazon EKS
- Funciones: configuración de funciones de Lambda
- Control de los estados del procesador de la instancia de Amazon EC2
- Deep Dive on Amazon ECS Cluster Auto Scaling
- Introducing Karpenter – An Open-Source High-Performance Kubernetes Cluster Autoscaler

Videos 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:

- Ejemplos de grupos de Amazon EC2 Auto Scaling
- Taller de Amazon EKS
- Scale your Amazon EKS workloads by running on IPv6

PERF02-BP06 Uso de aceleradores de computación optimizados basados en hardware

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

Patrones comunes de uso no recomendados:

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

Beneficios de establecer esta práctica recomendada: al utilizar aceleradores basados en hardware, como unidades de procesamiento gráfico (GPU) y matrices de puertas programables en campo (FPGA), puede poner en marcha 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 llevan a cabo 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 solo durante el tiempo que sea necesario y retírelo mediante automatización cuando no se requiera para mejorar la eficiencia del rendimiento general.

Pasos para la implementación

- Identifique qué instancias de computación acelerada pueden satisfacer sus requisitos.
- Para las cargas de trabajo de machine learning, utilice hardware personalizado específico para la carga de trabajo, como AWS Trainium, AWS Inferentia y Amazon EC2 DL1. AWS Las instancias de Inferentia, como las instancias Inf2, ofrecen hasta un 50 % más de rendimiento por vatio que las instancias de Amazon EC2 comparables.

- Recopile las métricas de uso de las instancias de computación acelerada. Por ejemplo, puede usar el agente de CloudWatch para recopilar métricas como `utilization_gpu` y `utilization_memory` para sus GPU, como se muestra en Recopilación de métricas de GPU NVIDIA con 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.
  - Optimización de las configuraciones de GPU
  - GPU Monitoring and Optimization in the Deep Learning AMI
  - Optimizing I/O for GPU performance tuning of deep learning training in Amazon SageMaker AI
- 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
- Instancias de GPU
- Instances with AWS Trainium
- Instances with AWS Inferentia
- Let's Architect! Architecting with custom chips and accelerators

- Computación acelerada
- Amazon EC2 VT1 Instances
- ¿Cómo elijo el tipo de instancia de Amazon EC2 apropiado para mi carga de trabajo?
- Choose the best AI accelerator and model compilation for computer vision inference with Amazon SageMaker AI

Videos 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 AI y NVIDIA GPU Cloud (NGC)
- Uso de SageMaker AI con Trainium e Inferentia para optimizar las cargas de trabajo de aprendizaje profundo, entrenamiento e inferencia
- Optimización de modelos de NLP con instancias Inf1 de Amazon Elastic Compute Cloud en Amazon SageMaker AI

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 de Well-Architected utilizan almacenes de datos diseñados específicamente 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 Uso de un almacén de datos personalizado que se adapte mejor a los requisitos de acceso y almacenamiento de datos
- PERF03-BP02 Evaluación de las opciones de configuración disponibles
- PERF03-BP03 Recopilación y registro de las métricas de rendimiento del almacén de datos
- PERF03-BP04 Implementación de estrategias para mejorar el rendimiento de las consultas en el almacén de datos
- PERF03-BP05 Implementación de patrones de acceso a datos que utilicen el almacenamiento en caché

PERF03-BP01 Uso 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).

Patrones comunes de uso no recomendados:

- 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: comprender las características y los requisitos de los datos le permite determinar la tecnología de almacenamiento más eficiente y funcional 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, escalado 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

- Haga 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)
  - Periodo de retención de datos
- Obtenga información sobre los diferentes almacenes de datos (servicios de almacenamiento y base de datos) disponibles en AWS para su carga de trabajo que se ajustan a las características de los datos, tal y como se describe en PERF01-BP01 Descubrimiento y comprensión de 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
Almacenamiento de objetos	Amazon S3	Escalabilidad ilimitada, alta disponibilidad y múltiples opciones de accesibilidad. La transferencia y el acceso a objetos dentro y fuera de Amazon S3 puede utilizar un servicio, como Aceleración de transferencias o Puntos de acceso, para respaldar su ubicación, sus necesidades de seguridad y sus patrones de acceso.
Almacenamiento de archivos	Amazon S3 Glacier	Diseñado para archivar datos.
Almacenamiento en streaming	Amazon Kinesis Amazon Managed Streaming para Apache Kafka (Amazon MSK)	Ingesta y almacenamiento eficientes de datos de streaming.
Sistema de archivos compartidos	Amazon Elastic File System (Amazon EFS)	Sistema de archivos montable al que pueden acceder varios tipos de soluciones de computación.

Tipo	Servicios de AWS	Características clave
Sistema de archivos compartidos	Amazon FSx	Se basa en las últimas soluciones de computación de AWS para admitir cuatro sistemas de archivos de uso común: NetApp ONTAP, OpenZFS, Windows File Server y Lustre. La latencia, el rendimiento y las E/S por segundo de Amazon FSx varían según el sistema de archivos y deben tenerse en cuenta a la hora de seleccion ar el sistema de archivos adecuado para sus necesidad es de carga de trabajo.
Almacenamiento en bloque	Amazon Elastic Block Store (Amazon EBS)	Servicio de almacenamiento en bloque de alto rendimien to, escalable y fácil de usar diseñado para Amazon Elastic Compute Cloud (Amazon EC2). Amazon EBS incluye almacenamiento respaldado por SSD para cargas de trabajo transacci onales y de IOPS intensiva s, así como almacenam iento respaldado por HDD para cargas de trabajo de rendimiento intensivo.

Tipo	Servicios de AWS	Características clave
Base de datos relacional	Amazon Aurora, Amazon RDS, Amazon Redshift.	Se han diseñado para respaldar las transacciones ACID (atomicidad, coherencia, aislamiento, durabilidad) y mantener la integridad referencial y una sólida coherencia de datos. Muchas aplicaciones tradicionales, la planificación de recursos empresariales (ERP), la administración de relaciones con los clientes (CRM) y el comercio electrónico utilizan bases de datos relacionales para almacenar sus datos.
Base de datos de clave-valor	Amazon DynamoDB	Optimizada para patrones de acceso comunes, normalmente para almacenar y recuperar grandes volúmenes de datos. Las aplicaciones web con mucho tráfico, los sistemas de comercio electrónico y las aplicaciones de juegos son casos de uso típicos para las bases de datos de clave-valor.

Tipo	Servicios de AWS	Características clave
Base de datos de documentos	Amazon DocumentDB	Diseñada para almacenar datos semiestructurados como documentos tipo JSON. Estas bases de datos ayudan a los desarrolladores a crear y actualizar de forma rápida aplicaciones como la administración de contenido, catálogos y perfiles de usuario.
Base de datos en memoria	Amazon ElastiCache, Amazon MemoryDB para Redis	Se utilizan para aplicaciones que requieren acceso a los datos en tiempo real, menor latencia y mayor rendimiento. Puede usar bases de datos en memoria para el almacenamiento en caché de aplicaciones, la administración de sesiones, las tablas de clasificación de juegos, el almacén de características de ML de baja latencia, el sistema de mensajería de microservicios y un mecanismo de streaming de alto rendimiento.

Tipo	Servicios de AWS	Características clave
Base de datos de gráficos	Amazon Neptune	Se utiliza para aplicaciones que deben navegar y consultar millones de relaciones entre conjuntos de datos de grafos con un alto grado de conexión y con una latencia de milisegundos a gran escala. Muchas empresas utilizan las bases de datos de gráficos para detección de fraude, redes sociales y motores de recomendaciones.
Base de datos de serie temporal	Amazon Timestream	Se usa para recopilar, sintetizar y obtener información de forma eficaz a partir de datos que cambian con el tiempo. Las aplicaciones de IoT, DevOps y telemetría industrial pueden utilizar bases de datos de serie temporal.

Tipo	Servicios de AWS	Características clave
Columna ancha	Amazon Keyspaces (para Apache Cassandra)	Utiliza tablas, filas y columnas, pero, a diferenci a de una base de datos relacional, los nombres y el formato de las columnas pueden variar de una fila a otra en la misma tabla. Por lo general, un almacén de columnas anchas está en aplicaciones industria les a gran escala para el mantenimiento de equipos, la administración de flotas y la optimización de rutas.
Libro mayor	Amazon Quantum Ledger Database (Amazon QLDB)	Proporciona una autoridad centralizada y de confianza para mantener un registro de transacciones escalable , inmutable y verificable criptográficamente para cada aplicación. Las bases de datos de libro mayor se utilizan para sistemas de registro, la cadena de suministro, registros e incluso transacciones bancarias.

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

Pregunta	Aspectos que deben tenerse en cuenta
¿Cómo se estructuran los datos?	• Si los datos no están estructurados, considere un almacén de objetos como Amazon S3 o una base de datos NoSQL como Amazon DocumentDB • Para los datos de clave-valor, considere DynamoDB, Amazon ElastiCache (Redis OSS) o Amazon MemoryDB
¿Qué nivel de integridad referencial se requiere?	• Para las restricciones de claves externas, 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 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.
¿Se requiere el cumplimiento de ACID (atomicidad, coherencia, aislamiento, durabilidad)?	• Si se requiere cumplir las propiedades ACID asociadas a las bases de datos relacionales, considere la posibilidad de usar una base de datos relacional como Amazon RDS y Aurora. • Si se requiere una coherencia sólida para la base de datos NoSQL, puede utilizar lecturas altamente coherentes con DynamoDB.

Pregunta	Aspectos que deben tenerse en cuenta
¿Cómo cambiarán los requisitos de almacenamiento con el tiempo? ¿Cómo afecta esto a la escalabilidad?	• Las bases de datos sin servidor, como DynamoDB y Amazon Quantum Ledger Database (Amazon QLDB), se escalarán de forma dinámica. • 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.
¿Cuál es la proporción de consultas de lectura en relación con las de escritura? ¿Es probable que el almacenamiento en caché mejore el rendimiento?	• Las cargas de trabajo de lectura intensiva pueden beneficiarse de una capa de almacenamiento en caché, como ElastiCac he o DAX si la base de datos es de DynamoDB. • Las lecturas también pueden descargarse en réplicas de lectura con bases de datos relacionales, como Amazon RDS.

Pregunta	Aspectos que deben tenerse en cuenta
¿Tiene mayor prioridad el almacenamiento y la modificación (OLTP, procesamiento de transacciones en línea) o la recuperación y la elaboración de informes (OLAP, procesami ento analítico en línea)?	• Para el procesamiento transaccional de lecturas de alto rendimiento sin hacer 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, considere utilizar una base de datos en columnas como Amazon Redshift o exportar los datos a Amazon S3 y llevar a cabo análisis con Athena o Amazon QuickSight.
¿Qué nivel de durabilidad requieren los datos?	• 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 periodos de tiempo.

Pregunta	Aspectos que deben tenerse en cuenta
¿Existe el deseo de evitar los motores de bases de datos comerciales o los costos de licencia?	• Considere la posibilidad de utilizar motores de código abierto como PostgreSQL y MySQL en Amazon RDS o Aurora. • Use AWS Database Migration Service y AWS Schema Conversion Tool para llevar a cabo migraciones de los motores de bases de datos comerciales a los de código abierto.
¿Cuál es la expectativa operativa de la base de datos? ¿El cambio a los servicios administrados es una preocupación principal?	• 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 costos operativos.
¿Cómo se accede actualmente a la base de datos? ¿Se trata solo del acceso a la aplicación, o hay usuarios de inteligencia empresarial (BI) y otras aplicaciones comerciales conectadas?	• Si tiene dependencias en herramientas externas, es posible que deba mantener la compatibilidad con las bases de datos que admiten. Amazon RDS es totalmente compatible con las diferentes versiones de motores que admite, como Microsoft SQL Server, Oracle, MySQL y PostgreSQL.

- Lleve a cabo 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:

- Tipos de volúmenes de Amazon EBS
- Opciones de almacenamiento para sus instancias de Amazon EC2
- Amazon EFS: Amazon EFS Performance
- Amazon FSx for Lustre Performance
- Amazon FSx for Windows File Server Performance

- Amazon S3 Glacier: documentación de S3 Glacier
- Amazon S3: consideraciones de la tasa de solicitudes y del rendimiento
- Almacenamiento en la nube en AWS
- Amazon EBS I/O Characteristics
- Bases de datos en la nube de AWS
- AWS Database Caching
- DynamoDB Accelerator
- Prácticas recomendadas de Amazon Aurora
- Rendimiento de Amazon Redshift
- Amazon Athena top 10 performance tips
- Amazon Redshift Spectrum best practices
- Prácticas recomendadas para Amazon DynamoDB
- Choose between Amazon EC2 and Amazon RDS
- Best Practices for Implementing Amazon ElastiCache

Videos 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
- Ejemplos de Amazon S3
- 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 Evaluación de 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.

Patrones comunes de uso no recomendados:

- Solo utiliza un tipo de almacenamiento, como, por ejemplo, Amazon EBS, para todas las cargas de trabajo.
- Utiliza IOPS aprovisionadas en todas las cargas de trabajo sin efectuar pruebas reales 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 lleva a cabo pruebas en las características de escalado de su almacén de datos.

Beneficios de establecer esta práctica recomendada: si explora y experimenta con las configuraciones de almacenamiento de datos, puede reducir el costo 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 almacenamientos de datos que se utilicen en función de los requisitos de almacenamiento y acceso. Para optimizar los costos 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 ser de ayuda para mejorar el rendimiento de su almacén de datos. Las principales opciones de almacenamiento de datos que debe tener en cuenta son las siguientes:

Opción de configuración	Ejemplos
Descarga de lecturas (como réplicas de lectura y almacenamiento en caché)	• En el caso de las tablas de DynamoDB, puede descargar las lecturas con DAX para el almacenamiento en caché. • Puede crear un clúster de Amazon ElastiCache (Redis OSS) 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 agregar réplicas de lectura para descargar las partes de lectura de la carga de trabajo.

Opción de configuración	Ejemplos
	• Las bases de datos sin servidor, como DynamoDB, se escalarán automátic amente. Asegúrese de que tenga suficient es unidades de capacidad de lectura (RCU) aprovisionadas para gestionar la carga de trabajo.

Opción de configuración	Ejemplos
Escalado de escrituras (como la fragmentación de claves de partición o la introducción de una cola)	• 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 al ajustar 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 si se elige una clave de partición distribuida de manera más uniforme o se particiona la escritura en función de la clave de partición.

Opción de configuración	Ejemplos
Políticas para administrar el ciclo de vida de los conjuntos de datos	• Puede usar Amazon S3 Lifecycle para administrar los objetos a lo largo de su ciclo de vida. Si sus patrones de acceso son desconocidos, cambiantes o impredecibles, puede utilizar Amazon S3 Intelligent-Tiering, que supervisa los patrones de acceso y mueve automáticamente los objetos a los que no se ha accedido a niveles de acceso de menor costo. Puede aprovechar las métricas de Lente de almacenamiento de Amazon S3 para identificar las oportunidades de optimización y las brechas en la administración del ciclo de vida. • La administración del ciclo de vida de Amazon EFS administra automáticamente el almacenamiento económico de los archivos para sus sistemas de archivos.
Administración y agrupación de conexiones	• 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 escalado automático para hacer frente a los picos de carga.

- Lleve a cabo 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 (como Amazon CloudWatch) y optimización (como Lente de almacenamiento de Amazon S3) de AWS para optimizar continuamente el almacén de datos utilizando patrones de uso del mundo real.

Recursos

Documentos relacionados:

- Almacenamiento en la nube en AWS
- Tipos de volúmenes de Amazon EBS
- Opciones de almacenamiento para sus instancias de Amazon EC2
- Amazon EFS: Amazon EFS Performance
- Amazon FSx for Lustre Performance
- Amazon FSx for Windows File Server Performance
- Amazon S3 Glacier: documentación de S3 Glacier
- Amazon S3: consideraciones de la tasa de solicitudes y del rendimiento
- Amazon EBS I/O Characteristics
- Bases de datos en la nube de AWS
- AWS Database Caching
- DynamoDB Accelerator
- Prácticas recomendadas de Amazon Aurora
- Rendimiento de Amazon Redshift
- Amazon Athena top 10 performance tips
- Amazon Redshift Spectrum best practices
- Prácticas recomendadas para Amazon DynamoDB

Videos 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
- Ejemplos de Amazon S3
- Ejemplos de Amazon DynamoDB
- AWS Database migration samples
- Database Modernization Workshop
- Working with parameters on your Amazon RDS for Postgress DB

PERF03-BP03 Recopilación y registro de 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 ser de ayuda para optimizar el almacén de datos, garantizar que se cumplen los requisitos de la carga de trabajo y 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 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 permite 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 el uso 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

- Identifique las métricas de rendimiento clave del almacén de datos que desee supervisar.
  - Métricas y dimensiones de Amazon S3
  - Supervisión de métricas en una instancia de Amazon RDS
  - Monitoreo de la carga de base de datos con Performance Insights en Amazon RDS
  - Descripción general de la supervisión mejorada
  - Dimensiones y métricas de DynamoDB
  - Supervisión de DynamoDB Accelerator
  - Supervisión de Amazon MemoryDB con Amazon CloudWatch
  - ¿Qué métricas debo monitorear?
  - Supervisión del rendimiento de clústeres de Amazon Redshift
  - Dimensiones y métricas de Timestream

  - Métricas de Amazon CloudWatch para Amazon Aurora
  - Supervisión de Amazon Keyspaces (para Apache Cassandra)
  - Supervisión de recursos de Amazon Neptune
- Use una solución de registro y supervisión aprobada para recopilar estas métricas. Amazon CloudWatch puede recopilar métricas entre 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 indiquen cuándo se superan los umbrales.
- Compruebe si la supervisión del almacén de datos puede beneficiarse de una solución de machine learning que detecte anomalías de rendimiento.
  - Amazon DevOps Guru para Amazon RDS brinda visibilidad sobre los problemas de rendimiento y recomienda acciones correctivas.
- Configure la retención de datos de la solución de supervisión y registro para que se ajuste a sus objetivos operativos y de seguridad.
  - Retención de datos predeterminada para las métricas de CloudWatch
  - Retención de datos predeterminada para Registros de CloudWatch

Recursos

Documentos relacionados:

- AWS Database Caching
- Amazon Athena top 10 performance tips
- Prácticas recomendadas con Amazon Aurora
- DynamoDB Accelerator
- Prácticas recomendadas para Amazon DynamoDB
- Amazon Redshift Spectrum best practices
- Rendimiento de Amazon Redshift
- Bases de datos en la nube con AWS
- Amazon RDS Performance Insights

Videos 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 Implementación de 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 costos 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 llevan a cabo en el almacén de datos.
- Identifique las consultas de procesamiento lento 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:
  - Usar un formato de archivo de 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.
    - Understanding materialized views
    - 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.
    - Llevar a cabo un mantenimiento periódico, como vacío, reindexación y ejecución de estadísticas.
- Experimente y pruebe estrategias en un entorno que no sea de producción.

Recursos

Documentos relacionados:

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

Videos 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:

- AWS Purpose Built Databases Workshop

PERF03-BP05 Implementación de 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 costos.

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 mediante la repetición de un cálculo anterior o mediante la recuperación 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 se pueden almacenar en caché en varios niveles de la aplicación, por ejemplo, dentro de la aplicación mediante llamadas remotas, lo que se conoce como almacenamiento en caché del cliente, o mediante un servicio secundario rápido para almacenar los datos, conocido como almacenamiento en caché remoto.

Almacenamiento en caché del cliente

Con el almacenamiento en caché 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 una 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

- 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.
  - Database Caching
  - Habilitación del almacenamiento en caché de la API para mejorar la capacidad de respuesta
- Identifique el tipo de estrategia de almacenamiento en caché adecuada que mejor se adapte a su patrón de acceso.
  - Estrategias de almacenamiento en caché
  - AWS Caching Solutions
- Siga las prácticas recomendadas de almacenamiento en caché para su almacén de datos.
- 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.

- 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.
  - Best practices: Redis clients and Amazon ElastiCache (Redis OSS)
- Supervise la tasa de aciertos de 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.
  - Which metrics should I monitor?
  - Best practices for monitoring Redis workloads on Amazon ElastiCache
  - Monitoring best practices with Amazon ElastiCache (Redis OSS) using Amazon CloudWatch
- 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:

- Using the Amazon ElastiCache Well-Architected Lens
- Monitoring best practices with Amazon ElastiCache (Redis OSS) using Amazon CloudWatch
- ¿Qué métricas debo monitorear?
- Documento técnico Performance at Scale with Amazon ElastiCache
- Desafíos y estrategias del almacenamiento en caché

Videos 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 (Redis OSS)

Ejemplos relacionados:

- Boosting MySQL database performance with Amazon ElastiCache (Redis OSS)

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 en las instalaciones, 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 Comprensión del efecto de las redes en el rendimiento
- PERF04-BP02 Evaluación de las características de las redes disponibles
- PERF04-BP03 Elección de la conectividad o VPN dedicadas adecuadas para la carga de trabajo
- PERF04-BP04 Uso del equilibrio de carga para distribuir el tráfico entre varios recursos
- PERF04-BP05 Elección de los protocolos de red para mejorar el rendimiento
- PERF04-BP06 Elección de la ubicación de la carga de trabajo en función de los requisitos de la red
- PERF04-BP07 Optimización de la configuración de red según las métricas

PERF04-BP01 Comprensión del efecto de las redes en el 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 en las instalaciones 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 lo 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 en las instalaciones, 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 implementación:

- 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.

- Obtenga información sobre los principales servicios de red de AWS, como las VPC, AWS Direct Connect, Elastic Load Balancing (ELB) y Amazon Route 53.
- Capture 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 • AWS Transit Gateway metrics • AWS PrivateLink metrics
Características de las redes de aplicaciones	• Elastic Fabric Adapter • AWS App Mesh metrics • Métricas de Amazon API Gateway
Características de las redes de periferia	• Métricas de Amazon CloudFront • Amazon Route 53 metrics • AWS Global Accelerator metrics
Características de las redes híbridas	• AWS Direct Connect metrics • AWS Site-to-Site VPN metrics • AWS Client VPN metrics • Nube de AWS WAN metrics
Características de las redes de seguridad	• AWS Shield, AWS WAF, and AWS Network Firewall metrics
Características de rastreo	• AWS X-Ray • VPC Reachability Analyzer • Analizador de acceso a la red • Amazon Inspector • Amazon CloudWatch RUM

- Comparación y prueba del rendimiento de la red:

- Lleve a cabo pruebas comparativas del 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.
- Haga pruebas de carga para experimentar con soluciones y opciones de redes.

Recursos

Documentos relacionados:

- Equilibrador de carga de aplicación
- Redes de EC2 mejoradas en Linux
- Redes de EC2 mejoradas en Windows
- Grupos de ubicación de EC2
- Habilitación de las redes mejoradas con Elastic Network Adapter (ENA) en las instancias Linux
- Network Load Balancer
- Productos de redes con AWS
- Transit Gateway
- Transitioning to latency-based routing in Amazon Route 53
- Puntos de enlace de la VPC

Videos 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
- Talleres de redes de AWS
- Hands-on Network Firewall Workshop
- Observing and Diagnosing your Network on AWS
- Finding and addressing Network Misconfigurations on AWS

PERF04-BP02 Evaluación de las características de las redes disponibles

Evalúe las características de la red en la nube que pueden aumentar el rendimiento. Mida el impacto de estas características a través de pruebas, métricas y análisis. Por ejemplo, aproveche las características 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 incluyen

características de producto (como la función Aceleración de transferencias de Amazon S3) para optimizar el tráfico de 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.
  - Considere la posibilidad de usar WAN en la Nube de AWS para diseñar, administrar y supervisar la red de su organización al crear una red global unificada.
  - Supervise sus redes globales y principales con las métricas de Registros de Amazon CloudWatch. Use Amazon CloudWatch RUM, que proporciona información para ayudar a identificar, comprender y mejorar la experiencia digital de los usuarios.
  - Consulte la latencia de red agregada entre Regiones de AWS y las zonas de disponibilidad, así como dentro de cada zona de disponibilidad, mediante AWS Network Manager para obtener información sobre la relación entre el rendimiento de su aplicación y 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 de su configuración.
- Si se trata de una carga de trabajo existente, identifique y documente el punto de referencia para sus métricas de rendimiento, y céntrese en los cuellos de botella y las áreas que debe 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 carga de trabajo nueva, ejecute pruebas de carga para identificar los cuellos de botella en el rendimiento.
- Para los cuellos de botella en el 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 el Analizador de acceso a la red para identificar rutas o rutas.
Protocolos de red	Consulte PERF04-BP05 Elección de los protocolos de red para mejorar el rendimiento
Topología de la red	Evalúe las ventajas y desventajas operativas de usar el emparejamiento de VPC frente a usar 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 en las instalaciones. Comparta su AWS Transit Gateway entre varias cuentas mediante AWS Resource Access Manager. Consulte PERF04-BP03 Elección de 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. Use 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 ayudarle a mejorar el rendimiento de su carga de trabajo para un público global. Para identificar qué opción de enrutamiento optimizaría el rendimiento de su carga de trabajo, revise su tráfico y la ubicación de los usuarios cuando la carga de trabajo se distribuya globalmente.

Oportunidad de mejora	Solución
Características de los recursos de almacenamiento	La Aceleración de transferencias de Amazon S3 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. Los puntos de acceso multirregionales de Amazon S3 replican el contenido en varias regiones y simplifica la carga de trabajo al proporcionar 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 red elásticas (ENI) utilizadas por las instancias de Amazon EC2, los contenedores y las funciones de Lambda están limitadas por el flujo. Revise sus grupos de ubicación para optimizar el rendimiento de sus redes 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. El comando `ethtool` se incluye en el controlador de ENA y expone métricas adicionales relacionadas con la red que pueden publicarse como una métrica personalizada en CloudWatch. Los Amazon Elastic Network Adapters (ENA) entregan una mayor optimización al proporcionar un mayor rendimiento para las instancias de un grupo con ubicación en clúster. Elastic Fabric Adapter (EFA) es una interfaz de red para instancias de Amazon EC2 que le permite ejecutar aplicaciones que requieren altos niveles de comunicaciones 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 las operaciones de E/S de Amazon EBS.

Recursos

Documentos relacionados:

- Equilibrador de carga de aplicación
- Redes de EC2 mejoradas en Linux
- Redes de EC2 mejoradas en Windows
- Grupos de ubicación de EC2
- Habilitación de las redes mejoradas con Elastic Network Adapter (ENA) en las instancias Linux
- Network Load Balancer
- Productos de redes con AWS
- Transitioning to Latency-Based Routing in Amazon Route 53
- Puntos de enlace de la VPC
- Logs de flujo de VPC

Videos 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
- Talleres de redes de AWS
- Observing and diagnosing your network
- Finding and addressing network misconfigurations on AWS

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

Cuando se requiera conectividad híbrida para conectar los recursos en las instalaciones 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 en función de los requisitos de ancho de banda. AWS Direct Connect le permite conectar su red en las instalaciones 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 costo 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 los 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

- Calcule los requisitos de ancho de banda y de latencia de sus aplicaciones actuales.
  - En el caso de cargas de trabajo existentes que se trasladan a AWS, utilice los datos de sus sistemas internos de supervisión de red.
  - 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.
- 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.
  - 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 costo (ECMP) con BGP.
  - 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.
- Siga la documentación de AWS para elegir la opción de conectividad adecuada:
  - Si decide usar AWS Direct Connect, seleccione el ancho de banda adecuado para su conectividad.
  - Si utiliza una conexión de AWS Site-to-Site VPN en varias ubicaciones para conectarse a una Región de AWS, utilice una conexión de VPN Site-to-Site acelerada para tener la oportunidad de mejorar el rendimiento de la red.
  - Si el diseño de su red consta de una conexión VPN IPsec a través de AWS Direct Connect, considere la posibilidad de utilizar una VPN con IP privada para mejorar la seguridad y lograr la segmentación. AWS La VPN con IP privada de Site-to-Site se implementa sobre la interfaz virtual (VIF) de tránsito.

    - 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, pasando por alto Regiones de AWS.
- Valide la configuración de la conectividad antes de la implementación en producción. Lleve a cabo pruebas de seguridad y rendimiento para asegurarse de que cumpla los requisitos de ancho de banda, fiabilidad, latencia y cumplimiento.
- 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
- VPC Endpoints
- Creación de una infraestructura de red de AWS multiVPC escalable y segura
- Client VPN

Videos 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

- Talleres de redes de AWS

PERF04-BP04 Uso del 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 dirigir el tráfico de manera eficaz.

Patrones comunes de uso no recomendados:

- No tiene en cuenta los requisitos de la carga de trabajo al elegir el tipo de equilibrador de carga.
- No aprovecha 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 escalado automático y un mejor uso 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 el uso.

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, zona local, Outpost o aislamiento de zona).

AWS proporciona varios modelos para que sus aplicaciones utilicen el equilibrio de carga. El equilibrador de carga de aplicación es el más adecuado para el equilibrio de carga del tráfico de HTTP y HTTPS y entrega un direccionamiento de solicitudes avanzado enfocado a la entrega de arquitecturas de aplicaciones modernas, incluidos los microservicios y los contenedores.

El equilibrador de carga de red es el más adecuado para el equilibrio de carga del tráfico de TCP donde se necesite un rendimiento extremo. Es capaz de gestionar millones de solicitudes por segundo a la vez que mantiene 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 hacer su backend para atender al tráfico.

Por ejemplo, al utilizar tanto el equilibrador de carga de aplicación (ALB) como el equilibrador de carga de red (NLB), puede llevar a cabo 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.

El equilibrador de carga de aplicación 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 el equilibrador de carga de red, que ofrece latencias extremadamente bajas. Como alternativa, puede decidir acercar su carga de trabajo a sus clientes con el equilibrador de carga de aplicación en las 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 el escalado automático 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 escalado automático, los destinos se agregarán 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: equilibrador de carga del servicio
- Equilibrador 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.
  - Use el equilibrador de carga de aplicación para cargas de trabajo HTTP/HTTPS.
  - Utilice el equilibrador de carga de red para cargas de trabajo distintas de HTTP que se ejecuten en TCP o UDP.
  - Utilice una combinación de ambos (ALB como objetivo de NLB) si desea aprovechar las características de ambos productos. Por ejemplo, puede hacerlo si desea utilizar las IP estáticas del equilibrador de carga de red junto con el enrutamiento basado en encabezado HTTP del equilibrador de carga de aplicación, o si desea exponer su carga de trabajo HTTP a un 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 oyentes HTTPS/TLS con un equilibrador de carga de aplicación y un equilibrador de carga de red 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 Aplicación del cifrado en tránsito.

- Seleccione el algoritmo de enrutamiento adecuado (solo para el 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, el equilibrador de carga de aplicación ofrece dos opciones para los algoritmos de enrutamiento:
  - Solicitudes menos pendientes: utilícelas 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.
  - Patrón rotativo: utilícelo cuando las solicitudes y los destinos sean similares, o si necesita distribuir las solicitudes equitativamente entre los destinos.
- Considere el aislamiento entre zonas o de zonas.
  - Desactive el aislamiento entre zonas (aislamiento de zonas) para mejorar la latencia y los dominios de error de zona. Está desactivado de forma predeterminada en el equilibrador de carga de red y en el equilibrador de carga de aplicación lo puede desactivar por grupo de destino.
  - Active el aislamiento entre zonas para aumentar la disponibilidad y flexibilidad. Está activado de forma predeterminada para el equilibrador de carga de aplicación y en el equilibrador de carga de red lo puede activar 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 más información sobre cómo hacer esto para Apache y Nginx, consulte ¿Cuál es la configuración óptima para usar Apache o NGINX como servidor de backend para ELB?.
- Active la supervisión de su equilibrador de carga.
  - Active los registros de acceso del equilibrador de carga de aplicación y el equilibrador de carga de red.
  - Los principales campos a tener en cuenta para el equilibrador de carga de aplicación son `request_processing_time`, `request_processing_time` y `response_processing_time`.
  - Los principales campos a tener en cuenta para el equilibrador de carga de red son `connection_time` y `tls_handshake_time`.
  - Esté preparado para consultar los registros cuando los necesite. Puede usar Amazon Athena para consultar tanto los registros del equilibrador de carga de aplicación como los registros del equilibrador de carga de red.

- Cree alarmas para las métricas relacionadas con el rendimiento, como `TargetResponseTime` para el ALB.

Recursos

Documentos relacionados:

- Comparación de productos de Elastic Load Balancing
- Infraestructura global de AWS
- Improving Performance and Reducing Cost Using Availability Zone Affinity
- Step by step for Log Analysis with Amazon Athena
- Consulta de los registros del Equilibrador de carga de aplicación
- 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

Videos 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 Elección de 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 mediante 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 agregar 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 agregada 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 agrega 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

- Use los servicios de 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.

- 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 la carga de trabajo sirve contenido dinámico a los usuarios, considere la posibilidad de usar 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.
- 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 estas cargas de trabajo, considere la posibilidad de usar la descarga de 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 el uso de la CPU en las instancias backend, lo que puede mejorar el rendimiento y aumentar la capacidad.
- Use el equilibrador de carga de red (NBT) para implementar 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 equilibrador de carga de red 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.
- Para sus cargas de trabajo de computación de alto rendimiento (HPC), considere la posibilidad de utilizar la funcionalidad Elastic Network Adapter (ENA) Express, que utiliza el protocolo SRD para mejorar el rendimiento de la red al proporcionar un ancho de banda de flujo único más alto (25 Gbps) y una latencia de cola más baja (percentil 99,9) para el tráfico de red entre las instancias de EC2.
- Utilice el equilibrador de carga de aplicación (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
- Equilibrador de carga de aplicación
- Redes de EC2 mejoradas en Linux
- Redes de EC2 mejoradas en Windows
- Grupos de ubicación de EC2
- Habilitación de las redes mejoradas con Elastic Network Adapter (ENA) en las instancias Linux
- Network Load Balancer
- Productos de redes con AWS
- Transitioning to Latency-Based Routing in Amazon Route 53
- Puntos de enlace de la VPC

Videos 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
- Talleres de redes de AWS

PERF04-BP06 Elección de 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.

Patrones comunes de uso no recomendados:

- 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 ubicación para la creación de redes. Un grupo de ubicación es una agrupación lógica de instancias para reducir la latencia. El uso de grupos de ubicación 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, como Amazon 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 de periferia 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

- Capture 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
  - How the client IP address is preserved in AWS Global Accelerator
- Analice los patrones de acceso a 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 la red.
  - Analice los datos para identificar el patrón de acceso a la red.
- Seleccione las regiones para la implementación de la carga de trabajo en función de los siguientes elementos clave:
  - Ubicación de los datos: en el caso de las aplicaciones con gran cantidad de datos (como macrodatos y machine learning), el código de la aplicación debe ejecutarse lo más cerca posible de los datos.
  - Ubicación de los usuarios: para las aplicaciones orientadas al usuario, elija una región (o regiones) cercana a los usuarios de su carga de trabajo.
  - Otras restricciones: tenga en cuenta las limitaciones, como el costo y el cumplimiento, tal y como se explica en What to Consider when Selecting a Region for your Workloads.
- Use Zonas locales de AWS para ejecutar cargas de trabajo como la renderización de video. 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 en las instalaciones 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 estas aplicaciones, considere la posibilidad de usar AWS Wavelength. AWS Wavelength integra los servicios de computación y almacenamiento de AWS en las redes 5G, proporcionando una infraestructura de computación periférica móvil para desarrollar, implementar y escalar aplicaciones de latencia ultrabaja.

- Utilice almacenamiento en caché local o soluciones de almacenamiento en caché de AWS para los activos de uso frecuente con el fin de mejorar el rendimiento, reducir el movimiento de datos y disminuir el impacto medioambiental.

Servicio	Cuándo se debe usar
Amazon CloudFront	Se usa para almacenar en caché el contenido estático como imágenes, scripts y videos, así como el contenido dinámico como respuestas de API y aplicaciones web.
Amazon ElastiCache	Se usa para almacenar en caché el contenido de las aplicaciones web.
DynamoDB Accelerator	Se usa para agregar 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 los siguientes:

Servicio	Cuándo se debe usar
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. Esta

aplicaciones pueden aprovechar servicios de redes que proporcionen direcciones IP estáticas de anycast y la terminación de TCP en las 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 multirregionales. 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 Implementación de regiones según los costos
- COST08-BP03 Implementación de servicios para reducir los costos de transferencia de datos
- REL10-BP01 Implementación de la carga de trabajo en varias ubicaciones
- REL10-BP02 Selección de las ubicaciones adecuadas para la implementación en varias ubicaciones
- SUS01-BP01 Selección de la región en función de los requisitos empresariales y los objetivos de sostenibilidad
- SUS02-BP04 Optimización de 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
- 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

Videos relacionados:

- AWS Local Zones Explainer Video
- 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 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 2021 - Building low-latency websites with Amazon CloudFront
- AWS re:Invent 2022 - Improve performance and availability with AWS Global Accelerator
- AWS re:Invent 2022 - Build your global wide area network using AWS
- AWS re:Invent 2020: Global traffic management with Amazon Route 53

Ejemplos relacionados:

- AWS Global Accelerator Custom Routing Workshop
- Handling Rewrites and Redirects using Edge Functions

PERF04-BP07 Optimización de 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 y optimizar las configuraciones de la misma.

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. Con las siguientes herramientas de la 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 dichas 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 AWS	Dónde se usa
Administrador de direcciones IP de Amazon VPC.	Utilice IPAM para planificar, seguir y supervisar las direcciones IP para sus cargas de trabajo de AWS y en las instalaciones. 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

Herramienta AWS	Dónde se usa
	en sus VPC. Con los registros de flujo de la VPC, puede diagnosticar reglas de grupos de seguridad excesivamente restrictivas o 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 flujo AWS Transit Gateway para recoger información sobre el tráfico IP que entra y sale de sus instancias de Transit Gateway.
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.
Analizador de accesibilidad	El Analizador de accesibilidad le ayuda a analizar y depurar la accesibilidad de la red. El Analizador de accesibilidad es una herramien ta de análisis de configuración que le permite hacer pruebas de conectividad entre un recurso de origen y uno de destino en las VPC. Esta herramienta le ayuda a verificar que la configuración de su red coincida con la conectividad prevista.

Herramienta AWS	Dónde se usa
Analizador de acceso a la red	Puede utilizar el Analizador de acceso a la red para comprobar el acceso de la red a los recursos. Puede utilizar el Analizador de acceso a la red para especificar los requisitos de acceso a la red e identificar posibles rutas de red que no cumplan los requisitos especific ados. Al optimizar su configuración de red correspondiente, puede comprender y verificar el estado de su red y demostrar si su red en AWS cumple con sus requisitos de conformid ad.
Amazon CloudWatch	Utilice 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 de 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 esta con base en sus observaciones.

Herramienta AWS	Dónde se usa
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. El Administrador de redes 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 Registros de flujo de AWS Transit Gateway.
- 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. El Analizador de acceso a la red puede ayudarle a hacer esto.

Recursos

Documentos relacionados:

- Registro de consultas de DNS públicas
- ¿Qué es IPAM?
- ¿Qué es Analizador de accesibilidad?
- ¿Qué es Analizador de acceso a la red?
- Métricas de CloudWatch para sus VPC

- Optimize performance and reduce costs for network analytics with VPC Flow Logs in Apache Parquet format (Optimización del rendimiento y reducción de los costos de análisis de red con registros de flujo de VPC en formato Apache Parquet)
- Supervisión de la red global con métricas de Amazon CloudWatch
- Supervisar de forma continua el tráfico y los recursos de red

Videos 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:

- Talleres de redes de AWS
- Supervisión de la red de AWS
- 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 con el fin de 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 a 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 al usar 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 implementación: utilice una canalización de integración continua/implementación continua (CI/CD), (por ejemplo, el repositorio del código fuente, los sistemas de diseño, la implementación y la automatización de pruebas) para implementar su infraestructura. Esto le permite implementar de manera repetible, coherente y por un bajo costo mientras itera.
- Métricas bien definidas: configure y supervise métricas para capturar 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 recopilación de elementos no utilizados 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 implementación, lance automáticamente las pruebas de rendimiento después de que las pruebas de ejecución más rápida se hayan aprobado 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: debería crear una serie de scripts de prueba que repliquen trayectos de usuario sintéticos o pregrabados. Estos scripts deben ser idempotentes y no acoplados; podría necesitar incluir scripts de precalentamiento para rendir resultados válidos. En la medida de lo posible, sus scripts de prueba deben replicar el comportamiento de uso en producción. Puede utilizar soluciones de software o de software como servicio (SaaS) para generar la carga. Considere la posibilidad de utilizar soluciones de AWS Marketplace e instancias de spot, ya que pueden ser formas rentables de generar 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 Establecimiento de indicadores clave de rendimiento (KPI) para medir el estado y el rendimiento de la carga de trabajo
- PERF05-BP02 Uso de soluciones de supervisión para saber en qué áreas es más crítico el rendimiento
- PERF05-BP03 Definición de un proceso para mejorar el rendimiento de la carga de trabajo
- PERF05-BP04 Pruebas de carga de la carga de trabajo
- PERF05-BP05 Uso de la automatización para solucionar de forma proactiva los problemas relacionados con el rendimiento
- PERF05-BP06 Mantenimiento de la carga de trabajo y los servicios actualizados
- PERF05-BP07 Revisión de las métricas a intervalos regulares

PERF05-BP01 Establecimiento de 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.

Patrones comunes de uso no recomendados:

- 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 publican y comparten como datos de métricas estándar.
- No tiene definido un KPI cuantitativo (que se pueda medir).
- Los KPI no se corresponden con los objetivos o estrategias empresariales.

Beneficios de establecer 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 unos resultados empresariales satisfactorios. 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. Además, indican cómo 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 se podría usar 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 por medio de datos granulares en tiempo real y datos históricos como referencia. Además, debe crear paneles en los que se hagan cálculos de métricas sobre los datos de los KPI para obtener información sobre las operaciones y la utilización. Los KPI se deben documentar e incluir umbrales que respalden los objetivos y las estrategias de la empresa, además de asignarse a las métricas que se supervisen. Los KPI deberían revisitarse siempre que cambien los objetivos empresariales, las estrategias o los requisitos del usuario final.

Pasos para la implementación

- Identificación de las partes interesadas: identifique y documente las partes interesadas clave de la empresa, como los equipos de desarrollo y operación.
- Definición de los objetivos: trabaje con estas partes interesadas para definir y documentar los objetivos de su 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 costo, así como los objetivos empresariales, como, por ejemplo, la satisfacción del usuario.
- Revisión de las prácticas recomendadas del sector: revise las prácticas sectoriales recomendadas para identificar los KPI relevantes que se ajusten a los objetivos de su carga de trabajo.
- Identificación de las métricas: identifique las métricas que estén alineadas con los objetivos de su 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.
- Definición y documentación de los KPI: utilice las prácticas recomendadas del sector y los objetivos de su carga de trabajo para establecer los objetivos del KPI de su 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.
- Implementación de la supervisión: utilice herramientas de supervisión como Amazon CloudWatch o AWS Config para recopilar métricas y medir los KPI.
- Comunicación de los KPI de forma visual: utilice herramientas de panel como Amazon QuickSight para visualizar y comunicar los KPI a las partes interesadas.
- Análisis y optimización: revise y analice periódicamente las métricas para identificar las áreas de la carga de trabajo que deben mejorarse. Colabore con las partes interesadas para implementar estas mejoras.

- Revisita y refinamiento: 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:

- Documentación de CloudWatch
- Supervisión, registro y rendimiento de los AWS Partner
- Herramientas de observabilidad de AWS
- La importancia de los indicadores clave de rendimiento (KPI) para las migraciones a gran escala a la nube
- How to track your cost optimization KPIs with the KPI Dashboard
- Documentación de X-Ray
- Uso de paneles de Amazon CloudWatch
- KPI de QuickSight

Videos 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 QuickSight

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

Comprenda y detecte las áreas en las que un aumento de 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.

Patrones comunes de uso no recomendados:

- 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 establecer esta práctica recomendada: el conocimiento de las áreas críticas de rendimiento ayuda a los propietarios de la carga de trabajo a supervisar los KPI y a priorizar las mejoras de alto 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 fragmentados y particionados de forma deficiente. Identifique las áreas restringidas de la carga de trabajo mediante pruebas de carga o supervisión.

Para aumentar la eficiencia del rendimiento, comprenda 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 implementar 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.

Servicio	Dónde se usa
Amazon CloudWatch Real-User Monitoring (RUM)	Para capturar las métricas de rendimiento de las aplicaciones a partir de las sesiones reales de los usuarios en el cliente y del frontend.
AWS X-Ray	Para hacer un seguimiento del tráfico a través de las capas de la aplicación e identific ar la latencia entre los componentes y las dependencias. Utilice los mapas de servicios de X-Ray para ver las relaciones y la latencia entre los componentes de la carga de trabajo.
Información sobre rendimiento de Amazon Relational Database Service	Para ver las métricas de rendimiento de la base de datos e identificar las mejoras de rendimiento.
Amazon RDS Enhanced Monitoring	Para ver las métricas de rendimiento del sistema operativo de la base de datos.
Amazon DevOps Guru	Para detectar patrones operativos anormales de forma que pueda identificar los problemas operativos antes de que afecten a sus clientes.

- 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 hacen las pruebas:
  - Configure Canarios sintéticos de CloudWatch para imitar las actividades de los usuarios en el navegador mediante programación con expresiones de frecuencia o tareas cron de Linux y generar métricas coherentes a lo largo del tiempo.
  - Use la solución Pruebas de carga distribuidas de AWS para generar picos de tráfico o probar la carga de trabajo con la tasa de crecimiento prevista.

- 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 nuevas mejoras y los impactos en el rendimiento de su carga de trabajo.

Recursos

Documentos relacionados:

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

Videos 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
- Cliente web de Amazon CloudWatch RUM
- SDK de X-Ray para Python
- Pruebas de carga distribuidas en AWS

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

Defina 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.

Patrones comunes de uso no recomendados:

- 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 basada en métricas.

Beneficios de establecer esta práctica recomendada: al definir el proceso para hacer 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 restricciones en el rendimiento de su carga de trabajo. Documente las restricciones de rendimiento de la carga de trabajo para que sepa los tipos de innovación que puedan mejorarlo.

Pasos para la implementación

- Identificación de los KPI: identifique los KPI de rendimiento de su carga de trabajo tal como se describe en PERF05-BP01 Establecimiento de indicadores clave de rendimiento (KPI) para medir el estado y el rendimiento de la carga de trabajo para basar su carga de trabajo.
- Implementación de la supervisión: utilice herramientas de observabilidad de AWS para recopilar métricas de rendimiento y medir los KPI.

- Análisis: haga 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 Uso de 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.
- Validación de las mejoras: utilice entornos de pruebas o de preproducción para validar la eficacia de la estrategia.
- Implementación de 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.
- Revisita y ajuste: revise periódicamente su proceso de mejora del rendimiento para identificar las áreas que se puedan optimizar.

Recursos

Documentos relacionados:

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

Videos 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:

- AWS GitHub

PERF05-BP04 Pruebas de carga de la carga de trabajo

Haga 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.

Patrones comunes de uso no recomendados:

- Hace pruebas de partes individuales de su carga de trabajo, pero no de la carga completa.
- Hace pruebas de carga en una infraestructura que no es la misma que su entorno de producción.
- Solo hace pruebas de carga hasta su carga prevista y no más allá, para ayudar a prever dónde puede tener problemas en el futuro.
- Hace 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 ponga en marcha, ya que parece un evento de denegación de servicio.

Beneficios de establecer esta práctica recomendada: calcular el rendimiento en una prueba de carga le mostrará qué áreas se verán afectadas a medida que 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 son un proceso que permite medir el rendimiento de la carga de trabajo en la nube bajo condiciones realistas y 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 ponerse en marcha con versiones sintéticas o saneadas de los datos de producción (debe eliminarse la información confidencial o de identificación). Haga automáticamente pruebas de carga en la canalización de entrega y compare los resultados con los KPI y los umbrales predefinidos. Este proceso le permitirá seguir alcanzando el rendimiento requerido.

Pasos para la implementación

- Definición de los objetivos de la prueba: identifique los aspectos de desempeño de su carga de trabajo que desea evaluar, como el rendimiento y el tiempo de respuesta.

- Selección de una herramienta para hacer la prueba: elija y configure la herramienta para hacer la prueba de carga que se ajuste a su carga de trabajo.
- Configuración del entorno: configure el entorno de prueba en función de su entorno de producción. Puede usar los servicios de AWS para poner en marcha entornos a escala de producción y poner a prueba su arquitectura.
- Implementación de la supervisión: utilice herramientas de supervisión como Amazon CloudWatch para recopilar métricas de todos los recursos de su arquitectura. También puede recopilar y publicar métricas personalizadas.
- Definición de escenarios: defina los escenarios y los parámetros de las pruebas de carga (como la duración de la prueba y el número de usuarios).
- Pruebas de carga: lleve a cabo escenarios de prueba a escala. Utilice la Nube de AWS para probar la carga de trabajo y detectar las áreas en las que el escalado no se hace correctamente o no se produce de forma lineal. Por ejemplo, utilice instancias de spot para generar cargas a bajo costo y descubrir obstáculos antes que se experimenten en la producción.
- Análisis de 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.
- Documentación y comunicación de 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.
- Repetición continua: las pruebas de carga deben hacerse con periodicidad, especialmente después de un cambio o actualización del sistema.

Recursos

Documentos relacionados:

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

Videos 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
- Demo of Amazon CloudWatch Synthetics

Ejemplos relacionados:

- Pruebas de carga distribuidas en AWS

PERF05-BP05 Uso de 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 forma proactiva los problemas relacionados con el rendimiento.

Patrones comunes de uso no recomendados:

- 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.

Beneficios de establecer esta práctica recomendada: la corrección proactiva de las acciones de alarma permite al personal de asistencia centrarse en aquellos elementos que no son accionables automáticamente. De este modo, el personal de operaciones podrá gestionar todas las alarmas sin sentirse abrumado y concentrarse exclusivamente en las críticas.

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

Guía para la implementación

Use alarmas para activar acciones automatizadas y corregir los problemas siempre que sea posible. Escale 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 las implementaciones si los KPI están fuera de los valores esperados.

Implemente procesos que informen el rendimiento cuando la carga de trabajo esté en marcha. 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

- Identificación del flujo de trabajo de corrección: identifique y estudie si el problema de rendimiento puede solucionarse automáticamente. Utilice soluciones de supervisión de AWS, como Amazon CloudWatch o AWS X-Ray, que le permitan entender mejor la causa raíz del problema.
- Definición de un proceso de automatización: cree un plan y un proceso de corrección paso a paso que pueda utilizar para solucionar el problema automáticamente.
- Configure 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.
- Automatización de la corrección: utilice los servicios y las tecnologías de AWS para automatizar el proceso de corrección. Por ejemplo, Automatización de AWS Systems Manager proporciona una forma segura 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.
- Prueba del flujo de trabajo: pruebe el proceso de corrección automatizado en un entorno de preproducción.
- Implementación del flujo de trabajo: implemente la corrección automática en el entorno de producción.
- Elaboración de un manual de estrategias: elabore y documente un manual 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.
- Revisión y perfeccionamiento: evalúe periódicamente la eficacia del flujo de trabajo de corrección automatizado. Ajuste los eventos de inicio y la lógica de corrección si es necesario.

Recursos

Documentos relacionados:

- Documentación de CloudWatch
- Socios de AWS Partner Network de supervisión, registro y rendimiento
- Documentación de X-Ray
- 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

Videos 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 Mantenimiento de 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.

Patrones comunes de uso no recomendados:

- 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 la carga de trabajo.

Beneficios de establecer 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, haga pruebas de rendimiento para medir el impacto que tienen en el rendimiento de la carga de trabajo.

Pasos para la implementación

- Inventario de la carga de trabajo: haga un inventario del software y la arquitectura de su carga de trabajo e identifique los componentes que deben actualizarse.
- Identificación de 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 de novedades de AWS para ver los productos que se adapten a su componente de carga de trabajo. Puede suscribirse a la fuente RSS o administrar sus suscripciones de correo electrónico.
- Definición de un calendario de actualización: establezca un calendario para evaluar nuevos servicios y características con su carga de trabajo.
  - Puede usar Inventario de AWS Systems Manager para recopilar los metadatos del sistema operativo (SO), las aplicaciones y los metadatos de instancias de sus instancias de Amazon EC2 y comprender rápidamente qué instancias están poniendo en marcha el software y las configuraciones requeridas por su política de software, así como las instancias que deben actualizarse.

- Evaluación de la nueva actualización: entienda cómo actualizar los componentes de su 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.
- Uso de la automatización: utilice la automatización del proceso de actualización a fin de reducir el nivel de esfuerzo para implementar nuevas funciones y limitar los errores causados por los procesos manuales.
  - Puede usar CI/CD para actualizar automáticamente las AMI, las imágenes de contenedor y otros artefactos relacionados con la aplicación en la nube.
  - Puede utilizar herramientas como AWS Systems Manager Patch Manager para automatizar el proceso de actualizaciones del sistema y programar la actividad mediante Ventanas de mantenimiento de AWS Systems Manager.
- Documentación del 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

Videos 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
- One Observability Workshop

PERF05-BP07 Revisión de las métricas a intervalos regulares

Revise qué métricas se recopilan 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 cuáles otras, en caso de que se les haga un seguimiento, podrían ayudar a identificar, abordar o prevenir problemas.

Patrones comunes de uso no recomendados:

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

Beneficios de establecer esta práctica recomendada: revise continuamente las métricas que se recopilan para verificar que puedan identificar, abordar o prevenir problemas correctamente. 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 métricas fueron útiles para abordar el problema y cuáles podrían haber ayudado, pero no se les da seguimiento 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 métricas fueron útiles para abordar el problema y cuáles podrían haber ayudado, pero no se les da seguimiento actualmente. Utilícelo para mejorar la calidad de la métrica que recopila, de modo que pueda prevenir o resolver más rápidamente incidentes futuros.

Pasos para la implementación

- Definición de las métricas: defina las métricas de rendimiento críticas para supervisar que estén adaptadas al objetivo de su carga de trabajo. Esto incluye métricas como el tiempo de respuesta y la utilización de los recursos.

- Establecimiento de 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.
- Configuración de una cadencia: establezca una cadencia (como semanal o mensual) para revisar las métricas críticas.
- Identificación de los problemas de rendimiento: durante cada revisión, evalúe las tendencias y la desviación de 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.
- Identificación de las acciones correctivas: utilice su análisis para identificar las acciones correctivas. Entre dichas medidas se pueden incluir el ajuste de parámetros, la corrección de errores y el escalado de los recursos.
- Documentación de los resultados: documente sus resultados, incluidos los problemas identificados, las causas raíz y las acciones correctivas.
- Iteración y mejora: 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:

- Documentación de CloudWatch
- Recopilación de métricas y registros de instancias de Amazon EC2 y en los servidores en las instalaciones con el agente de CloudWatch
- Consulte sus métricas con Información de métricas de CloudWatch
- Socios de AWS Partner Network de supervisión, registro y rendimiento
- Documentación de X-Ray

Videos 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 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 le ayuda a 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 garantizará que el rendimiento de su arquitectura no se degrade con el tiempo.

AWS se esfuerza por ayudarle a diseñar arquitecturas que rinden con eficiencia a la vez que entregan valor comercial. Utilice las herramientas y técnicas que se tratan en este documento para garantizar el éxito.

Colaboradores

Las siguientes personas y organizaciones han colaborado en 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

Documentación adicional

Para obtener más ayuda, consulte estas fuentes:

- Marco de AWS Well-Architected
- Centro de arquitectura de AWS

Revisiones del documento

Para recibir notificaciones sobre las actualizaciones de este documento técnico, suscríbase a la fuente RSS.

Cambio	Descripción	Fecha
Se han llevado a cabo actualizaciones menores en las prácticas recomendadas	PERF03-BP04 se ha actualizado con nuevas recomendaciones de servicio.	6 de noviembre de 2024
Actualización de las directrices de prácticas recomendadas	Varias pequeñas actualizaciones en todo el pilar.	27 de junio de 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 mejores prácticas son la selección de la arquitectura, la computación y el hardware, la administración de datos, las redes y la entrega de contenido, y los procesos y la cultura.	3 de octubre de 2023
Actualización menor	Eliminación del lenguaje no inclusivo.	13 de abril de 2023
Actualizaciones del nuevo marco	Se actualizaron las prácticas recomendadas con una guía prescriptiva y se agregaron nuevas prácticas recomendadas.	10 de abril de 2023

Documento técnico actualizado	Se actualizaron las prácticas recomendadas con una nueva guía de implementación.	15 de diciembre de 2022
Documento técnico actualizado	Se ampliaron las prácticas recomendadas y se agregaron planes de mejora.	20 de octubre de 2022
Actualización menor	Se eliminó el lenguaje no inclusivo.	22 de abril de 2022
Actualizaciones menores	Se actualizaron los enlaces.	10 de marzo de 2021
Actualizaciones menores	Se cambió el tiempo de espera de AWS Lambda a 900 segundos y se ha corregido el nombre de Amazon Keyspaces (para Apache Cassandra).	5 de octubre de 2020
Actualización menor	Corrección del enlace que no funciona.	15 de julio de 2020
Actualizaciones del nuevo marco	Revisión importante y actualización de contenidos	8 de julio de 2020
Documento técnico actualizado	Actualización menor por problemas gramaticales	1 de julio de 2018
Documento técnico actualizado	El documento técnico se ha actualizado para reflejar los cambios en AWS	1 de noviembre de 2017
Publicación inicial	Publicación del Pilar de eficiencia del rendimiento: Marco de AWS Well-Arch itected.	1 de noviembre de 2016

Avisos

Es responsabilidad de los clientes realizar su propia evaluación independiente de la información que contiene este documento. El presente documento: (a) tiene sólo fines informativos, (b) representa las ofertas y prácticas actuales de los productos de AWS, que están sujetas a cambios sin previo aviso, y (c) no supone ningún compromiso ni garantía por parte de AWS y sus filiales, proveedores o licenciantes. Los productos o servicios de AWS se proporcionan “tal cual”, sin garantías, afirmaciones ni condiciones de ningún tipo, ya sean expresas o implícitas. Las responsabilidades y obligaciones de AWS con respecto a sus clientes se controlan mediante los acuerdos de AWS y este documento no forma parte ni modifica ningún acuerdo entre AWS y sus clientes.

Glosario de AWS

Para ver la terminología más reciente de AWS, consulte el Glosario de AWS en la Referencia de Glosario de AWS.