Unable to locate subtitle

AWS Well-Architected Framework

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

AWS Well-Architected Framework: ***Unable to locate subtitle***

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon. All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.

Table of Contents

Resumen e introducción 1
Introducción 1
Definiciones 2
Sobre la arquitectura 5
Principios de diseño generales 6
Los pilares del marco 8
Excelencia operativa 8
Principios de diseño 9
Definición 10
Prácticas recomendadas 11
Recursos 20
Seguridad 21
Principios de diseño 21
Definición 22
Prácticas recomendadas 23
Recursos 30
Fiabilidad 31
Principios de diseño 31
Definición 32
Prácticas recomendadas 33
Recursos 39
Eficiencia del rendimiento 39
Principios de diseño 39
Definición 40
Prácticas recomendadas 41
Recursos 46
Optimización de costes 47
Principios de diseño 47
Definición 48
Prácticas recomendadas 49
Recursos 55
Sostenibilidad 56
Principios de diseño 56
Definición 57

Prácticas recomendadas 58
El proceso de revisión 67
Conclusión 70
Colaboradores 71
Otra documentación 72
Revisiones del documento 73
Apéndice: preguntas y prácticas recomendadas 77
Excelencia operativa 77
Organización 77
Prepárese 115
Operación 186
Evolucionar 222
Seguridad 240
Aspectos básicos de seguridad 240
Identity and Access Management 261
Detección 315
Protección de la infraestructura 326
Protección de los datos 346
Respuesta ante incidentes 378
Seguridad de las aplicaciones 401
Fiabilidad 422
Fundamentos 423
Arquitectura de la carga de trabajo 465
Administración de cambios 512
Administración de errores 553
Eficiencia del rendimiento 661
Selección de arquitectura 661
Computación y hardware 675
Gestión de datos 693
Redes y entrega de contenido 720
Proceso y cultura 751
Optimización de costes 767
Práctica de administración financiera en la nube 767
Conocimiento del gasto y del uso 792
Recursos rentables 835
Administración de la demanda y suministro de recursos 875

Optimización a lo largo del tiempo 889
Sostenibilidad 897
Selección de regiones 898
Alineación con la demanda 900
Software y arquitectura 914
Almacenamiento 926
Hardware y servicios 947
Proceso y cultura 958
Avisos 966

AWS Well-Architected Framework

Fecha de publicación: 3 de octubre de 2023 (Revisiones del documento)

AWS Well-Architected Framework le ayuda a comprender las ventajas y desventajas de las decisiones que toma al crear sistemas en AWS. El uso del marco le permitirá conocer las prácticas recomendadas de arquitectura para diseñar y operar sistemas en la nube que sean fiables, seguros, eficaces, rentables y sostenibles.

Introducción

AWS Well-Architected Framework le ayuda a comprender las ventajas y desventajas de las decisiones que toma al crear sistemas en AWS. Mediante el uso del marco, podrá conocer las prácticas recomendadas de arquitectura para diseñar y operar cargas de trabajo en la Nube de AWS que sean seguras, fiables, eficaces, rentables y sostenibles. Proporciona una forma de medir sus arquitecturas de forma constante en función de las prácticas recomendadas y de identificar áreas que se puedan mejorar. El proceso para revisar una arquitectura representa una conversación constructiva sobre decisiones arquitectónicas y no se trata de un mecanismo de auditoría. Creemos que contar con sistemas de buena arquitectura aumenta en gran medida la probabilidad de éxito empresarial.

AWS Solutions Architects cuenta con años de experiencia en soluciones de diseño de arquitecturas para una amplia variedad de sectores empresariales y casos de uso. Hemos ayudado a diseñar y revisar miles de arquitecturas de clientes en AWS. A partir de dicha experiencia, hemos identificado las prácticas recomendadas y estrategias centrales para sistemas de diseño de arquitecturas en la nube.

AWS Well-Architected Framework documenta un conjunto de cuestiones fundamentales que le permiten comprender si una arquitectura específica se corresponde con las prácticas recomendadas de la nube. El marco proporciona un enfoque coherente para evaluar los sistemas frente a las cualidades que espera de los sistemas modernos basados en la nube y la solución necesaria para lograr dichas cualidades. A medida que AWS continúa evolucionando y que seguimos aprendiendo más al trabajar con nuestros clientes, continuaremos perfeccionando la definición de una buena arquitectura.

Este marco está destinado a aquellos que ocupan puestos en tecnología, como los directores de tecnología (CTO), arquitectos, desarrolladores y miembros del equipo de operaciones. Describe las prácticas recomendadas y las estrategias de AWS para diseñar y operar una carga de trabajo en la

nube y proporciona enlaces a más detalles sobre implementación y patrones arquitectónicos. Para obtener más información, consulte la página de inicio de AWS Well-Architected.

AWS también proporciona un servicio para revisar sus cargas de trabajo de forma gratuita. La herramienta de AWS Well-Architected (AWS WA Tool) es un servicio en la nube que proporciona un proceso coherente para que revise y mida su arquitectura con AWS Well-Architected Framework. AWS WA Tool proporciona recomendaciones para lograr que sus cargas de trabajo sean más fiables, seguras, eficientes y rentables.

Para ayudarle a aplicar la prácticas recomendadas, hemos creado los laboratorios de AWS Well-Architected, que le proporcionan un repositorio de código y documentación para brindarle experiencia práctica en la implementación de las prácticas recomendadas. También colaboramos con socios selectos de la Red de socios de AWS (APN), que son miembros del programa para socios de AWS Well-Architected. Estos socios de AWS tienen exhaustivos conocimientos sobre AWS y pueden ayudarle a revisar y mejorar sus cargas de trabajo.

Definiciones

Cada día, los expertos de AWS ayudan a los clientes con la arquitectura de sistemas para aprovechar las prácticas recomendadas en la nube. Trabajamos con usted para lograr compensaciones arquitectónicas a medida que sus diseños evolucionan. A medida que implementa estos sistemas en entornos en directo, descubrimos el excelente rendimiento de dichos sistemas y las consecuencias de dichas compensaciones.

Con lo aprendido, hemos creado AWS Well-Architected Framework, que proporciona un conjunto coherente de prácticas recomendadas para que clientes y socios evalúen arquitecturas y un conjunto de preguntas que puede utilizar para evaluar en qué medida una arquitectura está alineada con las prácticas recomendadas de AWS.

AWS Well-Architected Framework se basa en seis pilares: excelencia operativa, seguridad, fiabilidad, eficiencia del rendimiento, optimización de costes y sostenibilidad.

Tabla 1. Los pilares de AWS Well-Architected Framework

Nombre	Descripción
Excelencia operativa	Capacidad de apoyar el desarrollo y ejecutar cargas de trabajo eficazmente, conocer sus operaciones y mejorar continuamente los

Nombre	Descripción
	procesos y procedimientos de soporte para ofrecer valor empresarial.
Seguridad	El pilar de seguridad describe cómo sacar partido de las tecnologías de nube para proteger datos, sistemas y recursos de una forma que pueda mejorar su nivel de seguridad .
Fiabilidad	El pilar de fiabilidad abarca la capacidad de una carga de trabajo para realizar su función prevista de forma correcta y coherente cuando se espera que lo haga. Esto incluye la capacidad de utilizar y probar la carga de trabajo a lo largo de todo su ciclo de vida. En este documento se incluye orientación de prácticas recomendadas para la implement ación de cargas de trabajo fiables en AWS.
Eficiencia del rendimiento	Es la capacidad de utilizar de forma eficaz los recursos informáticos para satisfacer los requisitos del sistema, así como de mantener la eficiencia a medida que la demanda cambia y las tecnologías evolucionan.
Optimización de costes	Capacidad de ejecutar sistemas para ofrecer valor empresarial al menor precio posible.
Sostenibilidad	Es la capacidad de mejorar constantemente el impacto en la sostenibilidad mediante la reducción del consumo de energía y el aumento de la eficiencia en todos los componentes de una carga de trabajo, maximizando los beneficios de los recursos aprovisionados y minimizando el número total de recursos necesarios.

En AWS Well-Architected Framework, usamos estos términos:

- Un componente es el código, la configuración y los recursos de AWS que cumplen con un requisito de forma conjunta. Un componente suele ser la unidad de responsabilidad técnica y está desacoplado de otros componentes.
- El término carga de trabajo se usa para identificar un grupo de componentes que, en conjunto, proporciona valor de negocio. Una carga de trabajo suele ser el nivel de detalle sobre el que hablan los líderes tecnológicos y comerciales.
- Pensamos en la arquitectura como la forma en que los componentes trabajan conjuntamente en una carga de trabajo. La forma en la que interactúan y se comunican los componentes es, a menudo, el foco de los diagramas de arquitectura.
- Los hitos marcan los cambios clave en su arquitectura a medida que evoluciona a lo largo del ciclo de vida del producto (diseño, implementación, prueba, lanzamiento y producción).
- Dentro de una organización, la cartera tecnológica es el conjunto de cargas de trabajo necesarias para que opere la empresa.
- El nivel de esfuerzo consiste en categorizar la cantidad de tiempo, esfuerzo y complejidad que requiere la implementación de una tarea. Cada organización tiene que considerar el tamaño y la experiencia del equipo y la complejidad de la carga de trabajo como contexto adicional a fin de determinar correctamente el nivel de esfuerzo de la organización.
  - Alto: el trabajo podría llevar varias semanas o meses. Esto podría desglosarse en múltiples historias, versiones y tareas.
  - Medio: el trabajo podría llevar varios días o semanas. Esto podría desglosarse en múltiples versiones y tareas.
  - Bajo: el trabajo podría llevar varias horas o días. Esto podría desglosarse en múltiples tareas.

Al diseñar cargas de trabajo, se hacen concesiones entre pilares según el contexto empresarial. Estas decisiones de negocios pueden impulsar sus prioridades de ingeniería. Podría optimizarlas para mejorar el impacto en la sostenibilidad y reducir los costes en detrimento de la fiabilidad en los entornos de desarrollo o, si se trata de soluciones fundamentales, podría optimizar la fiabilidad con incremento de costes e impacto en la sostenibilidad. En las soluciones de comercio electrónico, el rendimiento puede afectar a los ingresos y a la tendencia de los clientes a comprar. La seguridad y la excelencia operativa generalmente no se negocian contra los otros pilares.

Sobre la arquitectura

En entornos locales, los clientes suelen contar con un equipo central dedicado a la arquitectura tecnológica que está por encima de otros equipos de productos o características para verificar que sigan las prácticas recomendadas. En los equipos de arquitectura tecnológica suelen haber distintos roles como el de arquitecto técnico (infraestructura), arquitecto de soluciones (software), arquitecto de datos, arquitecto de redes y arquitecto de seguridad. A menudo, estos equipos usan TOGAF o el Zachman Framework como parte de una capacidad de arquitectura empresarial.

En AWS, preferimos distribuir capacidades dentro de los equipos en lugar de tener un equipo centralizado con esa capacidad. Existen riesgos cuando se elige distribuir la autoridad de la toma de decisiones, por ejemplo, al verificar que los equipos cumplan con los estándares internos. Mitigamos estos riesgos de dos maneras. En primer lugar, tenemos prácticas (formas de hacer las cosas, procesos, estándares y normas aceptadas) que se centran en permitir que cada equipo tenga esa capacidad, además de que contamos con expertos que verifican que los equipos eleven el nivel de los estándares que deben cumplir. En segundo lugar, implementamos mecanismos que realizan comprobaciones automáticas para verificar que se cumplan los estándares.

> «Las buenas intenciones nunca funcionan; hacen falta buenos mecanismos para que todo suceda», Jeff Bezos.

Esto significa reemplazar los esfuerzos humanos por mecanismos (a menudo automáticos) que controlan el cumplimiento de las normas y los procesos. Este enfoque distribuido está respaldado por los principios de liderazgo de Amazony establece una cultura en todos los roles que funciona a partir de las necesidades del cliente. Pensar en el cliente es esencial en nuestro proceso de innovación. Comenzamos con el cliente y lo que quiere, y dejamos que eso defina y guíe nuestros esfuerzos. Los equipos centrados en el cliente crean productos como respuesta a una necesidad del cliente.

Para la arquitectura, esto significa que esperamos que cada equipo tenga la capacidad de crear arquitecturas y seguir las prácticas recomendadas. Para ayudar a los nuevos equipos a obtener estas capacidades o a los equipos existentes a subir el listón, activamos el acceso a una comunidad virtual de ingenieros principales que pueden revisar sus diseños y ayudarlos a comprender cuáles son las prácticas recomendadas de AWS. La comunidad de ingenieros trabaja para que las prácticas recomendadas sean visibles y accesibles. Una forma de hacerlo es, por ejemplo, a través de charlas a la hora del almuerzo centradas en la aplicación de las prácticas recomendadas a casos reales.

Estas charlas se graban y pueden utilizarse como parte de los materiales de incorporación para los nuevos miembros del equipo.

Las prácticas recomendadas de AWS surgen de nuestra experiencia con miles de sistemas a escala de Internet. Preferimos usar datos para definir las prácticas recomendadas, pero también usamos expertos en la materia, como ingenieros principales, para establecerlas. A medida que los ingenieros principales ven emerger nuevas prácticas recomendadas, trabajan como una comunidad para verificar que los equipos las sigan. Con el tiempo, estas prácticas recomendadas se formalizan en nuestros procesos de revisión interna, y también en mecanismos que garantizan su cumplimiento. El Well-Architected Framework representa la implementación orientada al cliente de nuestro proceso de revisión interna, donde hemos codificado la forma de pensar de nuestros ingenieros principales en roles de campo, como los de la arquitectura de soluciones y de los equipos de ingeniería internos. El Well-Architected Framework es un mecanismo escalable que le permite aprovechar estos aprendizajes.

Al seguir el enfoque de una comunidad de ingeniería con propiedad distribuida de la arquitectura, creemos que puede surgir una arquitectura empresarial Well-Architected impulsada por las necesidades del cliente. Los líderes tecnológicos (como los CTO o los administradores de desarrollo) que realicen revisiones Well-Architected en todas sus cargas de trabajo podrán comprender mejor los riesgos de su cartera de tecnología. Con este enfoque, puede identificar temas en todos los equipos que su organización podría abordar mediante mecanismos, formaciones o charlas a la hora del almuerzo donde los ingenieros principales pueden compartir sus ideas sobre áreas específicas con múltiples equipos.

Principios de diseño generales

El Marco de Buena Arquitectura identifica un conjunto de principios generales de diseño que facilitan un buen diseño en la nube:

- No conjeture más sobre la capacidad que necesita: si opta por poca capacidad al implementar una carga de trabajo, puede terminar con recursos inactivos caros o lidiando con las implicaciones de rendimiento de una capacidad limitada. Con los servicios informáticos en la nube, estos problemas pueden desaparecer. Puede usar tanta capacidad como necesite y escalar automáticamente hacia arriba y hacia abajo.
- Ponga a prueba sistemas a escala de producción: en la nube, puede crear un entorno de prueba a escala de producción bajo demanda, completar las pruebas y retirar los recursos. Debido a que solo paga por el entorno de prueba cuando se ejecuta, puede simular su entorno en directo por una fracción del coste de las pruebas en las instalaciones.

- Automatice teniendo en cuenta la experimentación arquitectónica: la automatización le permite crear y replicar sus cargas de trabajo a bajo coste, además de evitarle los gastos generados por el esfuerzo manual. Puede rastrear cambios en su automatización, auditar su impacto y volver a los parámetros anteriores cuando sea necesario.
- Considere la posibilidad de usar arquitecturas evolutivas: en un entorno tradicional, las decisiones arquitectónicas a menudo se implementan como eventos estáticos puntuales, y solo se desarrollan algunas versiones principales de un sistema durante su vida útil. A medida que una empresa y su contexto evolucionan, estas decisiones iniciales pueden dificultar la capacidad del sistema para cumplir con los requisitos cambiantes de la empresa. En la nube, la capacidad de automatizar y probar bajo demanda reduce el riesgo de impacto de los cambios en el diseño. Esto permite que los sistemas evolucionen con el paso del tiempo para que las empresas puedan aprovechar las innovaciones como una práctica habitual.
- Impulse arquitecturas mediante el uso de datos: en la nube, puede recopilar datos sobre cómo sus decisiones arquitectónicas afectan al comportamiento de la carga de trabajo. Esto le permite tomar decisiones basadas en hechos sobre cómo mejorar su carga de trabajo. Su infraestructura en la nube es código, por lo que pueden utilizar esos datos para notificar sus elecciones de arquitectura y mejoras a lo largo del tiempo.
- Mejora mediante días de juego: ponga a prueba el rendimiento de su arquitectura y sus procesos al programar periódicamente días de juego para simular eventos en producción. Esto ayudará a comprender dónde se pueden realizar mejoras y a desarrollar la experiencia organizacional en la gestión de eventos.

Los pilares del marco

Crear un sistema de software se asemeja mucho a construir un edificio. Si los cimientos son endebles, pueden producirse problemas estructurales que minen la integridad y el funcionamiento del edificio. Al diseñar soluciones tecnológicas, si descuida los seis pilares de excelencia operativa, seguridad, fiabilidad, eficiencia del rendimiento, optimización de costes y sostenibilidad, puede resulta difícil crear un sistema que cumpla sus expectativas y requisitos. La incorporación de dichos pilares en su arquitectura ayudará a generar sistemas estables y eficientes. Esto le permitirá centrarse en otros aspectos del diseño, como los requisitos funcionales.

Pilares

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

Excelencia operativa

El pilar de excelencia operativa incluye la capacidad de promover el desarrollo y ejecutar cargas de trabajo de forma efectiva, comprender mejor sus operaciones y mejorar continuamente los procesos y procedimientos de soporte para aumentar el valor empresarial.

El pilar de excelencia operativa proporciona información general sobre los principios de diseño, prácticas recomendadas y preguntas. Encontrará una guía prescriptiva sobre implementación en el documento técnico Pilar de excelencia operativa.

Temas

- Principios de diseño
- Definición
- Prácticas recomendadas
- Recursos

Principios de diseño

A continuación, se presentan los principios de diseño para la excelencia operativa en la nube:

- Llevar a cabo operaciones como código: en la nube, puede aplicar la misma disciplina de ingeniería que usa para el código de aplicación a todo el entorno. Puede definir toda su carga de trabajo (aplicaciones, infraestructura, etc.) como código y actualizarla con código. Puede secuenciar los procedimientos operativos y automatizar su proceso al presentarlos como respuesta a eventos. Al llevar a cabo operaciones como código, limita los errores humanos y crea respuestas coherentes a los eventos.
- Realizar cambios frecuentes, pequeños y que pueda revertir: diseñe cargas de trabajo que sean escalables y tengan acoplamiento flexible para permitir que los componentes se actualicen con regularidad. Las técnicas de despliegue automatizadas, junto con cambios incrementales más pequeños, reducen el radio de repercusión y permiten revertir los cambios más rápido cuando se producen fallos. Esto aumenta la confianza para realizar cambios beneficiosos en su carga de trabajo y, al mismo tiempo, se mantiene la calidad y es posible adaptarse rápidamente a los cambios en las condiciones del mercado.
- Refinar los procedimientos de operaciones con frecuencia: a medida que evolucione sus cargas de trabajo, evolucione también sus operaciones de la forma correspondiente. a medida que vaya usando los procedimientos operativos, busque oportunidades para mejorarlos. Realice revisiones regulares y valide que todos los procedimientos sean efectivos y que los equipos estén familiarizados con ellos. Cuando se identifiquen lagunas, actualice los procedimientos en consecuencia. Comunique las actualizaciones de los procedimientos a todas las partes interesadas y equipos. Gamifique sus operaciones para compartir las prácticas recomendadas y formar a los equipos.
- Prever los errores: lleve a cabo ejercicios pre-mortem para identificar posibles fuentes de error a fin de poder eliminarlas o mitigarlas. Ponga a prueba las situaciones en las que se produzca un error y confirme que entiende su impacto. Ponga a prueba los procedimientos de respuesta para garantizar su eficacia, así como para asegurarse de que los equipos conocen su proceso. Configure días de juego habituales para poner a prueba la carga de trabajo y las respuestas del equipo ante eventos simulados.
- Aprender de los errores operativos: impulse las mejoras gracias a las lecciones que se aprendan después de todos los eventos operativos y errores. Comparta las enseñanzas con los equipos y con toda la organización.

- Utilizar servicios administrados: reduzca la carga operativa mediante el uso de servicios administrados de AWS siempre que sea posible. Desarrolle procedimientos operativos en torno a las interacciones con esos servicios.
- Implementar la observabilidad para obtener información práctica: conozca por completo el comportamiento, el rendimiento, el grado de fiabilidad, el coste y el estado de la carga de trabajo. Establezca indicadores clave de rendimiento (KPI) y utilice la telemetría de observabilidad para tomar decisiones informadas y medidas rápidas cuando los resultados empresariales estén en riesgo. Mejore proactivamente el rendimiento, la fiabilidad y el coste en función de datos de observabilidad procesables.

Definición

Hay cuatro áreas de prácticas recomendadas para la excelencia operativa en la nube:

- Organización
- Prepárese
- Operación
- Evolución

La dirección de la organización define los objetivos empresariales. La organización debe comprender los requisitos y prioridades, y usarlos para organizar y llevar a cabo su trabajo para lograr los objetivos de la empresa. Su carga de trabajo debe emitir la información necesaria para apoyarlos. Al implementar servicios para conseguir la integración, el despliegue y la entrega de su carga de trabajo, creará un flujo creciente de cambios positivos para la producción al automatizar los procesos repetitivos.

Puede haber riesgos inherentes a la operativa de la carga de trabajo. Debe comprender dichos riesgos y tomar una decisión informada para iniciar la producción. Sus equipos deben poder prestar asistencia a su carga de trabajo. Las métricas empresariales y operativas derivadas de los resultados empresariales deseados le permitirán entender el estado de su carga de trabajo y sus actividades de operaciones, así como responder a los incidentes. Las prioridades cambiarán a medida que cambien sus necesidades empresariales y su entorno empresarial. Úselas como referencia para introducir mejoras continuamente en su organización y en la operativa de la carga de trabajo.

Prácticas recomendadas

Temas

- Organización
- Prepárese
- Operación
- Evolución

Organización

Sus equipos deben disponer de un entendimiento compartido de toda la carga de trabajo, su rol en ella y los objetivos empresariales compartidos para establecer las prioridades que permitan conseguir el éxito empresarial. Unas prioridades bien definidas maximizarán los beneficios de sus esfuerzos. Evalúe las necesidades de los clientes internos y externos, e involucre a las partes interesadas clave, incluidos los equipos de negocios, desarrollo y operaciones, para determinar dónde se deben centrar los esfuerzos. La evaluación de las necesidades de los clientes le permitirá verificar que tiene una comprensión profunda de la asistencia que se necesita para lograr los resultados empresariales. Verifique que conoce las directrices y las obligaciones definidas por la gobernanza de su organización y los factores externos, tales como los requisitos normativos de cumplimiento y los estándares del sector, para asegurarse de que pueda exigir o aplicar un enfoque específico. Valide la existencia de mecanismos para identificar cambios en la gobernanza interna y en los requisitos de cumplimiento externos. Si no existen dichos requisitos, asegúrese de haber realizado una investigación exhaustiva para tomar las decisiones pertinentes. Revise sus prioridades con regularidad para poder actualizarlas a medida que cambien las necesidades.

Evalúe las amenazas a la empresa (por ejemplo, riesgos y responsabilidades empresariales, amenazas a la seguridad de la información) y mantenga dicha información en un registro de riesgos. Evalúe el impacto de los riesgos y las compensaciones entre los intereses opuestos o los enfoques alternativos. Por ejemplo, la aceleración de la velocidad de comercialización de las nuevas funciones puede primar sobre la optimización de los costes, o se puede elegir una base de datos relacional para los datos no relacionales para simplificar el esfuerzo de migración de un sistema. Gestione los beneficios y los riesgos para tomar decisiones fundamentadas a la hora de determinar dónde centrar sus esfuerzos. Algunos riesgos u opciones son aceptables durante un tiempo, incluso se podrían mitigar los riesgos asociados, pero también podría ser inaceptable permitir que un riesgo persista, en cuyo caso se tomarán medidas para abordarlo.

Sus equipos deben comprender su papel en la consecución de los resultados empresariales. Los equipos deben comprender el rol que desempeñan en el éxito de otros equipos, así como el que desempeñan los demás equipos en su propio éxito, además de tener objetivos en común. Comprender la responsabilidad, la propiedad, cómo se toman las decisiones y quién tiene autoridad para tomarlas ayudará a centrar los esfuerzos y a maximizar los beneficios de sus equipos. Las necesidades de un equipo se verán determinadas por el cliente al que dan soporte, su organización, la composición del equipo y las características de su carga de trabajo. No es razonable esperar que un único modelo operativo sea capaz de respaldar a todos los equipos y sus cargas de trabajo en la organización.

Verifique que se haya identificado a los encargados de cada aplicación, carga de trabajo, plataforma y componente de la infraestructura, y de que cada proceso y procedimiento disponga de un encargado responsable de su definición y de encargados responsables de su rendimiento.

Las acciones de los miembros del equipo se fundamentarán en la comprensión del valor empresarial de cada componente, proceso y procedimiento, el motivo por el cual se establecieron los recursos o se realizan determinadas actividades, y la razón por la que esa propiedad existe. Defina claramente las responsabilidades de los miembros del equipo para que puedan actuar de forma adecuada y disponer de mecanismos para identificar la responsabilidad y la propiedad. Cuente con mecanismos para solicitar adiciones, cambios y excepciones para no limitar la innovación. Defina acuerdos entre los equipos que describan el trabajo conjunto para darse apoyo entre sí y respaldar los resultados de la empresa.

Preste asistencia a los miembros de su equipo para que puedan ser más eficaces a la hora de actuar y apoyar los resultado empresariales. Los líderes comprometidos deben establecer expectativas y medir el éxito. Los directivos deben ser los patrocinadores, defensores e impulsores de la adopción de las prácticas recomendadas y de la evolución de la organización. Deje que los miembros del equipo actúen cuando los resultados corran algún riesgo para, así, minimizar el impacto, y anímelos a realizar escalamientos hacia los responsables de la toma de decisiones y las partes interesadas cuando crean que exista un riesgo, de manera que se pueda abordar y se eviten incidentes. Proporcione una comunicación oportuna, clara y procesable de los riesgos conocidos y de los eventos planificados para que los miembros del equipo puedan reaccionar de forma oportuna y adecuada.

Fomente la experimentación para acelerar el aprendizaje y mantener a los miembros del equipo interesados y comprometidos. Los equipos deben aumentar el conjunto de habilidades para adoptar nuevas tecnologías y para apoyar los cambios en la demanda y las responsabilidades. Debe apoyar y fomentar esto ofreciendo un horario estructurado dedicado a la formación. Verifique que los

miembros del equipo dispongan de los recursos (herramientas y miembros del equipo) para lograr el éxito y escalar con el fin de lograr los resultados empresariales. Aproveche la diversidad entre las organizaciones para buscar múltiples perspectivas únicas. Utilice esta perspectiva para aumentar la innovación, cuestionar sus suposiciones y reducir el riesgo de sesgo de confirmación. Fomente la inclusión, la diversidad y la accesibilidad en sus equipos para obtener perspectivas beneficiosas.

Si existen requisitos externos de regulación o conformidad que se aplican a su organización, debe utilizar los recursos proporcionados por el Centro de cumplimiento en la nube de AWS para ayudar a formar a sus equipos a fin de que puedan considerar el impacto en sus prioridades. Well-Architected Framework hace hincapié en aprender, medir y mejorar. Proporciona un enfoque coherente para evaluar las arquitecturas e implementar diseños que se escalarán con el tiempo. AWS proporciona la AWS Well-Architected Tool para ayudarle a revisar su enfoque antes del desarrollo, el estado de sus cargas de trabajo antes de la producción y el estado de sus cargas de trabajo durante la producción. Puede comparar las cargas de trabajo de las prácticas recomendadas de arquitectura de AWS más recientes, supervisar su estado global y obtener información sobre los riesgos potenciales. AWS Trusted Advisor es una herramienta que proporciona acceso a un conjunto básico de comprobaciones que recomiendan optimizaciones que pueden ayudar a definir sus prioridades. Los clientes de Business y Enterprise Support reciben acceso a comprobaciones adicionales centradas en la seguridad, la fiabilidad, el rendimiento, la optimización de los costes y la sostenibilidad que pueden ayudar a configurar sus prioridades.

AWS puede ayudarle a formar a sus equipos sobre AWS y sus servicios para aumentar su comprensión de cómo sus elecciones pueden tener un impacto en su carga de trabajo. Debe utilizar los recursos proporcionados por AWS Support (Centro de conocimiento de AWS, Foros de debate de AWS y Centro de AWS Support) y Documentación de AWS para formar a sus equipos. En caso de tener dudas sobre AWS, contacte con AWS Support a través del Centro de AWS Support. AWS también comparte los patrones y prácticas recomendadas que hemos aprendido a través del funcionamiento de AWS en la Amazon Builders' Library. Hay una gran variedad de información útil disponible a través del Blog de AWS y el Podcast oficial de AWS. AWS Training and Certification ofrece una formación gratuita a través de cursos digitales autoguiados sobre los conceptos básicos de AWS. También puede inscribirse en una formación adicional impartida por un instructor para respaldar el desarrollo de las habilidades de AWS de sus equipos.

Utilice herramientas o servicios que le permitan controlar de forma centralizada sus entornos en todas las cuentas, como, por ejemplo, AWS Organizations, para ayudarle a administrar los modelos operativos. Los servicios como AWS Control Tower amplían esta capacidad de administración al permitirle definir esquemas (que respaldan sus modelos operativos) para la configuración de las cuentas, aplicar una gobernanza continua mediante AWS Organizations y automatizar el

aprovisionamiento de nuevas cuentas. Los proveedores de servicios administrados tales como AWS Managed Services, los socios de AWS Managed Services o los proveedores de servicios administrados en la Red de socios de AWS proporcionan experiencia en la implementación de entornos en la nube y respaldan sus requisitos de seguridad y cumplimiento y sus objetivos empresariales. La incorporación de los servicios administrados a su modelo operativo puede ahorrarle tiempo y recursos, y le permite mantener a sus equipos internos centrados en los resultados estratégicos que diferenciarán a su empresa, en lugar de desarrollar nuevas competencias y capacidades.

Las siguientes preguntas se centran en estas consideraciones sobre la excelencia operativa. (Para ver una lista de preguntas y prácticas recomendadas sobre la excelencia operativa, consulte el Appendix).

OPS 1: How do you determine what your priorities are?
Everyone must understand their part in achieving business success. Have shared goals in order to set priorities for resources. This will maximize the benefits of your efforts.

OPS 2: How do you structure your organization to support your business outcomes?
Your teams must understand their part in achieving business outcomes. Teams must understand their roles in the success of other teams, the role of other teams in their success, and have shared goals. Understanding responsibility, ownership, how decisions are made, and who has authority to make decisions will help focus efforts and maximize the benefits from your teams.

OPS 3: How does your organizational culture support your business outcomes?
Provide support for your team members so that they can be more effective in taking action and supporting your business outcome.

Es posible que, en algún momento, quiera hacer énfasis en un pequeño subconjunto de prioridades. Utilice un enfoque equilibrado a largo plazo para verificar el desarrollo de las capacidades necesarias y la administración de riesgos. Revise las prioridades con regularidad y actualícelas a medida que cambien las necesidades. Cuando la responsabilidad y la propiedad no están definidas o se

desconocen, se corre el riesgo, tanto de no actuar a tiempo, como de que se hagan esfuerzos repetidos y potencialmente conflictivos para abordar dichas necesidades. La cultura organizativa tiene un impacto directo en la satisfacción laboral y la retención de los miembros del equipo. Estimule el compromiso y las capacidades de los miembros de su equipo para lograr el éxito de la empresa. La experimentación es necesaria para innovar y convertir las ideas en resultados. Debe saber que un resultado no deseado es un experimento exitoso que ha identificado un camino que no llevará al éxito.

Prepárese

Para prepararse para la excelencia operativa hay que entender las cargas de trabajo y sus comportamientos esperados. Entonces, podrá diseñarlas para que proporcionen información sobre su estado y crear los procedimientos para respaldarlas.

Diseñe la carga de trabajo para que proporcione la información necesaria para que pueda comprender el estado interno (por ejemplo, métricas, registros, eventos y rastreos) en todos los componentes en caso de problemas de investigación y observabilidad. La observabilidad va más allá de la simple supervisión, ya que proporciona una comprensión integral del funcionamiento interno de un sistema en función de sus resultados externos. Basada en métricas, registros y rastros, la observabilidad ofrece una visión profunda del comportamiento y la dinámica del sistema. Con una observabilidad eficaz, los equipos pueden discernir patrones, anomalías y tendencias, lo que les permite abordar de forma proactiva los posibles problemas y mantener un estado óptimo del sistema. La identificación de los indicadores clave de rendimiento (KPI) es fundamental para garantizar la alineación entre las actividades de supervisión y los objetivos empresariales. Esta alineación garantiza que los equipos tomen decisiones basadas en datos mediante la utilización de métricas que realmente sean relevantes, optimizando así tanto el rendimiento del sistema como los resultados empresariales. Además, la observabilidad permite que las empresas sean proactivas en lugar de reactivas. Los equipos pueden entender las relaciones de causa y efecto dentro de sus sistemas y predecir y prevenir los problemas en lugar de simplemente reaccionar ante ellos. A medida que las cargas de trabajo evolucionan, es esencial revisar y refinar la estrategia de observabilidad para garantizar que esta siga siendo relevante y eficaz.

Adopte enfoques que mejoren el flujo de cambios en la producción y que ayuden a la refactorización, a la respuesta rápida sobre la calidad y a la corrección de errores. Estos enfoques aceleran los cambios positivos que se introducen en la producción, limitan los problemas implementados y activan una rápida identificación y solución de los problemas introducidos a través de las actividades de despliegue o descubiertas en sus entornos.

Adopte enfoques que proporcionen una respuesta inmediata sobre la calidad y logren una recuperación rápida de los cambios que no muestran los resultados deseados. El uso de estas prácticas ayuda a mitigar el impacto de los problemas generados con la implementación de cambios. Planifique para hacer frente a los cambios fallidos para que pueda responder rápidamente si es necesario. Además, pruebe y valide los cambios que realice. Debe conocer las actividades planificadas en sus entornos para poder administrar el riesgo de que los cambios afecten a dichas actividades. Realice cambios frecuentes, pequeños y reversibles para limitar el alcance del cambio. Al hacerlo, los problemas se solucionan de forma más rápida con la opción de revertir un cambio. También significa que podrá beneficiarse de unos cambios valiosos de forma más frecuente.

Evalúe la disponibilidad operativa de la carga de trabajo, los procesos y procedimientos, y el personal para comprender los riesgos operativos relacionados con la carga de trabajo. Use un proceso coherente (que incluya listas de verificación manuales y automáticas) para saber cuándo una carga de trabajo o cambio estarán listos para lanzarse. Esto también le ayudará a detectar cualquier área para la que sea necesaria la elaboración de un plan de tratamiento. Debe disponer de runbooks que documenten las actividades rutinarias y guías de estrategias para aplicar los procesos de resolución de errores. Debe comprender los beneficios y los riesgos para tomar decisiones bien fundamentadas a fin de permitir que los cambios entren en la fase de producción.

AWS le permite ver toda su carga de trabajo (aplicaciones, infraestructura, política, gobernanza y operaciones) como código. Eso significa que puede aplicar la misma disciplina de ingeniería que usa para el código de las aplicaciones a cada elemento de su pila y compartirla entre los equipos u organizaciones para magnificar los beneficios de los esfuerzos de desarrollo. Use las operaciones como código en la nube y la capacidad de experimentar de manera segura para desarrollar la carga de trabajo, sus procedimientos operativos y poner en práctica los casos en los que se produzcan errores. Usar CloudFormation le permite tener entornos de producción, de pruebas y de desarrollo del entorno aislado coherentes y con formatos ya definidos, con un aumento de los niveles de control operativo.

Las siguientes preguntas se centran en estas consideraciones sobre la excelencia operativa.

OPS 4: How do you implement observability in your workload?

Implement observability in your workload so that you can understand its state and make data-driv en decisions based on business requirements.

OPS 5: How do you reduce defects, ease remediation, and improve flow into production?

Adopt approaches that improve flow of changes into production that achieve refactoring fast feedback on quality, and bug fixing. These accelerate beneficial changes entering production, limit issues deployed, and achieve rapid identification and remediation of issues introduced through deployment activities.

OPS 6: How do you mitigate deployment risks?

Adopt approaches that provide fast feedback on quality and achieve rapid recovery from changes that do not have desired outcomes. Using these practices mitigates the impact of issues introduce d through the deployment of changes.

OPS 7: How do you know that you are ready to support a workload?

Evaluate the operational readiness of your workload, processes and procedures, and personnel to understand the operational risks related to your workload.

Invierta en implementar actividades operativas como código para maximizar la productividad del personal de operaciones, minimizar las tasas de error y habilitar las respuestas automatizadas. Realice ensayos de fallas “pre-mortem” para anticipar el fracaso y crear procedimientos cuando sea apropiado. Aplique metadatos mediante etiquetas de registro y Grupos de recursos de AWS mediante una estrategia de etiquetado coherente para permitir la identificación de sus recursos. Etiquete sus recursos para la organización, la contabilidad de costes, los controles de acceso y el objetivo de ejecución de actividades de operaciones automatizadas. Adopte las prácticas de implementación que aprovechan la elasticidad de la nube a fin de facilitar las actividades de desarrollo y la implementación previa de sistemas para que la implementación sea más rápida. Cuando haga cambios en las listas de control que utiliza para evaluar sus cargas de trabajo, planifique lo que hará con los sistemas activos que ya no cumplen los requisitos.

Operación

La observabilidad le permite centrarse en datos significativos y comprender las interacciones y los resultados de su carga de trabajo. Al concentrarse en la información esencial y eliminar los datos

innecesarios, mantiene un enfoque sencillo para comprender el rendimiento de las cargas de trabajo. No solo es esencial recopilar datos, sino también interpretarlos correctamente. Defina puntos de referencia claros, establezca umbrales de alerta adecuados y supervise activamente cualquier desviación. Un cambio en una métrica clave, especialmente cuando se correlaciona con otros datos, puede identificar áreas problemáticas concretas. Con la observabilidad, está mejor preparado para prever y abordar los posibles desafíos, lo que garantiza que su carga de trabajo funcione sin problemas y satisfaga las necesidades empresariales.

El éxito operativo de una carga de trabajo se mide por los logros de los resultados del cliente y del negocio. Defina los resultados esperados, decida cómo se medirá el éxito e identifique las métricas que se usarán en los cálculos para determinar si su carga de trabajo y las operaciones se realizan con éxito. El estado de las operaciones incluye tanto el estado de la carga de trabajo como el éxito de las operaciones que se realizan para llevarlas a cabo (por ejemplo, la implementación y la respuesta frente a incidencias). Establezca puntos de referencia de métricas para las mejoras, la investigación y la intervención, y recopile y analice las métricas. A continuación, corrobore si comprende el éxito de las operaciones y cómo cambia con el tiempo. Utilice métricas recopiladas para determinar si satisface las necesidades del cliente y del negocio. Identifique también las áreas a mejorar.

Se requiere eficacia y eficiencia en la gestión de los eventos operativos para lograr excelencia operativa. Se aplica tanto a los eventos operativos planificados como a los no planificados. Utilice los runbooks establecidos para eventos bien conocidos y las guías de estrategia como ayuda para investigar y para resolver otros problemas. Priorice aquellos eventos que tengan mayor repercusión en el negocio y en el cliente. Verifique que, si se genera una alerta como respuesta a un evento, se ejecutará un proceso asociado con un encargado identificado de forma específica. Defina con antelación el personal necesario para resolver un evento e incluya procesos de escalamiento para que participe personal adicional, si es necesario, en función de la urgencia y el impacto. Identifique e implique a aquellos individuos que tengan autoridad para decidir sobre las acciones en aquellos casos en los que la respuesta a un evento que no se haya abordado previamente repercuta en el negocio.

Comunique el estado operativo de las cargas de trabajo mediante paneles y notificaciones adaptadas a la audiencia de destino (por ejemplo, cliente, negocio, desarrolladores, operaciones) para que puedan llevar a cabo las medidas adecuadas, gestionen sus expectativas y se les informe cuando se reanuden las operaciones habituales.

En AWS, puede generar vistas de panel de las métricas recopiladas a partir de cargas de trabajo y de AWS de forma nativa. Puede aprovechar CloudWatch o aplicaciones de terceros para añadir y

presentar vistas de la empresa, la carga de trabajo y las operaciones de las actividades operativas. AWS proporciona información sobre cargas de trabajo mediante capacidades de registro, como AWS X-Ray, CloudWatch, CloudTrail y registros de flujo de VPC para identificar problemas de las cargas de trabajo a fin de ofrecer apoyo a la hora de analizar y corregir la causa raíz.

Las siguientes preguntas se centran en estas consideraciones sobre la excelencia operativa.

OPS 8: How do you utilize workload observability in your organization?

Ensure optimal workload health by leveraging observability. Utilize relevant metrics, logs, and traces to gain a comprehensive view of your workload's performance and address issues efficient ly.

OPS 9: How do you understand the health of your operations?

Define, capture, and analyze operations metrics to gain visibility to operations events so that you can take appropriate action.

OPS 10: How do you manage workload and operations events?

Prepare and validate procedures for responding to events to minimize their disruption to your workload.

Todas las métricas que recopile deben estar alineadas con una necesidad empresarial y los resultados que apoyan. Desarrolle respuestas con scripts para los eventos bien conocidos y automatice su rendimiento en respuesta al reconocimiento del evento.

Evolución

Aprenda, comparta y mejore continuamente para mantener la excelencia operativa. Dedique ciclos de trabajo a hacer mejoras graduales de forma casi continua. Realice análisis posteriores al incidente de todos los eventos que afecten a los clientes. Identifique los factores que han contribuido a ello y actúe de forma preventiva para limitar o impedir que se repita. Comunique los factores que han contribuido a ello a las comunidades afectadas, según proceda. Evalúe y priorice las oportunidades

de mejora de forma gradual (por ejemplo, solicitudes de características, solución de problemas y requisitos de conformidad), entre ellos, los procedimientos operativos y de cargas de trabajo.

Incluya bucles de comentarios en los procedimientos para identificar rápidamente las áreas de mejora y recoger lo aprendido durante la ejecución de operaciones.

Comparta lo aprendido con los equipos para enseñar los beneficios de dichas lecciones. Analice las tendencias de las lecciones aprendidas y realice un análisis retrospectivo de las métricas de las operaciones entre equipos para identificar oportunidades y métodos de mejora. Aplique aquellos cambios que traigan consigo mejoras y evalúe los resultados para determinar el éxito.

En AWS, puede exportar los datos de registro a Amazon S3 o enviar registros directamente a Amazon S3 para un almacenamiento a largo plazo. Con AWS Glue, puede descubrir y preparar los datos de registro en Amazon S3 para realizar análisis y almacenar los metadatos asociados en AWS Glue Data Catalog. Amazon Athena, a través de su integración nativa con AWS Glue, se puede usar para analizar sus datos de registro, haciendo consultas con SQL estándar. Con una herramienta de inteligencia empresarial como Amazon Quick, puede visualizar, explorar y analizar sus datos. También puede descubrir las tendencias y los eventos de interés que pueden fomentar las mejoras.

La siguiente pregunta se centra en estas consideraciones acerca de la excelencia operativa.

OPS 11: How do you evolve operations?

Dedicate time and resources for nearly continuous incremental improvement to evolve the effectiveness and efficiency of your operations.

La correcta evolución de las operaciones se basa en cambios pequeños pero frecuentes; entornos seguros y tiempo para experimentar, desarrollar y probar mejoras, así como entornos en los que se anima a aprender a partir de los errores. La asistencia operativa en entornos de producción, pruebas, desarrollo y zonas de pruebas, con un nivel creciente de controles operativos, facilita el desarrollo y aumenta la predictibilidad de resultados exitosos a partir de los cambios que se implementen en la producción.

Recursos

Consulte los siguientes recursos para obtener más información sobre nuestras prácticas recomendadas de excelencia operativa.

Documentación

- DevOps y AWS

Documento técnico

- Pilar de excelencia operativa

Vídeo

- DevOps en Amazon

Seguridad

El pilar de seguridad engloba la capacidad de proteger datos, sistemas y activos para sacar partido de las tecnologías de nube con el fin de mejorar su nivel de seguridad.

El pilar de seguridad ofrece una visión general de principios de diseño, prácticas recomendadas y preguntas. Encontrará recomendaciones de implementación en el documento técnico Pilar de seguridad.

Temas

- Principios de diseño
- Definición
- Prácticas recomendadas
- Recursos

Principios de diseño

Existen siete principios de diseño de seguridad en la nube:

- Implemente sólidas bases de identidad: aplique el principio del privilegio mínimo y haga cumplir la separación de funciones con la autorización adecuada para cada interacción con los recursos de AWS. Centralice la administración de identidades y establézcase como objetivo eliminar la dependencia de las credenciales estáticas a largo plazo.

- Posibilite la trazabilidad: supervise, cree alertas y audite acciones y cambios en su entorno en tiempo real. Integre la recopilación de registros y métricas con sistemas para investigar y tomar medidas automáticamente.
- Implemente seguridad en todos las capas: aplique un enfoque de defensa exhaustivo con varios controles de seguridad. Impleméntelo en todas las capas (por ejemplo, red periférica, VPC, equilibrio de carga, cada instancia y servicio de computación, sistema operativo, aplicación y código).
- Automatice las prácticas recomendadas de seguridad: los mecanismos de seguridad automatizados basados en software mejoran la capacidad de escalar de forma segura, más rápida y más rentable. Cree arquitecturas seguras, como la implementación de controles definidos y administrados como código en plantillas controladas por versión.
- Cifre datos en tránsito y en reposo: clasifique sus datos en niveles de confidencialidad y utilice mecanismos como el cifrado, la tokenización y el control de acceso cuando corresponda.
- Mantenga a las personas alejadas de los datos: use mecanismos y herramientas para reducir o eliminar la necesidad de acceso directo o de procesamiento manual de datos. De esta forma, se reducen los errores humanos y el riesgo de una mala praxis o modificación al tratar con información confidencial.
- Prepárese para eventos de seguridad: prepárese para un incidente teniendo a su disposición procesos y políticas de investigación y administración de incidentes que se ajusten a los requisitos de su organización. Ejecute simulaciones de respuesta frente a incidencias y use herramientas con automatización para aumentar la velocidad de detección, investigación y recuperación.

Definición

Existen seis áreas de prácticas recomendadas para la seguridad en la nube:

- Seguridad
- Identity and Access Management
- Detección
- Protección de la infraestructura
- Protección de los datos
- respuesta frente a incidencias

Antes de diseñar una carga de trabajo, hay que adoptar prácticas que influyen en la seguridad. Debe controlar quién puede hacer qué. Además, debe poder identificar los incidentes de seguridad, proteger sus sistemas y servicios, y mantener la confidencialidad e integridad de los datos a través de la protección de datos. Debe tener un proceso ben definido y practicado para responder a los incidentes de seguridad. Estas herramientas y técnicas son importantes porque respaldan objetivos como la prevención de pérdidas económicas o la conformidad con las obligaciones reglamentarias.

El modelo de responsabilidad compartida de AWS permite que las organizaciones que adoptan la nube logren alcanzar sus metas de seguridad y cumplimiento. Puesto que AWS protege físicamente la infraestructura que sustenta nuestros servicios en la nube, como cliente de AWS puede centrarse en utilizar servicios para alcanzar sus metas. La nube de AWS también ofrece más acceso a los datos de seguridad y un enfoque automatizado para responder a los eventos de seguridad.

Prácticas recomendadas

Temas

- Seguridad
- Identity and Access Management
- Detección
- Protección de la infraestructura
- Protección de los datos
- Respuesta ante incidentes

Seguridad

Para utilizar la carga de trabajo de forma segura, debe adoptar prácticas recomendadas globales en cada área de seguridad. Lleve los requisitos y los procesos que ha definido en la excelencia operativa a un nivel de organización y carga de trabajo, y aplíquelo a todas las áreas.

Mantenerse al día con las recomendaciones de AWS y el sector y la inteligencia de amenazas le ayudan a desarrollar el modelo de amenaza y controlar los objetivos. La automatización de los procesos de seguridad, las pruebas y la validación le ayudan a escalar las operaciones de seguridad.

Las siguientes preguntas se centran en las consideraciones de seguridad. (Para ver una lista de preguntas y prácticas recomendadas sobre la seguridad, consulte el Apéndice).

SEC 1: ¿Cómo utiliza la carga de trabajo de forma segura?

Para utilizar la carga de trabajo de forma segura, debe adoptar prácticas recomendadas globales en cada área de seguridad. Lleve los requisitos y los procesos que ha definido en la excelencia operativa a un nivel de organización y carga de trabajo, y aplíquelo a todas las áreas. Manteners e al día con las recomendaciones de AWS, las fuentes del sector y la inteligencia de amenazas le ayudan a desarrollar el modelo de amenaza y controlar los objetivos. La automatización de los procesos de seguridad, las pruebas y la validación le ayudan a escalar las operaciones de seguridad.

En AWS, se recomienda separar las distintas cargas de trabajo por cuenta, según su función y los requisitos de cumplimiento o confidencialidad de los datos.

Identity and Access Management

La administración de identidades y accesos representa una parte clave de un programa de seguridad de la información, ya que garantiza que solo los usuarios y componentes autorizados e identificados puedan acceder a sus recursos (y solo de la forma prevista). Por ejemplo, debería definir las entidades principales (es decir, cuentas, usuarios, roles y servicios que pueden intervenir en su cuenta), crear políticas que hagan referencia a estas entidades e implementar una administración sólida de credenciales. Estos elementos de administración de privilegios constituyen el núcleo de la autenticación y la autorización.

En AWS, la administración de privilegios se apoya principalmente en el servicio AWS Identity and Access Management (IAM), que permite controlar el acceso de usuarios y programas a los servicios y recursos de AWS. Procure aplicar políticas granulares que asignen permisos a cada usuario, grupo, función o recurso. También puede exigir prácticas de contraseñas seguras; por ejemplo, puede establecer el nivel de complejidad, impedir la reutilización y emplear autenticación multifactor (MFA). Puede usar la federación con su servicio de directorio existente. Cuando las cargas de trabajo requieren que los sistemas tengan acceso a AWS, IAM posibilita un acceso seguro mediante la asignación de roles, perfiles de instancia, federación de identidades y credenciales temporales.

Las siguientes preguntas se centran en estas consideraciones de seguridad.

SEC 2: ¿Cómo administra las identidades de personas y máquinas?

Hay dos tipos de identidades que debe administrar cuando tenga que utilizar cargas de trabajo de AWS seguras. Entender el tipo de identidad que necesita administrar y a la que debe otorgar acceso ayuda a garantizar que las identidades adecuadas tengan acceso a los recursos correctos bajo las condiciones adecuadas.

Identidades humanas: los administradores, desarrolladores, operadores y clientes finales requieren una identidad para acceder a sus aplicaciones y entornos de AWS. Estos son miembros de la organización o usuarios externos con los que colabora y que interactúan con sus recursos de AWS a través de un navegador web, una aplicación cliente o herramientas de línea de comandos interactivas.

Identidades de máquinas: las aplicaciones de servicio, las herramientas operativas y las cargas de trabajo requieren una identidad para realizar solicitudes a los servicios de AWS, por ejemplo, para leer datos. Entre estas identidades se incluyen máquinas que se ejecutan en su entorno de AWS, como instancias de Amazon EC2 o funciones de AWS Lambda. También puede administr ar identidades de máquinas para terceros que necesiten acceso. Además, es posible que también tenga máquinas fuera de AWS que necesiten acceso a su entorno de AWS.

SEC 3: ¿Cómo administra los permisos de personas y máquinas?

Administre permisos para controlar el acceso a identidades de personas y de máquinas que requieran acceso a AWS y sus cargas de trabajo. Los permisos controlan quién puede acceder a qué y en qué condiciones.

Las credenciales no se deben compartir entre usuarios o sistemas. El acceso de los usuarios se debe conceder empleando un enfoque de privilegio mínimo con prácticas recomendadas, como los requisitos de contraseña y la obligatoriedad de usar MFA. El acceso programático, incluidas las llamadas a la API de los servicios de AWS, debe realizarse mediante credenciales temporales y de privilegio limitado como las emitidas por AWS Security Token Service.

AWS ofrece recursos que pueden ayudar a administrar la identidad y el acceso. Para conocer las prácticas recomendadas, explore los laboratorios prácticos sobre la administración de credenciales y autentificación, el control del acceso humanoy el control del acceso programático.

Detección

Puede usar los controles detectores para identificar una incidencia o amenaza potencial de seguridad. Son una parte fundamental de los marcos de gobernanza y se pueden usar como complemento de procesos de calidad, para una obligación legal o de conformidad y para la identificación de amenazas y respuestas. Hay distintos tipos de controles de detección. Por ejemplo, realizar un inventario de activos y de sus atributos detallados facilita la toma de decisiones (y los controles del ciclo de vida) para ayudar a establecer líneas de base operativas. También puede usar auditorías internas, un examen de los controles relacionados con los sistemas de información, para garantizar que las prácticas cumplan con las políticas y los requisitos, así como que haya establecido las notificaciones correctas de alertas automatizadas basadas en las condiciones definidas. Estos controles son factores reactivos importantes que ayudan a su organización a identificar la actividad anómala y comprender sus repercusiones.

AWS le permite realizar controles de detección mediante procesamiento de registros y eventos y supervisión que posibilitan aplicar auditorías, análisis automatizados y alarmas. Los registros de CloudTrail, las llamadas a la API de AWS y CloudWatch permiten supervisar las métricas con alarmas, y AWS Config proporciona el historial de configuración. Amazon GuardDuty es un servicio administrado de detección de amenazas que supervisa de forma continua comportamientos malintencionados o no autorizados y ayuda a proteger sus cargas de trabajo y cuentas de AWS. También dispone de registros de nivel de servicio; por ejemplo, puede utilizar Amazon Simple Storage Service (Amazon S3) para registrar las solicitudes de acceso.

Las siguientes preguntas se centran en las consideraciones de seguridad.

SEC 4: ¿Cómo detecta e investiga los eventos de seguridad?

Capte y analice eventos de registros y métricas para obtener visibilidad. Tome medidas sobre eventos de seguridad y posibles amenazas para ayudar a proteger su carga de trabajo.

La administración de registros es una parte importante de una carga de trabajo de Well-Architected por razones que van desde la seguridad y el análisis forense hasta los requisitos normativos o legales. Es fundamental que analice los registros y actúe al respecto para identificar posibles incidentes de seguridad. AWS proporciona una funcionalidad que facilita la puesta en práctica de la administración de registros, ya que permite definir un ciclo de vida de retención de datos y decidir dónde se conservarán, archivarán y eliminarán los datos. Así es más sencillo y rentable lograr que el manejo de datos sea predecible y fiable.

Protección de la infraestructura

La protección de la infraestructura abarca las metodologías de control, como la defensa en profundidad, necesarias para ajustarse a las prácticas recomendadas y las obligaciones organizativas o normativas. El uso de estas metodologías es fundamental para el éxito de las operaciones en curso, ya sea en la nube o en las instalaciones.

En AWS puede implementar la inspección de paquetes con y sin estado, ya sea mediante tecnologías incluidas en AWS o mediante los productos y servicios de socios disponibles en AWS Marketplace. Se recomienda usar Amazon Virtual Private Cloud (Amazon VPC) para crear un entorno privado, seguro y escalable en el que pueda definir su topología, incluidas puertas de enlace, tablas de enrutamiento y subredes públicas y privadas.

Las siguientes preguntas se centran en estas consideraciones de seguridad.

SEC 5: ¿Cómo protege sus recursos de red?

Cualquier carga de trabajo que tenga forma de conexión de red, ya sea internet o una red privada, requiere varias capas de defensa para protegerse de amenazas internas y externas basadas en la red.

SEC 6: ¿Cómo protege sus recursos informáticos?

Los recursos informáticos en su carga de trabajo requieren varias capas de defensa para ayudar a protegerse de amenazas externas e internas. Entre los recursos informáticos se incluyen instancia s EC2, contenedores, funciones de AWS Lambda, servicios de bases de datos, dispositivos IoT, etc.

Tener varias capas de defensa es aconsejable en cualquier tipo de entorno. En el caso de la protección de la infraestructura, muchos de los conceptos y métodos son válidos para los modelos de nube y para los locales. En un plan eficaz de seguridad de la información, es esencial imponer la protección de los límites, monitorear los puntos de entrada y salida y aplicar de forma exhaustiva registros, monitoreo y alertas.

Los clientes de AWS pueden adaptar o reforzar la configuración de Amazon Elastic Compute Cloud (Amazon EC2), un contenedor de Amazon Elastic Container Service (Amazon ECS) o una instancia

de AWS Elastic Beanstalk, y mantener esta configuración de forma persistente en una imagen de máquina de Amazon (AMI) inmutable. Así, todos los nuevos servidores virtuales (instancias) que se inicien con esta AMI reciben la configuración reforzada, tanto si los activa Auto Scaling como si se activan de forma manual.

Protección de los datos

Antes de diseñar un sistema, hay que adoptar prácticas que influyen en la seguridad. Por ejemplo, la clasificación de datos ofrece una manera de categorizar los datos de la organización según los niveles de confidencialidad, mientras que el cifrado protege los datos haciéndolos ininteligibles para el acceso no autorizado. Estas herramientas y técnicas son importantes porque respaldan objetivos como la prevención de pérdidas económicas o la conformidad con las obligaciones reglamentarias.

En AWS, las siguientes prácticas facilitan la protección de datos:

- Como cliente de AWS, controla por completo sus datos.
- AWS simplifica el cifrado de los datos y la administración de claves, incluyendo la rotación regular de las claves, que puede realizar de forma manual o automatizar fácilmente con AWS.
- Dispone de un registro detallado con contenidos importantes, como el acceso a los archivos y los cambios.
- AWS ha diseñado sistemas de almacenamiento con una resiliencia excepcional. Por ejemplo, Amazon S3 Standard, S3 Standard–IA, S3 One Zone-IA y Amazon Glacier están todos diseñados para proporcionar una durabilidad de objetos del 99,999999999 % en un año determinado. Este nivel de durabilidad corresponde a una pérdida media anual esperada del 0,000000001 % de los objetos.
- El control de versiones, que puede formar parte de un proceso más amplio de administración del ciclo de vida de los datos, puede protegerlos contra sobrescrituras o eliminaciones accidentales y daños similares.
- AWS nunca inicia un traslado de datos a otras regiones. El contenido situado en una región permanece en esa región a menos que se habilite explícitamente una característica o se emplee un servicio que proporcione esa funcionalidad.

Las siguientes preguntas se centran en estas consideraciones de seguridad.

SEC 7: ¿Cómo clasifica sus datos?

La clasificación proporciona una forma de categorizar los datos basada en el nivel de importanc ia y la confidencialidad para ayudarle a determinar los controles de protección y conservación adecuados.

SEC 8: ¿Cómo protege los datos en reposo?

Para proteger los datos en reposo debe implementar varios controles para reducir el riesgo de acceso no autorizado o mala gestión.

SEC 9: ¿Cómo protege sus datos en tránsito?

Para proteger los datos en tránsito debe implementar varios controles para reducir el riesgo de acceso no autorizado o pérdida.

AWS proporciona múltiples medios para cifrar los datos en reposo y en tránsito. Nuestros servicios incorporan características que facilitan el cifrado de los datos. Por ejemplo, hemos implementado el cifrado del servidor (SSE) en Amazon S3 para facilitar el almacenamiento de datos de forma cifrada. También puede hacer que Elastic Load Balancing (ELB) maneje todo el proceso de cifrado y descifrado HTTPS (generalmente conocido como terminación SSL).

Respuesta ante incidentes

Incluso con controles detectores y de prevención extremadamente eficaces, la organización debería continuar aplicando procesos para responder a incidencias de seguridad y mitigar su posible impacto. La arquitectura de la carga de trabajo afecta considerablemente a la capacidad de los equipos de operar de forma eficaz durante una incidencia, aislar o contener sistemas y restaurar operaciones a un estado conocido correcto. Si prepara las herramientas y el acceso en previsión de un incidente de seguridad, y practica periódicamente la respuesta a incidentes durante los días de juego, se asegurará de que su arquitectura posibilite una investigación y una recuperación sin demoras.

En AWS, las siguientes prácticas facilitan una respuesta efectiva frente a incidencias:

- Dispone de un registro detallado con contenidos importantes, como el acceso a los archivos y los cambios.
- Los eventos pueden procesarse automáticamente y activan herramientas que automatizan las respuestas mediante el uso de las API de AWS.
- Puede preaprovisionar herramientas y una sala blanca con AWS CloudFormation. Esto permite realizar análisis forenses en un ambiente seguro y aislado.

Las siguientes preguntas se centran en las consideraciones de seguridad.

SEC 10: ¿Cómo anticipa, responde y se recupera de los incidentes?
La preparación es fundamental para investigar de forma oportuna y efectiva, dar respuesta a incidentes de seguridad, así como recuperarse para minimizar las interrupciones en su organizac ión.

Asegúrese de tener una forma de conceder rápidamente el acceso a su equipo de seguridad y automatice tanto el aislamiento de las instancias como la captura de datos y estado para el análisis forense.

Recursos

Consulte los siguientes recursos para obtener más información sobre nuestras prácticas recomendadas en materia de seguridad.

Documentación

- Seguridad en la nube de AWS
- Conformidad de AWS
- Blog de seguridad de AWS

Documento técnico

- Pilar de seguridad
- Información general de seguridad de AWS
- Riesgo y conformidad de AWS

Vídeo

- Informe de seguridad de AWS
- Información general sobre responsabilidad compartida

Fiabilidad

El pilar de fiabilidad abarca la capacidad de una carga de trabajo para realizar su función prevista de forma correcta y coherente cuando se espera que lo haga. Esto incluye la capacidad de utilizar y probar la carga de trabajo a lo largo de todo su ciclo de vida. En este documento se incluyen consejos exhaustivos y de prácticas recomendadas para la implementación de cargas de trabajo fiables en AWS.

El pilar de fiabilidad proporciona información general sobre los principios de diseño, prácticas recomendadas y preguntas. Encontrará recomendaciones de implementación en el documento técnico Pilar de fiabilidad.

Temas

- Principios de diseño
- Definición
- Prácticas recomendadas
- Recursos

Principios de diseño

Existen cinco principios de diseño de fiabilidad en la nube:

- Recuperación automática de errores: al supervisar una carga de trabajo para los indicadores clave de rendimiento (KPI), puede desencadenar la automatización cuando se supere un umbral. Estos KPI deben ser una medida del valor de negocio, no de los aspectos técnicos del funcionamiento del servicio. De este modo, se permite la notificación y el seguimiento automático de los errores, así como los procesos de recuperación automatizada que pueden solucionar o corregir el error. Con una automatización más sofisticada, es posible anticipar y solucionar errores antes de que sucedan.
- Prueba de los procedimientos de recuperación: en un entorno local, a menudo se realizan pruebas para ver si una carga de trabajo funciona en una situación concreta. Normalmente, las pruebas no

se usan para comprobar estrategias de recuperación. En la nube, puede probar los errores de la carga de trabajo y validar los procedimientos de recuperación. Puede usar la automatización para simular diferentes errores o recrear escenarios que anteriormente han producido algún error. Esto expone vías de error que puede probar y arreglar antes de que se produzca un escenario de error real, lo que reduce el riesgo.

- Escalar horizontalmente para aumentar la disponibilidad de la carga de trabajo de agregación:reemplace un recurso grande por varios recursos pequeños para reducir el efecto de un solo error en todo la carga de trabajo. Distribuya las solicitudes a través de varios recursos más pequeños para garantizar que no compartan el mismo error.
- No más conjeturas sobre la capacidad: una causa común de los errores en los sistemas locales es la saturación de recursos, cuando las demandas que se le asignan a una carga de trabajo superan su capacidad (este es a menudo el objetivo de los ataques de denegación de servicio). En la nube, se puede supervisar la demanda y el uso de la carga de trabajo, además de automatizar la incorporación o eliminación de recursos de forma automatizada para mantener un nivel óptimo y satisfacer la demanda sin tener un aprovisionamiento excesivo o insuficiente. Aún hay límites, pero algunas cuotas se pueden controlar, mientras que otras se pueden administrar (consulte Administración de Service Quotas y restricciones).
- Administración de cambios en la automatización: los cambios que se apliquen a la infraestructura deben realizarse con automatización. Entre los cambios que se deben administrar se encuentran los de la automatización, de los que, posteriormente, se puede hacer un seguimiento y una revisión.

Definición

Existen cuatro áreas de prácticas recomendadas para la fiabilidad en la nube:

- Fundamentos
- Arquitectura de la carga de trabajo
- Administración de cambios
- Administración de errores

Para lograr fiabilidad hay que empezar por los cimientos: un entorno en el que las cuotas de servicio y la topología de la red se adapten a la carga de trabajo. La arquitectura de la carga de trabajo del sistema distribuido debe estar diseñada para prevenir y mitigar los errores. La carga de trabajo debe

gestionar los cambios en la demanda o los requisitos, y debe estar diseñada para detectar errores y repararse de forma automática.

Prácticas recomendadas

Temas

- Fundamentos
- Arquitectura de la carga de trabajo
- Administración de cambios
- Administración de errores

Fundamentos

Los requisitos fundamentales son aquellos cuyo alcance va más allá de una única carga de trabajo o proyecto. Antes de diseñar la arquitectura de cualquier sistema, los requisitos básicos que afectan a la fiabilidad deberían estar aplicados. Por ejemplo, es preciso que haya suficiente ancho de banda de la red en el centro de datos.

AWS ya incorpora la mayor parte de estos requisitos básicos. Si no lo están, permite que estos se traten según corresponda. La nube está diseñada para que sea, en esencia, ilimitada, por lo que la responsabilidad de brindar suficiente capacidad de cómputo y de red recae en AWS. Al mismo tiempo, será libre de cambiar la asignación y el tamaño de los recursos según lo necesite.

Las siguientes preguntas se centran en estas consideraciones de fiabilidad. (Para ver una lista de preguntas y prácticas recomendadas sobre la optimización de costos, consulte el Apéndice).

REL 1 ¿Cómo administra las cuotas de servicio y las restricciones?

Para las arquitecturas de carga de trabajo basadas en la nube, existen cuotas de servicio (que también se denominan límites de servicio). Estas cuotas existen para evitar aprovisionar por accidente más recursos de los necesarios y para limitar las tasas de solicitud en las operaciones de la API de modo que los servicios queden protegidos ante posibles usos inadecuados. También existen restricciones de recursos, por ejemplo, la velocidad a la que se pueden introducir bits en un cable de fibra óptica o la cantidad de almacenamiento de un disco físico.

REL 2 ¿Cómo planifica la topología de la red?

Suele haber cargas de trabajo en distintos entornos. Entre estos se incluyen los entornos de la nube (tanto públicamente accesibles como privados), y posiblemente la infraestructura del centro de datos existente. Los planes deben incluir consideraciones, como la conectividad dentro de los sistemas y entre ellos, la administración de las direcciones IP públicas, la administración de las direcciones IP privadas y la resolución de nombres de dominio.

Para las arquitecturas de carga de trabajo basadas en la nube, existen cuotas de servicio (que también se denominan límites de servicio). Estas cuotas existen para evitar aprovisionar por accidente más recursos de los necesarios y para limitar las tasas de solicitud en las operaciones de la API, de modo que los servicios queden protegidos ante posibles usos inadecuados. Suele haber cargas de trabajo en distintos entornos. Debe supervisar y administrar estas cuotas para todos los entornos de la carga de trabajo. Entre estos se incluyen los entornos de la nube (de acceso público y privado) y pueden incluir la infraestructura del centro de datos existente. Los planes deben incluir consideraciones, como la conectividad dentro de los sistemas y entre ellos, la administración de las direcciones IP públicas, la administración de las direcciones IP privadas y la resolución de nombres de dominio.

Arquitectura de la carga de trabajo

Una carga de trabajo fiable comienza por tomar decisiones de diseño anticipadas tanto para el software como para la infraestructura. Sus elecciones respecto a la arquitectura tendrán un impacto sobre el comportamiento de su carga de trabajo en los seis pilares de Well-Architected. Para la fiabilidad, hay patrones específicos que debe seguir.

En AWS, los desarrolladores de la carga de trabajo pueden elegir los idiomas y las tecnologías a usar. Los SDK de AWS simplifican la codificación al proporcionar API específicas de idioma para los servicios de AWS. Estos SDK, más la elección del idioma, permiten a los desarrolladores implementar las prácticas recomendadas de fiabilidad enumeradas aquí. Los desarrolladores también pueden informarse y aprender sobre el modo en que Amazon crea y opera software en la Amazon Builders' Library.

Las siguientes preguntas se centran en estas consideraciones de fiabilidad.

REL 3 ¿Cómo diseña la arquitectura de servicio de su carga de trabajo?

Desarrolle cargas de trabajo escalables y fiables utilizando una arquitectura orientada a servicios (SOA) o una arquitectura de microservicios. La arquitectura orientada a servicios (SOA) es la práctica de hacer que los componentes de software se puedan reutilizar mediante interfaces de servicio. La arquitectura de microservicios va más allá, para hacer que los componentes sean más pequeños y sencillos.

REL 4 ¿Cómo diseña las interacciones en un sistema distribuido para evitar errores?

Los sistemas distribuidos dependen de las redes de comunicaciones para interconectar componentes, como servidores o servicios. Su carga de trabajo debe funcionar de manera fiable aunque se pierdan datos o haya latencia en estas redes. Los componentes del sistema distribuido deben funcionar de forma que no repercutan negativamente en otros componentes ni en la carga de trabajo. Estas prácticas recomendadas evitan que se produzcan errores y mejoran el tiempo medio entre errores (MTBF).

REL 5 ¿Cómo diseña las interacciones en un sistema distribuido para mitigar o tolerar errores?

Los sistemas distribuidos dependen de las redes de comunicaciones para interconectar componentes, como servidores o servicios. Su carga de trabajo debe funcionar de manera fiable aunque se pierdan datos o haya latencia en estas redes. Los componentes del sistema distribuido deben funcionar de forma que no repercutan negativamente en otros componentes ni en la carga de trabajo. Estas prácticas recomendadas permiten que las cargas de trabajo toleren el estrés o los errores, se recuperen más rápidamente de ellos y mitiguen el impacto de dichos errores. El resultado es un tiempo medio de recuperación (MTTR) mejor.

Administración de cambios

Se deben prever y ajustar los cambios en su carga de trabajo o su entorno para poder conseguir un funcionamiento fiable de la carga de trabajo. Los cambios incluyen aquellos que se imponen a su carga de trabajo, como los picos de demanda, además de los inherentes a ella, como los despliegues de funciones o los parches de seguridad.

Con AWS puede monitorear el comportamiento de una carga de trabajo y automatizar la respuesta a los KPI. Por ejemplo, la carga de trabajo puede añadir servidores adicionales a medida que la carga de trabajo gane más usuarios. Puede controlar quién tiene permisos para realizar cambios en la carga de trabajo e inspeccionar el historial de cambios.

Las siguientes preguntas se centran en estas consideraciones de fiabilidad.

REL 6 ¿Cómo supervisa los recursos de las cargas de trabajo?

Los registros y las métricas son una potente herramienta para obtener información sobre el estado de sus cargas de trabajo. Puede configurar su carga de trabajo de forma que supervise registros y métricas, y envíe notificaciones cuando se crucen ciertos umbrales o se produzcan eventos importantes. La supervisión permite que su carga de trabajo reconozca cuándo se cruzan umbrales de bajo rendimiento o se producen errores, para que pueda recuperarse de los errores rápidamente una vez recibida una respuesta.

REL 7 ¿Cómo diseña su carga de trabajo para que se adapte a los cambios en la demanda?

Una carga de trabajo escalable proporciona elasticidad para agregar y eliminar recursos de forma automática a fin de que coincidan estrechamente con la demanda actual en cualquier momento dado.

REL 8 ¿Cómo implementa los cambios?

Los cambios controlados son necesarios para implementar nueva funcionalidad y garantizar que las cargas de trabajo y el entorno operativo ejecuten software conocido y puedan recibir parches o reemplazos de manera predecible. Si estos cambios no se controlan, puede ser difícil prever su efecto o abordar los problemas que surjan a raíz de ellos.

Cuando diseña la arquitectura de una carga de trabajo para agregar y eliminar recursos de forma automática como respuesta a los cambios solicitados, se aumenta la fiabilidad a la par que se garantiza que el éxito del negocio no se convierta en una carga. Al contar con monitoreo, el equipo recibirá alertas automáticas cuando los KPI se desvíen de las reglas esperadas. Los registros automáticos de los cambios realizados en el entorno le permiten inspeccionar e identificar

rápidamente aquellas medidas que hayan repercutido en la fiabilidad. Controlar la administración de cambios garantiza que se puedan aplicar reglas que ayuden a alcanzar el grado de fiabilidad deseado.

Administración de errores

De cualquier sistema con una complejidad razonable se esperan errores. La fiabilidad requiere que su carga de trabajo conozca los errores a medida que ocurren y que actúe para evitar que afecten a la disponibilidad. Las cargas de trabajo deben ser capaces de tolerar errores y de repararlos de forma automática.

Gracias a AWS, podrá aprovechar la automatización para reaccionar a los datos de monitoreo. Por ejemplo, cuando una métrica concreta pasa un umbral, podrá desencadenar una acción automática para solucionar el problema. Además, puede reemplazar un recurso que genere un error y forme parte del entorno de producción por uno nuevo y analizar dicho recurso fuera de banda en lugar de intentar diagnosticar y arreglar el recurso del error. Ya que la nube permite soportar versiones temporales de todo un sistema a bajo costo, puede usar las pruebas automáticas para comprobar los procesos de recuperación completos.

Las siguientes preguntas se centran en estas consideraciones de fiabilidad.

REL 9 ¿Cómo realiza una copia de seguridad de los datos?

Realice una copia de seguridad de los datos, las aplicaciones y la configuración para satisfacer sus requisitos de objetivos de tiempo de recuperación (RTO) y objetivos de punto de recuperación (RPO).

REL 10 ¿Cómo usa el aislamiento de errores para proteger su carga de trabajo?

Los límites aislados de los errores acotan el efecto de un error en una carga de trabajo a un número limitado de componentes. Los componentes fuera del límite no resultan afectados por el error. Mediante el uso de varios límites aislados de error, puede acotar el impacto en su carga de trabajo.

REL 11 ¿Cómo diseña su carga de trabajo para que soporte los errores de los componentes?

Las cargas de trabajo con un requisito de alta disponibilidad y un tiempo de recuperación (MTTR) bajo deben diseñarse para que sean resilientes.

REL 12 ¿Cómo pone a prueba la fiabilidad?

Una vez diseñada la carga de trabajo para que sea resiliente al estrés de producción, las pruebas son la única forma de garantizar que funcionará según lo previsto y proporcionará la resiliencia esperada.

REL 13 ¿Cómo planifica la recuperación de desastres (DR)?

Disponer de copias de seguridad y de componentes de cargas de trabajo redundantes es el principio de su estrategia de DR. El RTO y el RPO son sus objetivos para la restauración de su carga de trabajo. Estos se definen en función de las necesidades del negocio. Implemente una estrategia para satisfacer estos objetivos teniendo en cuenta las ubicaciones y la función de los recursos de las cargas de trabajo y los datos. La probabilidad de una interrupción y el coste de recuperación son también factores clave que ayudan a conocer el valor empresarial de proporcio nar recuperación de desastres para una carga de trabajo.

Haga una copia de seguridad de los datos de forma regular y ponga a prueba estos archivos para garantizar que pueda recuperarse tanto de los errores físicos como de los lógicos. Un factor clave para administrar los errores es probar de forma frecuente y automática las cargas de trabajo que causan error para después observar cómo se recuperan. Haga esto de manera regular y asegúrese de que dichas pruebas también se desencadenen tras realizar cambios importantes en la carga de trabajo. Realice un seguimiento activo de los KPI, así como el objetivo de tiempo de recuperación (RTO) y el objetivo de punto de recuperación (RPO) para evaluar la resiliencia de la carga de trabajo (especialmente, cuando se pongan a prueba situaciones en las que se produzca un error). Realizar el seguimiento de los KPI será de ayuda para identificar y mitigar los puntos únicos de error. El objetivo es someter los procesos de recuperación de la carga de trabajo a pruebas exhaustivas para que sepa que puede recuperar todos los datos y continuar brindando servicios a los clientes, aunque se experimenten problemas prolongados. Los procesos de recuperación deberían realizarse igual de bien que los procesos de producción normales.

Recursos

Consulte los siguientes recursos para obtener más información sobre nuestras prácticas recomendadas de fiabilidad.

Documentación

- Documentación de AWS
- Infraestructura global de AWS
- AWS Auto Scaling: cómo funcionan los planes de escalado
- ¿Qué es AWS Backup?

Documento técnico

- Pilar de fiabilidad: AWS Well-Architected
- Implementación de microservicios en AWS

Eficiencia del rendimiento

El pilar de eficiencia del rendimiento incluye la capacidad de utilizar de forma eficaz los recursos de computación para satisfacer los requisitos del sistema, así como el mantenimiento de esta eficiencia a medida que la demanda cambia y las tecnologías evolucionan.

El pilar de eficiencia del rendimiento ofrece una visión general de principios de diseño, prácticas recomendadas y preguntas. Encontrará una guía prescriptiva de implementación en el documento técnico Pilar de eficiencia del rendimiento.

Temas

- Principios de diseño
- Definición
- Prácticas recomendadas
- Recursos

Principios de diseño

Existen cinco principios de diseño para la eficiencia del rendimiento en la nube:

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

Definición

Existen cinco áreas de prácticas recomendadas para la eficiencia del rendimiento en la nube:

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

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

Revisar periódicamente sus opciones validará que aprovecha la continua evolución de la nube de AWS. Mediante la supervisión verifica que usted es consciente de 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 coherencia.

Prácticas recomendadas

Temas

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

Selección de la arquitectura

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

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

La siguiente pregunta se centra en estas consideraciones para mejorar la eficacia del rendimiento. (Para ver una lista de preguntas sobre la eficiencia del rendimiento y las prácticas recomendadas, consulte el Appendix).

PERF 1: How do you select appropriate cloud resources and architecture patterns for your workload?

Often, multiple approaches are required for more effective performance across a workload. Well-Architected systems use multiple solutions and features to improve performance.

Computación y hardware

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

En AWS, la computación está disponible de tres formas: instancias, contenedores y funciones.

- Las instancias son servidores virtualizados, lo que le permite cambiar sus funcionalidades con un botón o una llamada a la API. Como las decisiones sobre los recursos en la nube no son fijas, puede experimentar con diferentes tipos de servidores. En AWS, estas instancias de servidor virtual se presentan en diferentes familias y tamaños, y ofrecen una amplia variedad de capacidades, incluidas unidades de estado sólido (SSD) y unidades de procesamiento gráfico (GPU).
- Los contenedores son un método de virtualización del sistema operativo que le permite ejecutar una aplicación y sus dependencias en procesos aislados de recursos. AWS Fargate es un sistema de computación sin servidor para contenedores o puede utilizarse Amazon EC2 si necesita controlar la instalación, la configuración y la administración de su entorno de computación. También puede elegir entre varias plataformas de orquestación de contenedores: Amazon Elastic Container Service (ECS) o Amazon Elastic Kubernetes Service (EKS).
- Las funciones extraen el entorno de ejecución del código que desea aplicar. Por ejemplo, AWS Lambda permite ejecutar código sin ejecutar una instancia.

La siguiente pregunta se centra en estas consideraciones para mejorar la eficacia del rendimiento.

PERF 2: How do you select and use compute resources in your workload?

The more efficient compute solution for a workload varies based on application design, usage patterns, and configuration settings. Architectures can use different compute solutions for various components and turn on different features to improve performance. Selecting the wrong compute solution for an architecture can lead to lower performance efficiency.

Administración de datos

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

En AWS, el almacenamiento está disponible en tres formas: objeto, bloque y archivo.

- El almacenamiento de objetos proporciona una plataforma escalable y duradera para que se pueda acceder a los datos desde cualquier lugar de Internet para el contenido generado por el usuario, el archivo activo, la computación sin servidor, el almacenamiento de macrodatos o la copia de seguridad y la recuperación. Amazon Simple Storage Service (Amazon S3) es un servicio de almacenamiento de objetos que ofrece una escalabilidad, una disponibilidad de datos, una seguridad y un rendimiento líderes en el sector. Amazon S3 se ha diseñado para ofrecer una durabilidad del 99,999999999 % (11 nueves) y almacena los datos de millones de aplicaciones de empresas de todo el mundo.
- El almacenamiento en bloques proporciona un almacenamiento de alta disponibilidad, coherente y de baja latencia para cada host virtual y es análogo al almacenamiento de conexión directa (DAS) o a una red de área de almacenamiento (SAN). Amazon Elastic Block Store (Amazon EBS) se ha diseñado para cargas de trabajo que requieren un almacenamiento persistente al que pueden acceder las instancias de EC2 y que le ayuda a optimizar las aplicaciones con la capacidad de almacenamiento, el rendimiento y el coste adecuados.
- El almacenamiento de archivos proporciona acceso a un sistema de archivos compartido en varios sistemas. Las soluciones de almacenamiento de archivos como Amazon Elastic File System (Amazon EFS) son ideales para casos de uso como grandes repositorios de contenido, entornos de desarrollo, almacenes de medios o directorios de inicio de usuario. Amazon FSx facilita y

rentabiliza el lanzamiento y la ejecución de sistemas de archivos populares para poder aprovechar los completos conjuntos de características y el rápido rendimiento de los sistemas de archivos de código abierto y con licencia comercial más utilizados.

La siguiente pregunta se centra en estas consideraciones para mejorar la eficacia del rendimiento.

PERF 3: How do you store, manage, and access data in your workload?

The more efficient storage solution for a system varies based on the kind of access operation (block, file, or object), patterns of access (random or sequential), required throughput, frequency of access (online, offline, archival), frequency of update (WORM, dynamic), and availability and durability constraints. Well-architected systems use multiple storage solutions and turn on different features to improve performance and use resources efficiently.

Redes y entrega de contenido

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

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

La siguiente pregunta se centra en estas consideraciones para mejorar la eficacia del rendimiento.

PERF 4: How do you select and configure networking resources in your workload?

This question includes guidance and best practices to design, configure, and operate efficient networking and content delivery solutions in the cloud.

Proceso y cultura

Al diseñar cargas de trabajo, hay principios y prácticas que puede adoptar para ayudarle a ejecutar mejor cargas de trabajo en la nube eficientes y de alto rendimiento. Para adoptar una cultura que fomente la eficiencia del rendimiento de las cargas de trabajo en la nube, tenga en cuenta estos principios y prácticas clave.

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 despliegue: utilice una canalización de integración continua o de despliegue continuo (CI/CD), como, por ejemplo, el repositorio del código fuente, los sistemas de diseño, el despliegue y la automatización de pruebas, para desplegar su infraestructura. Esto le permite desplegar de manera repetible, coherente y por un bajo coste mientras itera.
- Métricas bien definidas: configure y supervise las métricas para recoger indicadores clave de rendimiento (KPI). Recomendamos que utilice tanto métricas técnicas, como comerciales. Para aplicaciones móviles o sitios web, las métricas clave registran el tiempo para el primer byte o la renderización. Otras métricas que generalmente se aplican incluyen el recuento de subprocesos, la tasa de recolección de basura y los estados de espera. Las métricas comerciales, como el costo acumulado agregado por solicitud, puede alertarle sobre formas de reducir costos. Considere con cuidado cómo planifica interpretar las métricas. Por ejemplo, podría elegir el percentil máximo o el 99.º, en vez del promedio.
- Prueba de rendimiento automática: como parte de su proceso de despliegue, inicie automáticamente las pruebas de rendimiento después de que las pruebas de ejecución más rápida se hayan superado con éxito. La automatización debería crear un nuevo entorno, establecer condiciones iniciales como datos de prueba y luego ejecutar una serie de puntos de referencia y pruebas de carga. Los resultados de estas pruebas deberían estar vinculados al diseño, para que pueda seguir los cambios del rendimiento en el tiempo. Para las pruebas de larga ejecución, puede hacer que esta parte de la canalización sea asíncrona al resto del diseño. Alternativamente, podría ejecutar las pruebas de rendimiento durante la noche con instancias de spot de Amazon EC2.
- Generación de cargas: debe crear una serie de scripts de prueba que repliquen trayectos de usuario sintéticos o pregrabados. Estos scripts deben ser idempotentes y no acoplados, y podría

necesitar incluir scripts de precalentamiento para obtener resultados válidos. En la medida de lo posible, sus scripts de prueba deben replicar el comportamiento de uso en la producción. Puede utilizar soluciones de software o de software como servicio (SaaS) para generar la carga. Considere usar soluciones de AWS Marketplace e instancias de spot, que pueden ser formas rentables de generar la carga.

- Visibilidad de rendimiento: las métricas clave deben ser visibles para su equipo, especialmente las métricas para cada versión de diseño. Esto le permite ver cualquier tendencia significativa, sea positiva o negativa, con el paso del tiempo. También debería exponer métricas en la cantidad de errores o excepciones para garantizar que está poniendo a prueba un sistema de trabajo.
- Visualización: utilice técnicas de visualización que dejen claro dónde se presentan problemas de rendimiento, puntos críticos, estados de espera o un uso bajo. Superponga las métricas de rendimiento sobre los diagramas de arquitectura: los gráficos de llamadas o el código pueden ayudar a identificar problemas con mayor rapidez.
- Proceso de revisión periódico: el mal funcionamiento de las arquitecturas suele ser el resultado de un proceso de revisión del rendimiento inexistente o deficiente. Si su arquitectura tiene un bajo rendimiento, la implementación de un proceso de revisión del rendimiento le permitirá impulsar la mejora iterativa.
- Optimización continua: adopte una cultura que optimice continuamente la eficiencia del rendimiento de su carga de trabajo en la nube.

La siguiente pregunta se centra en estas consideraciones para mejorar la eficacia del rendimiento.

PERF 5: What process do you use to support more performance efficiency for your workload?

When architecting workloads, there are principles and practices that you can adopt to help you better run efficient high-performing cloud workloads. To adopt a culture that fosters performance efficiency of cloud workloads, consider these key principles and practices.

Recursos

Consulte los siguientes recursos para obtener más información sobre nuestras prácticas recomendadas en la eficiencia del rendimiento.

Documentación

- Optimización del rendimiento de Amazon S3

- Rendimiento del volumen en Amazon EBS

Documento técnico

- Pilar de eficiencia del rendimiento

Vídeo

- AWS re:Invent 2019: Amazon EC2 foundations (CMP211-R2)
- AWS re:Invent 2019: Leadership session: Storage state of the union (STG201-L)
- AWS re:Invent 2019: Leadership session: AWS purpose-built databases (DAT209-L)
- AWS re:Invent 2019: Connectivity to AWS and hybrid AWS network architectures (NET317-R1)
- AWS re:Invent 2019: Powering next-gen Amazon EC2: Deep dive into the Nitro system (CMP303-R2)
- AWS re:Invent 2019: Scaling up to your first 10 million users (ARC211-R)

Optimización de costes

El pilar de optimización de costes incluye la capacidad de ejecutar sistemas para ofrecer valor empresarial al precio más bajo posible.

El pilar de optimización de costos proporciona información general sobre los principios de diseño, prácticas recomendadas y preguntas. Encontrará recomendaciones de implementación en el documento técnico Pilar de optimización de costes.

Temas

- Principios de diseño
- Definición
- Prácticas recomendadas
- Recursos

Principios de diseño

Existen cinco principios de diseño para la optimización de costes en la nube:

- Implementación de la administración financiera en la nube: para lograr el éxito financiero y acelerar la materialización del valor empresarial en la nube, debe invertir en la administración financiera en la nube y en la optimización de costes. Su organización debe dedicar tiempo y recursos para dominar este nuevo ámbito de la tecnología y la administración del uso. De forma similar a su capacidad de seguridad o de excelencia operativa, necesita desarrollar capacidades a través de la creación de conocimientos, programas, recursos y procesos que le ayuden a convertirse en una organización rentable.
- Adopción de un modelo de consumo: pague solo por los recursos informáticos que necesite; aumente o reduzca el uso según los requisitos del negocio, no con previsiones complejas. Por ejemplo, los entornos de desarrollo y pruebas se utilizan normalmente solo ocho horas al día durante la semana laboral. Puede interrumpir estos recursos cuando no se utilicen y obtener así un posible ahorro en costos del 75 % (40 horas frente a 168 horas).
- Medición de la eficiencia general: mida el resultado empresarial de la carga de trabajo y los costes relacionados con la entrega. Use esta medición para conocer las ganancias que obtiene al aumentar la producción y reducir los costos.
- Eliminación del gasto en trabajos pesados no diferenciados: AWS se ocupa del trabajo pesado de las operaciones del centro de datos, como el apilamiento y el suministro de energía a los servidores. También elimina la carga operativa de administrar sistemas operativos y aplicaciones con servicios administrados. De este modo, podrá centrarse en sus clientes y proyectos empresariales en lugar de hacerlo en la infraestructura de TI.
- Análisis y atribución de gastos: la nube facilita la identificación precisa del uso y el coste de los sistemas, lo que permite atribuir de forma transparente los costes de TI a los propietarios de cargas de trabajo individuales. De este modo, le ayuda a medir el retorno de la inversión (ROI) y da a los propietarios de cargas de trabajo la oportunidad de optimizar sus recursos y reducir costos.

Definición

Existen cinco áreas de prácticas recomendadas para la optimización de costes en la nube:

- Práctica de administración financiera en la nube
- Conciencia del gasto y del uso
- Recursos rentables
- Administración de la demanda y suministro de recursos
- Optimización a lo largo del tiempo

Al igual que con los otros pilares dentro de Well-Architected Framework, hay compensaciones que se deben tomar en cuenta; por ejemplo, si se debe optimizar el tiempo de comercialización o el coste. En algunos casos, es mejor optimizar el tiempo de comercialización (salida rápida al mercado, envío de nuevas funciones o simplemente el cumplimiento de una fecha límite) en lugar de invertir en la optimización de costos anticipados. Las decisiones de diseño a veces se guían por la prisa en lugar de basarse en los datos y siempre existe la tentación de sobrecompensar “por si acaso” en lugar de dedicar tiempo para realizar un punto de referencia a fin de definir cuál es la implementación más rentable. Esto puede dar como resultado implementaciones poco optimizadas y con un aprovisionamiento excesivo. Sin embargo, puede ser una opción razonable cuando usted necesite migrar recursos sin modificarlos, desde su entorno en las instalaciones hacia la nube y optimizarlos más adelante. La inversión del esfuerzo adecuado en una estrategia de optimización de costes por adelantado le permite obtener los beneficios económicos de la nube con mayor facilidad. Ello se logra mediante una adhesión constante a las prácticas recomendadas y evitar un aprovisionamiento excesivo innecesario. Las siguientes secciones proporcionan técnicas y prácticas recomendadas para la implementación inicial y continua de la administración financiera de la nube y la optimización de costes para sus cargas de trabajo.

Prácticas recomendadas

Temas

- Práctica de administración financiera en la nube
- Conciencia del gasto y del uso
- Recursos rentables
- Administración de la demanda y suministro de recursos
- Optimización a lo largo del tiempo

Práctica de administración financiera en la nube

Con la adopción de la nube, los equipos de tecnología innovan más rápido porque los ciclos de aprobación, aprovisionamiento y despliegue de la infraestructura son más cortos. Se necesita un nuevo enfoque de la administración financiera en la nube para obtener valor empresarial y éxito financiero. Este enfoque es la administración financiera en la nube que permite desarrollar capacidades en toda la organización implementando su amplio conocimiento organizativo a la hora de diseñar programas, recursos y procesos.

Muchas empresas constan de distintas unidades con distintas prioridades. La habilidad de que la organización siga una serie acordada de objetivos financieros y que disponga de los mecanismos necesarios para cumplirlos hará que sea mucho más eficiente. Una organización capaz innovará y diseñará más rápidamente, será más ágil y se adaptará a factores internos o externos.

En AWS puede usar Cost Explorer, y también Amazon Athena y Amazon QuickSight con el informe de costes y uso (CUR), para que toda la organización sea consciente de los costes y el uso. AWS Budgets proporciona notificaciones proactivas sobre los costes y el uso. Los blogs de AWS aportan información sobre nuevos servicios y funciones para asegurarle de que esté actualizado con las nuevas versiones de los servicios.

Las siguientes preguntas se centran en las consideraciones de la optimización de costes. (Para ver una lista de preguntas y prácticas recomendadas sobre la optimización de costes, consulte el Apéndice).

COST 1: ¿Cómo implementa la administración financiera en la nube?
Implementar la administración financiera en la nube permite a las empresas obtener valor empresarial y éxito financiero al optimizar su coste y uso, y al escalar en AWS.

Al diseñar una función de optimización de costes, trabaje con los miembros y sume expertos de CFM y optimización de costes a los equipos. Los miembros existentes del equipo comprenderán cómo funciona la empresa actualmente y cómo implementar mejoras rápidamente. Piense también en incluir a personas con habilidades adicionales o específicas, como analistas y gestores de proyectos.

A la hora de crear conciencia de los costes en toda la organización, mejore o aproveche los programas y procesos existentes. Es más rápido añadir a lo que ya existe que diseñar procesos y programas nuevos. De este modo verá resultados mucho antes.

Conciencia del gasto y del uso

La mayor flexibilidad y agilidad que permite la nube fomenta la innovación, además del desarrollo y la implementación rápidos. Elimina los procesos manuales y el tiempo asociados al aprovisionamiento de la infraestructura en las instalaciones, incluida la identificación de especificaciones de hardware, la negociación de presupuestos de precios, la administración de órdenes de compra, la programación de envíos y la posterior implementación de los recursos. Sin embargo, la facilidad de uso y la capacidad bajo demanda prácticamente ilimitada requiere una nueva forma de pensar sobre los gastos.

Muchos negocios constan de varios sistemas ejecutados por varios equipos. La capacidad de atribuir costos de recursos a los propietarios de organizaciones individuales o productos impulsa el comportamiento de uso eficiente y ayuda a reducir el desperdicio. La atribución de costos precisa le ayuda a saber qué productos son realmente rentables, y le permite tomar decisiones más informadas sobre dónde asignar presupuesto.

En AWS, puede crear una estructura de cuentas con AWS Organizations o AWS Control Tower, que permite separar los costes y le ayuda a la hora de asignar los costes y el uso. También puede etiquetar los recursos para aplicar información de la organización y empresarial al uso y los costes. Use AWS Cost Explorer para tener mayor visibilidad de los costes y el uso, o cree análisis y paneles personalizados con Amazon Athena y Amazon QuickSight. El control de los costes y el uso se hace mediante notificaciones a través de AWS Budgets y mediante AWS Identity and Access Management (IAM), y Service Quotas.

Las siguientes preguntas se centran en estas consideraciones para la optimización de costes.

COST 2: ¿Cómo controla el uso?

Establezca políticas y mecanismos para garantizar que se incurra en costes apropiados mientras se alcanzan los objetivos. Cuando emplea un enfoque de evaluar la situación, puede innovar sin gastar de más.

COST 3: ¿Cómo supervisa el uso y el coste?

Establezca políticas y procedimientos para monitorear y asignar adecuadamente sus costes. Esto le permite medir y mejorar la rentabilidad de esta carga de trabajo.

COST 4: ¿Cómo retira los recursos?

Implemente control de cambios y administración de recursos desde el inicio del proyecto hasta su finalización. De este modo, garantiza el cierre o la terminación de recursos no utilizados para reducir el desperdicio.

Puede usar etiquetas de asignación de coste para clasificar y hacer un seguimiento del uso y los costes de AWS. Cuando aplica etiquetas a sus recursos de AWS (como instancias EC2 o buckets de

S3), AWS genera un informe de costes y uso con su uso y sus etiquetas. Puede aplicar etiquetas que representen categorías de la organización (como centros de costos, nombres de cargas de trabajo o propietarios) para organizar sus costos en varios servicios.

Asegúrese de usar el nivel de detalle y especificación adecuados al supervisar y crear informes de los costes y el uso. En el caso de la información de alto nivel y las tendencias, use la información diaria con AWS Cost Explorer. Si quiere inspección y análisis en profundidad, use la información por hora en AWS Cost Explorer, o en Amazon Athena y Amazon Quick con el informe de costes y uso (CUR) con información por hora.

La combinación de recursos etiquetados con el seguimiento del ciclo de vida de las entidades (empleados, proyectos) hace posible la identificación de recursos o proyectos huérfanos que ya no generan valor para la organización y deberían retirarse. Puede establecer alertas de facturación para que se le notifique cuando se supere el gasto previsto.

Recursos rentables

El uso de las instancias y los recursos adecuados para su carga de trabajo es clave para ahorrar costos. Por ejemplo, un proceso de informes puede tardar cinco horas en ejecutarse en un servidor pequeño, pero una hora en un servidor más grande que es el doble de caro. Ambos servidores proporcionan el mismo resultado, pero el servidor más pequeño acarrea más costo a lo largo del tiempo.

Una carga de trabajo Well-Architected usa los recursos más rentables, lo que puede suponer un impacto económico positivo notable. También tiene la oportunidad de usar servicios administrados para reducir los costos. Por ejemplo, en lugar de mantener servidores para entregar correos electrónicos, puede usar un servicio que cobre por mensaje.

AWS ofrece una variedad de opciones de precios rentables y flexibles para adquirir instancias de Amazon EC2 y otros servicios de la manera que mejor se adapte a sus necesidades. Las instancias bajo demanda le permiten pagar capacidad informática por horas, sin que exista una tarifa mínima necesaria. Los Savings Plans y las instancias reservadas permiten ahorrar hasta un 75 % en relación con los precios bajo demanda. Con las instancias de spot, puede aprovechar la capacidad de Amazon EC2 que no se utilice y ahorrar hasta un 90 % en relación con los precios bajo demanda. Instancias de spot son adecuadas cuando el sistema puede tolerar el uso de una flota de servidores en la que los servidores pueden aparecer y desaparecer dinámicamente a nivel individual, como los servidores web sin estado, el procesamiento por lotes o al usar HPC y macrodatos.

La selección del servicio apropiado también puede reducir el uso y los costes, como CloudFront para minimizar la transferencia de datos, o eliminar los costes por completo, como el uso de Amazon Aurora en RDS para eliminar los caros costes de licencias de bases de datos.

Las siguientes preguntas se centran en estas consideraciones para la optimización de costes.

COST 5: ¿Cómo evalúa el coste cuando selecciona servicios?

Amazon EC2, Amazon EBS y Amazon S3 son servicios de AWS básicos. Los servicios administrados, como Amazon RDS y Amazon DynamoDB, son servicios de AWS de nivel superior o de aplicación. Cuando selecciona los bloques de creación y los servicios administrados apropiados, puede optimizar esta carga de trabajo para el coste. Por ejemplo, cuando usa servicios administrados, puede reducir o eliminar gran parte de sus gastos administrativos y operativos, lo que le permite trabajar en aplicaciones y actividades relacionadas con el negocio.

COST 6: ¿Cómo cumple los objetivos de costes cuando selecciona el tipo, el tamaño y el número de recursos?

Asegúrese de elegir el tamaño y el número de recursos apropiados para la tarea en cuestión. Al seleccionar el tipo, el tamaño y el número más rentables, minimiza el desperdicio.

COST 7: ¿Cómo utiliza los modelos de precios para reducir los costes?

Use el modelo de fijación de precios más apropiado para sus recursos a fin de minimizar los gastos.

COST 8: ¿Cómo planifica los gastos de transferencia de datos?

Asegúrese de planificar y monitorear los cargos de transferencia de datos para que pueda tomar decisiones de diseño y minimizar los costes. Un cambio de diseño pequeño, pero efectivo, puede reducir drásticamente sus costos operativos con el tiempo.

Al tener en cuenta el coste durante la selección del servicio y usar herramientas como Cost Explorer y AWS Trusted Advisor para revisar regularmente su uso de AWS, puede supervisar activamente su uso y ajustar sus despliegues de acuerdo con ello.

Administración de la demanda y suministro de recursos

Al migrar a la nube, paga solo por lo que necesita. Puede proporcionar recursos para que se adapten a la demanda de carga de trabajo en el momento que se requieran, eliminando así la necesidad de sobreaprovisionamiento, que es algo caro e inútil. También puede modificar la demanda con un límite, un búfer o una cola para suavizar la demanda y usar menos recursos, lo que bajará el coste. También puede procesarla más tarde con un servicio por lotes.

En AWS, puede aprovisionar los recursos automáticamente para que coincidan con la demanda de la carga de trabajo. Auto Scaling y los enfoques basados en la demanda y el tiempo le permiten agregar y quitar recursos según sea necesario. Si puede anticipar los cambios en la demanda, puede ahorrar más dinero y asegurarse de que sus recursos coincidan con las necesidades de su carga de trabajo. Puede usar Amazon API Gateway para implementar una limitación o Amazon SQS para implementar una cola en la carga de trabajo. Ambas herramientas le permitirán modificar la demanda en los componentes de la carga de trabajo.

Las siguientes preguntas se centran en las consideraciones de la optimización de costes.

COST 9: ¿Cómo administra la demanda y aprovisiona los recursos?

Para una carga de trabajo que tenga un gasto y un rendimiento equilibrados, asegúrese de que se use todo lo que paga y evite las instancias de infrautilización significativas. Una métrica de utilizaci ón sesgada en cualquier dirección tiene un efecto adverso en su organización, ya sea en los costes operativos (rendimiento degradado debido a la sobreutilización) o en los gastos desperdic iados de AWS (debido al sobreaprovisionamiento).

Al pensar en modificar la demanda y aprovisionar recursos, tenga muy en cuenta los patrones de uso, el tiempo que se tarda en aprovisionar nuevos recursos y en la predictibilidad del patrón de demanda. Al administrar la demanda, asegúrese de disponer de un búfer o una cola de tamaño adecuado, y de que responde a la demanda de carga de trabajo en el plazo requerido.

Optimización a lo largo del tiempo

A medida que AWS lanza nuevos servicios y funciones, se recomienda revisar sus decisiones de diseño existentes para asegurarse de que sigan siendo las más rentables. A medida que cambian sus requisitos, retire de forma agresiva recursos, servicios enteros y sistemas que ya no necesite.

Implementar nuevas funciones o tipos de recurso puede optimizar poco a poco la carga de trabajo, a la vez que minimiza el esfuerzo requerido para implementar el cambio. Esto proporciona mejoras continuas en la eficiencia a lo largo de tiempo y garantiza que dispone de la tecnología más avanzada para reducir los costes operativos. También puede reemplazar o añadir componentes nuevos a la carga de trabajo con servicios nuevos. Esto puede mejorar significativamente la eficiencia, por lo que es esencial que revise regularmente su carga de trabajo e implemente nuevos servicios y funciones.

Las siguientes preguntas se centran en las consideraciones de la optimización de costes.

COST 10: ¿Cómo evalúa los servicios nuevos?

A medida que AWS lanza nuevos servicios y funciones, se recomienda revisar sus decisiones de diseño existentes para asegurarse de que sigan siendo las más rentables.

Cuando revise regularmente sus implementaciones, evalúe como le pueden ayudar los nuevos servicios a ahorrar dinero. Por ejemplo, Amazon Aurora en RDS puede reducir los costes de las bases de datos relacionales. Usar un servicio sin servidor como Lambda puede eliminar la necesidad de operar y administrar instancias para ejecutar código.

Recursos

Consulte los siguientes recursos para obtener más información sobre nuestras prácticas recomendadas sobre la optimización de costes.

Documentación

- Documentación de AWS

Documento técnico

- Pilar de optimización de costos

Sostenibilidad

El pilar de sostenibilidad se centra en los impactos medioambientales, sobre todo en la eficiencia y el consumo energéticos, ya que son impulsores importantes que ayudan a los arquitectos a promover la adopción de medidas directas destinadas a reducir el uso de los recursos. Encontrará recomendaciones de implementación en el documento técnico Pilar de sostenibilidad.

Temas

- Principios de diseño
- Definición
- Prácticas recomendadas

Principios de diseño

Hay seis principios de diseño para la sostenibilidad en la nube:

- Analice su impacto: Mida la repercusión de su carga de trabajo en la nube y modele el impacto futuro de dicha carga. Incluya todas las fuentes de impacto, incluidas las repercusiones resultantes del uso de sus productos por parte de los clientes y de su eventual cierre y retirada. Revise los recursos y las emisiones que se requieren por unidad de trabajo para comparar el rendimiento productivo con el impacto total de sus cargas de trabajo en la nube. Use estos datos para establecer indicadores clave de rendimiento (KPI), evaluar formas de mejorar la productividad a la vez que se reduce el impacto y calcular la repercusión de los cambios propuestos a lo largo del tiempo.
- Establezca objetivos de sostenibilidad: Establezca objetivos de sostenibilidad a largo plazo, como la reducción de los recursos de computación y almacenamiento requeridos por transacción, para cada una de las cargas de trabajo en la nube. Modele el rendimiento de la inversión de las mejoras de sostenibilidad para las cargas de trabajo existentes y proporcione a los propietarios los recursos que necesitan para invertir en objetivos de sostenibilidad. Planifique el crecimiento y diseñe la arquitectura de sus cargas de trabajo para que ese crecimiento se traduzca en una reducción de la intensidad del impacto y elija una unidad de medida adecuada, por ejemplo, por usuario o por transacción. Los objetivos en general ayudan a respaldar el cumplimiento de los objetivos de sostenibilidad más amplios de su organización o negocio, a identificar las regresiones y a priorizar las áreas susceptibles de mejora.
- Maximice el uso: Aplique el tamaño adecuado a sus cargas de trabajo e implemente un diseño eficaz para garantizar una alta utilización y maximizar la eficiencia energética del hardware

subyacente. Dos hosts que se ejecutan al 30 % son menos eficientes que uno solo que se ejecute al 60 %, debido al consumo energético base por host. Al mismo tiempo, elimine o minimice los recursos inactivos, el procesamiento y el almacenamiento para reducir la energía total necesaria para ejecutar su carga de trabajo.

- Anticípese y adopte nuevas ofertas de hardware y software más eficaces: Admita las mejoras que hagan sus socios y proveedores como ayuda para reducir el impacto de sus cargas de trabajo en la nube. Supervise y evalúe de forma continua las nuevas ofertas de hardware y software más eficaces. Aporte flexibilidad al diseño para permitir una adopción rápida de tecnologías nuevas y eficaces.
- Use servicios administrados: El uso compartido de servicios en una amplia base de clientes ayuda a maximizar la utilización de los recursos, lo cual reduce la cantidad de infraestructura necesaria para admitir las cargas de trabajo en la nube. Por ejemplo, los clientes pueden compartir el impacto de los componentes de un centro de datos común, como la potencia y las redes, mediante la migración de las cargas de trabajo a la Nube de AWS y la adopción de servicios administrados, como AWS Fargate para contenedores sin servidor, en los que AWS opera a escala y es responsable de su funcionamiento eficiente. Use servicios administrados que le ayuden a minimizar su impacto, como pasar datos a los que no se accede con frecuencia a almacenamiento en frío de forma automática con configuraciones del ciclo de vida de Amazon S3 o Amazon EC2 Auto Scaling, a fin de ajustar la capacidad para satisfacer la demanda.
- Reduzca el impacto ulterior de sus cargas de trabajo en la nube: Reduzca la cantidad de energía o los recursos necesarios para usar sus servicios. Reduzca o elimine la necesidad de que los clientes tengan que actualizar sus dispositivos para usar sus servicios. Pruebe a usar granjas de dispositivos para comprender el impacto esperado y realice pruebas con los clientes para que entiendan el impacto real del uso de sus servicios.

Definición

Existen seis áreas de prácticas recomendadas para la sostenibilidad en la nube:

- Selección de regiones
- Patrones de comportamiento de los usuarios
- Patrones de software y arquitectura
- Patrones de datos
- Patrones de hardware
- Proceso de desarrollo e implementación

La sostenibilidad en la nube es un esfuerzo continuo centrado principalmente en la reducción de la energía y la eficiencia en todos los componentes de una carga de trabajo, con lo que se logra el máximo beneficio de los recursos aprovisionados y se minimizan los recursos totales necesarios. Este esfuerzo puede abarcar desde la selección inicial de un lenguaje de programación eficiente, la adopción de algoritmos modernos, el uso de técnicas eficientes de almacenamiento de datos, el despliegue de una infraestructura de computación eficiente y de tamaño correcto, y la minimización de los requisitos de un hardware de alta potencia para el usuario final.

Prácticas recomendadas

Temas

- Selección de regiones
- Patrones de comportamiento de los usuarios
- Patrones de software y arquitectura
- Patrones de datos
- Patrones de hardware
- Proceso de desarrollo y patrones de despliegue
- Recursos

Selección de regiones

Elija las regiones en las que va a implementar sus cargas de trabajo en función tanto de sus requisitos empresariales como de sus objetivos de sostenibilidad.

Las siguientes preguntas se centran en las consideraciones para la sostenibilidad. (Para ver una lista de preguntas y prácticas recomendadas sobre la optimización de costos, consulte el Apéndice).

SUS 1 ¿Cómo selecciona las regiones para respaldar el cumplimiento de sus objetivos de sostenibilidad?

Elija regiones cerca de proyectos de energías renovables de Amazon y regiones en las que la intensidad de carbono recogida en la cuadrícula sea más baja que en otras ubicaciones (o regiones).

Patrones de comportamiento de los usuarios

La forma en que los usuarios consumen sus cargas de trabajo y otros recursos puede ayudarle a identificar las mejoras necesarias para alcanzar sus objetivos de sostenibilidad. Escale la infraestructura para que se ajuste a la carga del usuario de forma continua y asegúrese de que solo se implementen los recursos mínimos necesarios para respaldar a los usuarios. Alinee los niveles de servicio con las necesidades de los clientes. Posicione los recursos de forma que se limite el uso de red necesario para que los usuarios puedan consumirlos. Elimine los recursos existentes que no se utilicen. Identifique los recursos creados que no se utilicen y detenga su generación. Proporcione a los miembros de su equipo dispositivos que satisfagan sus necesidades con un impacto mínimo en la sostenibilidad.

Las siguientes preguntas se centran en las consideraciones para la sostenibilidad:

SUS 2 ¿Cómo puede sacar partido de los patrones de comportamiento de los usuarios para admitir sus objetivos de sostenibilidad?

La forma en que los usuarios consumen sus cargas de trabajo y otros recursos puede ayudarle a identificar las mejoras necesarias para alcanzar sus objetivos de sostenibilidad. Escale la infraestr uctura para que se ajuste a la carga del usuario de forma continua y asegúrese de que solo se implementen los recursos mínimos necesarios para respaldar a los usuarios. Alinee los niveles de servicio con las necesidades de los clientes. Posicione los recursos de forma que se limite el uso de red necesario para que los usuarios puedan consumirlos. Elimine los recursos existente s que no se utilicen. Identifique los recursos creados que no se utilicen y detenga su generació n. Proporcione a los miembros de su equipo dispositivos que satisfagan sus necesidades con un impacto mínimo en la sostenibilidad.

Escalar la infraestructura con la carga del cliente: identifique los períodos de uso reducido o inexistente y escale verticalmente los recursos en consonancia para eliminar el exceso de capacidad y mejorar la eficiencia.

Alinear los acuerdos de nivel de servicio (SLA) con los objetivos de sostenibilidad: defina y actualice los SLA, por ejemplo, para los períodos de retención de datos o la disponibilidad, a fin de minimizar el número de recursos necesarios para admitir la carga de trabajo sin, por ello, dejar de satisfacer los requisitos empresariales.

Eliminar la creación y el mantenimiento de activos que no se utilizan: analice los recursos de aplicaciones (como los informes precompilados, los conjuntos de datos y las imágenes estáticas) y

los patrones de acceso a los recursos para identificar cualquier tipo de redundancia, infrautilización y los posibles objetivos de retirada. Consolide los recursos generados con contenido redundante (por ejemplo, informes mensuales con conjuntos de datos o resultados superpuestos o comunes) para eliminar los recursos consumidos cuando se duplican las salidas. Retire los recursos que no se utilicen (por ejemplo, imágenes de productos que ya no se venden) para liberar recursos consumidos y reducir el número de recursos que se usan para admitir la carga de trabajo.

Optimizar la ubicación geográfica de las cargas de trabajo para las ubicaciones de los usuarios: analice los patrones de acceso a la red para identificar la ubicación geográfica desde la que se conectan los clientes. Seleccione regiones y servicios que acorten la distancia que debe recorrer el tráfico de red a fin de reducir el total de recursos de red necesarios para admitir su carga de trabajo.

Optimizar los recursos de los miembros del equipo para las actividades realizadas: optimice los recursos proporcionados a los miembros del equipo para minimizar el impacto en la sostenibilidad a la vez que se cubren sus necesidades. Por ejemplo, realice las operaciones complejas (como la representación y la compilación) en escritorios en la nube compartidos con un uso intensivo, en lugar de hacerlo en sistemas de usuarios únicos de gran potencia infrautilizados.

Patrones de software y arquitectura

Implemente patrones que permitan suavizar la carga y mantener un uso elevado consistente de los recursos implementados para minimizar los recursos consumidos. Puede haber componentes que queden inactivos debido a la falta de uso relacionada con los cambios en el comportamiento de los usuarios a lo largo del tiempo. Revise los patrones y la arquitectura para consolidar los componentes infrautilizados a fin de incrementar el uso general. Retire los componentes que ya no son necesarios. Analice el rendimiento de los componentes de su carga de trabajo y optimice aquellos que consumen la mayor cantidad de recursos. Tenga en cuenta los dispositivos que usan los clientes para acceder a sus servicios e implemente patrones para minimizar la necesidad de realizar actualizaciones de los dispositivos.

Las siguientes preguntas se centran en las consideraciones para la sostenibilidad.

SUS 3 ¿Cómo puede sacar partido de los patrones de software y de arquitectura para respaldar sus objetivos de sostenibilidad?

Implemente patrones que permitan suavizar la carga y mantener un uso elevado consistente de los recursos implementados para minimizar los recursos consumidos. Puede haber componentes que queden inactivos debido a la falta de uso relacionada con los cambios en el comportamiento

SUS 3 ¿Cómo puede sacar partido de los patrones de software y de arquitectura para respaldar sus objetivos de sostenibilidad?

de los usuarios a lo largo del tiempo. Revise los patrones y la arquitectura para consolidar los componentes infrautilizados a fin de incrementar el uso general. Retire los componentes que ya no son necesarios. Analice el rendimiento de los componentes de su carga de trabajo y optimice aquellos que consumen la mayor cantidad de recursos. Tenga en cuenta los dispositivos que usan los clientes para acceder a sus servicios e implemente patrones para minimizar la necesidad de realizar actualizaciones de los dispositivos.

Optimizar el software y la arquitectura para tareas asíncronas y planificadas: use arquitecturas y diseños de software eficaces para minimizar el promedio de recursos necesarios por unidad de trabajo. Implemente mecanismos que deriven en un uso equilibrado de los componentes para reducir el número de recursos inactivos entre tareas y minimizar el impacto de los picos de carga.

Eliminar o refactorizar los componentes de la carga de trabajo que se usan poco o nada: supervise la actividad de la carga de trabajo para identificar posibles cambios en el uso de los componentes individuales a lo largo del tiempo. Elimine los componentes que ya no se usan ni se necesitan y refactorice aquellos con un uso reducido para limitar los recursos desperdiciados.

Optimizar área de código que consumen más tiempo o recursos: monitoree la actividad de la carga de trabajo para identificar los componentes de aplicación que consuman más recursos. Optimice el código que se ejecuta en estos componentes para minimizar el uso de los recursos y, a la vez, maximizar el rendimiento.

Optimice el efecto en los dispositivos y los equipos de los clientes: analice los dispositivos y equipos que usan los clientes para consumir sus servicios, el ciclo de vida que se espera que tengan y el impacto económico y en la sostenibilidad que supondría reemplazar esos componentes. Implemente patrones de software y arquitecturas que reduzcan al mínimo la necesidad de que los clientes tengan que reemplazar los dispositivos y actualizar los equipos. Por ejemplo, implemente características nuevas que usen código compatible con versiones de sistemas operativos y hardware anteriores o administre el tamaño de las cargas para que no superen la capacidad de almacenamiento del dispositivo de destino.

Usar los patrones de software y las arquitecturas que mejor admitan los datos de acceso y los patrones de almacenamiento: comprender cómo se usan los datos dentro de la carga de trabajo, cómo los consumen los usuarios, se transfieren y se almacenan. Seleccione las tecnologías adecuadas para minimizar los requisitos de almacenamiento y procesamiento de los datos.

Patrones de datos

Implemente patrones que permitan suavizar la carga y mantener un uso elevado consistente de los recursos implementados para minimizar los recursos consumidos. Puede haber componentes que queden inactivos debido a la falta de uso relacionada con los cambios en el comportamiento de los usuarios a lo largo del tiempo. Revise los patrones y la arquitectura para consolidar los componentes infrautilizados a fin de incrementar el uso general. Retire los componentes que ya no son necesarios. Analice el rendimiento de los componentes de su carga de trabajo y optimice aquellos que consumen la mayor cantidad de recursos. Tenga en cuenta los dispositivos que usan los clientes para acceder a sus servicios e implemente patrones para minimizar la necesidad de realizar actualizaciones de los dispositivos.

Las siguientes preguntas se centran en las consideraciones para la sostenibilidad:

SUS 4 ¿Cómo puede sacar partido de los patrones de uso y acceso a los datos para admitir sus objetivos de sostenibilidad?

Implemente prácticas de administración de datos para reducir el almacenamiento aprovisionado que se necesita para admitir la carga de trabajo y los recursos necesarios para su uso. Analice sus datos y use las configuraciones y tecnologías de almacenamiento que mejor admitan el valor empresarial de los datos y la forma en que se usan. Haga que el ciclo de vida de los datos incluya un almacenamiento más eficaz y de menor rendimiento cuando disminuyan los requisitos y elimine los datos que ya no se requieran.

Implementar una política de clasificación de datos: clasifique los datos para entender su importancia con respecto a los resultados empresariales. Use esta información para determinar cuándo puede mover los datos a un almacenamiento de más bajo consumo o bien eliminarlos de forma segura.

Usar tecnologías que respalden el acceso a los datos y los patrones de almacenamiento: use el almacenamiento que mejor respalde la forma en que accede y guarda sus datos a fin de minimizar los recursos aprovisionados para admitir la carga de trabajo. Por ejemplo, los dispositivos de estado sólido (SSD) requieren mucha más energía que las unidades magnéticas y solo deben utilizarse para los casos de uso de datos activos. Use almacenamiento de tipo de archivo de bajo consumo para los datos a los que se accede con poca frecuencia.

Usar políticas de ciclo de vida para eliminar los datos innecesarios: administre el ciclo de vida de todos sus datos e imponga plazos de eliminación de forma automática para minimizar los requisitos de almacenamiento totales de su carga de trabajo.

Minimizar el aprovisionamiento excesivo en el almacenamiento en bloque: para minimizar el almacenamiento total aprovisionado, cree almacenamiento en bloque con asignaciones de tamaño adecuadas para la carga de trabajo. Use volúmenes elásticos para expandir el almacenamiento a medida que crezcan los datos sin necesidad de ajustar el tamaño de almacenamiento asociado a los recursos informáticos. Revise periódicamente los volúmenes elásticos y contraiga los volúmenes con un aprovisionamiento excesivo para adaptarlos al tamaño de datos actual.

Eliminar los datos redundantes e innecesarios: duplique los datos solo cuando sea necesario para minimizar el almacenamiento total consumido. Use tecnologías de copia de seguridad que desdupliquen los datos en el nivel de archivo y de bloque. Limite el uso de configuraciones de matriz redundante de discos independientes (RAID), excepto cuando sea necesario para cumplir los SLA.

Usar sistemas de archivos compartidos o el almacenamiento de objetos para acceder a los datos comunes: adopte el almacenamiento compartido y fuentes de confianza únicas para evitar la duplicación de datos y reducir los requisitos de almacenamiento total de su carga de trabajo. Recupere datos del almacenamiento compartido solo cuando sea necesario. Desconecte los volúmenes que no se utilizan para liberar recursos. Minimice el movimiento de datos entre las redes: use el almacenamiento compartido y acceda a los datos de los almacenes regionales correspondientes para minimizar el total de recursos de redes necesarios para admitir el movimiento de los datos de su carga de trabajo.

Respalde solo los datos que sean difíciles de volver a crear: para minimizar el consumo de almacenamiento, realice copias de seguridad únicamente de aquellos datos que tengan valor empresarial o que sean necesarios para satisfacer los requisitos de cumplimiento. Examine las políticas de copia de seguridad y excluya el almacenamiento efímero que no proporcione valor alguno en un escenario de recuperación.

Patrones de hardware

Realice cambios en sus prácticas de administración de hardware como forma de reducir el impacto en la sostenibilidad de las cargas de trabajo. Minimice la cantidad de hardware necesario para aprovisionar e implementar y seleccione el hardware más eficaz para su carga de trabajo individual.

Las siguientes preguntas se centran en las consideraciones para la sostenibilidad:

SUS 5 ¿Cómo respaldan sus prácticas de uso y de administración de hardware sus objetivos de sostenibilidad?

Realice cambios en sus prácticas de administración de hardware como forma de reducir el impacto en la sostenibilidad de las cargas de trabajo. Minimice la cantidad de hardware necesario para aprovisionar e implementar y seleccione el hardware más eficaz para su carga de trabajo individual.

Usar la cantidad mínima de hardware para cubrir sus necesidades: use las capacidades de la nube para hacer cambios frecuentes en las implementaciones de su carga de trabajo. Actualice los componentes implementados a medida que cambian sus necesidades.

Usar tipos de instancia con el menor impacto: supervise de forma continuada el lanzamiento de nuevos tipos de instancia y aproveche las mejoras de la eficiencia energética; se incluyen los tipos de instancia diseñados para admitir cargas de trabajo específicas, como el entrenamiento y la inferencia en machine learning y la transcodificación de vídeo.

Usar los servicios administrados: los servicios administrados traspasan a AWS la responsabilidad del mantenimiento de un uso medio elevado y la optimización de la sostenibilidad del hardware implementado. Use servicios administrados para distribuir el impacto en la sostenibilidad del servicio entre todos los inquilinos del mismo, lo que reduce su contribución individual.

Optimizar el uso de las unidades de procesamiento gráfico (GPU): las GPU pueden ser el origen de un consumo de alta potencia y muchas de las cargas de trabajo de GPU son sumamente variables, como la representación, la transcodificación, el entrenamiento y el modelado de machine learning. Ejecute las instancias de GPU solo durante el tiempo que sea necesario y retírelas mediante automatización cuando no se requieran para minimizar los recursos consumidos.

Proceso de desarrollo y patrones de despliegue

Realice cambios en sus prácticas de desarrollo, prueba e implementación como forma de reducir el impacto en la sostenibilidad.

Las siguientes preguntas se centran en las consideraciones para la sostenibilidad:

SUS 6 ¿Cómo respaldan sus procesos de desarrollo e implementación sus objetivos de sostenibilidad?

Realice cambios en sus prácticas de desarrollo, prueba e implementación como forma de reducir el impacto en la sostenibilidad.

Adoptar métodos que puedan introducir mejoras de sostenibilidad rápidamente: pruebe y valide las mejoras potenciales antes de implementarlas en producción. Tenga en cuenta el coste de las pruebas al calcular las posibles ventajas futuras de una mejora. Desarrolle métodos de prueba de bajo coste para poder ofrecer pequeñas mejoras.

Mantener la carga de trabajo actualizada: la actualización de sistemas operativos, bibliotecas y aplicaciones puede mejorar la eficiencia de la carga de trabajo y permitir una adopción más sencilla de tecnologías más eficaces. Un software actualizado también puede incluir características que midan el impacto de su carga de trabajo en la sostenibilidad de forma más precisa, ya que los proveedores ofrecen características para cumplir sus objetivos de sostenibilidad propios.

Incrementar el uso de entornos de diseño: use la automatización y la infraestructura como código para incorporar los entornos de preproducción cuando sea necesario, y retirarlos cuando no se utilicen. Un patrón común consiste en programar períodos de disponibilidad que coincidan con las horas de trabajo de los miembros del equipo de desarrollo. La hibernación es una herramienta útil para preservar el estado y habilitar las instancias en línea de forma rápida solo cuando sea necesario. Use tipos de instancia con capacidad de ampliación, instancias de spot, servicios elásticos de base de datos, contenedores y otras tecnologías para alinear la capacidad de desarrollo y prueba con el uso.

Use granjas de dispositivos administrados para pruebas: las granjas de dispositivos administrados reparten el impacto en la sostenibilidad de la fabricación de hardware y del uso de los recursos en varios inquilinos. Las granjas de dispositivos administrados ofrecen diversidad en los tipos de dispositivos para que pueda ofrecer compatibilidad con hardware más antiguo y menos popular y evitar el impacto en la sostenibilidad para el cliente que tienen las actualizaciones innecesarias de los dispositivos.

Recursos

Consulte los siguientes recursos para obtener más información sobre nuestras prácticas recomendadas para la sostenibilidad.

Documento técnico

- Pilar de sostenibilidad

Vídeo

- The Climate Pledge

El proceso de revisión

La revisión de las arquitecturas debe realizarse de manera consistente, con un enfoque sin culpa que fomente la inmersión profunda. Debe ser un proceso rápido (de horas, no días), es decir, una conversación y no una auditoría. El propósito de revisar una arquitectura consiste en identificar cualquier problema crítico que deba abordarse o áreas que podrían mejorarse. El resultado de la revisión es un conjunto de acciones que deberían mejorar la experiencia de un cliente que utiliza la carga de trabajo.

Como se debatió en la sección «Arquitectura local», querrá que cada miembro del equipo asuma la responsabilidad de la calidad de su arquitectura. Recomendamos que los miembros del equipo que construyen una arquitectura usen el Marco de Buena Arquitectura para comprobar continuamente su arquitectura, en lugar de realizar una reunión de revisión formal. Un enfoque continuo permite a los miembros de su equipo actualizar las respuestas a medida que evoluciona la arquitectura y mejorar la arquitectura a medida que ofrece funciones.

AWS Well-Architected Framework se corresponde con la forma en que AWS revisa los sistemas y servicios internamente. Se basa en un conjunto de principios de diseño que influyen en el enfoque arquitectónico y en cuestiones que aseguran que las personas no descuiden las áreas que suelen aparecer en el análisis de causa raíz (RCA). Siempre que haya un problema importante con un sistema interno, un servicio de AWS o un cliente, observamos el RCA para comprobar si podemos mejorar los procesos de revisión que utilizamos.

Las revisiones deben aplicarse a hitos clave en el ciclo de vida del producto, al principio de la fase de diseño, para evitar caminos de un solo sentido difíciles de cambiar, y justo antes de la fecha de lanzamiento. (Muchas decisiones son reversibles, son caminos bidireccionales. Estas decisiones pueden basarse en un proceso superficial. Las caminos unidireccionales son difíciles o imposibles de revertir y requieren más inspección antes de iniciarlos). Tras entrar en producción, su carga de trabajo continuará evolucionando a medida que añada nuevas funciones y cambie las implementaciones de tecnología. La arquitectura de una carga de trabajo cambia con el tiempo. Deberá seguir buenas prácticas de higiene para evitar que sus características arquitectónicas se degraden a medida que evoluciona. Si realiza cambios significativos en la arquitectura, debe seguir un conjunto de procesos de higiene, incluida una revisión Well-Architected.

Si desea utilizar la revisión como una instantánea única o una medición independiente, querrá asegurarse de incluir a todas las personas adecuadas en la reunión. A menudo, descubrimos que las revisiones representan la primera vez que un equipo realmente comprende lo que ha implementado. Un enfoque que funciona bien cuando se revisa la carga de trabajo de otro equipo consiste en

organizar una serie de reuniones informales sobre su arquitectura, donde pueda obtener respuestas a la mayoría de las preguntas. Puede realizar un seguimiento con una o dos reuniones para aclarar o profundizar en áreas de ambigüedad o riesgo percibido.

A continuación, se presentan algunos elementos sugeridos para facilitar sus reuniones:

- Una sala de reuniones con pizarras blancas
- Impresión de diagramas o notas de diseño
- Lista de acciones de preguntas que requieren una investigación innovadora para responder (por ejemplo, ¿habilitamos el cifrado o no?)

Tras haber realizado una revisión, debe tener una lista de problemas que puede priorizar según el contexto de su negocio. También querrá tener en cuenta el impacto de dichos problemas en el trabajo diario de su equipo. Si aborda estos problemas con antelación, podría dedicarle más tiempo a trabajar en la creación de valor empresarial en lugar de resolver problemas recurrentes. A medida que aborde los problemas, puede actualizar su revisión para ver cómo mejora la arquitectura.

Dado que el valor de una revisión es claro tras completarla, es posible que, al principio, un nuevo equipo sea reticente. Aquí hay algunas objeciones que se pueden manejar formando al equipo sobre los beneficios de una revisión:

- «¡Estamos demasiado ocupados!» (Se suele decir cuando el equipo se está preparando para un gran lanzamiento).
  - Si se está preparando para un gran lanzamiento, deseará que vaya perfectamente. La revisión le permitirá detectar cualquier problema que pueda haber pasado por alto.
  - Recomendamos que realice revisiones al principio del ciclo de vida del producto para descubrir riesgos y desarrollar un plan de mitigación conforme a la hoja de ruta de entrega de características.
- «¡No tenemos tiempo para tener en cuenta los resultados!» (Se suele decir cuando el objetivo es un evento inamovible, como la Super Bowl).
  - Estos eventos no se pueden mover. ¿Realmente quiere entrar sin conocer los riesgos para su arquitectura? Incluso si no aborda todos estos problemas, puede tener manuales de estrategias para encargarse de ellos si se materializan.
- «¡No queremos que otros conozcan los secretos de la implementación de nuestra solución!»
  - Si muestra al equipo las preguntas en Well-Architected Framework, verá que ninguna de las preguntas revela información de propiedad comercial ni técnica.

A medida que realiza múltiples revisiones con los equipos de su organización, puede identificar problemas temáticos. Por ejemplo, es posible que descubra que un grupo de equipos tiene problemas en un pilar o tema en particular. Querrá ver todas sus revisiones de manera integral e identificar cualquier mecanismo, formación o charla de ingeniería que puedan ayudar a abordar esas cuestiones temáticas.

Conclusión

AWS Well-Architected Framework proporciona prácticas recomendadas sobre arquitectura en los seis pilares para diseñar y utilizar sistemas en la nube fiables, seguros, eficaces, rentables y sostenibles. El marco proporciona un conjunto de preguntas que le permiten revisar una arquitectura existente o propuesta. También proporciona un conjunto de prácticas recomendadas de AWS para cada pilar. El uso del marco de trabajo en su arquitectura le ayudará a producir sistemas estables y eficaces, lo que le permite centrarse en sus requisitos funcionales.

Colaboradores

Las siguientes personas y organizaciones contribuyeron a redactar este documento:

- Brian Carlson, director de operaciones de Well-Architected, Amazon Web Services
- Ben Potter, Jefe de Seguridad de Well-Architected, Amazon Web Services
- Seth Eliot, jefe de fiabilidad de Well-Architected, Amazon Web Services
- Eric Pullen, arquitecto de soluciones senior, Amazon Web Services
- Rodney Lester, arquitecto principal de soluciones, Amazon Web Services
- Jon Steele, director técnico de cuentas senior, Amazon Web Services
- Max Ramsay, arquitecto principal de soluciones de seguridad, Amazon Web Services
- Sam Elmalak, arquitecto de soluciones, Amazon Web Services
- Aden Leirer, gerente del programa de contenidos de Well-Architected, Amazon Web Services

Otra documentación

Centro de arquitectura de AWS

Cumplimiento en la nube de AWS

Programa de socios de AWS Well-Architected

AWS Well-Architected Tool

Página de inicio de AWS Well-Architected

Documento técnico Pilar de excelencia operativa

Documento técnico Pilar de seguridad

Documento técnico Pilar de fiabilidad

Documento técnico Pilar de eficiencia del rendimiento

Documento técnico Pilar de optimización de costes

Documento técnico Pilar de sostenibilidad

Amazon Builders' Library

Revisiones del documento

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

Cambio	Descripción	Fecha
Actualización importante	Reestructuración importante del pilar de rendimiento para llevar el número de áreas de prácticas recomendadas a cinco. Gran actualización de las prácticas recomendadas y las guías del pilar de seguridad en la Respuesta ante incidentes (SEC 10). Importantes cambios de contenido y consolidación en las áreas de excelencia operativa OPS 04, 05, 06, 08 y 09. Actualizaciones de las guías en los pilares de optimización de costes y fiabilidad . Actualizaciones menores en los niveles de riesgo del pilar de sostenibilidad .	October 3, 2023
Actualizaciones del nuevo marco	Prácticas recomendadas actualizadas con guía prescriptiva y prácticas recomendadas añadidas. Se han añadido nuevas preguntas a los pilares de seguridad y optimización de costes.	April 10, 2023

Actualización menor	Se ha añadido la definició n de nivel de esfuerzo y se han actualizado las prácticas recomendadas del apéndice.	October 20, 2022
Documento técnico actualiza do	Se ha añadido el pilar de sostenibilidad y se han actualizado los enlaces.	December 2, 2021
Actualización importante	Se ha añadido el pilar de sostenibilidad al marco.	November 20, 2021
Actualización menor	Se ha eliminado el lenguaje no inclusivo.	April 22, 2021
Actualización menor	Se han corregido numerosos enlaces.	March 10, 2021
Actualización menor	Se han realizado cambios editoriales mínimos en todo el documento.	July 15, 2020
Actualizaciones del nuevo marco	Se han revisado y reescrito casi todas las preguntas y respuestas.	July 8, 2020
Documento técnico actualiza do	Se han añadido AWS Well-Architected Tool, enlaces a AWS Well-Architected Labs y AWS Well-Architected Partners, y correcciones menores para permitir la versión del marco en múltiples idiomas.	July 1, 2019

Documento técnico actualizado	Revisión y reescritura de la mayoría de las preguntas y respuestas para garantizar que las preguntas se centren en un único tema. Esto provocó que algunas preguntas anteriores se dividieran en preguntas múltiples. Se agregaron términos comunes a las definiciones (carga de trabajo, componente, etc.). Se modificó la presentación de la pregunta en el cuerpo principal para incluir texto descriptivo.	November 1, 2018
Documento técnico actualizado	Actualizaciones para simplificar el texto de las preguntas, estandarizar las respuestas y mejorar la legibilidad.	June 1, 2018
Documento técnico actualizado	La excelencia operativa se trasladó al frente de los pilares y se reescribió para incluir otros pilares. Se han actualizado otros pilares para reflejar la evolución de AWS.	November 1, 2017
Documento técnico actualizado	Se actualizó el marco para incluir el pilar de excelencia operativa y se revisaron y actualizaron los otros pilares para reducir la duplicación e incorporar los aprendizajes de las revisiones a miles de clientes.	November 1, 2016

Actualizaciones menores	Se ha actualizado el apéndice con la información actual de Amazon CloudWatch Logs.	November 1, 2015
Publicación inicial	Publicación de AWS Well-Arch itected Framework.	October 1, 2015

Apéndice: preguntas y prácticas recomendadas

En este apéndice se resumen todas las preguntas y prácticas recomendadas de AWS Well-Architected Framework.

Pilares

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

Excelencia operativa

El pilar de excelencia operativa incluye la capacidad de apoyar el desarrollo y ejecutar cargas de trabajo de forma efectiva, conocer sus operaciones y mejorar continuamente los procesos y procedimientos de soporte para ofrecer valor empresarial. Encontrará recomendaciones de implementación en el documento técnico Pilar de excelencia operativa.

Áreas de prácticas recomendadas

- Organización
- Prepárese
- Operación
- Evolucionar

Organización

Preguntas

- OPERACIÓN 1. ¿Cómo determina cuáles son sus prioridades?
- OPERACIÓN 2. ¿Cómo estructura su organización para respaldar los resultados empresariales?
- OPERACIÓN 3. ¿Cómo ayuda la cultura de su organización a lograr los resultados empresariales?

OPERACIÓN 1. ¿Cómo determina cuáles son sus prioridades?

Todos deben comprender su parte para lograr el éxito empresarial. Tenga objetivos compartidos para establecer prioridades en cuanto a los recursos. Esto maximizará los beneficios de sus esfuerzos.

Prácticas recomendadas

- OPS01-BP01 Evaluar las necesidades externas del cliente
- OPS01-BP02 Evaluar las necesidades internas del cliente
- OPS01-BP03 Evaluar los requisitos de gobernanza
- OPS01-BP04 Evaluar los requisitos de cumplimiento
- OPS01-BP05 Evaluar el panorama de amenazas
- OPS01-BP06 Evaluar compensaciones
- OPS01-BP07 Gestionar beneficios y riesgos

OPS01-BP01 Evaluar las necesidades externas del cliente

Involucre a las partes interesadas clave, incluidos los equipos de negocios, desarrollo y operaciones, para determinar dónde centrar los esfuerzos en las necesidades externas del cliente. Esto asegurará que comprenda a fondo el respaldo operativo que se requiere para lograr los resultados empresariales deseados.

Antipatrones usuales:

- Ha decidido no ofrecer asistencia a los clientes fuera de las horas laborables centrales, pero no ha revisado los datos de solicitud de asistencia históricos. No sabe si esto afectará a sus clientes.
- Está desarrollando una función nueva pero no ha involucrado a sus clientes para saber si les interesa, de qué forma les interesa y sin experiencia para validar la necesidad y la forma de la entrega.

Beneficios de establecer esta práctica recomendada: Es mucho más probable que los clientes cuyas necesidades están satisfechas sigan siendo clientes. La evaluación y la comprensión de las necesidades externas de los clientes le permitirán priorizar sus esfuerzos para aportar valor a la empresa.

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

Guía para la implementación

- Comprender los objetivos empresariales: el éxito de un negocio se debe a metas compartidas y a un entendimiento entre las partes interesadas que abarca a los equipos de negocios, desarrollo y operaciones.
  - Revise los objetivos comerciales, las necesidades y las prioridades de los clientes externos: involucre a las partes interesadas clave, incluidos los equipos de negocios, desarrollo y operaciones, para analizar los objetivos, las necesidades y las prioridades de los clientes externos. Esto garantiza que comprenda a fondo el soporte operativo que se requiere para lograr los resultados de la empresa y de los clientes.
  - Establecer un entendimiento compartido: establezca una comprensión compartida de los roles comerciales de la carga de trabajo, los roles de cada uno de los equipos el manejo de la carga de trabajo y cómo estos apoyan los objetivos comerciales compartidos entre los clientes internos y externos.

Recursos

Documentos relacionados:

- Conceptos de AWS Well-Architected Framework: bucle de retroalimentación

OPS01-BP02 Evaluar las necesidades internas del cliente

Al determinar dónde centrar los esfuerzos en las necesidades internas del cliente, involucre a las partes interesadas clave, incluidos los equipos de negocios, desarrollo y operaciones. Esto asegurará que comprenda exhaustivamente el soporte operacional que se requiere para lograr resultados comerciales.

Utilice sus prioridades establecidas para centrar sus esfuerzos de mejora en las que tendrán el mayor impacto (por ejemplo, el desarrollo de las habilidades del equipo, la mejora del rendimiento de la carga de trabajo, la reducción de los costes, la automatización de los runbooks o la mejora de la monitorización). Actualice sus prioridades a medida que cambien las necesidades.

Antipatrones usuales:

- Ha decidido cambiar las asignaciones de direcciones IP de sus equipos de productos, sin consultarlos, para facilitar la administración de su red. No sabe el impacto que tendrá en sus equipos de productos.

- Está implementando una nueva herramienta de desarrollo, pero no ha involucrado a sus clientes internos para averiguar si es necesaria o si es compatible con sus prácticas actuales.
- Está implementando un nuevo sistema de supervisión, pero no se ha puesto en contacto con sus clientes internos para averiguar si tienen necesidades de supervisión o de elaboración de informes que deban tenerse en cuenta.

Beneficios de establecer esta práctica recomendada: La evaluación y la comprensión de las necesidades de los clientes internos le permitirán priorizar sus esfuerzos para aportar valor a la empresa.

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

Guía para la implementación

- Comprender los objetivos empresariales: el éxito de un negocio se debe a metas compartidas y a un entendimiento entre las partes interesadas que abarca a los equipos de negocios, desarrollo y operaciones.
  - Revise los objetivos comerciales, las necesidades y las prioridades de los clientes internos: involucre a las partes interesadas clave, incluidos los equipos de negocios, desarrollo y operaciones, para analizar los objetivos, las necesidades y las prioridades de los clientes internos. Esto garantiza que comprenda a fondo el soporte operativo que se requiere para lograr los resultados de la empresa y de los clientes.
  - Establecer un entendimiento compartido: establezca una comprensión compartida de los roles comerciales de la carga de trabajo, los roles de cada uno de los equipos el manejo de la carga de trabajo y cómo estos apoyan los objetivos comerciales compartidos entre los clientes internos y externos.

Recursos

Documentos relacionados:

- Conceptos de AWS Well-Architected Framework: bucle de retroalimentación

OPS01-BP03 Evaluar los requisitos de gobernanza

La gobernanza es el conjunto de políticas, normas o marcos que utiliza una empresa para conseguir sus objetivos empresariales. Los requisitos de gobernanza se generan en su organización. Pueden

afectar a los tipos de tecnologías que elija o influir en la forma de utilizar su carga de trabajo. Incorpore los requisitos de gobernanza de la organización a su carga de trabajo. La conformidad es la capacidad de demostrar que ha implementado los requisitos de gobernanza.

Resultado deseado:

- Los requisitos de gobernanza se incorporan al diseño arquitectónico y al funcionamiento de su carga de trabajo.
- Puede aportar pruebas de que ha seguido los requisitos de gobernanza.
- Los requisitos de gobernanza se revisan y actualizan periódicamente.

Patrones comunes de uso no recomendados:

- Su organización exige que la cuenta raíz disponga de autenticación multifactor. No ha implementado este requisito y la cuenta raíz está comprometida.
- Durante el diseño de la carga de trabajo, elige un tipo de instancia que no ha aprobado el departamento de TI. No puede lanzar la carga de trabajo y debe llevar a cabo un rediseño.
- Debe disponer de un plan de recuperación de desastres. No ha creado uno y la carga de trabajo sufre una interrupción prolongada.
- Su equipo quiere utilizar nuevas instancias pero sus requisitos de gobernanza no se han actualizado para permitirlo.

Beneficios de establecer esta práctica recomendada:

- Seguir los requisitos de gobernanza alinea su carga de trabajo con las políticas de la organización.
- Los requisitos de gobernanza reflejan los estándares del sector y las prácticas recomendadas para su organización.

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

Guía para la implementación

Colabore con las partes interesadas y las organizaciones de gobernanza para identificar los requisitos de gobernanza. Incluya los requisitos de gobernanza en su carga de trabajo. Sea capaz de demostrar que ha seguido los requisitos de gobernanza.

Ejemplo de cliente

En AnyCompany Retail, el equipo de operaciones en la nube colabora con las partes interesadas de toda la organización para desarrollar los requisitos de gobernanza. Por ejemplo, prohíben el acceso SSH a las instancias de Amazon EC2. Si los equipos necesitan acceso al sistema, deberán utilizar AWS Systems Manager Session Manager. El equipo de operaciones en la nube actualiza periódicamente los requisitos de gobernanza a medida que hay disponibles nuevos servicios.

Pasos para la implementación

1. Identifique a las partes interesadas para su carga de trabajo, incluidos los equipos centralizados.
2. Colabore con las partes interesadas para identificar los requisitos de gobernanza.
3. Una vez generada la lista, priorice los elementos de mejora y comience a implementarlos en su carga de trabajo.
   a. Utilice servicios como AWS Config para crear gobernanza como código y validar que se siguen los requisitos de gobernanza.
   b. Si utiliza AWS Organizations, puede aprovechar las políticas de control de servicios para implementar los requisitos de gobernanza.
4. Proporcione documentación que valide la implementación.

Nivel de esfuerzo para el plan de implementación: medio. La implementación de los requisitos de gobernanza que faltan puede dar lugar a un reajuste de su carga de trabajo.

Recursos

Prácticas recomendadas relacionadas:

- OPS01-BP04 Evaluar los requisitos de cumplimiento: el cumplimiento es como la gobernanza pero viene de fuera de una organización.

Documentos relacionados:

- AWS Management and Governance Cloud Environment Guide (Guía del entorno de la nube de la administración y la gobernanza de AWS)
- Best Practices for AWS Organizations Service Control Policies in a Multi-Account Environment (Prácticas recomendadas para las políticas de control de servicios de AWS Organizations en un entorno de varias cuentas)
- Governance in the Nube de AWS: The Right Balance Between Agility and Safety (Gobernanza en Nube de AWS: el equilibrio adecuado entre agilidad y seguridad)

- What is Governance, Risk, And Compliance (GRC)? (¿Qué es la gobernanza, el riesgo y el cumplimiento [GRC]?)

Vídeos relacionados:

- AWS Management and Governance: Configuration, Compliance, and Audit - AWS Online Tech Talks (Administración y gobernanza de AWS: configuración, cumplimiento y auditoría - AWS Online Tech Talks)
- AWS re:Inforce 2019: Governance for the Cloud Age (DEM12-R1) (AWS re:Inforce 2019: Gobernanza para la era de la nube [DEM12-R1])
- AWS re:Invent 2020: Achieve compliance as code using AWS Config(AWS re:Invent 2020: Conseguir el cumplimiento como código mediante AWS Config)
- AWS re:Invent 2020: Agile governance on AWS GovCloud (US)(AWS re:Invent 2020: Gobernanza ágil en AWS GovCloud (US))

Ejemplos relacionados:

- AWS Config Conformance Pack Samples (Muestras de paquetes de conformidad de AWS Config)

Servicios relacionados:

- AWS Config
- AWS Organizations: políticas de control de servicios

OPS01-BP04 Evaluar los requisitos de cumplimiento

Los requisitos de cumplimiento normativo, sectorial e interno son un motor importante para definir las prioridades de su organización. Es posible que su marco de cumplimiento le impida utilizar determinadas tecnologías o ubicaciones geográficas. Aplique la diligencia debida si no se identifican marcos de cumplimiento externos. Genere auditorías o informes que validen el cumplimiento.

Si indica que su producto se ajusta a estándares de conformidad específicos, debe tener un proceso interno que garantice el cumplimiento continuo. Algunos ejemplos de estándares de cumplimiento son PCI DSS, FedRAMP e HIPAA. Las estándares de conformidad aplicables se determinan en función de diversos factores, como los tipos de datos que la solución almacena o transmite, o las regiones geográficas compatibles con la solución.

Resultado deseado:

- Los requisitos de cumplimiento normativo, sectorial e interno se incorporan a la selección de arquitectura.
- Puede validar el cumplimiento y generar informes de auditoría.

Patrones comunes de uso no recomendados:

- Algunas partes de su carga de trabajo entran dentro del marco del estándar de seguridad de datos de la industria de tarjetas de pago (PCI DSS), pero su carga de trabajo almacena los datos de las tarjetas de crédito sin cifrar.
- Sus desarrolladores y arquitectos de software desconocen el marco de cumplimiento al que debe adherirse su organización.
- La auditoría anual de sistemas y organizaciones de control (SOC2) de tipo II tendrá lugar en breve y no puede verificar que los controles están aplicados.

Beneficios de establecer esta práctica recomendada:

- Evaluar y comprender los requisitos de cumplimiento se aplican a su carga de trabajo determinarán cómo priorizar sus esfuerzos para ofrecer valor empresarial.
- Elige las ubicaciones y las tecnologías adecuadas que sean congruentes con su marco de cumplimiento.
- Diseñar su carga de trabajo para que pueda auditar le permite demostrar que se atiene a su marco de cumplimiento.

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

Guía para la implementación

La implementación de esta práctica recomendada significa que se incorporan los requisitos de cumplimiento a su proceso de diseño de la arquitectura. Los miembros de su equipo conocen el marco de cumplimiento necesario. Valida el cumplimiento de acuerdo con el marco.

Ejemplo de cliente

AnyCompany Retail almacena la información de las tarjetas de crédito de los clientes. Los desarrolladores del equipo de almacenamiento de tarjetas saben que deben cumplir el marco PCI-

DSS. Han tomado medidas para verificar que la información de las tarjetas de crédito se almacena y se accede a ella de forma segura de acuerdo con el marco PCI-DSS. Cada año colaboran con su equipo de seguridad para validar el cumplimiento.

Pasos para la implementación

1. Colabore con sus equipos de seguridad y gobernanza para determinar qué marcos de cumplimiento sectorial, normativo o interno debe cumplir su carga de trabajo. Incorpore los marcos de cumplimiento a su carga de trabajo.
   a. Valide el cumplimiento continuo de los recursos de AWS con servicios como AWS Compute Optimizer y AWS Security Hub CSPM.
2. Informe a los miembros de su equipo sobre los requisitos de cumplimiento para que puedan utilizar y hacer evolucionar la carga de trabajo de acuerdo con ellos. Los requisitos de cumplimiento deben incluirse en las opciones de arquitectura y tecnología.
3. En función del marco de cumplimiento, puede que deba generar un informe de auditoría o de cumplimiento. Colabore con su organización para automatizar este proceso en la medida de lo posible.
   a. Use servicios como AWS Audit Manager para validar el cumplimiento y generar informes de auditoría.
   b. Puede descargar documentos de seguridad y cumplimiento de AWS con AWS Artifact.

Nivel de esfuerzo para el plan de implementación: medio. La implementación de marcos de cumplimiento puede suponer un desafío. La generación de informes de auditoría o documentos de cumplimiento añade complejidad adicional.

Recursos

Prácticas recomendadas relacionadas:

- SEC01-BP03 Identificar y validar objetivos de control: los objetivos de control de seguridad son una parte importante del cumplimiento general.
- SEC01-BP06 Automatizar la comprobación y validación de controles de seguridad en canalizaciones: como parte de sus canalizaciones, valide los controles de seguridad. También puede generar documentación de cumplimiento para los nuevos cambios.
- SEC07-BP02 Definir controles de protección de datos: muchos marcos de cumplimiento se basan en políticas de gestión y almacenamiento de datos.

- SEC10-BP03: Preparar capacidades forenses: a veces, las capacidades forenses pueden utilizarse para auditar el cumplimiento.

Documentos relacionados:

- Centro de cumplimiento de AWS
- Recursos de cumplimiento de AWS
- Documento técnico de riesgo y conformidad de AWS
- Modelo de responsabilidad compartida de AWS
- Servicios de AWS en el ámbito de los programas de cumplimiento

Vídeos relacionados:

- AWS re:Invent 2020: Achieve compliance as code using AWS Compute Optimizer(AWS re:Invent 2020: Conseguir el cumplimiento como código mediante AWS Compute Optimizer)
- AWS re:Invent 2021 - Cloud compliance, assurance, and auditing (AWS re:Invent 2021: Cumplimiento, garantía y auditoría de la nube)
- AWS Summit ATL 2022 - Implementing compliance, assurance, and auditing on AWS (COP202) (AWS Summit ATL 2022: Implementación del cumplimiento, la garantía y la auditoría en AWS [COP202])

Ejemplos relacionados:

- PCI DSS and AWS Foundational Security Best Practices on AWS(PCI DSS y prácticas recomendadas de seguridad básicas de AWS en AWS)

Servicios relacionados:

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

OPS01-BP05 Evaluar el panorama de amenazas

Evalúe las amenazas de la empresa (por ejemplo, competencia, riesgos y responsabilidades comerciales, riesgos operativos y amenazas a la seguridad de la información) y mantenga la información actual en un registro de riesgos. Incluya el impacto de los riesgos a la hora de determinar dónde centrar los esfuerzos.

El Well-Architected Framework hace hincapié en aprender, medir y mejorar. Proporciona un enfoque coherente para evaluar las arquitecturas e implementar diseños que se escalarán con el tiempo. AWS proporciona AWS Well-Architected Tool para ayudarle a revisar su enfoque antes del desarrollo, el estado de sus cargas de trabajo antes de la producción y el estado de sus cargas de trabajo durante la producción. Puede compararlos con las prácticas recomendadas de arquitectura de AWS más recientes, supervisar el estado general de sus cargas de trabajo y obtener información sobre posibles riesgos.

Los clientes de AWS pueden optar a una revisión guiada de sus cargas de trabajo de misión crítica para medir sus arquitecturas frente a las prácticas recomendadas de AWS Los clientes de Enterprise Support son elegibles para una revisión de operaciones en la nubediseñada para ayudarles a identificar las lagunas en su enfoque para operar en la nube.

La participación de todos los equipos en estas revisiones ayuda a establecer un entendimiento común de sus cargas de trabajo y de cómo los roles del equipo contribuyen al éxito. Las necesidades identificadas a través de la revisión pueden ayudar a dar forma a sus prioridades.

AWS Trusted Advisor es una herramienta que proporciona acceso a un conjunto básico de comprobaciones que recomiendan optimizaciones que pueden ayudar a definir sus sus prioridades. Los clientes de Business y Enterprise Support reciben acceso a comprobaciones adicionales centradas en la seguridad, la fiabilidad, el rendimiento y la optimización de los costos que pueden ayudar a configurar sus prioridades.

Antipatrones usuales:

- En su producto, está utilizando una versión antigua de una biblioteca de software. No está al tanto de las actualizaciones de seguridad de la biblioteca por problemas que pueden tener un impacto no deseado en su carga de trabajo.
- Su competidor acaba de lanzar una versión de su producto que resuelve muchas de las quejas de sus clientes sobre su producto. No ha priorizado la resolución de ninguno de estos problemas conocidos.

- Los reguladores han estado persiguiendo a empresas como la suya que no cumplen con los requisitos legales de conformidad de la normativa. No ha priorizado la conformidad de ninguno de sus requisitos pendientes.

Beneficios de establecer esta práctica recomendada: La identificación y la comprensión de las amenazas para su organización y su carga de trabajo le permiten determinar qué amenazas debe abordar, su prioridad y los recursos necesarios para hacerlo.

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

Guía para la implementación

- Evaluar el panorama de amenazas: evalúe las amenazas a la empresa (por ejemplo, competencia, riesgos y responsabilidades comerciales, riesgos operativos y amenazas a la seguridad de la información), de modo que pueda incluir su impacto al determinar dónde centrar los esfuerzos operativos.
  - Últimos boletines de seguridad de AWS
  - AWS Trusted Advisor
  - Mantener un modelo de amenazas: establezca y mantenga un modelo de amenazas que identifique las amenazas potenciales, las mitigaciones planificadas y en curso, y su prioridad. Revise la probabilidad de que las amenazas se manifiesten en forma de incidentes, el coste de recuperación de dichos incidentes y el daño esperado causado, así como el coste de prevención de los incidentes. Revise las prioridades a medida que cambie el contenido del modelo de amenazas.

Recursos

Documentos relacionados:

- Conformidad de Nube de AWS
- Últimos boletines de seguridad de AWS
- AWS Trusted Advisor

OPS01-BP06 Evaluar compensaciones

Evalúe el impacto de las compensaciones entre los intereses de métodos alternativos en competencia para ayudar a tomar decisiones fundamentadas a la hora de determinar dónde

centrar los esfuerzos o elegir una vía de acción. Por ejemplo, la aceleración de la velocidad de comercialización de las nuevas funciones puede primar sobre la optimización de los costes, o se puede elegir una base de datos relacional para los datos no relacionales para simplificar el esfuerzo de migración de un sistema en lugar de migrar a una base de datos optimizada para su tipo de datos y actualizar su aplicación.

AWS puede ayudarle a instruir a sus equipos sobre AWS y los servicios que ofrece para aumentar la comprensión que tengan sobre cómo las elecciones que hagan pueden tener un impacto en la carga de trabajo. Debe utilizar los recursos proporcionados por AWS Support (Centro de conocimientos de AWS, Foros de debate de AWS y Centro de AWS Support) y Documentación de AWS para instruir a sus equipos. Póngase en contacto con AWS Support a través del Centro de AWS Support para que le ayude con sus preguntas sobre AWS.

AWS también comparte los patrones y prácticas recomendadas que hemos aprendido a través del funcionamiento de AWS en Amazon Builders' Library. Hay una gran variedad de información útil disponible a través del Blog de AWS y del podcast oficial de AWS.

Patrones de uso no recomendados comunes:

- Está utilizando una base de datos relacional para gestionar series temporales y datos no relacionales. Hay opciones de bases de datos que están optimizadas para admitir los tipos de datos que utilice, pero no es consciente de los beneficios porque no ha evaluado las compensaciones entre las soluciones.
- Sus inversores le piden que demuestre el cumplimiento de las normas de seguridad de datos del sector de las tarjetas de pago (PCI DSS). No considera las ventajas y desventajas de satisfacer su solicitud y continuar con sus esfuerzos actuales de desarrollo. En lugar de eso, sigue adelante con sus esfuerzos de desarrollo sin demostrar la conformidad. Sus inversores dejan de apoyar a su empresa por la preocupación sobre la seguridad de la plataforma y las inversiones realizadas.

Beneficios de establecer esta práctica recomendada: comprender las implicaciones y consecuencias de sus elecciones le permite priorizar sus opciones.

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

Guía para la implementación

- Evaluar compensaciones: evalúe el impacto de las compensaciones entre los intereses en competencia para ayudar a tomar decisiones fundamentadas a la hora de determinar dónde

centrar los esfuerzos. Por ejemplo, la aceleración de la velocidad de comercialización de nuevas características podría enfatizarse en la optimización de costes.

- AWS puede ayudarle a instruir a sus equipos sobre AWS y los servicios que ofrece para aumentar la comprensión que tengan sobre cómo las elecciones que hagan pueden tener un impacto en la carga de trabajo. Debe utilizar los recursos proporcionados por AWS Support (Centro de conocimiento de AWS, Foros de debate de AWS, y Centro de AWS Support) y Documentación de AWS para instruir a sus equipos. Póngase en contacto con AWS Support a través del Centro de AWS Support para que le ayude con sus preguntas sobre AWS.
- AWS también comparte los patrones y prácticas recomendadas que hemos aprendido a través del funcionamiento de AWS en la Amazon Builders' Library Hay una gran variedad de información útil disponible a través del Blog de AWS y el Podcast oficial de AWS

Recursos

Documentos relacionados:

- Blog de AWS
- Conformidad de Nube de AWS
- Foros de debate de AWS
- Documentación de AWS
- Centro de conocimientos de AWS
- AWS Support
- Centro de AWS Support
- Amazon Builders' Library
- Podcast oficial de AWS

OPS01-BP07 Gestionar beneficios y riesgos

Gestione beneficios y riesgos para tomar decisiones fundamentadas a la hora de determinar dónde centrar los esfuerzos. Por ejemplo, puede resultar beneficioso implementar una carga de trabajo con problemas sin resolver para que las nuevas características importantes estén disponibles para los clientes. Puede ser posible mitigar los riesgos asociados, o puede ser inaceptable permitir que un riesgo persista, en cuyo caso se tomarán medidas para eliminar el riesgo.

Es posible que, en algún momento, quiera hacer énfasis en un pequeño subconjunto de sus prioridades. Utilice un enfoque equilibrado a largo plazo para asegurar el desarrollo de las

capacidades necesarias y la gestión de riesgos. Actualice sus prioridades a medida que cambien las necesidades.

Antipatrones usuales:

- Ha decidido incluir una biblioteca que hace todo lo que necesita y que uno de sus desarrolladores encontró en Internet. No ha evaluado los riesgos de adoptar esta biblioteca de una fuente desconocida y no sabe si contiene vulnerabilidades o código malicioso.
- Ha decidido desarrollar y desplegar una nueva función en lugar de solucionar un problema existente. No ha evaluado los riesgos de dejar el problema hasta que se despliegue la función y no sabe cuál será el impacto en sus clientes.
- Ha decidido no desplegar una función solicitada con frecuencia por los clientes debido a preocupaciones no especificadas de su equipo de conformidad.

Beneficios de establecer esta práctica recomendada: Identificar los beneficios disponibles de sus opciones y ser consciente de los riesgos para su organización le permite tomar decisiones fundamentadas.

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

Guía para la implementación

- Gestionar beneficios y riesgos: equilibre los beneficios de las decisiones con los riesgos involucrados.
  - Identificar beneficios: identifique los beneficios en función de los objetivos, las necesidades y las prioridades comerciales. Los ejemplos incluyen tiempo de comercialización, seguridad, fiabilidad, rendimiento y costo.
  - Identificar riesgos: identifique los riesgos en función de los objetivos, las necesidades y las prioridades de la empresa. Los ejemplos incluyen tiempo de comercialización, seguridad, fiabilidad, rendimiento y costo.
  - Evaluar los beneficios frente a los riesgos y tomar decisiones fundamentadas: determine el impacto de los beneficios y los riesgos en función de los objetivos, las necesidades y las prioridades de sus partes interesadas clave, incluidos los negocios, el desarrollo y las operaciones. Evalúe el valor del beneficio frente a la probabilidad de que el riesgo se materialice y el costo de su impacto. Por ejemplo, enfatizar en el tiempo de comercialización más que en la fiabilidad puede suponer una ventaja competitiva. Sin embargo, puede dar lugar a una reducción del tiempo de actividad si hay problemas de fiabilidad.

OPERACIÓN 2. ¿Cómo estructura su organización para respaldar los resultados empresariales?

Sus equipos deben comprender su papel en la consecución de los resultados empresariales. Los equipos deben comprender la función que desempeñan en el éxito de otros equipos, así como la que desempeñan los demás equipos en su propio éxito, y tener objetivos en común. Comprender la responsabilidad, la propiedad, cómo se toman las decisiones y quién tiene autoridad para tomarlas ayudará a centrar los esfuerzos y a maximizar los beneficios de sus equipos.

Prácticas recomendadas

- OPS02-BP01 Los recursos han identificado a los propietarios
- OPS02-BP02 Los procesos y procedimientos han identificado a los propietarios
- OPS02-BP03 Las actividades operativas han identificado a los propietarios responsables de su rendimiento
- OPS02-BP04 Los miembros del equipo saben de qué son responsables
- OPS02-BP05 Existen mecanismos para identificar la responsabilidad y la propiedad
- OPS02-BP06 Existen mecanismos para solicitar adiciones, cambios y excepciones
- OPS02-BP07 Las responsabilidades entre los equipos están predefinidas o negociadas

OPS02-BP01 Los recursos han identificado a los propietarios

Los recursos para su carga de trabajo deben tener propietarios identificados para el control de cambios, la resolución de problemas y otras funciones. Se asignan propietarios para las cargas de trabajo, las cuentas, la infraestructura, las plataformas y las aplicaciones. La propiedad se registra mediante herramientas como un registro central o metadatos adjuntos a los recursos. El valor empresarial de los componentes determina los procesos y los procedimientos que se les aplican.

Resultado deseado:

- Los recursos tienen propietarios identificados mediante metadatos o un registro central.
- Los miembros del equipo pueden identificar a quién pertenecen los recursos.
- Las cuentas tienen un único propietario siempre que sea posible.

Patrones comunes de uso no recomendados:

- Los contactos alternativos para sus Cuentas de AWS no están asignados.

- Los recursos carecen de etiquetas que identifiquen a qué equipos pertenecen.
- Tiene una cola ITSM sin una asignación de correo electrónico.
- Dos equipos tienen la propiedad solapada de un elemento fundamental de la infraestructura.

Beneficios de establecer esta práctica recomendada:

- El control de cambios de los recursos resulta sencillo con una propiedad asignada.
- Puede implicar a los propietarios adecuados a la hora de solucionar problemas.

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

Guía para la implementación

Defina qué significa la propiedad para los casos de uso de los recursos en su entorno. La propiedad puede indicar quién supervisa los cambios en el recurso, quién lo respalda durante la resolución de problemas o quién es responsable desde el punto de vista financiero. Especifique y registre los propietarios de los recursos, incluido el nombre, la información de contacto, la organización y el equipo.

Ejemplo de cliente

AnyCompany Retail define la propiedad como el equipo o la persona que se encarga de los cambios y de la asistencia a los recursos. Usa AWS Organizations para administrar sus Cuentas de AWS. Los contactos alternativos de la cuenta se configuran mediante bandejas de entrada de grupo. Cada cola de ITSM se asigna a un alias de correo electrónico. Las etiquetas identifican a quién pertenecen los recursos de AWS. Para otras plataformas e infraestructuras, tiene una página wiki que identifica la propiedad y la información de contacto.

Pasos para la implementación

1. Empiece con la definición de la propiedad de su organización. La propiedad puede implicar quién es el propietario del riesgo del recurso, quién es el propietario de los cambios en el recurso o quién presta asistencia al recurso cuando se solucionan problemas. La propiedad también podría implicar la propiedad financiera o administrativa del recurso.
2. Use AWS Organizations para administrar cuentas. Puede administrar los contactos alternativos de sus cuentas de forma centralizada.
   a. Si utiliza las direcciones de correo electrónico y los números de teléfono de la empresa como información de contacto, será posible comunicarse aunque la persona a la que pertenezca

dicha dirección o teléfono haya abandonado la organización. Por ejemplo, cree listas de correo electrónico diferentes para la facturación, las operaciones y la seguridad, y configure estos contactos como Facturación, Seguridad y Operaciones en cada Cuenta de AWS activa. Las notificaciones de AWS llegarán a diversas personas y se responderán incluso si alguien está de vacaciones, cambia de rol o deja la empresa.

   b. Si una cuenta no está administrada por AWS Organizations, los contactos alternativos de la cuenta permitirán que AWS pueda ponerse en contacto con las personas adecuadas en caso necesario. Configure los contactos alternativos de la cuenta para que apunten a un grupo en lugar de a una persona.

3. Utilice etiquetas para identificar a los propietarios de los recursos de AWS. Puede especificar tanto los propietarios como su información de contacto en etiquetas independientes.
   a. Puede utilizar reglas de AWS Config para imponer que los recursos tengan las etiquetas de propiedad obligatorias.
   b. Si desea una orientación detallada acerca de cómo crear una estrategia de etiquetado para su organización, consulte el documento técnico AWS Tagging Best Practices (Prácticas recomendadas de etiquetado de AWS).
4. Para otros recursos, plataformas e infraestructuras, cree documentación que identifique la propiedad. Todos los miembros del equipo deben poder acceder a ella.

Nivel de esfuerzo para el plan de implementación: bajo. Utilice la información de contacto de la cuenta y las etiquetas para asignar la propiedad de los recursos de AWS. Para otros recursos puede utilizar algo tan simple como una tabla en un wiki para registrar la propiedad y la información de contacto o una herramienta ITSM para asignar la propiedad.

Recursos

Prácticas recomendadas relacionadas:

- OPS02-BP02 Los procesos y procedimientos han identificado a los propietarios: los procesos y los procedimientos para respaldar los recursos dependen de la propiedad de los mismos.
- OPS02-BP04 Los miembros del equipo saben de qué son responsables: los miembros del equipo deben saber de qué recursos son propietarios.
- OPS02-BP05 Existen mecanismos para identificar la responsabilidad y la propiedad: la propiedad se debe poder descubrir mediante mecanismos como las etiquetas o los contactos de la cuenta.

Documentos relacionados:

- Administración de cuentas de AWS: actualización de la información de contacto
- Reglas de AWS Config: etiquetas obligatorias
- AWS Organizations: actualización de contactos alternativos en su organización
- Documento técnico AWS Tagging Best Practices (Prácticas recomendadas de etiquetado de AWS)

Ejemplos relacionados:

- Reglas de AWS Config: Amazon EC2 con etiquetas obligatorias y valores válidos

Servicios relacionados:

- AWS Config
- AWS Organizations

OPS02-BP02 Los procesos y procedimientos han identificado a los propietarios

Comprenda quién tiene la propiedad de la definición de los procesos y procedimientos individuales, por qué se utilizan esos procesos y procedimientos específicos, y por qué existe esa propiedad. Comprender las razones por las que se utilizan procesos y procedimientos específicos permite identificar las oportunidades de mejora.

Beneficios de establecer esta práctica recomendada: comprender la propiedad identifica quién puede aprobar las mejoras, aplicarlas o ambas cosas.

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

Guía para la implementación

- Los procesos y procedimientos tienen propietarios identificados responsables de su definición: registre los procesos y procedimientos utilizados en su entorno y la persona o equipo responsable de la definición de estos.
  - Identificar los procesos y procedimientos: identifique las actividades operativas realizadas en apoyo de sus cargas de trabajo. Documente estas actividades en un lugar accesible.
  - Definir a quién corresponde la definición de un proceso o procedimiento: identifique de forma exclusiva a la persona o equipo responsable de la especificación de una actividad. Son responsables de garantizar que un miembro del equipo con la debida capacitación y con los permisos, el acceso y las herramientas correctas pueda realizarlo con éxito. Si hay

problemas con la realización de la actividad, los miembros del equipo que la llevan a cabo son responsables de proporcionar la información detallada necesaria para mejorar la actividad.

- Capture la propiedad en los metadatos del artefacto de actividad: los procedimientos automatizados en servicios como AWS Systems Manager, a través de documentos, y AWS Lambda, como funciones, admiten la captura de información de metadatos como etiquetas. Capture la propiedad de los recursos mediante etiquetas o grupos de recursos, especificando la propiedad y la información de contacto. Utilice AWS Organizations para crear políticas de etiquetas y asegurar que se captura la información de contacto y de propiedad.

OPS02-BP03 Las actividades operativas han identificado a los propietarios responsables de su rendimiento

Comprenda quién tiene la responsabilidad de realizar actividades específicas en las cargas de trabajo definidas y por qué existe esa responsabilidad. Comprender quién tiene la responsabilidad de realizar las actividades sirve para saber quién llevará a cabo la actividad, validará el resultado y proporcionará información al propietario de la actividad.

Beneficios de establecer esta práctica recomendada: Comprender quién es el responsable de realizar una actividad permite saber a quién hay que notificar cuando se necesita una acción y quién realizará la acción, validará el resultado y proporcionará información al propietario de la actividad.

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

Guía para la implementación

- Las actividades operativas tienen propietarios identificados como responsables de su realización: capture la responsabilidad de realizar los procesos y procedimientos utilizados en su entorno
  - Identificar los procesos y procedimientos: identifique las actividades operativas realizadas en apoyo de sus cargas de trabajo. Documente estas actividades en un lugar accesible.
  - Defina quién es el responsable de realizar cada actividad: identifique al equipo responsable de una actividad. Asegúrese de que disponen de los detalles de la actividad, así como de las habilidades necesarias y de los permisos, accesos y herramientas correctos para realizar la actividad. Deben entender la condición en la que se va a realizar (por ejemplo, en un evento o en una programación). Haga que esta información sea detectable para que los miembros de su organización puedan identificar con quién necesitan ponerse en contacto, ya sea un equipo o una persona, para necesidades específicas.

OPS02-BP04 Los miembros del equipo saben de qué son responsables

Comprender las responsabilidades de su rol y cómo contribuye a los resultados de la empresa sirve para priorizar sus tareas y saber por qué su rol es importante. Esto permite a los miembros del equipo reconocer las necesidades y responder adecuadamente.

Beneficios de establecer esta práctica recomendada: comprender las responsabilidades sirve a la hora de tomar decisiones, emprender acciones y entregar las actividades a sus propietarios.

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

Guía para la implementación

- Asegurarse de que los miembros del equipo conocen sus roles y responsabilidades: identifique los roles y responsabilidades de los miembros del equipo y asegúrese de que comprenden las expectativas de su rol. Haga que esta información sea detectable para que los miembros de su organización puedan identificar con quién necesitan ponerse en contacto, ya sea un equipo o una persona, para necesidades específicas.

OPS02-BP05 Existen mecanismos para identificar la responsabilidad y la propiedad

Cuando no se identifica a ninguna persona o equipo, existen vías de derivación definidas para llegar a alguien con autoridad para asignar la propiedad o planificar que se aborde esa necesidad.

Beneficios de establecer esta práctica recomendada: Comprender quién tiene la responsabilidad o la propiedad le permite dirigirse al equipo o al miembro del equipo adecuado para hacer una solicitud o una transición de una tarea. Tener una persona identificada que tenga la autoridad para asignar la responsabilidad, la propiedad o el plan para abordar las necesidades reduce el riesgo de inacción y de que las necesidades no se resuelvan.

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

Guía para la implementación

- Existen mecanismos para identificar la responsabilidad y la propiedad: proporcione mecanismos accesibles para que los miembros de su organización descubran e identifiquen la propiedad y la responsabilidad. Estos mecanismos les permitirán identificar a quién dirigirse, ya sea un equipo o una persona, para necesidades específicas.

OPS02-BP06 Existen mecanismos para solicitar adiciones, cambios y excepciones

Puede realizar solicitudes a los propietarios de los procesos, procedimientos y recursos. Las solicitudes incluyen adiciones, cambios y excepciones. Estas solicitudes pasan por un proceso de administración de cambios. Tome decisiones fundamentadas para aprobar las solicitudes cuando sean viables y se determine que son adecuadas tras una evaluación de los beneficios y los riesgos.

Resultado deseado:

- Puede realizar solicitudes para cambiar procesos, procedimientos y recursos en función de la propiedad asignada.
- Los cambios se realizan de forma deliberada, valorando las ventajas y los riesgos.

Patrones comunes de uso no recomendados:

- Debe actualizar la forma de desplegar su aplicación, pero no hay forma de solicitar un cambio en el proceso de despliegue al equipo de operaciones.
- El plan de recuperación de desastres debe actualizarse, pero no hay ningún propietario identificado al que solicitar cambios.

Beneficios de establecer esta práctica recomendada:

- Los procesos, los procedimientos y los recursos pueden evolucionar a medida que cambian los requisitos.
- Los propietarios pueden tomar decisiones informadas sobre el momento de realizar cambios.
- Los cambios se realizan de forma deliberada.

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

Guía para la implementación

Para implementar esta práctica recomendada, debe poder solicitar cambios en los procesos, los procedimientos y los recursos. El proceso de administración de los cambios puede ser ligero. Documente el proceso de administración de cambios.

Ejemplo de cliente

AnyCompany Retail utiliza una matriz de asignación de responsabilidades (RACI) para identificar a quién corresponden los cambios en los procesos, los procedimientos y los recursos. Disponen de un proceso de administración de cambios documentado, ligero y fácil de seguir. Con la matriz RACI y el proceso, cualquiera puede enviar solicitudes de cambio.

Pasos para la implementación

1. Identifique los procesos, los procedimientos y los recursos de su carga de trabajo y los responsables de cada uno de ellos. Documéntelos en su sistema de administración de conocimientos.
   a. Si no ha implementado OPS02-BP01 Los recursos han identificado a los propietarios, OPS02-BP02 Los procesos y procedimientos han identificado a los propietarios o OPS02-BP03 Las actividades operativas han identificado a los propietarios responsables de su rendimiento, empiece con ellos en primer lugar.
2. Colabore con las partes interesadas de su organización para desarrollar un proceso de administración de cambios. El proceso debe abarcar las incorporaciones, los cambios y las excepciones de recursos, procesos y procedimientos.
   a. Puede utilizar el Administrador de cambios de AWS Systems Manager como plataforma de administración de cambios para los recursos de carga de trabajo.
3. Documente el proceso de administración de cambios en su sistema de administración de conocimientos.

Nivel de esfuerzo para el plan de implementación: medio. El desarrollo de un proceso de administración de cambios requiere la coordinación con las múltiples partes interesadas de toda la organización.

Recursos

Prácticas recomendadas relacionadas:

- OPS02-BP01 Los recursos han identificado a los propietarios: es necesario identificar a los propietarios de los recursos antes de crear un proceso de administración de cambios.
- OPS02-BP02 Los procesos y procedimientos han identificado a los propietarios: es necesario identificar a los propietarios de los procesos antes de crear un proceso de administración de cambios.

- OPS02-BP03 Las actividades operativas han identificado a los propietarios responsables de su rendimiento: es necesario identificar a los propietarios de las actividades de operaciones antes de crear un proceso de administración de cambios.

Documentos relacionados:

- Recomendaciones de AWS - Guía de estrategias básicas de migraciones grandes de AWS: creación de matrices RACI
- Documento técnico Administración de cambios en la nube

Servicios relacionados:

- Administrador de cambios de AWS Systems Manager

OPS02-BP07 Las responsabilidades entre los equipos están predefinidas o negociadas

Posibilite que se definan o negocien acuerdos entre equipos que describan cómo trabajan y se apoyan mutuamente (por ejemplo, tiempos de respuesta, objetivos de nivel de servicio o acuerdos de nivel de servicio). Los canales de comunicación entre equipos están documentados. Comprender el impacto del trabajo de los equipos en los resultados de la empresa y los resultados de otros equipos y organizaciones fundamenta la priorización de sus tareas y contribuye a que respondan adecuadamente.

Cuando la responsabilidad y la propiedad no están definidas o se desconocen, se corre el riesgo de que no se aborden las actividades necesarias a tiempo y de que se hagan esfuerzos repetidos y potencialmente conflictivos para satisfacer esas necesidades.

Resultado deseado:

- Se acuerdan y documentan los acuerdos de trabajo o asistencia entre equipos.
- Los equipos que se prestan asistencia o colaboran entre sí tienen definidos los canales de comunicación y las expectativas de respuesta.

Patrones comunes de uso no recomendados:

- Se produce un problema en producción y dos equipos distintos empiezan a solucionar los problemas independientemente el uno del otro. Sus esfuerzos aislados prolongan la interrupción.

- El equipo de operaciones necesita ayuda del equipo de desarrollo pero no hay un tiempo de respuesta acordado. La solicitud está atascada en la lista de tareas pendientes.

Beneficios de establecer esta práctica recomendada:

- Los equipos saben cómo interactuar y prestarse asistencia mutua.
- Se conocen las expectativas de capacidad de respuesta.
- Los canales de comunicación están claramente definidos.

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

Guía para la implementación

La implementación de esta práctica recomendada significa que no existe ninguna ambigüedad sobre cómo colaboran los equipos. Los acuerdos formales codifican la forma en que los equipos trabajan juntos o se prestan asistencia mutua. Los canales de comunicación entre equipos están documentados.

Ejemplo de cliente

El equipo de SRE de AnyCompany Retail tiene un acuerdo de nivel de servicio con su equipo de desarrollo. Cada vez que el equipo de desarrollo realiza una solicitud en su sistema de tickets, puede esperar una respuesta en quince minutos. Si se produce una interrupción en el sitio, el equipo de SRE toma la iniciativa en la investigación con la ayuda del equipo de desarrollo.

Pasos para la implementación

1. En colaboración con las partes interesadas de toda su organización, desarrolle acuerdos entre los equipos basados en procesos y procedimientos.
   a. Si dos equipos comparten un proceso o un procedimiento, elabore un runbook sobre cómo colaborarán.
   b. Si existen dependencias entre los equipos, acuerde un SLA de respuesta para las solicitudes.
2. Documente las responsabilidades en su sistema de administración de conocimientos.

Nivel de esfuerzo para el plan de implementación: medio. Si no existen acuerdos entre los equipos, puede resultar difícil llegar a un acuerdo con las partes interesadas de toda la organización.

Recursos

Prácticas recomendadas relacionadas:

- OPS02-BP02 Los procesos y procedimientos han identificado a los propietarios: es necesario identificar la propiedad del proceso antes de establecer acuerdos entre los equipos.
- OPS02-BP03 Las actividades operativas han identificado a los propietarios responsables de su rendimiento: es necesario identificar la propiedad de las actividades de operaciones antes de establecer acuerdos entre los equipos.

Documentos relacionados:

- AWS Executive Insights: potenciar la innovación con el equipo de dos pizzas
- Introducción a DevOps en AWS: equipos de dos pizzas

OPERACIÓN 3. ¿Cómo ayuda la cultura de su organización a lograr los resultados empresariales?

Preste ayuda a los miembros de su equipo para que puedan ser más eficaces a la hora de actuar y respaldar el resultado empresarial.

Prácticas recomendadas

- OPS03-BP01 Respaldo del área ejecutiva
- OPS03-BP02 Los miembros del equipo están capacitados para actuar cuando los resultados están en riesgo
- OPS03-BP03 Se fomenta el traslado a una instancia superior
- OPS03-BP04 Las comunicaciones son oportunas, claras y procesables
- OPS03-BP05 Se fomenta la experimentación
- OPS03-BP06 Los miembros del equipo están capacitados y se les anima a mantener y aumentar sus habilidades
- OPS03-BP07 Dotar a los equipos de los recursos adecuados
- OPS03-BP08 Se fomenta y se busca la diversidad de opiniones en los equipos y entre ellos

OPS03-BP01 Respaldo del área ejecutiva

Los directivos establecen de forma clara las expectativas de la organización y evalúan el éxito. Los directivos son los patrocinadores, defensores e impulsores de la adopción de las prácticas recomendadas y de la evolución de la organización

Beneficios de establecer esta práctica recomendada: directivos comprometidos, expectativas comunicadas con claridad y objetivos compartidos garantizan que los miembros del equipo sepan lo que se espera de ellos. La evaluación del éxito permite identificar los obstáculos que lo impiden para que se puedan abordar mediante la intervención del promotor del respaldo o de sus delegados.

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

Guía para la implementación

- Respaldo del área ejecutiva: los directivos establecen de forma clara las expectativas de la organización y evalúan el éxito. Los directivos son los patrocinadores, defensores e impulsores de la adopción de las prácticas recomendadas y de la evolución de la organización
  - Establezca las expectativas: defina y publique los objetivos de su organización, incluida la forma en que se medirán.
  - Haga un seguimiento de la consecución de los objetivos: mida la consecución incremental de los objetivos con regularidad y comparta los resultados para poder adoptar las medidas adecuadas si los resultados están en peligro.
  - Proporcione los recursos necesarios para alcanzar sus objetivos: revise periódicamente si los recursos siguen siendo adecuados o si son necesarios recursos adicionales en función de nueva información, cambios en los objetivos, responsabilidades o el entorno empresarial.
  - Apoye a sus equipos: manténgase en contacto con sus equipos para entender cómo lo están haciendo y si hay factores externos que les afecten. Cuando sus equipos se vean afectados por factores externos, vuelva a evaluar los objetivos y ajústelos según convenga. Identifique los obstáculos que impiden el progreso de su equipo. Actúe en nombre de sus equipos para resolver los obstáculos y eliminar las cargas innecesarias.
  - Sea un impulsor de la adopción de las prácticas recomendadas: confirme las prácticas recomendadas que han proporcionado beneficios cuantificables y exprese su reconocimiento a los creadores y a los que las han adoptado. Fomente una mayor adopción para magnificar los beneficios conseguidos.
  - Sea impulsor de la evolución de sus equipos: cree una cultura de mejora continua. Fomente el crecimiento y el desarrollo tanto personal como de la organización. Proporcione objetivos a

largo plazo a los que aspirar y que requieran una consecución incremental a lo largo del tiempo. Ajuste esta visión para que se adapte a sus necesidades, a los objetivos empresariales y al entorno empresarial a medida que vayan cambiando.

OPS03-BP02 Los miembros del equipo están capacitados para actuar cuando los resultados están en riesgo

El propietario de la carga de trabajo ha definido la orientación y el alcance facultando a los miembros del equipo a responder cuando los resultados están en riesgo. Los mecanismos de derivación se utilizan para obtener indicaciones cuando los eventos están fuera del ámbito definido.

Beneficios de establecer esta práctica recomendada: al hacer pruebas y validar los cambios con antelación, podrá abordar los problemas con mínimos costes y limitar el impacto en sus clientes. Cuando se hacen pruebas antes del despliegue, se minimizan los errores.

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

Guía para la implementación

- Los miembros del equipo están capacitados para actuar cuando los resultados están en riesgo: proporcione a los miembros de su equipo los permisos, las herramientas y la oportunidad de practicar las habilidades necesarias para responder con eficacia.
  - Proporcionar a los miembros de su equipo la oportunidad de practicar las habilidades necesarias para responder: ofrezca entornos alternativos seguros en los que se puedan probar y entrenar los procesos y procedimientos de forma segura. Realice jornadas de juego para que los miembros del equipo adquieran experiencia en la respuesta a incidentes del mundo real en entornos simulados y seguros.
  - Definir y reconocer la autoridad de los miembros del equipo para actuar: defina específicamente la autoridad de los miembros del equipo para actuar asignando permisos y acceso a las cargas de trabajo y los componentes que soportan. Reconozca que están facultados para actuar cuando los resultados están en riesgo.

OPS03-BP03 Se fomenta el traslado a una instancia superior

Los miembros del equipo disponen de mecanismos y se les anima a trasladar sus preocupaciones a los responsables de la toma de decisiones y a las partes interesadas si creen que los resultados están en peligro. El traslado a una instancia superior debe realizarse de forma temprana y frecuente para poder identificar los riesgos y evitar que provoquen incidentes.

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

Guía para la implementación

- Fomente el traslado a una instancia superior de forma temprana y frecuente: reconozca en el nivel de organización que dicho traslado temprano y frecuente es la práctica recomendada. Reconozca y acepte en el nivel de organización que los traslados pueden ser infundados y que es mejor tener la oportunidad de prevenir un incidente que perder esa oportunidad por no haber hecho el traslado.
  - Disponga de un mecanismo de traslado a un nivel superior: disponga de procedimientos documentados que definan cuándo y cómo debe producirse el traslado. Documente la serie de personas con autoridad en orden ascendente para tomar medidas o aprobar acciones y su información de contacto. El traslado debe continuar hasta que el miembro del equipo esté satisfecho de haber trasladado el riesgo a una persona capaz de abordarlo o haya contactado con la persona a la que pertenece el riesgo y la responsabilidad del funcionamiento de la carga de trabajo. Es esa persona la que en última instancia es propietaria de todas las decisiones con respecto a su carga de trabajo. Los traslados deben incluir la naturaleza del riesgo, la criticidad de la carga de trabajo, quién se ve afectado, cuál es el impacto y la urgencia, es decir, cuándo se espera el impacto.
  - Proteja a los empleados que hacen el traslado a un nivel superior: disponga de una política que proteja a los miembros del equipo de las represalias si realizan el traslado con respecto a un responsable de la toma de decisiones o a una parte interesada que no dé respuesta. Disponga de mecanismos para identificar si esto ocurre y responder de forma adecuada.

OPS03-BP04 Las comunicaciones son oportunas, claras y procesables

Existen y se utilizan mecanismos para avisar a tiempo a los miembros del equipo de los riesgos conocidos y de los eventos planificados. Se proporcionan el contexto, los detalles y el tiempo necesarios (cuando sea posible) para respaldar la determinación de si la acción es necesaria, qué acción se requiere y para tomar medidas de manera oportuna. Por ejemplo, avisar de las vulnerabilidades de software para que se puedan acelerar las revisiones o avisar de las promociones de ventas previstas para que se pueda implementar una congelación de cambios a fin de evitar el riesgo de interrupción del servicio. Los eventos planificados pueden registrarse en un calendario de cambios o de mantenimiento para que los miembros del equipo conozcan las actividades pendientes.

Resultado deseado:

- Las comunicaciones proporcionan contexto, detalles y expectativas de tiempo.

- Los miembros del equipo entienden claramente cuándo y cómo actuar en respuesta a las comunicaciones.
- Aproveche los calendarios de cambios para avisar de los cambios previstos.

Patrones comunes de uso no recomendados:

- Varias veces por semana se produce una alerta que es un falso positivo. Silencia la notificación cada vez que se produce.
- Se le pide que realice un cambio en sus grupos de seguridad pero no se le da una expectativa de cuándo debería producirse.
- Recibe notificaciones constantes en el chat cuando los sistemas se escalan verticalmente pero no es necesaria ninguna acción. Evita el canal de chat y se pierde una notificación importante.
- Se realiza un cambio en la producción sin informar al equipo de operaciones. El cambio desencadena una alerta y se activa el equipo de guardia.

Beneficios de establecer esta práctica recomendada:

- Su organización evita la fatiga por exceso de alertas.
- Los miembros del equipo pueden actuar con el contexto y las expectativas necesarias.
- Los cambios pueden realizarse durante los periodos de cambio, lo que reduce el riesgo.

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

Guía para la implementación

Para implementar esta práctica recomendada, debe colaborar con las partes interesadas de toda su organización para acordar unos estándares de comunicación. Dé a conocer esos estándares a su organización. Identifique y elimine las alertas que sean falsos positivos o que estén siempre activadas. Utilice calendarios de cambios para que los miembros del equipo sepan cuándo se pueden emprender acciones y qué actividades están pendientes. Verifique que las comunicaciones conlleven acciones claras con el contexto necesario.

Ejemplo de cliente

AnyCompany Retail utiliza el chat como principal medio de comunicación. Las alertas y otras informaciones se incluyen en canales específicos. Cuando alguien debe actuar, el resultado deseado

se expone claramente y, en muchos casos, se le proporciona un runbook o una guía de estrategias para que lo utilice. Usa un calendario para programar los cambios importantes en los sistemas de producción.

Pasos para la implementación

1. Analice las alertas para detectar falsos positivos o alertas que se desencadenan continuamente. Elimínelas o modifíquelas para que se desencadenen cuando sea necesaria la intervención manual. Si se desencadena una alerta, proporcione un runbook o una guía de estrategias.
   a. Puede usar Documentos de AWS Systems Manager para crear guías de estrategias y runbooks para las alertas.
2. Existen mecanismos para notificar los riesgos o los acontecimientos previstos de forma clara y procesable, con suficiente antelación para permitir las respuestas adecuadas. Utilice listas de correo electrónico o canales de chat para enviar notificaciones antes de los eventos previstos.
   a. Amazon Q Developer in chat applications puede utilizarse para enviar alertas y responder a eventos desde la plataforma de mensajería de su organización.
3. Proporcionar una fuente de información accesible en la que se puedan consultar los actos programados. Proporcione notificaciones de eventos planificados desde el mismo sistema.
   a. El calendario de cambios de AWS Systems Manager se puede utilizar para crear períodos en los que se pueden producir cambios. De este modo, se avisa a los miembros del equipo de que pueden realizar cambios con seguridad.
4. Supervise las notificaciones de vulnerabilidad y la información sobre revisiones para comprender las vulnerabilidades existentes y los riesgos potenciales asociados a los componentes de su carga de trabajo. Notifique a los miembros del equipo para que puedan actuar.
   a. Puede suscribirse a los boletines de seguridad de AWS para recibir notificaciones sobre vulnerabilidades en AWS.

Recursos

Prácticas recomendadas relacionadas:

- OPS07-BP03 Uso de runbooks para realizar los procedimientos: proporcione un runbook cuando se conozca el resultado para que las comunicaciones sean procesables.
- OPS07-BP04 Usar guías de estrategias para investigar problemas: en el caso de que se desconozca el resultado, las guías de estrategias pueden convertir las comunicaciones en procesables.

Documentos relacionados:

- Boletines de seguridad de AWS
- OpenCVE

Ejemplos relacionados:

- Laboratorios de Well-Architected: administración de inventario y parches (nivel 100)

Servicios relacionados:

- Amazon Q Developer in chat applications
- Calendario de cambios de AWS Systems Manager
- Documentos de AWS Systems Manager

OPS03-BP05 Se fomenta la experimentación

La experimentación es un catalizador para convertir nuevas ideas en productos y características. Acelera el aprendizaje y mantiene a los miembros del equipo interesados y comprometidos. Se anima a los miembros del equipo a experimentar con frecuencia para impulsar la innovación. Incluso cuando se produce un resultado no deseado, tiene valor saber lo que no hay que hacer. No se castiga a los miembros del equipo por experimentos realizados correctamente con resultados no deseados.

Resultado deseado:

- Su organización fomenta la experimentación para impulsar la innovación.
- Los experimentos se utilizan como una oportunidad de aprender.

Patrones comunes de uso no recomendados:

- Desea realizar una prueba A/B pero no existe ningún mecanismo para llevar a cabo el experimento. Despliega un cambio en la interfaz de usuario sin poder probarlo. El resultado es una experiencia negativa para el cliente.
- Su empresa solo tiene un entorno de prueba y producción. No existe un entorno aislado para experimentar con nuevas características o productos, por lo que deberá experimentar en el entorno de producción.

Beneficios de establecer esta práctica recomendada:

- La experimentación impulsa la innovación.
- Puede reaccionar más rápidamente a los comentarios de los usuarios mediante la experimentación.
- Su organización desarrolla una cultura de aprendizaje.

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

Guía para la implementación

Los experimentos se deben realizar de forma segura. Utilice múltiples entornos para experimentar sin poner en peligro los recursos de producción. Utilice las pruebas A/B y las marcas de características para probar experimentos. Proporcione a los miembros del equipo la posibilidad de realizar experimentos en un entorno aislado.

Ejemplo de cliente

AnyCompany Retail fomenta la experimentación. Los miembros del equipo pueden utilizar el 20 % de su semana laboral para experimentar o aprender nuevas tecnologías. Disponen de un entorno aislado en el que pueden innovar. Las pruebas A/B se utilizan para las nuevas características con el fin de validarlas con comentarios de usuarios reales.

Pasos para la implementación

1. Colabore con los directivos de su organización para respaldar la experimentación. Se debe animar a los miembros del equipo a realizar los experimentos de forma segura.
2. Proporcione a los miembros del equipo un entorno en el que puedan experimentar con seguridad. Deben tener acceso a un entorno similar al de producción.
   a. Puede utilizar una Cuenta de AWS independiente para crear un entorno aislado de experimentación. Puede utilizar AWS Control Tower para aprovisionar estas cuentas.
3. Utilice marcas de características y pruebas A/B para experimentar con seguridad y recopilar los comentarios de los usuarios.
   a. AWS AppConfig Feature Flags ofrece la posibilidad de crear marcas de características.
   b. Amazon CloudWatch Evidently se puede usar para ejecutar pruebas A/B en un despliegue limitado.
   c. Puede utilizar las versiones de AWS Lambda para desplegar una nueva versión de una función para pruebas beta.

Nivel de esfuerzo para el plan de implementación: alto. Proporcionar a los miembros del equipo un entorno en el que experimentar y una forma segura de llevar a cabo los experimentos puede requerir una inversión significativa. También es posible que deba modificar el código de la aplicación para utilizar las marcas de características o admitir pruebas A/B.

Recursos

Prácticas recomendadas relacionadas:

- OPS11-BP02 Realizar un análisis después del incidente: aprender de los incidentes es un motor importante para la innovación junto con la experimentación.
- OPS11-BP03 Implementar bucles de retroalimentación: los bucles de comentarios son una parte importante de la experimentación.

Documentos relacionados:

- An Inside Look at the Amazon Culture: Experimentation, Failure, and Customer Obsession (Una mirada al interior de la cultura de Amazon: experimentación, error y obsesión por el cliente)
- Best practices for creating and managing sandbox accounts in AWS(Prácticas recomendadas para crear y administrar cuentas de entorno aislado en AWS)
- Create a Culture of Experimentation Enabled by the Cloud (Crear una cultura de experimentación facilitada por la nube)
- Enabling experimentation and innovation in the cloud at SulAmérica Seguros (Facilitar la experimentación y la innovación en la nube en SulAmérica Seguros)
- Experiment More, Fail Less (Experimentar más, fracasar menos)
- Organización de su entorno de AWS mediante varias cuentas: unidad organizativa de entorno aislado
- Using AWS AppConfig Feature Flags (Uso de AWS AppConfig Feature Flags)

Vídeos relacionados:

- AWS On Air ft. Amazon CloudWatch Evidently | AWS Events (AWS On Air, presentación de Amazon CloudWatch Evidently | Eventos de AWS)
- AWS On Air San Fran Summit 2022 ft. AWS AppConfig Feature Flags integration with Jira

- AWS re:Invent 2022 - A deployment is not a release: Control your launches w/feature flags (BOA305-R) (AWS re:Invent 2022 - Un despliegue no es un lanzamiento: controle sus lanzamientos con marcas de características [BOA305-R])
- Programmatically Create an Cuenta de AWS with AWS Control Tower(Crear mediante programación una Cuenta de AWS con AWS Control Tower)
- Set Up a Multi-Account AWS Environment that Uses Best Practices for AWS Organizations (Configurar un entorno de AWS de varias cuentas que utilice las prácticas recomendadas para AWS Organizations)

Ejemplos relacionados:

- Entorno aislado de innovación de AWS
- Fundamentos de la personalización integral para el comercio electrónico

Servicios relacionados:

- Amazon CloudWatch Evidently
- AWS AppConfig
- AWS Control Tower

OPS03-BP06 Los miembros del equipo están capacitados y se les anima a mantener y aumentar sus habilidades

Los equipos deben aumentar el conjunto de habilidades para adoptar nuevas tecnologías, y para apoyar los cambios en la demanda y las responsabilidades en apoyo de sus cargas de trabajo. El aumento de las competencias en las nuevas tecnologías suele ser una fuente de satisfacción para los miembros del equipo y apoya la innovación. Apoye a los miembros de su equipo para que obtengan y mantengan certificaciones del sector que validen y reconozcan sus crecientes habilidades. Realice una formación interdisciplinar para promover la transferencia de conocimientos y reducir el riesgo de que se produzca un impacto significativo cuando pierda a miembros del equipo cualificados y experimentados con conocimiento institucional. Proporcione un tiempo estructurado dedicado al aprendizaje.

AWS proporciona recursos, que incluyen el Centro de recursos introductorios de AWS, Blogs de AWS, Charla técnica en línea de AWS, Eventos y seminarios web de AWSy Laboratorios de Well-

Architected de AWS, que proporcionan orientación, ejemplos y explicaciones detalladas para formar a sus equipos.

AWS también comparte los patrones y prácticas recomendadas que hemos aprendido a través del funcionamiento de AWS en la Amazon Builders' Library y una gran variedad de material formativo útil a través del Blog de AWS y Podcast oficial de AWS.

Debe aprovechar los recursos formativos que ofrece AWS como los laboratorios de Well-Architected, AWS Support (Centro de conocimientos de AWS, Foros de discusión de AWS y Centro de AWS Support) y Documentación de AWS para instruir a sus equipos Póngase en contacto con AWS Support a través del Centro de AWS Support para que le ayude con sus preguntas sobre AWS.

Formación de AWS and Certification ofrece una formación gratuita a través de cursos digitales autodidactas sobre los fundamentos de AWS. También puede inscribirse en una capacitación adicional dirigida por un instructor para apoyar el desarrollo de las habilidades de AWS de sus equipos.

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

Guía para la implementación

- A los miembros del equipo se les permite y se les anima a mantener y aumentar sus habilidades: Para adoptar nuevas tecnologías, apoyar la innovación y respaldar los cambios en la demanda y las responsabilidades en apoyo de sus cargas de trabajo es necesaria la formación continua.
  - Proporcionar recursos para la formación: proporcione tiempo estructurado dedicado, acceso a materiales de formación, recursos de laboratorio, y apoye la participación en conferencias y organizaciones profesionales que proporcionen oportunidades para aprender, tanto a los educadores como a los compañeros. Proporcione a los miembros del equipo sin experiencia acceso a los miembros del equipo experimentados como mentores o permitirles seguir su trabajo y acceder a sus métodos y habilidades. Fomente el aprendizaje de contenidos no relacionados directamente con el trabajo para tener una perspectiva más amplia.
  - Formación de equipos y compromiso entre equipos: planifique las necesidades de formación continua de los miembros de su equipo. Proporcione oportunidades para que los miembros del equipo se unan a otros equipos (temporal o permanentemente) para compartir habilidades y prácticas recomendadas que beneficien a toda la organización
  - Apoyar la obtención y el mantenimiento de las certificaciones del sector: apoye a los miembros de su equipo para que adquieran y mantengan certificaciones del sector que validen lo que han aprendido y reconozca sus logros.

Recursos

Documentos relacionados:

- Centro de recursos introductorios de AWS
- Blogs de AWS
- Conformidad de Nube de AWS
- Foros de discusión de AWS
- Documentación de AWS
- Charla técnica en línea de AWS
- Eventos y seminarios web de AWS
- Centro de conocimientos de AWS
- AWS Support
- Formación de AWS and Certification
- Laboratorios de Well-Architected de AWS,
- la Amazon Builders' Library
- Podcast oficial de AWS.

OPS03-BP07 Dotar a los equipos de los recursos adecuados

Mantenga la capacidad de los miembros del equipo y proporcione herramientas y recursos para apoyar sus necesidades de carga de trabajo. La sobrecarga de tareas de los miembros del equipo aumenta el riesgo de incidentes derivados de errores humanos. Las inversiones en herramientas y recursos (por ejemplo, aumentar las actividades frecuentes) pueden aumentar la eficacia de su equipo, lo que les permite admitir actividades adicionales.

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

Guía para la implementación

- Dotar a los equipos de los recursos necesarios: asegúrese de conocer el éxito de sus equipos y los factores que contribuyen a su éxito o al fracaso. Actúe para proveer a los equipos de los recursos adecuados.
  - Comprender el rendimiento de los equipos: mida la consecución de resultados operativos y el desarrollo de recursos por parte de sus equipos. Siga los cambios en la producción y la tasa de error a lo largo del tiempo. Implíquese con los equipos para entender los retos relacionados con

el trabajo que les afectan (por ejemplo, el aumento de las responsabilidades, los cambios en la tecnología, la pérdida de personal o el aumento de los clientes a los que se presta asistencia).

- Comprender el impacto en el rendimiento del equipo: manténgase en contacto con sus equipos para entender cómo lo están haciendo y si hay factores externos que les afecten. Cuando sus equipos se vean afectados por factores externos, vuelva a evaluar los objetivos y ajústelos según convenga. Identifique los obstáculos que impiden el progreso de su equipo. Actúe en nombre de sus equipos para ayudar a resolver los obstáculos y eliminar las cargas innecesarias.
- Proporcionar los recursos necesarios para que los equipos tengan éxito: revise periódicamente si los recursos siguen siendo adecuados o si se necesitan recursos adicionales, y haga los ajustes oportunos para apoyar a los equipos.

OPS03-BP08 Se fomenta y se busca la diversidad de opiniones en los equipos y entre ellos

Aproveche la diversidad en toda la organización para buscar múltiples perspectivas únicas. Utilice esta perspectiva para aumentar la innovación, cuestionar sus suposiciones y reducir el riesgo de caer en sesgos de confirmación. Fomente la inclusión, la diversidad y la accesibilidad en sus equipos para obtener perspectivas beneficiosas.

La cultura organizativa tiene un impacto directo en la satisfacción laboral y la retención de los miembros del equipo. Potencie el compromiso y las capacidades de los miembros de su equipo para lograr el éxito de su negocio.

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

Guía para la implementación

- Buscar opiniones y perspectivas diversas: fomente las contribuciones de todos. Dé voz a los grupos infrarrepresentados. Rote los roles y las responsabilidades en las reuniones.
  - Ampliar los roles y las responsabilidades: ofrezca a los miembros del equipo la oportunidad de asumir funciones que de otro modo no podrían desempeñar. Ganarán experiencia y perspectiva gracias al rol, y a las interacciones con nuevos miembros del equipo con los que, de otro modo, no podrían interactuar. Aportarán su experiencia y perspectiva al nuevo rol y a los miembros del equipo con los que interactúen. A medida que aumenta la perspectiva, pueden surgir nuevas oportunidades de negocio o identificarse nuevas oportunidades de mejora. Haga que los miembros de un equipo se turnen en tareas comunes que suelen realizar otros para comprender las exigencias y el impacto de su realización.
  - Proporcionar un entorno seguro y acogedor: disponga de una política y unos controles que protejan la seguridad mental y física de los miembros del equipo dentro de su organización.

Los miembros del equipo deben poder interactuar sin miedo a las represalias. Cuando los miembros del equipo se sienten seguros y acogidos, es más probable que se comprometan y sean productivos. Cuanto más diversa sea su organización, mejor comprenderá a las personas a las que apoya, incluidos sus clientes. Cuando los miembros de su equipo están cómodos, se sienten libres para hablar y confían en que se les escuchará, es más probable que compartan ideas valiosas (por ejemplo, oportunidades de marketing, necesidades de accesibilidad, segmentos de mercado no atendidos, riesgos no reconocidos en su entorno).

- Permitir que los miembros del equipo participen plenamente: proporcione los recursos necesarios para que sus empleados participen plenamente en todas las actividades relacionadas con el trabajo. Los miembros del equipo que se enfrentan a retos diarios han desarrollado habilidades para trabajar en torno a ellos. Estas habilidades desarrolladas de forma única pueden proporcionar un beneficio significativo a su organización. Apoyar a los miembros del equipo con las adaptaciones necesarias aumentará los beneficios que puede recibir de sus contribuciones.

Prepárese

Preguntas

- OPERACIÓN 4. ¿Cómo implementa la observabilidad en su carga de trabajo?
- OPERACIÓN 5. ¿Cómo reduce los defectos, facilita la reparación y mejora el flujo en la producción?
- OPERACIÓN 6. ¿Cómo mitiga los riesgos de implementación?
- Operación 7. ¿Cómo sabe que está listo para soportar una carga de trabajo?

OPERACIÓN 4. ¿Cómo implementa la observabilidad en su carga de trabajo?

Implemente la observabilidad en su carga de trabajo para que pueda comprender su estado y tomar decisiones basadas en datos en función de los requisitos empresariales.

Prácticas recomendadas

- OPS04-BP01 Identificar los indicadores clave de rendimiento
- OPS04-BP02 Implementar telemetría de aplicaciones
- OPS04-BP03 Implementar la telemetría de la experiencia del usuario
- OPS04-BP04 Implementar telemetría de dependencias

- OPS04-BP05 Implementar el rastreo distribuido

OPS04-BP01 Identificar los indicadores clave de rendimiento

La implementación de la observabilidad en su carga de trabajo comienza con la comprensión de su estado y la toma de decisiones basadas en datos en función de los requisitos empresariales. Una de las formas más eficaces de garantizar la alineación entre las actividades de supervisión y los objetivos empresariales consiste en definir y supervisar los indicadores clave de rendimiento (KPI).

Resultado deseado: prácticas de observabilidad eficientes que están estrechamente alineadas con los objetivos empresariales, lo que garantiza que los esfuerzos de supervisión siempre estén al servicio de resultados comerciales tangibles.

Patrones comunes de uso no recomendados:

- Indicadores clave de rendimiento indefinidos: trabajar sin indicadores clave de rendimiento claros puede llevar a una supervisión excesiva o insuficiente y a la pérdida de señales vitales.
- KPI estáticos: no se retienen ni refinan los KPI a medida que evolucionan la carga de trabajo o los objetivos empresariales.
- Desalineación: centrarse en las métricas técnicas que no se correlacionan directamente con los resultados empresariales o que son más difíciles de correlacionar con problemas de la vida real.

Beneficios de establecer esta práctica recomendada:

- Facilidad de identificación de problemas: los KPI empresariales suelen mostrar los problemas con más claridad que las métricas técnicas. Una caída en un KPI empresarial puede identificar un problema de forma más eficaz que analizar numerosas métricas técnicas.
- Alineación empresarial: garantiza que las actividades de supervisión respalden directamente los objetivos empresariales.
- Eficiencia: da prioridad a los recursos de supervisión y presta atención a las métricas que importan.
- Proactividad: detecta y aborda los problemas antes de que tengan implicaciones comerciales más amplias.

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

Guía para la implementación

Para definir de forma eficaz los KPI de la carga de trabajo:

1. Comience con los resultados empresariales: antes de profundizar en las métricas, comprenda los resultados empresariales deseados. ¿Se trata de un aumento de las ventas, una mayor participación de los usuarios o unos tiempos de respuesta más rápidos?
2. Correlacione las métricas técnicas con los objetivos empresariales: No todas las métricas técnicas tienen un impacto directo en los resultados empresariales. Identifique los que sí lo tienen, pero a menudo es más sencillo identificar un problema mediante un KPI empresarial.
3. Utilice Amazon CloudWatch: utilice CloudWatch para definir y supervisar las métricas que representan sus KPI.
4. Revise y actualice periódicamente los KPI: a medida que su carga de trabajo y su empresa evolucionen, mantenga la relevancia de sus KPI.
5. Implique a las partes interesadas: implique a los equipos técnicos y empresariales en la definición y revisión de los KPI.

Nivel de esfuerzo para el plan de implementación: Medio

Recursos

Prácticas recomendadas relacionadas:

- the section called “OPS04-BP02 Implementar telemetría de aplicaciones”
- the section called “OPS04-BP03 Implementar la telemetría de la experiencia del usuario”
- the section called “OPS04-BP04 Implementar telemetría de dependencias”
- the section called “OPS04-BP05 Implementar el rastreo distribuido”

Documentos relacionados:

- AWS Observability Best Practices
- Guía del usuario de CloudWatch
- AWS Observability Skill Builder Course

Vídeos relacionados:

- Developing an observability strategy

Ejemplos relacionados:

- Taller sobre observabilidad

OPS04-BP02 Implementar telemetría de aplicaciones

La telemetría de aplicaciones sirve de base de la observabilidad de su carga de trabajo. Es crucial emitir telemetría que ofrezca información procesable sobre el estado de la aplicación y el logro de los resultados técnicos y empresariales. Desde la solución de problemas hasta la medición del impacto de una nueva característica o la garantía de la alineación con los indicadores clave de rendimiento (KPI) de la empresa, la telemetría de las aplicaciones informa sobre la forma de crear, operar y hacer evolucionar su carga de trabajo.

Las métricas, los registros y los rastreos forman los tres pilares principales de la observabilidad. Sirven como herramientas de diagnóstico que describen el estado de la aplicación. Con el tiempo, ayudan a crear puntos de referencia e identificar anomalías. Sin embargo, para garantizar la alineación entre las actividades de supervisión y los objetivos empresariales, es fundamental definir y supervisar los KPI. Los KPI empresariales suelen facilitar la identificación de los problemas en comparación con las métricas técnicas únicamente.

Otros tipos de telemetría, como la supervisión de usuarios reales (RUM) y las transacciones sintéticas, complementan estos orígenes de datos principales. RUM ofrece información sobre las interacciones de los usuarios en tiempo real, mientras que las transacciones sintéticas simulan los posibles comportamientos de los usuarios, lo que ayuda a detectar los cuellos de botella antes de que los usuarios reales los encuentren.

Resultado deseado: obtenga información útil sobre el rendimiento de su carga de trabajo. Estos conocimientos le permiten tomar decisiones proactivas sobre la optimización del rendimiento, lograr una mayor estabilidad de la carga de trabajo, optimizar los procesos de CI/CD y utilizar los recursos de manera eficaz.

Patrones comunes de uso no recomendados:

- Observabilidad incompleta: no incorporar la observabilidad en todos los niveles de la carga de trabajo, lo que resulta en puntos ciegos que pueden ocultar información vital sobre el rendimiento y el comportamiento del sistema.

- Vista de datos fragmentada: cuando los datos están dispersos en varias herramientas y sistemas, resulta difícil mantener una visión integral del estado y el rendimiento de la carga de trabajo.
- Problemas informados por los usuarios: una señal de que falta una detección proactiva de los problemas mediante la telemetría y la supervisión de los KPI empresariales.

Beneficios de establecer esta práctica recomendada:

- Toma de decisiones informadas: con la información de la telemetría y los KPI empresariales, puede tomar decisiones basadas en datos.
- Mejora de la eficiencia operativa: la utilización de los recursos basada en datos conduce a la rentabilidad.
- Mejora de la estabilidad de la carga de trabajo: detección y resolución de problemas más rápidas, lo que mejora el tiempo de actividad.
- Procesos de CI/CD simplificados: la información obtenida de los datos de telemetría facilita el refinamiento de los procesos y la entrega fiable de código.

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

Guía para la implementación

Para implementar la telemetría de aplicaciones para su carga de trabajo, utilice servicios de AWS como Amazon CloudWatch y AWS X-Ray. Amazon CloudWatch proporciona un conjunto completo de herramientas de supervisión que le permiten observar sus recursos y aplicaciones en entornos locales y de AWS. Recopila, sigue y analiza las métricas, consolida y supervisa los datos de registro y responde a los cambios en los recursos, lo que mejora su comprensión de cómo funciona su carga de trabajo. Al mismo tiempo, AWS X-Ray le permite rastrear, analizar y depurar sus aplicaciones, lo que le proporciona una comprensión profunda del comportamiento de su carga de trabajo. Con características como los mapas de servicios, las distribuciones de latencia y la cronología de rastreo, X-Ray proporciona información sobre el rendimiento de su carga de trabajo y los cuellos de botella que la afectan.

Pasos para la implementación

1. Identifique qué datos debe recopilar: determine las métricas, los registros y los rastreos esenciales que podrían ofrecer información sustancial sobre el estado, el rendimiento y el comportamiento de su carga de trabajo.

2. Despliegue el agente de CloudWatch : el agente de CloudWatch es fundamental a la hora de obtener métricas y registros del sistema y las aplicaciones de su carga de trabajo y su infraestructura subyacente. El agente de CloudWatch también se puede utilizar para recopilar rastreos de X-Ray o OpenTelemetry y enviarlos a X-Ray.
3. Defina y supervise los KPI empresariales: Definir métricas personalizadas que se alineen con sus resultados empresariales.
4. Instrumente su aplicación con AWS X-Ray: además de desplegar el agente de CloudWatch, es crucial que instrumente su aplicación para emitir datos de rastreo. Este proceso puede proporcionar más información sobre el comportamiento y el rendimiento de su carga de trabajo.
5. Estandarice la recopilación de datos en toda su aplicación: estandarice las prácticas de recopilación de datos en toda la aplicación. La uniformidad ayuda a correlacionar y analizar los datos y proporciona una vista completa del comportamiento de la aplicación.
6. Analice los datos y actúe en función de ellos: una vez establecida la recopilación de datos y la normalización, utilice Amazon CloudWatch para el análisis de métricas y registros, y AWS X-Ray para el análisis de rastreos. Este análisis puede proporcionar información crucial sobre el estado, el rendimiento y el comportamiento de su carga de trabajo, lo que guiará su proceso de toma de decisiones.

Nivel de esfuerzo para el plan de implementación: Alto

Recursos

Prácticas recomendadas relacionadas:

- OPS04-BP01 Identificar los indicadores clave de rendimiento
- OPS04-BP03 Implementar la telemetría de la experiencia del usuario
- OPS04-BP04 Implementar telemetría de dependencias
- OPS04-BP05 Implementar el rastreo distribuido

Documentos relacionados:

- AWS Observability Best Practices
- Guía del usuario de CloudWatch
- Guía para desarrolladores de AWS X-Ray

- Instrumenting distributed systems for operational visibility (Instrumentación de los sistemas distribuidos para la visibilidad de las operaciones)
- AWS Observability Skill Builder Course
- Novedades de Amazon CloudWatch
- Novedades de AWS X-Ray

Vídeos relacionados:

- AWS re:Invent 2022 - Observability best practices at Amazon
- AWS re:Invent 2022 - Developing an observability strategy

Ejemplos relacionados:

- Taller sobre observabilidad
- Biblioteca de soluciones de AWS: Monitoreo de aplicaciones con Amazon CloudWatch

OPS04-BP03 Implementar la telemetría de la experiencia del usuario

Es crucial obtener información detallada sobre las experiencias de los clientes y las interacciones con su aplicación. La supervisión de usuarios reales (RUM) y las transacciones sintéticas sirven como herramientas poderosas para este propósito. La RUM proporciona datos sobre las interacciones reales de los usuarios, lo que ofrece una perspectiva sin filtrar de la satisfacción del usuario, mientras que las transacciones sintéticas simulan las interacciones de los usuarios, lo que ayuda a detectar posibles problemas incluso antes de que afecten a los usuarios reales.

Resultado deseado: Una visión integral de la experiencia del cliente, detección proactiva de problemas y optimización de las interacciones de los usuarios para ofrecer experiencias digitales fluidas.

Patrones comunes de uso no recomendados:

- Aplicaciones sin supervisión de usuarios reales (RUM):
  - Retraso en la detección de problemas: sin RUM, es posible que no se dé cuenta de los cuellos de botella o problemas de rendimiento hasta que los usuarios se quejen. Este enfoque reactivo puede provocar la insatisfacción de los clientes.

    - Falta de información sobre la experiencia del usuario: no usar RUM significa perder datos cruciales que muestran cómo los usuarios reales interactúan con su aplicación, lo que limita su capacidad de optimizar la experiencia del usuario.
- Aplicaciones sin transacciones sintéticas:
    - Omisión de casos de periferia: las transacciones sintéticas le ayudan a probar rutas y funciones que los usuarios habituales no suelen utilizar con frecuencia, pero que son fundamentales para determinadas funciones empresariales. Sin ellos, estas rutas podrían funcionar mal y el problema podría pasar desapercibido.
    - Comprobación de problemas cuando no se utiliza la aplicación: las pruebas sintéticas periódicas pueden simular momentos en los que los usuarios reales no interactúan activamente con la aplicación, lo que garantiza que el sistema siempre funcione correctamente.

Beneficios de establecer esta práctica recomendada:

- Detección proactiva de problemas: identifique y aborde los posibles problemas antes de que afecten a los usuarios reales.
- Experiencia de usuario optimizada: los comentarios continuos de la RUM ayudan a refinar y mejorar la experiencia general del usuario.
- Información sobre el rendimiento de los dispositivos y navegadores: comprenda el rendimiento de su aplicación en varios dispositivos y navegadores, lo que permitirá una mayor optimización.
- Flujos de trabajo empresariales validados: las transacciones sintéticas periódicas garantizan que las funcionalidades básicas y las rutas cruciales permanezcan operativas y eficientes.
- Mejora del rendimiento de las aplicaciones: utilice la información recopilada a partir de datos de usuarios reales para mejorar la capacidad de respuesta y la fiabilidad de las aplicaciones.

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

Guía para la implementación

Para utilizar la RUM y las transacciones sintéticas para la telemetría de la actividad del usuario, AWS ofrece servicios como Amazon CloudWatch RUM y Amazon CloudWatch Synthetics. Las métricas, los registros y los rastros, junto con los datos de actividad de los usuarios, proporcionan una vista completa tanto del estado operativo de la aplicación como de la experiencia del usuario.

Pasos para la implementación

1. Despliegue Amazon CloudWatch RUM: integre su aplicación con CloudWatch RUM para recopilar, analizar y presentar datos de usuarios reales.
   a. Utilice la biblioteca de JavaScript de CloudWatch RUM para integrar la RUM con su aplicación.
   b. Configure paneles para visualizar y supervisar los datos de los usuarios reales.
2. Configure CloudWatch Synthetics: cree valores controlados, o rutinas con scripts, que simulen las interacciones de los usuarios con su aplicación.
   a. Defina los flujos de trabajo y las rutas de las aplicaciones fundamentales.
   b. Diseñe valores controlados mediante scripts de CloudWatch Synthetics para simular las interacciones de los usuarios en estas rutas.
   c. Programe y supervise los valores controlados para que se ejecuten a intervalos específicos, lo que garantiza controles de rendimiento coherentes.
3. Analice los datos y actúe en función de ellos: utilice los datos de la RUM y las transacciones sintéticas para obtener información y tomar medidas correctivas cuando se detecten anomalías. Utilice paneles y alarmas de CloudWatch para mantenerse informado.

Nivel de esfuerzo para el plan de implementación: Medio

Recursos

Prácticas recomendadas relacionadas:

- OPS04-BP01 Identificar los indicadores clave de rendimiento
- OPS04-BP02 Implementar telemetría de aplicaciones
- OPS04-BP04 Implementar telemetría de dependencias
- OPS04-BP05 Implementar el rastreo distribuido

Documentos relacionados:

- Guía de Amazon CloudWatch RUM
- Guía de Amazon CloudWatch Synthetics

Vídeos relacionados:

- Optimize applications through end user insights with Amazon CloudWatch RUM

- AWS on Air ft. Real-User Monitoring for Amazon CloudWatch

Ejemplos relacionados:

- Taller sobre observabilidad
- Repositorio Git para Amazon CloudWatch RUM Web Client
- Using Amazon CloudWatch Synthetics to measure page load time

OPS04-BP04 Implementar telemetría de dependencias

La telemetría de dependencias es esencial para supervisar el estado y el rendimiento de los servicios y componentes externos de los que depende su carga de trabajo. Proporciona información valiosa sobre la accesibilidad, los tiempos de espera y otros eventos cruciales relacionados con dependencias como DNS, bases de datos o API de terceros. Al instrumentar su aplicación para que emita métricas, registros y rastreos sobre estas dependencias, obtendrá una comprensión más clara de los posibles cuellos de botella, problemas de rendimiento o errores que podrían afectar a su carga de trabajo.

Resultado deseado: las dependencias en las que se basa su carga de trabajo funcionan según lo previsto, lo que le permite abordar los problemas de forma proactiva y garantizar un rendimiento óptimo de la carga de trabajo.

Patrones comunes de uso no recomendados:

- Pasar por alto las dependencias externas: centrarse únicamente en las métricas internas de las aplicaciones y descuidar las métricas relacionadas con las dependencias externas.
- Falta de supervisión proactiva: esperar a que surjan problemas en lugar de supervisar continuamente el estado y el rendimiento de la dependencia.
- Supervisión en silos: uso de numerosas herramientas de supervisión dispares que pueden generar vistas fragmentadas e incoherentes del estado de la dependencia.

Beneficios de establecer esta práctica recomendada:

- Mejora de la fiabilidad de la carga de trabajo: al garantizar que las dependencias externas estén siempre disponibles y funcionen de manera óptima.
- Detección y resolución de problemas más rápidas: identificar y abordar de forma proactiva los problemas relacionados con las dependencias antes de que afecten a la carga de trabajo.

- Panorámica completa: obtener una visión integral de los componentes internos y externos que influyen en el estado de la carga de trabajo.
- Mejora de la escalabilidad de la carga de trabajo: mediante la comprensión de los límites de escalabilidad y las características de rendimiento de las dependencias externas.

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

Guía para la implementación

Para implementar la telemetría de dependencias, empiece por identificar los servicios, la infraestructura y los procesos de los que depende su carga de trabajo. Cuantifique qué aspecto tienen las buenas condiciones cuando esas dependencias funcionan según lo esperado y, a continuación, determine qué datos se necesitan para medirlas. Con esa información, puede crear paneles y alertas que proporcionen información a sus equipos de operaciones sobre el estado de esas dependencias. Use herramientas de AWS para detectar y cuantificar el efecto cuando las dependencias no pueden satisfacer las necesidades. Revise continuamente su estrategia para que tenga en cuenta los cambios en las prioridades, los objetivos y los conocimientos adquiridos.

Pasos para la implementación

Para implementar la telemetría de dependencias de manera eficaz:

1. Identifique las dependencias externas: colabore con las partes interesadas para identificar las dependencias externas de las que depende su carga de trabajo. Las dependencias externas pueden abarcar servicios como bases de datos externas, API de terceros, rutas de conectividad de red a otros entornos y servicios de DNS. El primer paso para lograr una telemetría de dependencias eficaz es comprender a la perfección cuáles son esas dependencias.
2. Desarrolle una estrategia de supervisión: una vez que tenga una idea clara de sus dependencias externas, diseñe una estrategia de supervisión adaptada a ellas. Esto implica comprender la importancia de cada dependencia, su comportamiento esperado y cualquier acuerdo u objetivo de nivel de servicio (SLA o SLT) asociado. Configure alertas proactivas que le notifiquen los cambios de estado o las desviaciones del rendimiento.
3. Utilice Amazon CloudWatch Internet Monitor: ofrece información sobre Internet global, lo que ayuda a comprender los cortes o interrupciones que podrían afectar a sus dependencias externas.
4. Manténgase informado con Panel de AWS Health: proporciona alertas y guías de corrección cuando se producen eventos en AWS que podrían afectar a sus servicios.

5. Instrumente su aplicación con AWS X-Ray: AWS X-Ray proporciona información sobre el rendimiento de las aplicaciones y sus dependencias subyacentes. Al rastrear las solicitudes de principio a fin, puede identificar cuellos de botella o errores en los servicios o componentes externos en los que se basa su aplicación.
6. Utilice Amazon DevOps Guru: este servicio basado en machine learning identifica problemas operativos, predice cuándo pueden producirse problemas críticos y recomienda medidas concretas. Tiene un valor incalculable para obtener información sobre las dependencias y determinar que no son el origen de los problemas operativos.
7. Supervise periódicamente: supervise continuamente las métricas y los registros relacionados con las dependencias externas. Configure alertas en caso de que se produzca un comportamiento inesperado o una degradación del rendimiento.
8. Valide después de los cambios: siempre que se produzca una actualización o un cambio en alguna de las dependencias externas, valide su rendimiento y compruebe su conformidad con los requisitos de la aplicación.

Nivel de esfuerzo para el plan de implementación: Medio

Recursos

Prácticas recomendadas relacionadas:

- OPS04-BP01 Identificar los indicadores clave de rendimiento
- OPS04-BP02 Implementar telemetría de aplicaciones
- OPS04-BP03 Implementar la telemetría de la experiencia del usuario
- OPS04-BP05 Implementar el rastreo distribuido

Documentos relacionados:

- What is AWS Health?
- Using Amazon CloudWatch Internet Monitor
- Guía para desarrolladores de AWS X-Ray
- Guía del usuario de Amazon DevOps Guru

Vídeos relacionados:

- Visibility into how internet issues impact app performance

- Introduction to Amazon DevOps Guru

Ejemplos relacionados:

- Gaining operational insights with AIOps using Amazon DevOps Guru
- AWS Health Aware

OPS04-BP05 Implementar el rastreo distribuido

El rastreo distribuido ofrece una forma de supervisar y visualizar las solicitudes a medida que atraviesan varios componentes de un sistema distribuido. Al obtener datos de rastreo de numerosos orígenes y analizarlos en una vista unificada, los equipos pueden comprender mejor cómo fluyen las solicitudes, dónde existen los cuellos de botella y dónde deben centrarse los esfuerzos de optimización.

Resultado deseado: obtenga una visión integral de las solicitudes que fluyen por su sistema distribuido, lo que permite una depuración precisa, un rendimiento optimizado y una mejor experiencia del usuario.

Patrones comunes de uso no recomendados:

- Instrumentación incoherente: no todos los servicios de un sistema distribuido están instrumentados para el rastreo.
- Hacer caso omiso de la latencia: centrarse únicamente en los errores y no tener en cuenta la latencia o las degradaciones graduales del rendimiento.

Beneficios de establecer esta práctica recomendada:

- Descripción general completa del sistema: visualización de toda la ruta de las solicitudes, desde la entrada hasta la salida.
- Depuración mejorada: identificación rápida de dónde se producen errores o problemas de rendimiento.
- Mejora de la experiencia del usuario: supervisión y optimización en función de los datos reales del usuario, lo que garantiza que el sistema satisfaga las demandas de la vida real.

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

Guía para la implementación

Comience por identificar todos los elementos de la carga de trabajo que requieren instrumentación. Una vez contabilizados todos los componentes, utilice herramientas como AWS X-Ray y OpenTelemetry para recopilar datos y analizarlos con herramientas como X-Ray y Amazon CloudWatch ServiceLens Map. Realice revisiones periódicas con los desarrolladores y complemente estas conversaciones con herramientas como Amazon DevOps Guru, X-Ray Analytics y X-Ray Insights para sacar a la luz resultados más profundos. Establezca alertas a partir de los datos de rastreo para notificar cuando los resultados, tal como se definen en el plan de supervisión de la carga de trabajo, estén en peligro.

Pasos para la implementación

Para implementar el rastreo distribuido de manera eficaz:

1. Adopte AWS X-Ray: integre X-Ray en su aplicación para obtener información sobre su comportamiento, comprender su rendimiento e identificar los cuellos de botella. Utilice X-Ray Insights para el análisis automático de rastreos.
2. Instrumente sus servicios: compruebe que todos los servicios, desde una función de AWS Lambda a una Instancia de EC2, envíen datos de rastreo. Cuantos más servicios instrumente, más clara será la vista de principio a fin.
3. incorpore supervisión de usuarios reales de CloudWatch y la supervisión sintética: integre la supervisión de usuarios reales (RUM) y la supervisión sintética con X-Ray. Esto permite recoger experiencias de usuario de la vida real y simular las interacciones de los usuarios para identificar posibles problemas.
4. Utilice la agente de CloudWatch: el agente puede enviar rastreos tanto de X-Ray como de OpenTelemetry, lo que mejora la profundidad de la información obtenida.
5. Utilice Amazon DevOps Guru: DevOps Guru utiliza datos de X-Ray, CloudWatch, AWS Config y AWS CloudTrail para proporcionar recomendaciones prácticas.
6. Analice los rastreos: revise periódicamente los datos de rastreo para detectar patrones, anomalías o cuellos de botella que podrían afectar al rendimiento de su aplicación.
7. Configure alertas: configure las alarmas de CloudWatch para detectar patrones inusuales o latencias prolongadas, lo que permite abordar los problemas de forma proactiva.
8. Mejora continua: revise su estrategia de rastreo a medida que se añadan o modifiquen servicios para recoger todos los puntos de datos pertinentes.

Nivel de esfuerzo para el plan de implementación: Medio

Recursos

Prácticas recomendadas relacionadas:

- OPS04-BP01 Identificar los indicadores clave de rendimiento
- OPS04-BP02 Implementar telemetría de aplicaciones
- OPS04-BP03 Implementar la telemetría de la experiencia del usuario
- OPS04-BP04 Implementar telemetría de dependencias

Documentos relacionados:

- Guía para desarrolladores de AWS X-Ray
- Guía del usuario del agente de Amazon CloudWatch
- Guía del usuario de Amazon DevOps Guru

Vídeos relacionados:

- Utilice AWS X-Ray Insights
- AWS on Air ft. Observability: Amazon CloudWatch and AWS X-Ray

Ejemplos relacionados:

- Instrumenting your Application with AWS X-Ray

OPERACIÓN 5. ¿Cómo reduce los defectos, facilita la reparación y mejora el flujo en la producción?

Adopte enfoques que mejoren el flujo de cambios en la producción, que activen la refactorización, la respuesta rápida sobre la calidad y la corrección de errores. Estos aceleran los cambios beneficiosos que se introducen en la producción, limitan los problemas desplegados y logran una rápida identificación y solución de los problemas introducidos a través de las actividades de despliegue.

Prácticas recomendadas

- OPS05-BP01 Usar el control de versiones

- OPS05-BP02 Probar y validar los cambios
- OPS05-BP03 Utilizar sistemas de administración de la configuración
- OPS05-BP04 Utilizar sistemas de administración de compilación y despliegue
- OPS05-BP05 Administrar parches
- OPS05-BP06 Compartir estándares de diseño
- OPS05-BP07 Adoptar prácticas para mejorar la calidad del código
- OPS05-BP08 Usar varios entornos
- OPS05-BP09 Realizar cambios frecuentes, pequeños y reversibles
- OPS05-BP10 Automatizar completamente la integración y el despliegue

OPS05-BP01 Usar el control de versiones

Use el control de versiones para activar el seguimiento de cambios y versiones.

Muchos servicios de AWS ofrecen capacidades de control de versiones. Utilice un sistema de control de revisiones o de orígenes como AWS CodeCommit para administrar el código y otros artefactos, como las plantillas de AWS CloudFormation controladas por versiones de la infraestructura.

Resultado deseado: sus equipos colaboran en el código. Cuando se fusiona, el código es coherente y no se pierde ningún cambio. Los errores se revierten fácilmente mediante el correcto control de versiones.

Patrones comunes de uso no recomendados:

- Ha estado desarrollando y almacenando el código en su estación de trabajo. Ha sufrido un error de almacenamiento irrecuperable en la estación de trabajo y el código se ha perdido.
- Después de sobrescribir el código existente con sus cambios, reinicia la aplicación y ya no está operativa. No puede revertir el cambio.
- Tiene un bloqueo de escritura en un archivo de informe que tiene que editar otra persona. Se pone en contacto con usted para pedirle que deje de trabajar en él para que puedan completar sus tareas.
- Su equipo de investigación ha estado trabajando en un análisis detallado que modela su trabajo futuro. Alguien ha guardado accidentalmente su lista de la compra sobre el informe final. No puede revertir el cambio y tiene que volver a crear el informe.

Beneficios de establecer esta práctica recomendada: Mediante el uso de las capacidades de control de versiones puede revertir fácilmente los estados buenos conocidos y las versiones anteriores, y limitar el riesgo de que se pierdan los activos.

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

Guía para la implementación

Mantenga los activos en repositorios con control de versiones. Esto permite hacer un seguimiento de los cambios, implementar versiones nuevas, detectar cambios en las versiones existentes y volver a versiones anteriores (por ejemplo, revertir a un estado conocido correcto en caso de error). Integre en sus procedimientos las capacidades de control de versiones de sus sistemas de administración de la configuración.

Recursos

Prácticas recomendadas relacionadas:

- OPS05-BP04 Utilizar sistemas de administración de compilación y despliegue

Documentos relacionados:

- What is AWS CodeCommit?

Vídeos relacionados:

- Introduction to AWS CodeCommit

OPS05-BP02 Probar y validar los cambios

Cada cambio desplegado se debe probar para evitar errores en producción. Esta práctica recomendada se centra en probar los cambios desde el control de versiones hasta la creación de artefactos. Además de los cambios en el código de la aplicación, las pruebas deben incluir la infraestructura, la configuración, los controles de seguridad y los procedimientos operativos. Las pruebas adoptan muchas formas, desde las pruebas unitarias hasta el análisis de componentes de software (SCA). Mover las pruebas más a la izquierda en el proceso de integración y entrega del software se traduce en una mayor certeza de la calidad de los artefactos.

Su organización debe desarrollar estándares de prueba para todos los artefactos de software. Las pruebas automatizadas reducen el trabajo y evitan los errores de las pruebas manuales. En algunos

casos puede ser necesario realizar pruebas manuales. Los desarrolladores deben tener acceso a los resultados de las pruebas automatizadas para crear bucles de comentarios que mejoren la calidad del software.

Resultado deseado: los cambios en el software se prueban antes de su entrega. Los desarrolladores tienen acceso a los resultados de las pruebas y las validaciones. Su organización tiene un estándar de pruebas que se aplica a todos los cambios de software.

Patrones comunes de uso no recomendados:

- Despliega un nuevo cambio de software sin realizar ninguna prueba. No funciona en producción, lo que provoca una interrupción del servicio.
- Los nuevos grupos de seguridad se despliegan con CloudFormation sin haberse probado en un entorno de preproducción. Los grupos de seguridad hacen que la aplicación sea inaccesible para los clientes.
- Se modifica un método, pero no hay pruebas unitarias. El software no funciona cuando se despliega en producción.

Beneficios de establecer esta práctica recomendada: Se reduce la tasa de errores en los despliegues de software. Se mejora la calidad del software. Los desarrolladores son más conscientes de la viabilidad de su código. Las políticas de seguridad se pueden desplegar con confianza para respaldar el cumplimiento de la organización. Los cambios en la infraestructura, como las actualizaciones automáticas de las políticas de escalamiento, se prueban con antelación para satisfacer las necesidades de tráfico.

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

Guía para la implementación

Las pruebas se realizan en todos los cambios, desde el código de la aplicación hasta la infraestructura, como parte de su práctica de integración continua. Los resultados de las pruebas se publican para que los desarrolladores tengan comentarios rápidos. Su organización tiene un estándar de pruebas que deben superar todos los cambios.

Ejemplo de cliente

Como parte de su canalización de integración continua, AnyCompany Retail realiza varios tipos de pruebas en todos los artefactos de software. Practica el desarrollo basado en pruebas, por lo que todo el software tiene pruebas unitarias. Una vez creado el artefacto, ejecuta pruebas integrales.

Una vez completada esta primera ronda de pruebas, ejecuta un examen estático de la seguridad de la aplicación, que busca vulnerabilidades conocidas. Los desarrolladores reciben mensajes a medida que se supera cada puerta de prueba. Una vez completadas todas las pruebas, el artefacto de software se almacena en un repositorio de artefactos.

Pasos para la implementación

1. Colabore con las partes interesadas de su organización en el desarrollo de un estándar de pruebas para los artefactos de software. ¿Qué pruebas estándar deben superar todos los artefactos? ¿Hay requisitos de cumplimiento o gobernanza que deban incluirse en la cobertura de las pruebas? ¿Necesita realizar pruebas de calidad del código? Cuando finalicen las pruebas, ¿quién tiene que saberlo?
   a. La arquitectura de referencia de canalizaciones de despliegue de AWS incluye una lista autorizada de tipos de pruebas que pueden realizarse en artefactos de software como parte de una canalización de integración.
2. Instrumente su aplicación con las pruebas necesarias en función de su estándar de pruebas de software. Cada conjunto de pruebas debería completarse en menos de diez minutos. Las pruebas deben ejecutarse como parte de una canalización de integración.
   a. Amazon CodeGuru Reviewer puede probar el código de su aplicación para detectar defectos.
   b. Puede usar el AWS CodeBuild para realizar pruebas en artefactos de software.
   c. AWS CodePipeline puede orquestar sus pruebas de software en una canalización.

Recursos

Prácticas recomendadas relacionadas:

- OPS05-BP01 Usar el control de versiones
- OPS05-BP06 Compartir estándares de diseño
- OPS05-BP10 Automatizar completamente la integración y el despliegue

Documentos relacionados:

- Adopt a test-driven development approach (Adoptar un enfoque de desarrollo basado en pruebas)
- Automated CloudFormation Testing Pipeline with TaskCat and CodePipeline
- Building end-to-end AWS DevSecOps CI/CD pipeline with open source SCA, SAST, and DAST tools

- Getting started with testing serverless applications (Introducción a las pruebas de aplicaciones sin servidor)
- My CI/CD pipeline is my release captain (Mi canalización CI/CD es mi capitán de lanzamiento)
- Documento técnico Practicing Continuous Integration and Continuous Delivery on AWS

Vídeos relacionados:

- AWS re:Invent 2020: Testable infrastructure: Integration testing on AWS
- AWS Summit ANZ 2021 - Driving a test-first strategy with CDK and test driven development
- Testing Your Infrastructure as Code with AWS CDK

Recursos relacionados:

- AWS Deployment Pipeline Reference Architecture - Application
- AWS Kubernetes DevSecOps Pipeline
- Policy as Code Workshop – Test Driven Development (Taller de política como código: desarrollo basado en pruebas)
- Run unit tests for a Node.js application from GitHub by using AWS CodeBuild
- Use Serverspec for test-driven development of infrastructure code (Usar Serverspec para el desarrollo basado en pruebas del código de la infraestructura)

Servicios relacionados:

- Amazon CodeGuru Reviewer
- AWS CodeBuild
- AWS CodePipeline

OPS05-BP03 Utilizar sistemas de administración de la configuración

Utilice sistemas de administración de la configuración para efectuar modificaciones en la configuración y realizar un seguimiento de ellas. Estos sistemas reducen tanto los errores causados por los procesos manuales como el nivel de esfuerzo requerido para implementar los cambios.

La administración de la configuración estática establece valores al inicializar un recurso que se espera que permanezcan constantes durante toda la vida del recurso. Algunos ejemplos son el

establecimiento de la configuración de un servidor web o de aplicaciones en una instancia o la definición de la configuración de un servicio de AWS en la Consola de administración de AWS a través de la AWS CLI.

La administración de la configuración dinámica establece valores en la inicialización que pueden cambiar o se espera que cambien durante la vida de un recurso. Por ejemplo, podría establecer un conmutador de características para activar la funcionalidad en su código a través de un cambio de configuración o cambiar el nivel de detalle del registro durante un incidente para recoger más datos y, después, volver a cambiar tras el incidente, con lo que se eliminan los registros ahora innecesarios y su gasto asociado.

En AWS, puede usar AWS Config para supervisar continuamente las configuraciones de sus recursos de AWS en las cuentas y las regiones. Le ayuda a hacer un seguimiento de su historial de configuración, comprender cómo un cambio de configuración afectaría a otros recursos y auditarlos con respecto a las configuraciones esperadas o deseadas mediante Reglas de AWS Config y paquetes de conformidad de AWS Config.

Si tiene configuraciones dinámicas en sus aplicaciones que se ejecutan en instancias de Amazon EC2, AWS Lambda, contenedores, aplicaciones móviles o dispositivos de IoT, puede utilizar AWS AppConfig para configurarlas, validarlas, desplegarlas y supervisarlas en todos sus entornos.

En AWS, puede crear canalizaciones de integración continua/despliegue continuo (CI/CD) con servicios como Herramientas para desarrolladores de AWS (por ejemplo, AWS CodeCommit, AWS CodeBuild, AWS CodePipeline, AWS CodeDeployy AWS CodeStar).

Resultado deseado: configura, valida y despliega como parte de su proceso de integración continua y entrega continua (CI/CD). Supervisa para validar que las configuraciones sean correctas. Esto minimiza cualquier impacto en los usuarios finales y los clientes.

Patrones comunes de uso no recomendados:

- Actualiza manualmente la configuración del servidor web en toda su flota y varios servidores dejan de responder debido a errores de actualización.
- Actualiza manualmente su flota de servidores de aplicaciones en el transcurso de muchas horas. La incoherencia en la configuración durante el cambio provoca comportamientos inesperados.
- Alguien ha actualizado sus grupos de seguridad y ya no se puede acceder a los servidores web. Sin saber lo que ha cambiado, se pierde mucho tiempo investigando el problema, lo que prolonga el tiempo de recuperación.

- Una configuración de preproducción se introduce en producción a través de CI/CD sin validación. Expone a los usuarios y clientes a datos y servicios incorrectos.

Beneficios de establecer esta práctica recomendada: La adopción de sistemas de administración de la configuración reduce el nivel de esfuerzo para realizar cambios y hacer un seguimiento de los mismos, así como la frecuencia de los errores provocados por los procedimientos manuales. Los sistemas de administración de la configuración ofrecen garantías con respecto a la gobernanza, el cumplimiento y los requisitos reglamentarios.

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

Guía para la implementación

Los sistemas de administración de la configuración se utilizan para seguir e implementar cambios en las configuraciones de las aplicaciones y el entorno. Los sistemas de administración de la configuración también se utilizan para reducir los errores causados por los procesos manuales, hacer que los cambios de configuración sean repetibles y auditables y reducir el nivel de esfuerzo.

Pasos para la implementación

1. Identifique a los propietarios de la configuración.
    a. Haga que los propietarios de las configuraciones estén al tanto de cualquier necesidad de cumplimiento, gobernanza o normativa.
2. Identifique los elementos de configuración y los resultados.
    a. Los elementos de configuración son todas las configuraciones de las aplicaciones y los entornos afectadas por un despliegue dentro de su canalización de CI/CD.
    b. Los resultados incluyen los criterios de éxito, la validación y lo que se debe supervisar.
3. Seleccione herramientas para la administración de la configuración en función de los requisitos empresariales y el proceso de entrega.
4. Considere la posibilidad de utilizar despliegues ponderados, como los despliegues de valores controlados, para realizar cambios de configuración significativos a fin de minimizar el impacto de las configuraciones incorrectas.
5. Integre la administración de la configuración en su canalización de CI/CD.
6. Valide todos los cambios introducidos.

Recursos

Prácticas recomendadas relacionadas:

- OPS06-BP01 Planificar para hacer frente a los cambios infructuosos
- OPS06-BP02 Despliegues de prueba
- OPS06-BP03 Emplear estrategias de despliegue seguros
- OPS06-BP04 Automatizar las pruebas y la reversión

Documentos relacionados:

- AWS Control Tower
- Acelerador de zonas de aterrizaje de AWS
- AWS Config
- What is AWS Config?
- AWS AppConfig
- What is AWS CloudFormation?
- Herramientas para desarrolladores de AWS

Vídeos relacionados:

- AWS re:Invent 2022 - Proactive governance and compliance for AWS workloads
- AWS re:Invent 2020: Achieve compliance as code using AWS Config
- Manage and Deploy Application Configurations with AWS AppConfig

OPS05-BP04 Utilizar sistemas de administración de compilación y despliegue

Utilice sistemas de administración del desarrollo y la implementación. Estos sistemas reducen tanto los errores causados por los procesos manuales como el nivel de esfuerzo requerido para implementar los cambios.

En AWS, puede crear canalizaciones de integración continua/despliegue continuo (CI/CD) utilizando servicios como las Herramientas para desarrolladores de AWS (por ejemplo, AWS CodeCommit, AWS CodeBuild, AWS CodePipeline, AWS CodeDeployy AWS CodeStar).

Resultado deseado: sus sistemas de administración de compilación y despliegue respaldan el sistema de integración continua y entrega continua (CI/CD) de su organización, que proporciona capacidades para automatizar implementaciones seguras con las configuraciones correctas.

Patrones comunes de uso no recomendados:

- Después de compilar su código en el sistema de desarrollo, copia el ejecutable en los sistemas de producción y no se inicia. Los archivos de registro locales indican que ha fallado debido a la falta de dependencias.
- Crea con éxito su aplicación con nuevas características en su entorno de desarrollo y proporciona el código a control de calidad. No pasa el control de calidad porque le faltan activos estáticos.
- El viernes, después de mucho esfuerzo, crea con éxito su aplicación manualmente en su entorno de desarrollo incluyendo las funcionalidades recién codificadas. El lunes, no puede repetir los pasos que le permitieron crear con éxito su aplicación.
- Realiza las pruebas que ha creado para su nueva versión. A continuación, dedica la siguiente semana a configurar un entorno de pruebas y a realizar todas las pruebas de integración existentes, seguidas de las pruebas de rendimiento. El nuevo código tiene un impacto inaceptable en el rendimiento y debe desarrollarse y probarse de nuevo.

Beneficios de establecer esta práctica recomendada: Al proporcionar mecanismos para gestionar las actividades de desarrollo e implementación, se reduce el nivel de esfuerzo para realizar tareas repetitivas, se libera a los miembros del equipo para que se centren en sus tareas creativas de alto valor y se limita la introducción de errores de procedimientos manuales.

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

Guía para la implementación

Los sistemas de administración de compilación y despliegue se utilizan para seguir e implementar cambios, reducir los errores causados por los procesos manuales y reducir el nivel de esfuerzo requerido para un despliegue seguro. Automatice completamente el proceso de integración e implementación, desde el registro del código hasta la construcción, prueba, despliegue y validación. Esto reduce el tiempo de entrega, disminuye los costes, fomenta una mayor frecuencia de cambios, reduce el nivel de esfuerzo y aumenta la colaboración.

Pasos para la implementación

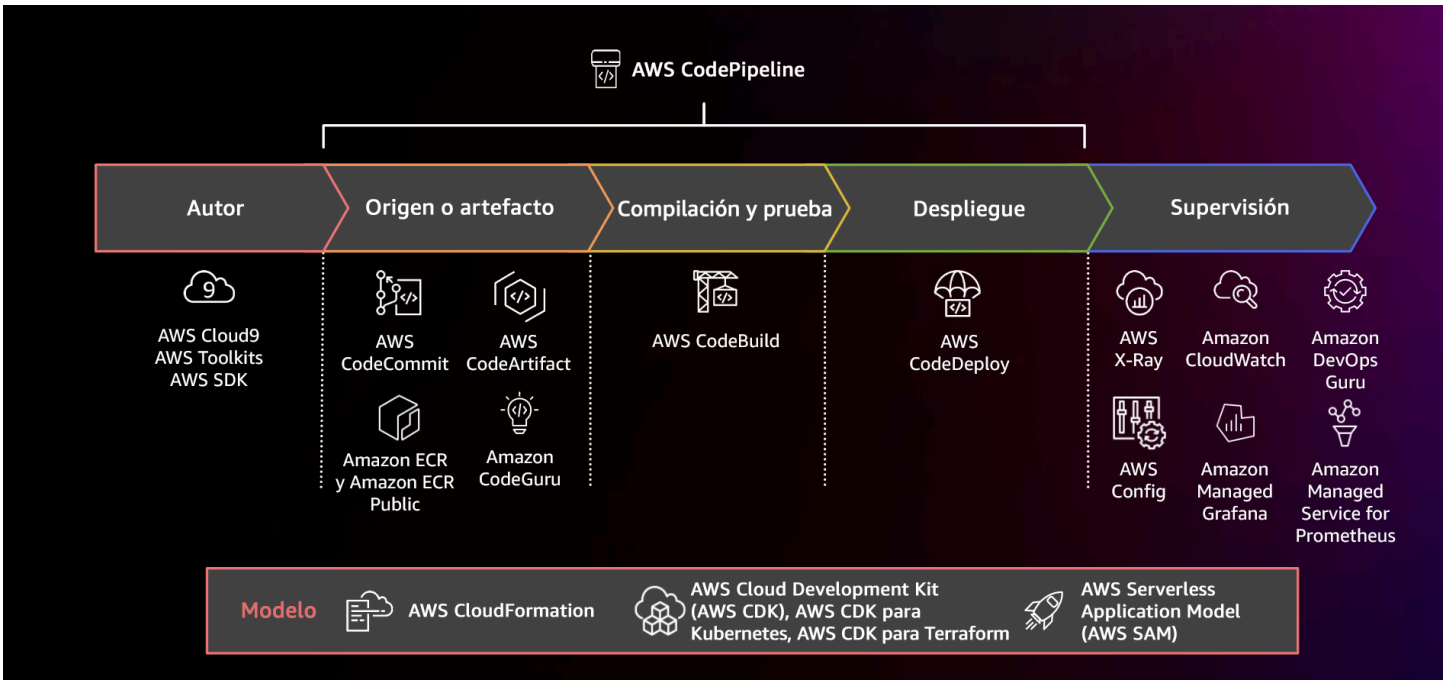


Diagrama que muestra el uso de una canalización de CI/CD con AWS CodePipeline y los servicios relacionados

1. Utilice AWS CodeCommit para controlar versiones, almacenar y administrar activos (como documentos, código fuente y archivos binarios).
2. Utilice CodeBuild para compilar el código fuente, ejecutar pruebas unitarias y producir artefactos listos para su despliegue.
3. Utilice CodeDeploy como un servicio de despliegue que automatiza los despliegues de aplicaciones para instancias de Amazon EC2 , instancias locales, funciones de AWS Lambda sin servidoro Amazon ECS.
4. Supervise sus despliegues.

Recursos

Prácticas recomendadas relacionadas:

- OPS06-BP04 Automatizar las pruebas y la reversión

Documentos relacionados:

- Herramientas para desarrolladores de AWS
- What is AWS CodeCommit?
- What is AWS CodeBuild?
- AWS CodeBuild
- What is AWS CodeDeploy?

Vídeos relacionados:

- AWS re:Invent 2022 - AWS Well-Architected best practices for DevOps on AWS

OPS05-BP05 Administrar parches

Administre parches para ampliar las características, resolver problemas y mantener la conformidad con la gobernanza. Automatice la administración de parches para reducir los errores causados por los procesos manuales, la escala y el nivel de esfuerzo requerido para aplicarlos.

La administración de parches y vulnerabilidades forma parte de sus actividades de administración de beneficios y riesgos. Es preferible tener infraestructuras inmutables y desplegar las cargas de trabajo en estados en buenas condiciones conocidos y verificados. Cuando esto no es viable, la opción que queda es el parcheado in situ.

Amazon EC2 Image Builder proporciona canalizaciones para actualizar las imágenes de las máquinas. Como parte de la administración de parches, considere las imágenes de máquina de Amazon (AMI) con una canalización de AMI o las imágenes de contenedores con una canalización de imágenes de Docker, mientras que AWS Lambda proporciona patrones para versiones ejecutables personalizadas y bibliotecas adicionales para eliminar las vulnerabilidades.

Debe administrar las actualizaciones de las imágenes de máquina de Amazon para imágenes de Linux o Windows Server con Amazon EC2 Image Builder. Puede usar el Amazon Elastic Container Registry (Amazon ECR) con su canalización actual para administrar imágenes de Amazon ECS y administrar imágenes de Amazon EKS. Lambda incluye características de administración de versiones.

La aplicación de parches no debe realizarse en los sistemas de producción sin antes realizar pruebas en un entorno seguro. Los parches solo deben aplicarse si sirven para mejorar los resultados operativos o empresariales. En AWS, puede usar AWS Systems Manager Patch Manager para automatizar el proceso de aplicación de parches en los sistemas administrados y programar la actividad con ventanas de mantenimiento de Systems Manager.

Resultado deseado: las imágenes del contenedor y AMI están parcheadas, actualizadas y listas para su lanzamiento. Puede realizar un seguimiento del estado de todas las imágenes desplegadas y determinar el cumplimiento de los parches. Puede informar sobre el estado actual y disponer de un proceso que satisfaga sus necesidades de cumplimiento.

Patrones comunes de uso no recomendados:

- Se le encomienda la aplicación de todos los nuevos parches de seguridad en un plazo de dos horas, lo que da lugar a numerosas interrupciones debido a la incompatibilidad de las aplicaciones con los parches.
- Una biblioteca sin parches tiene consecuencias no deseadas, ya que partes desconocidas utilizan las vulnerabilidades de la misma para acceder a su carga de trabajo.
- Aplica parches a los entornos de los desarrolladores sin avisarles. Recibe múltiples quejas de los desarrolladores porque su entorno ha dejado de funcionar tal como se esperaba.
- No se ha parcheado el software comercial disponible en el mercado en una instancia persistente. Cuando tiene un problema con el software y se pone en contacto con el proveedor, le notifican que la versión no es compatible y que tiene que aplicar un parche en un nivel específico para recibir asistencia.
- Ha utilizado un parche para el software de cifrado publicado recientemente que tiene importantes mejoras de rendimiento. Su sistema sin parches tiene problemas de rendimiento que continúan como resultado de no aplicar los parches.
- Se le notifica una vulnerabilidad de día cero que requiere una solución de emergencia y tiene que parchar todos sus entornos manualmente.

Beneficios de establecer esta práctica recomendada: Al establecer un proceso de administración de parches, que incluya sus criterios de aplicación de parches y la metodología de distribución en sus entornos, puede escalar e informar sobre los niveles de parches. Esto proporciona garantías en torno a la aplicación de parches de seguridad y garantiza una visibilidad clara del estado de las correcciones conocidas que se están aplicando. Esto fomenta la adopción de las características y capacidades deseadas, la rápida eliminación de problemas y el cumplimiento sostenido de la gobernanza. Implante sistemas de administración de parches y automatización para reducir el nivel de esfuerzo en el despliegue de parches y limitar los errores causados por los procesos manuales.

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

Guía para la implementación

Aplique parches a los sistemas para solucionar problemas, para obtener las características o capacidades deseadas y para mantener la conformidad con la política de gobernanza y los requisitos de soporte de los proveedores. En sistemas inmutables, implemente con el conjunto de parches adecuados para lograr el resultado deseado. Automatice el mecanismo de administración de parches para reducir el tiempo que tarda en aplicarlos, evitar los errores causados por los procesos manuales y reducir el nivel de esfuerzo requerido para aplicar los parches.

Pasos para la implementación

Para Amazon EC2 Image Builder:

1. Con Amazon EC2 Image Builder, especifique los detalles de la canalización:
    a. Cree una canalización de imágenes y asígnele un nombre.
    b. Defina el horario y la zona horaria de la canalización.
    c. Configure las dependencias.
2. Elija una receta:
    a. Seleccione una receta existente o cree una nueva.
    b. Seleccione el tipo de imagen.
    c. Asigne un nombre y versión a la receta.
    d. Seleccione la imagen base.
    e. Añada componentes de compilación y añádalos al registro de destino.
3. Opcional: defina la configuración de la infraestructura.
4. Opcional: defina los ajustes de configuración.
5. Revise la configuración.
6. Mantenga la higiene de las recetas con regularidad.

Para Systems Manager Patch Manager:

1. Cree un punto de referencia de parches.
2. Seleccione un método de operaciones de creación de rutas.
3. Habilite el análisis y la generación de informes de cumplimiento.

Recursos

Prácticas recomendadas relacionadas:

- OPS06-BP04 Automatizar las pruebas y la reversión

Documentos relacionados:

- What is Amazon EC2 Image Builder
- Create an image pipeline using the Amazon EC2 Image Builder
- Create a container image pipeline
- AWS Systems Manager Patch Manager
- Working with Patch Manager
- Working with patch compliance reports
- Herramientas para desarrolladores de AWS

Vídeos relacionados:

- CI/CD for Serverless Applications on AWS
- Diseñar con las operaciones en mente

Ejemplos relacionados:

- Well-Architected Labs - Inventory and Patch Management (Laboratorios de Well-Architected: administración de inventario y parches)
- AWS Systems Manager Patch Manager tutorials

OPS05-BP06 Compartir estándares de diseño

Comparta las prácticas recomendadas entre los equipos para aumentar la conciencia y maximizar los beneficios del trabajo de desarrollo. Documéntelas y manténgalas actualizadas a medida que evoluciona su arquitectura. Si se aplican los estándares compartidos en su organización, es fundamental que existan mecanismos para solicitar adiciones, cambios y excepciones a los estándares. Sin esta opción, los estándares se convierten en un obstáculo para la innovación.

Resultado deseado: Los estándares de diseño se comparten entre los equipos de sus organizaciones. Se documentan y actualizan a medida que evolucionan las prácticas recomendadas.

Patrones comunes de uso no recomendados:

- Dos equipos de desarrollo distintos han creado, cada uno, un servicio de autenticación de usuarios. Sus usuarios tienen que mantener un conjunto de credenciales diferente para cada parte del sistema a la que quieran acceder.
- Cada equipo administra su propia infraestructura. Un nuevo requisito de conformidad obliga a cambiar la infraestructura y cada equipo lo aplica de forma distinta.

Beneficios de establecer esta práctica recomendada: El uso de estándares compartidos favorece la adopción de las prácticas recomendadas y maximiza las ventajas de los esfuerzos de desarrollo. La documentación y actualización de los estándares de diseño mantiene a su organización al día de las prácticas recomendadas y de los requisitos de seguridad y cumplimiento.

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

Guía para la implementación

Comparta entre los equipos las prácticas recomendadas, los estándares de diseño, las listas de verificación, los procedimientos operativos, las orientaciones y los requisitos de gobernanza. Disponga de procedimientos para solicitar cambios, adiciones y excepciones a los estándares de diseño para apoyar la mejora y la innovación. Asegúrese de que los equipos estén al tanto del contenido publicado. Disponga de un mecanismo para mantener al día los estándares de diseño a medida que surgen nuevas prácticas recomendadas.

Ejemplo de cliente

AnyCompany Retail cuenta con un equipo de arquitectura interfuncional que crea patrones de arquitectura de software. Este equipo construye la arquitectura con la conformidad y la gobernanza integradas. Los equipos que adoptan estos estándares compartidos se benefician de la conformidad y la gobernanza integradas. Pueden construir rápidamente sobre el estándar de diseño. El equipo de arquitectura se reúne trimestralmente para evaluar los patrones de arquitectura y actualizarlos en caso necesario.

Pasos para la implementación

1. Identifique un equipo interfuncional que se encargue de desarrollar y actualizar los estándares de diseño. Este equipo debe trabajar con las partes interesadas de toda la organización a fin de desarrollar estándares de diseño, procedimientos operativos, listas de verificación, guías

y requisitos de gobernanza. Documente los estándares de diseño y compártalos dentro de su organización.

   a. AWS Service Catalog puede utilizarse para crear carteras que representen los estándares de diseño utilizando la infraestructura como código. Puede compartir carteras entre cuentas.

2. Disponga de un mecanismo para mantener al día los estándares de diseño a medida que se identifiquen nuevas prácticas recomendadas.
3. Si los estándares de diseño se aplican de forma centralizada, cuente con un proceso para solicitar cambios, actualizaciones y exenciones.

Nivel de esfuerzo para el plan de implementación: Medio. El desarrollo de un proceso para crear y compartir estándares de diseño precisa de coordinación y cooperación con las partes interesadas de toda la organización.

Recursos

Prácticas recomendadas relacionadas:

- OPS01-BP03 Evaluar los requisitos de gobernanza - Los requisitos de gobernanza influyen en los estándares de diseño.
- OPS01-BP04 Evaluar los requisitos de cumplimiento - La conformidad es un elemento vital de la creación de estándares de diseño.
- OPS07-BP02 Garantizar una revisión sistemática de la preparación operativa - Las listas de verificación de preparación operativa son un mecanismo para implementar los estándares de diseño a la hora de diseñar la carga de trabajo.
- OPS11-BP01 Tener un proceso de mejora continua - La actualización de los estándares de diseño forma parte de la mejora continua.
- OPS11-BP04 Realizar la administración de conocimientos - Como parte de su práctica de administración del conocimiento, documente y comparta los estándares de diseño.

Documentos relacionados:

- Automate AWS Backups with AWS Service Catalog
- AWS Service Catalog Account Factory-Enhanced
- How Expedia Group built Database as a Service (DBaaS) offering using AWS Service Catalog

- Maintain visibility over the use of cloud architecture patterns (Mantener la visibilidad sobre el uso de patrones de arquitectura de la nube)
- Simplify sharing your AWS Service Catalog portfolios in an AWS Organizations setup

Vídeos relacionados:

- AWS Service Catalog – Getting Started
- AWS re:Invent 2020: Manage your AWS Service Catalog portfolios like an expert

Ejemplos relacionados:

- AWS Service Catalog Reference Architecture
- AWS Service Catalog Workshop

Servicios relacionados:

- AWS Service Catalog

OPS05-BP07 Adoptar prácticas para mejorar la calidad del código

Adopte prácticas para mejorar la calidad del código y minimizar los defectos. Algunos ejemplos son el desarrollo basado en pruebas, las revisiones de código, la adopción de estándares y la programación en pareja. Incorpore estas prácticas a su proceso de integración y entrega continuas.

Resultado deseado: Su organización utiliza las prácticas recomendadas, como las revisiones de código o la programación en pareja, para mejorar la calidad del código. Los desarrolladores y operadores adoptan las prácticas recomendadas de calidad del código como parte del ciclo de vida de desarrollo del software.

Patrones comunes de uso no recomendados:

- Usted envía código a la rama principal de su aplicación sin una revisión del código. El cambio se despliega automáticamente en producción y provoca una interrupción del servicio.
- Se desarrolla una nueva aplicación sin pruebas de unidad, integrales o de integración. No hay forma de probar la aplicación antes del despliegue.

- Los equipos realizan cambios manuales en producción para corregir defectos. Los cambios no se someten a pruebas ni revisiones de código y no se capturan ni registran en los procesos de integración y entrega continuas.

Beneficios de establecer esta práctica recomendada: Al adoptar prácticas para mejorar la calidad del código, puede ayudar a minimizar los problemas introducidos en la producción. La calidad del código aumenta gracias a las prácticas recomendadas, como la programación en pareja y las revisiones del código.

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

Guía para la implementación

Adopte prácticas para mejorar la calidad del código y minimizar los defectos antes de la implementación. Utilice prácticas como desarrollo basado en pruebas, revisiones de código y programación en pareja para mejorar la calidad de su proceso.

Ejemplo de cliente

AnyCompany Retail adopta diversas prácticas para mejorar la calidad del código. Ha adoptado el desarrollo basado en pruebas como norma para escribir aplicaciones. Para algunas características nuevas, hace que los desarrolladores programen en pareja durante un sprint. Cada solicitud de extracción se somete a una revisión de código por parte de un desarrollador sénior antes de que se integre y despliegue.

Pasos para la implementación

1. Adopte prácticas que fomenten la calidad del código, como el desarrollo basado en pruebas, las revisiones del código y la programación en parejas, en su proceso de integración y entrega continuas. Utilice estas técnicas para mejorar la calidad del software.
   a. Amazon CodeGuru Reviewer puede proporcionar recomendaciones de programación para código Java y Python mediante el uso de machine learning.
   b. Puede crear entornos de desarrollo compartidos con AWS Cloud9 donde puede colaborar en el desarrollo del código.

Nivel de esfuerzo para el plan de implementación: Medio. Existen numerosas formas de implementar esta práctica recomendada, pero conseguir que la organización la adopte puede suponer un reto.

Recursos

Prácticas recomendadas relacionadas:

- OPS05-BP06 Compartir estándares de diseño - Puede compartir los estándares de diseño como parte de su práctica de calidad del código.

Documentos relacionados:

- Agile Software Guide (Guía del software Agile)
- My CI/CD pipeline is my release captain (Mi canalización CI/CD es mi capitán de lanzamiento)
- Automate code reviews with Amazon CodeGuru Reviewer (Revisiones automáticas de código con Amazon CodeGuru Reviewer)
- Adopt a test-driven development approach (Adoptar un enfoque de desarrollo basado en pruebas)
- How DevFactory builds better applications with Amazon CodeGuru
- On Pair Programming (Programación en pareja)
- RENGA Inc. automates code reviews with Amazon CodeGuru
- The Art of Agile Development: Test-Driven Development (El arte del desarrollo ágil: desarrollo basado en pruebas)
- Why code reviews matter (and actually save time!) (Por qué son importantes las revisiones del código [¡y ahorran tiempo!])

Vídeos relacionados:

- AWS re:Invent 2020: Continuous improvement of code quality with Amazon CodeGuru
- AWS Summit ANZ 2021 - Driving a test-first strategy with CDK and test driven development

Servicios relacionados:

- Amazon CodeGuru Reviewer
- Amazon CodeGuru Profiler
- AWS Cloud9

OPS05-BP08 Usar varios entornos

Use diversos entornos para experimentar, desarrollar y poner a prueba su carga de trabajo. Utilice niveles crecientes de controles a medida que los entornos se acerquen a la fase de producción para asegurarse de que su carga de trabajo funcione según lo previsto cuando se despliegue.

Resultado deseado: tiene varios entornos que reflejan sus necesidades de cumplimiento y gobernanza. Prueba y hace avanzar el código a través de entornos en su ruta hasta producción.

Patrones comunes de uso no recomendados:

- Está realizando el desarrollo en un entorno compartido y otro desarrollador sobrescribe sus cambios de código.
- Los controles de seguridad restrictivos de su entorno de desarrollo compartido le impiden experimentar con nuevos servicios y características.
- Realiza pruebas de carga en sus sistemas de producción y provoca una interrupción a los usuarios.
- Se ha producido un error crítico que ha provocado la pérdida de datos en producción. En el entorno de producción, se intenta recrear las condiciones que condujeron a la pérdida de datos para poder identificar cómo ocurrió y evitar que vuelva a suceder. Para evitar más pérdida de datos durante las pruebas, se ve obligado a hacer que la aplicación no esté disponible para los usuarios.
- Utiliza un servicio de inquilino múltiple y no puede atender la solicitud de un cliente de tener un entorno dedicado.
- Puede que no siempre pruebe, pero cuando lo hace, lo hace en su entorno de producción.
- Cree que la simplicidad de un entorno único anula el alcance del impacto de los cambios en el entorno.

Beneficios de establecer esta práctica recomendada: puede dar respaldo a varios entornos simultáneos de desarrollo, de pruebas y de producción sin crear conflictos entre los desarrolladores o las comunidades de usuarios.

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

Guía para la implementación

Use varios entornos y proporcione a los desarrolladores entornos aislados con controles minimizados para ayudar con la experimentación. Proporcione entornos de desarrollo individuales para ayudar al trabajo en paralelo, que aumenta la agilidad del desarrollo. Implemente controles más rigurosos en

los entornos que están cercanos a la producción para que los desarrolladores puedan innovar. Utilice infraestructura como código y sistemas de administración de la configuración para implementar entornos que estén configurados de forma coherente con los controles presentes en la producción y asegurarse de que los sistemas funcionarán como se espera cuando se implementen. Cuando los entornos no estén en uso (por ejemplo, sistemas de desarrollo durante la noche y los fines de semana), apáguelos para evitar los costos asociados a los recursos inactivos. Cuando realice pruebas de carga, despliegue entornos semejantes al de producción para mejorar los resultados válidos.

Recursos

Documentos relacionados:

- Instance Scheduler en AWS
- ¿Qué es AWS CloudFormation?

OPS05-BP09 Realizar cambios frecuentes, pequeños y reversibles

Los cambios frecuentes, pequeños y reversibles tienen menos alcance y menos repercusiones. Cuando se utilizan junto con sistemas de administración de cambios, sistemas de administración de la configuración y sistemas de compilación y entrega, los cambios frecuentes, pequeños y reversibles reducen el alcance y el impacto de un cambio. Al hacerlo, los problemas se solucionan de forma más eficaz y rápida con la opción de revertir los cambios.

Patrones comunes de uso no recomendados:

- Despliega una nueva versión de su aplicación trimestralmente con una ventana de cambios que significa que un servicio principal está desactivado.
- Realiza cambios frecuentes en el esquema de su base de datos sin realizar un seguimiento de los cambios en sus sistemas de administración.
- Realiza actualizaciones manuales in situ, sobrescribiendo las instalaciones y configuraciones existentes y no tiene un plan de reversión claro.

Beneficios de establecer esta práctica recomendada: los esfuerzos de desarrollo son más rápidos pues despliega pequeños cambios con frecuencia. Cuando los cambios son pequeños, es mucho más fácil identificar si tienen consecuencias no deseadas y es más fácil revertirlos. Cuando los cambios son reversibles, hay menos riesgo de aplicar el cambio, ya que la recuperación se simplifica. El proceso de cambio tiene un menor riesgo y el impacto de un cambio erróneo se reduce.

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

Guía para la implementación

Utilice cambios frecuentes, pequeños y reversibles para reducir el alcance y las repercusiones del cambio. Esto facilita la resolución de problemas, ayuda a realizar correcciones rápidamente y permite revertir los cambios. También aumenta el ritmo con el que entrega valor a la empresa.

Recursos

Prácticas recomendadas relacionadas:

- OPS05-BP03 Utilizar sistemas de administración de la configuración
- OPS05-BP04 Utilizar sistemas de administración de compilación y despliegue
- OPS06-BP04 Automatizar las pruebas y la reversión

Documentos relacionados:

- Implementing Microservices on AWS (Implementación de microservicios en AWS)
- Microservices - Observability

OPS05-BP10 Automatizar completamente la integración y el despliegue

Compilación, despliegue, y comprobación automáticas de la carga de trabajo Esto reduce tanto los errores causados por los procesos manuales como el esfuerzo requerido para implementar los cambios.

Aplique metadatos utilizando etiquetas de recursos y Grupos de recursos de AWS siguiendo una estrategia coherente de etiquetado para ayudar a identificar sus recursos. Etiquete sus recursos para la organización, la contabilidad de costes, los controles de acceso y el objetivo de ejecución de actividades de operaciones automatizadas.

Resultado deseado: los desarrolladores utilizan herramientas para entregar código y progresar hasta producción. Los desarrolladores no tienen que iniciar sesión en la Consola de administración de AWS para realizar actualizaciones. Existe una pista de auditoría completa de los cambios y la configuración, que satisface las necesidades de gobernanza y cumplimiento. Los procesos son repetibles y están estandarizados en todos los equipos. Los desarrolladores pueden centrarse en el desarrollo y en la introducción de código, lo que aumenta la productividad.

Patrones comunes de uso no recomendados:

- El viernes finaliza con la creación del nuevo código para la ramificación de características. El lunes, después de ejecutar los scripts de pruebas de calidad del código y cada uno de los scripts de pruebas unitarias, comprueba el código para la siguiente versión programada.
- Se le asigna la tarea de codificar una solución para un problema crítico que afecta a un gran número de clientes en producción. Después de probar la corrección, confirma el código y envía un correo electrónico a la administración de cambios para solicitar la aprobación de su despliegue en producción.
- Como desarrollador, debe iniciar sesión en la Consola de administración de AWS para crear un nuevo entorno de desarrollo utilizando métodos y sistemas no estándar.

Beneficios de establecer esta práctica recomendada: Al implementar sistemas automatizados de administración de compilación y despliegue, se reducen los errores causados por los procesos manuales y se reduce el esfuerzo para desplegar los cambios, lo que ayuda a los miembros de su equipo a centrarse en la entrega de valor empresarial. Aumenta la velocidad de entrega a medida que progresa hasta producción.

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

Guía para la implementación

Utilice sistemas de administración de compilación y despliegue para realizar un seguimiento e implementar el cambio, a fin de reducir tanto los errores causados por los procesos manuales como el nivel de esfuerzo. Automatice completamente el proceso de integración e implementación, desde el registro del código hasta la construcción, prueba, despliegue y validación. Esto reduce el tiempo de entrega, fomenta una mayor frecuencia de cambios, reduce el nivel de esfuerzo, aumenta la velocidad de comercialización, se traduce en un aumento de la productividad y aumenta la seguridad del código a medida que progresa hasta producción.

Recursos

Prácticas recomendadas relacionadas:

- OPS05-BP03 Utilizar sistemas de administración de la configuración
- OPS05-BP04 Utilizar sistemas de administración de compilación y despliegue

Documentos relacionados:

- What is AWS CodeBuild?
- What is AWS CodeDeploy?

Vídeos relacionados:

- AWS re\:Invent 2022 - AWS Well-Architected best practices for DevOps on AWS

OPERACIÓN 6. ¿Cómo mitiga los riesgos de implementación?

Adopte enfoques que proporcionen una respuesta inmediata sobre la calidad y logren una recuperación rápida de los cambios que no muestran los resultados deseados. El uso de estas prácticas ayuda a mitigar el impacto de los problemas generados con la implementación de cambios.

Prácticas recomendadas

- OPS06-BP01 Planificar para hacer frente a los cambios infructuosos
- OPS06-BP02 Despliegues de prueba
- OPS06-BP03 Emplear estrategias de despliegue seguros
- OPS06-BP04 Automatizar las pruebas y la reversión

OPS06-BP01 Planificar para hacer frente a los cambios infructuosos

Planifique la reversión a un estado óptimo conocido o la corrección en el entorno de producción si un despliegue causa un resultado no deseado. Tener una política para establecer un plan de este tipo ayuda a todos los equipos a desarrollar estrategias para recuperarse de los cambios fallidos. Algunos ejemplos de estrategias son los pasos de despliegue y reversión, las políticas de cambio, los indicadores de características, el aislamiento del tráfico y el cambio de tráfico. Una sola versión puede incluir varios cambios de componentes relacionados. La estrategia debe proporcionar la capacidad de resistir o recuperarse de un error de cualquier cambio de componente.

Resultado deseado: ha preparado un plan de recuperación detallado para su cambio en caso de que no tenga éxito. Además, ha reducido el tamaño de su versión para minimizar el impacto potencial en otros componentes de la carga de trabajo. Como resultado, ha reducido su impacto empresarial al acortar el posible tiempo de inactividad causado por un cambio infructuoso y ha aumentado la flexibilidad y la eficiencia de los tiempos de recuperación.

Patrones comunes de uso no recomendados:

- Ha realizado una implementación y la aplicación se comporta de forma inestable, aunque parece que hay usuarios activos en el sistema. Debe decidir si deshacer el cambio, lo que afectará a los usuarios activos, o esperar a revertir el cambio sabiendo que los usuarios pueden verse afectados igualmente.
- Después de hacer un cambio de rutina, sus nuevos entornos son accesibles, pero una de sus subredes ha quedado inaccesible. Tiene que decidir si revertirlo todo o intentar reparar la subred inaccesible. Mientras toma esa decisión, no se podrá acceder a la subred.
- Sus sistemas no tienen una arquitectura que permita actualizarlos con versiones más pequeñas. Como resultado, tiene dificultades para revertir esos cambios masivos durante un despliegue infructuoso.
- No utiliza la infraestructura como código (IaC) y ha realizado actualizaciones manuales en su infraestructura que han dado lugar a una configuración no deseada. No puede realizar un seguimiento eficaz de los cambios manuales ni revertirlos.
- Como no ha medido el aumento de la frecuencia de sus despliegues, su equipo no tiene incentivos para reducir el tamaño de los cambios y mejorar los planes de reversión para cada cambio, lo que genera más riesgos y mayores tasas de errores.
- No se mide la duración total de una interrupción provocada por cambios infructuosos. Su equipo no puede establecer prioridades ni mejorar la eficacia del proceso de despliegue y del plan de recuperación.

Beneficios de establecer esta práctica recomendada: tener un plan para recuperarse de cambios fallidos minimiza el tiempo medio de recuperación (MTTR) y reduce el impacto en la organización.

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

Guía para la implementación

La adopción por parte de los equipos de lanzamiento de políticas y prácticas coherentes permite a la organización planificar lo que debe suceder si se producen cambios infructuosos. La política debe permitir aplicar correcciones temporales en circunstancias concretas. En cualquier situación, un plan de corrección temporal o reversión debe estar bien documentado y probado antes de desplegarlo en producción en vivo para minimizar el tiempo que lleva revertir un cambio.

Pasos para la implementación

1. Documente las políticas que requieren que los equipos tengan planes efectivos para revertir los cambios dentro de un período específico.

   a. Las políticas deben especificar cuándo se permite una situación de corrección temporal.
   b. Exija un plan de reversión documentado al que puedan acceder todas las partes involucradas.
   c. Especifique los requisitos para la reversión (por ejemplo, cuando se descubra que se han desplegado cambios no autorizados).
2. Analice el grado de impacto de todos los cambios relacionados con cada componente de una carga de trabajo.
   a. Permita que los cambios repetibles se estandaricen, se diseñen con plantillas y se autoricen previamente si siguen un flujo de trabajo coherente que aplique las políticas de cambio.
   b. Reduzca el impacto potencial de cualquier cambio mediante la reducción del tamaño del cambio para que la recuperación lleve menos tiempo y cause menos repercusión en la empresa.
   c. Asegúrese de que los procedimientos de reversión reviertan el código al estado correcto conocido para evitar incidentes siempre que sea posible.
3. Integre herramientas y flujos de trabajo para aplicar sus políticas mediante programación.
4. Haga que los datos sobre los cambios sean visibles para otros propietarios de cargas de trabajo para mejorar la velocidad de diagnóstico de cualquier cambio infructuoso que no se pueda revertir.
   a. Mida el éxito de esta práctica a través de datos de cambios visibles e identifique las mejoras iterativas.
5. Utilice herramientas de supervisión para verificar el éxito o el fracaso de un despliegue a fin de acelerar la toma de decisiones sobre la reversión.
6. Mida la duración de la interrupción durante un cambio infructuoso para mejorar continuamente sus planes de recuperación.

Nivel de esfuerzo para el plan de implementación: Medio

Recursos

Prácticas recomendadas relacionadas:

- OPS06-BP04 Automatizar las pruebas y la reversión

Documentos relacionados:

- AWS Builders' Library | Asegurar la seguridad en las restauraciones durante las implementaciones
- AWS Documento técnico | Change Management in the Cloud

Vídeos relacionados:

- re:Invent 2019 | El enfoque de Amazon para el despliegue de alta disponibilidad

OPS06-BP02 Despliegues de prueba

Pruebe los procedimientos de lanzamiento en preproducción utilizando la misma configuración de despliegue, controles de seguridad, pasos y procedimientos que en producción. Valide que todos los pasos desplegados se completen según lo esperado, como la inspección de archivos, configuraciones y servicios. Pruebe más a fondo todos los cambios con pruebas funcionales, de integración y de carga, junto con cualquier supervisión, como la comprobación de estado. Al realizar estas pruebas, puede identificar los problemas de despliegue con prontitud y tiene la oportunidad de planificarlos y mitigarlos antes de llegar a producción.

Puede crear entornos paralelos temporales para probar cada cambio. Automatice el despliegue de los entornos de prueba mediante la infraestructura como código (IaC) para reducir la cantidad de trabajo que implica y garantizar la estabilidad, la coherencia y una entrega de características más rápida.

Resultado deseado: su organización adopta una cultura de desarrollo basada en pruebas que incluye el despliegue de pruebas. Esto garantiza que los equipos se centren en ofrecer valor empresarial en lugar de administrar las versiones. Los equipos participan desde el principio de la identificación de los riesgos del despliegue para determinar el curso de mitigación adecuado.

Patrones comunes de uso no recomendados:

- Durante las versiones de producción, los despliegues no probados provocan problemas frecuentes que requieren la solución de problemas y la escalada.
- Su versión contiene infraestructura como código (IaC) que actualiza los recursos existentes. No está seguro de si IaC se ejecuta correctamente o si afecta a los recursos.
- Despliega una característica nueva en su aplicación. No funciona según lo previsto y no hay visibilidad hasta que los usuarios afectados lo denuncien.
- Actualiza sus certificados. Instala accidentalmente los certificados en los componentes incorrectos, lo que pasa desapercibido y afecta a los visitantes del sitio web porque no se puede establecer una conexión segura con el sitio web.

Beneficios de establecer esta práctica recomendada: las exhaustivas pruebas en la preproducción de los procedimientos de despliegue y los cambios introducidos por ellos minimizan la posible

repercusión en producción causada por las etapas de despliegue. Esto aumenta la confianza durante el lanzamiento de producción y minimiza el soporte operativo sin ralentizar la velocidad de los cambios que se introducen.

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

Guía para la implementación

Probar el proceso de despliegue es tan importante como probar los cambios que resultan del despliegue. Esto se puede lograr probando los pasos de despliegue en un entorno de preproducción que refleje la producción lo más fielmente posible. Los problemas más habituales, como pasos de despliegue incompletos o incorrectos o configuraciones incorrectas, pueden detectarse antes de pasar a producción. Además, puede poner a prueba sus pasos de recuperación.

Ejemplo de cliente

Como parte de su canalización de integración y entrega continuas (CI/CD), AnyCompany Retail lleva a cabo los pasos definidos necesarios para lanzar actualizaciones de infraestructura y software para sus clientes en un entorno similar al de producción. La canalización se compone de comprobaciones previas para detectar desviaciones (detectar cambios en los recursos realizados fuera de su IaC) en los recursos antes del despliegue, así como para validar las acciones que la IaC emprende tras su inicio. Valida los pasos de despliegue, como verificar que determinados archivos y configuraciones estén en su sitio y que los servicios estén en estado de ejecución y respondan correctamente a las comprobaciones de estado del host local antes de volver a registrarse en el equilibrador de carga. Además, todos los cambios se someten a una serie de pruebas automatizadas, como pruebas funcionales, de seguridad, de regresión, de integración y de carga.

Pasos para la implementación

1. Realice comprobaciones previas a la instalación para reflejar el entorno de preproducción en producción.
   a. Utilice la detección de desviaciones para detectar cuándo se han cambiado los recursos fuera de CloudFormation.
   b. Utilice los conjuntos de cambios para validar que la intención de una actualización de la pila coincida con las acciones que CloudFormation lleva a cabo cuando se inicia el conjunto de cambios.
2. Esto desencadena un paso de aprobación manual en AWS CodePipeline para autorizar el despliegue en el entorno de preproducción.

3. Utilice configuraciones de despliegue como archivos de AWS CodeDeploy AppSpec para definir los pasos de despliegue y validación.
4. Cuando proceda, integre AWS CodeDeploy con otros servicios de AWS o bien integre AWS CodeDeploy con productos y servicios de los socios.
5. Supervise los despliegues con Amazon CloudWatch, AWS CloudTrail y las notificaciones de eventos de Amazon SNS.
6. Realice pruebas automatizadas posteriores al despliegue, incluidas pruebas funcionales, de seguridad, de regresión, de integración y de carga.
7. Solucione los problemas de despliegue.
8. La validación correcta de los pasos precedentes debería iniciar un flujo de trabajo de aprobación manual para autorizar el despliegue en producción.

Nivel de esfuerzo para el plan de implementación: Alto

Recursos

Prácticas recomendadas relacionadas:

- OPS05-BP02 Probar y validar los cambios

Documentos relacionados:

- AWS Builders' Library | Automatización de implementaciones seguras y sin intervención | Implementaciones de prueba
- Documento técnico de AWS | Práctica de integración y entrega continuas en AWS
- La historia de Apollo: el motor de despliegue de Amazon
- How to test and debug AWS CodeDeploy locally before you ship your code
- Integración de las pruebas de conectividad de red con el despliegue de la infraestructura

Vídeos relacionados:

- re:Invent 2020 | Pruebas de software y sistemas en Amazon

Ejemplos relacionados:

- Tutorial | Deploy and Amazon ECS service with a validation test

OPS06-BP03 Emplear estrategias de despliegue seguros

Los despliegues de producción seguros controlan el flujo de cambios beneficiosos con el objetivo de minimizar cualquier impacto percibido por los clientes como consecuencia de dichos cambios. Los controles de seguridad proporcionan mecanismos de inspección para validar los resultados deseados y limitar el alcance del impacto de cualquier defecto introducido por los cambios o por errores en el despliegue. Los despliegues seguros incluyen estrategias como: indicadores de características, caja individual, continuas (versiones de valores controlados), inmutables, división del tráfico y despliegues azul-verde.

Resultado deseado: su organización utiliza un sistema de entrega continua e integración continua (CI/CD) que proporciona capacidades para automatizar despliegues seguros. Los equipos deben utilizar estrategias adecuadas para despliegues seguros.

Patrones comunes de uso no recomendados:

- Implementa un cambio sin éxito en toda la producción de una sola vez. Como resultado, todos los clientes resultan afectados simultáneamente.
- Un defecto introducido en un despliegue simultáneo en todos los sistemas requiere una versión de emergencia. Corregirlo para todos los clientes lleva varios días.
- La administración del lanzamiento de producción requiere la planificación y la participación de varios equipos. Esto limita su capacidad de actualizar con frecuencia las características para sus clientes.
- Realiza una implementación mutable al modificar los sistemas existentes. Tras descubrir que el cambio no ha tenido éxito, se ve obligado a modificar de nuevo los sistemas para restaurar la versión antigua, lo que prolonga el tiempo de recuperación.

Beneficios de establecer esta práctica recomendada: los despliegues automatizados equilibran la velocidad de las implementaciones con la entrega de cambios beneficiosos de manera coherente a los clientes. Limitar el impacto evita caros errores de despliegue y maximiza la capacidad de los equipos de responder de manera eficiente a los errores.

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

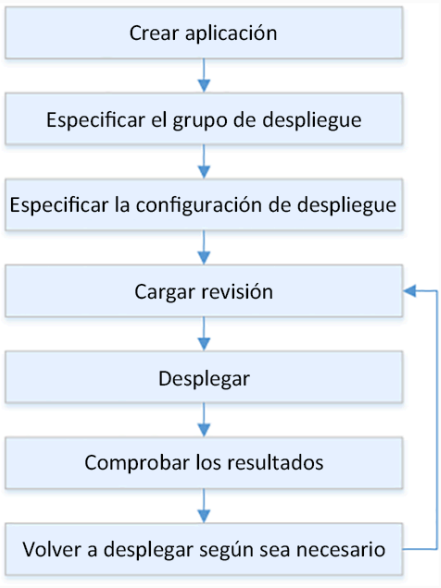
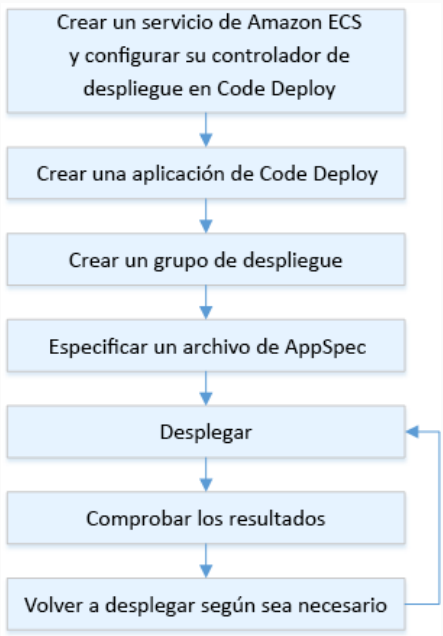
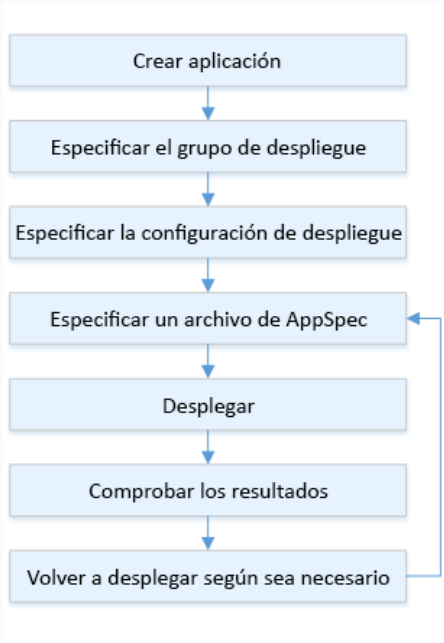
Guía para la implementación

Los errores continuos en la entrega pueden provocar una reducción de la disponibilidad del servicio y una mala experiencia para los clientes. Para maximizar la tasa de despliegues satisfactorios,

implemente controles de seguridad en el proceso de lanzamiento de principio a fin de minimizar los errores de despliegue, con el objetivo de lograr despliegues sin ningún error.

Ejemplo de cliente

AnyCompany Retail tiene la misión de lograr despliegues con un tiempo de inactividad mínimo o nulo, lo que significa que los usuarios no perciban ningún impacto durante los despliegues. Para lograrlo, la empresa ha establecido patrones de despliegue (consulte el siguiente diagrama de flujo de trabajo), como despliegues azul-verde y continuos. Todos los equipos adoptan uno o más de estos patrones en su canalización de CI/CD.

Flujo de trabajo de CodeDeploy para Amazon EC2	Flujo de trabajo de CodeDeploy para Amazon ECS	Flujo de trabajo de CodeDeploy para Lambda
		

Pasos para la implementación

1. Use un flujo de trabajo de aprobación para iniciar la secuencia de pasos de despliegue de producción, de la promoción a la producción.
2. Utilice un sistema de despliegue automatizado como AWS CodeDeploy. Entre las opciones de despliegue de AWS CodeDeploy se incluyen despliegues locales para EC2/local y despliegues azul-verde para EC2/local, AWS Lambda y Amazon ECS (consulte el diagrama de flujo de trabajo anterior).

   a. Cuando proceda, integre AWS CodeDeploy con otros servicios de AWS o bien integre AWS CodeDeploy con productos y servicios de los socios.
3. Use despliegues azul-verde para bases de datos como Amazon Aurora y Amazon RDS.
4. Supervise los despliegues con Amazon CloudWatch, AWS CloudTrail y las notificaciones de eventos de Amazon SNS.
5. Realice pruebas automatizadas posteriores al despliegue, incluidas pruebas funcionales, de seguridad, de regresión, de integración y cualquier prueba de carga.
6. Solucione los problemas de despliegue.

Nivel de esfuerzo para el plan de implementación: Medio

Recursos

Prácticas recomendadas relacionadas:

- OPS05-BP02 Probar y validar los cambios
- OPS05-BP09 Realizar cambios frecuentes, pequeños y reversibles
- OPS05-BP10 Automatizar completamente la integración y el despliegue

Documentos relacionados:

- AWS Builders' Library | Automatización de implementaciones seguras y sin intervención | Despliegues de producción
- AWS Builders' Library | My CI/CD pipeline is my release captain | Safe, automatic production releases
- Documento técnico de AWS | Práctica de integración y entrega continuas en AWS | Métodos de despliegue
- AWS CodeDeploy User Guide
- Working with deployment configurations in AWS CodeDeploy
- Configuración de una implementación de un lanzamiento canary de API Gateway
- Tipos de implementación de Amazon ECS
- Fully Managed Blue/Green Deployments in Amazon Aurora and Amazon RDS
- Blue/Green deployments with AWS Elastic Beanstalk

Vídeos relacionados:

- re:Invent 2020 | Sin intervención: automatización de canalizaciones de entrega continua en Amazon
- re:Invent 2019 | El enfoque de Amazon para el despliegue de alta disponibilidad

Ejemplos relacionados:

- Try a Sample Blue/Green Deployment in AWS CodeDeploy
- Worlshop | Buiding CI/CD pipelines for Lambda canary deployments using AWS CDK
- Taller | Despliegues de valores controlados y azul-verde para EKS y ECS
- Taller | Creación de una canalización de CI/CD multicuenta

OPS06-BP04 Automatizar las pruebas y la reversión

Para aumentar la velocidad, la fiabilidad y la confianza de su proceso de despliegue, tenga una estrategia para automatizar las capacidades de prueba y reversión en los entornos de preproducción y producción. Automatice las pruebas al desplegar en producción para simular las interacciones entre humanos y sistemas que verifican los cambios que se despliegan. Automatice la reversión para volver rápidamente a un estado válido anterior conocido. La reversión debe iniciarse automáticamente en condiciones predefinidas, como cuando no se logra el resultado deseado del cambio o cuando la prueba automatizada fracasa. La automatización de estas dos actividades mejora la tasa de éxito de los despliegues, minimiza el tiempo de recuperación y reduce el impacto potencial en la empresa.

Resultado deseado: sus pruebas automatizadas y sus estrategias de reversión se integran en el proceso de integración y entrega continuas (CI/CD). Su supervisión puede validarse según sus criterios de éxito e iniciar una reversión automática en caso de error. Esto minimiza cualquier impacto en los usuarios finales y los clientes. Por ejemplo, cuando se satisfacen todos los resultados de las pruebas, promociona el código al entorno de producción donde se inician las pruebas de regresión automatizadas, utilizando los mismos casos de prueba. Si los resultados de la prueba de regresión no coinciden con las expectativas, se inicia una reversión automática en el flujo de trabajo de la canalización.

Patrones comunes de uso no recomendados:

- Sus sistemas no tienen una arquitectura que permita actualizarlos con versiones más pequeñas. Como resultado, tiene dificultades para revertir esos cambios masivos durante un despliegue infructuoso.
- El proceso de despliegue consta de una serie de pasos manuales. Tras desplegar los cambios en la carga de trabajo, se inician las pruebas posteriores al despliegue. Tras las pruebas, se da cuenta de que no puede utilizar la carga de trabajo y los clientes están desconectados. A continuación, empieza a revertir a la versión anterior. Todos estos pasos manuales retrasan la recuperación general del sistema y provocan un impacto prolongado en sus clientes.
- Ha dedicado tiempo a desarrollar casos de prueba automatizados para funciones que no se utilizan con frecuencia en su aplicación, lo que minimiza el retorno de la inversión en su capacidad de realización de pruebas automatizadas.
- Su versión se compone de actualizaciones de aplicaciones, infraestructura, parches y configuración que son independientes entre sí. Sin embargo, tiene una única canalización de CI/CD que introduce todos los cambios a la vez. Un error en un componente le obliga a revertir todos los cambios, lo que hace que la reversión sea compleja e ineficiente.
- Su equipo completa el trabajo de codificación en el primer sprint y comienza el trabajo en el segundo, pero el plan no incluía las pruebas hasta el tercer sprint. Como resultado, las pruebas automatizadas revelaron defectos en el primer sprint que tenían que haberse resuelto antes de empezar a probar los resultados del segundo sprint, con lo que se retrasa todo el lanzamiento y se devalúan las pruebas automatizadas.
- Los casos de las pruebas de regresión automatizadas para el lanzamiento de producción se han completado, pero no está supervisando el estado de la carga de trabajo. Como no puede saber si el servicio se ha reiniciado o no, no está seguro de si es necesaria una reversión o si ya se ha producido.

Beneficios de establecer esta práctica recomendada: las pruebas automatizadas aumentan la transparencia del proceso de pruebas y su capacidad para abarcar más características en un intervalo más reducido. Al probar y validar los cambios en la producción, puede identificar los problemas de forma inmediata. La mejora de la coherencia con herramientas de prueba automatizadas permite una mejor detección de los defectos. Al revertir automáticamente a la versión anterior, se minimiza el impacto en los clientes. La reversión automatizada, en última instancia, inspira más confianza en sus capacidades de despliegue al reducir el impacto empresarial. En general, estas capacidades reducen el tiempo de entrega y, al mismo tiempo, garantizan la calidad.

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

Guía para la implementación

Automatice las pruebas de los entornos desplegados para confirmar los resultados deseados con más rapidez. Automatice la reversión a un estado conocido correcto anterior cuando no se logren resultados predefinidos para minimizar el tiempo de recuperación y reducir los errores causados por los procesos manuales. Integre las herramientas de prueba con el flujo de trabajo de la canalización para probar y minimizar las entradas manuales de manera coherente. Dé prioridad a la automatización de los casos de prueba, como aquellos que mitigan los mayores riesgos y que deben probarse con frecuencia con cada cambio. Además, automatice la reversión en función de las condiciones específicas predefinidas en su plan de pruebas.

Pasos para la implementación

1. Establezca un ciclo de vida de pruebas para su ciclo de vida de desarrollo que defina cada etapa del proceso de prueba, desde la planificación de los requisitos hasta el desarrollo de los casos de prueba, la configuración de las herramientas, las pruebas automatizadas y el cierre de los casos de prueba.
   a. Cree un enfoque de pruebas específico para la carga de trabajo a partir de su estrategia general de pruebas.
   b. Considere una estrategia de pruebas continuas cuando sea apropiado durante todo el ciclo de vida de desarrollo.
2. Seleccione herramientas automatizadas para realizar pruebas y reversiones en función de sus requisitos empresariales y de las inversiones en curso.
3. Decida qué casos de prueba quiere automatizar y cuáles se deberán realizar manualmente. Estos se pueden definir en función de la prioridad de valor empresarial de la característica que se está probando. Alinee a todos los miembros del equipo con este plan y verifique la responsabilidad de realizar las pruebas manuales.
   a. Aplique capacidades de pruebas automatizadas a casos de prueba específicos que tengan sentido para la automatización, como los casos repetibles o que se ejecutan con frecuencia, los que requieren tareas repetitivas o los que se requieren en varias configuraciones.
   b. Defina los scripts de automatización de pruebas, así como los criterios de éxito en la herramienta de automatización, de modo que se pueda iniciar la automatización continua del flujo de trabajo cuando fracasan casos específicos.
   c. Defina criterios de error concretos para la reversión automática.

4. Dé prioridad a la automatización de las pruebas para obtener resultados coherentes con un desarrollo exhaustivo de casos de prueba en los que la complejidad y la interacción humana tengan un mayor riesgo de fracaso.
5. Integre las herramientas automatizadas de pruebas y reversión en la canalización de CI/CD.
   a. Desarrolle criterios de éxito claros para los cambios.
   b. Supervise y observe para detectar estos criterios y revertir automáticamente los cambios cuando se cumplan criterios de reversión específicos.
6. Realice diferentes tipos de pruebas automatizadas de producción, como:
   a. Pruebas A/B para mostrar los resultados en comparación con la versión actual entre dos grupos de pruebas de usuarios.
   b. Pruebas de valor controlado que permiten desplegar el cambio en un subconjunto de usuarios antes de lanzarlo para todos.
   c. Pruebas de marca de características que permiten activar y desactivar las características de la nueva versión de una en una desde fuera de la aplicación para que cada característica nueva se pueda validar por sí sola.
   d. Pruebas de regresión para verificar la nueva funcionalidad con los componentes interrelacionados ya existentes.
7. Supervise los aspectos operativos de la aplicación, las transacciones y las interacciones con otras aplicaciones y componentes. Redacte informes que muestren el éxito de los cambios por carga de trabajo, de modo que pueda identificar qué partes de la automatización y el flujo de trabajo se pueden optimizar aún más.
   a. Elabore informes de resultados de pruebas que le ayuden a tomar decisiones rápidas sobre si se deben invocar o no los procedimientos de reversión.
   b. Implemente una estrategia que permita la reversión automática en función de condiciones de error predefinidas que resulten de uno o más de sus métodos de prueba.
8. Desarrolle los casos de prueba automatizados para poder volver a usarlos en futuros cambios repetibles.

Nivel de esfuerzo para el plan de implementación: Medio

Recursos

Prácticas recomendadas relacionadas:

- OPS06-BP01 Planificar para hacer frente a los cambios infructuosos

- OPS06-BP02 Despliegues de prueba

Documentos relacionados:

- AWS Builders' Library | Asegurar la seguridad en las restauraciones durante las implementaciones
- Redeploy and rollback a deployment with AWS CodeDeploy
- 8 best practices when automating your deployments with AWS CloudFormation

Ejemplos relacionados:

- Serverless UI testing using Selenium, AWS Lambda, AWS Fargate, and AWS Developer Tools

Vídeos relacionados:

- re:Invent 2020 | Sin intervención: automatización de canalizaciones de entrega continua en Amazon
- re:Invent 2019 | El enfoque de Amazon para el despliegue de alta disponibilidad

Operación 7. ¿Cómo sabe que está listo para soportar una carga de trabajo?

Evalúe la disponibilidad operativa de la carga de trabajo, los procesos y procedimientos, y el personal para comprender los riesgos operativos relacionados con la carga de trabajo.

Prácticas recomendadas

- OPS07-BP01 Garantizar la capacidad del personal
- OPS07-BP02 Garantizar una revisión sistemática de la preparación operativa
- OPS07-BP03 Uso de runbooks para realizar los procedimientos
- OPS07-BP04 Usar guías de estrategias para investigar problemas
- OPS07-BP05 Tomar decisiones fundamentadas para desplegar sistemas y cambios
- OPS07-BP06 Habilitar planes de asistencia para cargas de trabajo de producción

OPS07-BP01 Garantizar la capacidad del personal

Posea un mecanismo para comprobar que cuenta con la cantidad adecuada de personal formado para atender la carga de trabajo. Deben recibir formación sobre la plataforma y los servicios que

componen su carga de trabajo. Ofrézcales los conocimientos necesarios para operar la carga de trabajo. Debe disponer de suficiente personal formado para atender el funcionamiento normal de la carga de trabajo y solucionar las incidencias que se produzcan. Cuente con suficiente personal para que pueda rotar durante las guardias y vacaciones, a fin de evitar el síndrome de burnout.

Resultado deseado:

- Hay suficiente personal formado para atender la carga de trabajo cuando esta se encuentre disponible.
- El personal recibe formación sobre el software y los servicios que componen la carga de trabajo.

Antipatrones usuales:

- Se despliega una carga de trabajo sin miembros del equipo formados para operar la plataforma y los servicios en uso.
- Se carece de suficiente personal para facilitar las rotaciones de guardia o para que el personal se tome tiempo libre.

Beneficios de establecer esta práctica recomendada:

- Contar con miembros del equipo cualificados permite un apoyo eficaz para su carga de trabajo.
- Si hay suficientes miembros en el equipo, es posible atender la carga de trabajo y las rotaciones de guardia, al tiempo que disminuye el riesgo de síndrome de burnout.

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

Guía para la implementación

Compruebe que haya suficiente personal formado para atender la carga de trabajo. Compruebe que cuenta con suficientes miembros del equipo para cubrir las actividades operativas, incluidas las rotaciones de guardia.

Ejemplo de cliente

AnyCompany Retail se asegura de que los equipos que atienden la carga de trabajo cuentan con la formación y el personal adecuados. Tienen suficientes ingenieros para tolerar una rotación de guardia. El personal recibe formación sobre el software y la plataforma en los que se basa la carga

de trabajo y se le anima a obtener certificaciones. Hay personal suficiente para que los empleados puedan tomarse tiempo libre sin dejar de atender la carga de trabajo y la rotación de guardia.

Pasos para la implementación

1. Asigne un número adecuado de personal para operar y atender la carga de trabajo, incluidas las tareas de guardia.
2. Forme a su personal sobre el software y las plataformas que componen su carga de trabajo.
   a. AWS Training and Certification tiene una biblioteca de cursos acerca de AWS. Ofrece cursos gratuitos y de pago, en línea y presenciales.
   b. AWSorganiza eventos y seminarios web en los que aprenderá de la mano de expertos de AWS.
3. Evalúe periódicamente el tamaño y las competencias del equipo a medida que cambien las condiciones operativas y la carga de trabajo. Ajuste el tamaño y las competencias del equipo para que se ciñan a los requisitos operativos.

Nivel de esfuerzo para el plan de implementación: alto. La contratación y la formación de un equipo que atienda la carga de trabajo puede suponer un esfuerzo considerable, pero promete importantes ventajas a largo plazo.

Recursos

Prácticas recomendadas relacionadas:

- OPS11-BP04 Realizar la administración de conocimientos - Los miembros del equipo deben disponer de la información necesaria para operar y atender la carga de trabajo. La administración del conocimiento es la clave para alcanzar este objetivo.

Documentos relacionados:

- AWS Events and Webinars (Eventos y seminarios web de AWS)
- AWS Training and Certification (Formación y certificación de AWS)

OPS07-BP02 Garantizar una revisión sistemática de la preparación operativa

Utilice las revisiones de la preparación operativa (ORR) para validar que puede utilizar su carga de trabajo. ORR es un mecanismo desarrollado en Amazon para validar que los equipos puedan utilizar con seguridad sus cargas de trabajo. Una ORR es un proceso de revisión e inspección que

utiliza una lista de verificación de requisitos. Una ORR es una experiencia de autoservicio que los equipos utilizan para certificar sus cargas de trabajo. Las ORR incluyen las prácticas recomendadas procedentes de las lecciones aprendidas en nuestros años de creación de software.

Una lista de verificación de ORR se compone de recomendaciones de arquitectura, proceso operativo, administración de eventos y calidad de lanzamiento. Nuestro proceso de corrección de errores (CoE) es uno de los principales impulsores de estos elementos. Su análisis posterior al incidente debe impulsar la evolución de su propia ORR. Una ORR no solo consiste en seguir las prácticas recomendadas, sino en evitar que se repitan sucesos ya vistos. Por último, los requisitos de seguridad, gobernanza y conformidad también pueden incluirse en una ORR.

Ejecute las ORR antes de que una carga de trabajo se lance a la disponibilidad general y, después, a lo largo del ciclo de vida de desarrollo del software. Ejecutar la ORR antes del lanzamiento aumenta su capacidad para utilizar la carga de trabajo de forma segura. Vuelva a ejecutar periódicamente su ORR en la carga de trabajo para detectar cualquier desviación de las prácticas recomendadas. Puede tener listas de verificación de ORR para el lanzamiento de nuevos servicios y ORR para las revisiones periódicas. Esto le ayuda a mantenerse al día en cuanto a las nuevas prácticas recomendadas que surgen y a incorporar las lecciones aprendidas del análisis posterior al incidente. A medida que madure su uso de la nube, podrá incorporar los requisitos de ORR en su arquitectura de forma predeterminada.

Resultado deseado: tiene una lista de verificación de ORR con las prácticas recomendadas para su organización. Las ORR se realizan antes de lanzar las cargas de trabajo. Las ORR se realizan periódicamente a lo largo del ciclo de vida de la carga de trabajo.

Patrones comunes de uso no recomendados:

- Lanza una carga de trabajo sin saber si puede utilizarla.
- Los requisitos de gobernanza y seguridad no se incluyen en la certificación de una carga de trabajo para su lanzamiento.
- Las cargas de trabajo no se revalúan periódicamente.
- Las cargas de trabajo se lanzan sin los procedimientos necesarios.
- Observa la repetición de los mismos errores de causa raíz en varias cargas de trabajo.

Beneficios de establecer esta práctica recomendada:

- Sus cargas de trabajo incluyen las prácticas recomendadas de arquitectura, procesos y administración.

- Las lecciones aprendidas se incorporan al proceso de ORR.
- Se aplican los procedimientos necesarios cuando se lanzan las cargas de trabajo.
- Las ORR se ejecutan a lo largo del ciclo de vida del software de sus cargas de trabajo.

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

Guía para la implementación

Una ORR es dos cosas: un proceso y una lista de verificación. Su organización debe adoptar el proceso de ORR y contar con la asistencia de un patrocinador ejecutivo. Como mínimo, las ORR deben realizarse antes de que una carga de trabajo se lance a la disponibilidad general. Ejecute la ORR durante todo el ciclo de vida del desarrollo del software para mantenerla actualizada con las prácticas recomendadas o los nuevos requisitos. La lista de verificación de ORR debe incluir elementos de configuración, requisitos de seguridad y gobernanza, y las prácticas recomendadas de su organización. Con el tiempo, puede utilizar servicios, como AWS Config, AWS Security Hub CSPMy Barreras de protección de AWS Control Towerpara incorporar las prácticas recomendadas de la ORR en barreras de protección para la detección automática de las prácticas recomendadas.

Ejemplo de cliente

Tras varios incidentes de producción, AnyCompany Retail decidió implementar un proceso de ORR. Elaboró una lista de verificación compuesta de prácticas recomendadas, requisitos de gobernanza y conformidad, y lecciones aprendidas de las interrupciones. Las nuevas cargas de trabajo llevan a cabo las ORR antes de su lanzamiento. Cada carga de trabajo realiza una ORR anual con un subconjunto de prácticas recomendadas para incorporar nuevas prácticas y requisitos que se agregan a la lista de verificación de ORR. Con el tiempo, AnyCompany Retail utilizó AWS Config para detectar algunas prácticas recomendadas, lo que agilizó el proceso de ORR.

Pasos para la aplicación

Para saber más sobre las ORR, lea el documento técnico sobre las revisiones de la preparación operativa (ORR). En él se ofrece información detallada sobre la historia del proceso ORR, cómo crear su propia práctica ORR y cómo desarrollar su lista de verificación de ORR. Los siguientes pasos son una versión abreviada de ese documento. Para conocer en profundidad qué son las ORR y cómo crear las suyas, le recomendamos que lea ese documento técnico.

1. Reúna a las principales partes interesadas, incluidos los representantes de seguridad, operaciones y desarrollo.

2. Pida a cada parte interesada que aporte al menos un requisito. Para la primera iteración, intente limitar el número de elementos a treinta o menos.
   - El Apéndice B: Ejemplo de preguntas de ORR del documento técnico sobre las revisiones de la preparación operativa (ORR) contiene las preguntas de ejemplo que puede usar para empezar.
3. Recopile sus requisitos en una hoja de cálculo.
   - Puede usar enfoques personalizados en AWS Well-Architected Tool para desarrollar su ORR y compartirlos entre sus cuentas y su organización de AWS.
4. Identifique una carga de trabajo para realizar la ORR en ella. Lo ideal es una carga de trabajo previa al lanzamiento o una carga de trabajo interna.
5. Repase la lista de verificación de ORR y tome nota de los descubrimientos realizados. Los descubrimientos pueden no ser correctos si existe una mitigación. Agregue cualquier descubrimiento que carezca de una mitigación a su lista de tareas pendientes e impleméntelas antes de lanzarlas.
6. Siga agregando las prácticas recomendadas y los requisitos a su lista de verificación ORR con el tiempo.

Los clientes de Soporte con asistencia empresarial pueden solicitar el taller de revisión de la preparación operativa a su gerente técnico de cuentas. El taller es una sesión de trabajo en sentido inverso interactiva para desarrollar su propia lista de verificación de ORR.

Nivel de esfuerzo para el plan de implementación: Alto. La adopción de una práctica de ORR en su organización requiere el patrocinio ejecutivo y la aceptación de las partes interesadas. Cree y actualice la lista de verificación con las aportaciones de toda su organización.

Recursos

Prácticas recomendadas relacionadas:

- OPS01-BP03 Evaluar los requisitos de gobernanza : los requisitos de gobernanza encajan de forma natural en una lista de verificación de ORR.
- OPS01-BP04 Evaluar los requisitos de cumplimiento : los requisitos de conformidad se incluyen a veces en una lista de verificación de ORR. Otras veces son un proceso independiente.
- OPS03-BP07 Dotar a los equipos de los recursos adecuados : la capacidad del equipo es un buen candidato para un requisito de ORR.
- OPS06-BP01 Planificar para hacer frente a los cambios infructuosos : antes de lanzar la carga de trabajo, debe establecerse un plan de restauración o de avance.

- OPS07-BP01 Garantizar la capacidad del personal : para respaldar una carga de trabajo hay que contar con el personal necesario.
- SEC01-BP03 Identificar y validar objetivos de control : los objetivos de control de seguridad son excelentes requisitos de ORR.
- REL13-BP01 Definir objetivos de recuperación para la inactividad y la pérdida de datos : los planes de recuperación de desastres son un buen requisito de ORR.
- COST02-BP01 Desarrollar políticas basadas en los requisitos de su organización : las políticas de administración de costes son adecuadas para incluirlas en su lista de verificación de ORR.

Documentos relacionados:

- AWS Control Tower - Guardrails in AWS Control Tower (AWS Control Tower: Barreras de protección en AWS Control Tower)
- AWS Well-Architected Tool - Custom Lenses (AWS Well-Architected Tool: enfoques personalizados)
- Plantilla de revisión de la preparación operativa de Adrian Hornsby
- Documento técnico sobre las revisiones de la preparación operativa (ORR)

Vídeos relacionados:

- AWS Supports You | Building an Effective Operational Readiness Review (ORR) (AWS Supports You | Elaboración de una revisión de la preparación operativa [ORR])

Ejemplos relacionados:

- Sample Operational Readiness Review (ORR) Lens (Enfoque de muestra de revisión de la preparación operativa [ORR])

Servicios relacionados:

- AWS Config
- AWS Control Tower
- AWS Security Hub CSPM
- AWS Well-Architected Tool

OPS07-BP03 Uso de runbooks para realizar los procedimientos

Un runbook es un proceso documentado para lograr un resultado específico. Los runbooks consisten en una serie de pasos que alguien sigue para conseguir algo. Los runbooks se han utilizado en operaciones que se remontan a los primeros días de la aviación. En las operaciones en la nube, utilizamos runbooks para reducir el riesgo y lograr los resultados deseados. En su forma más simple, un runbook es una lista de verificación para completar una tarea.

Los runbooks son una parte esencial del funcionamiento de su carga de trabajo. Desde la incorporación de un nuevo miembro del equipo hasta el despliegue de una versión importante, los runbooks son los procesos codificados que proporcionan resultados coherentes independientemente de quién los utilice. Los runbooks deben publicarse en una ubicación central y actualizarse a medida que el proceso evolucione, ya que la actualización de los runbooks es un componente clave de un proceso de administración de cambios. También deben incluir directrices sobre la gestión de errores, las herramientas, los permisos, las excepciones y las escalaciones en caso de que se produzca un problema.

A medida que su organización madure, comience a automatizar los runbooks. Comience con runbooks que sean cortos y se utilicen con frecuencia. Utilice lenguajes de scripting para automatizar pasos o facilitar su realización. A medida que automatice los primeros runbooks, dedicará tiempo a automatizar runbooks más complejos. Con el tiempo, la mayoría de sus runbooks deberían estar automatizados de alguna manera.

Resultado deseado: Su equipo dispone de una colección de guías paso a paso para realizar las tareas de la carga de trabajo. Los runbooks contienen el resultado deseado, las herramientas y los permisos necesarios, y las instrucciones para la gestión de errores. Se almacenan en una ubicación central y se actualizan con frecuencia.

Patrones comunes de uso no recomendados:

- Depender de la memoria para completar cada paso de un proceso.
- Desplegar manualmente los cambios sin una lista de verificación.
- Diferentes miembros del equipo realizan el mismo proceso pero con diferentes pasos o resultados.
- Dejar que los runbooks se desincronicen con los cambios del sistema y la automatización.

Beneficios de establecer esta práctica recomendada:

- Reducción de los índices de error en las tareas manuales.

- Las operaciones se realizan de forma coherente.
- Los nuevos miembros del equipo pueden empezar a realizar tareas antes.
- Los runbooks pueden automatizarse para reducir el trabajo.

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

Guía para la implementación

Los runbooks pueden adoptar varias formas en función del nivel de madurez de su organización. Como mínimo, deben consistir en un documento de texto paso a paso. El resultado deseado debe indicarse claramente. Documente claramente los permisos o herramientas especiales necesarios. Proporcione directrices detalladas sobre la gestión de errores y las escalaciones en caso de que algo vaya mal. Indique el propietario del runbook y publíquelo en una ubicación central. Una vez que el runbook esté documentado, valídelo haciendo que otra persona de su equipo lo ejecute. A medida que los procedimientos evolucionen, actualice sus runbooks de acuerdo con su proceso de administración de cambios.

Sus runbooks deben automatizarse a medida que su organización madura. Con servicios como las automatizaciones de AWS Systems Manager, puede transformar un texto plano en automatizaciones que pueden ejecutarse contra su carga de trabajo. Estas automatizaciones pueden ejecutarse en respuesta a eventos, reduciendo la carga operativa para mantener su carga de trabajo.

Ejemplo de cliente

AnyCompany Retail debe realizar actualizaciones del esquema de la base de datos durante despliegues de software. El equipo de operaciones en la nube trabajó con el equipo de administración de bases de datos para crear un runbook para desplegar manualmente estos cambios. El runbook enumeraba cada paso del proceso en forma de lista de verificación. Incluía una sección sobre la gestión de errores en caso de que algo saliera mal. Publicaron el runbook en su wiki interna junto con sus otros runbooks. El equipo de operaciones en la nube tiene previsto automatizar el runbook en un futuro sprint.

Pasos para la aplicación

Si no tiene un repositorio de documentos, un repositorio de control de versiones es un buen lugar para empezar a crear su biblioteca de runbooks. Puede crear sus runbooks usando Markdown. Hemos proporcionado una plantilla de runbook de ejemplo que puede utilizar para empezar a crear runbooks.

```
# Runbook Title ## Runbook Info | Runbook ID | Description | Tools Used
 | Special Permissions | Runbook Author | Last Updated | Escalation POC |
 |-------|-------|-------|-------|-------|-------|-------| | RUN001 | What is this
 runbook for? What is the desired outcome? | Tools | Permissions | Your Name |
 2022-09-21 | Escalation Name | ## Steps 1. Step one 2. Step two
```

1. Si no tiene un repositorio de documentación o un wiki, cree un nuevo repositorio de control de versiones en su sistema de control de versiones.
2. Identifique un proceso que no tenga un runbook. Un proceso ideal es aquel que se lleva a cabo de forma semirregular, es corto en número de pasos y tiene errores de bajo impacto.
3. En su repositorio de documentos, cree un nuevo borrador de documento Markdown utilizando la plantilla. Introduzca `Runbook Title` y los campos necesarios en `Runbook Info`.
4. Empezando por el primer paso, rellene la parte `Steps` del runbook.
5. Asigne el runbook a un miembro del equipo. Pídales que utilicen el runbook para validar los pasos. Si falta algo o hay que aclararlo, actualice el runbook.
6. Publique el runbook en su almacén de documentación interno. Una vez publicado, comuníquelo a su equipo y a otras partes interesadas.
7. Con el tiempo, creará una biblioteca de runbooks. A medida que esa biblioteca crezca, comience a trabajar para automatizar los runbooks.

Nivel de esfuerzo para el plan de implementación: Bajo El estándar mínimo para un runbook es una guía de texto paso a paso. La automatización de runbooks puede aumentar el esfuerzo de implementación.

Recursos

Prácticas recomendadas relacionadas:

- OPS02-BP02 Los procesos y procedimientos han identificado a los propietarios: los runbooks deben tener un propietario encargado de su mantenimiento.
- OPS07-BP04 Usar guías de estrategias para investigar problemas: los runbooks y guías de categorías son semejantes pero tienen una diferencia clave y es que un runbook tiene un resultado deseado. En muchos casos los runbooks se activan una vez que una guía de categorías ha identificado una causa raíz.
- OPS10-BP01 Uso de un proceso para la administración de eventos, incidentes y problemas: los runbooks forman parte de una buena práctica de gestión de eventos, incidentes y problemas.

- OPS10-BP02 Tener un proceso por alerta: los runbooks y las guías de categorías deben usarse como respuesta a alertas. Con el tiempo, estas reacciones deberían automatizarse.
- OPS11-BP04 Realizar la administración de conocimientos: el mantenimiento de los runbooks es una parte fundamental de la administración de conocimientos.

Documentos relacionados:

- Achieving Operational Excellence using automated playbook and runbook (Lograr la excelencia operativa mediante la guía de estrategias y runbook automatizados)
- AWS Systems Manager: Working with runbooks (AWS Systems Manager: trabajar con runbooks)
- Migration playbook for AWS large migrations - Task 4: Improving your migration runbooks (Guía de categorías de migración para grandes migraciones de AWS - Tarea 4: Mejora de los runbooks de la migración)
- Utilice AWS Systems Manager Automation runbooks to resolve operational tasks (Uso de runbooks de automatización de AWS Systems Manager para resolver tareas operativas)

Vídeos relacionados:

- AWS re:Invent 2019: DIY guide to runbooks, incident reports, and incident response (SEC318-R1) (Guía paso a paso sobre runbooks, informes de incidentes y respuesta a incidentes (SEC318-R1))
- How to automate IT Operations on AWS | Amazon Web Services (Cómo automatizar las operaciones de TI en AWS | Amazon Web Services)
- Integrate Scripts into AWS Systems Manager (Integrar scripts en AWS Systems Manager)

Ejemplos relacionados:

- AWS Systems Manager: Automation walkthroughs (AWS Systems Manager: Tutoriales paso a paso de automatización)
- AWS Systems Manager: Restore a root volume from the latest snapshot runbook (AWS Systems Manager: Restaurar un volumen raíz desde el último runbook de instantáneas)
- Building an AWS incident response runbook using Jupyter notebooks and CloudTrail Lake (Crear un runbook de respuesta a incidentes de AWS con cuadernos de Jupyter y CloudTrail Lake)
- Gitlab: Runbooks

- Rubix - A Python library for building runbooks in Jupyter Notebooks (Rubix: Una biblioteca de Python para crear runbooks en cuadernos de Jupyter)
- Using Document Builder to create a custom runbook (Uso de Document Builder para crear un runbook personalizado)
- Well-Architected Labs: automatización de operaciones con guías de estrategias y runbooks

Servicios relacionados:

- AWS Systems Manager Automation (Automatización de AWS Systems Manager)

OPS07-BP04 Usar guías de estrategias para investigar problemas

Las guías de estrategias son guías paso a paso que se utilizan para investigar un incidente. Cuando se producen incidentes, se usan para investigar, determinar el impacto e identificar la causa raíz. Las guías de estrategias se utilizan en diversas situaciones, desde despliegues erróneos hasta incidentes de seguridad. En numerosos casos, identifican la causa raíz que un runbook sirve para mitigar. Las guías de estrategias son un componente esencial de los planes de respuesta a incidentes de su organización.

Una buena guía de estrategias tiene varias características clave. Orienta al usuario, paso a paso, a través del proceso de descubrimiento. Viéndolo desde fuera, ¿qué pasos debería seguir alguien para diagnosticar un incidente? Defina de forma clara en la guía de estrategias si se necesitan herramientas especiales o permisos de alto nivel en ella. El hecho de contar con un plan de comunicación para informar a las partes interesadas sobre el estado de la investigación es un componente clave. En las situaciones en las que no se pueda identificar la causa raíz, la guía de estrategias debe tener un plan de traslado a una instancia superior. Si se identifica la causa raíz, la guía de estrategias debe señalar un runbook que describa cómo resolverla. Las guías de estrategias deben almacenarse de forma centralizada y se debe realizar un mantenimiento periódico de ellas. Si se utilizan para alertas específicas, facilite a su equipo indicaciones sobre cada guía de estrategias en cada alerta.

A medida que madure su organización, automatice las guías de estrategias. Empiece con guías de estrategias que cubran incidentes de poco riesgo. Utilice scripting para automatizar los pasos de descubrimiento. Asegúrese de que dispone de runbooks complementarios para mitigar las causas raíz más habituales.

Resultado deseado: su organización dispone de guías de estrategias para incidentes comunes. Dichas guías de estrategias se almacenan en una ubicación central y están a disposición de los

miembros del equipo y se actualizan con frecuencia. Se crean runbooks complementarios para cualquier causa raíz conocida.

Patrones comunes de uso no recomendados:

- No existe una forma estándar de investigar un incidente.
- Los miembros del equipo confían en la memoria muscular o en el conocimiento institucional para solucionar un despliegue con errores.
- Los nuevos miembros del equipo aprenden a investigar los problemas con el método de ensayo y error.
- Las prácticas recomendadas para investigar los problemas no se comparten entre los equipos.

Beneficios de establecer esta práctica recomendada:

- Las guías de estrategias impulsan sus esfuerzos para mitigar los incidentes.
- Los distintos miembros del equipo pueden utilizar la misma guía de estrategias para identificar la causa raíz de forma coherente.
- Las causas raíz conocidas pueden tener runbooks desarrollados para ellas, lo que acelera el tiempo de recuperación.
- Las guías de estrategias permiten a los miembros del equipo empezar a contribuir antes.
- Los equipos pueden escalar sus procesos con guías de estrategias repetibles.

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

Guía para la implementación

La forma de crear y utilizar las guías de estrategias depende de la madurez de su organización. Si es la primera vez que utiliza la nube, cree guías de estrategias en formato de texto en un repositorio de documentos central. A medida que madure su organización, las guías de estrategias pueden semiautomatizarse con lenguajes de scripting como Python. Estos scripts pueden ejecutarse en un cuaderno de Jupyter para acelerar el descubrimiento. Las organizaciones avanzadas cuentan con guías de estrategias completamente automatizadas para los problemas más habituales que se solucionan de forma automática con runbooks.

Elabore una lista de incidentes comunes que afectan a la carga de trabajo para empezar a crear las guías de estrategias. Como punto de partida, elija guías para incidentes con poco riesgo y en los

que la causa raíz se haya reducido a unos pocos problemas. Una vez que disponga de guías de estrategias para las situaciones más sencillas, continúe con las de mayor riesgo o cuya causa raíz no se conozca bien.

Sus guías de estrategias en texto deben automatizarse a medida que su organización madura. Con servicios como las automatizaciones de AWS Systems Manager, el texto sin formato puede transformarse en automatizaciones. Estas automatizaciones pueden ejecutarse en la carga de trabajo para acelerar las investigaciones. Se pueden activar en respuesta a los incidentes, lo que reduce el tiempo medio para descubrir y resolver los incidentes.

Los clientes pueden usar Administrador de incidentes de AWS Systems Manager para responder a los incidentes. Este servicio proporciona una interfaz única para clasificar los incidentes, informar a las partes interesadas durante el descubrimiento y la mitigación y colaborar durante todo el incidente. Utiliza las automatizaciones de AWS Systems Manager para acelerar la detección y la recuperación.

Ejemplo de cliente

La empresa AnyCompany Retail se ha visto afectada por un incidente de producción. El ingeniero de guardia utilizó una guía de estrategias para investigar el problema. A medida que iba realizando los pasos, informaba a las partes interesadas clave identificadas en la guía de estrategias. El ingeniero identificó la causa raíz como una condición de secuencia (race condition) en un servicio backend. Mediante un runbook, el ingeniero relanzó el servicio, con lo que AnyCompany Retail volvió a estar en línea.

Pasos para la aplicación

Si no tiene un repositorio de documentos, le sugerimos que cree uno de control de versiones para su biblioteca de guías de estrategias. Puede crear las guías de estrategias con Markdown, que es compatible con la mayoría de los sistemas de automatización de este tipo de guías. Si está empezando desde cero, utilice la siguiente plantilla de guía de estrategias de ejemplo.

```
# Título de la guía de estrategias ## Información de la guía de estrategias |
 ID de la guía de estrategias | Descripción | Herramientas usadas | Permisos
 especiales | Autor de la guía de estrategias | Última actualización | Punto
 de contacto de derivación | Partes interesadas | Plan de comunicación |
 |-------|-------|-------|-------|-------|-------|-------|-------|-------| | RUN001
 | ¿Cuál es la finalidad de esta guía de estrategias? ¿Para qué incidente se usa? |
 Herramientas | Permisos | Su nombre | 21-09-2022 | Nombre de derivación | Nombre de
 parte interesada | ¿Cómo se comunicarán las actualizaciones durante la investigación?
 | ## Pasos 1. Paso uno 2. Paso dos
```

1. Si no tiene un repositorio de documentos o un wiki, cree un nuevo repositorio de control de versiones para las guías de instrucciones en su sistema de control de versiones.
2. Identifique un problema común que requiera una investigación. Este debería ser un escenario en el que la causa raíz se limita a unos pocos problemas y la resolución conlleva poco riesgo.
3. Con la plantilla Markdown, rellene la sección `Título de la guía de estrategias` y los campos situados debajo de `Información de la guía de estrategias`.
4. Rellene los pasos de solución adicionales. Indique con la mayor claridad posible las acciones que se deben realizar o las áreas que debe investigar.
5. Entregue a un miembro del equipo la guía de estrategias y pídale que la revise para validarla. Si falta algo o no está claro, actualice la guía de estrategias.
6. Publique la guía de estrategias en el repositorio de documentos e informe al equipo y a las partes interesadas.
7. Esta biblioteca de guías de estrategias crecerá a medida que vaya agregando más guías. Una vez que tenga varias guías de estrategias, empiece a automatizarlas con herramientas como AWS Systems Manager Automations para sincronizar la automatización y las guías de estrategias.

Nivel de esfuerzo para el plan de implementación: bajo. Las guías de estrategias deben ser documentos de texto almacenados en una ubicación central. Las organizaciones más maduras se inclinarán por la automatización de las guías de estrategias.

Recursos

Prácticas recomendadas relacionadas:

- OPS02-BP02 Los procesos y procedimientos han identificado a los propietarios: las guías de estrategias deben tener un propietario encargado de su mantenimiento.
- OPS07-BP03 Uso de runbooks para realizar los procedimientos: los runbooks y las guías de estrategias son similares, pero la diferencia clave es que un runbook tiene un resultado deseado. En muchos casos, los runbooks se usan una vez que una guía de estrategias ha identificado una causa raíz.
- OPS10-BP01 Uso de un proceso para la administración de eventos, incidentes y problemas: las guías de estrategias forman parte de una buena práctica de administración de eventos, incidentes y problemas.
- OPS10-BP02 Tener un proceso por alerta: los runbooks y las guías de estrategias deben usarse como respuesta a alertas. Con el tiempo, estas reacciones deberían automatizarse.

- OPS11-BP04 Realizar la administración de conocimientos: el mantenimiento de las guías de estrategias es una parte fundamental de la administración de conocimientos.

Documentos relacionados:

- Achieving Operational Excellence using automated playbook and runbook (Lograr la excelencia operativa mediante la guía de estrategias y runbook automatizados)
- AWS Systems Manager: Working with runbooks (AWS Systems Manager: trabajar con runbooks)
- Use AWS Systems Manager Automation runbooks to resolve operational tasks (Utilizar runbooks de AWS Systems Manager Automation para resolver tareas operativas)

Vídeos relacionados:

- AWS re:Invent 2019: DIY guide to runbooks, incident reports, and incident response (SEC318-R1) (Guía paso a paso sobre runbooks, informes de incidentes y respuesta a incidentes [SEC318-R1])
- Administrador de incidentes de AWS Systems Manager - AWS Virtual Workshops (Administrador de incidentes de AWS Systems Manager: talleres virtuales de AWS)
- Integrate Scripts into AWS Systems Manager (Integrar scripts en AWS Systems Manager)

Ejemplos relacionados:

- AWS Customer Playbook Framework (Marco de trabajo de guía de estrategias de cliente de AWS)
- AWS Systems Manager: Automation walkthroughs (AWS Systems Manager: tutoriales paso a paso de automatización)
- Building an AWS incident response runbook using Jupyter notebooks and CloudTrail Lake (Crear un runbook de respuesta a incidentes de AWS con cuadernos de Jupyter y CloudTrail Lake)
- Rubix - A Python library for building runbooks in Jupyter Notebooks (Rubix: Una biblioteca de Python para crear runbooks en cuadernos de Jupyter)
- Using Document Builder to create a custom runbook (Uso de Document Builder para crear un runbook personalizado)
- Well-Architected Labs: automatización de operaciones con guías de estrategias y runbooks
- Well-Architected Labs: guía de estrategias de respuesta ante incidentes con Jupyter

Servicios relacionados:

- AWS Systems Manager Automation
- Administrador de incidentes de AWS Systems Manager

OPS07-BP05 Tomar decisiones fundamentadas para desplegar sistemas y cambios

Disponga de procesos en caso de cambios fructíferos e infructuosos de su carga de trabajo. Un pre mortem es un ejercicio en el que un equipo simula un error para desarrollar estrategias de mitigación. Realice ensayos de errores pre mortem para anticipar el fracaso y crear procedimientos cuando sea apropiado. Evalúe las ventajas y los riesgos de desplegar cambios en la carga de trabajo. Verifique que todos los cambios cumplan con la gobernanza.

Resultado deseado:

- Tomará decisiones informadas cuando despliegue cambios en la carga de trabajo.
- Los cambios cumplirán con la gobernanza.

Antipatrones usuales:

- Despliegue de un cambio en la carga de trabajo sin un proceso para gestionar un despliegue infructuoso.
- Cambios en el entorno de producción que incumplen los requisitos de gobernanza.
- Despliegue de una nueva versión de la carga de trabajo sin establecer una línea de referencia para la utilización de recursos.

Beneficios de establecer esta práctica recomendada:

- Estará preparado para cambios infructuosos en su carga de trabajo.
- Los cambios en la carga de trabajo cumplirán las políticas de gobernanza.

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

Guía para la implementación

Utilice ejercicios pre mortem para desarrollar procesos en caso de cambios infructuosos. Documente los procesos para los cambios infructuosos. Asegúrese de que todos los cambios se ajusten a la gobernanza. Evalúe las ventajas y los riesgos de desplegar cambios en la carga de trabajo.

Ejemplo de cliente

AnyCompany Retail realiza con regularidad ejercicios pre mortem para validar los procesos en caso de cambios infructuosos. Documenta los procesos en una wiki compartida y la actualiza con frecuencia. Todos los cambios se ajustan a los requisitos de gobernanza.

Pasos para la implementación

1. Tome decisiones informadas cuando despliegue cambios en la carga de trabajo. Establezca y revise los criterios para un despliegue fructífero. Desarrolle escenarios o criterios que desencadenen la reversión de un cambio. Sopese las ventajas del despliegue de cambios frente a los riesgos de un cambio infructuoso.
2. Verifique que todos los cambios cumplan las políticas de gobernanza.
3. Utilice ejercicios pre mortem para desarrollar planes en caso de cambios infructuosos y documentar las estrategias de mitigación. Lleve a cabo un ejercicio de simulación para modelar un cambio infructuoso y validar los procedimientos de reversión.

Nivel de esfuerzo para el plan de implementación: moderado. La implementación de una práctica de pre mortem exige la coordinación y el esfuerzo de las partes interesadas de toda la organización.

Recursos

Prácticas recomendadas relacionadas:

- OPS01-BP03 Evaluar los requisitos de gobernanza - Los requisitos de gobernanza son un factor clave para determinar si se debe desplegar un cambio.
- OPS06-BP01 Planificar para hacer frente a los cambios infructuosos - Establezca planes para mitigar un despliegue infructuosos y utilice actividades pre mortem para validarlos.
- OPS06-BP02 Despliegues de prueba - Cada cambio de software debe probarse adecuadamente antes de su despliegue, a fin de reducir los defectos en producción.
- OPS07-BP01 Garantizar la capacidad del personal - Disponer de suficiente personal formado para atender la carga de trabajo es esencial para tomar una decisión informada sobre el despliegue de un cambio en el sistema.

Documentos relacionados:

- Amazon Web Services: Riesgos y conformidad

- Modelo de responsabilidad compartida de AWS
- Governance in the Nube de AWS: The Right Balance Between Agility and Safety (Gobernanza en Nube de AWS: el equilibrio entre agilidad y seguridad)

OPS07-BP06 Habilitar planes de asistencia para cargas de trabajo de producción

Facilite la asistencia de cualquier software y servicio del que dependa su carga de trabajo de producción. Seleccione un nivel de asistencia adecuado para satisfacer sus necesidades de nivel de servicio de producción. Los planes de asistencia para estas dependencias son necesarios en caso de que se produzca una interrupción del servicio o un problema con el software. Documente los planes de asistencia y cómo solicitar asistencia de todos los proveedores de servicios y software. Implemente mecanismos que verifiquen que los puntos de asistencia de los contactos se mantienen actualizados.

Resultado deseado:

- Implemente planes de asistencia para el software y los servicios de los que dependen las cargas de trabajo de producción.
- Elija un plan de asistencia adecuado en función de las necesidades del nivel de servicio.
- Documente los planes de asistencia, los niveles de asistencia y la forma de solicitarla.

Patrones comunes de uso no recomendados:

- No dispone de un plan de asistencia para un proveedor de software fundamental. Su carga de trabajo se ve afectada por su proveedor y no puede hacer nada para acelerar una solución u obtener actualizaciones puntuales de él.
- Un desarrollador que era el principal punto de contacto para un proveedor de software ha abandonado la empresa. No puede ponerse en contacto directamente con el equipo de asistencia del proveedor. Debe dedicar tiempo a investigar y recorrer los sistemas de contacto genéricos, lo que aumenta el tiempo necesario para responder cuando sea necesario.
- Se produce una interrupción de la producción con un proveedor de software. No hay documentación sobre cómo presentar un caso de asistencia.

Beneficios de establecer esta práctica recomendada:

- Con el nivel de asistencia adecuado, podrá obtener una respuesta en el plazo necesario para satisfacer las necesidades de servicio.
- Como cliente con asistencia puede remitir a un nivel superior si hay problemas de producción.
- Los proveedores de software y servicios pueden ayudar en la resolución de problemas durante un incidente.

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

Guía para la implementación

Facilite planes de asistencia de cualquier proveedor de software y servicio del que dependa su carga de trabajo de producción. Configure planes de asistencia adecuados para satisfacer las necesidades de nivel de servicio. Para los clientes de AWS, esto significa habilitar AWS Business Support o superior en cualquier cuenta en la que tenga cargas de trabajo de producción. Reúnase con los proveedores de asistencia con regularidad para obtener información actualizada sobre las ofertas de asistencia, los procesos y los contactos. Documente cómo solicitar asistencia a los proveedores de software y servicios, incluida la forma de remitir a un nivel superior si se produce una interrupción. Implemente mecanismos para mantener actualizados los contactos de asistencia.

Ejemplo de cliente

En AnyCompany Retail, todas las dependencias de software y servicios comerciales disponen de planes de asistencia. Por ejemplo, tienen habilitado AWS Enterprise Support en todas las cuentas con cargas de trabajo de producción. Cualquier desarrollador puede abrir un caso de asistencia cuando surja un problema. Hay una página wiki con información sobre cómo solicitar asistencia, a quién notificarlo y las prácticas recomendadas para agilizar un caso.

Pasos para la implementación

1. Colabore con las partes interesadas de su organización para identificar a los proveedores de software y servicios en los que se basa su carga de trabajo. Documente estas dependencias.
2. Determine las necesidades de nivel de servicio de su carga de trabajo. Seleccione un plan de asistencia que se ajuste a ellas.
3. Para el software y los servicios comerciales, establezca un plan de asistencia con los proveedores.
   a. Al suscribirse a AWS Business Support o un plan superior en todas las cuentas de producción, disfrutará de tiempos de respuesta más rápidos por parte de AWS Support, lo que resulta muy recomendable. Si no dispone de Premium Support, deberá tener un plan de acción para

administrar los problemas que requieran la ayuda de AWS Support. AWS Support le ofrece una combinación de herramientas, tecnología, personal y programas diseñados para ayudarle de forma proactiva a optimizar el rendimiento, rebajar los costes e innovar rápidamente. AWS Business Support proporciona ventajas adicionales, como el acceso a AWS Trusted Advisor y AWS Personal Health Dashboard, así como tiempos de respuesta más rápidos.

4. Documente el plan de asistencia en su herramienta de administración de conocimientos. Incluya la forma de solicitar asistencia, a quién notificar si se presenta un caso de asistencia y cómo remitir a un nivel superior durante un incidente. Un wiki es un buen mecanismo para que cualquiera pueda realizar las actualizaciones necesarias en la documentación cuando tenga conocimiento de cambios en los procesos de asistencia o en los contactos.

Nivel de esfuerzo para el plan de implementación: bajo. La mayoría de los proveedores de software y servicios ofrecen planes de asistencia opcionales. Documentar y compartir las prácticas recomendadas de asistencia en su sistema de administración de conocimientos verifican que su equipo sabe qué hacer cuando se produce un problema de producción.

Recursos

Prácticas recomendadas relacionadas:

- OPS02-BP02 Los procesos y procedimientos han identificado a los propietarios

Documentos relacionados:

- Planes de AWS Support

Servicios relacionados:

- AWS Business Support
- AWS Enterprise Support

Operación

Preguntas

- Operación 8. ¿Cómo utiliza la observabilidad de la carga de trabajo en su organización?
- Operación 9. ¿Cómo hace para comprender el estado de las operaciones?

- OPERACIÓN 10. ¿Cómo administra la carga de trabajo y los eventos de operaciones?

Operación 8. ¿Cómo utiliza la observabilidad de la carga de trabajo en su organización?

Garantice un estado óptimo de la carga de trabajo al utilizar la observabilidad. Utilice métricas, registros y rastros pertinentes para obtener una visión integral del rendimiento de su carga de trabajo y abordar los problemas de manera eficiente.

Prácticas recomendadas

- OPS08-BP01 Analizar las métricas de la carga de trabajo
- OPS08-BP02 Analizar los registros de la carga de trabajo
- OPS08-BP03 Analizar los rastreos de la carga de trabajo
- OPS08-BP04 Crear alertas procesables
- OPS08-BP05 Crear paneles

OPS08-BP01 Analizar las métricas de la carga de trabajo

Después de implementar la telemetría de la aplicación, analice periódicamente las métricas recopiladas. Si bien la latencia, las solicitudes, los errores y la capacidad (o las cuotas) proporcionan información sobre el rendimiento del sistema, es fundamental dar prioridad la revisión de las métricas de resultados empresariales. Esto garantiza que tome decisiones basadas en datos alineadas con sus objetivos empresariales.

Resultado deseado: información veraz sobre el rendimiento de la carga de trabajo que genera decisiones basadas en datos y garantiza la alineación con los objetivos empresariales.

Patrones comunes de uso no recomendados:

- Analizar las métricas de forma aislada sin tener en cuenta su impacto en los resultados empresariales.
- Confiar de forma excesiva en las métricas técnicas y, al mismo tiempo, dejar de lado las métricas empresariales.
- Revisar infrecuentemente las métricas, lo que hace que se pierdan oportunidades de toma de decisiones en tiempo real.

Beneficios de establecer esta práctica recomendada:

- Comprensión mejorada de la correlación entre el rendimiento técnico y los resultados empresariales.
- Proceso de toma de decisiones mejorado basado en datos en tiempo real.
- Identificación y mitigación proactivas de los problemas antes de que afecten a los resultados empresariales.

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

Guía para la implementación

Utilice herramientas como Amazon CloudWatch para realizar análisis de métricas. Los servicios de AWS como AWS Cost Anomaly Detection y Amazon DevOps Guru pueden utilizarse para detectar anomalías, especialmente cuando se desconocen los umbrales estáticos o cuando los patrones de comportamiento son más adecuados para la detección de anomalías.

Pasos para la implementación

1. Analice y revise: revise e interprete periódicamente las métricas de carga de trabajo.
   a. Dé prioridad a las métricas de resultados empresariales sobre las métricas puramente técnicas.
   b. Comprenda la importancia de los picos, las caídas o los patrones en sus datos.
2. Utilice Amazon CloudWatch: utilice Amazon CloudWatch para obtener una vista centralizada y un análisis profundo.
   a. Configure paneles de CloudWatch para visualizar sus métricas y compararlas a lo largo del tiempo.
   b. Utilice percentiles en CloudWatch para obtener una vista clara de la distribución de métricas, lo que puede ayudar a definir los SLA y comprender los valores atípicos.
   c. Configure AWS Cost Anomaly Detection para identificar patrones inusuales sin depender de umbrales estáticos.
   d. Implemente la observabilidad multicuenta de CloudWatch para supervisar y solucionar problemas de las aplicaciones que abarcan varias cuentas dentro de una región.
   e. Utilice CloudWatch Metric Insights para consultar y analizar datos de métricas en cuentas y regiones, identificando tendencias y anomalías.
   f. Utilice la calculadora de métricas de CloudWatch para transformar, añadir o realizar cálculos en sus métricas a fin de obtener información más detallada.

3. Emplee Amazon DevOps Guru: incorpore Amazon DevOps Guru por su detección de anomalías mejorada con machine learning para identificar los primeros signos de problemas operativos en sus aplicaciones sin servidor y solucionarlos antes de que afecten a sus clientes.
4. Optimice en función de los conocimientos: tome decisiones informadas en función de su análisis de métricas para ajustar y mejorar sus cargas de trabajo.

Nivel de esfuerzo para el plan de implementación: Medio

Recursos

Prácticas recomendadas relacionadas:

- OPS04-BP01 Identificar los indicadores clave de rendimiento
- OPS04-BP02 Implementar telemetría de aplicaciones

Documentos relacionados:

- The Wheel Blog - Emphasizing the importance of continually reviewing metrics
- Percentile are important
- Using AWS Cost Anomaly Detection
- observabilidad multicuenta de CloudWatch
- Query your metrics with CloudWatch Metrics Insights

Vídeos relacionados:

- Enable Cross-Account Observability in Amazon CloudWatch
- Introduction to Amazon DevOps Guru
- Continuously Analyze Metrics using AWS Cost Anomaly Detection

Ejemplos relacionados:

- Taller sobre observabilidad
- Gaining operation insights with AIOps using Amazon DevOps Guru

OPS08-BP02 Analizar los registros de la carga de trabajo

El análisis periódico de los registros de la carga de trabajo es esencial para adquirir una comprensión exhaustiva de los aspectos operativos de su aplicación. Al examinar, visualizar e interpretar de manera eficiente los datos de registro, puede optimizar continuamente el rendimiento y la seguridad de las aplicaciones.

Resultado deseado: amplios conocimientos sobre el comportamiento y las operaciones de las aplicaciones derivados de un análisis exhaustivo de los registros, lo que garantiza la detección y mitigación proactivas de los problemas.

Patrones comunes de uso no recomendados:

- Descuidar el análisis de los registros hasta que surja un problema crítico.
- No utilizar el conjunto completo de herramientas disponibles para el análisis de registros, lo que significa perder información crucial.
- Confiar únicamente en la revisión manual de los registros sin utilizar las capacidades de automatización y consulta.

Beneficios de establecer esta práctica recomendada:

- Identificación proactiva de los cuellos de botella operativos, las amenazas a la seguridad y otros posibles problemas.
- Utilización eficiente de los datos de registro para la optimización continua de las aplicaciones.
- Mejor comprensión del comportamiento de las aplicaciones, lo que ayuda a depurar y solucionar problemas.

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

Guía para la implementación

Amazon CloudWatch Logs es una poderosa herramienta para el análisis de registros. Las características integradas, como CloudWatch Logs Insights e Información de colaboradores, hacen que el proceso de obtener información significativa de los registros sea intuitivo y eficiente.

Pasos para la implementación

1. Configure CloudWatch Logs: configure las aplicaciones y los servicios para que enviar los registros a CloudWatch Logs.

2. Configure CloudWatch Logs Insights: Utilice CloudWatch Logs Insights para buscar y analizar sus datos de registro de forma interactiva.
   a. Cree consultas para extraer patrones, visualizar datos de registro y obtener información procesable.
3. Utilice la información de los colaboradores: Utilice Información de colaboradores de CloudWatch para identificar a los principales interlocutores en dimensiones de alta cardinalidad, como las direcciones IP o los agentes de usuario.
4. Implemente filtros de métricas de CloudWatch Logs: Configure filtros de métricas de registro de CloudWatch para convertir los datos de registro en métricas procesables. Esto le permite configurar alarmas o analizar más a fondo los patrones.
5. Revisión y perfeccionamiento periódicos: revise periódicamente sus estrategias de análisis de registros para recoger toda la información pertinente y optimizar continuamente el rendimiento de las aplicaciones.

Nivel de esfuerzo para el plan de implementación: Medio.

Recursos

Prácticas recomendadas relacionadas:

- OPS04-BP01 Identificar los indicadores clave de rendimiento
- OPS04-BP02 Implementar telemetría de aplicaciones
- OPS08-BP01 Analizar las métricas de la carga de trabajo

Documentos relacionados:

- Análisis de los datos de registros con CloudWatch Logs Insights
- Using CloudWatch Contributor Insights
- Creating and Managing CloudWatch Logs Log Metric Filters

Vídeos relacionados:

- Analyze Log Data with CloudWatch Logs Insights
- Use CloudWatch Contributor Insights to Analyze High-Cardinality Data

Ejemplos relacionados:

- Consultas de ejemplo de CloudWatch Logs
- Taller sobre observabilidad

OPS08-BP03 Analizar los rastreos de la carga de trabajo

El análisis de los datos de rastreo es crucial para lograr una visión integral del recorrido operativo de una aplicación. Al visualizar y comprender las interacciones entre varios componentes, se puede ajustar el rendimiento, identificar los cuellos de botella y mejorar las experiencias de los usuarios.

Resultado deseado: logre una visibilidad clara de las operaciones distribuidas de su aplicación, lo que permite una resolución de problemas más rápida y una mejor experiencia del usuario.

Patrones comunes de uso no recomendados:

- Pasar por alto los datos de rastreo y confiar únicamente en los registros y las métricas.
- No se correlacionan los datos de rastreo con los registros asociados.
- Hacer caso omiso de las métricas derivadas de los rastreos, como la latencia y las tasas de errores.

Beneficios de establecer esta práctica recomendada:

- Mejore la solución de problemas y reduzca el tiempo medio de resolución (MTTR).
- Obtenga información sobre las dependencias y su impacto.
- Identifique y rectifique rápidamente los problemas de rendimiento.
- Utilice las métricas derivadas de los rastreos para tomar decisiones informadas.
- Mejore la experiencia del usuario mediante interacciones de componentes optimizadas.

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

Guía para la implementación

AWS X-Ray ofrece un conjunto completo para el análisis de datos de rastreo, que proporciona una visión integral de las interacciones del servicio, supervisa las actividades de los usuarios y detecta problemas de rendimiento. Características como ServiceLens, X-Ray Insights, X-Ray Analytics y

Amazon DevOps Guru mejoran la profundidad de la información procesable derivada de los datos de rastreo.

Pasos para la implementación

Los siguientes pasos ofrecen un enfoque estructurado para implementar de manera eficaz el análisis de datos de rastreo mediante servicios de AWS:

1. Integre AWS X-Ray: asegúrese de que X-Ray esté integrado con sus aplicaciones para obtener datos de rastreo.
2. Analice métricas de X-Ray: profundice en las métricas derivadas de los rastreos de X-Ray, como la latencia, las tasas de solicitudes, las tasas de errores y las distribuciones del tiempo de respuesta mediante el mapa de servicios para supervisar el estado de las aplicaciones.
3. Utilice ServiceLens: utilice el mapa de ServiceLens para mejorar la observabilidad de sus servicios y aplicaciones. Esto permite la visualización integrada de rastreos, métricas, registros, alarmas y otra información de estado.
4. Habilite X-Ray Insights:
   a. Active X-Ray Insights para que detecte automáticamente las anomalías en los rastreos.
   b. Examine la información para identificar patrones y determinar las causas raíz, como el aumento de tasas de errores o latencias.
   c. Consulte el cronograma de información para obtener un análisis cronológico de los problemas detectados.
5. Utilice X-Ray Analytics: X-Ray Analytics le permite explorar a fondo los datos de rastreo, identificar patrones y extraer información.
6. Use grupos en X-Ray: cree grupos en X-Ray para filtrar los rastreos en función de criterios como la alta latencia, lo que permite un análisis más específico.
7. Incorpore Amazon DevOps Guru: utilice Amazon DevOps Guru para beneficiarse de los modelos de machine learning que identifican anomalías operativas en los rastreos.
8. Utilice CloudWatch Synthetics: Utilice CloudWatch Synthetics para crear valores controlados para supervisar continuamente sus puntos de conexión y flujos de trabajo. Estos valores controlados pueden integrarse con X-Ray para proporcionar datos de rastreo para un análisis en profundidad de las aplicaciones que se están probando.
9. Utilice la supervisión de usuarios reales (RUM): Con AWS X-Ray y CloudWatch RUM, puede analizar y depurar la ruta de solicitud desde los usuarios finales de la aplicación hasta los servicios downstream administrados por AWS. Esto le ayuda a identificar las tendencias de latencia y los errores que afectan a sus usuarios.

10Establezca una correlación con los registros: correlacione los datos de rastreo con los registros relacionados dentro de la vista de rastreos de X-Ray para obtener una perspectiva detallada del comportamiento de la aplicación. Esto le permite ver los eventos de registro directamente asociados con las transacciones rastreadas.

Nivel de esfuerzo para el plan de implementación: Medio.

Recursos

Prácticas recomendadas relacionadas:

- OPS08-BP01 Analizar las métricas de la carga de trabajo
- OPS08-BP02 Analizar los registros de la carga de trabajo

Documentos relacionados:

- Using ServiceLens to Monitor Application Health
- Exploring Trace Data with X-Ray Analytics
- Detecting Anomalies in Traces with X-Ray Insights
- Continuous Monitoring with CloudWatch Synthetics

Vídeos relacionados:

- Analyze and Debug Applications Using Amazon CloudWatch Synthetics and AWS X-Ray
- Utilice AWS X-Ray Insights

Ejemplos relacionados:

- Taller sobre observabilidad
- Implementing X-Ray with AWS Lambda
- Plantillas de CloudWatch Synthetics Canary

OPS08-BP04 Crear alertas procesables

Es crucial detectar y responder rápidamente a las desviaciones en el comportamiento de su aplicación. Es especialmente vital reconocer cuándo están en peligro los resultados basados en los

indicadores clave de rendimiento (KPI) o cuándo surgen anomalías inesperadas. Basar las alertas en los KPI garantiza que las señales que reciba estén directamente relacionadas con el impacto empresarial u operativo. Este enfoque de alertas procesables promueve respuestas proactivas y ayuda a mantener el rendimiento y la fiabilidad del sistema.

Resultado deseado: recibe alertas oportunas, pertinentes y procesables para identificar y mitigar rápidamente los posibles problemas, especialmente cuando los resultados de los KPI están en peligro.

Patrones comunes de uso no recomendados:

- Configurar demasiadas alertas que no son cruciales, lo que provoca un exceso de alertas.
- No se da prioridad a las alertas en función de los KPI, lo que dificulta la comprensión del impacto empresarial de los problemas.
- Si no se abordan las causas raíz, se generan alertas repetitivas sobre el mismo problema.

Beneficios de establecer esta práctica recomendada:

- Se reduce el exceso de alertas al poner el foco en las alertas pertinentes y procesables.
- Mejora del tiempo de actividad y la fiabilidad del sistema gracias a la detección y mitigación proactivas de problemas.
- Mejora de la colaboración en equipo y resolución de problemas más rápida mediante la integración con herramientas de alerta y comunicación populares.

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

Guía para la implementación

Para crear un mecanismo de alerta eficaz, es fundamental utilizar métricas, registros y datos de rastreo que indiquen cuándo los resultados basados en los KPI están en peligro o se detectan anomalías.

Pasos para la implementación

1. Defina los indicadores clave de rendimiento (KPI). Identifique los KPI de su aplicación. Las alertas deben estar vinculadas a estos KPI para reflejar el impacto empresarial con precisión.
2. Implemente la detección de anomalías:

- Utilice AWS Cost Anomaly Detection: Configure AWS Cost Anomaly Detection para detectar automáticamente patrones inusuales, lo que garantiza que las alertas solo se generen en caso de auténticas anomalías.
- Utilice X-Ray Insights:
  a. Configure X-Ray Insights para que detecte anomalías en los datos de rastreo.
  b. Configure notificaciones para X-Ray Insights para recibir alertas sobre los problemas detectados.
- Integre con DevOps Guru:
  a. Utilice Amazon DevOps Guru por sus capacidades de machine learning para detectar anomalías operativas con los datos existentes.
  b. Navegue hasta la configuración de notificaciones en DevOps Guru para configurar alertas de anomalías.

3. Implemente alertas procesables: diseñe alertas que proporcionen la información adecuada para tomar medidas de inmediato.
4. Reduzca el exceso de alarmas: minimice las alertas que no son cruciales. Abrumar a los equipos con numerosas alertas insignificantes puede llevar a que se acaben pasando por algo problemas críticos y a una reducción de la eficacia general del mecanismo de alerta.
5. Configure alarmas compuestas: Utilice alarmas compuestas de Amazon CloudWatch para consolidar varias alarmas.
6. Realice integraciones con herramientas de alerta: incorpore herramientas como Ops Genie y PagerDuty.
7. Utilice Amazon Q Developer in chat applications Integrar Amazon Q Developer in chat applicationspara transmitir alertas a Chime, Microsoft Teams y Slack.
8. Alerta basada en registros: Utilice los filtros de métricas de registro en CloudWatch para crear alarmas basadas en eventos de registro concretos.
9. Revise e itere: revise y perfeccione periódicamente las configuraciones de las alertas.

Nivel de esfuerzo para el plan de implementación: Medio.

Recursos

Prácticas recomendadas relacionadas:

- OPS04-BP01 Identificar los indicadores clave de rendimiento

- OPS04-BP02 Implementar telemetría de aplicaciones
- OPS04-BP03 Implementar la telemetría de la experiencia del usuario
- OPS04-BP04 Implementar telemetría de dependencias
- OPS04-BP05 Implementar el rastreo distribuido
- OPS08-BP01 Analizar las métricas de la carga de trabajo
- OPS08-BP02 Analizar los registros de la carga de trabajo
- OPS08-BP03 Analizar los rastreos de la carga de trabajo

Documentos relacionados:

- Using Amazon CloudWatch Alarms
- Create a composite alarm
- Create a CloudWatch alarm based on anomaly detection
- DevOps Guru Notifications
- X-Ray Insights notifications
- Monitoree, opere y resuelva problemas en sus recursos de AWS con ChatOps interactivos
- Amazon CloudWatch Integration Guide | PagerDuty
- Integrate OpsGenie with Amazon CloudWatch

Vídeos relacionados:

- Create Composite Alarms in Amazon CloudWatch
- Amazon Q Developer in chat applications Overview
- AWS on Air ft. Mutative Commands in Amazon Q Developer in chat applications

Ejemplos relacionados:

- Alarms, incident management, and remediation in the cloud with Amazon CloudWatch
- Tutorial: Creating an Amazon EventBridge rule that sends notifications to Amazon Q Developer in chat applications
- Taller sobre observabilidad

OPS08-BP05 Crear paneles

Los paneles son la perspectiva centrada en las personas de los datos de telemetría de sus cargas de trabajo. Si bien proporcionan una interfaz visual vital, no deben reemplazar los mecanismos de alerta, sino complementarlos. Cuando se diseñan con cuidado, no solo pueden ofrecer información rápida sobre el estado y el rendimiento del sistema, sino que también pueden presentar a las partes interesadas información en tiempo real sobre los resultados empresariales y el impacto de los problemas.

Resultado deseado: información clara y procesable sobre el estado del sistema y la empresa mediante representaciones visuales.

Patrones comunes de uso no recomendados:

- Paneles demasiado complicados con demasiadas métricas.
- Confiar en los paneles sin alertas de detección de anomalías.
- No actualizar los paneles a medida que evolucionan las cargas de trabajo.

Beneficios de establecer esta práctica recomendada:

- Visibilidad inmediata de las métricas y los KPI cruciales del sistema.
- Mejora de la comunicación y la comprensión de las partes interesadas.
- Información rápida sobre el impacto de los problemas operativos.

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

Guía para la implementación

Paneles centrados en la empresa

Los paneles adaptados a los KPI empresariales implican a un mayor número de partes interesadas. Si bien es posible que estas personas no estén interesadas en las métricas del sistema, están interesadas en comprender las implicaciones empresariales de estas cifras. Un panel centrado en la empresa garantiza que todas las métricas técnicas y operativas que se supervisan y analizan estén en sintonía con los objetivos empresariales generales. Esta alineación proporciona claridad y garantiza que todo el mundo coincida en lo que es esencial y lo que no. Además, los paneles que destacan los KPI empresariales suelen ser más procesables. Las partes interesadas pueden

comprender rápidamente el estado de las operaciones, las áreas que requieren atención y el impacto potencial en los resultados empresariales.

Con esto en mente, al crear sus paneles, asegúrese de que haya un equilibrio entre las métricas técnicas y los KPI empresariales. Ambos son vitales, pero se dirigen a públicos diferentes. Lo ideal sería disponer de paneles que proporcionen una visión integral del estado y el rendimiento del sistema y, al mismo tiempo, hagan hincapié en los resultados empresariales clave y sus implicaciones.

Los paneles de Amazon CloudWatch son páginas de inicio personalizables de la consola de CloudWatch que puede usar para supervisar los recursos en una sola vista, incluso aquellos que están repartidos por diferentes cuentas y Regiones de AWS.

Pasos para la implementación

1. Cree un panel básico: cree un panel nuevo en CloudWatch, y asígnele un nombre descriptivo.
2. Use widgets de Markdown: antes de profundizar en las métricas, utilice widgets de Markdown para añadir un contexto textual en la parte superior del panel. Debe explicar lo que cubre el panel, la importancia de las métricas representadas y también puede contener enlaces a otros paneles y herramientas de solución de problemas.
3. Cree variables de panel: incorpore variables de panel cuando proceda, para ofrecer vistas dinámicas y flexibles del panel.
4. Cree widgets de métricas: añada widgets de métricas para visualizar las diversas métricas que emite su aplicación. Adapte estos widgets para que representen de forma eficaz el estado del sistema y los resultados empresariales.
5. Consultas de Log Insights: utilice CloudWatch Logs Insights para obtener métricas procesables de sus registros y mostrar esta información en su panel.
6. Configure alarmas: Integrar las alarmas de CloudWatch en el panel para ver rápidamente cualquier métrica que supere sus umbrales.
7. Utilice Información de colaboradores: incorpore Información de colaboradores de CloudWatch para analizar los campos de alta cardinalidad y obtener una comprensión más clara de los principales contribuyentes de su recurso.
8. Diseñe widgets personalizados: para necesidades concretas que los widgets estándar no satisfacen, considere la posibilidad de crear widgets personalizados. Pueden proceder de varios orígenes de datos o representar los datos de formas únicas.
9. Itere y refine: a medida que evolucione la aplicación, revise periódicamente el panel para asegurarse de que siga siendo relevante.

Recursos

Prácticas recomendadas relacionadas:

- OPS04-BP01 Identificar los indicadores clave de rendimiento
- OPS08-BP01 Analizar las métricas de la carga de trabajo
- OPS08-BP02 Analizar los registros de la carga de trabajo
- OPS08-BP03 Analizar los rastreos de la carga de trabajo
- OPS08-BP04 Crear alertas procesables

Documentos relacionados:

- La creación de paneles para la visibilidad operativa
- Using Amazon CloudWatch Dashboards

Vídeos relacionados:

- Create Cross Account & Cross Region CloudWatch Dashboards
- AWS re:Invent 2021 - Gain enterprise visibility with Nube de AWS operation dashboards

Ejemplos relacionados:

- Taller sobre observabilidad
- Monitoreo de aplicaciones con Amazon CloudWatch

Operación 9. ¿Cómo hace para comprender el estado de las operaciones?

Defina, capture y analice las métricas de las operaciones para obtener visibilidad de los eventos de operaciones y poder tomar las medidas adecuadas.

Prácticas recomendadas

- OPS09-BP01 Medir los objetivos operativos y los KPI con métricas
- OPS09-BP02 Comunicar el estado y las tendencias para garantizar la visibilidad de la operación
- OPS09-BP03 Revisar las métricas de las operaciones y dar prioridad a las mejoras

OPS09-BP01 Medir los objetivos operativos y los KPI con métricas

Obtenga objetivos y KPI que definan el éxito de las operaciones de su organización y determine las métricas que los reflejen. Establezca líneas de base como puntos de referencia y reevalúelas periódicamente. Desarrolle mecanismos para recopilar estas métricas de los equipos para su evaluación.

Resultado deseado:

- Se han publicado y compartido los objetivos y los KPI de los equipos de operaciones de la organización.
- Se establecen métricas que reflejan estos KPI. Algunos ejemplos podrían ser:
  - Profundidad de la cola de tickets o antigüedad media de los tickets.
  - Recuento de tickets agrupado por tipo de problema.
  - Tiempo dedicado a resolver problemas con o sin un procedimiento operativo estandarizado (SOP).
  - Cantidad de tiempo empleado en recuperarse de un error producido al introducir código.
  - Volumen de llamadas.

Patrones comunes de uso no recomendados:

- No se cumplen los plazos de despliegue porque los desarrolladores se ven obligados a realizar tareas de solución de problemas. Los equipos de desarrollo abogan por más personal, pero no pueden indicar cuántas personas necesitan porque no se puede medir el tiempo empleado.
- Se configuró un servicio de asistencia de nivel 1 para gestionar las llamadas de los usuarios. Con el tiempo, se añadieron más cargas de trabajo, pero no se asignó personal al servicio de asistencia de nivel 1. La satisfacción de los clientes se resiente a medida que aumenta la duración de las llamadas y los problemas tardan más en resolverse, pero la administración no ve ningún indicador de ello, lo que impide tomar medidas.
- Una carga de trabajo problemática se ha transferido a un equipo de operaciones independiente para su gestión. A diferencia de otras cargas de trabajo, esta nueva carga no se suministró con la documentación y los runbooks adecuados. Por lo tanto, los equipos dedican más tiempo a solucionar problemas y hacer frente a errores. Sin embargo, no hay métricas que lo documenten, lo que dificulta la rendición de cuentas.

Beneficios de establecer esta práctica recomendada: mientras que la supervisión de la carga de trabajo muestra el estado de nuestras aplicaciones y servicios, la supervisión de los equipos de operaciones permite a los propietarios obtener información sobre los cambios que se producen entre los consumidores de esas cargas de trabajo, como los cambios en las necesidades empresariales. Mida la eficacia de estos equipos y evalúelos con respecto a los objetivos empresariales mediante la creación de métricas que puedan reflejar el estado de las operaciones. Las métricas pueden resaltar los problemas de asistencia o identificar cuándo se producen desviaciones respecto a un objetivo de nivel de servicio.

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

Guía para la implementación

Programe tiempo con la dirección empresarial y las partes interesadas para determinar los objetivos generales del servicio. Determine cuáles deberían ser las tareas de los distintos equipos de operaciones y qué desafíos podrían presentárseles. Con estos, haga una lluvia de ideas sobre los indicadores clave de rendimiento (KPI) para reflejar los objetivos operativos. Podría ser la satisfacción del cliente, el tiempo transcurrido desde la concepción de la característica hasta el despliegue o el tiempo promedio de resolución de problemas, entre otras cosas.

A partir de los KPI, identifique las métricas y los orígenes de datos que podrían reflejar mejor estos objetivos. La satisfacción del cliente podría ser una combinación de varios indicadores, como los tiempos de espera o respuesta de las llamadas, las puntuaciones de satisfacción y los tipos de problemas planteados. Los tiempos de despliegue podrían ser la suma del tiempo necesario para las pruebas y el despliegue, además de las correcciones posteriores al despliegue que deban añadirse. Las estadísticas que muestran el tiempo dedicado a diferentes tipos de problemas (o el recuento de esos problemas) pueden proporcionar una panorámica de dónde se necesita un esfuerzo específico.

Recursos

Documentos relacionados:

- Quick - Using KPIs
- Amazon CloudWatch - Using Metrics
- Creación de paneles
- How to track your cost optimization KPIs with KPI Dashboard

OPS09-BP02 Comunicar el estado y las tendencias para garantizar la visibilidad de la operación

Es necesario conocer el estado de sus operaciones y la dirección de sus tendencias para identificar qué resultados corren peligro, si se puede respaldar o no el trabajo adicional o los efectos que los cambios han tenido en sus equipos. Durante los eventos de operaciones, disponer de páginas de estado que los usuarios y los equipos de operaciones puedan consultar para obtener información puede reducir la presión sobre los canales de comunicación y difundir la información de forma proactiva.

Resultado deseado:

- La dirección de operaciones puede ver de un vistazo el volumen de llamadas que reciben sus equipos y las actividades que se están llevando a cabo, como los despliegues.
- Las alertas se difunden a las partes interesadas y las comunidades de usuarios cuando se producen repercusiones en las operaciones normales.
- La dirección de la organización y las partes interesadas pueden consultar una página de estado en respuesta a una alerta o una repercusión y obtener información sobre un evento operativo, como puntos de contacto, información de tickets y tiempos de recuperación estimados.
- Los informes se ponen a disposición de la dirección y otras partes interesadas para mostrar las estadísticas de las operaciones, como el volumen de llamadas durante un período de tiempo, las puntuaciones de satisfacción de los usuarios, el número de entradas pendientes y su antigüedad.

Patrones comunes de uso no recomendados:

- Una carga de trabajo deja de funcionar y un servicio no está disponible. El volumen de llamadas aumenta a medida que los usuarios quieren saber qué pasa. Los administradores contribuyen al aumento del volumen de solicitudes pues quieren saber quién está trabajando en el problema. Varios equipos de operaciones duplican sus esfuerzos al tratar de investigar.
- El interés por una nueva capacidad lleva a la reasignación de varios miembros del personal a actividades de ingeniería. No se proporcionan refuerzos y los tiempos de resolución de problemas aumentan. Esta información no se recopila, y la dirección no se da cuenta del problema hasta después de varias semanas y de que los usuarios muestren su insatisfacción.

Beneficios de establecer esta práctica recomendada: durante los eventos operativos que afectan a la empresa, se puede desperdiciar mucho tiempo y energía solicitando información a varios equipos para intentar comprender la situación. Al establecer paneles y páginas de estado ampliamente difundidos, las partes interesadas pueden obtener rápidamente información sobre si se detectó o

no un problema, quién se encarga del problema o cuándo se espera que las operaciones vuelvan a la normalidad. Esto evita que los miembros del equipo dediquen demasiado tiempo a comunicar su estado a los demás y dediquen más tiempo a abordar los problemas.

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

Guía para la implementación

Cree paneles que muestren las métricas clave actuales de sus equipos de operaciones y póngalos a disposición tanto de la dirección de operaciones como de la administración.

Cree páginas de estado que se puedan actualizar rápidamente para mostrar cuándo se produce un incidente o evento, quién es el propietario y quién coordina la respuesta. Comparta en esta página todos los pasos o soluciones que los usuarios deberían tener en cuenta y difunda ampliamente la ubicación. Anime a los usuarios a comprobar primero esta ubicación cuando se enfrenten a un problema desconocido.

Recopile y proporcione informes que muestren el estado de las operaciones a lo largo del tiempo y distribúyalos entre la dirección y los responsables de la toma de decisiones para ilustrar el trabajo de operaciones junto con los desafíos y las necesidades.

Comparta con los equipos las métricas e informes que mejor reflejen los objetivos y los KPI y en qué aspectos han influido a la hora de impulsar el cambio. Dedique tiempo a estas actividades para aumentar la importancia de las operaciones dentro de los equipos y entre ellos.

Recursos

Documentos relacionados:

- Measure Progress
- La creación de paneles para la visibilidad operativa

Soluciones relacionadas:

- Data Operations

OPS09-BP03 Revisar las métricas de las operaciones y dar prioridad a las mejoras

Destinar tiempo y recursos dedicados a revisar el estado de las operaciones garantiza que atender la línea empresarial diaria siga siendo una prioridad. Reúna a la dirección de operaciones y las partes

interesadas para revisar periódicamente las métricas, reafirmar o modificar las metas y los objetivos y dar prioridad a las mejoras.

Resultado deseado:

- La dirección y el personal de operaciones se reúnen periódicamente para revisar las métricas durante un período de informe determinado. Se comunican los desafíos, se celebran las victorias y se comparten las lecciones aprendidas.
- Las partes interesadas y la dirección empresarial reciben información periódica sobre el estado de las operaciones y se les pide su opinión sobre los objetivos, los KPI y las iniciativas futuras. Se analizan y contextualizan las compensaciones entre la prestación de servicios, las operaciones y el mantenimiento.

Patrones comunes de uso no recomendados:

- Se lanza un nuevo producto, pero los equipos de operaciones de nivel 1 y nivel 2 no están adecuadamente capacitados para ofrecer respaldo ni cuentan con personal adicional. La dirección no ve las métricas que muestran el empeoramiento de los tiempos de resolución de los tickets y el aumento del volumen de incidentes. No se toman medidas hasta que han transcurrido varias semanas, cuando el número de suscriptores comienza a caer porque los usuarios descontentos abandonan la plataforma.
- Hace mucho tiempo que existe un proceso manual para realizar el mantenimiento de una carga de trabajo. Si bien había interés por automatizar, esta era una prioridad baja dada la poca importancia del sistema. Sin embargo, con el tiempo, el sistema ha ido ganando importancia y ahora estos procesos manuales consumen la mayor parte del tiempo de las operaciones. No hay recursos programados para proporcionar más herramientas a las operaciones, lo que provoca el agotamiento del personal a medida que aumentan las cargas de trabajo. La dirección se da cuenta cuando se les informa que el personal se va a la competencia.

Beneficios de establecer esta práctica recomendada: en algunas organizaciones, puede ser desafiante asignar el mismo tiempo y atención que se dedica a la prestación de servicios y a los nuevos productos u ofertas. Cuando esto ocurre, la línea empresarial puede resentirse a medida que el nivel de servicio esperado se deteriora lentamente. Esto se debe a que las operaciones no cambian ni evolucionan con el crecimiento de la empresa y pronto pueden quedarse rezagadas. Sin una revisión periódica de la información que recopilan las operaciones, es posible que el riesgo para la empresa solo resulte evidente cuando sea demasiado tarde. Al asignar tiempo para revisar las métricas y los procedimientos tanto entre el personal de operaciones como con la dirección, el

papel crucial que desempeñan las operaciones permanece visible y los riesgos se pueden identificar mucho antes de que alcancen niveles críticos. Los equipos de operaciones obtienen una mejor perspectiva de los cambios e iniciativas empresariales inminentes, lo que permite realizar esfuerzos proactivos. La visibilidad de la dirección de las métricas de las operaciones muestra el papel que desempeñan estos equipos en la satisfacción del cliente, tanto interna como externa, y les permite sopesar mejor las opciones en función de las prioridades, o garantizar que las operaciones tengan el tiempo y los recursos para cambiar y evolucionar con las nuevas iniciativas empresariales y de carga de trabajo.

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

Guía para la implementación

Dedique tiempo a revisar las métricas de las operaciones entre las partes interesadas y los equipos de operaciones y a revisar los datos de los informes. Analice estos informes en el contexto de las metas y los objetivos de la organización para determinar si se están cumpliendo. Identifique las fuentes de ambigüedad en las que las metas no estén claras o en las que pueda haber conflictos entre lo que se pide y lo que se da.

Identifique dónde el tiempo, las personas y las herramientas pueden ayudar a obtener resultados operativos. Determine a qué KPI afectaría esto y cuáles deberían ser los objetivos de éxito. Vuelva a examinar todo esto periódicamente a fin de garantizar que las operaciones cuenten con los recursos suficientes para respaldar la línea empresarial.

Recursos

Documentos relacionados:

- Amazon Athena
- Referencia de métricas y dimensiones de Amazon CloudWatch
- Amazon Quick
- AWS Glue
- AWS Glue Data Catalog
- Collect metrics and logs from Amazon EC2 instances and on-premises servers with the Amazon CloudWatch Agent
- Using Amazon CloudWatch metrics

OPERACIÓN 10. ¿Cómo administra la carga de trabajo y los eventos de operaciones?

Prepare y valide los procedimientos de respuesta a los eventos para minimizar la interrupción de la carga de trabajo.

Prácticas recomendadas

- OPS10-BP01 Uso de un proceso para la administración de eventos, incidentes y problemas
- OPS10-BP02 Tener un proceso por alerta
- OPS10-BP03 Prioridad de los eventos operativos según el impacto empresarial
- OPS10-BP04 Definir rutas de escalado
- OPS10-BP05 Definir un plan de comunicación con los clientes en caso de interrupciones del servicio
- OPS10-BP06 Comunicar el estado a través de paneles
- OPS10-BP07 Automatizar las respuestas a eventos

OPS10-BP01 Uso de un proceso para la administración de eventos, incidentes y problemas

Su organización tiene procesos para gestionar eventos, incidentes y problemas. Los eventos son cosas que ocurren en su carga de trabajo pero que podrían no necesitar intervención. Los incidentes son eventos que requieren intervención. Los problemas son eventos recurrentes que requieren una intervención o que no pueden resolverse. Necesita procesos para mitigar el impacto de estos eventos en su negocio y asegurarse de que responde adecuadamente.

Cuando se producen incidentes y problemas en su carga de trabajo, necesita procesos para gestionarlos. ¿Cómo va a comunicar el estado del evento a las partes interesadas? ¿Quién supervisa la dirección de la respuesta? ¿Cuáles son las herramientas que utiliza para mitigar el evento? Estos son ejemplos de algunas de las preguntas que debe responder para tener un proceso de respuesta sólido.

Los procesos deben estar documentados en un lugar central y a disposición de cualquier persona involucrada en su carga de trabajo. Si no tiene un wiki central o un almacén de documentos, se puede utilizar un repositorio de control de versiones. Mantendrá estos planes actualizados a medida que sus procesos evolucionen.

Los problemas son candidatos a la automatización. Estos eventos le restan tiempo a su capacidad de innovar. Empiece por crear un proceso repetible para mitigar el problema. Con el tiempo, céntrese

en automatizar la mitigación o en solucionar el problema subyacente. Esto libera tiempo para dedicarlo a hacer mejoras en su carga de trabajo.

Resultado deseado: Su organización tiene un proceso para gestionar eventos, incidentes y problemas. Estos procesos se documentan y almacenan en un lugar central. Se actualizan a medida que cambian los procesos.

Patrones comunes de uso no recomendados:

- Se produce un incidente en el fin de semana y el ingeniero de guardia no sabe qué hacer.
- Un cliente le envía un correo electrónico diciendo que la aplicación no funciona. Se reinicia el servidor para solucionarlo. Esto ocurre con frecuencia.
- Hay un incidente en el que varios equipos trabajan de forma independiente para intentar resolverlo.
- Los despliegues ocurren en su carga de trabajo sin registrarse.

Beneficios de establecer esta práctica recomendada:

- Tiene una pista de auditoría de los eventos en su carga de trabajo.
- Su tiempo para recuperarse de un incidente disminuye.
- Los miembros del equipo pueden resolver incidentes y problemas de manera coherente.
- Hay un esfuerzo más consolidado cuando se investiga un incidente.

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

Guía para la implementación

La implementación de esta práctica recomendada implica el seguimiento de los eventos de la carga de trabajo. Dispone de procesos para gestionar las incidencias y los problemas. Los procesos se documentan, se comparten y se actualizan con frecuencia. Los problemas se identifican, se priorizan y se solucionan.

Ejemplo de cliente

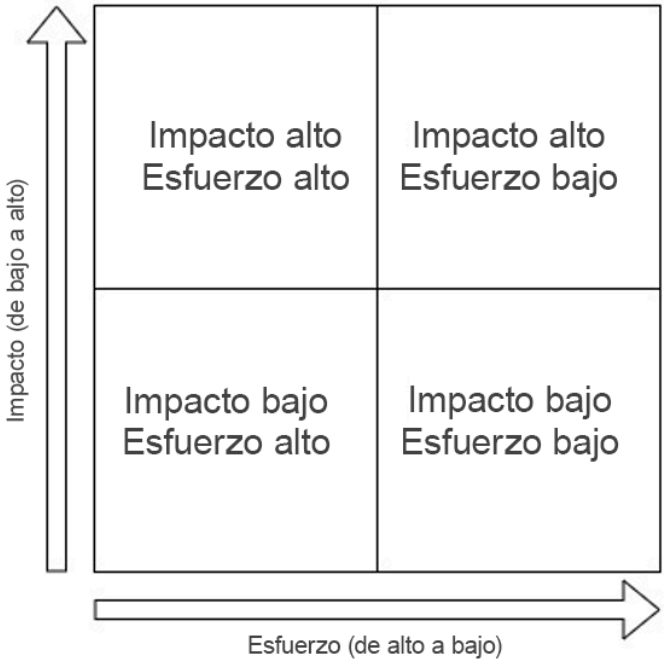
AnyCompany Retail tiene una parte de su wiki interna dedicada a los procesos de gestión de eventos, incidentes y problemas. Todos los eventos se envían a Amazon EventBridge. Los problemas se identifican como OpsItems en AWS Systems Manager OpsCenter y su solución se

prioriza, reduciendo la mano de obra no diferenciada. A medida que los procesos cambian, se actualizan en su wiki interna. Utilizan Administrador de incidentes de AWS Systems Manager para gestionar los incidentes y coordinar los esfuerzos de mitigación.

Pasos para la aplicación

1. Eventos
   - Realice un seguimiento de los eventos que se producen en su carga de trabajo, aunque no sea necesaria la intervención humana.
   - Trabaje con las partes interesadas en la carga de trabajo para desarrollar una lista de eventos que deben rastrearse. Algunos ejemplos son los despliegues completados o la aplicación de parches con éxito.
   - Puede utilizar servicios como Amazon EventBridge o bien Amazon Simple Notification Service para generar eventos personalizados para el seguimiento.
2. Los incidentes
   - Comience por definir el plan de comunicación de incidentes. ¿Qué partes interesadas deben ser informadas? ¿Cómo los mantendrá informados? ¿Quién supervisa los esfuerzos de coordinación? Recomendamos establecer un canal de chat interno para la comunicación y la coordinación.
   - Defina rutas de derivación para los equipos que apoyan su carga de trabajo, especialmente si el equipo no tiene una rotación de guardia. En función de su nivel de soporte, también puede registrar un caso con Soporte.
   - Cree una guía de estrategias para investigar el incidente. Debe incluir el plan de comunicación y los pasos detallados de la investigación. Incluya la comprobación del Panel de AWS Health en su investigación.
   - Documente su plan de respuesta a incidentes. Comunique el plan de gestión de incidentes para que los clientes internos y externos comprendan las normas de actuación y lo que se espera de ellos. Forme a los miembros de su equipo en cómo usarlo.
   - Los clientes pueden usar Incident Manager para establecer y gestionar su plan de respuesta a incidentes.
   - Los clientes de Enterprise Support pueden solicitar el Taller de gestión de incidentes a su gerente técnico de cuentas. Este taller guiado pone a prueba su actual plan de respuesta a incidentes y le ayuda a identificar áreas de mejora.
3. Problemas
   - Los problemas deben identificarse y seguirse en el sistema ITSM.

- Identifique todos los problemas conocidos y priorícelos según el esfuerzo para solucionarlos y según el impacto en la carga de trabajo.



- Resuelva primero los problemas de alto impacto y bajo esfuerzo. Una vez resueltos estos, pase a los problemas que entran en el cuadrante de bajo impacto y bajo esfuerzo.
- Puede usar Systems Manager OpsCenter para identificar estos problemas, adjuntarles runbooks y hacer un seguimiento de los mismos.

Nivel de esfuerzo para el plan de implementación: Medio Se necesita tanto un proceso como herramientas para implementar esta práctica recomendada. Documente sus procesos y póngalos a disposición de cualquier persona relacionada con la carga de trabajo. Actualícelos con frecuencia. Tiene un proceso para gestionar los problemas y mitigarlos o solucionarlos.

Recursos

Prácticas recomendadas relacionadas:

- OPS07-BP03 Uso de runbooks para realizar los procedimientos: los problemas conocidos necesitan un runbook asociado para que los esfuerzos de mitigación sean coherentes.
- OPS07-BP04 Usar guías de estrategias para investigar problemas: los incidentes deben investigarse utilizando guías de estrategias.
- OPS11-BP02 Realizar un análisis después del incidente: realice siempre una autopsia después de recuperarse de un incidente.

Documentos relacionados:

- Atlassian - Incident management in the age of DevOps (Atlassian: gestión de incidentes en la era de DevOps)
- AWS Security Incident Response Guide (Guía de respuesta ante incidentes de seguridad de AWS)
- Incident Management in the Age of DevOps and SRE (Gestión de incidentes en la era de DevOps y SRE)
- PagerDuty - What is Incident Management? (PagerDuty: ¿Qué es la gestión de incidentes?)

Vídeos relacionados:

- AWS re:Invent 2020: Incident management in a distributed organization (Gestión de incidencias en una organización distribuida)
- AWS re:Invent 2021 - Building next-gen applications with event-driven architectures (Creación de aplicaciones de nueva generación con arquitecturas basadas en eventos)
- AWS Supports You | Exploring the Incident Management Tabletop Exercise (AWS le apoya | Ejercicio práctico de exploración de gestión de incidentes)
- Administrador de incidentes de AWS Systems Manager - AWS Virtual Workshops (Administrador de incidentes de AWS Systems Manager: talleres virtuales de AWS)
- AWS What's Next ft. Incident Manager | AWS Events (Novedades de AWS - Incident Manager | Eventos de AWS)

Ejemplos relacionados:

- AWS Management and Governance Tools Workshop - OpsCenter (Taller de herramientas de administración y gobernanza de AWS - OpsCenter)
- AWS Proactive Services – Incident Management Workshop (Servicios proactivos de AWS: taller de gestión de incidencias)
- Building an event-driven application with Amazon EventBridge (Creación de una aplicación basada en eventos con Amazon EventBridge)
- Building event-driven architectures on AWS (Desarrollo de arquitecturas basadas en eventos en AWS)

Servicios relacionados:

- Amazon EventBridge
- Amazon SNS
- Panel de AWS Health
- Administrador de incidentes de AWS Systems Manager
- AWS Systems Manager OpsCenter

OPS10-BP02 Tener un proceso por alerta

Tenga una respuesta bien definida (runbook o guía de estrategia) con un propietario identificado de forma específica para cualquier evento del que se alerte. Esto garantiza respuestas rápidas y eficaces a eventos operativos y previene que los eventos procesables queden ocultos por notificaciones menos importantes.

Patrones de uso no recomendados comunes:

- Su sistema de supervisión le presenta un flujo de conexiones aprobadas junto con otros mensajes. El volumen de mensajes es tan grande que pasa por alto los mensajes de error periódicos que requieren su intervención.
- Recibe una alerta de que el sitio web está inactivo. No hay un proceso definido para cuando sucede esto. Se ve obligado a adoptar un enfoque ad hoc para diagnosticar y resolver el problema. El desarrollo de este proceso sobre la marcha alarga el tiempo de recuperación.

Beneficios de establecer esta práctica recomendada: Al alertar solo cuando es necesario actuar, se evita que las alertas de bajo valor oculten las de alto valor. Al contar con un proceso para cada alerta procesable, permite una respuesta coherente y rápida a los eventos de su entorno.

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

Guía para la implementación

- Procese por alerta: cualquier evento del que se alerte debe tener una respuesta bien definida (runbook o guía de estrategia) con un propietario identificado de manera específica (por ejemplo, un individuo, un equipo o un rol) responsable de una realización correcta. Una respuesta puede llevarse a cabo de forma automática o no (otro equipo puede ejecutarla); sin embargo, el propietario es el responsable de garantizar que el proceso obtenga los resultados esperados. Al contar con estos procesos, se asegura de disponer de respuestas a eventos operativos eficaces y rápidas y, además, podrá prevenir que los eventos procesables queden ocultos por notificaciones

menos importantes. Por ejemplo, Auto Scaling puede aplicarse para escalar el front-end de una web, pero el equipo operativo puede ser responsable de garantizar que las normas y los límites de Auto Scaling sean apropiados para las necesidades de la carga de trabajo.

Recursos

Documentos relacionados:

- Características de Amazon CloudWatch
- ¿Qué es Amazon CloudWatch Events?

Vídeos relacionados:

- Diseñe un plan de monitoreo

OPS10-BP03 Prioridad de los eventos operativos según el impacto empresarial

Asegúrese de que, cuando varios eventos requieran una intervención, se aborden primero los más importantes para el negocio. Hay diversos tipos de impactos, como muertes o daños físicos, pérdidas económicas, así como daños a la reputación o confianza.

Antipatrones usuales:

- Recibe una solicitud de soporte para añadir una configuración de impresora para un usuario. Mientras trabaja en el problema, recibe una solicitud de soporte indicando que su sitio web de venta al por menor no funciona. Después de completar la configuración de la impresora para su usuario, comienza a trabajar en el problema del sitio web.
- Se le notifica que tanto su sitio web de venta al por menor como su sistema de nóminas no funcionan. No sabes cuál debe tener la máxima prioridad.

Beneficios de establecer esta práctica recomendada: La priorización de las respuestas a los incidentes con mayor impacto en la empresa permite gestionar dicho impacto.

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

Guía para la implementación

- Prioridad de los eventos operativos según el impacto empresarial: asegúrese de que, cuando varios eventos requieran una intervención, se aborden primero los más importantes para el negocio. Hay diversos tipos de impactos, como muertes o daños físicos, pérdidas económicas, infracciones de normas, así como daños a la reputación o confianza.

OPS10-BP04 Definir rutas de escalado

Defina las rutas de derivación en los runbooks y guías de estrategia, como, por ejemplo, aquello que desencadena una derivación y los procedimientos. Identifique a los titulares de cada acción de forma específica para garantizar respuestas rápidas y eficaces a los eventos operativos.

Identifique cuándo se requiere una decisión humana antes de realizar una acción. Trabaje con los responsables de la toma de decisiones para que esa decisión se tome con antelación y la acción se apruebe previamente, para que el tiempo medio de resolución no se prolongue esperando una respuesta.

Antipatrones usuales:

- Su sitio web de venta al por menor no funciona. No comprende el libro de instrucciones para recuperar el sitio. Empieza a llamar a sus colegas con la esperanza de que alguien pueda ayudarle.
- Recibe una incidencia de soporte para una aplicación inalcanzable. No tiene permisos para administrar el sistema. No sabe quién lo hace. Se intenta contactar con el propietario del sistema que abrió el incidente y no hay respuesta. No tiene contactos para el sistema y sus colegas no están familiarizados con él.

Beneficios de establecer esta práctica recomendada: Al definir los escalados, los desencadenantes y los procedimientos de los escalados, se permite la adición sistemática de recursos a un incidente a un ritmo adecuado para el impacto.

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

Guía para la implementación

- Definir rutas de escalado: defina las rutas de escalado en los runbooks y guías de estrategia, como, por ejemplo, aquello que desencadena un escalado y los procedimientos. Por ejemplo, la derivación de un problema de los ingenieros de apoyo a los ingenieros de apoyo senior cuando

los runbooks no tengan la respuesta a un problema o cuando haya transcurrido un periodo de tiempo definido previamente. Otro ejemplo sería la derivación de una carga de trabajo de los ingenieros de apoyo senior al equipo de desarrollo cuando las guías de estrategia no puedan identificar qué ruta seguir para solucionar el problema o cuando haya transcurrido un periodo de tiempo definido previamente. Identifique a los titulares de cada acción de forma específica para garantizar respuestas rápidas y eficaces a los eventos operativos. Las derivaciones pueden incluir a terceros. Por ejemplo, a un proveedor de conectividad de red o de software. Las derivaciones pueden incluir a los responsables de la toma de decisiones en lo que respecta a los sistemas afectados autorizados e identificados

OPS10-BP05 Definir un plan de comunicación con los clientes en caso de interrupciones del servicio

Defina y pruebe un plan de comunicación para interrupciones del sistema en el que pueda confiar, a fin de mantener informados a sus clientes y partes interesadas durante las interrupciones del servicio. Comuníquese directamente con sus usuarios tanto cuando los servicios que utilizan se vean afectados como cuando vuelvan a la normalidad.

Resultado deseado:

- Dispone de un plan de comunicación para situaciones que van desde el mantenimiento programado hasta grandes errores inesperados, incluida la invocación de planes de recuperación de desastres.
- En sus comunicaciones, proporciona información clara y transparente sobre los problemas de los sistemas para ayudar a los clientes a no dudar del rendimiento de sus sistemas.
- Utiliza mensajes de error y páginas de estado personalizados para reducir el pico de solicitudes al servicio de asistencia y mantener informados a los usuarios.
- El plan de comunicación se pone a prueba periódicamente para garantizar que funcione según lo previsto cuando se produzca una interrupción real.

Antipatrones usuales:

- Se produce una interrupción de la carga de trabajo, pero no dispone de un plan de comunicación. Los usuarios saturan el sistema de tickets de problemas con solicitudes porque no tienen información sobre la interrupción del servicio.
- Envía una notificación por correo electrónico a los usuarios durante una interrupción del servicio. No incluye un plazo para el restablecimiento del servicio, por lo que los usuarios no pueden hacer planes para la interrupción.

- Existe un plan de comunicación para las interrupciones del servicio, pero no se ha probado nunca. Se produce una interrupción y el plan de comunicación no funciona porque se ha omitido un paso crucial que podría haberse detectado en las pruebas.
- Durante una interrupción, envía una notificación a los usuarios que contiene demasiados detalles técnicos e información según su acuerdo de confidencialidad de AWS.

Beneficios de establecer esta práctica recomendada:

- Mantener la comunicación durante las interrupciones del servicio garantiza que los clientes estén al tanto de la evolución de los problemas y el tiempo estimado para su resolución.
- El desarrollo de un plan de comunicaciones bien definido asegura que sus clientes y usuarios finales estén bien informados para que puedan tomar las medidas adicionales necesarias para mitigar la repercusión de las interrupciones del servicio.
- Con una comunicación adecuada y un mejor conocimiento de las interrupciones planificadas y no planificadas, puede mejorar la satisfacción del cliente, limitar las reacciones imprevistas y favorecer la retención de clientes.
- Una comunicación oportuna y transparente sobre las interrupciones del sistema estimula la confianza necesaria para mantener las relaciones entre usted y sus clientes.
- Una estrategia de comunicación sólida durante una interrupción o crisis del servicio reduce las conjeturas y habladurías que podrían obstaculizar su capacidad de recuperación.

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

Guía para la implementación

Los planes de comunicación que mantienen informados a los clientes durante las interrupciones del servicio son holísticos y abarcan numerosas interfaces, como páginas de error orientadas al cliente, mensajes de error de API personalizados, banners de estado del sistema y páginas de comprobación de estado. Si su sistema tiene usuarios registrados, puede comunicarse a través de canales de mensajería como correo electrónico, SMS o notificaciones push para enviar mensajes con contenido personalizado a los clientes.

Herramientas de comunicación con el cliente

Como primera línea de defensa, las aplicaciones web y móviles deben proporcionar mensajes de error amables e informativos durante una interrupción del servicio, así como tener la capacidad de

redirigir el tráfico a una página de estado. Amazon CloudFront es una red de entrega de contenido (CDN) totalmente administrada que incluye capacidades para definir y presentar contenido de error personalizado. Las páginas de error personalizadas de CloudFront son una buena primera capa de mensajería para clientes en caso de interrupciones de servicio en el nivel de componente. CloudFront también contribuye a simplificar la administración y activación de una página de estado para interceptar todas las solicitudes durante interrupciones planificadas o no planificadas.

Los mensajes de error de API personalizados pueden ayudar a detectar y reducir el efecto cuando las interrupciones se limitan a servicios discretos. Amazon API Gateway le permite configurar respuestas personalizadas para sus API de REST. Esto le permite proporcionar mensajes claros y significativos a los consumidores de la API cuando API Gateway no es capaz de llegar a los servicios de backend. Los mensajes personalizados también sirven para apoyar el contenido de banners de interrupción y notificaciones cuando una característica concreta del sistema se degrada debido a interrupciones en el nivel de servicio.

La mensajería directa es el tipo más personalizado de mensajería para clientes. Amazon Pinpoint es un servicio administrado para comunicaciones multicanal escalables. Amazon Pinpoint le permite crear campañas que pueden transmitir mensajes ampliamente a toda su base de clientes afectada mediante SMS, correo electrónico, voz, notificaciones push o canales personalizados que usted defina. Cuando administra la mensajería con Amazon Pinpoint, las campañas de mensajes están bien definidas, se pueden probar y se pueden aplicar de forma inteligente a segmentos de clientes específicos. Una vez establecidas, las campañas se pueden programar o activar por eventos y se pueden probar con toda facilidad.

Ejemplo de cliente

Cuando la carga de trabajo se ve afectada, AnyCompany Retail envía una notificación por correo electrónico a sus usuarios. En el correo electrónico se describe qué funcionalidad empresarial se ha visto afectada y se proporciona una estimación realista de cuándo se restablecerá el servicio. Además, dispone de una página de estado que muestra información en tiempo real sobre el estado de su carga de trabajo. El plan de comunicación se prueba en un entorno de desarrollo dos veces al año para comprobar que sea eficaz.

Pasos para la implementación

1. Determine los canales de comunicación de su estrategia de mensajería. Considere los aspectos arquitectónicos de su aplicación y determine la mejor estrategia para hacer llegar los comentarios a los clientes. Esto podría incluir una o más de las estrategias de orientación descritas, como páginas de error y estado, respuestas de error de API personalizadas o mensajería directa.

2. Diseñe páginas de estado para su aplicación. Si ha determinado que las páginas de estado o de error personalizadas son adecuadas para sus clientes, tendrá que diseñar el contenido y los mensajes para esas páginas. Las páginas de error explican a los usuarios por qué no está disponible una aplicación, cuándo podría volver a estar disponible y qué pueden hacer mientras tanto. Si su aplicación utiliza Amazon CloudFront puede presentar respuestas de error personalizadas o utilizar Lambda at Edge para traducir errores y reescribir el contenido de la página. CloudFront también permite intercambiar destinos del contenido de su aplicación a un origen de contenido estático de Amazon S3 que contenga su página de mantenimiento o estado de interrupción del servicio.
3. Diseñe el conjunto correcto de estados de error de la API para su servicio. Los mensajes de error producidos por API Gateway cuando no es posible llegar a los servicios de backend, así como las excepciones de nivel de servicio, podrían no contener mensajes amables adecuados para mostrar a los usuarios finales. Sin tener que realizar cambios de código en sus servicios de backend, puede configurar respuestas de error personalizadas de API Gateway para asignar códigos de respuesta HTTP a mensajes de error de API seleccionados.
4. Diseñe la mensajería desde una perspectiva empresarial de modo que sea relevante para los usuarios finales de su sistema y no contenga detalles técnicos. Tenga en cuenta su audiencia y adapte los mensajes. Por ejemplo, dirija a los usuarios internos hacia una solución o un proceso manual que aproveche sistemas alternativos. En cuanto a los usuarios externos, puede pedirles que esperen hasta que se restablezca el sistema o que se suscriban a las actualizaciones para recibir una notificación una vez que se restablezca el sistema. Defina la mensajería aprobada para numerosas situaciones, como interrupciones inesperadas del servicio, mantenimiento planificado y errores parciales del sistema en los que una característica concreta podría degradarse o no estar disponible.
5. Cree plantillas y automatice la mensajería para clientes. Una vez que haya establecido el contenido del mensaje, puede utilizar Amazon Pinpoint u otras herramientas para automatizar la campaña de mensajería. Con Amazon Pinpoint puede crear segmentos de clientes objetivo para usuarios específicos afectados y transformar los mensajes en plantillas. Revise el tutorial de Amazon Pinpoint para entender cómo configurar una campaña de mensajería.
6. Evite vincular estrechamente las capacidades de mensajería a su sistema de orientación al cliente. Su estrategia de mensajería no debe depender de servicios ni almacenes de datos del sistema para verificar que puede enviar mensajes correctamente cuando se produzcan interrupciones del servicio. Plantéese la posibilidad de crear la capacidad de enviar mensajes desde más de una región o zona de disponibilidad para asegurar la disponibilidad de la mensajería. Si utiliza servicios de AWS para enviar mensajes, aproveche las operaciones del plano de datos en lugar de las operaciones del plano de control para invocar la mensajería.

Nivel de esfuerzo para el plan de implementación: alto. El desarrollo de un plan de comunicación —y los mecanismos para enviarlo— puede demandar un esfuerzo considerable.

Recursos

Prácticas recomendadas relacionadas:

- OPS07-BP03 Uso de runbooks para realizar los procedimientos - El plan de comunicación debe ir acompañado de un runbook para que el personal sepa cómo responder.
- OPS11-BP02 Realizar un análisis después del incidente - Tras una interrupción, lleve a cabo un análisis posterior al incidente para identificar mecanismos que eviten otra interrupción.

Documentos relacionados:

- Error Handling Patterns in Amazon API Gateway and AWS Lambda (Patrones de gestión de errores en Amazon API Gateway y AWS Lambda)
- Amazon API Gateway responses (Respuestas de Amazon API Gateway)

Ejemplos relacionados:

- AWS Health Dashboard
- Summary of the AWS Service Event in the Northern Virginia (US-EAST-1) Region (Resumen del evento de servicio de AWS en la región del norte de Virginia [Este de EE. UU. 1])

Servicios relacionados:

- AWS Support
- Contrato de usuario de AWS
- Amazon CloudFront
- Amazon API Gateway
- Amazon Pinpoint
- Amazon S3

OPS10-BP06 Comunicar el estado a través de paneles

Proporcione paneles adaptados a las audiencias de destino (por ejemplo, equipos técnicos internos, liderazgo y clientes) para comunicar el estado operativo actual del negocio y facilitar métricas de interés.

Puede crear paneles mediante Amazon CloudWatch Dashboards en las páginas de inicio personalizables en la consola de CloudWatch. Mediante servicios de inteligencia empresarial como Quick puede crear y publicar paneles interactivos de su carga de trabajo y estado operativo (por ejemplo, índices de pedidos, usuarios conectados y tiempos de transacción). Cree paneles que presenten vistas a nivel de sistema y de empresa de sus métricas.

Patrones de uso no recomendados comunes:

- Ejecuta, a petición, un informe sobre la utilización actual de su aplicación para la administración.
- Durante un incidente, cada veinte minutos se pone en contacto con usted un propietario del sistema preocupado por saber si ya está solucionado.

Beneficios de establecer esta práctica recomendada: Mediante la creación de paneles, posibilita el acceso de autoservicio a la información, lo que permite a sus clientes informarse por sí mismos y determinar si necesitan tomar medidas.

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

Guía para la implementación

- Comunique el estado mediante paneles: proporcione paneles adaptados a las audiencias de destino (por ejemplo, equipos técnicos internos, liderazgo y clientes) para comunicar el estado operativo actual del negocio y facilitar métricas de interés. Proporcionar una opción de autoservicio para facilitar información sobre el estado hace que haya menos interrupciones cuando se solicita esta información al equipo operativo. Algunos ejemplos son los paneles de control de Amazon CloudWatch y Panel de AWS Health.
  - Los paneles de CloudWatch crean y usan vistas de métricas personalizadas

Recursos

Documentos relacionados:

- Quick

- Los paneles de CloudWatch crean y usan vistas de métricas personalizadas

OPS10-BP07 Automatizar las respuestas a eventos

Automatice las respuestas a los eventos para reducir los errores causados por los procesos manuales y garantizar respuestas coherentes y rápidas.

Hay varias formas de automatizar las acciones del runbook y de la guía de estrategias en AWS. Para responder a un evento de un cambio de estado en sus recursos de AWS o de sus propios eventos personalizados, debe crear reglas de CloudWatch Events para desencadenar respuestas a través de destinos de CloudWatch (por ejemplo, funciones de Lambda, temas de Amazon Simple Notification Service (Amazon SNS), tareas de Amazon ECS y AWS Systems Manager Automation).

Para responder a una métrica que cruza un umbral para un recurso (por ejemplo, el tiempo de espera), debe crear alarmas de CloudWatch para realizar una o más acciones mediante acciones de Amazon EC2, acciones de Auto Scaling, o para enviar una notificación a un tema de Amazon SNS. Si necesita realizar acciones personalizadas en respuesta a una alarma, invoque a Lambda a través de una notificación de Amazon SNS. Use Amazon SNS para publicar notificaciones de eventos y mensajes de derivación a fin de mantener a las personas informadas.

AWS también admite sistemas de terceros a través de las API y los SDK del servicio de AWS. Hay una serie de herramientas de supervisión proporcionados por los socios de AWS y terceros que permiten la supervisión, las notificaciones y las respuestas. Algunas de estas herramientas incluyen New Relic, Splunk, Loggly, SumoLogic y Datadog.

Debe tener los procedimientos manuales importantes disponibles para usarlos cuando los procedimientos automatizados fallen.

Antipatrones usuales:

- Un desarrollador comprueba su código. Este evento podría haberse utilizado para iniciar una compilación y luego realizar pruebas, pero en su lugar no ocurre nada.
- La aplicación registra un error específico antes de dejar de funcionar. El procedimiento de reinicio de la aplicación se entiende bien y puede programarse. Podría utilizar el evento de registro para invocar un script y reiniciar la aplicación. En cambio, cuando el error se produce a las 3 de la madrugada de un domingo, le despiertan como recurso de guardia, que es responsable de reparar el sistema.

Beneficios de establecer esta práctica recomendada: Al utilizar respuestas automatizadas a los eventos, se reduce el tiempo de respuesta y se limita la introducción de errores por actividades manuales.

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

Guía para la implementación

- Automatizar las respuestas a eventos: automatice las respuestas a los eventos para reducir los errores causados por los procesos manuales y garantizar respuestas coherentes y rápidas.
  - ¿Qué es Amazon CloudWatch Events?
  - Creación de una regla de CloudWatch Events que se desencadena con un evento
  - Creación de una regla de CloudWatch Events que se desencadena en una llamada a la API de AWS con AWS CloudTrail
  - Ejemplos de eventos de CloudWatch Events de los servicios admitidos

Recursos

Documentos relacionados:

- Funciones de Amazon CloudWatch
- Ejemplos de eventos de CloudWatch Events de los servicios admitidos
- Creación de una regla de CloudWatch Events que se desencadena en una llamada a la API de AWS con AWS CloudTrail
- Creación de una regla de CloudWatch Events que se desencadena con un evento
- ¿Qué es Amazon CloudWatch Events?

Vídeos relacionados:

- Diseñar un plan de supervisión

Ejemplos relacionados:

Evolucionar

Pregunta

- OPERACIÓN 11. ¿Cómo desarrolla las operaciones?

OPERACIÓN 11. ¿Cómo desarrolla las operaciones?

Dedique tiempo y recursos a la mejora gradual casi continua, para desarrollar la eficacia y la eficiencia de sus operaciones.

Prácticas recomendadas

- OPS11-BP01 Tener un proceso de mejora continua
- OPS11-BP02 Realizar un análisis después del incidente
- OPS11-BP03 Implementar bucles de retroalimentación
- OPS11-BP04 Realizar la administración de conocimientos
- OPS11-BP05 Definir los elementos que impulsan la mejora
- OPS11-BP06 Validar las informaciones
- OPS11-BP07 Realizar revisiones de métricas de operaciones
- OPS11-BP08 Documentar y compartir las lecciones aprendidas
- OPS11-BP09 Asignar tiempo para realizar mejoras

OPS11-BP01 Tener un proceso de mejora continua

Evalúe su carga de trabajo con respecto a las prácticas recomendadas de arquitectura interna y externa. Realice revisiones de la carga de trabajo al menos una vez al año. Dé prioridad a las oportunidades de mejora en su cadencia de desarrollo de software.

Resultado deseado:

- Analiza su carga de trabajo con respecto a las prácticas recomendadas de arquitectura al menos una vez al año.
- Las oportunidades de mejora tienen la misma prioridad en su proceso de desarrollo de software.

Antipatrones usuales:

- No ha realizado una revisión de la arquitectura de su carga de trabajo desde que se desplegó hace varios años.
- Las oportunidades de mejora reciben una prioridad menor y permanecen en la lista de tareas pendientes.
- No existe un estándar para implementar las modificaciones de las prácticas recomendadas para la organización.

Beneficios de establecer esta práctica recomendada:

- Su carga de trabajo se mantiene actualizada en cuanto a las prácticas recomendadas de arquitectura.
- La evolución de su carga de trabajo se realiza de forma deliberada.
- Puede aprovechar las prácticas recomendadas de la organización para mejorar todas las cargas de trabajo.

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

Guía para la implementación

Al menos una vez al año, realice una revisión de la arquitectura de su carga de trabajo. Mediante prácticas recomendadas internas y externas, evalúe su carga de trabajo e identifique las oportunidades de mejora. Dé prioridad a las oportunidades de mejora en su cadencia de desarrollo de software.

Ejemplo de cliente

Se realiza un proceso anual de revisión de la arquitectura de todas las cargas de trabajo en AnyCompany Retail. Ha desarrollado su propia lista de prácticas recomendadas que se aplican a todas las cargas de trabajo. Gracias a la característica de enfoque personalizado de AWS Well-Architected Tool, realiza revisiones mediante la herramienta y su enfoque personalizado de prácticas recomendadas. Las oportunidades de mejora que se generan a partir de las revisiones se priorizan en sus sprints de software.

Pasos para la implementación

1. Realice revisiones periódicas de la arquitectura de su carga de trabajo de producción al menos una vez al año. Utilice un estándar de arquitectura documentado que incluya prácticas recomendadas específicas de AWS.
   a. Le recomendamos que, para estas revisiones, utilice sus propios estándares definidos internamente. Si no dispone de un estándar interno, le recomendamos que utilice el marco AWS Well-Architected Framework.
   b. Puede usar AWS Well-Architected Tool para crear un enfoque personalizado de sus prácticas recomendadas internas y llevar a cabo la revisión de la arquitectura.
   c. Los clientes pueden contactar con su arquitecto de soluciones de AWS para realizar una revisión guiada del marco Well-Architected Framework de su carga de trabajo.

2. Dé prioridad a las oportunidades de mejora identificadas durante la revisión en su proceso de desarrollo de software.

Nivel de esfuerzo para el plan de implementación: bajo. Puede utilizar el marco AWS Well-Architected Framework para llevar a cabo la revisión anual de la arquitectura.

Recursos

Prácticas recomendadas relacionadas:

- OPS11-BP02 Realizar un análisis después del incidente - El análisis posterior al incidente es otro generador de elementos de mejora. Incorpore las lecciones aprendidas a su lista interna de prácticas recomendadas de arquitectura.
- OPS11-BP08 Documentar y compartir las lecciones aprendidas - A medida que desarrolle sus propias prácticas recomendadas de arquitectura, compártalas en toda su organización.

Documentos relacionados:

- AWS Well-Architected Tool - Custom Lenses (AWS Well-Architected Tool: enfoques personalizados)
- AWS Well-Architected Whitepaper - The review process (Documento técnico de AWS Well-Architected: el proceso de revisión)
- Customize Well-Architected Reviews using Custom Lenses and the AWS Well-Architected Tool(Personalizar las revisiones de Well-Architected mediante enfoques personalizados y AWS Well-Architected Tool)
- Implementing the AWS Well-Architected Custom Lens lifecycle in your organization (Implementación del ciclo de vida del enfoque personalizado de AWS Well-Architected en su organización)

Vídeos relacionados:

- Well-Architected Labs - Level 100: Custom Lenses on AWS Well-Architected Tool(Laboratorios de Well-Architected - Nivel 100: enfoques personalizados en AWS Well-Architected Tool)

Ejemplos relacionados:

- AWS Well-Architected Tool

OPS11-BP02 Realizar un análisis después del incidente

Revise los eventos que afectan a los clientes e identifique los factores que contribuyen al evento y las medidas preventivas. Use esta información para desarrollar un plan de mitigación que limite o evite la reaparición del problema. Desarrolle procedimientos para proporcionar respuestas rápidas y eficaces. Comunique los factores que han contribuido al problema y las medidas correctivas según corresponda, adaptados al público de destino.

Patrones de uso no recomendados comunes:

- Administra un servidor de aplicaciones. Aproximadamente cada 23 horas y 55 minutos se terminan todas sus sesiones activas. Ha tratado de identificar lo que va mal en su servidor de aplicaciones. Sospecha que podría tratarse de un problema de red, pero no consigue la colaboración del equipo de red porque están demasiado ocupados para ayudarle. Carece de un proceso predefinido que seguir para obtener asistencia y recopilar la información necesaria para determinar lo que está sucediendo.
- Ha tenido pérdidas de datos dentro de su carga de trabajo. Es la primera vez que ocurre y la causa no es evidente. Decide que no es importante porque puede recrear los datos. Comienza a producirse con mayor frecuencia la pérdida de datos afectando a sus clientes. Esto también supone una carga operativa adicional al restaurar los datos perdidos.

Beneficios de establecer esta práctica recomendada: disponer de un proceso predefinido para determinar los componentes, las condiciones, las acciones y los eventos que han contribuido a un incidente le permite identificar las oportunidades de mejora.

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

Guía para la implementación

- Usar un proceso para determinar los factores que han contribuido al problema: revise todos los incidentes que afectan a los clientes Disponga de un proceso para identificar y documentar los factores que han contribuido al incidente, de manera que se puedan elaborar medidas de mitigación para limitar o prevenir su repetición y se puedan desarrollar procedimientos para dar respuestas rápidas y eficaces. Comunique la causa raíz como sea apropiado, según el público de destino.

OPS11-BP03 Implementar bucles de retroalimentación

Los bucles de retroalimentación proporcionan información procesable que impulsa la toma de decisiones. Cree bucles de retroalimentación en sus procedimientos y cargas de trabajo. Le servirán para identificar los problemas y las áreas que necesitan mejoras. También validan las inversiones realizadas en las mejoras. Estos bucles de retroalimentación son la base para mejorar continuamente la carga de trabajo.

Los bucles de retroalimentación se dividen en dos categorías: retroalimentación inmediata y análisis retrospectivo. La retroalimentación inmediata se obtiene mediante la revisión del rendimiento y los resultados de las actividades operativas. Esta retroalimentación procede de los miembros del equipo, de los clientes o del resultado automático de la actividad. Se recibe retroalimentación inmediata de aspectos como las pruebas A/B y el envío de nuevas características. Es esencial responder rápido a los errores.

El análisis retrospectivo se realiza periódicamente para obtener retroalimentación de la revisión de resultados operativos y de las métricas a lo largo del tiempo. Estas retrospectivas tienen lugar al final de un sprint, en una cadencia, o después de lanzamientos o eventos importantes. Este tipo de bucle de retroalimentación valida las inversiones en operaciones o la carga de trabajo. Le ayuda a medir el éxito y valida su estrategia.

Resultado deseado: utilice la retroalimentación inmediata y el análisis retrospectivo para impulsar las mejoras. Existe un mecanismo para obtener la retroalimentación de los usuarios y de los miembros del equipo. El análisis retrospectivo se utiliza para identificar las tendencias que impulsan las mejoras.

Patrones comunes de uso no recomendados:

- Lanza una nueva característica pero no tiene forma de recibir la retroalimentación de los clientes sobre ella.
- Después de invertir en mejoras operativas, no realiza una retrospectiva para validarlas.
- Recopila la retroalimentación de los clientes pero no la revisa con regularidad.
- Los bucles de retroalimentación dan lugar a propuestas de acción, pero no se incluyen en el proceso de desarrollo del software.
- Los clientes no reciben retroalimentación sobre las mejoras que han propuesto.

Beneficios de establecer esta práctica recomendada:

- Puede hacer un recorrido inverso desde el cliente para impulsar nuevas características.
- La cultura de su organización puede reaccionar más rápidamente a los cambios.
- Las tendencias se utilizan para identificar las oportunidades de mejora.
- Las retrospectivas validan las inversiones realizadas en la carga de trabajo y las operaciones.

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

Guía para la implementación

Implementar esta práctica recomendada implica utilizar tanto la retroalimentación inmediata como el análisis retrospectivo. Estos bucles de retroalimentación impulsan las mejoras. Existen muchos mecanismos para obtener retroalimentación inmediata, como encuestas, sondeos de opinión de los clientes o formularios de retroalimentación. Su organización también utiliza las retrospectivas para identificar las oportunidades de mejora y validar las iniciativas.

Ejemplo de cliente

AnyCompany Retail ha creado un formulario web en el que los clientes pueden dar retroalimentación o informar de sus problemas. Durante el examen semanal, el equipo de desarrollo de software evalúa la retroalimentación de los usuarios. La retroalimentación se utiliza periódicamente para dirigir la evolución de la plataforma. Realizan una retrospectiva al final de cada sprint para identificar los elementos que quieren mejorar.

Pasos para la aplicación

1. Retroalimentación inmediata
    - Necesita un mecanismo para recibir retroalimentación de los clientes y de los miembros del equipo. Las actividades de sus operaciones también se pueden configurar para ofrecer retroalimentación automática.
    - Su organización necesita un proceso para revisar esta retroalimentación, determinar qué hay que mejorar y programar la mejora.
    - La retroalimentación debe agregarse a su proceso de desarrollo de software.
    - A medida que vaya incorporando mejoras, haga un seguimiento del remitente de la retroalimentación.
        - Puede usar AWS Systems Manager OpsCenter para crear estas mejoras y realizar su seguimiento como OpsItems.

2. Análisis retrospectivo
    - Realice retrospectivas al final de un ciclo de desarrollo, con una cadencia determinada o después de un lanzamiento importante.
    - Convoque a las partes implicadas en la carga de trabajo para una reunión retrospectiva.
    - Cree tres columnas en una pizarra u hoja de cálculo: Parar, Iniciar y Mantener.
        - Parar corresponde a lo que quiera que su equipo deje de hacer.
        - Iniciar corresponde a las ideas que quiere empezar a hacer.
        - Mantener corresponde a lo que desea seguir haciendo.
    - Recorra la sala y recopile la retroalimentación de las partes interesadas.
    - Dé prioridad a la retroalimentación. Asigne acciones y partes interesadas a los elementos de los apartados Iniciar o Mantener.
    - Agregue las acciones a su proceso de desarrollo de software y comunique las actualizaciones de estado a las partes interesadas a medida que realiza las mejoras.

Nivel de esfuerzo para el plan de implementación: Medio. Para implementar esta práctica recomendada, necesita un método para recibir retroalimentación inmediata y analizarla. Además, debe establecer un proceso de análisis retrospectivo.

Recursos

Prácticas recomendadas relacionadas:

- OPS01-BP01 Evaluar las necesidades externas del cliente: los bucles de retroalimentación son un mecanismo para recopilar las necesidades de los clientes externos.
- OPS01-BP02 Evaluar las necesidades internas del cliente: las partes interesadas internas pueden utilizar los bucles de retroalimentación para comunicar las necesidades y los requisitos.
- OPS11-BP02 Realizar un análisis después del incidente: los análisis posteriores a los incidentes son una forma importante de análisis retrospectivo que se realiza después de los incidentes.
- OPS11-BP07 Realizar revisiones de métricas de operaciones: las revisiones de las métricas de las operaciones identifican tendencias y áreas de mejora.

Documentos relacionados:

- 7 Pitfalls to Avoid When Building a CCOE (7 obstáculos que evitar al crear un CCOE)
- Atlassian Team Playbook - Retrospectivas

- Email Definitions: Feedback Loops (Definiciones del correo electrónico: bucles de retroalimentación)
- Establishing Feedback Loops Based on the AWS Well-Architected Framework Review (Establecimiento de bucles de retroalimentacións basados en la revisión de AWS Well-Architected Framework)
- IBM Garage Methodology - Hold a retrospective (Metodología de IBM Garage: realizar una retrospectiva)
- Investopedia – The PDCS Cycle (Investopedia: el ciclo PDCS)
- Maximizing Developer Effectiveness (Maximizar la eficacia de los desarrolladores) por Tim Cochran
- Operations Readiness Reviews (ORR) Whitepaper - Iteration (Documento técnico de revisiones de preparación de operaciones [ORR]: iteración)
- TIL CSI - Continual Service Improvement (TIL CSI: mejora continua del servicio)
- When Toyota met e-commerce: Lean at Amazon (Cuando Toyota se encontró con el comercio electrónico: eficiencia en Amazon)

Vídeos relacionados:

- Building Effective Customer Feedback Loops (Crear bucles de retroalimentación eficaces de los clientes)

Ejemplos relacionados:

- Astuto: herramienta de código abierto para la retroalimentación de los clientes
- Soluciones de AWS: QnABot en AWS
- Fider: una plataforma para organizar la retroalimentación de los clientes

Servicios relacionados:

- AWS Systems Manager OpsCenter

OPS11-BP04 Realizar la administración de conocimientos

La administración del conocimiento ayuda a los miembros del equipo a encontrar la información necesaria para cumplir con su cometido. En las organizaciones basadas en el aprendizaje, la

información se comparte libremente, lo que capacita a los individuos. La información puede descubrirse o buscarse. La información es precisa y está actualizada. Existen mecanismos para crear nueva información, actualizar la existente y archivar la obsoleta. Los ejemplos más frecuentes de plataforma de administración del conocimiento es un sistema de administración de contenido, como una wiki.

Resultado deseado:

- Los miembros del equipo tienen acceso a información oportuna y precisa.
- Se puede buscar información.
- Existen mecanismos para añadir, actualizar y archivar la información.

Antipatrones usuales:

- No existe un almacenamiento centralizado de conocimientos. Los miembros del equipo administran sus propias notas en sus máquinas locales.
- Dispone de una wiki autoalojada, pero no de mecanismos para administrar la información, lo que provoca que esta quede obsoleta.
- Alguien identifica información que falta, pero no existe un proceso para solicitar que se añada a la wiki del equipo. La añaden ellos mismos, pero se saltan un paso clave, lo que provoca una interrupción del servicio.

Beneficios de establecer esta práctica recomendada:

- Los miembros del equipo tienen más poder porque la información se comparte libremente.
- Los nuevos miembros del equipo se incorporan más rápidamente porque la documentación está actualizada y es posible hacer búsquedas en ella.
- La información es oportuna, precisa y procesable.

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

Guía para la implementación

La administración del conocimiento es una faceta importante de las organizaciones de aprendizaje. Para empezar, necesita un repositorio central para almacenar los conocimientos (un ejemplo habitual es una wiki autoalojada). Debe desarrollar procesos para añadir, actualizar y archivar conocimientos. Desarrolle estándares sobre lo que debe documentarse y permita que todos contribuyan.

Ejemplo de cliente

AnyCompany Retail alberga una wiki interna donde se almacenan todos los conocimientos. Se anima a los miembros del equipo a añadir información a la base de conocimientos mientras realizan sus tareas cotidianas. Cada trimestre, un equipo interfuncional evalúa las páginas menos actualizadas y determina si deben archivarse o actualizarse.

Pasos para la implementación

1. Empiece por identificar el sistema de administración de contenido en el que se almacenarán los conocimientos. Consiga el acuerdo de las partes interesadas de toda la organización.
   a. Si no dispone de un sistema de administración de contenido, considere la posibilidad de crear una wiki autoalojada o utilizar un repositorio de control de versiones como punto de partida.
2. Elabore runbooks para añadir, actualizar y archivar información. Forme a su equipo sobre estos procesos.
3. Identifique qué conocimientos deben almacenarse en el sistema de administración de contenido. Empiece por las actividades diarias (runbooks y guías) que realizan los miembros del equipo. Colabore con las partes interesadas para priorizar qué conocimientos se añaden.
4. Trabaje periódicamente con las partes interesadas para identificar la información obsoleta y archivarla o actualizarla.

Nivel de esfuerzo para el plan de implementación: medio. Si no dispone de un sistema de administración de contenido, puede crear una wiki autoalojada o un repositorio de documentos controlado por versiones.

Recursos

Prácticas recomendadas relacionadas:

- OPS11-BP08 Documentar y compartir las lecciones aprendidas - La administración del conocimiento facilita el intercambio de información sobre las lecciones aprendidas.

Documentos relacionados:

- Atlassian - Knowledge Management (Atlassian: Administración de conocimientos)

Ejemplos relacionados:

- DokuWiki
- Gollum
- MediaWiki
- Wiki.js

OPS11-BP05 Definir los elementos que impulsan la mejora

Identifique los elementos que impulsan mejoras para evaluar y priorizar las oportunidades.

En AWS, puede agregar los registros de todas sus actividades de operaciones, cargas de trabajo e infraestructuras para crear un historial de actividades detallado. A continuación, podrá utilizar las herramientas de AWS para analizar sus operaciones y el estado de la carga de trabajo a lo largo del tiempo (por ejemplo, identificar tendencias, correlacionar eventos y actividades con los resultados y comparar y contrastar entre entornos y entre sistemas) para revelar las oportunidades de mejora según sus elementos de impulso.

Debe utilizar CloudTrail para realizar el seguimiento de la actividad de la API (a través de la Consola de administración de AWS, CLI, SDK y API) para saber lo que está sucediendo en sus cuentas. Realice el seguimiento de las actividades de despliegue de las herramientas para desarrolladores de AWS con CloudTrail y CloudWatch. Esto agregará un historial de actividad detallado de sus despliegues y sus resultados a sus datos de registro de CloudWatch Logs.

Exporte sus datos de registro a Amazon S3 para realizar el almacenamiento a largo plazo. Con AWS Glue, descubre y prepara sus datos de registro en Amazon S3 para el análisis. Uso Amazon Athena, mediante su integración nativa con AWS Glue, para analizar sus datos de registro. Use una herramienta de inteligencia empresarial como Quick para visualizar, explorar y analizar sus datos.

Patrones de uso no recomendados comunes:

- Tiene un script que funciona pero no es elegante. Invierte tiempo en volver a escribirlo. Ahora es una obra de arte.
- Su empresa emergente está tratando de conseguir otra financiación de un capitalista de riesgo. Quiere que demuestre el cumplimiento de la norma PCI DSS. Quiere que esté satisfecho, así que documenta su cumplimiento y se salta una fecha de entrega para un cliente, por lo que lo pierde. No ha sido algo malo, pero ahora se pregunta si ha sido lo correcto.

Beneficios de establecer esta práctica recomendada: Al determinar los criterios que desea utilizar para mejorar, puede minimizar el impacto de las motivaciones basadas en los acontecimientos o la inversión emocional.

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

Guía para la implementación

- Conozca los elementos que impulsan la mejora: solo debe hacer cambios en un sistema cuando producen un resultado deseado.
  - Capacidades deseadas: evalúe las características y capacidades deseadas al evaluar las oportunidades de mejora.
    - Novedades de AWS
  - Problemas inaceptables: evalúe los problemas, errores y vulnerabilidades inaceptables al evaluar las oportunidades de mejora.
    - Últimos boletines de seguridad de AWS
    - AWS Trusted Advisor
  - Requisitos de cumplimiento: evalúe las actualizaciones y los cambios necesarios para mantener el cumplimiento de la normativa, la política o la asistencia de terceros cuando revise las oportunidades de mejora.
    - Conformidad de AWS
    - Programas de conformidad de AWS
    - Noticias recientes de conformidad de AWS

Recursos

Documentos relacionados:

- Amazon Athena
- Quick
- Conformidad de AWS
- Noticias recientes de conformidad de AWS
- Programas de conformidad de AWS
- AWS Glue
- Últimos boletines de seguridad de AWS

- AWS Trusted Advisor
- Exporte sus datos de registro a Amazon S3
- Novedades de AWS

OPS11-BP06 Validar las informaciones

Revise los resultados de los análisis y las respuestas con equipos multifuncionales y con los propietarios de la empresa. Use estas revisiones para establecer un entendimiento común, identificar repercusiones adicionales y determinar cursos de acción. Ajuste las respuestas cuando corresponda.

Patrones de uso no recomendados comunes:

- Ve que la utilización de la CPU está en el 95 % en un sistema y considera una prioridad encontrar una manera de reducir la carga en el sistema. Determina que el mejor procedimiento es escalar verticalmente. El sistema es un transcodificador y el sistema se ha escalado para que funcione al 95 % de utilización de la CPU todo el tiempo. El propietario del sistema podría haberle explicado la situación si se hubiera puesto en contacto con él. Su tiempo se ha desperdiciado.
- El propietario de un sistema sostiene que su sistema es crucial para su negocio. El sistema no se colocó en un entorno de alta seguridad. Para mejorar la seguridad, se implementan los controles adicionales de detección y prevención que se requieren para los sistemas importantes. Notifica al propietario del sistema que el trabajo está terminado y que se le cobrará por los recursos adicionales. En la discusión que sigue a esta notificación, el propietario del sistema se entera de que hay una definición formal para los sistemas importantes que este sistema no cumple.

Beneficios de establecer esta práctica recomendada: al validar las ideas con los propietarios de los negocios y los expertos en la materia, puede establecer un entendimiento común y orientar las mejoras de forma más eficaz.

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

Guía para la implementación

- Validar informaciones: colabore con los propietarios de las empresas y los expertos en la materia para asegurarse de que existe un entendimiento común y un acuerdo sobre el significado de los datos que ha recopilado. Identifique las preocupaciones adicionales, los impactos potenciales y determine un procedimiento.

OPS11-BP07 Realizar revisiones de métricas de operaciones

Realice análisis retrospectivos periódicos de las métricas de operaciones con participantes de diferentes equipos y áreas de la empresa. Use estas revisiones para identificar las oportunidades de mejora, los posibles cursos de acción y para compartir las lecciones aprendidas.

Busque oportunidades para mejorar en todos sus entornos (por ejemplo, desarrollo, pruebas y producción).

Patrones de uso no recomendados comunes:

- Hubo una importante promoción al por menor que se interrumpió por su periodo de mantenimiento. La empresa no es consciente de que existe un periodo de mantenimiento estándar que podría retrasarse si se producen otros eventos que afecten a la empresa.
- Ha sufrido una interrupción prolongada debido al uso de una biblioteca con errores que se utiliza habitualmente en su organización. Desde entonces, ha migrado a una biblioteca fiable. Los demás equipos de su organización no saben que están en peligro. Si se reunieran periódicamente y revisaran este incidente, serían conscientes del riesgo.
- El rendimiento de su transcodificador ha ido cayendo de forma constante y ha afectado al equipo de medios. Todavía no es terrible. No tendrá la oportunidad de averiguarlo hasta que no sea lo suficientemente grave como para provocar un incidente. Si revisara las métricas de sus operaciones con el equipo de medios, habría una oportunidad para que se reconociera el cambio en las métricas y su experiencia y se solucionara el problema.
- No está revisando su satisfacción de los SLA de los clientes. Tiende a no cumplir los SLA de sus clientes. Existen sanciones económicas relacionadas con el incumplimiento de los SLA de sus clientes. Si se reunieran periódicamente para revisar las métricas de estos SLA, tendrían la oportunidad de reconocer y abordar el problema.

Beneficios de establecer esta práctica recomendada: Al reunirse periódicamente para revisar las métricas de las operaciones, los eventos y los incidentes, se mantiene un entendimiento común entre los equipos, se comparten las lecciones aprendidas y se pueden priorizar y orientar las mejoras.

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

Guía para la implementación

- Revisiones de las métricas de las operaciones: realice análisis retrospectivos periódicos de las métricas de operaciones con participantes de diferentes equipos y áreas de la empresa. Involucre

a las partes interesadas, incluidos los equipos de negocio, desarrollo y operaciones, para confirmar los hallazgos obtenidos de la retroalimentación inmediata y el análisis retrospectivo, y para compartir las lecciones aprendidas. Use sus ideas para identificar las oportunidades de mejora y los posibles cursos de acción.

- Amazon CloudWatch
- Uso de métricas de Amazon CloudWatch
- Publique métricas personalizadas
- Referencia de métricas y dimensiones de Amazon CloudWatch

Recursos

Documentos relacionados:

- Amazon CloudWatch
- Referencia de métricas y dimensiones de Amazon CloudWatch
- Publique métricas personalizadas
- Uso de métricas de Amazon CloudWatch

OPS11-BP08 Documentar y compartir las lecciones aprendidas

Documente y comparta las lecciones aprendidas de las actividades de operaciones para poder aplicarlas internamente y entre los equipos.

Deber compartir lo que sus equipos aprenden para aumentar el beneficio en toda su organización. Es necesario compartir información y recursos para prevenir errores evitables y facilitar los esfuerzos de desarrollo. Esto le permitirá concentrarse en la entrega de las características deseadas.

Utilice AWS Identity and Access Management (IAM) para definir los permisos que proporcionan el acceso controlado a los recursos que desea compartir en las cuentas y a través de ellas. A continuación, debe utilizar los repositorios de AWS CodeCommit, controlados por versión, para compartir bibliotecas de aplicaciones, procedimientos en scripts, documentación de procedimientos y otra documentación del sistema. Comparta sus estándares de computación al dar acceso a sus AMI y autorizar el uso de sus funciones Lambda a través de las cuentas. También puede compartir sus estándares de infraestructura como plantillas de AWS CloudFormation.

A través de las API y los SDK de AWS, puede integrar herramientas y repositorios externos y de terceros (por ejemplo, GitHub, BitBucket y SourceForge). Cuando comparta lo que ha aprendido

y desarrollado, tenga cuidado de estructurar los permisos para asegurar la integridad de los repositorios compartidos.

Patrones de uso no recomendados comunes:

- Ha sufrido una interrupción prolongada debido al uso de una biblioteca con errores que se utiliza habitualmente en su organización. Desde entonces, ha migrado a una biblioteca fiable. Los demás equipos de su organización no saben que están en peligro. Si documentara y compartiera su experiencia con esta biblioteca, serían conscientes del riesgo.
- Ha identificado un caso límite en un microservicio compartido internamente que provoca la caída de las sesiones. Ha actualizado sus llamadas al servicio para evitar este caso límite. Los demás equipos de su organización no saben que están en peligro. Si documentara y compartiera su experiencia con esta biblioteca, serían conscientes del riesgo.
- Ha encontrado una forma de reducir significativamente los requisitos de utilización de la CPU para uno de sus microservicios. No sabe si otros equipos podrían aprovechar esta técnica. Si documentara y compartiera su experiencia con esta biblioteca, tendrían la oportunidad de hacerlo.

Beneficios de establecer esta práctica recomendada: Comparta las lecciones aprendidas para respaldar las mejoras y maximizar los beneficios de la experiencia.

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

Guía para la implementación

- Documente y comparta las lecciones aprendidas: disponga de procedimientos para documentar las lecciones aprendidas durante la ejecución de las actividades de operaciones y el análisis retrospectivo para que puedan servir a otros equipos.
  - Comparta aprendizajes: disponga de procedimientos para compartir las lecciones aprendidas y los artefactos asociados entre equipos. Por ejemplo, comparta los procedimientos actualizados, la orientación, la gobernanza y las mejores prácticas a través de un wiki accesible. Comparta los scripts, el código y las bibliotecas por medio de un repositorio común.
    - Delegating access to your AWS environment (Delegación del acceso a su entorno de AWS)
    - Share an AWS CodeCommit repository (Compartir un repositorio de AWS CodeCommit)
    - Easy authorization of AWS Lambda functions (Autorización fácil de las funciones AWS Lambda)
    - Compartir una AMI con cuentas de AWS específicas

- Speed template sharing with an AWS CloudFormation designer URL (Acelerar el intercambio de plantillas con un URL de diseñador de AWS CloudFormation)
- Uso de AWS Lambda con Amazon SNS

Recursos

Documentos relacionados:

- Easy authorization of AWS Lambda functions (Autorización fácil de las funciones AWS Lambda)
- Share an AWS CodeCommit repository (Compartir un repositorio de AWS CodeCommit)
- Compartir una AMI con cuentas de AWS específicas
- Speed template sharing with an AWS CloudFormation designer URL (Acelerar el intercambio de plantillas con un URL de diseñador de AWS CloudFormation)
- Uso de AWS Lambda con Amazon SNS

Vídeos relacionados:

- Delegating access to your AWS environment (Delegación del acceso a su entorno de AWS)

OPS11-BP09 Asignar tiempo para realizar mejoras

Dedique tiempo y recursos de sus procesos para hacer posibles las mejoras incrementales continuas.

En AWS, se pueden crear duplicados temporales de los entornos, lo que reduce el riesgo, el esfuerzo y el coste de la experimentación y las pruebas. Estos ambientes duplicados pueden usarse para probar las conclusiones de su análisis, experimentar, así como para desarrollar y probar las mejoras planeadas.

Patrones de uso no recomendados comunes:

- Hay un problema de rendimiento conocido en su servidor de aplicaciones. Se añade a las tareas pendientes existentes detrás de cada implementación de funciones programadas. Si el ritmo de adición de funciones previstas se mantiene constante, el problema del rendimiento nunca se solucionará.
- Para respaldar la mejora continua, aprueba que los administradores y desarrolladores utilicen todo su tiempo extra para seleccionar y aplicar las mejoras. Nunca se completan las mejoras.

Beneficios de establecer esta práctica recomendada: al dedicar tiempo y recursos en sus procesos hace posibles las mejoras incrementales continuas.

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

Guía para la implementación

- Asignar tiempo para realizar mejoras: dedique tiempo y recursos dentro de sus procesos para hacer posibles las mejoras incrementales continuas. Aplique cambios para mejorar y evalúe los resultados para determinar el éxito. Si los resultados no alcanzan los objetivos y la mejora sigue siendo una prioridad, busque cursos de acción alternativos.

Seguridad

El pilar de seguridad engloba la capacidad de proteger datos, sistemas y activos para sacar partido de las tecnologías de nube con el fin de mejorar su nivel de seguridad. Encontrará recomendaciones de implementación en el documento técnico Pilar de seguridad.

Áreas de prácticas recomendadas

- Aspectos básicos de seguridad
- Identity and Access Management
- Detección
- Protección de la infraestructura
- Protección de los datos
- Respuesta ante incidentes
- Seguridad de las aplicaciones

Aspectos básicos de seguridad

Pregunta

- SEGURIDAD 1. ¿Cómo utiliza la carga de trabajo de forma segura?

SEGURIDAD 1. ¿Cómo utiliza la carga de trabajo de forma segura?

Para utilizar la carga de trabajo de forma segura, debe adoptar prácticas recomendadas globales en cada área de seguridad. Tome los requisitos y los procesos que ha definido en la excelencia

operativa a nivel de organización y de carga de trabajo, y aplíquelos a todas las áreas. Mantenerse actualizado con AWS, las prácticas recomendadas del sector y la inteligencia de amenazas le ayudan a desarrollar el modelo de amenaza y los objetivos de control. La automatización de los procesos de seguridad, las pruebas y la validación le permite escalar las operaciones de seguridad.

Prácticas recomendadas

- SEC01-BP01 Separar cargas de trabajo utilizando cuentas
- SEC01-BP02 Proteger el usuario raíz y las propiedades de la cuenta
- SEC01-BP03 Identificar y validar objetivos de control
- SEC01-BP04 Mantenerse al día de las amenazas de seguridad
- SEC01-BP05 Mantenerse al día con las recomendaciones de seguridad
- SEC01-BP06 Automatizar la comprobación y validación de controles de seguridad en canalizaciones
- SEC01-BP07 Identificar amenazas y priorizar mitigaciones con un modelo de amenazas
- SEC01-BP08 Evaluar e implementar nuevos servicios y características de seguridad de forma periódica

SEC01-BP01 Separar cargas de trabajo utilizando cuentas

Establezca barreras de protección y medidas de aislamiento comunes entre los entornos (por ejemplo, producción, desarrollo y pruebas) y las cargas de trabajo mediante una estrategia de varias cuentas. Es muy recomendable que la separación se realice a nivel de cuenta, ya que así se consigue una barrera de aislamiento sólida para gestionar la seguridad, la facturación y el acceso.

Resultado deseado: una estructura de cuentas que aísle las operaciones en la nube, las cargas de trabajo no relacionadas y los entornos en cuentas separadas para aumentar la seguridad en toda la infraestructura en la nube.

Antipatrones usuales:

- Colocar en la misma cuenta varias cargas de trabajo no relacionadas con diferentes niveles de confidencialidad de los datos.
- Estructura de la unidad organizativa (OU) mal definida.

Beneficios de establecer esta práctica recomendada:

- Menor alcance del impacto si se accede inadvertidamente a una carga de trabajo
- Gobernanza central del acceso a los servicios, recursos y regiones de AWS.
- Mantenimiento de la seguridad de la infraestructura en la nube con políticas y una administración centralizada de los servicios de seguridad
- Proceso automatizado de creación y mantenimiento de las cuentas
- Auditoría centralizada de la infraestructura para los requisitos de conformidad y reglamentarios

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

Guía para la implementación

Las Cuentas de AWS proporcionan una barrera de aislamiento de seguridad entre cargas de trabajo o recursos que operan con distintos niveles de confidencialidad. AWS ofrece herramientas para administrar sus cargas de trabajo en la nube a escala mediante una estrategia de varias cuentas para aprovechar esta barrera de aislamiento. Para obtener orientación sobre los conceptos, los patrones y la implementación de una estrategia de varias cuentas en AWS, consulte Organizing Your AWS Environment Using Multiple Accounts (Organización del entorno de AWS utilizando varias cuentas).

Cuando tenga varias Cuentas de AWS con una administración central, sus cuentas deben organizarse en una jerarquía definida por capas de unidades organizativas (OU). Luego, pueden organizarse y aplicarse controles de seguridad a las OU y a las cuentas miembro mediante el establecimiento de controles preventivos uniformes en las cuentas miembros de la organización. Los controles de seguridad se heredan, lo que permite filtrar los permisos disponibles para las cuentas miembros situadas en niveles inferiores de una jerarquía de OU. Un buen diseño aprovecha esta herencia para reducir el número y la complejidad de las políticas de seguridad necesarias para lograr los controles de seguridad deseados para cada cuenta miembro.

AWS Organizations y AWS Control Tower son dos de los servicios que puede utilizar para implementar y administrar esta estructura de varias cuentas en su entorno de AWS. AWS Organizations le permite organizar las cuentas en una jerarquía definida por una o varias capas de OU, en la que cada OU contiene una serie de cuentas miembro. Las políticas de control de servicios (SCP) permiten al administrador de la organización establecer controles preventivos detallados en las cuentas miembros y AWS Config puede utilizarse para establecer controles proactivos y de detección en las cuentas miembro. Muchos servicios de AWS se integran con AWS Organizations para proporcionar controles administrativos delegados y realizar tareas específicas del servicio en todas las cuentas miembros de la organización.

Por encima de AWS Organizations, AWS Control Tower proporciona una configuración recomendada de un solo clic para un entorno de AWS de varias cuentas con una zona de aterrizaje. La zona de aterrizaje es el punto de entrada al entorno de varias cuentas que se establece por medio de Control Tower. Control Tower tiene varias ventajas con respecto a AWS Organizations. Estas son tres ventajas que mejoran la gobernanza de las cuentas:

- Barreras de protección de seguridad obligatorias integradas que se aplican automáticamente a las cuentas que se admiten en la organización.
- Barreras de protección opcionales que pueden activarse o desactivarse para un conjunto determinado de OU.
- AWS Control Tower Account Factory proporciona un despliegue automatizado de cuentas que contienen bases de referencia y opciones de configuración preaprobadas dentro de su organización.

Pasos para la implementación

1. Diseñe una estructura de unidades organizativas: si la estructura de unidades organizativas está diseñada correctamente, se reduce la carga administrativa necesaria para crear y mantener las políticas de control de los servicios y otros controles de seguridad. La estructura de su unidad organizativa debe ajustarse a sus necesidades empresariales, la confidencialidad de los datos y la estructura de la carga de trabajo.
2. Cree una zona de aterrizaje para su entorno de varias cuentas: una zona de aterrizaje proporciona una base de seguridad e infraestructura uniforme desde la que su organización puede desarrollar, iniciar y desplegar cargas de trabajo rápidamente. Puede utilizar una zona de aterrizaje personalizada o AWS Control Tower para organizar su entorno.
3. Establezca barreras de protección: implemente barreras de protección de seguridad uniformes para su entorno en toda su zona de aterrizaje. AWS Control Tower proporciona una lista de controles obligatorios y opcionales que pueden desplegarse. Los controles obligatorios se despliegan automáticamente al implementar Control Tower. Revise la lista de los controles más recomendables y opcionales, e implemente los controles que sean adecuados a sus necesidades.
4. Restrinja el acceso a las regiones añadidas recientemente: para las nuevas Regiones de AWS, los recursos de IAM, como los usuarios y los roles, solo se propagan a las regiones que especifique. Esta acción puede realizarse a través de la consola cuando se utiliza Control Tower o ajustando las políticas de permisos de IAM en AWS Organizations.

5. Considere la posibilidad de usar AWS CloudFormation StackSets: StackSets le ayuda a desplegar recursos como políticas, roles y grupos de IAM en diferentes regiones y Cuentas de AWS a partir de una plantilla aprobada.

Recursos

Prácticas recomendadas relacionadas:

- SEC02-BP04 Recurrir a un proveedor de identidades centralizado

Documentos relacionados:

- AWS Control Tower
- AWS Security Audit Guidelines (Directivas de auditoría de seguridad de AWS)
- Prácticas recomendadas de seguridad en IAM
- Use CloudFormation StackSets to provision resources across multiple Cuentas de AWS and regions (Utilice CloudFormation StackSets para aprovisionar recursos en varias cuentas y regiones de AWS)
- Preguntas frecuentes de AWS Organizations
- Terminología y conceptos de AWS Organizations
- Best Practices for AWS Organizations Service Control Policies in a Multi-Account Environment (Prácticas recomendadas para las políticas de control de servicios de AWS Organizations en un entorno de varias cuentas)
- AWS Account Management Reference Guide (Guía de referencia para la administración de cuentas de AWS)
- Organización de su entorno de AWS mediante varias cuentas

Vídeos relacionados:

- Enable AWS adoption at scale with automation and governance (Facilitar la adopción de AWS a escala con la automatización y la gobernanza)
- Security Best Practices the Well-Architected Way (Prácticas recomendadas de seguridad al estilo de Well-Architected)
- Building and Governing Multiple Accounts using AWS Control Tower (Creación y administración de varias cuentas mediante Control Tower)

- Enable Control Tower for Existing Organizations (Habilitar Control Tower para las organizaciones existentes)

Talleres relacionados:

- Control Tower Immersion Day (Día de inmersión en Control Tower)

SEC01-BP02 Proteger el usuario raíz y las propiedades de la cuenta

El usuario raíz es el usuario con más privilegios de una Cuenta de AWS. Tiene acceso administrativo completo a todos los recursos de la cuenta y, en algunos casos, no se puede limitar con políticas de seguridad. Deshabilitar el acceso programático al usuario raíz, establecer controles apropiados para este usuario y evitar su uso rutinario ayuda a reducir el riesgo de exposición inadvertida de las credenciales raíz y el consiguiente peligro que esto supone para el entorno de la nube.

Resultado deseado: proteger al usuario raíz ayuda a reducir la posibilidad de que se produzcan daños accidentales o intencionados por el uso indebido de las credenciales de usuario raíz. Establecer controles de detección también puede servir para alertar al personal adecuado cuando se realizan acciones con el usuario raíz.

Antipatrones usuales:

- Utilizar el usuario raíz para realizar tareas que no se encuentran entre las pocas que requieren credenciales de usuario raíz.
- Dejar de comprobar periódicamente los planes de contingencia para verificar el funcionamiento de las infraestructuras críticas, los procesos y el personal durante una emergencia.
- Considerar únicamente el flujo de inicio de sesión típico de la cuenta y olvidarse de considerar o probar métodos alternativos de recuperación de la cuenta.
- No ocuparse de DNS, servidores de correo electrónico y proveedores de telefonía como parte del perímetro crítico de seguridad, ya que estos se utilizan en el flujo de recuperación de la cuenta.

Beneficios de establecer esta práctica recomendada: proteger el acceso al usuario raíz genera confianza, ya que las acciones en su cuenta están controladas y auditadas.

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

Guía para la implementación

AWS dispone de muchas herramientas para proteger su cuenta. Sin embargo, dado que algunas de estas medidas no están habilitadas de forma predeterminada, deberá implementarlas directamente. Considere estas recomendaciones como pasos básicos para proteger su Cuenta de AWS. A medida que vaya implementando estos pasos, es importante que cree un proceso para evaluar y supervisar continuamente los controles de seguridad.

Cuando crea una Cuenta de AWS por primera vez, empieza con una sola identidad que tiene acceso completo a todos los servicios y recursos de AWS de la cuenta. Esta identidad se denomina usuario raíz de la Cuenta de AWS. Puede iniciar sesión como usuario raíz con la dirección de correo electrónico y la contraseña que se ha usado para crear la cuenta. Dado el elevado nivel de acceso que se concede al usuario raíz de AWS, debe limitar el uso de este usuario de AWS a la realización de tareas que lo requieran específicamente. Las credenciales de inicio de sesión del usuario raíz deben estar muy bien protegidas y siempre se debe habilitar la autenticación multifactor (MFA) para el usuario raíz de la Cuenta de AWS.

Además del flujo de autenticación normal para iniciar sesión con su usuario raíz mediante un nombre de usuario, una contraseña y un dispositivo de autenticación multifactor (MFA), existen flujos de recuperación de la cuenta para iniciar sesión con el usuario raíz de la Cuenta de AWS que tiene acceso a la dirección de correo electrónico y al número de teléfono asociados a su cuenta. Por lo tanto, también es muy importante proteger la cuenta de correo electrónico del usuario raíz a la que se envía el mensaje de recuperación y el número de teléfono asociado a la cuenta. Tenga en cuenta también las posibles dependencias circulares si la dirección de correo electrónico asociada al usuario raíz está alojada en servidores de correo electrónico o recursos del servicio de nombres de dominio (DNS) de la misma Cuenta de AWS.

Cuando se utiliza AWS Organizations, hay varias Cuentas de AWS y cada una de ellas tiene un usuario raíz. Se designa una cuenta como cuenta de administración y, a continuación, se pueden añadir varias capas de cuentas miembro por debajo de esa cuenta de administración. Dé prioridad a la seguridad del usuario raíz de su cuenta de administración y, luego, céntrese en los usuarios raíz de las cuentas miembros. La estrategia para proteger el usuario raíz de su cuenta de administración puede ser diferente a la de los usuarios raíz de sus cuentas miembro, y puede colocar controles de seguridad preventivos en los usuarios raíz de sus cuentas miembro.

Pasos para la implementación

Se recomienda realizar los siguientes pasos de implementación para establecer controles para el usuario raíz. Cuando sea oportuno, las recomendaciones hacen referencia a la referencia de CIS

AWS Foundations versión 1.4.0. Además de estos pasos, consulte las prácticas recomendadas de AWS para proteger sus recursos y su Cuenta de AWS.

Controles preventivos

1. Establezca información de contacto precisa para la cuenta.
    a. Esta información se utiliza para el flujo de recuperación de las contraseñas perdidas, el flujo de recuperación de cuentas de los dispositivos MFA perdidos y para comunicaciones críticas relacionadas con la seguridad con su equipo.
    b. Utilice una dirección de correo electrónico alojada en su dominio corporativo (preferiblemente una lista de distribución) como dirección de correo electrónico del usuario raíz. Al utilizar una lista de distribución en lugar de la cuenta de correo electrónico de una persona, se consigue redundancia y continuidad adicionales para acceder a la cuenta raíz durante largos periodos de tiempo.
    c. El número de teléfono que figure en la información de contacto debe ser un teléfono dedicado y seguro para este fin. El número de teléfono no debe figurar en ninguna parte ni compartirse con nadie.
2. No cree claves de acceso para el usuario raíz. Si existen claves de acceso, elimínelas (CIS 1.4).
    a. Elimine cualquier credencial programática de larga duración (claves de acceso y secretas) para el usuario raíz.
    b. Si ya existen claves de acceso para el usuario raíz, debe hacer que los procesos que utilizan dichas claves pasen a utilizar claves de acceso temporales de un rol de AWS Identity and Access Management (IAM) y, a continuación, eliminar las claves de acceso del usuario raíz.
3. Determine si necesita almacenar las credenciales del usuario raíz.
    a. Si utiliza AWS Organizations para crear nuevas cuentas miembro, la contraseña inicial del usuario raíz de esas nuevas cuentas miembro se establece en un valor aleatorio que no está expuesto a usted. Considere la posibilidad de utilizar el flujo de restablecimiento de las contraseñas desde su cuenta de administración de AWS Organization para obtener acceso a la cuenta miembro si es necesario.
    b. Para Cuentas de AWS independientes o la cuenta de administración de AWS, considere la posibilidad de crear y almacenar de forma segura credenciales para el usuario raíz. Habilite MFA para el usuario raíz.
4. Habilite controles preventivos para usuarios raíz de cuentas miembro en entornos de varias cuentas de AWS.

a. Considere la posibilidad de habilitar la barrera de protección preventiva No permitir la creación de claves de acceso para el usuario raíz para las cuentas miembros.

b. Considere la posibilidad de habilitar la barrera de protección preventiva No permitir acciones como usuario raíz para las cuentas miembros.

5. Si necesita credenciales para el usuario raíz:

a. Utilice una contraseña compleja.

b. Habilite la autenticación multifactor (MFA) para el usuario raíz, especialmente para las cuentas de administración de AWS Organizations (pagador) (CIS 1.5).

c. Considere la posibilidad de usar dispositivos MFA físicos para mejorar la resiliencia y la seguridad, ya que los dispositivos de un solo uso pueden reducir las posibilidades de que los dispositivos que contienen los códigos MFA puedan reutilizarse para otros fines. Verifique que los dispositivos MFA físicos que funcionan con baterías se sustituyan con regularidad. (CIS 1.6)

- Para configurar MFA para el usuario raíz, siga las instrucciones para habilitar un dispositivo MFA virtual o un dispositivo MFA físico.

d. Considere la posibilidad de inscribir varios dispositivos MFA de reserva. Se permiten hasta 8 dispositivos MFA por cuenta.

- Tenga en cuenta que, si inscribe más de un dispositivo MFA para el usuario raíz, se desactiva automáticamente el flujo para recuperar su cuenta si el dispositivo MFA se pierde.

e. Guarde la contraseña con todas las medidas de seguridad y tenga en cuenta las dependencias circulares si la guarda electrónicamente. No guarde la contraseña de forma que sea necesario acceder a la misma Cuenta de AWS para obtenerla.

6. Opcional: considere la posibilidad de establecer un programa de rotación periódica de contraseñas para el usuario raíz.

- Las prácticas recomendadas de administración de credenciales dependen de los requisitos de las normativas y políticas que tenga. Los usuarios raíz protegidos por MFA no dependen de una contraseña como único factor de autenticación.
- Cambiar la contraseña del usuario raíz de forma periódica reduce el riesgo de que se utilice de forma indebida si se ha expuesto de forma inadvertida.

Controles de detección

- Cree alarmas para detectar el uso de las credenciales del usuario raíz (CIS 1.7). Si se habilita Amazon GuardDuty, este supervisará el uso de credenciales de API del usuario raíz y alertará de ese uso mediante el hallazgo de RootCredentialUsage.

- Evalúe e implemente los controles de detección que se incluyen en el paquete de conformidad del pilar de seguridad de AWS Well-Architected para AWS Config o, si utiliza AWS Control Tower, los controles más recomendados que hay disponibles en Control Tower.

Guía operativa

- Determine qué persona de la organización debe tener acceso a las credenciales del usuario raíz.
  - Utilice la regla de dos personas para no haya una sola persona que tenga acceso a todas las credenciales y el dispositivo MFA necesarios para obtener acceso de usuario raíz.
  - Compruebe que sea la organización, y no un único individuo, quien mantenga un control del número de teléfono y el alias de correo electrónico asociados a la cuenta (que se utilizan para el restablecimiento de la contraseña y el flujo de restablecimiento de MFA).
- Utilice el usuario raíz únicamente de forma excepcional (CIS 1.7).
  - El usuario raíz de AWS no debe utilizarse para las tareas diarias, ni siquiera para las tareas administrativas. Inicie sesión únicamente como usuario raíz para realizar las tareas de AWS que requieran dicho usuario. Todas las demás acciones deben realizarlas otros usuarios que asuman los roles apropiados.
- Compruebe periódicamente que el acceso al usuario raíz funciona para poder probar los procedimientos antes de que se produzca una situación de emergencia que requiera el uso de las credenciales del usuario raíz.
- Compruebe periódicamente que la dirección de correo electrónico asociada a la cuenta y las que figuran en los contactos alternativos funcionan. Supervise las bandejas de entrada de estas direcciones de correo electrónico para comprobar si se reciben notificaciones de seguridad de `<abuse@amazon.com>`. Asegúrese también de que los números de teléfono asociados a la cuenta funcionan.
- Prepare procedimientos de respuesta a incidentes para responder al uso indebido de la cuenta raíz. Consulte la AWS Security Incident Response Guide (Guía de respuesta a incidentes de seguridad de AWS) y las prácticas recomendadas de la sección Incident Response (Respuesta a incidentes) del documento técnico sobre los pilares de seguridad para obtener más información sobre la creación de una estrategia de respuesta a incidentes para su Cuenta de AWS.

Recursos

Prácticas recomendadas relacionadas:

- SEC01-BP01 Separar cargas de trabajo utilizando cuentas

- SEC02-BP01 Usar mecanismos de inicio de sesión sólidos
- SEC03-BP02 Conceder acceso con privilegios mínimos
- SEC03-BP03 Establecer un proceso de acceso de emergencia
- SEC10-BP05: Aprovisionamiento previo del acceso

Documentos relacionados:

- AWS Control Tower
- AWS Security Audit Guidelines (Directivas de auditoría de seguridad de AWS)
- Prácticas recomendadas de seguridad en IAM
- Amazon GuardDuty: alerta sobre el uso de credenciales raíz
- Step-by-step guidance on monitoring for root credential use through CloudTrail (Guía paso a paso para supervisar el uso de credenciales raíz a través de Control Tower)
- MFA tokens approved for use with AWS (Tokens MFA aprobados para su uso con AWS)
- Implementación del acceso con rotura de cristales en AWS
- Top 10 security items to improve in your Cuenta de AWS (Los 10 elementos de seguridad que debe mejorar en su cuenta de AWS)
- What do I do if I notice unauthorized activity in my Cuenta de AWS? (¿Qué hago si observo actividad no autorizada en mi cuenta de AWS?

Vídeos relacionados:

- Enable AWS adoption at scale with automation and governance (Facilitar la adopción de AWS a escala con la automatización y la gobernanza)
- Security Best Practices the Well-Architected Way (Prácticas recomendadas de seguridad al estilo de Well-Architected)
- Limitación del uso de credenciales raíz de AWS de AWS re:inforce 2022 – Security best practices with AWS IAM (AWS re:inforce 2022: prácticas recomendadas de seguridad con AWS IAM)

Ejemplos relacionados y laboratorios:

- Laboratorio: Cuenta de AWS and root user (La cuenta de AWS y el usuario raíz)

SEC01-BP03 Identificar y validar objetivos de control

En función de sus requisitos de cumplimiento y los riesgos identificados a partir de su modelo de amenazas, derive y valide los objetivos de control y los controles que tiene que aplicar a su carga de trabajo. La validación continua de los objetivos de control y los controles le ayuda a medir la efectividad de la mitigación de riesgos.

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

Guía para la implementación

- Identificar los requisitos de cumplimiento: descubra los requisitos organizativos, legales y de conformidad que debe cumplir la carga de trabajo.
- Identificar los recursos de cumplimiento de AWS: identifique los recursos que AWS tiene disponibles para ayudarle en términos de cumplimiento.
  - https://aws.amazon.com/compliance/
  - https://aws.amazon.com/artifact/

Recursos

Documentos relacionados:

- Directrices de auditoría de seguridad de AWS
- Boletines de seguridad

Vídeos relacionados:

- AWS Security Hub CSPM: gestionar las alertas de seguridad y automatizar el cumplimiento
- Prácticas recomendadas de seguridad a la forma Well-Architected

SEC01-BP04 Mantenerse al día de las amenazas de seguridad

Para ayudarle a definir e implementar los controles adecuados, reconozca los vectores de ataque manteniéndose al día de las últimas amenazas de seguridad. Use AWS Managed Services para facilitar la recepción de notificaciones de comportamientos inesperados o inusuales en sus cuentas de AWS. Investigue mediante herramientas de socios de AWS o orígenes de información sobre amenazas de terceros como parte de su flujo de información de seguridad. La lista de

vulnerabilidades y exposiciones comunes (CVE) contiene vulnerabilidades de ciberseguridad divulgadas públicamente que puede utilizar para estar al día.

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

Guía para la implementación

- Suscríbase a orígenes de inteligencia sobre amenazas: consulte periódicamente información de inteligencia de distintos orígenes que sean relevantes para las tecnologías que se usan en su carga de trabajo.
  - Lista de vulnerabilidades y exposiciones comunes (CVE)
- Considere AWS Shield Advanced : proporciona visibilidad casi en tiempo real sobre los orígenes de inteligencia si se puede acceder a su carga de trabajo desde Internet.

Recursos

Documentos relacionados:

- AWS Security Audit Guidelines (Directrices de auditoría de seguridad de AWS)
- AWS Shield
- Boletines de seguridad

Vídeos relacionados:

- Security Best Practices the Well-Architected Way (Prácticas recomendadas de seguridad a la forma Well-Architected)

SEC01-BP05 Mantenerse al día con las recomendaciones de seguridad

Manténgase al día de las recomendaciones de seguridad de AWS y de todo el sector para hacer evolucionar la postura de seguridad de su carga de trabajo. Los boletines de seguridad de AWS contienen información importante sobre la seguridad y las notificaciones de privacidad.

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

Guía para la implementación

- Siga las actualizaciones de AWS: suscríbase o compruebe regularmente las nuevas recomendaciones, consejos y trucos.

  - Laboratorios de AWS Well-Architected
  - Blog de seguridad de AWS
  - Documentación de servicio de AWS
- Suscríbase a las noticias del sector: consulte habitualmente noticias de distintas fuentes que sean relevantes para las tecnologías que se utilicen en su carga de trabajo.
  - Ejemplo: lista de vulnerabilidades y exposiciones comunes

Recursos

Documentos relacionados:

- Boletines de seguridad

Videos relacionados:

- Prácticas recomendadas de seguridad a la forma Well-Architected

SEC01-BP06 Automatizar la comprobación y validación de controles de seguridad en canalizaciones

Establezca referencias y plantillas seguras para mecanismos de seguridad que se comprueben y validen como parte de sus compilaciones, canalizaciones y procesos. Utilice herramientas y automatización para probar y validar todos los controles de seguridad de forma continua. Por ejemplo, escanee elementos como imágenes de máquinas y plantillas de infraestructura como código en busca de vulnerabilidades de seguridad, irregularidades y divergencias respecto de una referencia establecida en cada etapa. AWS CloudFormation Guard puede ayudarle a verificar que las plantillas de CloudFormation sean seguras, ahorrarle tiempo y reducir el riesgo de errores de configuración.

Reducir el número de configuraciones incorrectas de seguridad introducidas en un entorno de producción es fundamental. De este modo, establecer un control de calidad más exhaustivo y reducir los defectos durante el proceso de compilación facilitará obtener mejores resultados. Cuando sea posible, diseñe canalizaciones de integración e implementación continuas (CI/CD) para probar si hay problemas de seguridad. Las canalizaciones de CI/CD ofrecen la oportunidad de mejorar la seguridad en cada etapa de la compilación y la entrega. Las herramientas de seguridad de CI/CD también deben mantenerse actualizadas para mitigar las amenazas en evolución.

Realice un seguimiento de los cambios en la configuración de su carga de trabajo para ayudar con la auditoría normativa, la gestión de cambios y las investigaciones que puedan afectarle. Puede utilizar AWS Config para registrar y evaluar sus recursos de AWS y de terceros. Le permite evaluar y auditar de forma continua el cumplimiento general de las reglas y los paquetes de conformidad, que son conjuntos de reglas con acciones de corrección.

Entre las medidas de seguimiento de los cambios deberían incluirse cambios planificados que formen parte del proceso de control de cambios de la organización (lo que a veces se denomina "MACD": "mover", "agregar", "cambiar" y "eliminar", por sus siglas en inglés), cambios ad hoc o cambios inesperados, como incidentes. Los cambios pueden producirse en la infraestructura, pero también pueden estar relacionados con otras categorías, como los cambios en los repositorios de código, en los inventarios de aplicaciones e imágenes de máquinas, en los procesos y políticas o en la documentación.

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

Guía para la implementación

- Automatice la administración de la configuración: aplique y valide configuraciones seguras de forma automática mediante el uso de un servicio o herramienta de administración de la configuración.
  - AWS Systems Manager
  - AWS CloudFormation
  - Configurar una canalización de CI/CD en AWS

Recursos

Documentos relacionados:

- Cómo utilizar las políticas de control de servicios para establecer barreras de protección de permisos entre cuentas de AWS Organizations

Videos relacionados:

- Administración de entornos de AWS con varias cuentas utilizando AWS Organizations
- Prácticas recomendadas de seguridad a la forma Well-Architected

SEC01-BP07 Identificar amenazas y priorizar mitigaciones con un modelo de amenazas

Utilice el modelado de amenazas para identificar y mantener un registro actualizado de las amenazas potenciales y las mitigaciones asociadas para su carga de trabajo. Priorice sus amenazas y adapte sus mitigaciones de controles de seguridad para evitarlas, detectarlas y responder a ellas. Revise y mantenga todo esto en el contexto de su carga de trabajo y de la evolución del panorama de seguridad.

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

Guía para la implementación

¿Qué es el modelado de amenazas?

«El modelado de amenazas sirve para identificar, comunicar y comprender las amenazas y mitigaciones dentro del contexto de la protección de algo de valor» – «Application Threat Modeling» de Open Web Application Security Project (OWASP)

¿Por qué debe modelar las amenazas?

Los sistemas son complejos y, con el tiempo, se hacen más complejos y potentes aún, por lo que aportan más valor empresarial y aumentan la satisfacción y el compromiso de los clientes. Esto significa que, en las decisiones de diseño de TI, se deben tener en cuenta un número cada vez mayor de casos de uso. Debido a esta complejidad y al número de combinaciones de casos de uso, los enfoques no estructurados suelen resultar ineficaces para encontrar y mitigar las amenazas. En su lugar, se necesita un enfoque sistemático para encontrar las amenazas potenciales para el sistema, pero también para concebir mitigaciones y priorizarlas para asegurarse de que los limitados recursos de la organización tengan el máximo impacto en la mejora de la postura de seguridad general del sistema.

El modelado de amenazas está diseñado para proporcionar este enfoque sistemático, con el objetivo de encontrar y abordar los problemas en las primeras fases del proceso de diseño, cuando las mitigaciones tienen un coste y un esfuerzo relativamente bajos en comparación con las fases posteriores del ciclo de vida. Este enfoque se ajusta al principio de seguridad shift-left (desplazamiento a la izquierda) del sector. En última instancia, el modelado de amenazas se integra en el proceso de administración de riesgos de una organización y ayuda a tomar decisiones sobre qué controles aplicar mediante un enfoque basado en las amenazas.

¿Cuándo debe realizarse el modelado de amenazas?

Empiece a modelar las amenazas lo antes posible en el ciclo de vida de su carga de trabajo, ya que así tendrá más flexibilidad para actuar en relación con las amenazas que identifique. Al igual que ocurre con los errores de software, cuanto antes identifique las amenazas, más rentable le resultará abordarlas. Un modelo de amenazas es un documento vivo y debe evolucionar a medida que cambien sus cargas de trabajo. Revise los modelos de amenazas a lo largo del tiempo, especialmente cuando se produzca un cambio importante, un cambio en el panorama de las amenazas o cuando adopte una nueva función o servicio.

Pasos para la implementación

¿Cómo podemos realizar el modelado de amenazas?

Hay muchas formas diferentes de realizar el modelado de amenazas. Al igual que ocurre con los lenguajes de programación, cada una tiene sus ventajas y sus inconvenientes, por lo que debe elegir la que mejor le convenga. Un enfoque es comenzar con Shostack's 4 Question Frame for Threat Modeling (Marco de 4 preguntas para el modelado de amenazas de Shostack), que plantea preguntas abiertas para proporcionar una estructura a su modelado de amenazas:

1. ¿En qué está trabajando?

   La finalidad de esta pregunta es ayudarle a comprender y acordar el sistema que está construyendo y los detalles de ese sistema que son relevantes para la seguridad. Lo más habitual es responder que se está creando un modelo o diagrama, ya esto ayuda a visualizar lo que está construyendo, por ejemplo, con un diagrama de flujo de datos. Anotar las suposiciones y los detalles importantes sobre su sistema también le ayuda a definir el alcance del trabajo. De esta manera, todas las personas que contribuyen al modelo de amenazas pueden centrarse en lo mismo, y evita dar largos rodeos hacia temas que están fuera del alcance (lo que incluye versiones desactualizadas de su sistema). Por ejemplo, si crea una aplicación web, probablemente no merezca la pena que modele la secuencia de arranque de confianza del sistema operativo para los clientes del navegador, ya que no tiene capacidad para influir en esto con su diseño.

2. ¿Qué puede salir mal?

   Aquí es donde usted identifica las amenazas que afectan a su sistema. Las amenazas son acciones o acontecimientos accidentales o intencionados que tienen repercusiones no deseadas y podrían afectar a la seguridad de su sistema. Si no tiene una idea clara de lo que podría salir mal, no podrá hacer nada al respecto.

No existe una lista formal de lo que puede salir mal. Para crear esta lista, todos los miembros de su equipo y las personas relevantes implicadas en el modelado de amenazas deben hacer una lluvia de ideas y colaborar. Para facilitar la lluvia de ideas, puede utilizar un modelo de identificación de amenazas, como STRIDE, que sugiere diferentes categorías para evaluar las siguientes amenazas: suplantación de identidad, manipulación, repudio, divulgación de información, denegación de servicio y elevación de privilegios. Además, puede facilitar la lluvia de ideas inspirándose en listas e investigaciones existentes, como OWASP Top 10 (Los 10 principales riesgos de seguridad de OWASP), HiTrust Threat Catalog (Catálogo de amenazas de HiTrust) y el propio catálogo de amenazas de su organización.

3. ¿Qué vamos a hacer al respecto?

   Igual que en la pregunta anterior, no existe una lista formal de todas las mitigaciones posibles. En este paso, tenemos las amenazas, los actores y las áreas de mejora identificados en el paso anterior.

   La seguridad y la conformidad constituyen una responsabilidad compartida entre usted y AWS. Es importante entender que, cuando se pregunta «¿Qué vamos a hacer al respecto?», también se está preguntando «¿Quién es responsable de hacer algo al respecto?». Comprender el reparto de responsabilidades entre usted y AWS le ayuda a delimitar su modelado de amenazas a las mitigaciones que están bajo su control, que suelen ser una combinación de opciones de configuración de los servicios de AWS y las mitigaciones específicas de su propio sistema.

   En lo que se refiere a la parte de AWS de esa responsabilidad compartida, descubrirá que los servicios de AWS están dentro del ámbito de muchos programas de conformidad. Estos programas le ayudan a conocer los sólidos controles que hay en AWS para mantener la seguridad y la conformidad de la nube. Los clientes de AWS pueden descargar informes de auditoría de estos programas desde AWS Artifact.

   Independientemente de los servicios de AWS que utilice, el cliente siempre tiene una parte de la responsabilidad y las mitigaciones que se corresponden con estas responsabilidades deben incluirse en su modelo de amenazas. En cuanto a las mitigaciones de los controles de seguridad de los propios servicios de AWS, debe considerar la posibilidad de implementar controles de seguridad en todos los dominios, como los de administración de identidades y accesos (autenticación y autorización), protección de datos (en reposo y en tránsito), seguridad de la infraestructura, registro y supervisión. En la documentación de cada servicio de AWS, hay un capítulo dedicado a la seguridad que ofrece orientación sobre los controles de seguridad que deben considerarse como mitigaciones. Y lo que es más importante, considere el código que está

escribiendo y sus dependencias, y piense en los controles que podría establecer para hacer frente a esas amenazas. Estos controles podrían ser cosas como la validación de entradas, la gestión de sesiones y la gestión de límites. Muchas veces, la mayoría de las vulnerabilidades se introducen en el código personalizado, así que céntrese en esta área.

4. ¿Hemos hecho un buen trabajo?

   El objetivo es que su equipo y su organización mejoren con el tiempo tanto la calidad de los modelos de amenazas como la velocidad a la que los realizan. Estas mejoras se deben a una combinación de práctica, aprendizaje, enseñanza y revisión. Para profundizar y ponerse manos a la obra, es recomendable que usted y su equipo completen el curso de formación el taller Threat modeling the right way for builders training course (Modelado de amenazas de la forma adecuada para constructores). Además, si busca orientación sobre cómo integrar el modelado de amenazas en el ciclo de vida de desarrollo de aplicaciones de su organización, consulte la publicación How to approach threat modeling (Cómo abordar el modelado de amenazas) en el blog de seguridad de AWS.

Threat Composer

Para ayudarle y guiarle en la creación de modelos de amenazas, considere la posibilidad de utilizar la herramienta Threat Composer, que tiene como objetivo obtener valor más rápido al modelar amenazas. La herramienta le ayuda a hacer lo siguiente:

- Escribir instrucciones de amenazas útiles adaptadas a la gramática de amenazas que funcionan en un flujo de trabajo no lineal natural
- Generar un modelo de amenazas legible por humanos
- Generar un modelo de amenazas legible por máquina que le permita tratar los modelos de amenazas como código
- Ayudarle a identificar rápidamente las áreas de mejora de la calidad y la cobertura mediante el panel de información

Para obtener más información, visite «Threat Composer» y cambie al espacio de trabajo de ejemplo definido por el sistema.

Recursos

Prácticas recomendadas relacionadas:

- SEC01-BP03 Identificar y validar objetivos de control
- SEC01-BP04 Mantenerse al día de las amenazas de seguridad
- SEC01-BP05 Mantenerse al día con las recomendaciones de seguridad
- SEC01-BP08 Evaluar e implementar nuevos servicios y características de seguridad de forma periódica

Documentos relacionados:

- How to approach threat modeling (Cómo abordar el modelado de amenazas) (Blog de seguridad de AWS)
- NIST: Guide to Data-Centric System Threat Modelling (Guía para el modelado de amenazas de sistemas centrados en datos)

Vídeos relacionados:

- AWS Summit ANZ 2021 - How to approach threat modelling (AWS Summit ANZ 2021 - Cómo abordar el modelado de amenazas)
- AWS Summit ANZ 2022 - Scaling security – Optimise for fast and secure delivery (AWS Summit ANZ 2022 - Escalar la seguridad - Optimizar para una entrega rápida y segura)

Formación relacionada:

- Threat modeling the right way for builders – AWS Skill Builder virtual self-paced training (Modelado de amenazas de la forma correcta para constructores - Formación virtual autodidacta de AWS Skill Builder)
- Threat modeling the right way for builders – AWS Workshop (Modelado de amenazas de la forma correcta para constructores - Taller)

Herramientas relacionadas:

- Threat Composer

SEC01-BP08 Evaluar e implementar nuevos servicios y características de seguridad de forma periódica

Evalúe e implemente servicios y características de seguridad de AWS y socios de AWS que le permitan desarrollar la postura de seguridad de su carga de trabajo. En el blog de seguridad de AWS se destacan nuevos servicios y características de AWS, guías de implementación y directrices de seguridad generales. Novedades de AWS es una forma ideal de estar al día de las nuevas características, servicios y anuncios de AWS.

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

Guía para la implementación

- Planifique revisiones periódicas: cree un calendario de actividades de revisión que incluya requisitos de cumplimiento, evaluación de nuevas características y servicios de seguridad de AWS y revisión de las noticias del sector.
- Descubra servicios y características de AWS: descubra las características de seguridad disponibles para los servicios que está utilizando y las características nuevas que se vayan lanzando.
  - Blog de seguridad de AWS
  - Boletines de seguridad de AWS
  - Documentación de servicio de AWS
- Defina el proceso de incorporación de servicios de AWS: defina procesos para incorporar nuevos servicios de AWS. Incluya la forma en que evalúa los nuevos servicios de AWS en cuanto a su funcionalidad, así como los requisitos de conformidad de su carga de trabajo.
- Pruebe nuevos servicios y características: pruebe nuevos servicios y características a medida que se publiquen en un entorno que no sea de producción y que replique de forma fidedigna uno de producción.
- Implemente otros mecanismos de defensa: ponga en marcha mecanismos automatizados para defender su carga de trabajo, explore las opciones disponibles.
  - Corrección de recursos de AWS disconformes con Reglas de AWS Config

Recursos

Videos relacionados:

- Prácticas recomendadas de seguridad a la forma Well-Architected

Identity and Access Management

Preguntas

- SEGURIDAD 2. ¿Cómo administra la autenticación para personas y máquinas?
- SEGURIDAD 3. ¿Cómo administra los permisos para las personas y las máquinas?

SEGURIDAD 2. ¿Cómo administra la autenticación para personas y máquinas?

Hay dos tipos de identidades que tiene que administrar cuando tenga que utilizar cargas de trabajo de AWS seguras. Entender el tipo de identidad que tiene que administrar y a la que otorgar acceso ayuda a comprobar que las identidades adecuadas tengan acceso a los recursos correctos bajo las condiciones adecuadas.

Identidades humanas: los administradores, desarrolladores, operadores y clientes finales requieren una identidad para acceder a sus aplicaciones y entornos de AWS. Estos son miembros de la organización o usuarios externos con los que colabora y que interactúan con sus recursos de AWS a través de un navegador web, una aplicación de cliente o herramientas de línea de comandos interactivas.

Identidades de máquinas: las aplicaciones de servicio, las herramientas operativas y las cargas de trabajo requieren una identidad para realizar solicitudes a los servicios de AWS, como por ejemplo, para leer datos. Entre estas identidades se incluyen máquinas que se ejecutan en el entorno de AWS, como, por ejemplo, instancias Amazon EC2 o funciones de AWS Lambda. También puede administrar identidades de máquinas para terceros que necesiten acceso. Además, es posible que también tenga máquinas fuera de AWS que necesiten acceso al entorno de AWS.

Prácticas recomendadas

- SEC02-BP01 Usar mecanismos de inicio de sesión sólidos
- SEC02-BP02 Usar credenciales temporales
- SEC02-BP03 Almacenar y usar secretos de forma segura
- SEC02-BP04 Recurrir a un proveedor de identidades centralizado
- SEC02-BP05 Auditar y rotar las credenciales periódicamente
- SEC02-BP06 Aprovechar los grupos y atributos de usuarios

SEC02-BP01 Usar mecanismos de inicio de sesión sólidos

Los inicios de sesión (autenticación mediante credenciales de inicio de sesión) pueden ser arriesgados si no se utilizan mecanismos como la autenticación multifactor (MFA), especialmente en situaciones en las que las credenciales de inicio de sesión se han revelado de forma inadvertida o son fáciles de adivinar. Utilice mecanismos de inicio de sesión sólidos para reducir estos riesgos. Para ello, exija que se cumplan políticas de contraseñas sólidas y se utilice MFA.

Resultado deseado: reducir los riesgos que supone el acceso involuntario a las credenciales en AWS utilizando mecanismos de inicio de sesión sólidos para los usuarios de AWS Identity and Access Management (IAM), el usuario raíz de la Cuenta de AWS AWS IAM Identity Center (sucesor de AWS Single Sign-On) y los proveedores de identidad de terceros. Esto significa exigir que se use MFA, aplicar políticas de contraseñas sólidas y detectar comportamientos de inicio de sesión anómalos.

Antipatrones usuales:

- No aplicar una política de contraseñas segura para sus identidades que incluya contraseñas complejas y MFA.
- Compartir las mismas credenciales entre diferentes usuarios.
- No utilizar controles de detección de inicios de sesión sospechosos.

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

Guía para la implementación

Existen muchas formas en que las identidades humanas pueden iniciar sesión en AWS. Una práctica recomendada de AWS es confiar en un proveedor de identidades centralizado que utilice la federación (federación directa o mediante AWS IAM Identity Center) a la hora de autenticarse en AWS. En ese caso, deberá establecer un proceso de inicio de sesión seguro con su proveedor de identidades o Microsoft Active Directory.

Cuando abre una Cuenta de AWS por primera vez, comienza con un usuario raíz de la Cuenta de AWS. Solo debe utilizar el usuario raíz de la cuenta para configurar el acceso de sus usuarios (y para las tareas que requieran el usuario raíz). Es importante habilitar MFA para el usuario raíz de la cuenta inmediatamente después de abrir su Cuenta de AWS y proteger ese usuario utilizando la guía de prácticas recomendadas de AWS.

Si crea usuarios en AWS IAM Identity Center, asegure el proceso de inicio de sesión en ese servicio. Para las identidades de consumidor, puede utilizar Amazon Cognito user pools y proteger el proceso

de inicio de sesión en ese servicio, o utilizar uno de los proveedores de identidades que admiten los Amazon Cognito user pools.

Si utiliza usuarios de AWS Identity and Access Management (IAM), debe asegurar el proceso de inicio de sesión mediante IAM.

Independientemente del método de inicio de sesión que se utilice, es fundamental aplicar una política de inicio de sesión sólida.

Pasos para la implementación

Estas son recomendaciones generales para un inicio de sesión sólido. Los ajustes reales que configure se deben establecer en la política de la empresa o se debe utilizar un estándar como NIST 800-63.

- Exija el uso de MFA. Es una práctica recomendada de IAM exigir que se utilice MFA para identidades y cargas de trabajo humanas. Si se habilita MFA, habrá una capa adicional de seguridad que requiere que los usuarios proporcionen credenciales de inicio de sesión y una contraseña de un solo uso (OTP) o una cadena que se verifica criptográficamente y se genera desde un dispositivo físico.
- Imponga una longitud mínima para la contraseña. Esto es un factor fundamental para la seguridad de la contraseña.
- Imponga una complejidad de las contraseñas para que sean más difíciles de adivinar.
- Permita que los usuarios cambien sus propias contraseñas.
- Cree identidades individuales en lugar de credenciales compartidas. Si crea identidades individuales, puede dar a cada usuario un conjunto único de credenciales de seguridad. Tener usuarios individuales permite auditar la actividad de cada uno de ellos.

Recomendaciones sobre IAM Identity Center

- IAM Identity Center proporciona una política de contraseñas predefinida cuando se utiliza el directorio predeterminado que establece los requisitos de longitud, complejidad y reutilización de las contraseñas.
- Habilite MFA y configure el ajuste contextual o continuo para MFA cuando la fuente de identidad sea el directorio predeterminado, AWS Managed Microsoft AD o AD Connector.
- Permita que los usuarios registren sus propios dispositivos MFA.

Recomendaciones sobre el directorio de Amazon Cognito user pools:

- Configure los ajustes de seguridad de la contraseña.
- Exija el uso de MFA a los usuarios.
- Utilice la configuración de seguridad avanzada de Amazon Cognito user pools para funciones como la autenticación adaptativa, que puede bloquear inicios de sesión sospechosos.

Recomendaciones de usuarios de IAM

- Lo ideal es que utilice IAM Identity Center o la federación directa. Sin embargo, es posible que necesite usuarios de IAM. En ese caso, establezca una política de contraseñas para los usuarios de IAM. Puede usar una política de contraseñas para definir requisitos, tales como la longitud mínima o si deben contener caracteres alfanuméricos.
- Cree una política de IAM para imponer el inicio de sesión MFA de modo que los usuarios puedan administrar sus propias contraseñas y dispositivos MFA.

Recursos

Prácticas recomendadas relacionadas:

- SEC02-BP03 Almacenar y usar secretos de forma segura
- SEC02-BP04 Recurrir a un proveedor de identidades centralizado
- SEC03-BP08 Compartir recursos de forma segura en su organización

Documentos relacionados:

- Política de contraseñas de AWS IAM Identity Center (sucesor de AWS Single Sign-On)
- Política de contraseñas de usuarios de IAM
- Configuración de la contraseña del usuario raíz de la Cuenta de AWS
- Política de contraseñas de Amazon Cognito
- Credenciales de AWS
- Prácticas recomendadas de seguridad en IAM

Vídeos relacionados:

- Managing user permissions at scale with AWS IAM Identity Center (Administración de permisos de usuario a escala con AWS SSO)
- Mastering identity at every layer of the cake (Dominar la identidad en cada capa del pastel)

SEC02-BP02 Usar credenciales temporales

Al realizar cualquier tipo de autenticación, es mejor utilizar credenciales temporales en lugar de credenciales de larga duración para reducir o eliminar riesgos, tales como que las credenciales se divulguen, compartan o roben de forma inadvertida.

Resultado deseado: para reducir el riesgo que implican las credenciales de larga duración, utilice credenciales temporales siempre que sea posible tanto para las identidades humanas como para las de las máquinas. Las credenciales de larga duración entrañan muchos riesgos; por ejemplo, pueden subirse en el código en repositorios públicos de GitHub. Al utilizar credenciales temporales, reducirá enormemente las posibilidades de que las credenciales se vean comprometidas.

Antipatrones usuales:

- Desarrolladores que utilizan claves de acceso de larga duración de IAM users en lugar de obtener credenciales temporales de la CLI mediante federación.
- Desarrolladores que incrustan claves de acceso de larga duración en su código y suben ese código a repositorios de Git públicos.
- Desarrolladores que incrustan claves de acceso de larga duración en aplicaciones móviles que luego se ponen a disposición de todo el mundo en las tiendas de aplicaciones.
- Usuarios que comparten claves de acceso de larga duración con otros usuarios, o empleados que abandonan la empresa con claves de acceso de larga duración aún en su poder.
- Utilizar claves de acceso de larga duración para identidades de máquinas cuando podrían utilizarse credenciales temporales.

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

Guía para la implementación

Utilice credenciales de seguridad temporales en lugar de credenciales de larga duración para todas las solicitudes de la API y la CLI de AWS. Las solicitudes de la API y la CLI a los servicios de AWS deben, en prácticamente todos los casos, firmarse utilizando claves de acceso de AWS. Estas solicitudes pueden firmarse con credenciales temporales o de larga duración. El único caso en

que debe utilizar credenciales de larga duración, que también se conocen como claves de acceso de larga duración, es cuando utiliza un usuario de IAM o el usuario raíz de la Cuenta de AWS. Si se federa a AWS o asume un rol de IAM a través de otros métodos, se generan credenciales temporales. Incluso cuando accede a la Consola de administración de AWS utilizando credenciales de inicio de sesión, se generan credenciales temporales para que pueda realizar llamadas a los servicios de AWS. Hay pocas situaciones en las que necesite credenciales de larga duración y casi todas las tareas se pueden realizar utilizando credenciales temporales.

Evitar el uso de credenciales de larga duración en favor de credenciales temporales debería acompañarse de una estrategia de reducción del uso de usuarios de IAM en favor de la federación y los roles de IAM. Aunque en el pasado se han utilizado usuarios de IAM tanto para identidades humanas como de máquinas, ahora recomendamos no utilizarlos para evitar los riesgos que conlleva el uso de claves de acceso de larga duración.

Pasos para la implementación

Para identidades humanas, como las de empleados, administradores, desarrolladores, operadores y clientes:

- Debe recurrir a un proveedor de identidades centralizado y exigir a los usuarios humanos que utilicen la federación con un proveedor de identidades para acceder a AWS utilizando credenciales temporales. La federación para sus usuarios puede realizarse con federación directa a cada Cuenta de AWS o mediante AWSIAM Identity Center (sucesor de AWS IAM Identity Center) y el proveedor de identidades de su elección. La federación tiene una serie de ventajas con respecto a los usuarios de IAM, además de eliminar las credenciales de larga duración. Sus usuarios también pueden solicitar credenciales temporales desde la línea de comandos para la federación directa o mediante IAM Identity Center. Esto significa que hay pocos casos de uso que requieran usuarios de IAM o credenciales de larga duración para sus usuarios.
- Cuando conceda a terceros (por ejemplo, proveedores de software como servicio [SaaS]), acceso a los recursos de su Cuenta de AWS, puede utilizar roles entre cuentas y políticas basadas en recursos.
- Si necesita conceder acceso a sus recursos de AWS a aplicaciones para consumidores o clientes, puede utilizar grupos de identidades de Amazon Cognito o grupos de usuarios de Amazon Cognito user pools para proporcionar credenciales temporales. Los permisos para las credenciales se configuran a través de roles de IAM También puede definir un rol de IAM separado con permisos limitados para los usuarios invitados que no se hayan autenticado.

En el caso de las identidades de máquina, puede que necesite utilizar credenciales de larga duración. En estos casos, debe exigir que las cargas de trabajo utilicen credenciales temporales con roles de IAM para acceder a AWS.

- Para Amazon Elastic Compute Cloud (Amazon EC2), puede utilizar roles para Amazon EC2.

- AWS Lambda le permite configurar un rol de ejecución de Lambda para conceder al servicio permisos para realizar acciones de AWS utilizando credenciales temporales. Existen muchos otros modelos similares para que los servicios de AWS concedan credenciales temporales utilizando roles de IAM.
- Para los dispositivos IoT, puede utilizar el proveedor de credenciales de AWS IoT Core para solicitar credenciales temporales.
- Para sistemas locales o sistemas que se ejecutan fuera de AWS que necesitan acceso a los recursos de AWS, puede utilizar Funciones de IAM en cualquier lugar.

Hay escenarios en los que las credenciales temporales no son una opción y puede que necesite utilizar credenciales de larga duración. En estas situaciones, audite y rote las credenciales periódicamente y rote las claves de acceso con regularidad para los casos de uso que requieran credenciales de larga duración. Algunos ejemplos que podrían requerir credenciales de larga duración son los plugins de WordPress y los clientes de AWS de terceros. En situaciones en las que deba utilizar credenciales de larga duración, o para credenciales que no sean claves de acceso de AWS, como inicios de sesión en bases de datos, puede utilizar un servicio diseñado para administrar secretos, como AWS Secrets Manager. Secrets Manager simplifica la administración, la rotación y el almacenamiento seguro de secretos cifrados mediante servicios compatibles. Si desea obtener más información sobre la rotación de las credenciales de larga duración, consulte Rotación de las claves de acceso.

Recursos

Prácticas recomendadas relacionadas:

- SEC02-BP03 Almacenar y usar secretos de forma segura
- SEC02-BP04 Recurrir a un proveedor de identidades centralizado
- SEC03-BP08 Compartir recursos de forma segura en su organización

Documentos relacionados:

- Credenciales de seguridad temporales
- Credenciales de AWS
- Prácticas recomendadas de seguridad en IAM
- Roles de IAM
- IAM Identity Center
- Federación y proveedores de identidades
- Rotación de las claves de acceso
- Soluciones de socios con competencia en seguridad: acceso y control de acceso
- Cuenta de AWS usuario raíz

Vídeos relacionados:

- Managing user permissions at scale with AWSIAM Identity Center (successor to AWS IAM Identity Center) (Administración de permisos de usuario a escala con AWS SSO [sucesor de AWS Single Sign-On])
- Mastering identity at every layer of the cake (Dominar la identidad en cada capa del pastel)

SEC02-BP03 Almacenar y usar secretos de forma segura

Una carga de trabajo necesita una capacidad automatizada para demostrar su identidad a bases de datos, recursos y servicios de terceros. Para ello, se utilizan credenciales de acceso secretas, como claves de acceso a API, contraseñas y tokens OAuth. El uso de un servicio creado específicamente para almacenar, administrar y rotar estas credenciales ayuda a reducir la probabilidad de que dichas credenciales se vean comprometidas.

Resultado deseado: implementar un mecanismo para administrar de forma segura las credenciales de las aplicaciones que logre los siguientes objetivos:

- Identificar qué secretos son necesarios para la carga de trabajo.
- Reducir el número de credenciales de larga duración necesarias y sustituirlas por credenciales de corta duración cuando sea posible.
- Establecer un almacenamiento seguro y una rotación automatizada de las credenciales restantes de larga duración.
- Auditar el acceso a los secretos que existen en la carga de trabajo.

- Supervisar de forma continua para verificar que no hay secretos incrustados en el código fuente durante el proceso de desarrollo.
- Reducir la probabilidad de que las credenciales se divulguen de forma inadvertida.

Antipatrones usuales:

- Credenciales no rotativas.
- Almacenar credenciales a largo plazo en el código fuente o en archivos de configuración.
- Almacenar credenciales en reposo sin cifrar.

Beneficios de establecer esta práctica recomendada:

- Los secretos se almacenan cifrados en reposo y en tránsito.
- El acceso a las credenciales se controla a través de una API (es algo parecido a una máquina expendedora de credenciales).
- El acceso a una credencial (tanto de lectura como de escritura) se audita y registra.
- Separación de preocupaciones: la rotación de credenciales la realiza un componente independiente, que puede separarse del resto de la arquitectura.
- Los secretos se distribuyen automáticamente bajo demanda a los componentes de software y la rotación se produce en una ubicación central.
- El acceso a las credenciales puede controlarse de forma detallada.

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

Guía para la implementación

En el pasado, las credenciales que se utilizaban para autenticarse en bases de datos, API de terceros, tokens y otros secretos podían estar incrustadas en el código fuente o en archivos del entorno. AWS proporciona varios mecanismos para almacenar estas credenciales de forma segura, rotarlas automáticamente y auditar su uso.

La mejor manera de abordar la administración de secretos es seguir la norma de eliminar, sustituir y rotar. La credencial más segura es aquella que no se tiene que almacenar, administrar ni manejar. Es posible que haya credenciales que ya no sean necesarias para el funcionamiento de la carga de trabajo y que, por tanto, puedan eliminarse de forma segura.

En el caso de las credenciales que siguen siendo necesarias para el correcto funcionamiento de la carga de trabajo, podría existir la oportunidad de sustituir una credencial de larga duración por una credencial temporal o de corta duración. Por ejemplo, en lugar de codificar una clave de acceso secreta de AWS, considere la posibilidad de sustituir esa credencial de larga duración por una credencial temporal utilizando roles de IAM.

Es posible que algunos secretos de larga duración no puedan eliminarse ni sustituirse. Estos secretos pueden almacenarse en un servicio como AWS Secrets Manager, donde pueden almacenarse, administrarse y rotarse de forma centralizada y periódica.

Una auditoría del código fuente y de los archivos de configuración de la carga de trabajo puede revelar muchos tipos de credenciales. La siguiente tabla resume las estrategias para manejar los tipos comunes de credenciales:

Credential type	Description	Suggested strategy
IAM access keys	AWS IAM access and secret keys used to assume IAM roles inside of a workload	Replace: Use Roles de IAM assigned to the compute instances (such as Amazon EC2 or AWS Lambda) instead. For interoperability with third parties that require access to resources in your Cuenta de AWS, ask if they support Acceso entre cuentas de AWS. For mobile apps, consider using temporary credentials through Grupos de identidades de Amazon Cognito (identidades federadas). For workloads running outside of AWS, consider Funciones de IAM en cualquier lugar or Activaciones híbridas de AWS Systems Manager.

Credential type	Description	Suggested strategy
SSH keys	Secure Shell private keys used to log into Linux EC2 instances, manually or as part of an automated process	Replace: Use AWS Systems Manager or EC2 Instance Connect to provide programmatic and human access to EC2 instances using IAM roles.
Application and database credentials	Passwords – plain text string	Rotate: Store credentials in AWS Secrets Manager and establish automated rotation if possible.
Amazon RDS and Aurora Admin Database credentials	Passwords – plain text string	Replace: Use the Integración de Secrets Manager con Amazon RDS or Amazon Aurora. In addition, some RDS database types can use IAM roles instead of passwords for some use cases (for more detail, see Autenticación de bases de datos de IAM).
OAuth tokens	Secret tokens – plain text string	Rotate: Store tokens in AWS Secrets Manager and configure automated rotation.
API tokens and keys	Secret tokens – plain text string	Rotate: Store in AWS Secrets Manager and establish automated rotation if possible.

Un antipatrón común es incrustar claves de acceso de IAM dentro del código fuente, los archivos de configuración o las aplicaciones móviles. Cuando se requiera una clave de acceso de IAM para comunicarse con un servicio de AWS, utilice credenciales de seguridad temporales (a corto plazo). Estas credenciales a corto plazo pueden proporcionarse a través de roles de IAM para instancias de EC2, roles de ejecución para funciones Lambda, roles de IAM de Cognito para el acceso de usuarios

móviles y políticas de IoT Core para dispositivos IoT. Cuando interactúe con terceros, es preferible que delegue el acceso a un rol de IAM con el acceso necesario a los recursos de su cuenta en lugar de configurar un usuario de IAM y enviar a ese tercero la clave de acceso secreta para ese usuario.

Hay muchos casos en los que la carga de trabajo requiere que se almacenen los secretos necesarios para interoperar con otros servicios y recursos. AWS Secrets Manager se ha creado específicamente para administrar de forma segura estas credenciales, así como el almacenamiento, el uso y la rotación de tokens de API, contraseñas y otras credenciales.

AWS Secrets Manager proporciona cinco capacidades clave para garantizar el almacenamiento y la gestión seguros de credenciales confidenciales: cifrado en reposo, cifrado en tránsito, auditoría exhaustiva, control de acceso detallado y rotación de credenciales extensible. También son aceptables otros servicios de administración de secretos de socios de AWS o soluciones desarrolladas localmente que proporcionen capacidades y garantías similares.

Pasos para la implementación

1. Identifique rutas de código que contengan credenciales codificadas mediante herramientas automatizadas como Amazon CodeGuru.
   - Utilice Amazon CodeGuru para analizar sus repositorios de código. Una vez finalizada la revisión, filtre `Type=Secrets` en CodeGuru para encontrar las líneas de código problemáticas.
2. Identifique las credenciales que pueden eliminarse o sustituirse.
   a. Identifique las credenciales que ya no sean necesarias y márquelas para eliminarlas.
   b. En el caso de las claves secretas de AWS que estén incrustadas en el código fuente, sustitúyalas por roles de IAM asociados a los recursos necesarios. Si parte de su carga de trabajo está fuera de AWS pero requiere credenciales de IAM para acceder a recursos de AWS, considere la posibilidad de usar Funciones de IAM en cualquier lugar o activaciones híbridas de AWSSystems Manager.
3. Para otros secretos de terceros de larga duración que requieran el uso de la estrategia de rotación, integre Secrets Manager en su código para recuperar secretos de terceros en tiempo de ejecución.
   a. La consola CodeGuru puede crear automáticamente un secreto en Secrets Manager utilizando las credenciales descubiertas.
   b. Integre la recuperación de secretos desde Secrets Manager en el código de su aplicación.
      - Las funciones Lambda sin servidor pueden utilizar una extensión de Lambda agnóstica del lenguaje.

    - Para instancias o contenedores EC2, AWS proporciona ejemplos de código del lado del cliente para recuperar secretos de Secrets Manager en varios lenguajes de programación populares.
4. Revise periódicamente su base de código y vuelva a analizarlo para verificar que no se hayan añadido nuevos secretos.
    - Considere la posibilidad de utilizar una herramienta como git-secrets para evitar que se envíen nuevos secretos a su repositorio de código fuente.
5. Supervise la actividad de Secrets Manager en busca de indicios de un uso inesperado, un acceso inapropiado a secretos o intentos de eliminar secretos.
6. Reduzca la exposición humana a las credenciales. Restrinja el acceso para leer, escribir y modificar credenciales a un rol de IAM dedicado a este fin, y solo proporcione acceso para asumir el rol a un pequeño subconjunto de usuarios operativos.

Recursos

Prácticas recomendadas relacionadas:

- SEC02-BP02 Usar credenciales temporales
- SEC02-BP05 Auditar y rotar las credenciales periódicamente

Documentos relacionados:

- Introducción a AWS Secrets Manager
- Federación y proveedores de identidades
- Amazon CodeGuru Introduces Secrets Detector (El revisor de Amazon CodeGuru presenta el detector de secretos)
- Cómo AWS Secrets Manager usa AWS Key Management Service
- Cifrado y descifrado de secretos en Secrets Manager
- Entradas del blog de Secrets Manager
- Amazon RDS presenta la integración con AWS Secrets Manager

Vídeos relacionados:

- Best Practices for Managing, Retrieving, and Rotating Secrets at Scale (Prácticas recomendadas para administrar, recuperar y rotar secretos a escala)

- Find Hard-Coded Secrets Using Amazon CodeGuru Secrets Detector (Encuentre secretos difíciles de descifrar utilizando el detector de secretos de Amazon CodeGuru)
- Securing Secrets for Hybrid Workloads Using AWS Secrets Manager (Asegurar secretos para cargas de trabajo híbridas utilizando AWS Secrets Manager)

Talleres relacionados:

- Almacene, recupere y administre credenciales confidenciales en AWS Secrets Manager
- Activaciones híbridas de AWS Systems Manager

SEC02-BP04 Recurrir a un proveedor de identidades centralizado

Para las identidades de la plantilla (empleados y contratistas), recurra a un proveedor de identidades que le permita administrar las identidades desde un lugar centralizado. De este modo se facilita la administración del acceso en varias aplicaciones y sistemas, pues crea, asigna, administra, revoca y audita el acceso desde un único lugar.

Resultado deseado: tiene un proveedor de identidades centralizado en el que administra de forma centralizada los usuarios de la plantilla, las políticas de autenticación (como la exigencia de la autenticación multifactor [MFA]) y la autorización de los sistemas y las aplicaciones (como la asignación del acceso en función de la pertenencia o los atributos del grupo del usuario). Los usuarios de la plantilla inician sesión en el proveedor de identidades central y se federan (inicio de sesión único) en aplicaciones internas y externas, lo que elimina la necesidad de que los usuarios recuerden varias credenciales. El proveedor de identidades está integrado con sus sistemas de recursos humanos (RR. HH.) para que los cambios de personal se sincronicen automáticamente con su proveedor de identidades. Por ejemplo, si alguien abandona la organización, puede revocar automáticamente el acceso a las aplicaciones y sistemas federados (incluido AWS). Ha habilitado el registro de auditoría detallado en su proveedor de identidades y supervisa estos registros para detectar comportamientos inusuales de los usuarios.

Patrones comunes de uso no recomendados:

- No se utiliza la federación ni el inicio de sesión único. Los usuarios de la plantilla crean cuentas de usuario y credenciales independientes en numerosas aplicaciones y sistemas.
- No ha automatizado el ciclo de vida de las identidades de los usuarios de la plantilla, por ejemplo, no ha integrado su proveedor de identidades con sus sistemas de recursos humanos. Cuando un

usuario abandona la organización o cambia de rol, se sigue un proceso manual para eliminar o actualizar sus registros en varias aplicaciones y sistemas.

Beneficios de establecer esta práctica recomendada: al usar un proveedor de identidades centralizado, hay un único lugar en el que se administran las identidades y políticas de los usuarios de la plantilla, la capacidad de asignar acceso a aplicaciones a los usuarios y grupos y la capacidad de supervisar la actividad de inicio de sesión de los usuarios. Al integrarse con sus sistemas de recursos humanos (RR. HH.), cuando un usuario cambia de rol, estos cambios se sincronizan con el proveedor de identidades y sus aplicaciones y permisos asignados se actualizan automáticamente. Cuando un usuario abandona la organización, su identidad se inhabilita automáticamente en el proveedor de identidades, lo que revoca su acceso a las aplicaciones y sistemas federados.

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

Guía para la implementación

Guía para el acceso a AWS de los usuarios de la plantilla

Es posible que los usuarios de la plantilla, como empleados y contratistas de su organización, tengan que acceder a AWS mediante la Consola de administración de AWS o la AWS Command Line Interface (AWS CLI) para desempeñar sus funciones laborales. Para conceder acceso a AWS, puede federar a los usuarios de la plantilla desde su proveedor de identidades centralizado en AWS en dos niveles: federación directa a cada Cuenta de AWS o federación de varias cuentas en su organización de AWS.

- Para federar a los usuarios de la plantilla directamente con cada Cuenta de AWS, puede utilizar un proveedor de identidades centralizado para federar a AWS Identity and Access Management en esa cuenta. La flexibilidad de IAM le permite habilitar un proveedor de identidades SAML 2.0 o Open ID Connect (OIDC) para cada Cuenta de AWS y utilizar atributos de usuario federados para el control de acceso. Para iniciar sesión en el proveedor de identidades, los usuarios de la plantilla utilizarán su navegador web y proporcionarán sus credenciales (como contraseñas y códigos de token de MFA). El proveedor de identidades envía una aserción SAML a su navegador que se envía a la URL de inicio de sesión de la Consola de administración de AWS para permitir que el usuario haga un inicio de sesión único en la Consola de administración de AWS asumiendo un rol de IAM. Los usuarios también pueden obtener credenciales de API de AWS temporales para usarlas en la AWS CLI o bien SDK de AWS de AWS STS asumiendo el rol de IAM mediante una aserción SAML del proveedor de identidades.

- Para federar a los usuarios de la plantilla con varias cuentas en su organización de AWS, puede usar AWS IAM Identity Center para administrar de forma centralizada el acceso de los usuarios de la plantilla a las aplicaciones y Cuentas de AWS. Habilite el centro de identidades para su organización y configure el origen de las identidades. IAM Identity Center proporciona un directorio de orígenes de identidades predeterminado que puede usar para administrar sus usuarios y grupos. Como alternativa, puede elegir un origen de identidades externo conectándose a su proveedor de identidades externo con SAML 2.0 y aprovisionando automáticamente usuarios y grupos con SCIM, o conectándose a su directorio de Microsoft AD con Directory Service. Una vez configurado un origen de identidades, puede asignar acceso a Cuentas de AWS a usuarios y grupos definiendo políticas de privilegios mínimos en sus conjuntos de permisos. Los usuarios de la plantilla pueden autenticarse a través de su proveedor de identidades central para iniciar sesión en el portal de acceso de AWS e iniciar sesión única en las aplicaciones en la nube y Cuentas de AWS que se les asignen. Los usuarios pueden configurar la AWS CLI v2 para autenticarse con el centro de identidades y obtener credenciales para ejecutar comandos de AWS CLI. El centro de identidades también permite el acceso mediante el inicio de sesión único a aplicaciones de AWS como Amazon SageMaker AI Studio y portales de Monitor de AWS IoT SiteWise.

Tras seguir las instrucciones anteriores, los usuarios de la plantilla ya no tendrán que usar grupos y IAM users para las operaciones normales al administrar las cargas de trabajo de AWS. En cambio, los usuarios y grupos se administran fuera de AWS y los usuarios pueden acceder a los recursos de AWS como una Identidad federada. Las identidades federadas utilizan los grupos definidos por su proveedor de identidades centralizado. Debe identificar y eliminar los grupos de IAM, los IAM users y las credenciales de usuario de larga duración (contraseñas y claves de acceso) que ya no sean necesarios en sus cuentas de Cuentas de AWS. Puede buscar credenciales no utilizadas con informes de credenciales de IAM, eliminar los IAM users correspondientes y eliminar los grupos de IAM. Puede aplicar una política de control de servicio (SCP) a su organización, lo que ayuda a evitar la creación de nuevos grupos y IAM users. Al hacerlo, exige que el acceso a AWS tenga lugar a través de identidades federadas.

Guía para los usuarios de sus aplicaciones

Puede administrar las identidades de los usuarios de sus aplicaciones, como una aplicación móvil, mediante Amazon Cognito como su proveedor de identidades centralizado. Amazon Cognito permite la autenticación, la autorización y la administración de usuarios para sus aplicaciones web y móviles. Amazon Cognito proporciona un almacén de identidades que se escala a millones de usuarios, admite la federación de identidades sociales y empresariales y ofrece características de seguridad avanzadas para ayudar a proteger a sus usuarios y su empresa. Puede integrar su aplicación web o

móvil personalizada con Amazon Cognito para añadir autenticación de usuarios y control de acceso a sus aplicaciones en cuestión de minutos. Basado en estándares de identidad abiertos, como SAML y Open ID Connect (OIDC), Amazon Cognito es compatible con varias normativas de cumplimiento y se integra con los recursos de desarrollo de frontend y backend.

Pasos para la implementación

Pasos para los usuarios de la plantilla que acceden a AWS

- Federe a los usuarios de la plantilla para AWS mediante un proveedor de identidades centralizado utilizando uno de los siguientes enfoques:
  - Utilice IAM Identity Center para habilitar el inicio de sesión único en varias Cuentas de AWS de su organización de AWS mediante la federación con su proveedor de identidades.
  - Utilice IAM para conectar su proveedor de identidades directamente a cada Cuenta de AWS, lo que permite un acceso federado y detallado.
- Identifique y elimine los grupos y IAM users que se sustituyan por identidades federadas.

Pasos para los usuarios de sus aplicaciones

- Utilice Amazon Cognito como proveedor de identidades centralizado para sus aplicaciones.
- Integre sus aplicaciones personalizadas con Amazon Cognito mediante OpenID Connect y OAuth. Puede desarrollar sus aplicaciones personalizadas mediante las bibliotecas de Amplify, que proporcionan interfaces sencillas para integrarse con una variedad de servicios de AWS, como Amazon Cognito para la autenticación.

Recursos

Prácticas recomendadas por Well-Architected:

- SEC02-BP06 Aprovechar los grupos y atributos de usuarios
- SEC03-BP02 Conceder acceso con privilegios mínimos
- SEC03-BP06 Administrar el acceso en función del ciclo de vida

Documentos relacionados:

- Identity federation in AWS
- Prácticas recomendadas de seguridad en IAM

- AWS Identity and Access Management Best practices
- Getting started with IAM Identity Center delegated administration
- How to use customer managed policies in IAM Identity Center for advanced use cases
- AWS CLI v2: IAM Identity Center credential provider

Vídeos relacionados:

- AWS re:Inforce 2022 - AWS Identity and Access Management (IAM) deep dive
- AWS re:Invent 2022 - Simplify your existing workforce access with IAM Identity Center
- AWS re:Invent 2018: Mastering Identity at Every Layer of the Cake

Ejemplos relacionados:

- Workshop: Using AWS IAM Identity Center to achieve strong identity management
- Workshop: Serverless identity

Herramientas relacionadas:

- Socios con competencia en seguridad de AWS: Identity and Access Management
- saml2aws

SEC02-BP05 Auditar y rotar las credenciales periódicamente

Audite y rote las credenciales periódicamente para limitar el tiempo que pueden utilizarse para acceder a sus recursos. Las credenciales de larga duración entrañan muchos riesgos, y estos riesgos pueden reducirse rotándolas regularmente.

Resultado deseado: implementar la rotación de credenciales para ayudar a reducir los riesgos asociados al uso de credenciales de larga duración. Auditar regularmente y corregir la no conformidad con las políticas de rotación de credenciales.

Antipatrones usuales:

- No auditar el uso de credenciales.
- Utilizar credenciales de larga duración de forma innecesaria.
- Utilizar credenciales de larga duración y no rotarlas regularmente.

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

Guía para la implementación

Cuando no pueda confiar en credenciales temporales y necesite credenciales de larga duración, audítelas para verificar que los controles definidos, por ejemplo, la autenticación multifactor (MFA), se aplican, se rotan periódicamente y tienen el nivel de acceso adecuado.

Es necesario realizar una validación periódica, preferiblemente mediante una herramienta automatizada, para verificar que se están aplicando los controles correctos. En el caso de las identidades humanas, debe exigir a los usuarios que cambien sus contraseñas periódicamente y retirar las claves de acceso para sustituirlas por credenciales temporales. Si pasa de usuarios de AWS Identity and Access Management (IAM) a identidades centralizadas, puede generar un informe de credenciales para auditar a sus usuarios.

También recomendamos que aplique y supervise una configuración de MFA en su proveedor de identidades. Puede configurar Reglas de AWS Config o utilizar los estándares de seguridad de AWS Security Hub CSPM para supervisar si los usuarios tienen habilitado MFA. Considere la posibilidad de utilizar Funciones de IAM en cualquier lugar para proporcionar credenciales temporales para identidades de máquinas. En situaciones en las que no sea posible utilizar roles de IAM y credenciales temporales, es necesario realizar auditorías frecuentes y rotar las claves de acceso.

Pasos para la implementación

- Audite las credenciales periódicamente: auditar las identidades que están configuradas en el proveedor de identidades e IAM le permite asegurarse de que las únicas identidades que pueden acceder a su carga de trabajo son aquellas que estén autorizadas. Dichas identidades pueden incluir, entre otras, usuarios de IAM, usuarios de AWS IAM Identity Center, usuarios de Active Directory o usuarios de un proveedor de identidades ascendente diferente. Por ejemplo, elimine a las personas que abandonen la organización y los roles entre cuentas que ya no sean necesarios. Implante un proceso para auditar periódicamente los permisos a los servicios a los que accede una entidad de IAM. Esto le ayudará a identificar las políticas que debe modificar para eliminar los permisos que no se utilizan. Utilice informes de credenciales y AWS Identity and Access Management Access Analyzer para auditar las credenciales y los permisos de IAM. Puede utilizar Amazon CloudWatch para configurar alarmas para llamadas a la API específicas que se realicen dentro de su entorno de AWS. Amazon GuardDuty también puede alertarle de actividades inesperadas, que podrían indicar que el acceso es demasiado permisivo o que se ha producido un acceso no intencionado a las credenciales de IAM.

- Rote las credenciales periódicamente: cuando no pueda utilizar credenciales temporales, rote las claves de acceso de larga duración de IAM de forma periódica (cada 90 días como máximo). Si se revela una clave de acceso de forma involuntaria sin su conocimiento, esto limita el tiempo durante el que se pueden utilizar las credenciales para acceder a sus recursos. Si desea obtener más información sobre la rotación de las claves de acceso para los usuarios de IAM, consulte Rotación de las claves de acceso.
- Revise los permisos de IAM: para mejorar la seguridad de su Cuenta de AWS, revise y supervise de forma regular cada una de sus políticas de IAM. Verifique que las políticas sigan el principio del privilegio mínimo.
- Considere la posibilidad de automatizar la creación y actualización de recursos de IAM: IAM Identity Center automatiza muchas tareas de IAM, como la administración de roles y políticas. Como alternativa, se puede utilizar AWS CloudFormation para automatizar el despliegue de los recursos de IAM, incluidos los roles y las políticas, para reducir la posibilidad de que se produzcan errores humanos, ya que las plantillas se pueden verificar y controlar por versiones.
- Utilice Funciones de IAM en cualquier lugar para sustituir a los usuarios de IAM en las identidades de máquina: Funciones de IAM en cualquier lugar le permite utilizar roles en áreas en las que tradicionalmente no podía, como los servidores locales. Funciones de IAM en cualquier lugar utiliza un certificado X.509 de confianza para autenticarse en AWS y recibir credenciales temporales. El uso de Funciones de IAM en cualquier lugar evita la necesidad de rotar estas credenciales, ya que las credenciales de larga duración ya no se almacenan en su entorno local. Tenga en cuenta que deberá supervisar y rotar el certificado X.509 a medida que se acerque su fecha de vencimiento.

Recursos

Prácticas recomendadas relacionadas:

- SEC02-BP02 Usar credenciales temporales
- SEC02-BP03 Almacenar y usar secretos de forma segura

Documentos relacionados:

- Introducción a AWS Secrets Manager
- Prácticas recomendadas de seguridad en IAM
- Federación y proveedores de identidades
- Soluciones de socios con competencia en seguridad: acceso y control de acceso

- Credenciales de seguridad temporales
- Obtener informes de credenciales para su Cuenta de AWS

Vídeos relacionados:

- Best Practices for Managing, Retrieving, and Rotating Secrets at Scale (Prácticas recomendadas para administrar, recuperar y rotar secretos a escala)
- Managing user permissions at scale with AWS IAM Identity Center (Administración de permisos de usuario a escala con AWS SSO)
- Mastering identity at every layer of the cake (Dominar la identidad en cada capa del pastel)

Ejemplos relacionados:

- Well-Architected Lab - Automated IAM User Cleanup (Laboratorio de AWS Well-Architected: Limpieza automatizada de usuarios de IAM)
- Well-Architected Lab - Automated Deployment of IAM Groups and Roles (Laboratorio de AWS Well-Architected: Despliegue automatizado de grupos y roles de IAM)

SEC02-BP06 Aprovechar los grupos y atributos de usuarios

A medida que crezca el número de usuarios que administra, tendrá que determinar formas de organizarlos para que pueda administrarlos a escala. Coloque a usuarios con requisitos de seguridad comunes en grupos definidos por su proveedor de identidades, y prepare mecanismos para garantizar que los atributos de usuarios que puedan usarse para controlar el acceso (por ejemplo, los de departamento o ubicación) sean correctos y estén actualizados. Use estos grupos y atributos para controlar el acceso, en lugar de usuarios individuales. Esto le permitirá administrar el acceso de forma centralizada cambiando la pertenencia a un grupo de un usuario o sus atributos una vez con un conjunto de permisos, en lugar de actualizar muchas políticas individuales cuando el acceso de un usuario tenga que cambiarse. Puede usar AWS IAM Identity Center (IAM Identity Center) para administrar grupos y atributos de usuarios. IAM Identity Center admite los atributos de usuarios utilizados más habitualmente, ya sea mediante introducción manual durante la creación del usuario o mediante un aprovisionamiento automático con un motor de sincronización, como lo que se define en el estándar Sistema para administración de identidades entre dominios (SCIM).

Coloque a usuarios con requisitos de seguridad comunes en grupos definidos por su proveedor de identidades, y prepare mecanismos para garantizar que los atributos de usuarios que puedan usarse

para controlar el acceso (por ejemplo, los de departamento o ubicación) sean correctos y estén actualizados. Use estos grupos y atributos en lugar de usuarios individuales para controlar el acceso. Esto le permite administrar el acceso de forma centralizada cambiando la pertenencia a un grupo de un usuario o sus atributos una vez, en lugar de tener que actualizar muchas políticas individuales cuando el acceso de un usuario necesita cambiarse.

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

Guía para la implementación

- Si utiliza AWS IAM Identity Center (IAM Identity Center), configure grupos: IAM Identity Center le proporciona la capacidad de configurar grupos de usuarios y asignar a los grupos el nivel deseado de permisos.
  - Inicio de sesión único de AWS: administración de identidades
- Descubra el control de acceso basado en atributos (ABAC): ABAC es una estrategia de autorización que define permisos basados en atributos.
  - ¿Qué es ABAC para AWS?
  - Laboratorio: Control de acceso basado en etiquetas de IAM para EC2

Recursos

Documentos relacionados:

- Introducción a AWS Secrets Manager
- Prácticas recomendadas de IAM
- Proveedores de identidades y federación
- Usuario raíz de una cuenta de AWS

Videos relacionados:

- Prácticas recomendadas para administrar, recuperar y rotar secretos a escala
- Administración de permisos de usuarios a escala con AWS IAM Identity Center
- Dominar la identidad en cada piso de la tarta

Ejemplos relacionados:

- Laboratorio: Control de acceso basado en etiquetas de IAM para EC2

SEGURIDAD 3. ¿Cómo administra los permisos para las personas y las máquinas?

Administre permisos para controlar el acceso a identidades de personas y de máquinas que requieran acceso a AWS y sus cargas de trabajo. Los permisos controlan a qué puede acceder cada usuario y en qué condiciones.

Prácticas recomendadas

- SEC03-BP01 Definir los requisitos de acceso
- SEC03-BP02 Conceder acceso con privilegios mínimos
- SEC03-BP03 Establecer un proceso de acceso de emergencia
- SEC03-BP04 Reducir continuamente los permisos
- SEC03-BP05 Definir las barreras de protección de los permisos para su organización
- SEC03-BP06 Administrar el acceso en función del ciclo de vida
- SEC03-BP07 Analizar el acceso público y entre cuentas
- SEC03-BP08 Compartir recursos de forma segura en su organización
- SEC03-BP09 Compartir recursos de forma segura con terceros

SEC03-BP01 Definir los requisitos de acceso

A cada componente o recurso de su carga de trabajo deben acceder administradores, usuarios finales u otros componentes. Tenga una definición clara de quién o qué debe tener acceso a cada componente, elija el tipo de identidad y el método de autenticación y autorización adecuados.

Patrones comunes de uso no recomendados:

- Codificación rígida o almacenamiento de secretos en la aplicación.
- Concesión de permisos personalizados para cada usuario.
- Uso de credenciales de larga duración.

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

Guía para la implementación

A cada componente o recurso de su carga de trabajo deben acceder administradores, usuarios finales u otros componentes. Tenga una definición clara de quién o qué debe tener acceso a cada componente, elija el tipo de identidad y el método de autenticación y autorización adecuados.

El acceso normal a las Cuentas de AWS en la organización debe proporcionarse mediante acceso federado o un proveedor de identidades centralizado. También debe centralizar su administración de identidades y asegurarse de que existe una práctica establecida para integrar el acceso de AWS al ciclo de vida de los empleados. Por ejemplo, cuando un empleado cambia a un cargo con un nivel de acceso distinto, su pertenencia al grupo también debe cambiar para reflejar sus nuevos requisitos de acceso.

Al definir los requisitos de acceso para las identidades que no son humanas, determine qué aplicaciones y componentes necesitan acceso y cómo se conceden los permisos. El enfoque recomendado es utilizar roles de IAM creados con el modelo de acceso de privilegio mínimo. Las políticas administradas de AWS proporcionan políticas de IAM predefinidas que cubren los casos de uso más comunes.

Los servicios de AWS, como AWS Secrets Manager y AWS Systems Manager Parameter Store, pueden servir para desacoplar los secretos de la aplicación o de la carga de trabajo de forma segura en los casos en los que no es factible utilizar roles de IAM. En Secrets Manager, puede establecer una rotación automática de sus credenciales. Puede utilizar Systems Manager para hacer referencia a los parámetros en sus scripts, comandos, documentos SSM, configuración y flujos de trabajo de automatización con el nombre único que especificó al crear el parámetro.

Puede usar Funciones de AWS Identity and Access Management en cualquier lugar para obtener credenciales de seguridad temporales en IAM para las cargas de trabajo que se ejecutan fuera de AWS. Sus cargas de trabajo puede usar las mismas políticas de IAM y roles de IAM que utiliza con las aplicaciones de AWS para acceder a los recursos de AWS.

Siempre que sea posible, se deben preferir las credenciales temporales a corto plazo en lugar de las credenciales estáticas a largo plazo. En las situaciones en las que necesite usuarios de IAM con acceso programático y credenciales a largo plazo, utilice información de la clave de acceso utilizada por última vez para rotar y retirar las claves de acceso.

Recursos

Documentos relacionados:

- Control de acceso basado en atributos (ABAC)

- AWS IAM Identity Center
- Funciones de IAM en cualquier lugar
- Políticas administradas de AWS para IAM Identity Center
- Condiciones de las políticas de AWS IAM
- Casos de uso de IAM
- Elimine credenciales innecesarias
- Administración de políticas
- How to control access to AWS resources based on Cuenta de AWS, OU, or organization (Cómo controlar el acceso a los recursos de AWS en función de la Cuenta de AWS, la unidad organizativa o la organización)
- Identify, arrange, and manage secrets easily using enhanced search in AWS Secrets Manager (Identificar, organizar y administrar fácilmente los secretos mediante la búsqueda mejorada en AWS Secrets Manager)

Vídeos relacionados:

- Become an IAM Policy Master in 60 Minutes or Less (Consiga dominar las políticas de IAM en 60 minutos o menos)
- Separation of Duties, Least Privilege, Delegation, and CI/CD (Separación de deberes, privilegio mínimo, delegación y CI/CD)
- Streamlining identity and access management for innovation (Optimizar la administración de identidades y accesos para la innovación)

SEC03-BP02 Conceder acceso con privilegios mínimos

Se recomienda conceder exclusivamente el acceso que las identidades necesitan para realizar acciones concretas en recursos específicos en determinadas condiciones. Utilice atributos de grupo y de identidad para configurar dinámicamente los permisos en función de las necesidades en lugar de configurarlos para cada usuario. Por ejemplo, puede conceder acceso a un grupo de desarrolladores para que solamente puedan administrar recursos de su proyecto. De este modo, si un desarrollador abandona el proyecto, su acceso se revoca automáticamente sin cambiar las políticas de acceso subyacentes.

Resultado esperado: los usuarios solo tienen los permisos necesarios para desempeñar su trabajo. A los usuarios solo se les concede acceso a entornos de productos para llevar a cabo una tarea

específica en un periodo de tiempo limitado y el acceso se debe revocar una vez terminada la tarea. Los permisos se deben revocar cuando no se necesiten, por ejemplo, cuando un usuario cambia de proyecto o de puesto. Los privilegios de administrador solo se deben conceder a un pequeño grupo de administradores de confianza. Los permisos se deben revisar periódicamente para evitar su acumulación. A las cuentas de máquinas o sistemas se les debe asignar el conjunto más reducido de permisos que sean necesarios para realizar sus tareas.

Antipatrones usuales:

- Concesión predeterminada de permisos de administrador a los usuarios.
- Uso del usuario raíz para las actividades cotidianas.
- Creación de políticas excesivamente permisivas, pero sin todos los privilegios de administrador.
- No revisar los permisos para averiguar si se les permite el acceso de privilegio mínimo.

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

Guía para la implementación

El principio de privilegio mínimo establece que a las identidades solo les debe permitir realizar el menor conjunto de acciones necesarias para completar una tarea específica. De este modo, se equilibra la facilidad de uso, la eficiencia y la seguridad. Operar según este principio contribuye a limitar el acceso involuntario y a realizar el seguimiento de quién tiene acceso a determinados recursos. Los usuarios y roles de IAM no tienen permisos de forma predeterminada. El usuario raíz tiene acceso total de forma predeterminada y se debe controlar y supervisar de forma estricta. Únicamente se debe usar para tareas que requieran acceso raíz.

Las políticas de IAM se usan para conceder permisos a roles de IAM o recursos específicos. Por ejemplo, las políticas basadas en la identidad se pueden adjuntar a grupos de IAM, mientras que los buckets de S3 se pueden controlar mediante políticas basadas en recursos.

Al crear una política de IAM, puede especificar las acciones de servicio, los recursos y las condiciones que se deben cumplir para que AWS permita o deniegue el acceso. AWS es compatible con una amplia variedad de condiciones que le ayudarán a acotar el acceso. Por ejemplo, mediante la clave de condición `PrincipalOrgID`, puede denegar acciones si el solicitante no forma parte de su organización de AWS.

También puede controlar las solicitudes que realicen los servicios de AWS en su nombre, como que AWS CloudFormation cree una función de AWS Lambda, mediante la clave de condición `CalledVia`. Debe estratificar los diferentes tipos de políticas para establecer una defensa en

profundidad y limitar los permisos generales de sus usuarios. También puede restringir qué permisos se pueden conceder y en qué condiciones. Por ejemplo, puede permitir que sus equipos de aplicaciones creen sus propias políticas de IAM para los sistemas que creen, pero también debe aplicar un límite de permiso para acotar el máximo de permisos que puede recibir el sistema.

Pasos para la implementación

- Implemente políticas de privilegio mínimo: asigne políticas de acceso con privilegio mínimo a grupos y roles de IAM para reflejar el rol o la función del usuario que haya definido.
  - Base las políticas en el uso de la API: una forma de determinar los permisos necesarios consiste en revisar los registros de AWS CloudTrail. Esta revisión le permite crear permisos adaptados a las acciones que el usuario realiza realmente en AWS. IAM Access Analyzer puede generar automáticamente una política de IAM basada en la actividad. Puede usar IAM Access Advisor en el nivel de organización o de cuenta para realizar el seguimiento de la información a la que se ha accedido por última vez para una política concreta.
- Considere la utilización de políticas administradas por AWS para funciones de trabajo. Cuando empiece a crear políticas de permisos detalladas, puede ser difícil saber por dónde empezar. AWS tiene políticas administradas para roles comunes, por ejemplo, facturación, administradores de bases de datos y científicos de datos. Estas políticas pueden servir para limitar el acceso que tienen los usuarios al mismo tiempo que se determina cómo implementar las políticas de privilegio mínimo.
- Elimine los permisos innecesarios: elimine los permisos que no son necesarios y limite las políticas excesivamente permisivas. La generación de políticas de IAM Access Analyzer puede ser de ayuda en la optimización de las políticas de permisos.
- Garantice que los usuarios cuenten con acceso limitado a los entornos de producción: los usuarios solo deben tener acceso a los entornos de producción con un motivo válido. Después de que el usuario lleve a cabo las tareas específicas que requieren el acceso a producción, se debe revocar el acceso. La limitación del acceso a los entornos de producción previene los eventos involuntarios que afectan a la producción y reduce el ámbito de las consecuencias del acceso involuntario.
- Considere el uso de límites de permisos: un límite de permisos es una característica para usar una política administrada que establece los permisos máximos que una política basada en identidad puede conceder a una entidad de IAM. El límite de permisos de una identidad le permite llevar a cabo únicamente las acciones autorizadas tanto por sus políticas basadas en la identidad como por sus límites de permisos.
- Considere el uso de etiquetas de recursos para los permisos: un modelo de control de acceso basado en atributos mediante etiquetas de recursos le permite conceder acceso según la finalidad

del recurso, el propietario, el entorno u otros criterios. Por ejemplo, puede usar etiquetas de recursos para diferenciar entre los entornos de desarrollo y de producción. Con estas etiquetas, puede limitar a los desarrolladores al entorno de desarrollo. Mediante la combinación de las políticas de etiquetado y de permisos, puede conseguir un acceso detallado a los recursos sin necesidad de definir políticas complicadas y personalizadas para cada puesto.

- Use las políticas de control de servicios para AWS Organizations. Las políticas de control de servicios controlan de forma centralizada el máximo de permisos disponibles para las cuentas de los miembros de su organización. Es importante destacar que las políticas de control de servicios le permiten restringir los permisos del usuario raíz en las cuentas de los miembros. Considere también la posibilidad de utilizar AWS Control Tower, que proporciona controles prescriptivos administrados que enriquecen AWS Organizations. También puede definir sus propios controles en Control Tower.
- Establezca una política de ciclo de vida del usuario para la organización: las políticas de este tipo definen las tareas que se realizan cuando los usuarios se incorporan en AWS, cambian de rol o ámbito, o ya no necesitan acceder a AWS. Las revisiones de permisos se deben realizar durante cada paso del ciclo de vida de un usuario para verificar son restrictivos de forma correcta y para evitar la acumulación de permisos.
- Establezca una programación periódica para revisar los permisos y eliminar los que no sean necesarios: debe revisar periódicamente el acceso de usuario para verificar que los usuarios no tengan permisos demasiado permisivos. AWS Config y IAM Access Analyzer pueden ser de ayuda al auditar los permisos de usuario.
- Establezca una matriz de roles de trabajo: con una matriz de roles de trabajo se visualizan los distintos roles y los niveles de acceso necesarios en su presencia de AWS. Con una matriz de roles de trabajo, puede definir y separar los permisos según las responsabilidades de usuario en su organización. Use grupos en vez de aplicar permisos directamente a usuarios o roles individuales.

Recursos

Documentos relacionados:

- Conceder privilegios mínimos
- Límites de permisos para las entidades de IAM
- Techniques for writing least privilege IAM policies (Técnicas para elaborar políticas de IAM de privilegio mínimo)

- IAM Access Analyzer makes it easier to implement least privilege permissions by generating IAM policies based on access activity (IAM Access Analyzer facilita la implementación de los permisos de privilegio mínimo al generar políticas de IAM basadas en la actividad de acceso)
- Delegate permission management to developers by using IAM permissions boundaries (Delegar la administración de permisos para desarrolladores mediante límites de permisos de IAM)
- Refining Permissions using last accessed information (Mejora de los permisos con la información del último acceso)
- IAM policy types and when to use them (Tipos de políticas de IAM y cuándo utilizarlas)
- Testing IAM policies with the IAM policy simulator (Prueba de las políticas de IAM con el simulador de políticas de IAM)
- Guardrails in AWS Control Tower (Barreras de protección en AWS Control Tower)
- Zero Trust architectures: An AWS perspective (Arquitecturas de confianza cero: una perspectiva de AWS)
- How to implement the principle of least privilege with CloudFormation StackSets (Cómo implementar el principio de privilegio mínimo con CloudFormation StackSets)
- Control de acceso basado en atributos (ABAC)
- Reducción del alcance de las políticas con datos de la actividad de los usuarios
- Ver accesos de rol
- Use el etiquetado para organizar el entorno y fomentar la responsabilidad
- Estrategias de etiquetado de AWS
- Etiquetado de recursos de AWS

Vídeos relacionados:

- Next-generation permissions management (Administración de permisos de nueva generación)
- Zero Trust: An AWS perspective (Confianza cero: una perspectiva de AWS)
- How can I use permissions boundaries to limit users and roles to prevent privilege escalation? (¿Cómo puedo utilizar los límites de los permisos para restringir a los usuarios y los roles para evitar la escalada de privilegios?)

Ejemplos relacionados:

- Laboratorio: Límites de permisos de IAM para delegar la creación de roles

- Laboratorio: Control de acceso basado en etiquetas de IAM para EC2

SEC03-BP03 Establecer un proceso de acceso de emergencia

Cree un proceso que permita el acceso de emergencia a sus cargas de trabajo en el caso improbable de que se produzca un problema con su proveedor de identidades centralizado.

Debe diseñar procesos para diferentes modos de error que puedan provocar un evento de emergencia. Por ejemplo, en circunstancias normales, los usuarios de la plantilla se federan en la nube mediante un proveedor de identidades centralizado (SEC02-BP04) para administrar sus cargas de trabajo. Sin embargo, si su proveedor de identidades centralizado no responde o se modifica la configuración de la federación en la nube, es posible que los usuarios de la plantilla no puedan federarse en esta. Un proceso de acceso de emergencia permite a los administradores autorizados acceder a los recursos de la nube a través de medios alternativos (como una forma alternativa de federación o acceso directo de los usuarios) para solucionar problemas con la configuración de la federación o las cargas de trabajo. El proceso de acceso de emergencia se utiliza hasta que se restablezca el mecanismo de federación normal.

Resultado deseado:

- Ha definido y documentado los modos de error que se consideran una emergencia: tenga en cuenta sus circunstancias normales y los sistemas de los que dependen los usuarios para administrar sus cargas de trabajo. Considere cómo cada una de estas dependencias puede no funcionar y provocar una situación de emergencia. Puede que las preguntas y las prácticas recomendadas en el Pilar de fiabilidad le resulten útiles para identificar los modos de error y diseñar sistemas más resilientes para minimizar la probabilidad de que se produzcan errores.
- Ha documentado los pasos que se deben seguir para confirmar que la avería se trata de un caso de emergencia. Por ejemplo, puede solicitar a sus administradores de identidades que comprueben el estado de sus proveedores de identidades principales y en espera y, si ninguno estuviera disponible, declarar un evento de emergencia por error en el proveedor de identidades.
- Ha definido un proceso de acceso de emergencia concreto para cada tipo de modo de emergencia o de error. La especificidad puede reducir la tentación de los usuarios de abusar de un proceso general para todo tipo de emergencias. Sus procesos de acceso de emergencia describen las circunstancias en las que se debe utilizar cada proceso y, por otra parte, las situaciones en las que no se debe utilizar el proceso y señala los procesos alternativos que podrían aplicarse.
- Sus procesos están bien documentados con instrucciones detalladas y guías de estrategia que se pueden seguir de forma rápida y eficiente. Recuerde que un evento de emergencia puede resultar

estresante para sus usuarios, ya que que pueden estar sometidos a una fuerte presión de plazos, por lo que debe diseñar su proceso de la manera más sencilla posible.

Patrones comunes de uso no recomendados:

- No tiene procesos de acceso de emergencia bien documentados y ensayados. Sus usuarios no están preparados para emergencias y siguen procesos improvisados cuando estas se producen.
- Sus procesos de acceso de emergencia dependen de los mismos sistemas (como un proveedor de identidades centralizado) que sus mecanismos de acceso normales. Esto significa que el error de un sistema de este tipo podría afectar tanto a sus mecanismos de acceso normales como a los de emergencia y repercutir en su capacidad para recuperarse del error.
- Sus procesos de acceso de emergencia se utilizan en situaciones que no son de emergencia. Por ejemplo, los usuarios suelen hacer un uso inapropiado de los procesos de acceso de emergencia, ya que les resulta más fácil realizar cambios directamente que enviarlos a través de una canalización.
- Sus procesos de acceso de emergencia no generan registros suficientes para auditar los procesos, o los registros no se supervisan para alertar de un posible uso indebido de los procesos.

Beneficios de establecer esta práctica recomendada:

- Si cuenta con procesos de acceso de emergencia bien documentados y ensayados, puede reducir el tiempo que tardan los usuarios en responder y resolver un evento de emergencia. Esto puede reducir el tiempo de inactividad y aumentar la disponibilidad de los servicios que presta a sus clientes.
- Puede realizar un seguimiento de cada solicitud de acceso de emergencia y detectar y alertar sobre intentos no autorizados de utilizar indebidamente el proceso para eventos que no sean de emergencia.

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

Guía para la implementación

Esta sección proporciona guías para crear procesos de acceso de emergencia para varios modos de error relacionados con las cargas de trabajo desplegadas en AWS, comenzando con una guía común que se aplica a todos los modos de error y siguiendo con una guía específica basada en el tipo de modo de error.

Guía común para todos los modos de error

Tenga en cuenta lo siguiente al diseñar un proceso de acceso de emergencia para un modo de error:

- Documente las condiciones previas y los supuestos del proceso, es decir, cuándo el proceso debe o no debe aplicarse. Esto ayuda a detallar el modo de error y a documentar los supuestos, como el estado de otros sistemas relacionados. Por ejemplo, el proceso del modo de error 2 supone que el proveedor de identidades está disponible, pero que la configuración activada en AWS se ha modificado o ha caducado.
- Cree de antemano los recursos necesarios para el proceso de acceso de emergencia (SEC10-BP05). Por ejemplo, cree de antemano el acceso de emergencia a la Cuenta de AWS con roles y IAM users, y los roles de IAM entre cuentas en todas las cuentas de la carga de trabajo. Esto asegura que estos recursos estén listos y disponibles cuando ocurra una emergencia. Al crear de antemano los recursos, no depende de las API del plano de control de AWS (utilizadas para crear y modificar los recursos de AWS) que podrían no estar disponibles en caso de emergencia. Además, al crear de antemano los recursos de IAM, no es necesario tener en cuenta los posibles retrasos debido a una coherencia eventual.
- Incluya los procesos de acceso de emergencia como parte de sus planes de administración de incidentes (SEC10-BP02). Documente cómo se realiza el seguimiento de los eventos de emergencia y cómo se comunican a otros miembros de su organización, como los equipos de compañeros o la dirección y, cuando corresponda, externamente a sus clientes y socios comerciales.
- Defina el proceso de solicitud de acceso de emergencia en su sistema de flujo de trabajo de solicitudes de servicio existente, si dispone de uno. Por lo general, estos sistemas de flujo de trabajo le permiten crear formularios de entrada para recopilar información sobre la solicitud, realizar un seguimiento de la solicitud en cada etapa del flujo de trabajo y añadir pasos de aprobación automatizados y manuales. Relacione cada solicitud con el correspondiente evento de emergencia registrado en su sistema de administración de incidentes. Disponer de un sistema uniforme para los accesos de emergencia le permite realizar un seguimiento de esas solicitudes en un solo sistema, analizar las tendencias de uso y mejorar sus procesos.
- Compruebe que solo los los usuarios autorizados puedan iniciar los procesos de acceso de emergencia y que estos procesos requieran la aprobación de los compañeros del usuario o de la dirección, según corresponda. El proceso de aprobación debe funcionar de manera eficaz tanto dentro como fuera del horario laboral. Defina cómo las solicitudes de aprobación admiten aprobadores secundarios si los principales no están disponibles y cómo se escalan en la cadena de administración hasta la aprobación.

- Compruebe que el proceso genere registros y eventos de auditoría detallados para los intentos correctos e infructuosos de obtener acceso de emergencia. Supervise tanto el proceso de solicitud como el mecanismo de acceso de emergencia para detectar el uso indebido o los accesos no autorizados. Correlacione la actividad con los eventos de emergencia en curso de su sistema de administración de incidentes y alerte cuando se produzcan acciones fuera de los períodos de tiempo esperados. Por ejemplo, debe supervisar y alertar si se produce actividad en la Cuenta de AWS de acceso de emergencia, ya que nunca debe usarse en operaciones normales.
- Pruebe los procesos de acceso de emergencia de manera periódica para comprobar que los pasos estén claros y para garantizar el nivel de acceso correcto de manera rápida y eficiente. Sus procesos de acceso de emergencia deben probarse como parte de las simulaciones de respuesta ante incidentes (SEC10-BP07) y pruebas de recuperación de desastres (REL13-BP03).

Modo de error 1: el proveedor de identidades utilizado para federarse en AWS no está disponible

Como se describe en SEC02-BP04 Recurrir a un proveedor de identidades centralizado, recomendamos confiar en un proveedor de identidades centralizado para federar a los usuarios de su plantilla y concederles acceso a las Cuentas de AWS. Puede federar varias Cuentas de AWS en su organización de AWS con IAM Identity Center, o puede federar Cuentas de AWS individuales con IAM. En ambos casos, los usuarios de la plantilla se autentican con su proveedor de identidades centralizado antes de que se les redirija a un punto de conexión de inicio de sesión de AWS para el inicio de sesión único.

En el caso poco probable de que su proveedor de identidades centralizado no esté disponible, los usuarios de la plantilla no podrán federarse en las Cuentas de AWS ni administrar sus cargas de trabajo. En este caso de emergencia, puede proporcionar un proceso de acceso de emergencia para que un pequeño grupo de administradores acceda a las Cuentas de AWS con el fin de realizar tareas cruciales que no puedan esperar a que sus proveedores de identidades centralizados vuelvan a estar disponibles. Por ejemplo, su proveedor de identidades no estará disponible durante 4 horas y, durante ese período, necesita modificar los límites superiores de un grupo de Amazon EC2 Auto Scaling en una cuenta de producción para gestionar un aumento inesperado en el tráfico de clientes. Los administradores de emergencias deben seguir el proceso de acceso de emergencia para acceder a la Cuenta de AWS de producción específica y realizar los cambios necesarios.

El proceso de acceso de emergencia se basa en una Cuenta de AWS de acceso de emergencia creada de antemano que se utiliza únicamente para el acceso de emergencia y dispone de recursos de AWS (como roles de IAM y IAM users) para respaldar el proceso de acceso de emergencia. Durante las operaciones normales, nadie debe acceder a la cuenta de acceso de emergencia y

usted debe supervisar y alertar sobre el uso indebido de esta cuenta (para obtener más información, consulte la sección anterior de guía común).

La cuenta de acceso de emergencia tiene roles de IAM de acceso de emergencia con permisos para asumir roles entre cuentas en las Cuentas de AWS que requieran acceso de emergencia. Estos roles de IAM se crean de antemano y se configuran con políticas de confianza que confían en los roles de IAM de la cuenta de emergencia.

El proceso de acceso de emergencia puede utilizar uno de los siguientes enfoques:

- Puede crear de antemano un conjunto de IAM users para los administradores de emergencias de la cuenta de acceso de emergencia con contraseñas seguras y tokens de MFA asociados. Estos IAM users tienen permisos para asumir los roles de IAM que, entonces, permiten el acceso entre cuentas a la Cuenta de AWS donde se requiere el acceso de emergencia. Recomendamos crear el menor número posible de usuarios y asignar cada usuario a un único administrador de emergencias. Durante una emergencia, un usuario administrador de emergencias inicia sesión en la cuenta de acceso de emergencia con su contraseña y el código de token de MFA, cambia el rol de IAM de acceso de emergencia en la cuenta de emergencia y, finalmente, cambia el rol de IAM de acceso de emergencia en la cuenta de carga de trabajo para realizar la acción de cambio de emergencia. La ventaja de este enfoque es que cada IAM user se asigna a un administrador de emergencias y usted puede saber qué usuario inició sesión revisando los eventos de CloudTrail. La desventaja es que hay que mantener varios IAM users con sus contraseñas de larga duración y los tokens de MFA asociados.
- Puede utilizar el usuario raíz de la Cuenta de AWS de acceso de emergencia para iniciar sesión en la cuenta de acceso de emergencia, asumir el rol de IAM de acceso de emergencia y asumir el rol entre cuentas en la cuenta de carga de trabajo. Recomendamos configurar una contraseña segura y varios tokens de MFA para el usuario raíz. También recomendamos almacenar la contraseña y los tokens de MFA en un almacén de credenciales empresarial seguro que aplique una autenticación y una autorización sólidas. Debe proteger los factores de restablecimiento de la contraseña y el token de MFA. Para ello, establezca la dirección de correo electrónico de la cuenta en una lista de distribución de correo electrónico supervisada por los administradores de seguridad en la nube y el número de teléfono de la cuenta en un número de teléfono compartido también supervisado por los administradores de seguridad. La ventaja de este enfoque es que solo hay que administrar un conjunto de credenciales de usuario raíz. La desventaja es que, dado que se trata de un usuario compartido, es posible que varios administradores inicien sesión como usuario raíz. Debe auditar los eventos de registro del almacén empresarial para identificar qué administrador extrajo la contraseña del usuario raíz.

Modo de error 2: la configuración del proveedor de identidades en AWS se ha modificado o ha caducado

Para permitir que los usuarios de la plantilla se federen en Cuentas de AWS, puede configurar el IAM Identity Center con un proveedor de identidades externo o crear un proveedor de identidades de IAM (SEC02-BP04). Por lo general, se configuran importando un documento XML de metadatos de SAML proporcionado por el proveedor de identidades. El documento XML de metadatos incluye un certificado X.509 correspondiente a una clave privada que el proveedor de identidades utiliza para firmar sus aserciones SAML.

Un administrador podría modificar o eliminar estas configuraciones de AWS de forma accidental. En otro escenario, el certificado X.509 importado a AWS podría caducar cuando aún no se ha importado a AWS un nuevo XML de metadatos con un certificado nuevo. Ambos escenarios pueden desbaratar la federación a AWS de los usuarios de la plantilla y provocar una emergencia.

En un caso de emergencia de este tipo, puede proporcionar a sus administradores de identidades acceso a AWS para solucionar los problemas de federación. Por ejemplo, el administrador de identidades utiliza el proceso de acceso de emergencia para iniciar sesión en la Cuenta de AWS de acceso de emergencia, cambia a un rol en la cuenta de administrador del centro de identidades y actualiza la configuración del proveedor de identidades externo importando el último documento XML de metadatos SAML de su proveedor de identidades para volver a habilitar la federación. Una vez que se corrija la federación, los usuarios de la plantilla seguirán utilizando el proceso operativo normal para federarse en sus cuentas de carga de trabajo.

Puede seguir los enfoques detallados en el modo de error 1 anterior para crear un proceso de acceso de emergencia. Puede conceder permisos con privilegios mínimos a sus administradores de identidades para que accedan únicamente a la cuenta de administrador del centro de identidades y realicen acciones en el centro de identidades en esa cuenta.

Modo de error 3: interrupción del centro de identidades

En el caso poco probable de que se produzca una interrupción en un IAM Identity Center o en una Región de AWS, le recomendamos que establezca una configuración que pueda utilizar para proporcionar acceso temporal a la Consola de administración de AWS.

El proceso de acceso de emergencia utiliza la federación directa desde su proveedor de identidades a IAM en una cuenta de emergencia. Para obtener información detallada sobre el proceso y las consideraciones de diseño, consulte la sección sobre la configuración del acceso de emergencia a la Consola de administración de AWS.

Pasos para la implementación

Pasos comunes para todos los modos de error

- Cree una Cuenta de AWS dedicada a los procesos de acceso de emergencia. Cree de antemano los recursos de IAM necesarios en la cuenta, como roles de IAM o IAM users, y opcionalmente, proveedores de identidades de IAM. Además, cree de antemano roles de IAM entre cuentas en la Cuentas de AWS de la carga de trabajo con relaciones de confianza con los roles de IAM correspondientes en la cuenta de acceso de emergencia. Puede usar el CloudFormation StackSets con AWS Organizations para crear dichos recursos en las cuentas de los miembros de su organización.

- Cree políticas de control de servicios (SCP) de AWS Organizations para denegar la eliminación y modificación de los roles de IAM entre cuentas en las Cuentas de AWS miembro.
- Habilite CloudTrail para la Cuenta de AWS de acceso de emergencia y envíe los eventos de ruta a un bucket de S3 central en su Cuenta de AWS de recopilación de registros. Si utiliza AWS Control Tower para configurar y gobernar su entorno multicuenta de AWS, cada cuenta que cree con AWS Control Tower o inscriba en AWS Control Tower tendrá CloudTrail habilitado de forma predeterminada y se enviará a un bucket de S3 en una Cuenta de AWS de archivo de registro dedicada.
- Supervise la actividad de la cuenta de acceso de emergencia mediante la creación de reglas de EventBridge que concuerden con el inicio de sesión de la consola y la actividad de la API por parte de los roles de IAM de emergencia. Envíe notificaciones a su centro de operaciones de seguridad cuando se produzca actividad fuera de un evento de emergencia continuo registrado en su sistema de administración de incidentes.

Pasos adicionales para el modo de error 1: el proveedor de identidades utilizado para federarse en AWS no está disponible y el modo de error 2: la configuración del proveedor de identidades en AWS se ha modificado o ha caducado

- Cree de antemano los recursos en función del mecanismo que elija para el acceso de emergencia:
  - Con IAM users: cree de antemano los IAM users con contraseñas seguras y los dispositivos MFA asociados.
  - Con el usuario raíz de la cuenta de emergencia: configure el usuario raíz con una contraseña segura y almacene la contraseña en el almacén de credenciales de su empresa. Asocie varios

dispositivos MFA físicos al usuario raíz y almacene los dispositivos en lugares a los que puedan acceder rápidamente los miembros de su equipo de administradores de emergencias.

Pasos adicionales para el modo de error 3: interrupción del centro de identidades

- Como se detalla en la configuración del acceso de emergencia a la Consola de administración de AWS, en la Cuenta de AWS de acceso de emergencia, cree un proveedor de identidades de IAM para habilitar la federación SAML directa desde su proveedor de identidades.
- Cree grupos de operaciones de emergencia en su IdP sin miembros.
- Cree los roles de IAM correspondientes a los grupos de operaciones de emergencia en la cuenta de acceso de emergencia.

Recursos

Prácticas recomendadas por Well-Architected:

- SEC02-BP04 Recurrir a un proveedor de identidades centralizado
- SEC03-BP02 Conceder acceso con privilegios mínimos
- SEC10-BP02: Desarrollar planes de administración de incidentes
- SEC10-BP07 Ejecutar los días de juego

Documentos relacionados:

- configuración del acceso de emergencia a la Consola de administración de AWS
- Concesión de acceso a la Consola de administración de AWS a los usuarios federados SAML 2.0
- Break glass access

Vídeos relacionados:

- AWS re:Invent 2022 - Simplify your existing workforce access with IAM Identity Center
- AWS re:Inforce 2022 - AWS Identity and Access Management (IAM) deep dive

Ejemplos relacionados:

- AWS Break Glass Role

- AWS customer playbook framework
- AWS incident response playbook samples

SEC03-BP04 Reducir continuamente los permisos

A medida que los equipos determinen qué acceso es necesario, elimine los permisos innecesarios y establezca procesos de revisión para conseguir permisos con privilegios mínimos. Supervise y elimine continuamente las identidades y los permisos que no se utilicen, tanto para el acceso humano como para el de las máquinas.

Resultado deseado: las políticas de permisos deben cumplir el principio del privilegio mínimo. A medida que se definan mejor las responsabilidades y los roles del trabajo, debe revisar sus políticas de permisos para eliminar los permisos innecesarios. Este enfoque reduce el alcance del impacto en caso de que las credenciales se expongan de forma inadvertida o se acceda a ellas sin autorización.

Antipatrones usuales:

- Conceder de forma predeterminada permisos de administrador a los usuarios.
- Crear políticas excesivamente permisivas, pero sin todos los privilegios de administrador.
- Mantener políticas de permisos después de que ya no son necesarias.

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

Guía para la implementación

Cuando los equipos y los proyectos están dando sus primeros pasos, utilizar unas políticas de permisos permisivas sirve para fomentar la innovación y la agilidad. Por ejemplo, en un entorno de desarrollo o de pruebas, se puede dar acceso a los desarrolladores a un amplio conjunto de servicios de AWS. Recomendamos que evalúe el acceso continuamente y lo restrinja únicamente a aquellos servicios y acciones de servicio que sean necesarios para realizar el trabajo actual. Recomendamos realizar esta evaluación tanto para las identidades humanas como para las de máquina. Las identidades de máquina, que a veces se denominan cuentas del sistema o del servicio, son identidades que dan acceso a AWS a aplicaciones o servidores. Este acceso es especialmente importante en un entorno de producción, donde unos permisos demasiado permisivos pueden tener un impacto enorme y el potencial de exponer los datos de los clientes.

AWS tiene numerosos métodos para ayudar a identificar a los usuarios, roles, permisos y credenciales no utilizados. AWS también puede ayudar a analizar la actividad de acceso de los

usuarios y roles de IAM, incluidas las claves de acceso asociadas, y el acceso a recursos de AWS, como los objetos de los buckets de Amazon S3. La generación de políticas de AWS Identity and Access Management Access Analyzer puede ayudarle a crear políticas de permisos restrictivas basadas en los servicios y acciones reales con los que interactúa una entidad principal. El control de acceso basado en atributos (ABAC) puede ayudar a simplificar la administración de permisos, ya que le permite proporcionar permisos a los usuarios utilizando sus atributos en lugar de tener que asociar políticas de permisos directamente a cada usuario.

Pasos para la implementación

- Utilice AWS Identity and Access Management Access Analyzer: IAM Access Analyzer le ayuda a identificar recursos de su organización y sus cuentas, como buckets de Amazon Simple Storage Service (Amazon S3) o roles de IAM, que se comparten con una entidad externa.
- Utilice la generación de políticas de IAM Access Analyzer: la generación de políticas de IAM Access Analyzer le ayuda a crear políticas de permisos detalladas basadas en la actividad de acceso de un usuario o rol de IAM.
- Determine un marco temporal y una política de uso aceptables para los usuarios y roles de IAM: utilice la marca de tiempo del último acceso para identificar a los usuarios y roles no utilizados y eliminarlos. Revise la información de último acceso a servicios y acciones para identificar y delimitar los permisos de usuarios y roles específicos. Por ejemplo, puede utilizar la información sobre el último acceso para identificar las acciones específicas de Amazon S3 necesarias para el rol de su aplicación y restringir el acceso únicamente a dichas acciones. Estas características de información sobre el último acceso están disponibles en la Consola de administración de AWS y de manera programática para permitirle incorporarlas en sus flujos de trabajo de infraestructura y sus herramientas automatizadas.
- Considere la posibilidad de registrar eventos de datos en AWS CloudTrail: de manera predeterminada, CloudTrail no registra eventos de datos como la actividad a nivel de objeto de Amazon S3 (por ejemplo, `GetObject` y `DeleteObject`) o las actividades de tabla de Amazon DynamoDB (por ejemplo, `PutItem` y `DeleteItem`). Considere la posibilidad de habilitar el registro de estos eventos para determinar qué usuarios y roles necesitan acceder a objetos de Amazon S3 o elementos de tabla de DynamoDB específicos.

Recursos

Documentos relacionados:

- Conceder privilegios mínimos

- Elimine credenciales innecesarias
- ¿Qué es AWS CloudTrail?
- Administración de políticas
- Registro y monitoreo en DynamoDB
- Habilitación del registro de eventos de CloudTrail para buckets y objetos de Amazon S3
- Obtener informes de credenciales para su Cuenta de AWS

Vídeos relacionados:

- Become an IAM Policy Master in 60 Minutes or Less (Consiga dominar las políticas de IAM en 60 minutos o menos)
- Separation of Duties, Least Privilege, Delegation, and CI/CD (Separación de deberes, privilegio mínimo, delegación y CI/CD)
- AWS re:Inforce 2022 - AWS Identity and Access Management (IAM) deep dive (AWS re:Inforce 2022: Profundización en AWS Identity and Access Management [IAM])

SEC03-BP05 Definir las barreras de protección de los permisos para su organización

Establezca controles comunes que restrinjan el acceso a todas las identidades de su organización. Por ejemplo, puede restringir el acceso a determinadas Regiones de AWS o impedir que sus operadores eliminen recursos comunes, como un rol de IAM que usa su equipo de seguridad central.

Patrones comunes de uso no recomendados:

- Ejecutar cargas de trabajo en su cuenta de administrador de la organización.
- Ejecutar cargas de trabajo de producción y de no producción en la misma cuenta.

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

Guía para la implementación

A medida que crezca y administre cargas de trabajo adicionales en AWS, deberá separarlas mediante cuentas y administrar dichas cuentas con AWS Organizations. Le recomendamos que establezca barreras de protección de permisos comunes que restrinjan el acceso a todas las identidades de su organización. Por ejemplo, puede restringir el acceso a determinadas Regiones de

AWS o impedir que su equipo elimine recursos comunes, como un rol de IAM que usa su equipo de seguridad central.

Puede empezar con la implementación de ejemplos de políticas de control de servicios, como impedir que los usuarios desactiven servicios clave. Las SCP utilizan el lenguaje de las políticas de IAM y le permiten establecer controles a los que se adhieren todas las entidades principales de IAM (usuarios y roles). Puede restringir el acceso a determinadas acciones de servicio, recursos y según una condición específica para satisfacer las necesidades de control de acceso de su organización. Si es necesario, puede definir excepciones a sus barreras de protección. Por ejemplo, puede restringir las acciones de servicio para todas las entidades de IAM de la cuenta excepto para un rol de administrador específico.

Le recomendamos que evite ejecutar cargas de trabajo en su cuenta de administración. Esta cuenta debe utilizarse para controlar y desplegar las barreras de protección que afectarán a las cuentas de los miembros. Algunos servicios de AWS admiten el uso de una cuenta de administrador delegada. Cuando esté disponible, deberá utilizar esta cuenta delegada en lugar de la cuenta de administración. Debe limitar firmemente el acceso a la cuenta de administrador de la organización.

El uso de una estrategia de varias cuentas le permite tener una mayor flexibilidad a la hora de aplicar las barreras de protección a sus cargas de trabajo. La Arquitectura de referencia de Seguridad de AWS ofrece recomendaciones sobre cómo diseñar la estructura de su cuenta. Los servicios de AWS como AWS Control Tower proporcionan capacidades para administrar de forma centralizada tanto los controles preventivos como los de detección en toda la organización. Defina un propósito claro para cada cuenta o unidad organizativa en su organización y limite los controles de acuerdo con dicho propósito.

Recursos

Documentos relacionados:

- AWS Organizations
- Service control policies (SCPs) (Políticas de control de servicios [SCP])
- Get more out of service control policies in a multi-account environment (Saque más partido a las políticas de control del servicio en un entorno de varias cuentas)
- Arquitectura de referencia de AWS (AWS SRA)

Vídeos relacionados:

- Enforce Preventive Guardrails using Service Control Policies (Aplicar las barreras de protección preventivas mediante políticas de control de servicios)
- Building governance at scale with AWS Control Tower (Consolidar la gobernanza a escala con AWS Control Tower)
- AWS Identity and Access Management deep dive (Profundización en AWS Identity and Access Management)

SEC03-BP06 Administrar el acceso en función del ciclo de vida

Integre los controles de acceso con el ciclo de vida de la aplicación y el operador, el proveedor de federación centralizado. Por ejemplo, quite el acceso a un usuario cuando abandone la organización o cambie de rol.

A medida que vaya administrando las cargas de trabajo con cuentas independientes, habrá casos en los que necesite compartir recursos entre esas cuentas. Le recomendamos que comparta los recursos con AWS Resource Access Manager (AWS RAM). Este servicio le permite compartir de forma sencilla y segura recursos de AWS en su AWS Organizations y las unidades organizativas. Con AWS RAM, el acceso a los recursos compartidos se concede o revoca automáticamente a medida que las cuentas entran y salen de la organización o unidad organizativa con la que se comparten. Esto ayuda a garantizar que los recursos se comparten solo con las cuentas que pretende.

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

Guía para la implementación

Ciclo de vida de acceso de los usuarios Implemente una política de ciclo de vida de acceso de los usuarios para la incorporación de nuevos usuarios, los cambios de puesto y la salida de usuarios, de modo que solo tengan acceso los usuarios actuales.

Recursos

Documentos relacionados:

- Control de acceso basado en atributos (ABAC)
- Conceder privilegios mínimos
- Analizador de acceso de IAM
- Elimine credenciales innecesarias
- Administración de políticas de IAM

Vídeos relacionados:

- Become an IAM Policy Master in 60 Minutes or Less (Consiga dominar las políticas de IAM en 60 minutos o menos)
- Separation of Duties, Least Privilege, Delegation, and CI/CD (Separación de deberes, privilegio mínimo, delegación y CI/CD)

SEC03-BP07 Analizar el acceso público y entre cuentas

Supervise continuamente los resultados que ponen de relieve el acceso público y entre cuentas. Reduzca el acceso público y el acceso entre cuentas solo a los recursos que requieran ese acceso.

Resultado deseado: saber cuáles de sus recursos de AWS se comparten y con quién. Supervisar y auditar continuamente sus recursos compartidos para verificar que solo se compartan con las entidades principales autorizadas.

Antipatrones usuales:

- No mantener un inventario de los recursos compartidos.
- No seguir un proceso para aprobar el acceso público o entre cuentas a los recursos.

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

Guía para la implementación

Si su cuenta pertenece a AWS Organizations, puede conceder acceso a los recursos a toda la organización, a unidades organizativas específicas o a cuentas individuales. Si su cuenta no es miembro de una organización, puede compartir recursos con cuentas individuales. Puede conceder acceso directo entre cuentas utilizando políticas basadas en recursos —por ejemplo, políticas de buckets de Amazon Simple Storage Service (Amazon S3)—- o permitiendo que una entidad principal de otra cuenta asuma un rol de IAM en su cuenta. Cuando utilice políticas de recursos, compruebe que solo se concede acceso a las entidades principales autorizadas. Defina un proceso para aprobar todos los recursos que deban estar disponibles públicamente.

AWS Identity and Access Management Access Analyzer utiliza la seguridad comprobable para identificar todas las rutas de acceso a un recurso desde fuera de su cuenta. Revisa continuamente las políticas de recursos e informa de los resultados del acceso público y entre cuentas para facilitarle el análisis de un acceso potencialmente amplio. Considere la posibilidad de configurar IAM

Access Analyzer con AWS Organizations para comprobar que tiene visibilidad de todas sus cuentas. IAM Access Analyzer también le permite previsualizar los resultados antes de desplegar los permisos de recursos. Esto le permite validar que sus cambios de política conceden solo el acceso público y entre cuentas previsto a sus recursos. Al diseñar el acceso de varias cuentas, puede utilizar políticas de confianza para controlar en qué casos se puede asumir un rol. Por ejemplo, podría utilizar la clave de condición `PrincipalOrgId` para denegar un intento de asumir un rol desde fuera de su AWS Organizations.

AWS Config puede informar de los recursos que están mal configurados y, a través de las comprobaciones de políticas de AWS Config, puede detectar los recursos que tienen configurado el acceso público. Servicios como AWS Control Tower y AWS Security Hub CSPM simplifican el despliegue de controles de detección y barreras de protección en AWS Organizations para identificar y corregir los recursos expuestos públicamente. Por ejemplo,AWS Control Tower dispone de una barrera de protección administrada que puede detectar si las Cuentas de AWS pueden restaurar alguna instantánea de Amazon EBS.

Pasos para la implementación

- Considere la posibilidad de habilitar AWS Config para AWS Organizations: AWS Config le permite agregar los hallazgos de varias cuentas que están dentro de una AWS Organizations a una cuenta de administrador delegado. Esto proporciona una visión global y le permite desplegar Reglas de AWS Config en todas las cuentas para detectar recursos de acceso público.
- Configure AWS Identity and Access Management Access Analyzer: IAM Access Analyzer le ayuda a identificar los recursos y cuentas de su organización, como los buckets de Amazon S3 o los roles de IAM, que se comparten con una entidad externa.
- Utilice la corrección automatizada en AWS Config para responder a los cambios en la configuración del acceso público de los buckets de Amazon S3: puede volver a habilitar automáticamente la configuración de acceso público en bloque para los buckets de Amazon S3.
- Implemente la supervisión y las alertas para identificar si los buckets de Amazon S3 se han hecho públicos: debe disponer de supervisión y alertas para identificar cuándo se desactiva el acceso público a bloques de Amazon S3 y si los buckets de Amazon S3 se hacen públicos. Además, si utiliza AWS Organizations, puede crear una política de control de servicios que impida realizar cambios en las políticas de acceso público de Amazon S3. AWS Trusted Advisor comprueba si hay buckets de Amazon S3 que tengan permisos de acceso abierto. Los permisos del bucket que otorgan, suben o eliminan el acceso para todo el mundo crean posibles vulnerabilidades de seguridad, ya que permiten que cualquiera añada, modifique o elimine elementos en un bucket. La comprobación de Trusted Advisor examina los permisos explícitos del bucket y las políticas

asociadas que podrían anular los permisos del bucket. También puede utilizar AWS Config para supervisar sus buckets de Amazon S3 para comprobar si tienen acceso público. Para obtener más información, consulte How to Use AWS Config to Monitor for and Respond to Amazon S3 Buckets Allowing Public Access (Cómo utilizar AWS Config para supervisar y responder a los buckets de Amazon S3 que permiten el acceso público). Al revisar el acceso, es importante tener en cuenta qué tipos de datos contienen los buckets de Amazon S3. Amazon Macie ayuda a detectar y proteger datos confidenciales, como PII, PHI y credenciales, además de claves privadas o de AWS.

Recursos

Documentos relacionados:

- Uso de AWS Identity and Access Management Access Analyzer
- AWS Control Tower controls library (Biblioteca de controles de AWS Tower Control)
- AWS Foundational Security Best Practices standard (Estándar de prácticas recomendadas de seguridad básicas de AWS)
- AWS Config Managed Rules (Reglas administradas de AWS Config)
- Referencia de verificaciones de AWS Trusted Advisor
- Supervisión de resultados de la verificación de AWS Trusted Advisor con Amazon EventBridge
- Managing AWS Config Rules Across All Accounts in Your Organization (Administración de reglas de AWS Config en todas las cuentas de su organización)
- AWS Config y AWS Organizations

Vídeos relacionados:

- Best Practices for securing your multi-account environment (Prácticas recomendadas para proteger su entorno de varias cuentas)
- Dive Deep into IAM Access Analyzer (Profundización en IAM Access Analyzer)

SEC03-BP08 Compartir recursos de forma segura en su organización

A medida que el número de cargas de trabajo va aumentando, es posible que necesite compartir el acceso a los recursos de esas cargas de trabajo o aprovisionar los recursos varias veces entre varias cuentas. Es posible que disponga de componentes para compartimentar el entorno, por ejemplo, en

entornos de desarrollo, pruebas y producción. Sin embargo, disponer de componentes de separación no le impide compartir de forma segura. Al compartir componentes que se solapan, puede reducir la sobrecarga operativa y conseguir una experiencia uniforme sin tener que adivinar qué podría haber pasado por alto al crear el mismo recurso varias veces.

Resultado deseado: reducir al mínimo el acceso involuntario mediante métodos seguros para compartir recursos dentro de su organización y facilitar su iniciativa de prevención de pérdida de datos. Reducir la sobrecarga operativa en comparación con la administración de componentes individuales, reducir los errores derivados de crear manualmente el mismo componente varias veces y aumentar la escalabilidad de las cargas de trabajo. Puede disminuir el tiempo de resolución en escenarios con varios puntos de fallo y aumentar su confianza a la hora de determinar cuándo un componente ya no es necesario. Para obtener orientación prescriptiva sobre el análisis de recursos que se comparten externamente, consulte SEC03-BP07 Analizar el acceso público y entre cuentas.

Antipatrones usuales:

- Falta de un proceso para supervisar continuamente y alertar automáticamente sobre un uso compartido externo inesperado.
- Falta de una referencia sobre lo que se debe compartir y lo que no.
- Adoptar de manera predeterminada una política muy abierta en lugar de compartir explícitamente cuando es necesario.
- Crear manualmente recursos fundamentales que se solapan cuando es necesario.

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

Guía para la implementación

Diseñe sus controles y patrones de acceso para que rijan el consumo de recursos compartidos de forma segura y solo con entidades de confianza. Supervise los recursos compartidos y revise el acceso a ellos de forma continua; además, reciba alertas sobre un uso compartido inapropiado o inesperado. Revise Analizar el acceso público y entre cuentas para ayudarle a establecer una gobernanza que reduzca el acceso externo solo a los recursos que lo requieran, además de a establecer un proceso para supervisar continuamente y alertar automáticamente.

El uso compartido entre cuentas dentro de AWS Organizations está respaldado por una serie de servicios de AWS, como AWS Security Hub CSPM, Amazon GuardDuty y AWS Backup. Estos servicios permiten compartir datos con una cuenta central, acceder a ellos desde una cuenta central o administrar recursos y datos desde una cuenta central. Por ejemplo, AWS Security Hub CSPM

puede transferir hallazgos desde cuentas individuales a una cuenta central en la que podrá verlos todos. AWS Backup puede realizar una copia de seguridad de un recurso y compartirlo entre varias cuentas. Puede utilizar AWS Resource Access Manager (AWS RAM) para compartir otros recursos comunes, como subredes de VPC y asociaciones de Transit Gateway, AWS Network Firewall o canalizaciones de Amazon SageMaker AI.

Para limitar su cuenta para que solo comparta recursos dentro de su organización, utilice políticas de control de servicios (SCP) para impedir el acceso a las entidades principales externas. Cuando comparta recursos, combine controles basados en identidades y controles de red para crear un perímetro de datos para su organización que le ayude a protegerse contra el acceso no intencionado. Un perímetro de datos es un conjunto de barreras de protección preventivas para ayudar a verificar que solo sus identidades de confianza accedan a los recursos de confianza desde las redes previstas. Estos controles ponen límites apropiados a los recursos que se pueden compartir y evitan que se compartan o expongan recursos que no deberían permitirse. Por ejemplo, como parte de su perímetro de datos, puede utilizar políticas de punto de conexión de VPC y la condición `AWS:PrincipalOrgId` para asegurarse de que las identidades que acceden a sus buckets de Amazon S3 pertenecen a su organización. Es importante tener en cuenta que los SCP no se aplican a los roles vinculados al servicio (LSR) ni a las entidades principales del servicio de AWS.

Cuando utilice Amazon S3, deshabilite las ACL para su bucket de Amazon S3 y utilice las políticas de IAM para definir el control de acceso. Para restringir el acceso a un origen de Amazon S3 desde Amazon CloudFront, migre de la identidad de acceso de origen (OAI) al control de acceso de origen (OAC), que admite características adicionales como el cifrado del servidor con AWS Key Management Service.

En algunos casos, es posible que desee permitir compartir recursos fuera de su organización o conceder a un tercero acceso a sus recursos. Para obtener orientación prescriptiva sobre la administración de permisos para compartir recursos externamente, consulte Administración de permisos.

Pasos para la implementación

1. Use AWS Organizations.

   AWS Organizations es un servicio de administración de cuentas que le permite consolidar varias Cuentas de AWS en una organización que usted crea y administra de manera centralizada. Puede agrupar sus cuentas en unidades organizativas (OU) y asociar diferentes políticas a cada OU para ayudarle a satisfacer sus necesidades presupuestarias, de seguridad y de conformidad. También puede controlar cómo los servicios de inteligencia artificial (IA) y machine learning (ML) de AWS

pueden recopilar y almacenar datos, y utilizar la administración de varias cuentas de los servicios de AWS integrada con Organizations.

2. Integre AWS Organizations con servicios de AWS.

   Cuando habilita un servicio de AWS para que realice tareas en su nombre en las cuentas miembros de su organización, AWS Organizations crea un rol vinculado al servicio de IAM para dicho servicio en cada cuenta miembro. Debe administrar el acceso de confianza mediante la Consola de administración de AWS, las API de AWS o la AWS CLI. Para obtener orientación prescriptiva sobre la habilitación del acceso de confianza, consulte Uso de AWS Organizations con otros servicios de AWS y Servicios de AWS que se pueden utilizar con Organizations.

3. Establezca un perímetro de datos.

   El perímetro de AWS suele representarse como una organización administrada por AWS Organizations. Junto con las redes y sistemas locales, el acceso a los recursos de AWS es lo que muchas personas consideran que es el perímetro de Mi AWS. El objetivo del perímetro es verificar que se permite el acceso si la identidad es de confianza, el recurso es de confianza y la red es la que se espera.

   a. Defina e implemente los perímetros.

      Siga los pasos que se describen en Perimeter implementation (Implementación del perímetro) del documento técnico Building a Perimeter on AWS (Construir un perímetro en AWS) para cada condición de autorización. Para obtener orientación prescriptiva sobre la protección de la capa de red, consulte Protección de redes.

   b. Supervise y alerte continuamente.

      AWS Identity and Access Management Access Analyzer ayuda a identificar los recursos y las cuentas de su organización que se comparten con entidades externas. Puede integrar IAM Access Analyzer con AWS Security Hub CSPM para enviar y agregar los hallazgos sobre un recurso desde IAM Access Analyzer a Security Hub CSPM para ayudarle a analizar la postura de seguridad de su entorno. Para permitir la integración, habilite tanto IAM Access Analyzer como Security Hub CSPM en cada región de cada cuenta. También puede utilizar Reglas de AWS Config para auditar la configuración y alertar a quien corresponda utilizando Amazon Q Developer in chat applications con AWS Security Hub CSPM. A continuación, puede utilizar los documentos de AWS Systems Manager Automation para corregir los recursos no conformes.

   c. Para obtener orientación prescriptiva sobre la supervisión y alerta continua de los recursos compartidos externamente, consulte Analizar el acceso público y entre cuentas.

4. Utilice el uso compartido de recursos en los servicios de AWS y restrínjalos de la forma oportuna.

Muchos servicios de AWS le permiten compartir recursos con otra cuenta o dirigirse a un recurso de otra cuenta, como las imágenes de máquina de Amazon (AMI) y AWS Resource Access Manager (AWS RAM). Restrinja la API `ModifyImageAttribute` para especificar las cuentas de confianza con las que compartir la AMI. Especifique la condición `ram:RequestedAllowsExternalPrincipals` cuando utilice AWS RAM para restringir el uso compartido únicamente a su organización; de esta forma, ayuda a evitar el acceso desde identidades que no sean de confianza. Para obtener orientación prescriptiva y conocer otras consideraciones, consulte Resource sharing and external targets (Uso compartido de recursos y destinos externos).

5. Utilice AWS RAM para compartir de forma segura en una cuenta o con otras Cuentas de AWS.

   AWS RAM le ayuda a compartir de forma segura los recursos que ha creado con roles y usuarios de su cuenta y con otras Cuentas de AWS. En un entorno de varias cuentas, AWS RAM le permite crear un recurso una vez y compartirlo con otras cuentas. Este enfoque ayuda a reducir su sobrecarga operativa a la vez que proporciona coherencia, visibilidad y auditabilidad en integraciones con Amazon CloudWatch y AWS CloudTrail, algo que no tiene cuando utiliza el acceso entre cuentas.

   Si tiene recursos que compartió anteriormente mediante una política basada en recursos, puede utilizar la API PromoteResourceShareCreatedFromPolicy o una equivalente para promover el recurso compartido a un recurso compartido completo de AWS RAM.

   En algunos casos, puede que tenga que dar pasos adicionales para compartir recursos. Por ejemplo, para compartir una instantánea cifrada, necesita compartir una clave AWS KMS.

Recursos

Prácticas recomendadas relacionadas:

- SEC03-BP07 Analizar el acceso público y entre cuentas
- SEC03-BP09 Compartir recursos de forma segura con terceros
- SEC05-BP01 Crear capas de red

Documentos relacionados:

- Bucket owner granting cross-account permission to objects it does not own (El propietario del bucket concede permisos entre varias cuentas a objetos que no son de su propiedad)

- How to use Trust Policies with IAM (Cómo utilizar las políticas de confianza con IAM)
- Building Data Perimeter on AWS (Creación de un perímetro de datos en AWS)
- Cómo utilizar un ID externo al otorgar acceso a los recursos de AWS a terceros
- Servicios de AWS que se pueden utilizar con AWS Organizations
- Establishing a data perimeter on AWS: Allow only trusted identities to access company data (Establecer un perímetro de datos en AWS: permitir que solo las identidades de confianza accedan a los datos de la empresa)

Vídeos relacionados:

- Granular Access with AWS Resource Access Manager (Acceso detallado con AWS Resource Access Manager)
- Securing your data perimeter with VPC endpoints (Protección del perímetro de datos con puntos de conexión de VPC)
- Establishing a data perimeter on AWS (Establecer un perímetro de datos en AWS)

Herramientas relacionadas:

- Data Perimeter Policy Examples (Ejemplos de políticas del perímetro de datos)

SEC03-BP09 Compartir recursos de forma segura con terceros

La seguridad de su entorno en la nube no se limita a su organización. Su organización puede recurrir a terceros para administrar una parte de sus datos. La administración de permisos para el sistema administrado por terceros debe seguir la práctica del acceso justo a tiempo utilizando el principio del privilegio mínimo con credenciales temporales. Si colabora estrechamente con un tercero, podrán reducir juntos el alcance del impacto y el riesgo de un acceso no intencionado.

Resultado deseado: cualquiera puede utilizar las credenciales de larga duración de AWS Identity and Access Management (IAM), las claves de acceso de IAM y las claves secretas que están asociadas a un usuario siempre que las credenciales sean válidas y estén activas. El uso de un rol de IAM y credenciales temporales le ayuda a mejorar su postura de seguridad general al reducir el esfuerzo que supone mantener credenciales de larga duración, incluida la sobrecarga de administración y operativa que entrañan esos datos confidenciales. Al utilizar un identificador único universal (UUID) para el ID externo en la política de confianza de IAM y mantener bajo su control las políticas de IAM asociadas al rol de IAM, puede auditar y verificar que el acceso concedido a un tercero no sea

demasiado permisivo. Para obtener orientación prescriptiva sobre el análisis de recursos que se comparten externamente, consulte SEC03-BP07 Analizar el acceso público y entre cuentas.

Antipatrones usuales:

- Utilizar la política de confianza de IAM predeterminada sin ninguna condición.
- Utilizar credenciales y claves de acceso de IAM de larga duración.
- Reutilizar ID externos.

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

Guía para la implementación

Es posible que desee permitir que se compartan recursos fuera de AWS Organizations o conceder a un tercero acceso a su cuenta. Por ejemplo, es posible que un tercero le proporcione una solución de supervisión que necesite acceder a los recursos de su cuenta. En esos casos, cree un rol entre cuentas de IAM con solo los privilegios que necesite el tercero. Además, defina una política de confianza utilizando la condición de ID externo. Cuando utilice un ID externo, usted o el tercero pueden generar un ID único para cada cliente, tercero o tenencia. El ID único no debe controlarlo nadie más que usted después de crearlo. El tercero debe implementar un proceso para relacionar el ID externo con el cliente de una forma segura, auditable y reproducible.

También puede utilizar Funciones de IAM en cualquier lugar para administrar roles de IAM para aplicaciones fuera de AWS que utilicen API de AWS.

Si el tercero ya no necesita acceder a su entorno, elimine el rol. Procure no proporcionar credenciales de larga duración a terceros. Manténgase al tanto de otros servicios de AWS que admiten el uso compartido. Por ejemplo, AWS Well-Architected Tool permite compartir una carga de trabajo con otras Cuentas de AWS y AWS Resource Access Manager le ayuda a compartir de forma segura un recurso de AWS de su propiedad con otras cuentas.

Pasos para la implementación

1. Utilice roles entre cuentas para proporcionar acceso a cuentas externas.

   Los roles entre cuentas reducen la cantidad de información confidencial que almacenan las cuentas externas y terceros para dar servicio a sus clientes. Los roles entre cuentas le permiten conceder acceso a los recursos de AWS de su cuenta de forma segura a un tercero, como AWS Partner u otras cuentas de su organización, al tiempo que puede mantener la capacidad de administrar y auditar dicho acceso.

El tercero podría estar proporcionándole un servicio desde una infraestructura híbrida o extrayendo datos a una ubicación externa. Funciones de IAM en cualquier lugar le ayuda a permitir que las cargas de trabajo de terceros interactúen de forma segura con sus cargas de trabajo de AWS y a reducir aún más la necesidad de utilizar credenciales de larga duración.

No debería utilizar credenciales de larga duración ni claves de acceso asociadas a usuarios para proporcionar acceso a cuentas externas. En su lugar, utilice roles entre cuentas para proporcionar el acceso entre cuentas.

2. Utilice un ID externo con terceros.

   El uso de un ID externo le permite designar quién puede asumir un rol en una política de confianza de IAM. La política de confianza puede exigir que el usuario que asume el rol reafirme la condición y el objetivo en el que opera. También proporciona un mecanismo para que el propietario de la cuenta permita que el rol se adopte únicamente en circunstancias específicas. La función principal del ID externo es abordar y prevenir el problema del suplente confundido.

   Utilice un ID externo si es propietario de una Cuenta de AWS y ha configurado un rol para un tercero que accede a otras Cuentas de AWS además de la suya, o cuando tenga que asumir roles en nombre de diferentes clientes. Trabaje con su tercero o AWS Partner para establecer una condición de ID externo que desee incluir en la política de confianza de IAM.

3. Utilice ID externos universalmente únicos.

   Implemente un proceso que genere un valor único aleatorio para un ID externo, como un identificador universalmente único (UUID). El hecho de que un tercero reutilice los ID externos para distintos clientes no resuelve el problema del suplente confundido, ya que el cliente A podría ver los datos del cliente B utilizando el ARN de rol del cliente B junto con el ID externo duplicado. En un entorno de varios inquilinos, en el que un tercero da soporte a varios clientes con diferentes Cuentas de AWS, el tercero debe utilizar un ID único diferente como ID externo para cada Cuenta de AWS. El tercero es responsable de detectar los ID externos duplicados y de asignar de forma segura cada cliente a su ID externo correspondiente. El tercero debe realizar pruebas para verificar que solo puede asumir el rol cuando se especifica el ID externo. El tercero debería abstenerse de almacenar el ARN del rol del cliente y el ID externo hasta que se requiera el ID externo.

   El ID externo no se trata como un secreto, pero no debe ser un valor fácil de adivinar, como un número de teléfono, un nombre o un ID de cuenta. Convierta el ID externo en un campo de solo lectura para que no pueda modificarse con el fin de suplantar la configuración.

El ID externo puede generarlo usted o el tercero. Defina un proceso para determinar quién es el responsable de generar el ID. Independientemente de la entidad que cree el ID externo, el tercero aplica la unicidad y los formatos de manera uniforme en todos los clientes.

4. Declare obsoletas las credenciales de larga duración proporcionadas por el cliente.

   Declare obsoleto el uso de credenciales de larga duración y utilice roles de cuentas cruzadas o Funciones de IAM en cualquier lugar. Si debe utilizar credenciales de larga duración, establezca un plan para migrar al acceso basado en roles. Para obtener información sobre la administración de claves, consulte Administración de identidades. Trabaje también con el equipo de su Cuenta de AWS y el tercero para establecer un runbook de mitigación de riesgos. Para obtener orientación prescriptiva sobre cómo responder y mitigar el impacto potencial de un incidente de seguridad, consulte Respuesta a incidentes.

5. Verifique que la configuración tenga una orientación prescriptiva o esté automatizada.

   La política que se cree para el acceso entre cuentas en sus cuentas debe seguir el principio del privilegio mínimo. El tercero debe proporcionarle un documento de políticas de roles o un mecanismo de configuración automatizado que utilice una plantilla de AWS CloudFormation o algo equivalente. Esto reduce la posibilidad de que se produzcan errores asociados a la creación manual de políticas y ofrece un registro de seguimiento auditable Para obtener más información sobre el uso de una plantilla de CloudFormation para crear roles entre cuentas, consulte Cross-Account Roles (Roles entre cuentas).

   El tercero debe proporcionar un mecanismo de configuración automatizado y auditable. Sin embargo, debería automatizar la configuración del rol con el documento de la política de roles que describe el acceso necesario. Con una plantilla de CloudFormation o algo equivalente, debería supervisar los cambios y utilizar la detección de desviaciones como parte de la práctica de auditoría.

6. Tenga en cuenta los cambios.

   La estructura de su cuenta, su necesidad de utilizar al tercero o la oferta de servicios que este le proporciona pueden cambiar. Debe anticiparse a los cambios y a los fallos y planificar en consecuencia las personas, los procesos y la tecnología adecuados. Audite de forma periódica el nivel de acceso que proporciona e implemente métodos de detección que le alerten de cambios inesperados. Supervise y audite el uso del rol y el almacén de datos de los ID externos. Debe estar preparado para revocar el acceso del tercero, de forma temporal o permanente, a causa de cambios o patrones de acceso inesperados. Asimismo, mida el impacto en su operación de

revocación, incluido el tiempo que lleva realizarla, las personas implicadas, el coste y el impacto en otros recursos.

Para obtener una orientación prescriptiva sobre los métodos de detección, consulte las prácticas recomendadas en Detección.

Recursos

Prácticas recomendadas relacionadas:

- SEC02-BP02 Usar credenciales temporales
- SEC03-BP05 Definir las barreras de protección de los permisos para su organización
- SEC03-BP06 Administrar el acceso en función del ciclo de vida
- SEC03-BP07 Analizar el acceso público y entre cuentas
- SEC04 Detección

Documentos relacionados:

- Bucket owner granting cross-account permission to objects it does not own (El propietario del bucket concede permisos entre varias cuentas a objetos que no son de su propiedad)
- How to use Trust Policies with IAM roles (Cómo utilizar las políticas de confianza con roles de IAM)
- Delegación del acceso entre Cuentas de AWS mediante roles de IAM
- How do I access resources in another Cuenta de AWS using AWS IAM? (¿Cómo accedo a los recursos en otra cuenta de AWS a través de AWS IAM?)
- Prácticas recomendadas de seguridad en IAM
- Lógica de evaluación de políticas entre cuentas
- Cómo utilizar un ID externo al otorgar acceso a los recursos de AWS a terceros
- Collecting Information from AWS CloudFormation Resources Created in External Accounts with Custom Resources (Recopilación de información de recursos de AWS CloudFormation creados en cuentas externas con recursos personalizados)
- Securely Using External ID for Accessing AWS Accounts Owned by Others (Uso seguro del ID externo para acceder a cuentas de AWS propiedad de terceros)
- Extend IAM roles to workloads outside of IAM with IAM Roles Anywhere (Extienda los roles de IAM de AWS a cargas de trabajo fuera de AWS con Funciones de IAM en cualquier lugar)

Vídeos relacionados:

- How do I allow users or roles in a separate Cuenta de AWS access to my Cuenta de AWS? (¿Cómo permito que los usuarios o roles de una cuenta de AWS independiente accedan a mi cuenta de AWS?
- AWS re:Invent 2018: Become an IAM Policy Master in 60 Minutes or Less (AWS re:Invent 2018: Consiga dominar las políticas de IAM en 60 minutos o menos)
- AWS Knowledge Center Live: IAM Best Practices and Design Decisions (Centro de conocimiento de AWS en directo: Prácticas recomendadas y decisiones de diseño de IAM)

Ejemplos relacionados:

- Well-Architected Lab - Lambda cross account IAM role assumption (Level 300) (Laboratorio de Well-Architected: Asunción de roles de IAM entre cuentas de Lambda [Nivel 300])
- Configure cross-account access to Amazon DynamoDB (Configurar el acceso entre cuentas a Amazon DynamoDB)
- AWS STS Network Query Tool (Herramienta de consulta de red AWS STS)

Detección

Pregunta

- SEGURIDAD 4. ¿Cómo detecta e investiga los eventos de seguridad?

SEGURIDAD 4. ¿Cómo detecta e investiga los eventos de seguridad?

Capte y analice eventos de registros y métricas para obtener visibilidad. Tome medidas sobre eventos de seguridad y posibles amenazas para ayudar a proteger su carga de trabajo.

Prácticas recomendadas

- SEC04-BP01 Configurar el registro de servicios y aplicaciones
- SEC04-BP02 Análisis centralizados de registros, hallazgos y métricas
- SEC04-BP03 Automatizar la respuesta a eventos
- SEC04-BP04 Implementar eventos de seguridad procesables

SEC04-BP01 Configurar el registro de servicios y aplicaciones

Retenga los registros de eventos de seguridad de servicios y aplicaciones. Se trata de un principio fundamental de seguridad en casos de uso de auditoría, investigación y uso operativo, y un requisito de seguridad común basado en las normas, políticas y procedimientos de gobernanza, riesgo y cumplimiento (GRC).

Resultado deseado: una organización debe ser capaz de recuperar de forma fiable y uniforme los registros de eventos de seguridad de los servicios y aplicaciones de AWS en el momento oportuno cuando sea necesario realizar un proceso o cumplir una obligación interna (por ejemplo, la respuesta a un incidente de seguridad). Considere la posibilidad de centralizar los registros para obtener mejores resultados operativos.

Antipatrones usuales:

- Los registros se almacenan para siempre o se eliminan demasiado pronto.
- Todo el mundo puede acceder a los registros.
- Depender por completo de procesos manuales para la gobernanza y el uso de los registros.
- Almacenar todos y cada uno de los tipos de registros por si fueran necesarios.
- Comprobar la integridad de los registros solo cuando es necesario.

Ventajas de esta práctica recomendada: implementar un mecanismo de análisis de causa raíz (RCA) para los incidentes de seguridad y una fuente de pruebas para sus obligaciones de gobernanza, riesgo y conformidad.

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

Guía para la implementación

Durante una investigación de seguridad u otros casos de uso basados en sus requisitos, necesita poder revisar los registros correspondientes para registrar y comprender todo el alcance y la cronología del incidente. También necesita los registros para generar alertas que indican que se han producido determinadas acciones de interés. Es fundamental seleccionar, habilitar, almacenar y configurar mecanismos de consulta y recuperación, así como de alerta.

Pasos para la implementación

- Seleccione y habilite las fuentes de registro. Antes de una investigación de seguridad, necesita obtener los registros relevantes para reconstruir de forma reatroctiva la actividad que se ha

producido en una Cuenta de AWS. Seleccione y habilite las fuentes de registros relevantes para sus cargas de trabajo.

Los criterios de selección de las fuentes de registros deben basarse en los casos de uso que requiera su negocio. Establezca un registro de seguimiento para cada Cuenta de AWS mediante AWS CloudTrail o un registro de seguimiento de AWS Organizations y, para ello, configure un bucket de Amazon S3.

AWS CloudTrail es un servicio de registro que rastrea las llamadas a la API que se realizan en una Cuenta de AWS y captura la actividad de los servicios de AWS. Está habilitado de manera predeterminada y retiene durante 90 días los eventos de administración que se pueden recuperar a través del historial de eventos de CloudTrail mediante la Consola de administración de AWS, la AWS CLI o un SDK de AWS. Si desea una retención y una visibilidad de los eventos de datos mayores, cree un registro de seguimiento de CloudTrail y asócielo a un bucket de Amazon S3 y, opcionalmente, a un grupo de registros de Amazon CloudWatch. Como alternativa, puede crear un CloudTrail Lake, que retiene los registros de CloudTrail hasta siete años y dispone de una utilidad de consulta basada en SQL.

AWS recomienda a los clientes que utilizan una VPC que habiliten los registros del tráfico de red y de DNS mediante los registros de flujo de VPC y los registros de consultas de solucionador de Amazon Route 53, respectivamente, y que los transmitan por streaming a un bucket de Amazon S3 o a un grupo de registros de CloudWatch. Puede crear un registro de flujo de VPC para una VPC, una subred o una interfaz de red. En el caso de los registros de flujo de VPC, puede elegir cómo y dónde utilizar los registros de flujo para reducir costes.

Los registros de AWS CloudTrail, los registros de flujo de VPC y los registros de consulta del solucionador de Route 53 son las fuentes de registros básicas que facilitan las investigaciones de seguridad en AWS. También puede utilizar Amazon Security Lake para recopilar, normalizar y almacenar estos datos de registros en los formatos Apache Parquet y Open Cybersecurity Schema Framework (OCSF), que están listos para su consulta. Security Lake también admite otros registros de AWS y registros de fuentes de terceros.

Los servicios de AWS pueden generar registros que no capturan las fuentes de registros básicas, como los registros de Elastic Load Balancing, los registros de AWS WAF, los registros del registrador de AWS Config, los hallazgos de Amazon GuardDuty, los registros de auditoría de Amazon Elastic Kubernetes Service (Amazon EKS) y los registros del sistema operativo y las aplicaciones de las instancias de Amazon EC2. Para obtener una lista completa de las opciones de registro y supervisión, consulte Appendix A: Cloud capability definitions – Logging and Events

(Apéndice A: Definiciones de las capacidades de la nube - Registro y eventos) de AWS Security Incident Response Guide (Guía de respuesta a incidentes de seguridad de AWS).

- Investigue las capacidades de registro de cada servicio y aplicación de AWS: cada servicio y aplicación de AWS le proporciona opciones para el almacenamiento de registros, cada una de las cuales tiene sus propias capacidades de retención y ciclo de vida. Los dos servicios de almacenamiento de registros más comunes son Amazon Simple Storage Service (Amazon S3) y Amazon CloudWatch. Para periodos de retención largos, se recomienda utilizar Amazon S3 por su rentabilidad y la flexibilidad de sus ciclos de vida. Si la opción de registro principal es Amazon CloudWatch Logs, quizá debería considerar la posibilidad de archivar los registros a los que se accede con menos frecuencia en Amazon S3.
- Seleccione el almacenamiento de registros: la elección del almacenamiento de registros suele estar relacionada con la herramienta de consulta que utilice, las capacidades de retención, la familiaridad con él y el coste. Las principales opciones para el almacenamiento de registros son un bucket de Amazon S3 o un grupo de CloudWatch Log.

  Un bucket de Amazon S3 es un almacenamiento rentable y duradero que tiene una política de ciclo de vida opcional. Los registros almacenados en buckets de Amazon S3 pueden consultarse a través de servicios como Amazon Athena.

  Un grupo de CloudWatch Logs ofrece un almacenamiento duradero y una utilidad de consulta integrada a través de CloudWatch Logs Insights.
- Identifique un periodo de retención de registros adecuado: cuando utilice un bucket de Amazon S3 o un grupo de CloudWatch Logs para almacenar registros, deberá establecer ciclos de vida adecuados para cada fuente de registros con el fin de optimizar los costes de almacenamiento y recuperación. Por lo general, los clientes tienen entre tres meses y un año de registros disponibles para su consulta, con un periodo de retención de hasta siete años. La elección de la disponibilidad y el periodo de retención debe ajustarse a sus requisitos de seguridad y a una combinación de requisitos legales, reglamentarios y empresariales.
- Habilite el registro para cada servicio y aplicación de AWS con las políticas de retención y ciclo de vida adecuadas: para cada servicio o aplicación de AWS de su organización, busque la guía de configuración de registro específica:
  - Configuración de registros de seguimiento de AWS CloudTrail
  - Configuración de registros de flujo de VPC
  - Configuración de exportaciones de hallazgos de Amazon GuardDuty
  - Configuración de grabaciones de AWS Config

  - Configuración del tráfico de ACL web de AWS WAF
  - Configuración de registros del tráfico de red de AWS Network Firewall
  - Configuración de registros de acceso de Elastic Load Balancing
  - Configuración de registros de consultas del solucionador de Amazon Route 53
  - Configuración de registros de Amazon RDS
  - Configuración de registros del plano de control de Amazon EKS
  - Configuración del agente de Amazon CloudWatch para instancias de Amazon EC2 y servidores locales
- Seleccione e implemente mecanismos de consulta para los registros: para las consultas de registros, puede utilizar CloudWatch Logs Insights para los datos almacenados en los grupos de CloudWatch Logs y Amazon Athena y Amazon OpenSearch Service para los datos almacenados en Amazon S3. También puede utilizar herramientas de consulta de terceros, como un servicio de administración de eventos e información de seguridad (SIEM).

  En el proceso de selección de una herramienta de consulta de registros, se deben tener en cuenta los aspectos relacionados con las personas, los procesos y la tecnología de sus operaciones de seguridad. Seleccione una herramienta que cumpla los requisitos operativos, empresariales y de seguridad, y que sea accesible y pueda mantenerse a largo plazo. Tenga en cuenta que las herramientas de consulta de registros funcionan de forma óptima cuando el número de registros a analizar se mantiene dentro de los límites de la herramienta. No es raro disponer de varias herramientas de consulta debido a limitaciones técnicas o de costes.

  Por ejemplo, podría utilizar una herramienta de administración de eventos e información de seguridad (SIEM) de terceros para realizar consultas en los últimos 90 días de datos, pero utilizar Athena para realizar consultas anteriores a esos 90 días debido al coste de la ingestión de registros de un SIEM. Independientemente de cuál sea la implementación, compruebe que su enfoque permite reducir al mínimo el número de herramientas necesarias para maximizar la eficiencia operativa, especialmente durante la investigación de un evento de seguridad.
- Utilice registros para las alertas: AWS proporciona alertas a través de varios servicios de seguridad:
  - AWS Config supervisa y registra las configuraciones de sus recursos de AWS y le permite automatizar la evaluación y la corrección con respecto a las configuraciones deseadas.
  - Amazon GuardDuty es un servicio de detección de amenazas que supervisa continuamente la actividad maliciosa y el comportamiento no autorizado para proteger sus Cuentas de AWS y cargas de trabajo. GuardDuty ingiere, agrega y analiza información de fuentes, como eventos

de administración y datos de AWS CloudTrail, registros DNS, registros de flujo de VPC y registros de auditoría de Amazon EKS. GuardDuty extrae secuencias de datos independientes directamente de CloudTrail, los registros de flujo de VPC, los registros de consultas de DNS y Amazon EKS. No es necesario que administre las políticas de los buckets de Amazon S3 ni que modifique la forma en que recopila y almacena los registros. Aun así, es recomendable que retenga estos registros para sus propios fines de investigación y conformidad.

- AWS Security Hub CSPM proporciona un único lugar en el que se agregan, organizan y priorizan las alertas de seguridad, o los hallazgos, desde varios servicios de AWS y productos de terceros opcionales para ofrecerle una vista completa de las alertas de seguridad y los estados de conformidad.

También puede utilizar motores de generación de alertas personalizados para alertas de seguridad que no cubran estos servicios o para alertas específicas relevantes para su entorno. Para obtener información sobre la creación de estas alertas y detecciones, consulte Detection (Detección) en AWS Security Incident Response Guide (Guía de respuesta a incidentes de seguridad de AWS).

Recursos

Prácticas recomendadas relacionadas:

- SEC04-BP02 Análisis centralizados de registros, hallazgos y métricas
- SEC07-BP04 Definir la administración del ciclo de vida de los datos
- SEC10-BP06: Desplegar las herramientas con anticipación

Documentos relacionados:

- AWS Security Incident Response Guide (Guía de respuesta ante incidentes de seguridad de AWS)
- Cómo comenzar a utilizar Amazon Security Lake
- Introducción a Amazon CloudWatch Logs
- Soluciones de socios de seguridad: registro y monitorización

Vídeos relacionados:

- AWS re:Invent 2022 - Introducing Amazon Security Lake (re:Invent 2022: Introducción a Amazon Security Lake)

Ejemplos relacionados:

- Assisted Log Enabler for AWS(Habilitador de registro asistido para AWS)
- AWS Security Hub CSPM Findings Historical Export (Exportación de hallazgos históricos de AWS Security Hub)

Herramientas relacionadas:

- Snowflake for Cybersecurity

SEC04-BP02 Análisis centralizados de registros, hallazgos y métricas

Los equipos de operaciones de seguridad confían en la recopilación de registros y el uso de herramientas de búsqueda para descubrir posibles eventos de interés, que podrían indicar una actividad no autorizada o un cambio no intencionado. Sin embargo, solo con analizar los datos recopilados y procesar manualmente la información no basta para satisfacer el volumen de información procedente de arquitecturas complejas. Solo con los análisis y los informes no se facilita la asignación de los recursos adecuados para trabajar en un evento a tiempo.

Una práctica recomendada para crear un equipo de operaciones de seguridad eficaz es integrar en profundidad el flujo de hallazgos y eventos de seguridad en un sistema de flujo de trabajo y notificación, como un sistema de emisión de tiques, un sistema de errores o problemas u otro sistema de administración de eventos e información de seguridad (SIEM, por sus siglas en inglés). De esta forma, se saca el flujo de trabajo de informes estáticos y de correo electrónico, y le permite enrutar, escalar y administrar eventos o hallazgos. Numerosas organizaciones ya integran también alertas de seguridad en sus plataformas de productividad de desarrolladores, de colaboración o de chats. Para las organizaciones que estén comenzando a incorporar la automatización, un sistema de tickets de baja latencia basado en API ofrece una flexibilidad considerable al planificar qué automatizar primero.

Esta práctica recomendada no se aplica solo a los eventos de seguridad generados a partir de mensajes de registro que muestran eventos de red o actividad del usuario, sino también a partir de cambios detectados en la propia infraestructura. La capacidad de detectar cambios, determinar su conveniencia y luego enrutar esa información al flujo de trabajo de corrección adecuado resulta esencial para mantener y validar una arquitectura segura en el contexto de los cambios en los que la naturaleza de su indeseabilidad es lo suficientemente sutil como para que su ejecución no pueda evitarse actualmente con una combinación de configuraciones de AWS Identity and Access Management (IAM) y AWS Organizations.

Amazon GuardDuty y AWS Security Hub CSPM ofrecen mecanismos de agregación, desduplicación y análisis para los registros que también están disponibles mediante otros servicios de AWS. GuardDuty ingiere, agrega y analiza información de fuentes como eventos de administración y datos de AWS CloudTrail, registros DNS de VPC y registros de flujo de VPC. Security Hub CSPM puede ingerir, agregar y analizar los resultados de GuardDuty, AWS Config, Amazon Inspector, Amazon Macie y AWS Firewall Manager, y un número significativo de productos de seguridad de terceros disponibles en AWS Marketplace y, si está convenientemente compilado, su propio código. Tanto GuardDuty como Security Hub CSPM tienen un modelo de administrador-miembro que puede combinar los hallazgos y los conocimientos de varias cuentas, los clientes con un SIEM local suelen utilizar Security Hub CSPM como registro del lado de AWS y un preprocesador y agregador de alertas a partir del que pueden ingerir Amazon EventBridge mediante un procesador y reenviador basado en AWS Lambda.

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

Guía para la implementación

- Evaluar las capacidades de procesamiento de registros: evalúe de las opciones disponibles para procesar registros.
  - Usar Amazon OpenSearch Service para registrar y supervisar (casi) todo
  - Búsqueda de un socio especializado en soluciones de registro y monitoreo
- Como punto de partida para el análisis de registros de CloudTrail, pruebe con Amazon Athena.
  - Configuración de Athena para analizar registros de CloudTrail
- Implementar el registro centralizado en AWS: consulte la siguiente solución de ejemplo de AWS para centralizar el registro procedente de varias fuentes.
  - Solución de centralización de registros
- Implementar el registro centralizado con un socio: los socios de APN tienen soluciones que le ayudarán a analizar los registros de forma centralizada.
  - Registro y supervisión

Recursos

Documentos relacionados:

- Soluciones de AWS: registro centralizado
- AWS Security Hub CSPM

- Amazon CloudWatch
- Amazon EventBridge
- Introducción: Amazon CloudWatch Logs
- Soluciones de socios de seguridad: registro y monitorización

Vídeos relacionados:

- Supervisión centralizada de la configuración de recursos y el cumplimiento
- Corrección de los hallazgos de Amazon GuardDuty y AWS Security Hub CSPM
- Administración de amenazas en la nube: Amazon GuardDuty y AWS Security Hub CSPM

SEC04-BP03 Automatizar la respuesta a eventos

El uso de la automatización para investigar y corregir eventos reduce el esfuerzo y los posibles errores humanos, y le permite escalar sus capacidades de investigación. Las revisiones frecuentes le ayudarán a ajustar sus herramientas de automatización y a aplicar iteraciones continuas.

En AWS, la investigación de eventos de interés y la información sobre cambios potencialmente inesperados en un flujo de trabajo automatizado se pueden lograr con Amazon EventBridge. Este servicio ofrece un motor de reglas escalable diseñado para gestionar tanto formatos de eventos nativos de AWS (p. ej., eventos de AWS CloudTrail) como eventos personalizados que puede generar a partir de su aplicación. Amazon GuardDuty también le permite enrutar eventos a un sistema de flujo de trabajo para esos sistemas de respuesta a incidentes de creación (AWS Step Functions) o a una cuenta de seguridad centralizada, o a un bucket para seguir analizándolos.

La detección de cambios y el enrutamiento de esta información al flujo de trabajo correcto también se puede llevar a cabo utilizando Reglas de AWS Config y paquetes de conformidad. AWS Config detecta cambios en los servicios del ámbito (aunque con una mayor latencia que EventBridge) y genera eventos que se pueden analizar con Reglas de AWS Config para restaurar, aplicar la política de conformidad y reenviar información a sistemas, como plataformas de administración de cambios y sistemas de emisión de tiques operativos. Además de escribir sus propias funciones de Lambda para responder a eventos de AWS Config, puede utilizar el kit de desarrollo de Reglas de AWS Configy una biblioteca de código abierto Reglas de AWS Config. Los paquetes de conformidad son una colección de Reglas de AWS Config y acciones de corrección que se despliegan como una entidad única elaborada como una plantilla YAML. A plantilla de paquete de conformidad de ejemplo está disponible para el pilar de seguridad del modelo Well-Architected.

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

Guía para la implementación

- Implementar alertas automatizadas con GuardDuty: GuardDuty es un servicio de detección de amenazas que supervisa sin descanso cualquier actividad malintencionada o comportamiento no autorizado para proteger sus cargas de trabajo y sus Cuentas de AWS. Active GuardDuty y configure alertas automatizadas.
- Automatizar los procesos de investigación: desarrolle procesos automatizados que investiguen un evento y envíen la información a un administrador para ganar tiempo.
  - Laboratorio: experiencia práctica con Amazon GuardDuty

Recursos

Documentos relacionados:

- Soluciones de AWS: registro centralizado
- AWS Security Hub CSPM
- Amazon CloudWatch
- Amazon EventBridge
- Introducción: Amazon CloudWatch Logs
- Soluciones de socios de seguridad: registro y monitorización
- Configuración de Amazon GuardDuty

Vídeos relacionados:

- Supervisión centralizada de la configuración de recursos y el cumplimiento
- Corrección de los hallazgos de Amazon GuardDuty y AWS Security Hub CSPM
- Administración de amenazas en la nube: Amazon GuardDuty y AWS Security Hub CSPM

Ejemplos relacionados:

- Laboratorio: Despliegue automatizado de controles de detección

SEC04-BP04 Implementar eventos de seguridad procesables

Cree alertas que se envíen a su equipo y que este pueda actuar sobre ellas. Asegúrese de que las alertas incluyen información relevante para que el equipo pueda actuar. Para cada mecanismo de detección que tenga, también debería tener un proceso, en forma de runbook o bien manual, para realizar la investigación. Por ejemplo, cuando se activa Amazon GuardDuty, se generan diferentes resultados. Debe tener una entrada de runbook para cada tipo de resultado, por ejemplo, si se detecta un troyano , su runbook tiene instrucciones simples que indican a alguien que debe investigarlo y solucionarlo.

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

Guía para la implementación

- Descubra las métricas disponibles para los servicios de AWS: descubra las métricas disponibles a través de Amazon CloudWatch para los servicios que está utilizando.
  - Documentación de servicio de AWS
  - Uso de métricas de Amazon CloudWatch
- Configure las alarmas de Amazon CloudWatch.
  - Uso de alarmas de Amazon CloudWatch

Recursos

Documentos relacionados:

- Amazon CloudWatch
- Amazon EventBridge
- Soluciones de socios de seguridad: registro y monitorización

Vídeos relacionados:

- Centrally Monitoring Resource Configuration and Compliance (Supervisión centralizada de la configuración de recursos y el cumplimiento)
- Remediating Amazon GuardDuty and AWS Security Hub CSPM Findings (Corrección de los resultados de Amazon GuardDuty y AWS Security Hub CSPM)
- Threat management in the cloud: Amazon GuardDuty and AWS Security Hub CSPM (Administración de amenazas en la nube: Amazon GuardDuty y AWS Security Hub CSPM)

Protección de la infraestructura

Preguntas

- SEGURIDAD 5. ¿Cómo protege los recursos de red?
- SEGURIDAD 6. ¿Cómo protege sus recursos informáticos?

SEGURIDAD 5. ¿Cómo protege los recursos de red?

Cualquier carga de trabajo que tenga forma de conexión de red, ya sea internet o una red privada, requiere varias capas de defensa para protegerse de amenazas internas y externas basadas en la red.

Prácticas recomendadas

- SEC05-BP01 Crear capas de red
- SEC05-BP02 Controlar el tráfico en todas las capas
- SEC05-BP03 Automatizar la protección de la red
- SEC05-BP04 Implementar inspección y protección

SEC05-BP01 Crear capas de red

Agrupe por capas los componentes que tienen los mismos requisitos de confidencialidad para minimizar el alcance potencial del impacto de un acceso no autorizado. Por ejemplo, un clúster de base de datos que está en una nube virtual privada (VPC) y no necesita acceso a Internet debe colocarse en subredes sin enrutamiento hacia Internet o desde Internet. El tráfico solo debe salir desde el siguiente recurso adyacente menos confidencial. Supongamos que tiene una aplicación web detrás de un equilibrador de carga. Su base de datos no debería ser accesible directamente desde este equilibrador de carga. Solo deberían tener acceso directo a su base de datos la lógica empresarial o el servidor web.

Resultado deseado: crear una red en capas. Las redes en capas ayudan a agrupar de forma lógica componentes de red similares. También reducen el alcance potencial del impacto que supondría un acceso no autorizado a la red. Una red que se haya configurado por capas de la forma adecuada hace más difícil que los usuarios no autorizados se dirijan a recursos adicionales dentro de su entorno de AWS. Además de asegurar las rutas de red internas, también debe proteger la periferia de su red, como las aplicaciones web y los puntos de conexión de API.

Antipatrones usuales:

- Crear todos los recursos en una única VPC o subred.
- Utilizar grupos de seguridad demasiado permisivos.
- No utilizar subredes.
- Permitir el acceso directo a almacenes de datos como bases de datos.

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

Guía para la implementación

Los componentes como instancias Amazon Elastic Compute Cloud (Amazon EC2), clústeres de base de datos de Amazon Relational Database Service (Amazon RDS) y funciones AWS Lambda que comparten requisitos de accesibilidad pueden segmentarse en capas formadas por subredes. Considere la posibilidad de desplegar cargas de trabajo sin servidor, como funciones Lambda, dentro de una VPC o detrás de un Amazon API Gateway. Las tareas de AWS Fargate que no tengan la necesidad de acceder a Internet deben colocarse en subredes sin una ruta hacia o desde Internet. Este enfoque por capas mitiga el impacto de una configuración errónea de una sola capa, que podría permitir un acceso no intencionado. Para AWS Lambda, puede ejecutar sus funciones en su VPC a fin de aprovechar los controles basados en VPC.

Si la conectividad de red incluye miles de VPC, Cuentas de AWS y redes locales, debe utilizar AWS Transit Gateway. Transit Gateway actúa como un centro que controla cómo se enruta el tráfico entre todas las redes conectadas, que actúan como radios. El tráfico entre Amazon Virtual Private Cloud (Amazon VPC) y Transit Gateway permanece en la red privada de AWS, lo que reduce la exposición externa a usuarios no autorizados y a posibles problemas de seguridad. El emparejamiento interregional de Transit Gateway también cifra el tráfico interregional sin ningún punto único de fallo ni cuello de botella en el ancho de banda.

Pasos para la implementación

- Utilice Reachability Analyzer para analizar la ruta entre un origen y un destino en función de la configuración: Reachability Analyzer le permite automatizar la verificación de la conectividad hacia y desde los recursos conectados a la VPC. Tenga en cuenta que, para realizar este análisis, se revisa la configuración (no se envían paquetes de red).
- Utilice el analizador de acceso de la red de Amazon VPC para identificar el acceso no intencionado de la red a los recursos: el analizador de acceso de la red de Amazon VPC le permite especificar sus requisitos de acceso a la red e identificar posibles rutas de la red.

- Considere si es necesario que los recursos se encuentren en una subred pública: no coloque recursos en subredes públicas de su VPC a menos que sea absolutamente necesario que reciban tráfico de red de fuentes públicas.
- Cree subredes en sus VPC: cree subredes para cada capa de red (en grupos que incluyan varias zonas de disponibilidad) para mejorar la microsegmentación. Compruebe también que ha asociado las tablas de enrutamiento correctas con sus subredes para controlar el enrutamiento y la conectividad a Internet.
- Utilice AWS Firewall Manager para administrar sus grupos de seguridad de VPC: AWS Firewall Manager ayuda a disminuir la carga de administración que supone el uso de varios grupos de seguridad.
- Utilice AWS WAF para protegerse contra vulnerabilidades web comunes: AWS WAF puede ayudar a mejorar la seguridad de la periferia inspeccionando el tráfico en busca de vulnerabilidades web comunes, como la inyección de código SQL. También le permite restringir el tráfico de direcciones IP procedentes de determinados países o ubicaciones geográficas.
- Utilice Amazon CloudFront como red de distribución de contenido (CDN): Amazon CloudFront puede ayudarle a acelerar su aplicación web al almacenar los datos más cerca de sus usuarios. También puede mejorar la seguridad de la periferia al utilizar HTTPS, restringir el acceso a zonas geográficas y garantizar que el tráfico de red solo pueda acceder a los recursos cuando se enrute a través de CloudFront.
- Utilice Amazon API Gateway cuando cree interfaces de programación de aplicaciones (API):Amazon API Gateway ayuda a publicar, supervisar y proteger las API de REST, HTTPS y WebSocket.

Recursos

Documentos relacionados:

- AWS Firewall Manager
- Amazon Inspector
- Seguridad de Amazon VPC
- Reachability Analyzer
- Amazon VPC Network Access Analyzer (Analizador de acceso de la red de Amazon Virtual Private Cloud)

Vídeos relacionados:

- AWS Transit Gateway reference architectures for many VPCs (Arquitecturas de referencia de AWS Transit Gateway para muchas VPC)
- Application Acceleration and Protection with Amazon CloudFront, AWS WAF, and AWS Shield (Aceleración y protección de aplicaciones con Amazon CloudFront, AWS WAF y AWS Shield)
- AWS re:Inforce 2022 - Validate effective network access controls on AWS (re:Inforce 2022: Validar la eficacia de los controles de acceso a la red en AWS)
- AWS re:Inforce 2022 - Advanced protections against bots using AWS WAF(re:Inforce 2022: Protecciones avanzadas contra bots con AWS WAF)

Ejemplos relacionados:

- Well-Architected Lab - Automated Deployment of VPC (Laboratorio de Well-Architected: Despliegue automatizado de VPC)
- Taller: Amazon VPC Network Access Analyzer (Analizador de acceso de la red de Amazon VPC)

SEC05-BP02 Controlar el tráfico en todas las capas

Al diseñar su topología de red, debería examinar los requisitos de conectividad de cada componente. Por ejemplo, si un componente requiere accesibilidad a Internet (entrante y saliente), conectividad a las VPC, servicios en la periferia y centros de datos externos.

Una VPC le permite definir su topología de red que abarca una Región de AWS, con un rango de direcciones IPv4 privadas que puede configurar o un rango de direcciones IPv6 que selecciona AWS. Debe aplicar múltiples controles con un enfoque de defensa en profundidad tanto para el tráfico entrante como para el saliente, incluido el uso de grupos de seguridad (firewall de inspección con estado), ACL de red, subredes y tablas de enrutamiento. Puede crear subredes en una zona de disponibilidad de una VPC. Cada subred puede tener una tabla de enrutamiento asociada que define las reglas para administrar las rutas que sigue el tráfico de la subred. Puede definir una subred que se pueda enrutar en Internet con una ruta que se dirija a una puerta de enlace de Internet o NAT asociada a la VPC o mediante otra VPC.

Cuando una instancia, una base de datos de Amazon Relational Database Service (Amazon RDS) u otro servicio se lanza en una VPC, tiene su propio grupo de seguridad por interfaz de red. Este firewall está situado fuera de la capa del sistema operativo y se puede usar para definir reglas sobre el tráfico entrante y saliente permitido. También puede definir las relaciones entre grupos de seguridad. Por ejemplo, las instancias de un grupo de seguridad de nivel de base de datos solo aceptan tráfico de instancias de nivel de aplicación, según la referencia de los grupos de seguridad

aplicados a las instancias implicadas. A no ser que utilice protocolos que no sean TCP, no debería ser necesario disponer de una instancia de Amazon Elastic Compute Cloud (Amazon EC2) que sea directamente accesible desde Internet (ni siquiera con puertos restringidos por grupos de seguridad) sin un equilibrador de carga o CloudFront. Esto ayuda en la protección ante el acceso no intencionado debido a un problema con el sistema operativo o la aplicación. Una subred también puede tener una ACL de red asociada, que actúa como firewall sin estado. Debería configurar la ACL de red para que estreche el ámbito del tráfico permitido entre capas; tenga en cuenta que tiene que definir reglas tanto para el tráfico entrante como para el saliente.

Algunos servicios de AWS requieren que los componentes accedan a Internet para realizar llamadas a la API, donde están ubicados los puntos de conexión de la API de AWS . Otros servicios de AWS usan Puntos de enlace de la VPC en sus Amazon VPC. Muchos servicios de AWS, incluidos Amazon S3 y Amazon DynamoDB, son compatibles con los puntos de conexión de VPC, y esta tecnología se ha generalizado en AWS PrivateLink. Le recomendamos que utilice este enfoque para acceder a servicios de AWS, servicios de terceros y sus propios servicios alojados en otras VPC de forma segura. Todo el tráfico de red de AWS PrivateLink permanece en la estructura global de AWS y nunca pasa por Internet. La conectividad solamente la puede iniciar el consumidor del servicio y no el proveedor de este. El uso de AWS PrivateLink para el acceso de un servicio externo le permite crear VPC aisladas por espacios vacíos sin acceso a Internet y ayuda a proteger sus VPC de vectores de amenaza externos. Los servicios de terceros pueden usar AWS PrivateLink para permitir que sus clientes se conecten a los servicios de sus VPC a través de direcciones IP privadas. Para activos de VPC que necesiten realizar conexiones salientes a Internet, se pueden establecer para que sean solo salientes (unidireccionales) mediante una puerta de enlace NAT administrada por AWS, puertas de enlace de Internet solo salientes o proxies web que cree y administre.

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

Guía para la implementación

- Controlar el tráfico de web en una VPC: implemente prácticas recomendadas de VPC para controlar el tráfico.
  - Seguridad de Amazon VPC
  - Puntos de enlace de la VPC
  - Grupo de seguridad de Amazon VPC
  - ACL de red
- Controlar el tráfico en la periferia: implemente servicios en la periferia, como Amazon CloudFront, para proporcionar una capa adicional de protección y otras características.

    - Casos de uso de Amazon CloudFront
    - AWS Global Accelerator
    - Firewall para aplicaciones web de AWS (AWS WAF)
    - Amazon Route 53
    - Enrutamiento de entrada de Amazon VPC
- Controlar el tráfico de red privado: implemente servicios que protejan el tráfico privado para su carga de trabajo.
    - Interconexión de Amazon VPC
    - Servicios de punto de conexión de Amazon VPC (AWS PrivateLink)
    - Amazon VPC Transit Gateway
    - AWS Direct Connect
    - AWS Site-to-site VPN
    - AWS Client VPN
    - Puntos de acceso de Amazon S3

Recursos

Documentos relacionados:

- AWS Firewall Manager
- Amazon Inspector
- Introducción a AWS WAF

Vídeos relacionados:

- Arquitecturas de referencia de AWS Transit Gateway para muchas VPC
- Aceleración y protección de aplicaciones con Amazon CloudFront, AWS WAF y AWS Shield

Ejemplos relacionados:

- Laboratorio: Despliegue automatizado de VPC

SEC05-BP03 Automatizar la protección de la red

Automatice los mecanismos de protección para proporcionar una red de autodefensa basada en la inteligencia de amenazas y la detección de anomalías. Por ejemplo, las herramientas de detección y prevención de intrusiones que puedan adaptarse a las amenazas actuales y reducir su impacto. Un firewall de una aplicación web es un ejemplo de dónde puede automatizar la protección de la red, por ejemplo, utilizando la solución AWS WAF Security Automations (https://github.com/awslabs/aws-waf-security-automations) para bloquear automáticamente las solicitudes que se originen en direcciones IP asociadas con actores de amenazas conocidos.

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

Guía para la implementación

- Automatice la protección para el tráfico basado en la web: AWS ofrece una solución que utiliza AWS CloudFormation para desplegar automáticamente un conjunto de reglas de AWS WAF diseñadas para filtrar ataques basados en web frecuentes. Los usuarios pueden seleccionar entre funciones de protección preconfiguradas que definen las reglas incluidas en una lista de control de acceso web de AWS WAF (ACL web).
  - Automatizaciones de seguridad de AWS WAF
- Plantéese soluciones de AWS Partner: los socios de AWS ofrecen cientos de productos destacados que son equivalentes o idénticos a los controles que ya utiliza en sus entornos locales o que pueden integrarse con ellos. Estos productos complementan a los servicios de AWS existentes y le permiten implementar una completa arquitectura de seguridad, así como disfrutar de una experiencia más coherente tanto en la nube como en los entornos locales.
  - Seguridad de la infraestructura

Recursos

Documentos relacionados:

- AWS Firewall Manager
- Amazon Inspector
- Seguridad del Amazon VPC
- Introducción a AWS WAF

Videos relacionados:

- Arquitecturas de referencia de AWS Transit Gateway para muchas VPC
- Aceleración y protección de aplicaciones con Amazon CloudFront, AWS WAF y AWS Shield

Ejemplos relacionados:

- Laboratorio: Despliegue automatizado de VPC

SEC05-BP04 Implementar inspección y protección

Inspeccione y filtre el tráfico en cada capa. Puede inspeccionar las configuraciones de VPC para buscar posibles accesos no deseados mediante Analizador de acceso de red de VPC. Puede especificar los requisitos de acceso a la red e identificar las posibles rutas de red que no los cumplan. En el caso de los componentes que realizan transacciones a través de protocolos basados en HTTP, un firewall de aplicaciones web puede ayudar a protegerlos de los ataques más habituales. AWS WAF es un firewall de aplicaciones web que le permite supervisar y bloquear las solicitudes HTTP(s) que coincidan con sus reglas configurables y que se reenvíen a una API de Amazon API Gateway, Amazon CloudFront o un Application Load Balancer. Para empezar con AWS WAF, puede usar Reglas administradas de AWS en combinación con sus propias integraciones socios o utilizar las existentes.

Para administrar AWS WAF, las protecciones de AWS Shield Advanced y los grupos de seguridad de Amazon VPC en AWS Organizations, puede utilizar AWS Firewall Manager. Le permite configurar y administrar de forma centralizada las reglas de firewall en todas sus cuentas y aplicaciones, lo que facilita escalar la aplicación de las reglas comunes. También le permite responder rápidamente a los ataques, con AWS Shield Advancedo soluciones que puede bloquear automáticamente las solicitudes no deseadas a sus aplicaciones web. Firewall Manager también funciona con AWS Network Firewall. AWS Network Firewall es un servicio administrado que utiliza un motor de reglas para ofrecerle un control detallado del tráfico de red con estado y sin estado. Es compatible con las especificaciones del sistema de prevención de intrusiones (IPS) de código abierto compatibles con Suricata para ayudar a proteger su carga de trabajo.

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

Guía para la implementación

- Configure Amazon GuardDuty: GuardDuty es un servicio de detección de amenazas que supervisa sin descanso cualquier actividad malintencionada o comportamiento no autorizado para proteger sus cargas de trabajo y sus Cuentas de AWS. Active GuardDuty y configure alertas automatizadas.

    - Amazon GuardDuty
    - Laboratorio: Despliegue automatizado de controles de detección
- Configure los registros de flujo de nube virtual privada (VPC): la característica de registros de flujo de VPC permite registrar información acerca del tráfico IP que entra y sale de las interfaces de red en la VPC. Los datos de registro de flujo se pueden publicar en Amazon CloudWatch Logs y Amazon Simple Storage Service (Amazon S3). Cuando cree un registro de flujo, podrá recuperar y ver sus datos en el destino elegido.
- Considere el reflejo de tráfico de VPC: el reflejo de tráfico es una característica de Amazon VPC que puede utilizar para copiar el tráfico de red de una interfaz de red elástica de instancias de Amazon Elastic Compute Cloud (Amazon EC2) y, a continuación, enviarlo a dispositivos de seguridad y supervisión fuera de banda para la inspección de contenido, la supervisión de amenazas y la resolución de problemas.
    - Reflejo de tráfico de VPC

Recursos

Documentos relacionados:

- AWS Firewall Manager
- Amazon Inspector
- Seguridad de Amazon VPC
- Introducción a AWS WAF

Vídeos relacionados:

- AWS Transit Gateway reference architectures for many VPCs (Arquitecturas de referencia de AWS Transit Gateway para muchas VPC)
- Application Acceleration and Protection with Amazon CloudFront, AWS WAF, and AWS Shield (Aceleración y protección de aplicaciones con Amazon CloudFront, AWS WAF y AWS Shield)

Ejemplos relacionados:

- Laboratorio: Despliegue automatizado de VPC

SEGURIDAD 6. ¿Cómo protege sus recursos informáticos?

Los recursos informáticos en su carga de trabajo requieren varias capas de defensa para ayudar a protegerse de amenazas externas e internas. Entre los recursos informáticos se incluyen instancias EC2, contenedores, funciones de AWS Lambda, servicios de bases de datos, dispositivos IoT, etc.

Prácticas recomendadas

- SEC06-BP01 Administrar las vulnerabilidades
- SEC06-BP02 Reducir la superficie expuesta a ataques
- SEC06-BP03 Implementar servicios administrados
- SEC06-BP04 Automatizar la protección informática
- SEC06-BP05 Permitir que los usuarios realicen acciones a distancia
- SEC06-BP06 Validar la integridad del software

SEC06-BP01 Administrar las vulnerabilidades

Analice con frecuencia su código, sus dependencias y su infraestructura en busca de vulnerabilidades, y aplique parches para solucionarlas, para ayudarle a protegerse contra las nuevas amenazas.

Resultado deseado: crear y mantener un programa de administración de vulnerabilidades. Analice periódicamente recursos como las instancias de Amazon EC2, los contenedores de Amazon Elastic Container Service (Amazon ECS) y las cargas de trabajo de Amazon Elastic Kubernetes Service (Amazon EKS) y aplique parches en ellos. Configure periodos de mantenimiento para los recursos administrados por AWS, como las bases de datos de Amazon Relational Database Service (Amazon RDS). Utilice el análisis de código estático para inspeccionar el código fuente de las aplicaciones en busca de problemas comunes. Considere la posibilidad de realizar pruebas de penetración en aplicaciones web si su organización cuenta con los conocimientos necesarios o puede contratar ayuda externa.

Antipatrones usuales:

- No disponer de un programa de administración de vulnerabilidades.
- Aplicar parches en el sistema sin tener en cuenta la gravedad o la forma de evitar riesgos.
- Utilizar software que haya superado la fecha de fin de vida útil (EOL) de su proveedor.
- Desplegar código en producción antes de analizarlo en busca de problemas de seguridad.

Beneficios de establecer esta práctica recomendada:

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

Guía para la implementación

Un programa de administración de vulnerabilidades incluye la evaluación de la seguridad, la identificación de problemas, el establecimiento de prioridades y la aplicación de parches como parte de la resolución de los problemas. La automatización es la clave para analizar continuamente las cargas de trabajo en busca de problemas y exposiciones no intencionadas a la red y para realizar correcciones. La automatización de la creación y actualización de recursos ahorra tiempo y reduce el riesgo de que los errores de configuración den lugar a más problemas. En un programa de administración de vulnerabilidades bien diseñado, también se debería considerar la realización de pruebas de vulnerabilidad durante las fases de desarrollo y despliegue del ciclo de vida del software. Implementar la administración de vulnerabilidades durante el desarrollo y el despliegue ayuda a disminuir la posibilidad de que una vulnerabilidad pueda abrirse camino en su entorno de producción.

Para implementar un programa de administración de vulnerabilidades, es necesario conocer bien el modelo de responsabilidad compartida de AWS y cómo se relaciona con sus cargas de trabajo específicas. En el modelo de responsabilidad compartida, AWS es responsable de proteger la infraestructura de la Nube de AWS. Esta infraestructura está compuesta por hardware, software, redes e instalaciones que ejecutan servicios de la Nube de AWS. Usted es responsable de la seguridad en la nube, por ejemplo, de los datos reales, de la configuración de seguridad y de las tareas de administración de las instancias de Amazon EC2, así como de verificar que sus objetos de Amazon S3 estén clasificados y configurados correctamente. Su enfoque de la administración de vulnerabilidades también puede variar en función de los servicios que consuma. Por ejemplo, AWS es quien administra la aplicación de parches de nuestro servicio de base de datos relacional administrado, Amazon RDS, pero usted es el responsable de aplicar los parches en las bases de datos autoalojadas.

AWS dispone de numerosos servicios para ayudarle con su programa de administración de vulnerabilidades. Amazon Inspector analiza continuamente las cargas de trabajo de AWS en busca de problemas de software y accesos no intencionados a la red. AWS Systems Manager Patch Manager ayuda a administrar la aplicación de parches en todas sus instancias de Amazon EC2. Amazon Inspector y Systems Manager pueden consultarse en AWS Security Hub CSPM, un servicio de administración de la postura de seguridad en la nube que ayuda a automatizar las comprobaciones de seguridad de AWS y a centralizar las alertas de seguridad.

Amazon CodeGuru puede ayudar a identificar posibles problemas en las aplicaciones Java y Python mediante el análisis estático del código.

Pasos para la implementación

- Configure Amazon Inspector: Amazon Inspector detecta automáticamente las instancias de Amazon EC2 recién lanzadas, las funciones Lambda y las imágenes de contenedor elegibles que se envían a Amazon ECR y las analiza inmediatamente en busca de problemas del software, defectos potenciales y una exposición no intencionada a la red.
- Analice el código fuente: analice bibliotecas y dependencias en busca de problemas y defectos. Amazon CodeGuru puede analizar y proporcionar recomendaciones para corregir problemas de seguridad comunes tanto para aplicaciones Java como Python. La Fundación OWASP publica una lista de herramientas de análisis del código fuente (también conocidas como herramientas SAST).
- Implemente un mecanismo para analizar y aplicar parches en su entorno existente, así como para incluir el análisis como parte de un proceso de desarrollo de la canalización CI/CD: implemente un mecanismo para analizar y aplicar parches para solucionar los problemas en sus dependencias y sistemas operativos y protegerse contra nuevas amenazas. Haga que ese mecanismo se ejecute de forma regular. La administración de vulnerabilidades de software es esencial para saber dónde hay que aplicar parches o solucionar problemas del software. Dé prioridad a la corrección de los posibles problemas de seguridad incorporando evaluaciones de vulnerabilidad en una fase temprana de la canalización de la integración continua y entrega continua (CI/CD). Su enfoque puede variar en función de los servicios de AWS que consuma. Para buscar posibles problemas en el software que se ejecuta en instancias de Amazon EC2, añada Amazon Inspector a su canalización para que le avise y detenga el proceso de desarrollo si se detectan problemas o posibles defectos. Amazon Inspector supervisa continuamente los recursos. También puede utilizar productos de código abierto como OWASP Dependency-Check, Snyk, OpenVAS, administradores de paquetes y herramientas de AWS Partner para la administración de vulnerabilidades.
- Utilice AWS Systems Manager: usted es el responsable de la administración de parches en sus recursos de AWS, incluidas las instancias de Amazon Elastic Compute Cloud (Amazon EC2), las imágenes de máquina de Amazon (AMI) y otros recursos de computación. AWS Systems Manager Patch Manager automatiza el proceso de aplicación de parches a instancias administradas con actualizaciones de seguridad y de otro tipo. Patch Manager puede utilizarse para aplicar parches en instancias de Amazon EC2 tanto para sistemas operativos como para aplicaciones, incluidas aplicaciones de Microsoft, paquetes de servicios de Windows y actualizaciones de versiones secundarias para instancias basadas en Linux. Además de Amazon EC2, Patch Manager también puede utilizarse para aplicar parches en servidores locales.

Para obtener una lista de los sistemas operativos compatibles, consulte Sistemas operativos compatibles en la Guía del usuario de Systems Manager. Puede analizar instancias para ver solamente un informe de los parches que faltan, o puede analizar e instalar automáticamente todos los parches que falten.

- Utilice AWS Security Hub CSPM: Security Hub CSPM proporciona una vista completa de su estado de seguridad en AWS. Recopila datos de seguridad en múltiples servicios de AWS y proporciona esos hallazgos en un formato estandarizado, que le permite priorizar los hallazgos de seguridad en todos los servicios de AWS.
- Utilice AWS CloudFormation: AWS CloudFormation es un servicio de infraestructura como código (IaC) que puede ayudarle a administrar las vulnerabilidades mediante la automatización del despliegue de recursos y la estandarización de la arquitectura de los recursos en diversas cuentas y entornos.

Recursos

Documentos relacionados:

- AWS Systems Manager
- Security Overview of AWS Lambda (Información general de seguridad de AWS Lambda)
- Amazon CodeGuru
- Improved, Automated Vulnerability Management for Cloud Workloads with a New Amazon Inspector (Administración de vulnerabilidades mejorada y automatizada para cargas de trabajo en la nube con un nuevo Amazon Inspector)
- Automate vulnerability management and remediation in AWS using Amazon Inspector and AWS Systems Manager – Part 1 (Automatice la administración y corrección de vulnerabilidades en AWS mediante Amazon Inspector y AWS Systems Manager - Primera parte)

Vídeos relacionados:

- Securing Serverless and Container Services (Protección de servicios de contenedor y sin servidor)
- Security best practices for the Amazon EC2 instance metadata service (Prácticas recomendadas de seguridad para el servicio de metadatos de instancias de Amazon EC2)

SEC06-BP02 Reducir la superficie expuesta a ataques

Reduzca su exposición al acceso no intencionado endureciendo los sistemas operativos y minimizando los componentes, bibliotecas y servicios consumibles externamente que se estén utilizando. Puede comenzar por reducir los componentes no utilizados, ya sean paquetes de sistemas operativos, aplicaciones para cargas de trabajo basadas en Amazon Elastic Compute Cloud (Amazon EC2) o módulos de software externos en su código, para todas las cargas de trabajo. Puede encontrar muchas guías de endurecimiento y configuración de seguridad para sistemas operativos y software de servidores comunes. Por ejemplo, puede empezar por el Centro para la seguridad de Internet e iterar a partir de ahí.

En Amazon EC2, puede crear sus propias imágenes de máquina de Amazon (AMI), a las que habrá aplicado parches y habrá endurecido, para ayudarle a cumplir los requisitos de seguridad específicos de su organización. Los parches y otros controles de seguridad que apliquen en la AMI serán efectivos en el momento en el que se crearon, no son dinámicos a no ser que los modifique tras el lanzamiento, por ejemplo con AWS Systems Manager.

Puede simplificar el proceso de creación de AMI seguras con EC2 Image Builder. EC2 Image Builder reduce significativamente el esfuerzo necesario para crear y mantener imágenes golden sin escribir ni mantener automatizaciones. Cuando hay disponibles actualizaciones de software, Image Builder produce automáticamente una nueva imagen sin exigir a los usuarios que inicien manualmente creaciones de imágenes. EC2 Image Builder le permite validar fácilmente la funcionalidad y seguridad de sus imágenes antes de usarlas en producción con pruebas propias o proporcionadas por AWS. También puede aplicar la configuración de seguridad facilitada por AWS para proteger aún más sus imágenes de modo que cumplan criterios de seguridad internos. Por ejemplo, puede producir imágenes que se ajusten al estándar Security Technical Implementation Guide (STIG) con plantillas proporcionadas por AWS.

Mediante el uso de herramientas de análisis de código estático de terceros, puede identificar problemas de seguridad comunes como enlaces de entrada de funciones sin comprobar, además de vulnerabilidades y exposiciones comunes aplicables (CVE). Puede usar Amazon CodeGuru para los lenguajes compatibles. Las herramientas de comprobación de dependencias también se pueden usar para determinar si las bibliotecas con las que se vincula su código están actualizadas a su última versión, si no tienen CVE y si tienen condiciones de licencia que se ajusten a sus requisitos de política del software.

Con Amazon Inspector, puede llevar a cabo evaluaciones de configuración de sus instancias en busca de CVE conocidos, evaluarlas en función de referencias de seguridad y automatizar la notificación de defectos. Amazon Inspector se ejecuta en instancias de producción o en una

canalización de compilación y notifica a los desarrolladores e ingenieros cuando detecten algún hallazgo. Puede acceder a los hallazgos programáticamente y dirigir el equipo a trabajos pendientes y sistemas de seguimiento de errores. EC2 Image Builder se puede usar para mantener imágenes de servidor (AMI) con aplicación de parches automática, aplicación de políticas de seguridad proporcionadas por AWS y otras personalizaciones. Al usar contenedores, implemente el análisis de imágenes de ECR en su canalización de compilación y aplíquelo de forma frecuente a su repositorio de imágenes para buscar CVE en sus contenedores.

Aunque Amazon Inspector y otras herramientas son eficaces a la hora de identificar configuraciones y cualquier CVE que pudiese constar, son necesarios otros métodos para comprobar su carga de trabajo en el nivel de la aplicación. El fuzzing es un método conocido de detección de errores utilizando la automatización para inyectar datos con un formato incorrecto en los campos de entrada y otras áreas de su aplicación.

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

Guía para la implementación

- Endurezca el sistema operativo: configure el sistema operativo para cumplir con las prácticas recomendadas.
  - Protección de Amazon Linux
  - Protección de Microsoft Windows Server
- Endurezca los recursos en contenedores: configure los recursos en contenedores para cumplir con las prácticas recomendadas de seguridad.
- Implemente prácticas recomendadas de AWS Lambda.
  - Prácticas recomendadas de AWS Lambda

Recursos

Documentos relacionados:

- AWS Systems Manager
- Sustitución de un host bastión con Amazon EC2 Systems Manager
- Información general de seguridad de AWS Lambda

Videos relacionados:

- Ejecución de cargas de trabajo de alta seguridad en Amazon EKS
- Protección de servicios de contenedor y sin servidor
- Prácticas recomendadas de seguridad para el servicio de metadatos de instancias Amazon EC2

Ejemplos relacionados:

- Laboratorio: Despliegue de un firewall para aplicaciones web

SEC06-BP03 Implementar servicios administrados

Implemente servicios que administren recursos, como Amazon Relational Database Service (Amazon RDS), AWS Lambda y Amazon Elastic Container Service (Amazon ECS), para reducir sus tareas de mantenimiento de seguridad como parte del modelo de responsabilidad compartida. Por ejemplo, Amazon RDS le ayuda a configurar, operar y escalar una base de datos relacional, y automatiza tareas administrativas, como el aprovisionamiento de hardware, la configuración de bases de datos, la aplicación de parches y la creación de copias de seguridad. Esto significa que tendrá más tiempo libre para centrarse en proteger su aplicación de otras maneras descritas en AWS Well-Architected Framework. Lambda le permite ejecutar código sin aprovisionar ni administrar servidores, de modo que solo tendrá que centrarse en la conectividad, invocación y seguridad en el nivel del código, y no en la infraestructura ni en el sistema operativo.

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

Guía para la implementación

- Explore los servicios disponibles: explore, pruebe e implemente servicios que administren recursos, como Amazon RDS, AWS Lambda y Amazon ECS.

Recursos

Documentos relacionados:

- Sitio web de AWS
- AWS Systems Manager
- Sustitución de un host bastión con Amazon EC2 Systems Manager
- Información general de seguridad de AWS Lambda

Vídeos relacionados:

- Ejecución de cargas de trabajo de alta seguridad en Amazon EKS
- Protección de servicios de contenedor y sin servidor
- Prácticas recomendadas de seguridad para el servicio de metadatos de instancias Amazon EC2

Ejemplos relacionados:

- Laboratorio: solicitar un certificado público en AWS Certificate Manager

SEC06-BP04 Automatizar la protección informática

Automatice sus mecanismos de protección informática, incluida la administración de vulnerabilidades, la reducción de superficies expuestas a ataques y la administración de recursos. La automatización le ayudará a dedicar tiempo a proteger otros aspectos de su carga de trabajo y a reducir el riesgo de errores humanos.

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

Guía para la implementación

- Automatizar la administración de la configuración: aplique y valide configuraciones seguras de forma automática mediante el uso de un servicio o herramienta de administración de la configuración.
  - AWS Systems Manager
  - AWS CloudFormation
  - Laboratorio: Despliegue automatizado de VPC
  - Laboratorio: Despliegue automatizado de una aplicación web en EC2
- Automatizar la aplicación de revisiones en instancias Amazon Elastic Compute Cloud (Amazon EC2): Patch Manager de AWS Systems Manager automatiza el proceso de aplicación de revisiones a instancias administradas con actualizaciones de seguridad y de otro tipo. Puede utilizar Patch Manager para aplicar revisiones tanto para sistemas operativos como para aplicaciones.
  - Patch Manager de AWS Systems Manager

    - Implementación de revisiones centralizadas para varias regiones y cuentas con Automatización de AWS Systems Manager

- Implementar la detección y prevención de intrusiones: implemente una herramienta de detección y prevención de intrusiones para supervisar y detener las actividades maliciosas en las instancias.
- Considerar soluciones de AWS Partner: los socios de AWS ofrecen cientos de productos destacados que son equivalentes o idénticos a los controles que ya utiliza en sus entornos locales o que pueden integrarse con ellos. Estos productos complementan a los servicios de AWS existentes y le permiten implementar una completa arquitectura de seguridad, así como disfrutar de una experiencia más coherente tanto en la nube como en los entornos locales.
    - Seguridad de la infraestructura

Recursos

Documentos relacionados:

- AWS CloudFormation
- AWS Systems Manager
- Patch Manager de AWS Systems Manager
- Implementación de revisiones centralizadas para varias regiones y cuentas con Automatización de AWS Systems Manager
- Seguridad de la infraestructura
- Sustitución de un host bastión con Amazon EC2 Systems Manager
- Información general de seguridad de AWS Lambda

Vídeos relacionados:

- Ejecución de cargas de trabajo de alta seguridad en Amazon EKS
- Protección de servicios de contenedor y sin servidor
- Prácticas recomendadas de seguridad para el servicio de metadatos de instancias de Amazon EC2

Ejemplos relacionados:

- Laboratorio: Despliegue de un firewall para aplicaciones web
- Laboratorio: Despliegue automatizado de una aplicación web en EC2

SEC06-BP05 Permitir que los usuarios realicen acciones a distancia

La eliminación de la capacidad de acceder de forma interactiva reduce el riesgo de errores humanos y el potencial de llevar a cabo configuración o administración manuales. Por ejemplo, utilice un flujo de trabajo de administración de cambios para desplegar instancias de Amazon Elastic Compute Cloud (Amazon EC2) utilizando infraestructura como código, y después administre instancias de Amazon EC2 utilizando herramientas como AWS Systems Manager en lugar de permitir el acceso directo o mediante un host bastión. AWS Systems Manager puede automatizar una variedad de tareas de mantenimiento y despliegue, utilizando características como automatización de automatización, documentos (guías de estrategias) y el comando de ejecución. Las pilas de AWS CloudFormation se generan a partir de las canalizaciones y pueden automatizar el despliegue de su infraestructura y las tareas de administración sin usar directamente la Consola de administración de AWS o las API.

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

Guía para la implementación

- Sustituir el acceso a la consola: sustituya el acceso a la consola (SSH o RDP) a instancias con Run Command de AWS Systems Manager para automatizar las tareas de administración.

- Run Command de AWS Systems Manager

Recursos

Documentos relacionados:

- AWS Systems Manager
- Run Command de AWS Systems Manager
- Sustitución de un host bastión con Amazon EC2 Systems Manager
- Información general de seguridad de AWS Lambda

Vídeos relacionados:

- Ejecución de cargas de trabajo de alta seguridad en Amazon EKS
- Protección de servicios de contenedor y sin servidor
- Prácticas recomendadas de seguridad para el servicio de metadatos de instancias de Amazon EC2

Ejemplos relacionados:

- Laboratorio: Despliegue de un firewall para aplicaciones web

SEC06-BP06 Validar la integridad del software

Implemente mecanismos (por ejemplo, firma de código) para validar que el software, el código y las bibliotecas que se utilizan en la carga de trabajo procedan de fuentes de confianza y no se hayan manipulado. Por ejemplo, debería verificar el certificado de firma de código de binarios y scripts para confirmar el autor y asegurarse de que no se haya manipulado desde que el autor lo creó. AWS Signer puede ayudar a garantizar la confianza e integridad de su código administrando de forma centralizada el ciclo de vida de la firma del código, incluida la certificación de la firma y las claves públicas y privadas. Puede aprender a usar patrones avanzados y prácticas recomendadas para la firma de código con AWS Lambda. Además, comparar la suma de comprobación de un software que haya descargado con la suma de comprobación del proveedor puede ayudar a garantizar que no haya existido manipulación alguna.

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

Guía para la implementación

- Investigue mecanismos: la firma de código es un mecanismo que se puede usar para validar la integridad del software.
  - NIST: consideraciones de seguridad para la firma de código

Recursos

Documentos relacionados:

- AWS Signer
- Nuevo: Firma de código, un control de confianza e integridad para AWS Lambda

Protección de los datos

Preguntas

- SEGURIDAD 7. ¿Cómo clasifica sus datos?
- SEGURIDAD 8. ¿Cómo protege los datos en reposo?
- SEGURIDAD 9. ¿Cómo protege sus datos en tránsito?

SEGURIDAD 7. ¿Cómo clasifica sus datos?

La clasificación proporciona una forma de categorizar los datos, basada en el nivel de importancia y la confidencialidad, para ayudarle a determinar los controles de protección y conservación adecuados.

Prácticas recomendadas

- SEC07-BP01 Identificar los datos en su carga de trabajo
- SEC07-BP02 Definir controles de protección de datos
- SEC07-BP03 Automatizar la identificación y la clasificación
- SEC07-BP04 Definir la administración del ciclo de vida de los datos

SEC07-BP01 Identificar los datos en su carga de trabajo

Es fundamental conocer el tipo y la clasificación de los datos que procesa su carga de trabajo, los procesos empresariales asociados, dónde se almacenan los datos y quién es su propietario. También debe conocer los requisitos legales y de conformidad aplicables a su carga de trabajo y qué controles deben aplicarse en los datos. Identificar los datos es el primer paso en el proceso de clasificación de los datos.

Beneficios de establecer esta práctica recomendada:

La clasificación de datos permite a los propietarios de las cargas de trabajo identificar las ubicaciones en las que se almacenan datos confidenciales y determinar cómo se debe acceder a esos datos y compartirlos.

La clasificación de los datos tiene como objetivo responder a las siguientes preguntas:

- ¿Qué tipo de datos tiene?

Podrían ser datos como los siguientes:

- Datos de propiedad intelectual (PI), como secretos comerciales, patentes o acuerdos contractuales.
- Información sanitaria protegida (PHI), como historiales médicos que contienen información sobre la historia clínica de una persona.
- Información de identificación personal (PII), como nombre, dirección, fecha de nacimiento y número de identificación nacional o de registro.
- Datos de tarjetas de crédito, como el número de cuenta principal (PAN), el nombre del titular de la tarjeta, la fecha de caducidad y el número de código de servicio.
- ¿Dónde se almacenan los datos confidenciales?
- ¿Quién puede acceder a los datos, modificarlos y borrarlos?
- Es esencial conocer los permisos de los usuarios para protegerse de posibles tratamientos indebidos de los datos.

- ¿Quién puede realizar operaciones de creación, lectura, actualización y eliminación (CRUD)?
  - Para tener en cuenta la posible escalada de privilegios, conozca quién puede administrar los permisos de los datos.
- ¿Qué impacto podría tener en la empresa que los datos se divulgasen involuntariamente, se alteraren o se eliminasen?
  - Conozca el riesgo que supone que los datos se modifiquen, eliminen o revelen de forma inadvertida.

Si conoce las respuestas a estas preguntas, podrá tomar las siguientes medidas:

- Reducir el alcance de los datos confidenciales (como el número de ubicaciones de datos confidenciales) y limitar el acceso a los datos confidenciales solo a los usuarios autorizados.
- Conocer los distintos tipos de datos para poder implementar los mecanismos y técnicas de protección de datos adecuados, como el cifrado, la prevención de la pérdida de datos y la administración de identidades y accesos.
- Optimizar los costes proporcionando los objetivos de control adecuados para los datos.
- Responder con confianza a las preguntas de los reguladores y auditores sobre el tipo y la cantidad de datos, y sobre cómo se aíslan entre sí los datos con distintos niveles de confidencialidad.

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

Guía para la implementación

La clasificación de los datos es el acto de identificar el nivel de confidencialidad de los datos. Podría ser necesario etiquetarlos para que la búsqueda y el seguimiento de esos datos sean más sencillos. La clasificación de los datos también reduce la duplicación de datos, lo que puede ayudar a disminuir los costes de almacenamiento y de copias de seguridad al tiempo que acelera el proceso de búsqueda.

Utilice servicios como Amazon Macie para automatizar a escala tanto la detección como la clasificación de datos confidenciales. Otros servicios, como Amazon EventBridge y AWS Config, pueden utilizarse para automatizar la corrección de problemas de seguridad de los datos, como los buckets sin cifrar de Amazon Simple Storage Service (Amazon S3) y volúmenes EBS de Amazon EC2 o recursos de datos sin etiquetar. Para obtener una lista completa de las integraciones de los servicios de AWS, consulte la documentación de EventBridge.

Es posible detectar PII en datos no estructurados, como correos electrónicos de clientes, tickets de soporte, reseñas de productos y redes sociales, mediante Amazon Comprehend, que es un servicio de procesamiento de lenguaje natural (NLP) que utiliza machine learning (ML) para encontrar información y relaciones, como personas, lugares, sentimientos y temas en un texto no estructurado. Para obtener una lista de los servicios de AWS que pueden ayudar a identificar los datos, consulte Common techniques to detect PHI and PII data using AWS services (Técnicas comunes para detectar datos PHI y PII utilizando los servicios de AWS).

Otro método que facilita la clasificación y protección de los datos es el etiquetado de recursos de AWS. El etiquetado le permite asignar metadatos a sus recursos de AWS y puede utilizarlo para administrar, identificar, organizar, buscar y filtrar recursos.

En algunos casos, puede optar por etiquetar recursos enteros (como un bucket de S3), especialmente cuando se espera que una carga de trabajo o un servicio específico almacene procesos o transmisiones de una clasificación de datos ya conocida.

Cuando sea apropiado, puede etiquetar un bucket de S3 en lugar de objetos individuales para facilitar la administración y el mantenimiento de la seguridad.

Pasos para la implementación

Detecte datos confidenciales dentro de Amazon S3:

1. Antes de empezar, asegúrese de que dispone de los permisos adecuados para acceder a la consola de Amazon Macie y a las operaciones de la API. Para obtener más información, consulte Getting started with Amazon Macie (Introducción a Amazon Macie).

2. Utilice Amazon Macie para detectar automáticamente los datos cuando los datos confidenciales residan en Amazon S3.
   - Utilice la guía Getting Started with Amazon Macie (Introducción a Amazon Macie) para configurar un repositorio de resultados de detección de datos confidenciales y crear un trabajo de detección de datos confidenciales.
   - Cómo utilizar Amazon Macie para previsualizar datos confidenciales en buckets de S3.

     De forma predeterminada, Macie analiza los objetos con el conjunto de identificadores de datos administrados que recomendamos para detectar automáticamente los datos confidenciales. Para adaptar el análisis, configure Macie para que utilice identificadores de datos administrados específicos, identificadores de datos personalizados y listas de permitidos cuando detecte automáticamente los datos confidenciales para su cuenta u organización. Para ajustar el alcance del análisis, puede excluir buckets específicos (por ejemplo, buckets de S3 que suelen almacenar datos de registro de AWS).
3. Para configurar y utilizar la detección automatizada de datos confidenciales, consulte Performing automated sensitive data discovery with Amazon Macie (Detección automática de datos confidenciales con Amazon Macie).
4. También puede consultar Automated Data Discovery for Amazon Macie (Detección automática de datos para Amazon Macie).

Detecte datos confidenciales dentro de Amazon RDS:

Para obtener más información sobre la detección de datos en bases de datos de Amazon Relational Database Service (Amazon RDS), consulte Enabling data classification for Amazon RDS database with Macie (Habilitación de la clasificación de datos para bases de datos de Amazon RDS con Macie).

Detecte datos confidenciales dentro de DynamoDB:

- En Detecting sensitive data in DynamoDB with Macie (Detección de datos confidenciales en DynamoDB con Macie), se explica cómo utilizar Amazon Macie para detectar datos confidenciales en tablas de Amazon DynamoDB exportando los datos a Amazon S3 para analizarlos.

Soluciones de los socios de AWS

- Considere la posibilidad de utilizar nuestra amplia AWS Partner Network. Los socios de AWS disponen de exhaustivas herramientas y marcos de conformidad que se integran directamente

con los servicios de AWS. Los socios pueden proporcionarle una solución de gobernanza y conformidad a medida para ayudarle a satisfacer sus necesidades organizativas.

- Para obtener soluciones personalizadas para la clasificación de datos, consulte Data governance in the age of regulation and compliance requirements (Gobernanza de datos en la era de la regulación y los requisitos de conformidad).

Para aplicar automáticamente las normas de etiquetado que adopte su organización, puede crear y desplegar políticas mediante AWS Organizations. Las políticas de etiquetado le permiten especificar reglas que definen los nombres válidos de las claves y qué valores son válidos para cada clave. Puede optar por supervisarlas únicamente, lo que le ofrece la oportunidad de evaluar y limpiar sus etiquetas existentes. Una vez que sus etiquetas cumplan las normas elegidas, puede activar la aplicación de la norma en las políticas de etiquetas para evitar que se creen etiquetas que no las cumplan. Para obtener más información, consulte Securing resource tags used for authorization using a service control policy in AWS Organizations (Proteger las etiquetas de recursos utilizadas para la autorización mediante una política de control de servicios en las organizaciones de AWS) y el ejemplo de política para evitar que las etiquetas se modifiquen, excepto por las entidades principales autorizadas.

- Para comenzar a utilizar las políticas de etiquetas en AWS Organizations, se recomienda encarecidamente seguir el flujo de trabajo de Introducción a las políticas de etiquetas antes de pasar a políticas de etiquetas más avanzadas. Conocer el efecto que tiene asociar una simple política de etiquetas a una sola cuenta antes de extenderla a toda una unidad organizativa (OU) u organización le permite ver los efectos que tiene antes de imponer su cumplimiento. En Introducción a las políticas de etiquetas, encontrará enlaces a instrucciones de tareas relacionadas con políticas más avanzadas.
- Considere la posibilidad de evaluar otros servicios y características de AWS compatibles con la clasificación de datos, que se enumeran en el documento técnico Data Classification (Clasificación de datos).

Recursos

Documentos relacionados:

- Getting Started with Amazon Macie (Introducción a Amazon Macie)
- Automated data discovery with Amazon Macie (Detección automática de datos con Amazon Macie)
- Introducción a las políticas de etiquetas

- Detecting PII entities (Detectar entidades PII)

Blogs relacionados:

- Cómo utilizar Amazon Macie para previsualizar datos confidenciales en buckets de S3.
- Performing automated sensitive data discovery with Amazon Macie (Detección automática de datos confidenciales con Amazon Macie)
- Common techniques to detect PHI and PII data using AWS Services (Técnicas comunes para detectar datos PHI y PII utilizando los servicios de AWS)
- Detecting and redacting PII using Amazon Comprehend (Detección y ocultación de PII utilizando Amazon Comprehend)
- Securing resource tags used for authorization using a service control policy in AWS Organizations (Protección de las etiquetas de recursos utilizadas para la autorización mediante una política de control de servicios en AWS Organizations)
- Enabling data classification for Amazon RDS database with Macie (Habilitación de la clasificación de datos para la base de datos de Amazon RDS con Macie)
- Detecting sensitive data in DynamoDB with Macie (Detección de datos confidenciales en DynamoDB con Macie)
-

Vídeos relacionados:

- Event-driven data security using Amazon Macie (Seguridad de datos basada en eventos utilizando Amazon Macie)
- Amazon Macie for data protection and governance (Amazon Macie para la protección y gobernanza de datos)
- Fine-tune sensitive data findings with allow lists (Optimización de los hallazgos de datos confidenciales con listas de permitidos)

SEC07-BP02 Definir controles de protección de datos

Proteja los datos de acuerdo con su nivel de clasificación. Por ejemplo, proteja los datos clasificados como públicos utilizando recomendaciones relevantes a la vez que protege los datos confidenciales con controles adicionales.

Mediante el uso de etiquetas de recursos, separando cuentas de AWS por nivel de confidencialidad (y potencialmente también según las reservas, enclaves o comunidades de intereses), políticas de IAM, SCP de AWS Organizations, AWS Key Management Service (AWS KMS) y AWS CloudHSM, puede definir e implementar sus políticas de clasificación y protección de datos con cifrado. Por ejemplo, si tiene un proyecto con buckets de S3 que contienen datos muy críticos o instancias de Amazon Elastic Compute Cloud (Amazon EC2) que procesen información confidencial, se les puede asignar la etiqueta `Project=ABC` . Solamente su equipo inmediato sabrá qué significa el código del proyecto; también proporciona una forma de utilizar el control de acceso basado en atributos. Puede definir los niveles de acceso a las claves de cifrado de AWS KMS con políticas y concesiones de claves para garantizar que los servicios adecuados tengan acceso al contenido confidencial a través de un mecanismo seguro. Si va a tomar decisiones de autorización en función de etiquetas, debería asegurarse de que los permisos de las etiquetas se definan convenientemente con políticas de etiquetas en AWS Organizations.

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

Guía para la implementación

- Definir el esquema de identificación y clasificación de datos: la identificación y clasificación de los datos se realiza para evaluar el impacto potencial, el tipo de datos que almacena y quién debe acceder a ellos.
  - Documentación de AWS
- Detectar los controles de AWS disponibles: para los servicios de AWS que está utilizando o pretende utilizar, detecte los controles de seguridad. Muchos servicios tienen una sección de seguridad en su documentación.
  - Documentación de AWS
- Identificar los recursos de cumplimiento de AWS: identifique los recursos que AWS tiene disponibles para ayudarle.
  - https://aws.amazon.com/compliance/

Recursos

Documentos relacionados:

- Documentación de AWS
- Documento técnico sobre clasificación de datos
- Introducción a Amazon Macie

- Falta el texto

Vídeos relacionados:

- Introducción al nuevo Amazon Macie

SEC07-BP03 Automatizar la identificación y la clasificación

La automatización de la identificación y clasificación de datos puede ayudarle a implementar los controles correctos. El uso de la automatización para esto, en lugar del acceso directo de una persona, reduce el riesgo de error y exposición humanos. Debería valorar el uso de una herramienta como Amazon Macie, que usa el machine learning para detectar, clasificar y proteger automáticamente los datos confidenciales en AWS. Amazon Macie reconoce la información confidencial, como la información de identificación personal (PII) o la propiedad intelectual, y le proporciona paneles y alertas que le permiten visualizar cómo se desplazan estos datos o cómo se obtiene acceso a ellos.

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

Guía para la implementación

- Use el inventario de Amazon Simple Storage Service (Amazon S3): el inventario de Amazon S3 es una de las herramientas que puede utilizar para realizar una auditoría y elaborar informes sobre el estado de cifrado y replicación de los objetos.
  - Inventario de Amazon S3
- Plantéese usar Amazon Macie: Amazon Macie utiliza el aprendizaje automático para descubrir y clasificar automáticamente los datos almacenados en Amazon S3.
  - Amazon Macie

Recursos

Documentos relacionados:

- Amazon Macie
- Inventario de Amazon S3
- Documento técnico sobre clasificación de datos
- Introducción a Amazon Macie

Videos relacionados:

- Introducción al nuevo Amazon Macie

SEC07-BP04 Definir la administración del ciclo de vida de los datos

Su estrategia de ciclo de vida definida debería basarse en el nivel de confidencialidad, además de en los requisitos jurídicos y organizativos. Debe tener en cuenta algunos aspectos, como la duración de retención, los procesos de destrucción, la administración del acceso, la transformación o el intercambio de los datos. Al seleccionar una metodología de clasificación de los datos, valore la facilidad de uso frente al acceso. También debería dar cabida a los distintos niveles de acceso y particularidades para implementar un enfoque seguro, pero utilizable, para cada nivel. Utilice siempre un enfoque de defensa en profundidad y reduzca el acceso humano a los datos y los mecanismos de transformación, eliminación o copia de los datos. Por ejemplo, exija a los usuarios que se autentiquen en una aplicación con métodos estrictos, y otorgue a la aplicación, en lugar de a los usuarios, el requisito de permiso de acceso para llevar a cabo una acción a distancia. Además, asegúrese de que los usuarios procedan de una ruta de red de confianza y requieran acceso a las claves de descifrado. Use herramientas como paneles e informes automáticos para proporcionar a los usuarios información de los datos en lugar de proporcionarles acceso directo a los datos.

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

Guía para la implementación

- Identifique los tipos de datos: identifique los tipos de datos que almacena o procesa en su carga de trabajo. Esos datos podrían ser textos, imágenes, bases de datos binarias, etc.

Recursos

Documentos relacionados:

- Documento técnico sobre clasificación de datos
- Introducción a Amazon Macie

Vídeos relacionados:

- Introducción al nuevo Amazon Macie

SEGURIDAD 8. ¿Cómo protege los datos en reposo?

Para proteger los datos en reposo debe implementar varios controles para reducir el riesgo de acceso no autorizado o mala gestión.

Prácticas recomendadas

- SEC08-BP01 Implementar una administración de claves segura
- SEC08-BP02 Aplicar el cifrado en reposo
- SEC08-BP03 Automatizar la protección de los datos en reposo
- SEC08-BP04: Aplicación del control de acceso
- SEC08-BP05: Uso de mecanismos para mantener a las personas alejadas de los datos

SEC08-BP01 Implementar una administración de claves segura

La administración segura de claves incluye el almacenamiento, la rotación, el control de acceso y la supervisión del material de claves necesario para proteger los datos en reposo para su carga de trabajo.

Resultado deseado: un mecanismo de administración de claves escalable, repetible y automatizado. El mecanismo debe proporcionar la capacidad de hacer cumplir el acceso con privilegios mínimos al material de claves y proporcionar el equilibrio correcto entre la disponibilidad, la confidencialidad y la integridad de las claves. Es preciso supervisar el acceso a las claves y el material de claves debe rotarse mediante un proceso automatizado. Las identidades humanas nunca deben tener acceso al material de claves.

Patrones comunes de uso no recomendados:

- Acceso humano a material de claves no cifrado.
- Creación de algoritmos criptográficos personalizados.
- Permisos demasiado amplios para acceder a material de claves.

Beneficios de establecer esta práctica recomendada: al establecer un mecanismo de administración de claves seguro para su carga de trabajo, puede ayudar a proteger su contenido contra el acceso no autorizado. Además, es posible que esté sujeto a requisitos reglamentarios de cifrado de datos. Una solución de administración de claves eficaz puede proporcionar mecanismos técnicos alineados con esas regulaciones para proteger el material de claves.

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

Guía para la implementación

Muchos requisitos reglamentarios y prácticas recomendadas incluyen el cifrado de los datos en reposo como un control de seguridad fundamental. Para satisfacer este control, su carga de trabajo necesita un mecanismo que almacene y administre de forma segura el material de claves utilizado para cifrar los datos en reposo.

AWS ofrece AWS Key Management Service (AWS KMS) para proporcionar un almacenamiento duradero, seguro y redundante para las claves de AWS KMS. Muchos servicios de AWS se integran con AWS KMS para respaldar el cifrado de sus datos. AWS KMS utiliza módulos de seguridad de hardware validados por la norma FIPS 140-2 de nivel 3 para proteger sus claves. No hay ningún mecanismo para exportar claves de AWS KMS en texto sin formato.

Si se despliegan cargas de trabajo mediante una estrategia de cuentas múltiples, se considera una práctica recomendada mantener las claves de AWS KMS en la misma cuenta que la carga de trabajo que las utiliza. En este modelo distribuido, la responsabilidad de administrar las claves de AWS KMS recae en el equipo de aplicaciones. En otros casos de uso, las organizaciones pueden optar por almacenar las claves de AWS KMS en una cuenta centralizada. Esta estructura centralizada requiere políticas adicionales para habilitar el acceso entre cuentas necesario para que la cuenta de carga de trabajo acceda a las claves almacenadas en la cuenta centralizada, pero puede ser más aplicable en casos de uso en los que una sola clave se comparte entre varias Cuentas de AWS.

Independientemente de dónde se almacene el material de claves, el acceso a la clave debe controlarse estrictamente mediante el uso de políticas de claves y políticas de IAM. Las políticas de claves son la forma principal de controlar el acceso a una clave de AWS KMS. Además, las concesiones de claves de AWS KMS pueden proporcionar acceso a servicios de AWS para cifrar y descifrar datos en su nombre. Tómese tiempo para revisar las prácticas recomendadas de control de acceso a sus claves de AWS KMS.

Se recomienda supervisar el uso de claves de cifrado para detectar patrones de acceso inusuales. Las operaciones realizadas con claves administradas por AWS y claves administradas por el cliente almacenadas en AWS KMS pueden registrarse en AWS CloudTrail y deben revisarse periódicamente. Debe prestarse especial atención a la supervisión de los eventos de destrucción de claves. Para mitigar la destrucción accidental o malintencionada de material de claves, los eventos de destrucción de claves no eliminan el material de claves inmediatamente. Los intentos de eliminar claves de AWS KMS están sujetos a un período de esperapredeterminado de 30 días, lo que da tiempo a los administradores para revisar estas acciones y anular la solicitud si es necesario.

La mayoría de los servicios de AWS utilizan AWS KMS de forma transparente para usted; su único requisito es decidir si desea utilizar una clave administrada por AWS o por el cliente. Si la carga de trabajo requiere el uso directo de AWS KMS para cifrar o descifrar datos, la práctica recomendada es utilizar cifrado de sobre para proteger los datos. La SDK de cifrado de AWS puede proporcionar a sus aplicaciones elementos básicos de cifrado del lado del cliente para implementar el cifrado de sobre e integrarse con AWS KMS.

Pasos para la implementación

1. Determine las opciones de administración de claves apropiadas (administradas por AWS o administradas por el cliente) para la clave.
   - Para facilitar el uso, AWS ofrece claves propias de AWS y administradas por AWS para la mayoría de los servicios, que proporcionan la capacidad de cifrado en reposo sin la necesidad de administrar el material de claves o las políticas de claves.
   - Si utiliza claves administradas por el cliente, considere el almacén de claves predeterminado para ofrecer el mejor equilibrio entre agilidad, seguridad, soberanía de datos y disponibilidad. Otros casos de uso podrían exigir el uso de almacenes de claves personalizados con AWS CloudHSM o el almacén de claves externo.
2. Revise la lista de servicios que utiliza para su carga de trabajo para comprender cómo AWS KMS se integra con el servicio. Por ejemplo, las instancias de EC2 pueden usar volúmenes de EBS cifrados, que verifican que las instantáneas de Amazon EBS creadas a partir de esos volúmenes también estén cifradas mediante una clave administrada por el cliente y mitigan la divulgación accidental de datos de instantáneas no cifrados.
   - Cómo los servicios de AWS utilizan AWS KMS
   - Para obtener información detallada sobre las opciones de cifrado que ofrece un servicio de AWS, consulte el tema de cifrado en reposo en la guía del usuario o en la guía para desarrolladores del servicio.
3. Implemente AWS KMS: AWS KMS le permite crear y administrar fácilmente las claves y controlar el uso del cifrado en una gran variedad de servicios de AWS y en sus aplicaciones.
   - Introducción: AWS Key Management Service (AWS KMS)
   - Revise las prácticas recomendadas de control de acceso a sus claves de AWS KMS.
4. Considere AWS Encryption SDK: utilice el AWS Encryption SDK con la integración de AWS KMS cuando la aplicación necesite cifrar datos en el lado del cliente.
   - AWS Encryption SDK

5. Habilite IAM Access Analyzer para revisar y notificar automáticamente si hay políticas de claves de AWS KMS demasiado amplias.
6. Habilite Security Hub CSPM para recibir notificaciones si hay políticas de claves mal configuradas, claves programadas para su eliminación o claves sin la rotación automática habilitada.
7. Determine el nivel de registro adecuado para sus claves de AWS KMS. Como las llamadas a AWS KMS, incluidos los eventos de solo lectura, se registran, los registros de CloudTrail asociados con AWS KMS pueden resultar voluminosos.
    - Algunas organizaciones prefieren segregar la actividad de registro de AWS KMS en una ruta distinta. Para obtener más detalles, consulte la sección de registro de llamadas a la API de AWS KMS con CloudTrail de la guía para desarrolladores de AWS KMS.

Recursos

Documentos relacionados:

- AWS Key Management Service
- Herramientas y servicios de criptografía de AWS
- Protección de los datos de Amazon S3 mediante el cifrado
- Cifrado de sobre
- Compromiso de soberanía digital de AWS
- Demystifying AWS KMS key operations, bring your own key, custom key store, and ciphertext portability
- Detalles criptográficos de AWS Key Management Service

Vídeos relacionados:

- How Encryption Works in AWS
- Securing Your Block Storage on AWS
- AWS data protection: Using locks, keys, signatures, and certificates

Ejemplos relacionados:

- Implement advanced access control mechanisms using AWS KMS

SEC08-BP02 Aplicar el cifrado en reposo

Debe obligar a usar el cifrado de los datos en reposo. El cifrado mantiene la confidencialidad de los datos confidenciales en caso de que se produzca un acceso no autorizado o se divulguen de manera accidental.

Resultado deseado: los datos privados deben cifrarse de forma predeterminada cuando estén en reposo. El cifrado ayuda a mantener la confidencialidad de los datos y proporciona una capa adicional de protección contra la divulgación o exfiltración de datos intencionada o inadvertida. No es posible leer los datos cifrados ni acceder a ellos sin antes descifrarlos. Hay que hacer un inventario y controlar todos los datos almacenados sin cifrar.

Antipatrones usuales:

- No utilizar configuraciones para que el cifrado se realice de forma predeterminada.
- Proporcionar un acceso demasiado permisivo a las claves de descifrado.
- No supervisar el uso de las claves de cifrado y descifrado.
- Almacenar datos sin cifrar.
- Utilizar la misma clave de cifrado para todos los datos, independientemente de su uso, tipos y clasificación.

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

Guía para la implementación

Asigne claves de cifrado a clasificaciones de datos en sus cargas de trabajo. Este enfoque ayuda a proteger contra un acceso excesivamente permisivo si utiliza una única clave de cifrado, o muy pocas, para sus datos (consulte SEC07-BP01 Identificar los datos en su carga de trabajo).

AWS Key Management Service (AWS KMS) se integra con muchos servicios de AWS para facilitar el cifrado de sus datos en reposo. Por ejemplo, en Amazon Simple Storage Service (Amazon S3) puede establecer el cifrado predeterminado en un bucket para que todos los objetos nuevos se cifren automáticamente. Cuando utilice AWS KMS, tenga en cuenta hasta qué punto es necesario restringir los datos. AWS administra y utiliza en su nombre las claves de AWS KMS predeterminadas y controladas por el servicio. En el caso de los datos confidenciales que requieren un acceso detallado a la clave de cifrado subyacente, considere la posibilidad de usar claves administradas por el cliente (CMK). Usted tiene el control total sobre las CMK, incluida la rotación y la administración del acceso mediante el uso de políticas de claves.

Además, Amazon Elastic Compute Cloud (Amazon EC2) y Amazon S3 admiten la aplicación del cifrado mediante la configuración del cifrado predeterminado. Puede utilizar Reglas de AWS Config para comprobar automáticamente que está utilizando cifrado, por ejemplo, para volúmenes de Amazon Elastic Block Store (Amazon EBS), instancias de Amazon Relational Database Service (Amazon RDS) y buckets de Amazon S3.

AWS también proporciona opciones para el cifrado del cliente, lo que le permite cifrar los datos antes de subirlos a la nube. AWS Encryption SDK proporciona una forma de cifrar sus datos mediante el cifrado de sobre. Usted proporciona la clave de encapsulado y AWS Encryption SDK genera una clave de datos única para cada objeto de datos que cifra. Considere la posibilidad de utilizar AWS CloudHSM si necesita un módulo de seguridad de hardware (HSM) administrado de un solo inquilino. AWS CloudHSM le permite generar, importar y administrar claves criptográficas en un HSM validado por FIPS 140-2 nivel 3. Entre los casos de uso de AWS CloudHSM, se incluye la protección de claves privadas para la emisión de una autoridad de certificación (CA) y la habilitación del cifrado de datos transparente (TDE) para bases de datos Oracle. El SDK de cliente de AWS CloudHSM proporciona software que le permite cifrar datos del cliente utilizando claves almacenadas dentro de AWS CloudHSM antes de subir sus datos a AWS. El Amazon DynamoDB Encryption Client también le permite cifrar y firmar elementos antes de subirlos a una tabla de DynamoDB.

Pasos para la implementación

- Aplique el cifrado en reposo para Amazon S3: implemente el cifrado predeterminado de buckets de Amazon S3.

  Configure el cifrado predeterminado para los nuevos volúmenes de Amazon EBS: especifique que desea que todos los volúmenes de Amazon EBS recién creados se creen de forma cifrada, con la opción de utilizar la clave predeterminada que proporciona AWS o una clave que usted cree.

  Configure imágenes de máquina de Amazon (AMI) cifradas: al copiar una AMI existente con cifrado habilitado, se cifran automáticamente las instantáneas y los volúmenes raíz.

  Configure el cifrado de Amazon RDS: configure el cifrado para sus clústeres de base de datos e instantáneas en reposo de Amazon RDS mediante la opción de cifrado.

  Cree y configure claves de AWS KMS con políticas que limiten el acceso a las entidades principales adecuadas para cada clasificación de datos: por ejemplo, cree una clave de AWS KMS para cifrar los datos de producción y otra distinta para cifrar los datos de desarrollo o de prueba. También puede proporcionar acceso a la clave a otras Cuentas de AWS. Considere la posibilidad de tener cuentas diferentes para sus entornos de desarrollo y de producción. Si en su entorno de

producción es necesario descifrar artefactos en la cuenta de desarrollo, puede editar la política de CMK que se utiliza para cifrar los artefactos de desarrollo para otorgar a la cuenta de producción la capacidad de descifrar dichos artefactos. Después, el entorno de producción puede ingerir los datos descifrados para usarlos en producción.

Configure el cifrado en servicios de AWS adicionales: para otros servicios de AWS que utilice, revise la documentación de seguridad de ese servicio para determinar las opciones de cifrado del servicio.

Recursos

Documentos relacionados:

- AWS Crypto Tools
- Documentación de AWS
- AWS Encryption SDK
- Documento técnico Introducción a los detalles criptográficos de AWS KMS
- AWS Key Management Service
- AWS cryptographic services and tools (Herramientas y servicios criptográficos de AWS)
- Cifrado de Amazon EBS
- Cifrado predeterminado para volúmenes de Amazon EBS
- Cifrado de recursos de Amazon RDS
- Habilitación del cifrado predeterminado de bucket de Amazon S3
- Protección de datos de Amazon S3 mediante cifrado

Vídeos relacionados:

- How Encryption Works in AWS (Cómo funciona el cifrado en AWS)
- Securing Your Block Storage on AWS (Protección del almacenamiento en bloque de AWS)

SEC08-BP03 Automatizar la protección de los datos en reposo

Utilice herramientas automatizadas para validar y aplicar continuamente los controles de datos en reposo; por ejemplo, verifique que solo hay recursos de almacenamiento cifrados. Puede automatizar la validación de que todos los volúmenes de EBS están cifrados con Reglas de AWS Config.

AWS Security Hub CSPM también puede verificar varios controles mediante comprobaciones automatizadas con respecto a los estándares de seguridad. Además, su Reglas de AWS Config puede corregir recursos no conformes automáticamente.

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

Guía para la implementación

Datos en reposo representan los datos que se conservan en un almacenamiento no volátil durante cualquier periodo de tiempo en una carga de trabajo. Esto incluye el almacenamiento en bloque, el almacenamiento de objetos, las bases de datos, los archivos, los dispositivos IoT y todo tipo de forma de almacenamiento en la que se conserven los datos. La protección de los datos en reposo reduce el riesgo de acceso no autorizado si se implementan el cifrado y los controles de acceso adecuados.

Aplique el cifrado en reposo: debe asegurarse de que la única forma de almacenar los datos sea mediante el cifrado. AWS KMS se integra perfectamente con muchos servicios de AWS para facilitarle el cifrado de todos sus datos en reposo. Por ejemplo, en Amazon Simple Storage Service (Amazon S3), puede establecer el cifrado predeterminado en un bucket de modo que todos los objetos nuevos se cifran automáticamente. Además, Amazon EC2 y Amazon S3 admiten la aplicación del cifrado mediante la configuración del cifrado predeterminado. Puede usar las reglas de administradas de AWS Config para comprobar automáticamente que está utilizando el cifrado, por ejemplo, para volúmenes de EBS, instancias de Amazon Relational Database Service (Amazon RDS)y buckets de Amazon S3.

Recursos

Documentos relacionados:

- AWS Crypto Tools
- SDK de cifrado de AWS

Vídeos relacionados:

- How Encryption Works in AWS (Cómo funciona el cifrado en AWS)
- Securing Your Block Storage on AWS (Protección del almacenamiento en bloque de AWS)

SEC08-BP04: Aplicación del control de acceso

Para ayudarle a proteger sus datos en reposo, aplique el control de acceso mediante mecanismos como el aislamiento y el control de versiones, y utilice el principio del privilegio mínimo. Impida que se conceda acceso público a sus datos.

Resultado deseado: verificar que solo los usuarios autorizados puedan acceder a los datos en función de su necesidad de utilizarlos. Proteja sus datos con copias de seguridad periódicas y el control de versiones para evitar que se modifiquen o eliminen de forma intencionada o involuntaria. Aísle los datos críticos de otros datos para proteger su confidencialidad e integridad.

Antipatrones usuales:

- Almacenar juntos datos con diferentes requisitos de confidencialidad o clasificación.
- Utilizar permisos demasiado permisivos en las claves de descifrado.
- Clasificar incorrectamente los datos.
- No conservar copias de seguridad detalladas de los datos importantes.
- Proporcionar acceso persistente a los datos de producción.
- No auditar el acceso a los datos ni revisar periódicamente los permisos.

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

Guía para la implementación

Hay muchos controles que pueden ayudar a proteger sus datos en reposo, como el control de acceso (con el privilegio mínimo), el aislamiento y el control de versiones. El acceso a sus datos debe auditarse utilizando mecanismos de detección, como AWS CloudTrail, y registros de nivel de servicio, como los registros de acceso de Amazon Simple Storage Service (Amazon S3). Debe realizar un inventario de los datos a los que se puede acceder públicamente y crear un plan para reducir la cantidad de datos disponibles públicamente a lo largo del tiempo.

El bloqueo de almacenes de Amazon Glacier y el bloqueo de objetos de Amazon S3 proporcionan un control de acceso obligatorio para los objetos de Amazon S3: una vez bloqueada una política de almacenes con la opción de conformidad, ni siquiera el usuario raíz puede cambiarla hasta que venza el bloqueo.

Pasos para la implementación

- Aplique el control de acceso: aplique el control de acceso con privilegios mínimos, incluido el acceso a las claves de cifrado.
- Separe los datos en función de diferentes niveles de clasificación: utilice diferentes Cuentas de AWS para los niveles de clasificación de los datos y administre dichas cuentas mediante AWS Organizations.
- Revise las políticas de AWS Key Management Service (AWS KMS): revise el nivel de acceso concedido en las políticas de AWS KMS.
- Revise los permisos de los objetos y buckets de Amazon S3: revise periódicamente el nivel de acceso otorgado por las políticas de buckets de S3. La práctica recomendada es evitar el uso de buckets de lectura o escritura pública. Plantéese utilizar AWS Config para detectar buckets que están disponibles al público y Amazon CloudFront para ofrecer contenido de Amazon S3. Verifique que los buckets que no deben permitir el acceso público estén configurados correctamente para impedirlo. De manera predeterminada, todos los buckets de S3 son privados y solo permiten el acceso a los usuarios que cuentan con una autorización explícita.
- Habilite AWS IAM Access Analyzer: IAM Access Analyzer analiza los buckets de Amazon S3 y genera un hallazgo cuando una política de S3 concede acceso a una entidad externa.
- Habilite el control de versiones de Amazon S3 y el bloqueo de objetos cuando corresponda.
- Utilice el inventario de Amazon S3: el inventario de Amazon S3 puede utilizarse para auditar e informar sobre el estado de replicación y cifrado de sus objetos de S3.
- Revise los permisos de uso compartido de Amazon EBS y https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/sharing-amis.html: los permisos de uso compartido pueden permitir que las imágenes y los volúmenes se compartan con Cuentas de AWS externas a su carga de trabajo.
- Revise periódicamente los recursos compartidos de AWS Resource Access Manager para determinar si los recursos deben seguir compartiéndose. Resource Access Manager le permite compartir recursos, como las políticas de AWS Network Firewall, las reglas de resolución de Amazon Route 53 y las subredes, dentro de sus Amazon VPC. Audite periódicamente los recursos compartidos y deje de compartir los recursos que ya no sea necesario.

Recursos

Prácticas recomendadas relacionadas:

- SEC03-BP01 Definir los requisitos de acceso

- SEC03-BP02 Conceder acceso con privilegios mínimos

Documentos relacionados:

- Documento técnico Introducción a los detalles criptográficos de AWS KMS
- Introducción a la administración de permisos de acceso a los recursos de Amazon S3
- Información general sobre la administración de acceso a sus recursos de AWS KMS
- Reglas de AWS Config
- Amazon S3 + Amazon CloudFront: A Match Made in the Cloud (Amazon S3 + Amazon CloudFront: una combinación perfecta en la nube)
- Uso del control de versiones
- Usar Bloqueo de objetos de Amazon S3
- Compartir una instantánea de Amazon EBS
- AMI compartidas
- Hosting a single-page application on Amazon S3 (Alojar una aplicación de una sola página en Amazon S3)

Vídeos relacionados:

- Securing Your Block Storage on AWS (Protección del almacenamiento en bloque de AWS)

SEC08-BP05: Uso de mecanismos para mantener a las personas alejadas de los datos

Impida a todos los usuarios acceder directamente a sistemas e información confidenciales en circunstancias de funcionamiento normales. Por ejemplo, utilice un flujo de trabajo de administración de cambios para administrar instancias de Amazon Elastic Compute Cloud (Amazon EC2) con herramientas en lugar de permitir el acceso directo o un host bastión. Esto se puede lograr mediante la automatización de AWS Systems Manager, que usa documentos de automatización que incluyen los pasos que han de seguirse para realizar tareas. Estos documentos se pueden almacenar en el control de código fuente, otros compañeros pueden revisarlos antes de su publicación, y pueden probarse de forma exhaustiva para reducir al mínimo el riesgo en comparación con el acceso al shell. Los usuarios empresariales podrían tener un panel en lugar de acceso directo a un almacén de datos para ejecutar consultas. Allí donde no se utilicen canalizaciones de CI/CD, determine qué controles y procesos son necesarios para ofrecer correctamente un mecanismo de acceso instantáneo que suele estar deshabilitado.

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

Guía para la implementación

- Implemente mecanismos para mantener a las personas alejadas de los datos: entre estos mecanismos se incluyen el uso de paneles, como Quick, para mostrar datos a los usuarios en lugar de realizar consultas directas.
  - Quick
- Automatice la administración de la configuración: realice acciones a distancia, y aplique y valide configuraciones seguras de forma automática mediante el uso de un servicio o herramienta de administración de la configuración. Evite usar hosts bastión o acceder directamente a instancias de EC2.
  - AWS Systems Manager
  - AWS CloudFormation
  - Canalización de CI/CD para plantillas de AWS CloudFormation en AWS

Recursos

Documentos relacionados:

- Documento técnico Detalles criptográficos de AWS KMS

Vídeos relacionados:

- Cómo funciona el cifrado en AWS
- Protección del almacenamiento en bloque de AWS

SEGURIDAD 9. ¿Cómo protege sus datos en tránsito?

Para proteger los datos en tránsito debe implementar varios controles para reducir el riesgo de acceso no autorizado o pérdida.

Prácticas recomendadas

- SEC09-BP01: Implementación de la administración segura de claves y certificados
- SEC09-BP02 Aplicar el cifrado en tránsito
- SEC09-BP03: Automatización de la detección del acceso involuntario a los datos

- SEC09-BP04: Autenticar las comunicaciones de red

SEC09-BP01: Implementación de la administración segura de claves y certificados

Los certificados de seguridad de la capa de transporte (TLS) se utilizan para proteger las comunicaciones de red y establecer la identidad de los sitios web, los recursos y las cargas de trabajo a través de Internet, así como de las redes privadas.

Resultado deseado: un sistema de administración de certificados seguro que puede aprovisionar, desplegar, almacenar y renovar certificados en una infraestructura de clave pública (PKI). Un mecanismo seguro de administración de claves y certificados evita que se divulgue el material de claves privadas del certificado y renueva automáticamente el certificado de forma periódica. También se integra con otros servicios para proporcionar comunicaciones de red e identidad seguras para los recursos de la máquina dentro de su carga de trabajo. Las identidades humanas nunca deben tener acceso al material de claves.

Patrones comunes de uso no recomendados:

- Realizar pasos manuales durante los procesos de despliegue o renovación del certificado.
- No prestar suficiente atención a la jerarquía de la autoridad de certificación (CA) al diseñar una CA privada.
- Usar certificados autofirmados para recursos públicos.

Beneficios de establecer esta práctica recomendada:

- Simplificar la administración de certificados mediante el despliegue y la renovación automatizadas.
- Fomentar el cifrado de los datos en tránsito mediante certificados TLS.
- Aumentar la seguridad y auditabilidad de las medidas de certificación adoptadas por la autoridad de certificación.
- Organizar las tareas de administración en los diferentes capas de la jerarquía de CA.

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

Guía para la implementación

Las cargas de trabajo modernas hacen un uso extensivo de las comunicaciones de red cifradas mediante protocolos PKI como TLS. La administración de certificados de PKI puede ser compleja,

pero el aprovisionamiento, el despliegue y la renovación automatizados de los certificados pueden reducir la fricción asociada con la administración de certificados.

AWS proporciona dos servicios para administrar los certificados de PKI de uso general: AWS Gestor de certificados y AWS Autoridad de certificación privada (AWS Privada CA). ACM es el servicio principal que los clientes utilizan para aprovisionar, administrar y desplegar certificados para su uso tanto en cargas de trabajo de AWS tanto públicas como privadas. ACM emite certificados mediante AWS Privada CA y se integra con muchos otros servicios administrados de AWS para proporcionar certificados TLS seguros para las cargas de trabajo.

AWS Privada CA le permite establecer su propia autoridad de certificación raíz o subordinada y emitir certificados TLS a través de una API. Puede usar este tipo de certificados en situaciones en las que controla y administra la cadena de confianza en el lado del cliente de la conexión TLS. Además de los casos de uso de TLS, AWS Privada CA se puede utilizar para emitir certificados para pods de Kubernetes, atestaciones de productos de dispositivos Matter, firma de código y otros casos de uso con una plantilla personalizada. También puede utilizar Funciones de IAM en cualquier lugar para proporcionar credenciales temporales de IAM a las cargas de trabajo locales a las que se les hayan emitido certificados X.509 firmados por su CA privada.

Además de ACM y AWS Privada CA, AWS IoT Core proporciona soporte especializado para el aprovisionamiento, la administración y el despliegue de certificados de PKI en dispositivos IoT. AWS IoT Core proporciona mecanismos especializados para incorporar dispositivos IoT en su infraestructura de clave pública a escala.

Consideraciones para establecer una jerarquía de CA privada

Si tiene que establecer una CA privada, es importante prestar especial atención para diseñar correctamente la jerarquía de CA desde el principio. Se recomienda desplegar cada nivel de jerarquía de CA en Cuentas de AWS independientes al crear una jerarquía de CA privada. Este paso deliberado reduce el área de superficie de cada nivel de la jerarquía de CA, lo que facilita la detección de anomalías en los datos de registro de CloudTrail y reduce el alcance del acceso o el impacto si se produce un acceso no autorizado a una de las cuentas. La CA raíz debe residir en su propia cuenta independiente y solo debe usarse para emitir uno o más certificados de CA intermedios.

A continuación, cree una o más CA intermedias en cuentas independientes de la cuenta de la CA raíz para emitir certificados para los usuarios finales, los dispositivos u otras cargas de trabajo. Por último, emita certificados desde su CA raíz a las CA intermedias, que a su vez emitirán certificados para sus usuarios finales o dispositivos. Para obtener más información sobre la planificación del

despliegue de la CA y el diseño de la jerarquía de las CA, incluida la planificación de la resiliencia, la replicación entre regiones, el uso compartido de las CA en toda la organización y mucho más, consulte Planificación de la implementación de AWS Privada CA.

Pasos para la implementación

1. Determine los servicios de AWS pertinentes que necesita para su caso de uso:
   - Muchos casos de uso pueden utilizar la infraestructura de clave pública existente de AWS mediante AWS Gestor de certificados. ACM se puede usar para desplegar certificados TLS para servidores web, equilibradores de carga u otros usos para certificados de confianza pública.
   - Considere AWS Privada CA cuando necesite establecer su propia jerarquía de autoridades de certificación privadas o necesite acceder a certificados exportables. ACM se puede utilizar entonces para emitir muchos tipos de certificados de entidad final mediante la AWS Privada CA.
   - Para los casos de uso en los que los certificados se deben aprovisionar a escala para dispositivos de Internet de las cosas (IoT) integrados, considere AWS IoT Core.
2. Implemente la renovación automática de certificados siempre que sea posible:
   - Utilice la renovación administrada de ACM para los certificados emitidos por ACM junto con los servicios administrados de AWS integrados.
3. Establezca registros y registros de auditoría:
   - Habilite los registros de CloudTrail para hacer un seguimiento del acceso a las cuentas que tienen autoridades de certificación. Considere configurar la validación de integridad del archivo de registro en CloudTrail para verificar la autenticidad de los datos de registro.
   - Genere y revise periódicamente informes de auditoría que enumeren los certificados que su CA privada ha emitido o revocado. Estos informes se pueden exportar a un bucket de S3.
   - Al desplegar una CA privada, también tendrá que establecer un bucket de S3 para almacenar la lista de revocación de certificados (CRL). Para obtener instrucciones sobre cómo configurar este bucket de S3 en función de los requisitos de su carga de trabajo, consulte Planificación de una lista de revocación de certificados (CRL).

Recursos

Prácticas recomendadas relacionadas:

- SEC02-BP02 Usar credenciales temporales
- SEC08-BP01 Implementar una administración de claves segura
- SEC09-BP04: Autenticar las comunicaciones de red

Documentos relacionados:

- How to host and manage an entire private certificate infrastructure in AWS
- How to secure an enterprise scale ACM Private CA hierarchy for automotive and manufacturing
- Prácticas recomendadas de CA privada
- How to use AWS RAM to share your ACM Private CA cross-account

Vídeos relacionados:

- Activating AWS Gestor de certificados Private CA (taller)

Ejemplos relacionados:

- Taller de CA privada
- Taller de IoT Device Management (incluido el aprovisionamiento de dispositivos)

Herramientas relacionadas:

- Complemento para el administrador de certificados de Kubernetes para usar AWS Privada CA

SEC09-BP02 Aplicar el cifrado en tránsito

Aplique los requisitos de cifrado definidos en función de las políticas, las obligaciones reglamentarias y las normas de su organización para ayudar a cumplir los requisitos organizativos, legales y de conformidad. Utilice únicamente protocolos con cifrado cuando transmita datos confidenciales fuera de su nube virtual privada (VPC). El cifrado ayuda a mantener la confidencialidad de los datos incluso cuando transitan por redes que no son de confianza.

Resultado deseado: todos los datos deben cifrarse en tránsito utilizando protocolos TLS seguros y conjuntos de cifrado. El tráfico de red entre sus recursos e Internet debe cifrarse para mitigar el acceso no autorizado a los datos. El tráfico de red de su entorno interno de AWS únicamente debe cifrarse utilizando TLS siempre que sea posible. La red interna de AWS se cifra de manera predeterminada y el tráfico de red dentro de una VPC no se puede suplantar ni espiar a menos que una parte no autorizada haya obtenido acceso a cualquier recurso que esté generando tráfico (como las instancias de Amazon EC2 y los contenedores de Amazon ECS). Considere la posibilidad de proteger el tráfico entre redes con una red privada virtual (VPN) IPsec.

Antipatrones usuales:

- Utilizar versiones de SSL, TLS y componentes del conjunto de cifrado obsoletos (por ejemplo, SSL v3.0, claves RSA de 1024 bits y cifrado RC4).
- Permitir tráfico no cifrado (HTTP) hacia o desde recursos destinados al público.
- No supervisar y sustituir los certificados X.509 antes de que caduquen.
- Utilizar certificados X.509 autofirmados para TLS.

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

Guía para la implementación

Los servicios de AWS facilitan puntos de conexión HTTPS con TLS para la comunicación, lo que proporciona cifrado en tránsito al comunicarse con las API de AWS. Los protocolos inseguros, como HTTP, se pueden auditar y bloquear en una VPC mediante el uso de grupos de seguridad. Las solicitudes HTTP también se pueden redirigir automáticamente a HTTPS en Amazon CloudFront o en un Application Load Balancer. Dispone de un control total sobre los recursos informáticos para implementar el cifrado en tránsito en los servicios. También puede usar la conectividad de VPN en la VPC desde una red externa o AWS Direct Connect para facilitar el cifrado de tráfico. Compruebe que sus clientes realizan llamadas a las API de AWS utilizando al menos TLS 1.2, ya que AWS va a dejar de utilizar TLS 1.0 y 1.1 en junio de 2023. Hay soluciones de terceros disponibles en AWS Marketplace si tiene requisitos especiales.

Pasos para la implementación

- Aplique el cifrado en tránsito: los requisitos de cifrado definidos deben basarse en los últimos estándares y prácticas recomendadas, y solo permitir protocolos seguros. Por ejemplo, configure un grupo de seguridad para permitir solamente el protocolo HTTPS a una instancia del equilibrador de carga de aplicaciones o una instancia de Amazon EC2.
- Configure protocolos seguros en los servicios de periferia: configure HTTPS con Amazon CloudFront y utilice un perfil de seguridad apropiado para su postura de seguridad y su caso de uso.
- Utilice una VPN para la conectividad externa: considere la posibilidad de utilizar una VPN IPsec para proteger las conexiones punto a punto o de red a red para ofrecer tanto privacidad como integridad de los datos.

- Configure protocolos seguros en los equilibradores de carga: seleccione una política de seguridad que proporcione los conjuntos de cifrado más seguros que admitan los clientes que se conectarán al agente de escucha. Cree un agente de escucha HTTPS para su Application Load Balancer.
- Configure protocolos seguros en Amazon Redshift: configure su clúster para que requiera una conexión de capa de sockets seguros (SSL) o de seguridad de la capa de transporte (TLS).
- Configure protocolos seguros: revise la documentación del servicio de AWS para determinar las capacidades de cifrado en tránsito.
- Configure el acceso seguro al realizar subidas en los buckets de Amazon S3: utilice los controles de políticas de buckets de Amazon S3 para aplicar el acceso seguro a los datos.
- Considere la posibilidad de utilizar AWS Gestor de certificados: ACM le permite aprovisionar, administrar y desplegar certificados TLS públicos para utilizarlos con los servicios de AWS.
- Considere la posibilidad de utilizar AWS Autoridad de certificación privada para las necesidades de PKI privadas: AWS Privada CA le permite crear jerarquías de autoridades de certificación (CA) privadas para emitir certificados X.509 de entidad final que pueden utilizarse para crear canales TLS cifrados.

Recursos

Documentos relacionados:

- Documentación de AWS
- Uso de HTTPS con CloudFront
- Conectar la VPC a redes remotas mediante AWS Virtual Private Network
- Create an HTTPS listener for your Application Load Balancer (Creación de un agente de escucha HTTPS para Application Load Balancer)
- Tutorial: configure SSL/TLS en Amazon Linux 2
- Uso de SSL/TLS para cifrar una conexión a una instancia de base de datos
- Configuración de las opciones de seguridad para las conexiones

SEC09-BP03: Automatización de la detección del acceso involuntario a los datos

Utilice herramientas, tales como Amazon GuardDuty, para detectar automáticamente actividades sospechosas o intentos de trasladar datos fuera de los límites definidos. Por ejemplo, GuardDuty puede detectar actividad de lectura de Amazon Simple Storage Service (Amazon S3) que sea inusual con la búsqueda Exfiltration:S3/AnomalousBehavior. Además de GuardDuty, los registros de flujo

de Amazon VPC, que capturan información sobre el tráfico de red, pueden utilizarse con Amazon EventBridge para desencadenar la detección de conexiones anómalas, tanto las que se han llegado a establecer como las denegadas. El analizador de acceso de Amazon S3 puede ayudar a evaluar a qué datos pueden acceder ciertos usuarios en los buckets de Amazon S3.

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

Guía para la implementación

- Automatización de la detección del acceso involuntario a los datos: utilice una herramienta o mecanismo de detección para detectar automáticamente los intentos de trasladar datos fuera de los límites definidos; por ejemplo, para detectar un sistema de base de datos que esté copiando datos a un host desconocido.
  - Registros de flujo de VPC
- Plantéese utilizar Amazon Macie: Amazon Macie es un servicio de privacidad y seguridad de datos totalmente administrado que utiliza el machine learning y la coincidencia de patrones para detectar y proteger los datos confidenciales en AWS.
  - Amazon Macie

Recursos

Documentos relacionados:

- Registros de flujo de VPC
- Amazon Macie

SEC09-BP04: Autenticar las comunicaciones de red

Verifique la identidad de las comunicaciones mediante el uso de protocolos que admiten la autenticación, como la seguridad de la capa de transporte (TLS) o IPsec.

Diseñe su carga de trabajo para utilizar protocolos de red seguros y autenticados siempre que haya una comunicación entre servicios, aplicaciones o usuarios. El uso de protocolos de red que admiten autenticación y autorización proporciona un mayor control sobre los flujos de red y reduce la repercusión del acceso no autorizado.

Resultado deseado: una carga de trabajo con flujos de tráfico entre servicios bien definidos en el plano de datos y en el plano de control. Los flujos de tráfico utilizan protocolos de red autenticados y cifrados cuando es técnicamente posible.

Antipatrones usuales:

- Tener tráfico no cifrado o no autenticado en la carga de trabajo.
- Reutilizar credenciales de autenticación para varios usuarios o entidades.
- Confiar únicamente en los controles de red como mecanismo de control de acceso.
- Crear un mecanismo de autenticación personalizado en lugar de confiar en los mecanismos de autenticación estándar del sector.
- Tener un tráfico excesivamente permisivo entre los componentes del servicio u otros recursos de la VPC.

Beneficios de establecer esta práctica recomendada:

- Limita el alcance de la repercusión del acceso no autorizado a una parte de la carga de trabajo.
- Proporciona un nivel de garantía mayor de que las acciones solo las realizan entidades autenticadas.
- Mejora el desacoplamiento de los servicios al definir claramente las interfaces de transferencia de datos previstas y obligar a usarlas.
- Mejora la supervisión, el registro y la respuesta a los incidentes mediante la atribución de solicitudes y unas interfaces de comunicación bien definidas.
- Proporciona una defensa en profundidad para las cargas de trabajo al combinar los controles de red con los controles de autenticación y autorización.

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

Guía para la implementación

Los patrones de tráfico de red de la carga de trabajo se pueden clasificar en dos categorías:

- El tráfico este-oeste representa los flujos de tráfico entre los servicios que constituyen una carga de trabajo.
- El tráfico norte-sur representa los flujos de tráfico entre su carga de trabajo y los consumidores.

Aunque es una práctica común cifrar el tráfico norte-sur, es menos común proteger el tráfico este-oeste mediante protocolos autenticados. Las prácticas de seguridad modernas recomiendan que el diseño de red por sí solo no garantice una relación de confianza entre dos entidades. Cuando dos servicios pueden residir dentro de un límite de red común, sigue siendo una buena práctica recomendada cifrar, autenticar y autorizar las comunicaciones entre esos servicios.

Por ejemplo, las API del servicio de AWS utilizan el protocolo de firma AWS Signature Version 4 (SigV4) para autenticar a la persona que llama, independientemente de la red en la que se origine la solicitud. Esta autenticación garantiza que las API de AWS puedan verificar la identidad que solicitó la acción y, a continuación, esa identidad se pueda combinar con políticas para tomar una decisión de autorización que determine si la acción debe permitirse o no.

Servicios como Amazon VPC Lattice y Amazon API Gateway le permiten usar el mismo protocolo de firma SigV4 para incorporar autenticación y autorización al tráfico este-oeste en sus propias cargas de trabajo. Si los recursos fuera de su entorno de AWS necesitan comunicarse con servicios que requieren autenticación y autorización basadas en Sigv4, puede usar AWS Identity and Access Management (IAM) Roles Anywhere en el recurso que no es deAWS para adquirir credenciales de AWS temporales. Estas credenciales se pueden usar para firmar solicitudes de los servicios que utilizan SigV4 para autorizar el acceso.

Otro mecanismo común para autenticar el tráfico este-oeste es la autenticación mutua de TLS (mTLS). Muchas aplicaciones de internet de las cosas (IoT), aplicaciones de empresa a empresa y microservicios utilizan mTLS para validar la identidad de ambos lados de una comunicación TLS mediante el uso de certificados X.509 del lado del cliente y del lado del servidor. Estos certificados puede emitirlos AWS Autoridad de certificación privada (AWS Privada CA). Puede utilizar servicios como Amazon API Gateway y AWS App Mesh para proporcionar autenticación mTLS para la comunicación entre cargas de trabajo o dentro de ellas. Aunque mTLS proporciona información de autenticación para ambos lados de una comunicación TLS, no tiene un mecanismo de autorización.

Por último, OAuth 2.0 y OpenID Connect (OIDC) son dos protocolos que se suelen utilizar para controlar el acceso de los usuarios a los servicios, pero ahora también se están popularizando para el tráfico de servicio a servicio. API Gateway proporciona un autorizador de token web JSON (JWT) que permite a las cargas de trabajo restringir el acceso a las rutas de la API mediante JWT emitidas por proveedores de identidad OIDC u OAuth 2.0. Los ámbitos OAuth2 pueden utilizarse como fuente para tomar las decisiones de autorización básicas, pero las comprobaciones de autorizaciones siguen teniendo que implementarse en la capa de aplicación, y los ámbitos OAuth2 por sí solos no pueden satisfacer necesidades de autorización más complejas.

Pasos para la implementación

- Defina y documente los flujos de red de su carga de trabajo: el primer paso para implementar una estrategia de defensa en profundidad es definir los flujos de tráfico de la carga de trabajo.
  - Cree un diagrama de flujo de datos en el que se defina claramente cómo se transmiten los datos entre los diferentes servicios que componen su carga de trabajo. Este diagrama es el primer paso para imponer esos flujos a través de canales de red autentificados.
  - Instrumente su carga de trabajo en las fases de desarrollo y prueba para validar que el diagrama de flujo de datos refleje con precisión el comportamiento de la carga de trabajo en la versión ejecutable.
  - Un diagrama de flujo de datos también puede ser útil cuando se realiza un ejercicio de modelado de amenazas, como se describe en «SEC01-BP07 Identificar amenazas y priorizar mitigaciones con un modelo de amenazas».
- Establezca controles de red: considere la posibilidad de usar las capacidades de AWS para establecer controles de red que se ajusten a sus flujos de datos. Aunque los límites de la red no deberían ser el único control de seguridad, estos proporcionan una capa en la estrategia de defensa en profundidad para proteger su carga de trabajo.
  - Use grupos de seguridad para establecer, definir y restringir los flujos de datos entre los recursos.
  - Considere la posibilidad de usar AWS PrivateLink para comunicarse con servicios de AWS y de terceros compatibles con AWS PrivateLink. Los datos que se envían a través de un punto de conexión de la interfaz de AWS PrivateLink permanecen en la estructura de red de AWS y no atraviesan la Internet pública.
- Implemente autenticación y autorización en todos los servicios de su carga de trabajo: elija el conjunto de servicios de AWS más adecuado para proporcionar flujos de tráfico autenticados y cifrados en su carga de trabajo.
  - Considere la posibilidad de usar Amazon VPC Lattice para proteger la comunicación de servicio a servicio. VPC Lattice puede usar la autenticación SigV4 combinada con políticas de autenticación para controlar el acceso de un servicio a otro.
  - Para la comunicación de servicio a servicio mediante mTLS, considere la posibilidad de usar API Gateway o App Mesh. AWS Privada CA se puede usar para establecer una jerarquía de CA privada capaz de emitir certificados para su uso con los mTLs.
  - Al realizar la integración con servicios que utilizan OAuth 2.0 u OIDC, considere la posibilidad de usar API Gateway con el autorizador JWT.

    - Para la comunicación entre la carga de trabajo y los dispositivos de IoT, considere la posibilidad de usar AWS IoT Core, que ofrece varias opciones para el cifrado y la autenticación del tráfico de red.
- Supervise el acceso no autorizado: supervise continuamente los canales de comunicación no deseados, las entidades principales no autorizadas que intentan acceder a los recursos protegidos y otros patrones de acceso inadecuados.
    - Si utiliza VPC Lattice para administrar el acceso a sus servicios, piense en la posibilidad de habilitar y supervisar registros de acceso de VPC Lattice. Estos registros de acceso incluyen información sobre la entidad solicitante, información de red, incluida la VPC de origen y destino, y los metadatos de la solicitud.
    - Considere la posibilidad de habilitar registros de flujo de VPC para capturar los metadatos de los flujos de red y revisarlos periódicamente para detectar anomalías.
    - Consulte la AWS Security Incident Response Guide y la sección «Respuesta ante incidentes» del pilar de seguridad de AWS Well-Architected Framework para obtener más información sobre la planificación, la simulación y la respuesta a los incidentes de seguridad.

Recursos

Prácticas recomendadas relacionadas:

- SEC03-BP07 Analizar el acceso público y entre cuentas
- SEC02-BP02 Usar credenciales temporales
- SEC01-BP07 Identificar amenazas y priorizar mitigaciones con un modelo de amenazas

Documentos relacionados:

- «Evaluating access control methods to secure Amazon API Gateway APIs»
- «Configuración de la autenticación TLS mutua para una API de REST»
- «How to secure API Gateway HTTP endpoints with JWT authorizer»
- «Authorizing direct calls to AWS services using AWS IoT Core credential provider»
- AWS Security Incident Response Guide (Guía de respuesta ante incidentes de seguridad de AWS)

Vídeos relacionados:

- «AWS re:invent 2022: Introducing VPC Lattice»

- «AWS re:invent 2020: Serverless API authentication for HTTP APIs on AWS»

Ejemplos relacionados:

- «Amazon VPC Lattice Workshop»
- Taller «Zero-Trust Episode 1 – The Phantom Service Perimeter»

Respuesta ante incidentes

Pregunta

- SEGURIDAD 10. ¿Cómo anticipa, responde a y se recupera de los incidentes?

SEGURIDAD 10. ¿Cómo anticipa, responde a y se recupera de los incidentes?

Incluso con controles eficaces de detección y prevención, la organización debería continuar implementando mecanismos para responder ante incidentes de seguridad y mitigar su posible impacto. Su preparación afecta considerablemente a la capacidad de los equipos de operar de forma eficaz durante un incidente, de aislar, contener y realizar una investigación forense de los problemas y de restaurar operaciones a un estado conocido correcto. La preparación de las herramientas y el acceso en previsión de un incidente de seguridad, así como la práctica de manera periódica de la respuesta ante incidentes durante simulacros, le ayudarán a asegurarse de que podrá recuperarse con una interrupción mínima en el negocio.

Prácticas recomendadas

- SEC10-BP01 Identificación del personal clave y los recursos externos
- SEC10-BP02: Desarrollar planes de administración de incidentes
- SEC10-BP03: Preparar capacidades forenses
- SEC10-BP04 Desarrollar y probar guías estratégicas de respuesta a incidentes de seguridad
- SEC10-BP05: Aprovisionamiento previo del acceso
- SEC10-BP06: Desplegar las herramientas con anticipación
- SEC10-BP07 Ejecutar simulaciones
- SEC10-BP08 Establecer un marco de trabajo para aprender de los incidentes

SEC10-BP01 Identificación del personal clave y los recursos externos

Identifique las obligaciones legales, el personal y los recursos internos y externos que ayudarían a su organización a responder ante un incidente.

Al definir su enfoque a la hora de responder ante un incidente en la nube, de forma conjunta con otros equipos (como el consejo jurídico, el equipo directivo, las partes interesadas de la empresa y los servicios de asistencia de AWS, entre otros), debe identificar el personal clave, las partes interesadas y los contactos pertinentes. Con el fin de reducir la dependencia y el tiempo de respuesta, asegúrese de que su equipo, los equipos especializados en seguridad y los equipos de intervención cuenten con los conocimientos adecuados sobre los servicios que utiliza y que tengan la oportunidad de realizar una formación práctica.

Le animamos a identificar a los socios de seguridad externos de AWS que puedan ofrecerle una experiencia externa y una perspectiva diferente para aumentar sus capacidades de respuesta. Sus socios de seguridad de confianza pueden ayudarle a identificar posibles riesgos o amenazas con los que puede que no esté familiarizado.

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

Guía para la implementación

- Identifique al personal clave de la organización: Mantenga una lista de contactos del personal de su organización al que tendría que involucrar para responder y recuperarse ante un incidente.
- Identifique a los socios externos: Si es necesario, póngase en contacto con socios externos que puedan ayudarle a responder y a recuperarse ante un incidente.

Recursos

Documentos relacionados:

- AWS Incident Response Guide (Guía de respuesta ante incidentes de AWS)

Vídeos relacionados:

- Prepare for and respond to security incidents in your AWS environment (Cómo prepararse y responder ante incidentes de seguridad en el entorno de AWS)

Ejemplos relacionados:

SEC10-BP02: Desarrollar planes de administración de incidentes

El primer documento que se desarrolla para la respuesta a incidentes es el plan de respuesta a incidentes. El plan de respuesta a incidentes está diseñado para ser la base de su programa y estrategia de respuesta a incidentes.

Beneficios de establecer esta práctica recomendada: desarrollar procesos de respuesta a incidentes exhaustivos y claramente definidos es clave para que el programa de respuesta a incidentes sea satisfactorio y escalable. Cuando se produce un evento de seguridad, tener unos pasos y flujos de trabajo claros le ayudará a responder a tiempo. Es posible que ya tenga procesos de respuesta a incidentes. Independientemente de su estado actual, es importante actualizar, iterar y probar sus procesos de respuesta a incidentes con regularidad.

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

Guía para la implementación

Un plan de administración de incidentes es fundamental para responder y mitigar el impacto potencial de los incidentes de seguridad y recuperarse de él. Un plan de administración de incidentes es un proceso estructurado para identificar y solucionar los incidentes de seguridad y responder a ellos en el momento oportuno.

La nube tiene muchos de los mismos roles y requisitos operativos que se encuentran en un entorno local. A la hora de crear un plan de administración de incidentes, es importante tener en cuenta las estrategias de respuesta y recuperación que mejor se ajusten al resultado empresarial y a los requisitos de conformidad. Por ejemplo, si trabaja con cargas de trabajo en AWS que cumplen con la normativa FedRAMP en Estados Unidos, es útil cumplir con la guía de administración de seguridad informática NIST SP 800-61. Del mismo modo, cuando opere con cargas de trabajo con datos de información de identificación personal (PII) de Europa, considere situaciones como la forma en que podría proteger y responder a los problemas relacionados con la residencia de datos según lo dispuesto por la normativa del Reglamento General de Protección de Datos (RGPD).

Al diseñar un plan de administración de incidentes para sus cargas de trabajo en AWS, comience con el modelo de responsabilidad compartida de AWS para crear un enfoque de defensa en profundidad en la respuesta a incidentes. En este modelo, AWS administra la seguridad de la nube y usted es responsable de la seguridad en la nube. Esto significa que retiene el control y es responsable de los controles de seguridad que decida implementar. La AWS Security Incident Response Guide (Guía de respuesta ante incidentes de seguridad de AWS) expone en detalle los conceptos clave y las orientaciones básicas para crear un plan de administración de incidentes centrado en la nube.

Un plan eficaz de administración de incidentes debe iterarse continuamente, manteniéndose al día con su objetivo de operaciones en la nube. Considere la posibilidad de utilizar los planes de implementación que se detallan a continuación cuando cree y haga evolucionar su plan de administración de incidentes.

Pasos para la implementación

Definición de roles y responsabilidades

La gestión de los eventos de seguridad requiere disciplina en toda la organización y una buena disposición a entrar en acción. Dentro de la estructura organizativa, debe haber muchas personas que tengan responsabilidades y obligaciones, que se consulten o que se mantengan informadas durante un incidente, como los representantes de Recursos Humanos (RR. HH.), el equipo directivo y el departamento legal. Tenga en cuenta estas funciones y responsabilidades y piense si debe participar algún tercero. Tenga en cuenta que, en muchas zonas geográficas, hay leyes locales que rigen lo que se debe y lo que no se debe hacer. Aunque parezca un mero trámite burocrático, elaborar un gráfico de las personas con responsabilidades y obligaciones, las personas que hay que consultar y las personas a las que hay que informar (RACI) para sus planes de respuesta en materia de seguridad facilita una comunicación rápida y directa y deja claro quiénes son los líderes en las diferentes etapas del evento.

Durante un incidente, es fundamental incluir a los propietarios y desarrolladores de las aplicaciones y los recursos afectados, ya que son los expertos en la materia (SME) que pueden proporcionar información y contexto para ayudar a medir el impacto. Asegúrese de establecer relaciones con los desarrolladores y propietarios de las aplicaciones antes de confiar en su experiencia para responder a los incidentes. Es posible que los propietarios de aplicaciones o SME, como los administradores o ingenieros de la nube, tengan que actuar en situaciones en las que el entorno no sea familiar o sea complejo, o a los que las personas encargadas de la respuesta no tengan acceso.

Por último, en la investigación o la respuesta pueden participar socios de confianza, ya que pueden proporcionar experiencia adicional y un control muy valioso. Si no dispone de estas habilidades en su propio equipo, tal vez sea conveniente contratar a una persona externa para que le ayude.

Conozca a los equipos de asistencia y respuesta de AWS

- AWS Support
  - Soporte dispone de una amplia variedad de planes que ofrecen acceso a herramientas y experiencia que respalda el éxito y el buen estado operativo de sus soluciones de AWS. Si

necesita asistencia técnica y más recursos para planificar, desplegar y optimizar su entorno de AWS, puede seleccionar el plan de asistencia que mejor se adapte a su caso de uso de AWS.

  - Utilice el Centro de soporte de Consola de administración de AWS (es necesario iniciar sesión) como punto de contacto central para obtener asistencia en caso de problemas que afecten a sus recursos de AWS. El acceso a Soporte está controlado por AWS Identity and Access Management. Para obtener más información sobre el acceso a las características de Soporte, consulte Getting started with Soporte(Introducción a AWS Support).

- Equipo de respuesta a incidentes de clientes (CIRT) de AWS
  - El equipo de respuesta a incidentes de clientes (CIRT) de AWS es un equipo global de AWS especializado que ofrece asistencia a los clientes las 24 horas del día y los 7 días de la semana durante eventos de seguridad activos en el lado del cliente del modelo de responsabilidad compartida de AWS.
  - Cuando el CIRT de AWS le ofrece asistencia, le ayuda en la clasificación y la recuperación de un evento de seguridad activo en AWS. Puede ayudarle a analizar la causa raíz mediante el uso de registros de servicio de AWS y ofrecerle recomendaciones para la recuperación. También puede proporcionar recomendaciones de seguridad y mejores prácticas para ayudarle a evitar eventos de seguridad en el futuro.
  - Los clientes de AWS pueden interactuar con el CIRT de AWS a través de un caso de Soporte.
- Asistencia en respuestas a DDoS
  - AWS ofrece AWS Shield, que proporciona un servicio de protección contra ataques de denegación de servicio distribuidos (DDoS) administrado que protege las aplicaciones web que se ejecutan en AWS. Shield ofrece detección permanente y mitigaciones automáticas en línea que pueden minimizar el tiempo de inactividad y la latencia de las aplicaciones, por lo que no es necesario utilizar Soporte para beneficiarse de la protección contra ataques DDoS. Hay dos niveles de Shield: AWS Shield Standard y AWS Shield Advanced. Para conocer las diferencias entre estos dos niveles, consulte la documentación de características de Shield.
- AWS Managed Services (AMS)
  - AWS Managed Services (AMS) proporciona una administración continua de su infraestructura de AWS para que pueda centrarse en sus aplicaciones. Al implementar las prácticas recomendadas para mantener su infraestructura, AMS ayuda a reducir sus gastos y riesgos operativos. AMS automatiza actividades comunes, como solicitudes de cambios, monitorización, administración de parches, seguridad y servicios de copia de seguridad, y ofrece servicios de ciclo de vida completo para aprovisionar, ejecutar y ofrecer asistencia a su infraestructura.

- AMS asume la responsabilidad de desplegar un conjunto de controles de detección de seguridad y proporciona una primera línea de respuesta a las alertas las 24 horas del día y los 7 días de la semana. Cuando se inicia una alerta, AMS sigue un conjunto estándar de guías automáticas y manuales para verificar una respuesta coherente. Estas guías de estrategias se comparten con los clientes de AMS durante la incorporación para que puedan desarrollar y coordinar una respuesta con AMS.

Desarrolle el plan de respuesta a incidentes

El plan de respuesta a incidentes está diseñado para ser la base de su programa y estrategia de respuesta a incidentes. El plan de respuesta a incidentes debe figurar en un documento formal. Un plan de respuesta a incidentes suele incluir las siguientes secciones:

- Descripción general del equipo de respuesta a incidentes: describe los objetivos y las funciones del equipo de respuesta a incidentes.
- Funciones y responsabilidades: enumera las partes interesadas de la respuesta a los incidentes y detalla sus funciones cuando se produce un incidente.
- Un plan de comunicación: detalla la información de contacto y cómo se comunica durante un incidente.
- Métodos de comunicación auxiliares: se recomienda tener un método de comunicación auxiliar fuera de banda para informar de los incidentes. Un ejemplo de una aplicación que proporciona un canal de comunicaciones fuera de banda seguro es AWS Wickr.
- Fases de la respuesta a un incidente y medidas a tomar: se enumeran las fases de la respuesta a un incidente (por ejemplo, detección, análisis, erradicación, contención y recuperación), incluidas las medidas de alto nivel que se deben tomar en esas fases.
- Definiciones de gravedad y priorización del incidente: se detalla cómo clasificar la gravedad de un incidente, cómo priorizar el incidente y, a continuación, cómo las definiciones de gravedad afectan a los procedimientos de escalamiento.

Aunque estas secciones son comunes en empresas de diferentes tamaños y de diferentes sectores, el plan de respuesta a incidentes de cada organización es único. Debe elaborar un plan de respuesta a incidentes que mejor se adapte a su organización.

Recursos

Prácticas recomendadas relacionadas:

- SEC04 (¿Cómo detecta e investiga los eventos de seguridad?)

Documentos relacionados:

- AWS Security Incident Response Guide (Guía de respuesta ante incidentes de seguridad de AWS)
- NIST: guía de administración de incidentes de seguridad informática

SEC10-BP03: Preparar capacidades forenses

Antes de que se produzca un incidente de seguridad, considere la posibilidad de desarrollar capacidades forenses que le ayuden a investigar los eventos de seguridad.

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

Los conceptos de la ciencia forense tradicional que se utiliza en el entorno local también son aplicables a AWS Para obtener información clave sobre cómo comenzar a desarrollar capacidades forenses en la Nube de AWS, consulte Forensic investigation environment strategies in the Nube de AWS.

Una vez que haya configurado la estructura del entorno y la Cuenta de AWS para el análisis forense, defina las tecnologías necesarias para ejecutar de forma eficaz unas metodologías sólidas desde el punto de vista forense en las cuatro fases:

- Recopilación: recopile registros de AWS pertinentes, como los registros de AWS CloudTrail, AWS Config, de flujo de VPC y de nivel de host. Siempre que sea posible, recopile instantáneas, copias de seguridad y volcados de memoria de los recursos de AWS afectados.
- Examen: examine los datos recopilados mediante la extracción y la evaluación de la información importante.
- Análisis: analice los datos recopilados para comprender el incidente y sacar conclusiones.
- Informes: presente la información resultante de la fase de análisis.

Pasos para la implementación

Prepare el entorno forense

AWS Organizations le permite administrar y gobernar un entorno de AWS de forma centralizada a medida que aumentan y se escalan los recursos de AWS. Una organización de AWS se encarga

de agrupar las cuentas de Cuentas de AWS para que pueda administrarlas como una sola unidad. Puede usar unidades organizativas (OU) para agrupar las cuentas y administrarlas como una sola unidad.

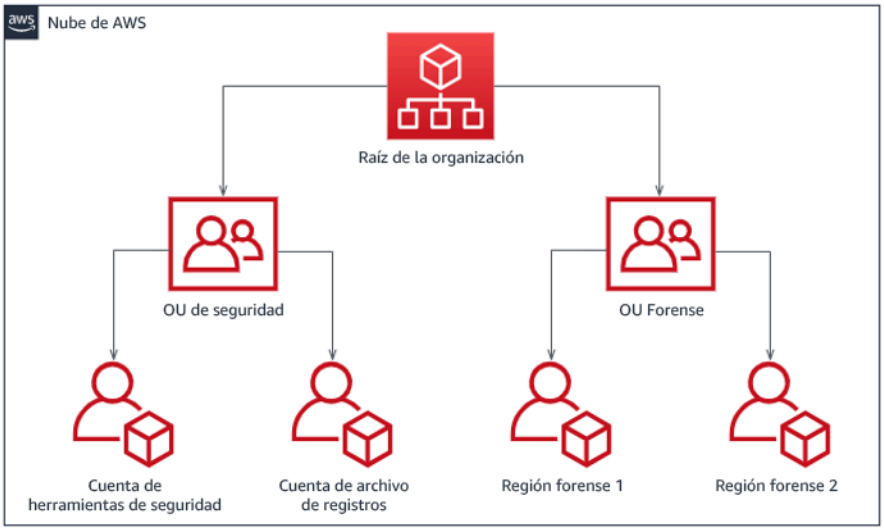
Para la respuesta a incidentes, es útil contar con una estructura de Cuenta de AWS que respalde las funciones de respuesta ante incidentes, lo que incluye una OU de seguridad y una OU forense. Dentro de la unidad organizativa de seguridad, debe tener cuentas para:

- Archivo de registros: agregue los registros en una Cuenta de AWS de archivo de registros con permisos limitados.
- Herramientas de seguridad: centralice los servicios de seguridad en una Cuenta de AWS de herramientas de seguridad. Esta cuenta funciona como un administrador delegado de los servicios de seguridad.

Dentro de la unidad organizativa forense, tiene la opción de implementar una o varias cuentas forenses diferentes para cada una de las regiones en las que opera, en función de lo que le venga mejor a su modelo empresarial y operativo. Si crea una cuenta forense para cada región, puede impedir que se creen recursos de AWS fuera de esa región y reducir el riesgo de que esos recursos se copien en una región no deseada. Por ejemplo, si solo opera en US East (N. Virginia) Region (`us-east-1`) y US West (Oregon) (`us-west-2`), entonces tendría dos cuentas en la OU forense: una para `us-east-1` y otra para `us-west-2`.

Puede crear una Cuenta de AWS forense para varias regiones. Debe tener cuidado al copiar los recursos de AWS en esa cuenta y asegurarse de que cumple los requisitos de soberanía de datos. Dado que aprovisionar nuevas cuentas lleva tiempo, es imperativo crear e instrumentar las cuentas forenses mucho antes de que se produzca un incidente para que los responsables puedan estar preparados y utilizarlas eficazmente en su respuesta.

En el siguiente diagrama, se muestra un ejemplo de una estructura de cuentas que incluye una unidad organizativa forense con cuentas forenses para cada región:



Estructura de cuentas por región para la respuesta a incidentes

Capture copias de seguridad e instantáneas

Crear copias de seguridad de los principales sistemas y bases de datos es fundamental para poder recuperarse de un incidente de seguridad y para fines forenses. Con las copias de seguridad, puede restaurar los sistemas a su estado seguro anterior. En AWS, puede realizar instantáneas de diversos recursos. Las instantáneas le proporcionan copias de seguridad puntuales de esos recursos. Hay muchos servicios de AWS que pueden ayudarle con la copia de seguridad y la recuperación. Para obtener más detalles sobre estos servicios y enfoques de copia de seguridad y recuperación, consulte Backup and Recovery Prescriptive Guidance y Use backups to recover from security incidents.

Es esencial que las copias de seguridad estén bien protegidas, especialmente en ciertas situaciones, como el ransomware. Para obtener instrucciones sobre cómo proteger las copias de seguridad, consulte Top 10 security best practices for securing backups in AWS. Además de proteger las copias de seguridad, debe probar periódicamente los procesos de copia de seguridad y restauración para comprobar que la tecnología y los procesos que tiene implementados funcionan según lo previsto.

Automatice los análisis forenses

Durante un evento de seguridad, es necesario que el equipo de respuesta a incidentes pueda recopilar y analizar las pruebas rápidamente y, al mismo tiempo, mantener la precisión durante todo el tiempo que rodee al evento (por ejemplo, capturar registros relacionados con un evento o recurso específico, o recopilar un volcado de memoria de una instancia de Amazon EC2). Para el equipo de respuesta a incidentes, resulta difícil y lleva mucho tiempo recopilar manualmente

las pruebas pertinentes, especialmente en una gran cantidad de instancias y cuentas. Además, la recopilación manual puede ser más propensa a errores humanos. Por estas razones, debe desarrollar e implementar la automatización del análisis forense en la medida que sea posible.

AWS ofrece una serie de recursos de automatización para el análisis forense, que se enumeran en la sección de recursos siguiente. Estos recursos son ejemplos de patrones forenses que hemos desarrollado y que los clientes han implementado. Aunque pueden resultar útiles como arquitectura de referencia al empezar, valore la posibilidad de modificarlos o crear nuevos patrones de automatización forense en función del entorno, los requisitos, las herramientas y los procesos forenses.

Recursos

Documentos relacionados:

- AWS Security Incident Response Guide - Develop Forensics Capabilities
- AWS Security Incident Response Guide - Forensics Resources
- Forensic investigation environment strategies in the Nube de AWS
- How to automate forensic disk collection in AWS
- AWS Prescriptive Guidance - Automate incident response and forensics

Vídeos relacionados:

- Automating Incident Response and Forensics

Ejemplos relacionados:

- Automated Incident Response and Forensics Framework
- Automated Forensics Orchestrator for Amazon EC2

SEC10-BP04 Desarrollar y probar guías estratégicas de respuesta a incidentes de seguridad

Una parte esencial de la preparación de los procesos de respuesta a incidentes es desarrollar unas guías estratégicas. Las guías estratégicas de respuesta a incidentes recogen una serie de directrices y pasos prescriptivos que deben seguirse cuando se produce un evento de seguridad. Contar con una estructura y unos pasos claros simplifica la respuesta y reduce la probabilidad de que se produzcan errores humanos.

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

Guía para la implementación

Deben crearse guías estratégicas para escenarios de incidentes, por ejemplo:

- Incidentes esperados: deben crearse guías estratégicas para los incidentes que anticipe. Esto puede incluir amenazas como la denegación de servicio (DoS), el ransomware y las amenazas de las credenciales.
- Alertas o resultados de seguridad conocidos: deben crearse guías estratégicas para las alertas y los resultados de seguridad conocidos, como los resultados de GuardDuty. Podría recibir un resultado de GuardDuty y pensar: «¿Y ahora qué?». Si desea evitar que un resultado de GuardDuty se ignore o no se gestione del modo correcto, cree una guía estratégica para cada posible resultado de GuardDuty. Puede encontrar información e instrucciones sobre los procesos de corrección en la documentación de GuardDuty. Conviene señalar que GuardDuty no está habilitado de forma predeterminada y que tiene un coste. Para obtener más detalles sobre GuardDuty, consulte Apéndice A: Definiciones de capacidades en la nube: visibilidad y alertas.

Las guías estratégicas deben incluir los pasos técnicos que los analistas de seguridad deben completar para investigar y responder adecuadamente a un posible incidente de seguridad.

Pasos para la implementación

Algunos de los elementos que deben incluirse en una guía estratégica son:

- Descripción general de la guía estratégica: ¿qué escenario de riesgo o incidente se aborda en este manual de estrategias? ¿Cuál es el objetivo del manual de estrategias?
- Requisitos previos: ¿qué registros, mecanismos de detección y herramientas automatizadas se necesitan en el escenario de este incidente? ¿Cuál es la notificación esperada?
- Información sobre la comunicación y la remisión a instancias superiores: ¿quién participa y cuál es su información de contacto? ¿Cuáles son las responsabilidades de cada una de las partes interesadas?
- Medidas de respuesta: en las diferentes fases de respuesta a un incidente, ¿qué medidas tácticas se deben tomar? ¿Qué consultas deben ejecutar los analistas? ¿Qué código debe ejecutarse para lograr el resultado deseado?
  - Detección: ¿cómo se va a detectar el incidente?
  - Análisis: ¿cómo se va a determinar el alcance del impacto?

  - Contención: ¿cómo se va a aislar el incidente para limitar el alcance?
  - Erradicación: ¿cómo se va a eliminar la amenaza del entorno?
  - Recuperación: ¿cómo se va a conseguir que el sistema o recurso afectado vuelva a ser productivo?
- Resultados esperados: después de ejecutar las consultas y el código, ¿cuál es el resultado esperado de la guía estratégica?

Recursos

Prácticas recomendadas por Well-Architected:

- SEC10-BP02 - Desarrolle planes de gestión de incidentes

Documentos relacionados:

- Framework for Incident Response Playbooks
- Develop your own Incident Response Playbooks
- Incident Response Playbook Samples
- Building an AWS incident response runbook using Jupyter playbooks and CloudTrail Lake

SEC10-BP05: Aprovisionamiento previo del acceso

Verifique que ha aprovisionado previamente el acceso correcto a los equipos de intervención de incidentes en AWS para reducir el tiempo necesario de investigación hasta la recuperación.

Patrones comunes de uso no recomendados:

- Uso de la cuenta raíz para la respuesta ante incidentes.
- Alterar las cuentas de usuario existentes.
- Manipular los permisos de IAM directamente al proporcionar un aumento puntual de los privilegios.

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

Guía para la implementación

AWS recomienda reducir o eliminar la dependencia de credenciales de larga duración siempre que sea posible, en favor de credenciales temporales y mecanismos de aumento puntual de escalada de privilegios. Las credenciales de larga duración están expuestas a riesgos de seguridad y aumentan la carga operativa. Para la mayoría de las tareas de administración, así como para las de respuesta ante incidentes, le recomendamos que implemente la federación de identidades junto con el escalado temporal del acceso administrativo. En este modelo, un usuario solicita el aumento a un nivel superior de privilegios (como un rol de respuesta ante incidentes) y, siempre que el usuario reúna los requisitos para el aumento, se envía una solicitud a un aprobador. Si la solicitud es aprobada, el usuario recibe un conjunto de credenciales de AWS temporales que puede usar para completar sus tareas. Una vez que caducan estas credenciales, el usuario debe enviar una nueva solicitud de aumento.

Recomendamos el uso del escalado temporal de privilegios en la mayoría de las situaciones de respuesta ante incidentes. La forma correcta de hacerlo es utilizar el AWS Security Token Service y políticas de sesión para definir el alcance del acceso.

Hay situaciones en las que las identidades federadas no están disponibles; por ejemplo:

- Interrupción relacionada con un proveedor de identidades (IdP) comprometido.
- Una configuración deficiente o un error humano provocan la ruptura del sistema de administración de acceso federado.
- Actividad maliciosa como un evento de denegación de servicio distribuido (DDoS) o un sistema no disponible.

En los casos anteriores, debe haber un acceso inmediato de emergencia configurado para permitir la investigación y la reparación puntual de los incidentes. También le recomendamos que utilice un usuario de IAM con los permisos adecuados para realizar tareas y acceder a los recursos de AWS. Utilice las credenciales del usuario raíz solo para tareas que requieren el acceso del usuario raíz. Para verificar que los equipos de intervención de incidentes disponen del nivel correcto de acceso a AWS y otros sistemas pertinentes, recomendamos el aprovisionamiento previo de cuentas de usuario exclusivas. Las cuentas de usuario requieren un acceso con privilegios y se deben controlar y supervisar de forma estricta. Las cuentas deben crearse con el menor número de privilegios requeridos para realizar las tareas necesarias y el nivel de acceso debe basarse en las guías de estrategias creadas como parte del plan de administración de incidentes.

La práctica recomendada es crear usuarios y roles personalizados y exclusivos. El hecho de escalar temporalmente el acceso de los usuarios o de los roles mediante la incorporación de políticas de IAM provoca que no esté claro qué acceso tenían los usuarios durante el incidente y se corre el riesgo de que los privilegios escalados no se revoquen.

Es importante eliminar tantas dependencias como sea posible para verificar que se puede acceder en el mayor número posible de escenarios de error. Como medida de apoyo, cree una guía de estrategias para verificar que los usuarios de respuesta ante incidentes se crean como usuarios de AWS Identity and Access Management en una cuenta de seguridad exclusiva y no se administran a través de una federación existente o una solución de inicio de sesión único (SSO). Cada miembro del equipo de intervención debe tener su propia cuenta con nombre. La configuración de la cuenta debe aplicar una política de contraseñas seguras y la autenticación multifactor (MFA). Si las guías de estrategias de respuesta ante incidentes solo requieren acceso a la Consola de administración de AWS, el usuario no debería tener configuradas las claves de acceso y se le debería prohibir explícitamente la creación de claves de acceso. Esto se puede configurar con políticas de IAM o políticas de control de servicios (SCP) como se menciona en las prácticas recomendadas de seguridad de AWS para SCP de AWS Organizations. Los usuarios solo deben tener el privilegio de poder asumir roles de respuesta ante incidentes en otras cuentas.

Durante un incidente, podría ser necesario conceder acceso a otras personas internas o externas para respaldar las actividades de investigación, reparación o recuperación. En este caso, utilice el mecanismo de guía de estrategias mencionado anteriormente. Debe haber un proceso para verificar que cualquier acceso adicional se revoque inmediatamente después de que finalice el incidente.

Para verificar que el uso de los roles de respuesta ante incidentes se puede supervisar y auditar de forma adecuada, es esencial que las cuentas de usuario de IAM creadas para este fin no se compartan con otras personas y que el usuario raíz de Cuenta de AWS no se utilice a menos que se requiera para una tarea específica. Si el usuario raíz es necesario (por ejemplo, no está disponible el acceso de IAM a una cuenta específica), utilice un proceso aparte con una guía de estrategias disponible para verificar la disponibilidad de la contraseña y el token MFA del usuario raíz.

Para configurar las políticas de IAM de los roles de respuesta ante incidentes, utilice IAM Access Analyzer para generar políticas basadas en los registros de AWS CloudTrail. Para ello, conceda acceso de administrador al rol de respuesta ante incidentes en una cuenta que no sea de producción y ejecute las guías de estrategias. Una vez completado, se puede crear una política que únicamente permita las acciones realizadas. Esta política se puede aplicar a los roles de respuesta ante incidentes en todas las cuentas. Es recomendable crear una política de IAM independiente para cada guía de estrategias a fin de facilitar la administración y la auditoría. Entre los ejemplos de guías de

estrategias se podrían incluir planes de respuesta para ransomware, vulneraciones de datos, pérdida de acceso a la producción y otras situaciones.

Utilice las cuentas de usuario de respuesta ante incidentes para asumir los roles de IAM de respuesta ante incidentes exclusivas en otras Cuentas de AWS. Estos roles se deben configurar para que solo puedan asumirlos los usuarios de la cuenta de seguridad. La relación de confianza debe requerir que la entidad principal de llamada se haya autenticado mediante MFA. Los roles deben utilizar políticas de IAM de ámbito estricto para controlar el acceso. Asegúrese de que todas las solicitudes `AssumeRole` para estos roles estén registradas en CloudTrail y se haya alertado de ellas y que se registre cualquier acción realizada con estos roles.

Se recomienda que tanto las cuentas de usuario de IAM como los roles de IAM tengan nombres claros para poder encontrarlos fácilmente en los registros de CloudTrail. Un ejemplo sería asignar a las cuentas de IAM el nombre *<ID_USUARIO>*`-BREAK-GLASS` y los roles de IAM `BREAK-GLASS-ROLE`.

CloudTrail se utiliza para registrar la actividad de API en sus cuentas de AWS y debe utilizarse para configurar alertas sobre el uso de los roles de respuesta ante incidentes. Consulte la publicación del blog sobre la configuración de alertas cuando se utilizan claves de usuario raíz. Las instrucciones se pueden modificar para configurar la métrica Amazon CloudWatch filtro a filtro en los eventos `AssumeRole` relacionados con el rol IAM de respuesta ante incidentes:

```
{ $.eventName = "AssumeRole" && $.requestParameters.roleArn =
 "<ARN_DE_ROL_DE_RESPUESTA_ANTE_INCIDENTES>" && $.userIdentity.invokedBy NOT EXISTS &&
 $.eventType != "AwsServiceEvent" }
```

Como es probable que los roles de respuesta ante incidentes tengan un nivel de acceso alto, es importante que estas alertas lleguen a un grupo amplio y se actúe con rapidez.

Durante un incidente, es posible que un miembro del equipo de intervención necesite acceder a sistemas que no están directamente protegidos por IAM. Pueden ser instancias de Amazon Elastic Compute Cloud, bases de datos de Amazon Relational Database Service o plataformas de software como servicio (SaaS). Se recomienda que en lugar de utilizar protocolos nativos como SSH o RDP, se use AWS Systems Manager Session Manager para todos los accesos administrativos a las instancias de Amazon EC2. Este acceso se puede controlar mediante IAM, que es seguro y está auditado. También se podrían automatizar partes de sus guías de estrategias mediante documentos de AWS Systems Manager Run Command, lo que puede reducir los errores del usuario y mejorar el tiempo de recuperación. Para el acceso a las bases de datos y a las herramientas de terceros,

recomendamos almacenar las credenciales de acceso en AWS Secrets Manager y conceder el acceso a los roles de equipos de intervención ante incidentes.

Por último, la administración de las cuentas de usuario de IAM de respuesta ante incidentes debe agregarse a sus procesos de incorporación, traslado y abandono de los empleados y revisarse y probarse periódicamente para verificar que solo se permite el acceso previsto.

Recursos

Documentos relacionados:

- Managing temporary elevated access to your AWS environment (Administrar el acceso de alto nivel temporal al entorno de AWS)
- AWS Security Incident Response Guide (Guía de respuesta ante incidentes de seguridad de AWS)
- AWS Elastic Disaster Recovery
- Administrador de incidentes de AWS Systems Manager
- Setting an account password policy for IAM users (Establecer una política de contraseñas de cuenta para los usuarios de IAM)
- Using multi-factor authentication (MFA) in AWS (Uso de la autenticación multifactor [MFA] en AWS)
- Configuring Cross-Account Access with MFA (Configuración del acceso entre cuentas con MFA)
- Using IAM Access Analyzer to generate IAM policies (Uso de IAM Access Analyzer para generar políticas de IAM)
- Best Practices for AWS Organizations Service Control Policies in a Multi-Account Environment (Prácticas recomendadas para las políticas de control de servicios de AWS Organizations en un entorno de varias cuentas)
- How to Receive Notifications When Your AWS Account's Root Access Keys Are Used (Cómo recibir notificaciones cuando se utilizan las claves de acceso raíz de su cuenta de AWS)
- Create fine-grained session permissions using IAM managed policies (Crear permisos de sesión detallados mediante políticas administradas de IAM)

Vídeos relacionados:

- Automating Incident Response and Forensics in AWS (Automatización de la respuesta ante incidentes y el análisis forense en AWS)

- DIY guide to runbooks, incident reports, and incident response (Guía paso a paso sobre runbooks, informes de incidentes y respuesta a incidentes)
- Prepare for and respond to security incidents in your AWS environment (Cómo prepararse y responder ante incidentes de seguridad en el entorno de AWS)

Ejemplos relacionados:

- Laboratorio: Configuración de la cuenta y usuario raíz de AWS
- Laboratorio: Respuesta ante incidentes con la consola de AWS y la CLI

SEC10-BP06: Desplegar las herramientas con anticipación

Asegúrese de que el personal de seguridad despliega las herramientas correctas con anticipación para reducir el plazo de investigación hasta conseguir la recuperación.

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

Guía para la implementación

Para automatizar las funciones de operaciones y de respuesta de seguridad, puede utilizar un completo conjunto de API y herramientas de AWS. Puede automatizar totalmente las funcionalidades de administración de identidades, seguridad de red, protección de datos y supervisión, y hacer que estén disponibles a través de métodos de desarrollo de software populares que ya tenga establecidos. Al crear procesos de automatización de seguridad, el sistema podrá supervisar, revisar e iniciar una respuesta, y no necesitará empleados que supervisen el nivel de seguridad y reaccionen manualmente a los eventos.

Si los equipos de intervención de incidentes siguen respondiendo a alertas de la misma forma, corren el riesgo de fatigarse por el excesivo número de alertas. Con el paso del tiempo, el equipo puede llegar a no reaccionar ante las alertas e incluso cometer errores durante la gestión de situaciones habituales o pasar por alto alertas inusuales. La automatización ayuda a evitar este problema con funciones que procesan alertas repetitivas y habituales, dejando a las personas que gestionen los incidentes extraordinarios y delicados. La integración de sistemas de detección de anomalías, como Amazon GuardDuty, AWS CloudTrail Insights y Amazon CloudWatch Anomaly Detection, puede reducir la carga de alertas comunes basadas en umbrales.

Puede mejorar los procesos manuales automatizando los pasos del proceso mediante programación. Después de definir el patrón de solución de un evento, puede descomponer dicho patrón en una lógica procesable y escribir el código que ejecute dicha lógica. A continuación, los equipos de

intervención pueden ejecutar ese código para solucionar el problema. Con el paso del tiempo, puede automatizar cada vez más pasos y, en última instancia, gestionar automáticamente todas las clases de incidentes comunes.

Durante una investigación de seguridad, es necesario que pueda revisar los registros pertinentes para registrar y comprender el alcance completo y la cronología del incidente. También necesita registros para generar alertas que indiquen que se han producido determinadas acciones de interés. Es fundamental seleccionar, habilitar, almacenar y configurar mecanismos de consulta y recuperación, así como de alerta. Además, una forma eficaz de proporcionar herramientas para buscar datos de registro es usar Amazon Detective.

AWS tiene a su disposición más de 200 servicios en la nube y miles de características. Le recomendamos que revise los servicios que pueden respaldar y simplificar su estrategia de respuesta a incidentes.

Además de los registros, debe desarrollar e implementar una estrategia coherente de etiquetado. El etiquetado puede ayudarle a proporcionar contexto en relación con el propósito de un recurso de AWS. El etiquetado también se puede utilizar en la automatización.

Pasos para la implementación

Seleccione y configure registros de análisis y alertas

Consulte la siguiente documentación sobre la configuración de registros para la respuesta a incidentes:

- Estrategias de registro para la respuesta a incidentes de seguridad
- SEC04-BP01 Configurar el registro de servicios y aplicaciones

Habilite los servicios de seguridad para respaldar la detección y la respuesta

AWS ofrece funcionalidades nativas de detección, prevención y respuesta, y se pueden utilizar otros servicios para diseñar soluciones de seguridad personalizadas. Para obtener una lista de los servicios más relevantes para la respuesta a incidentes de seguridad, consulte Definiciones de las capacidades de la nube.

Desarrolle e implemente una estrategia de etiquetado

Puede resultar difícil obtener información contextual sobre el caso de uso empresarial y las partes interesadas internas pertinentes en relación con un recurso de AWS. Una forma de hacerlo es

mediante etiquetas, que asignan metadatos a los recursos de AWS y se componen de una clave y un valor definidos por el usuario. Puede crear etiquetas para clasificar los recursos en función de su propósito, propietario, entorno, tipo de datos procesados y otros criterios de su elección.

Una estrategia de etiquetado coherente puede acelerar los tiempos de respuesta y minimizar el tiempo que se invierte en el contexto de la organización al permitirle identificar y discernir rápidamente la información contextual sobre un recurso de AWS. Las etiquetas también pueden servir como un mecanismo para iniciar automatizaciones de respuesta. Para obtener más detalles sobre qué etiquetar, consulte Tagging your AWS resources. Primero tendrá que definir las etiquetas que desea implementar en toda la organización. Después, implementará y hará cumplir la estrategia de etiquetado. Para obtener más detalles sobre la implementación y su aplicación, consulte Implement AWS resource tagging strategy using AWS Tag Policies and Service Control Policies (SCPs).

Recursos

Prácticas recomendadas por Well-Architected:

- SEC04-BP01 Configurar el registro de servicios y aplicaciones
- SEC04-BP02 Análisis centralizados de registros, hallazgos y métricas

Documentos relacionados:

- Estrategias de registro para la respuesta a incidentes de seguridad
- Incident response cloud capability definitions

Ejemplos relacionados:

- Threat Detection and Response with Amazon GuardDuty and Amazon Detective
- Security Hub Workshop
- Vulnerability Management with Amazon Inspector

SEC10-BP07 Ejecutar simulaciones

Las organizaciones crecen y evolucionan con el tiempo, pero también las amenazas, por lo que es importante que revise continuamente sus capacidades de respuesta a los incidentes. Ejecutar simulaciones (también conocidas como días de juego) es uno de los métodos que se pueden

utilizar para realizar esta evaluación. En las simulaciones, se utilizan escenarios de eventos de seguridad reales diseñados para imitar las tácticas, técnicas y procedimientos (TTP) del actor de una amenaza y permiten a la organización probar y evaluar sus capacidades de respuesta a los incidentes respondiendo a estos simulacros de ataques cibernéticos tal y como podría ocurrir en la realidad.

Ventajas de aplicar esta práctica recomendada: las simulaciones brindan una serie de ventajas:

- Comprobar si se está preparado para un ataque cibernético y mejorar la confianza de los equipos de respuesta a los incidentes.
- Probar la precisión y la eficiencia de las herramientas y los flujos de trabajo.
- Perfeccionar los métodos de comunicación y escalamiento en consonancia con su plan de respuesta a incidentes.
- Ofrecer la oportunidad de responder a vectores menos comunes.

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

Guía para la implementación

Hay tres tipos principales de simulaciones:

- Ejercicios prácticos: el enfoque de los ejercicios prácticos consiste en realizar una sesión de debate en la que participen las diversas partes interesadas en la respuesta a incidentes para practicar las funciones y responsabilidades, y utilizar las herramientas de comunicación y las guías estratégicas establecidas. Por lo general, este ejercicio se puede realizar durante un día completo en un lugar virtual o físico, o bien en una combinación de ambos. Como se trata de un debate, el ejercicio de simulación se centra en los procesos, las personas y la colaboración. La tecnología forma parte integral del debate, pero en este tipo de ejercicio no se hace un uso real de las herramientas o los guiones de respuesta a incidentes.
- Ejercicios del equipo morado: los ejercicios del equipo morado aumentan el nivel de colaboración entre las personas que se encargan de la respuesta a los incidentes (equipo azul) y los actores de las amenazas simuladas (equipo rojo). El equipo azul está compuesto por miembros del centro de operaciones de seguridad (SOC), pero también puede incluir a otras partes interesadas que participarían durante un ataque cibernético real. El equipo rojo está compuesto por un equipo de pruebas de penetración o partes interesadas clave que cuentan con formación en seguridad ofensiva. El equipo rojo trabaja en colaboración con los facilitadores del ejercicio para diseñar un escenario que sea preciso y factible. Durante los ejercicios del equipo morado, la atención se

centra en los mecanismos de detección, las herramientas y los procedimientos operativos estándar (SOP) que facilitan las iniciativas de respuesta a los incidentes.

- Ejercicios del equipo rojo: durante un ejercicio del equipo rojo, el atacante (equipo rojo) realiza una simulación para lograr un determinado objetivo o un conjunto de objetivos en un ámbito predeterminado. Los defensores (equipo azul) no conocen necesariamente el ámbito y la duración del ejercicio; de esta manera, se consigue una evaluación más realista de cómo responderían ante un incidente real. Dado que los ejercicios de equipo rojo pueden ser pruebas invasivas, tenga cuidado e implemente controles para verificar que el ejercicio no produzca un daño real en su entorno.

Considere la posibilidad de realizar simulaciones de ataques cibernéticos con regularidad. Cada tipo de ejercicio puede aportar ventajas únicas para los participantes y la organización en su conjunto, por lo que puede optar por empezar con tipos de simulaciones menos complejos (como los ejercicios prácticos) y pasar luego a los más complejos (ejercicios de equipo rojo). El tipo de simulación se debe elegir en función de su nivel de madurez en seguridad, sus recursos y los resultados deseados. Es posible que algunos clientes opten por no realizar los ejercicios de equipo rojo por su complejidad y su coste.

Pasos para la implementación

Independientemente del tipo de simulación que elija, las simulaciones suelen tener estos pasos de implementación:

1. Defina los elementos básicos del ejercicio: defina el escenario y los objetivos de la simulación. Ambos deben contar con la aceptación de los directivos.
2. Identifique a las principales partes interesadas: como mínimo, en un ejercicio se necesitan facilitadores y participantes. Dependiendo del escenario, podrían participar otras partes interesadas, como los directivos del departamento legal, de comunicaciones o ejecutivo.
3. Cree y pruebe el escenario: es posible que haya que redefinir el escenario a medida que se va creando si algunos elementos específicos no son factibles. Se espera que, al final de esta etapa, haya un escenario definitivo.
4. Facilite la simulación: el tipo de simulación determina la forma de realizarla (un escenario en papel o un escenario simulado muy técnico). Los facilitadores deben adaptar sus tácticas de facilitación a los objetivos del ejercicio y, siempre que sea posible, involucrar a todos los participantes del ejercicio para obtener la mayor ventaja.

5. Desarrolle el informe posterior a la acción (AAR): identifique las áreas que funcionaron bien, las que pueden mejorar y las posibles carencias. El AAR debe medir la eficacia de la simulación, así como la respuesta del equipo al evento simulado, de modo que se pueda seguir su progreso a lo largo del tiempo con futuras simulaciones.

Recursos

Documentos relacionados:

- AWS Incident Response Guide

Vídeos relacionados:

- AWS GameDay - Security Edition

SEC10-BP08 Establecer un marco de trabajo para aprender de los incidentes

La implementación de un marco de trabajo sobre las lecciones aprendidas y una funcionalidad de análisis de la causa raíz no solo ayudará a mejorar las capacidades de respuesta a los incidentes, sino también a evitar que el incidente se repita. Al aprender de cada incidente, puede ayudar a evitar que se repitan los mismos errores, exposiciones o configuraciones incorrectas, lo que no solo mejorará el nivel de seguridad, sino también minimizará el tiempo que se pierde en situaciones evitables.

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

Guía para la implementación

Es importante implementar un marco de trabajo sobre las lecciones aprendidas que establezca y logre, al más alto nivel, los siguientes puntos:

- ¿Cuándo se imparte una lección aprendida?
- ¿Qué implica el proceso de lecciones aprendidas?
- ¿Cómo se lleva a cabo una lección aprendida?
- ¿Quién participa en el proceso y cómo?
- ¿Cómo se van a identificar las áreas de mejora?
- ¿Cómo se va a garantizar que las mejoras se supervisan e implementan de manera efectiva?

El marco no debe centrarse en las personas ni en buscar culpables, sino en mejorar las herramientas y los procesos.

Pasos para la implementación

Además de los resultados generales enumerados anteriormente, es importante asegurarse de que se hacen las preguntas correctas para obtener el máximo valor del proceso (información que conduzca a mejoras viables). Considere la posibilidad de usar estas preguntas para fomentar el debate sobre las lecciones aprendidas:

- ¿Cuál fue el incidente?
- ¿Cuándo se identificó por primera vez el incidente?
- ¿Cómo se identificó?
- ¿Qué sistemas alertaron sobre la actividad?
- ¿Qué sistemas, servicios y datos estaban involucrados?
- ¿Qué ocurrió exactamente?
- ¿Qué funcionó correctamente?
- ¿Qué no funcionó correctamente?
- ¿Qué procesos o procedimientos fallaron o no lograron escalar para responder al incidente?
- ¿Qué se puede mejorar en las siguientes áreas?:
  - Personal
    - ¿Las personas a las que había que contactar estaban realmente disponibles y la lista de contactos estaba actualizada?
    - ¿A las personas les faltaba formación o capacidades necesarias para responder e investigar el incidente de manera eficaz?
    - ¿Los recursos adecuados estaban listos y disponibles?
  - Procesar
    - ¿Se siguieron los procesos y los procedimientos?
    - ¿Los procesos y procedimientos para este (tipo de) incidente estaban documentados y disponibles?
    - ¿Faltaba algún proceso y procedimiento necesario?
    - ¿Los encargados de responder al incidente pudieron acceder oportunamente a la información necesaria para responder al problema?
  - Tecnología

- ¿Los sistemas de alerta existentes identificaron la actividad y alertaron sobre ella eficazmente?
- ¿Cómo podríamos haber reducido el tiempo de detección en un 50 %?
- ¿Es necesario mejorar las alertas existentes o crear nuevas alertas para este (tipo de) incidente?
- ¿Las herramientas existentes permitían investigar (buscar/analizar) el incidente de forma eficaz?
- ¿Qué se puede hacer para poder identificar antes este (tipo de) incidente?
- ¿Qué se puede hacer para ayudar a evitar que este (tipo de) incidente vuelva a ocurrir?
- ¿Quién es el responsable del plan de mejora y cómo comprobará que se ha implementado?
- ¿Qué plazos hay para implementar y probar otros procesos y controles preventivos o de supervisión?

Esta lista no incluye todas las posibilidades. Solo pretende servir como punto de partida para identificar cuáles son las necesidades de la organización y la empresa, y cómo se pueden analizar para aprender lo mejor posible de los incidentes y aumentar continuamente el nivel de seguridad. Lo más importante es empezar incorporando las lecciones aprendidas como un componente estándar del proceso de respuesta a incidentes, la documentación y las expectativas de las partes interesadas.

Recursos

Documentos relacionados:

- AWS Security Incident Response Guide - Establish a framework for learning from incidents
- NCSC CAF guidance - Lessons learned

Seguridad de las aplicaciones

Pregunta

- SEGURIDAD 11. ¿Cómo incorpora y valida las propiedades de seguridad de las aplicaciones durante el ciclo de vida de diseño, desarrollo y despliegue?

SEGURIDAD 11. ¿Cómo incorpora y valida las propiedades de seguridad de las aplicaciones durante el ciclo de vida de diseño, desarrollo y despliegue?

La formación de los usuarios, las pruebas mediante automatización, el conocimiento de las dependencias y la validación de las propiedades de seguridad de herramientas y aplicaciones contribuyen a reducir la probabilidad de que se produzcan problemas de seguridad en las cargas de trabajo de producción.

Prácticas recomendadas

- SEC11-BP01 Formar en seguridad de las aplicaciones
- SEC11-BP02 Automatizar las pruebas a lo largo del ciclo de vida de desarrollo y lanzamiento
- SEC11-BP03 Realizar pruebas de penetración periódicas
- SEC11-BP04 Revisiones manuales del código
- SEC11-BP05 Centralizar los servicios para paquetes y dependencias
- SEC11-BP06 Desplegar software mediante programación
- SEC11-BP07 Evaluar periódicamente las propiedades de seguridad de las canalizaciones
- SEC11-BP08 Crear un programa que integre la propiedad de la seguridad en los equipos de la carga de trabajo

SEC11-BP01 Formar en seguridad de las aplicaciones

Ofrezca formación a los creadores de su organización sobre las prácticas habituales para el desarrollo y el funcionamiento seguros de las aplicaciones. La adopción de prácticas de desarrollo centradas en la seguridad contribuye a reducir la probabilidad de que surjan problemas que solo se detectan en la fase de revisión de la seguridad.

Resultado deseado: El software debe diseñarse y crearse teniendo en cuenta la seguridad. Cuando los creadores de una organización reciben formación sobre prácticas de desarrollo seguras que parten de un modelo de amenazas, mejora la calidad y la seguridad general del software producido. Este planteamiento puede acortar el tiempo hasta la entrega del software o de las características, ya que se reduce la necesidad de tener que volver a repetir los procesos tras la fase de revisión de la seguridad.

A efectos de esta práctica recomendada, el desarrollo seguro se refiere al software que se está escribiendo y a las herramientas o sistemas que prestan soporte al ciclo de vida de desarrollo del software (SDLC).

Patrones comunes de uso no recomendados:

- Esperar a una revisión de seguridad para estudiar las propiedades de seguridad de un sistema.
- Dejar todas las decisiones de seguridad en manos del equipo de seguridad.
- No comunicar claramente cómo se relacionan las decisiones tomadas en el SDLC con las expectativas o políticas generales de seguridad de la organización.
- Intervenir demasiado tarde en el proceso de revisión de la seguridad.

Beneficios de establecer esta práctica recomendada:

- Entender mejor los requisitos de la organización en materia de seguridad en una fase temprana del ciclo de desarrollo.
- Poder identificar y corregir más rápidamente los posibles problemas de seguridad, lo que se traduce en una entrega más rápida de las características.
- Mejora de la calidad del software y los sistemas.

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

Guía para la implementación

Proporcione formación a los creadores de su organización. Una buena base para iniciar la formación sobre seguridad es empezar con un curso sobre modelado de amenazas. Lo ideal sería que los creadores pudieran acceder por sí mismos a la información relevante para sus cargas de trabajo. Este acceso les ayuda a tomar decisiones informadas sobre las propiedades de seguridad de los sistemas que crean sin necesidad de preguntar a otro equipo. El proceso de solicitud de revisiones al equipo de seguridad debe estar claramente definido y ser fácil de seguir. Los pasos del proceso de revisión deben incluirse en la formación sobre seguridad. Cuando se disponga de patrones o plantillas de implementación, deben ser fáciles de encontrar y vincular a los requisitos generales de seguridad. Plantéese el uso de AWS CloudFormation, Componentes de Kit de desarrollo de la nube de AWS (AWS CDK), Service Catalog u otras herramientas basadas en plantillas para reducir la necesidad de configuración personalizada.

Pasos para la aplicación

- Empiece por ofrecer a los creadores un curso sobre modelado de amenazas para sentar una buena base y ayudarles a formarse en cómo pensar en la seguridad.
- Ofrezca acceso a formación para socios de AWS, sector o Formación de AWS y Certification.

- Ofrezca formación sobre el proceso de revisión de la seguridad de su organización, que aclare el reparto de responsabilidades entre el equipo de seguridad, los equipos de carga de trabajo y otras partes interesadas.
- Publique guías de autoservicio sobre cómo cumplir sus requisitos de seguridad, incluidos ejemplos de código y plantillas, si están disponibles.
- Obtenga comentarios periódicamente de los equipos de creadores sobre su experiencia con el proceso de formación y revisión de la seguridad, y utilícelos para mejorar.
- Utilice días de juegos o campañas de detección de errores para reducir el número de problemas y mejorar las competencias de los creadores.

Recursos

Prácticas recomendadas relacionadas:

- SEC11-BP08 Crear un programa que integre la propiedad de la seguridad en los equipos de la carga de trabajo

Documentos relacionados:

- Formación de AWS y Certification
- How to think about cloud security governance (Cómo concebir la gobernanza de la seguridad en la nube)
- How to approach threat modeling (Cómo abordar el modelado de amenazas)
- Accelerating training – The AWS Skills Guild (Acelere la formación: AWS Skills Guild)

Vídeos relacionados:

- Proactive security: Considerations and approaches (Seguridad proactiva: consideraciones y estrategias)

Ejemplos relacionados:

- Workshop on threat modeling (Taller de modelado de amenazas)
- Industry awareness for developers (Concienciación del sector para desarrolladores)

Servicios relacionados:

- AWS CloudFormation
- Componentes de Kit de desarrollo de la nube de AWS (AWS CDK) (AWS CDK)
- Service Catalog
- AWS BugBust

SEC11-BP02 Automatizar las pruebas a lo largo del ciclo de vida de desarrollo y lanzamiento

Automatice las pruebas de las propiedades de seguridad a lo largo del ciclo de vida de desarrollo y lanzamiento. La automatización facilita la identificación coherente y repetible de posibles problemas en el software antes de su lanzamiento, lo que reduce el riesgo de problemas de seguridad en el software suministrado.

Resultado deseado: El objetivo de las pruebas automatizadas es proporcionar una forma programática de detectar problemas de forma temprana y frecuente a lo largo del ciclo de vida de desarrollo. Al automatizar las pruebas de regresión, puede volver a ejecutar pruebas funcionales y no funcionales para verificar que el software probado previamente siga funcionando como se esperaba después de un cambio. Cuando se definen pruebas unitarias de seguridad para detectar errores de configuración habituales, como autenticación dañada o ausente, es posible identificar y solucionar estos problemas en una fase temprana del proceso de desarrollo.

La automatización de pruebas utiliza casos de prueba creados específicamente para la validación de aplicaciones, basados en los requisitos de la aplicación y la funcionalidad deseada. El resultado de las pruebas automatizadas se basa en la comparación de los resultados de las pruebas generados con los resultados esperados, lo que agiliza el ciclo de vida de las pruebas. Las metodologías de pruebas como las pruebas de regresión y los conjuntos de pruebas unitarias son las más adecuadas para la automatización. La automatización de las pruebas de las propiedades de seguridad permite a los creadores recibir información automatizada sin tener que esperar a una revisión de seguridad. Las pruebas automatizadas en forma de análisis de código estático o dinámico permiten aumentar la calidad del código y contribuyen a detectar posibles problemas de software en una fase temprana del ciclo de vida de desarrollo.

Patrones comunes de uso no recomendados:

- No comunicar los casos de prueba y los resultados de las pruebas automatizadas.
- Realizar las pruebas automatizadas solo justo antes del lanzamiento.
- Automatizar casos de prueba con requisitos que cambian con frecuencia.

- No proporcionar orientación sobre cómo abordar los resultados de las pruebas de seguridad.

Beneficios de establecer esta práctica recomendada:

- Menor dependencia de las personas que evalúan las propiedades de seguridad de los sistemas.
- La obtención de resultados coherentes en numerosos flujos de trabajo mejora la coherencia general.
- Menos probabilidades de que se introduzcan problemas de seguridad en el software de producción.
- Reducción del intervalo de tiempo entre la detección y la corrección gracias a la detección temprana de los problemas de software.
- Mayor visibilidad del comportamiento sistémico o repetido en numerosos flujos de trabajo, que puede servir para impulsar mejoras en toda la organización.

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

Guía para la implementación

A medida que crea el software, adopte diversos mecanismos de prueba de software para asegurarse de probar tanto los requisitos funcionales, basados en la lógica empresarial, como los requisitos no funcionales, que se centran en la fiabilidad, el rendimiento y la seguridad de su aplicación.

Las pruebas de seguridad de aplicaciones estáticas (SAST) analizan el código fuente para revelar patrones de seguridad anómalos y proporcionan indicios de código propenso a errores. Las pruebas SAST se basan en datos estáticos, como la documentación (especificación de requisitos, documentación de diseño y especificaciones de diseño) y el código fuente de la aplicación, con objeto de encontrar una serie de problemas de seguridad conocidos. Los analizadores de código estático pueden ayudar a agilizar el análisis de grandes volúmenes de código. El NIST Quality Group ofrece una comparación de analizadores de seguridad del código fuente, que abarca herramientas de código abierto para analizadores de código de bytes y analizadores de código binario.

Complemente las pruebas estáticas con metodologías de pruebas de seguridad de análisis dinámico (DAST), que efectúan pruebas de la aplicación en ejecución a fin de identificar comportamiento potencialmente inesperado. Las pruebas dinámicas pueden utilizarse para detectar problemas potenciales que no son evidentes mediante el análisis estático. Las pruebas en las etapas de repositorio de código, compilación y canalización le permiten comprobar si existen diferentes tipos de problemas potenciales que podrían introducirse en el código. Amazon CodeWhisperer proporciona

recomendaciones de código, incluido el análisis de seguridad, en el IDE del creador. Amazon CodeGuru Reviewer puede identificar problemas cruciales, problemas de seguridad y errores difíciles de detectar durante el desarrollo de la aplicación, y proporciona recomendaciones para mejorar la calidad del código.

El taller de seguridad para desarrolladores usa herramientas de desarrollo de AWS, como AWS CodeBuild, AWS CodeCommit y AWS CodePipeline, para la automatización de canalizaciones de lanzamiento que incluyen las metodologías de prueba SAST y DAST.

A medida que avance en el SDLC, establezca un proceso iterativo que incorpore revisiones periódicas de las aplicaciones con su equipo de seguridad. Los comentarios recogidos en estas revisiones de seguridad deben abordarse y validarse como parte de la revisión de la preparación para el lanzamiento. Estas revisiones establecen una sólida postura de seguridad de la aplicación y proporcionan a los desarrolladores información práctica para afrontar posibles problemas.

Pasos para la aplicación

- Implemente herramientas coherentes de IDE, revisión de código y CI/CD que incluyan pruebas de seguridad.
- Considere en qué momento del SDLC es apropiado bloquear las canalizaciones en lugar de limitarse a notificar a los creadores que es necesario solucionar los problemas.
- El taller de seguridad para desarrolladores ofrece un ejemplo de integración de pruebas estáticas y dinámicas en un proceso de lanzamiento.
- La realización de pruebas o análisis de código mediante herramientas automatizadas, como Amazon CodeWhisperer integrado con los IDE de los desarrolladores y Amazon CodeGuru Reviewer para escanear código al confirmar, ayuda a los desarrolladores a obtener información en el momento adecuado.
- Si usa AWS Lambda para la compilación, puede utilizar Amazon Inspector para analizar el código de la aplicación en sus funciones.
- El taller de CI/CD de AWS proporciona un punto de partida la crear canalizaciones de CI/CD en AWS.
- Cuando se incluyen pruebas automatizadas en las canalizaciones de CI/CD, es preciso utilizar un sistema de tickets para realizar un seguimiento de la notificación y corrección de problemas de software.
- En el caso de las pruebas de seguridad que puedan generar hallazgos, la vinculación a orientaciones para la corrección ayuda a los creadores a mejorar la calidad del código.

- Analice periódicamente los resultados de las herramientas automatizadas para dar prioridad a la siguiente automatización, la formación de los creadores o la campaña de concienciación.

Recursos

Documentos relacionados:

- Entrega continua e implementación continua

- Socios con competencias en DevOps de AWS
- Socios con competencia en seguridad de AWS para la seguridad de las aplicaciones
- Choosing a Well-Architected CI/CD approach (Elección de un enfoque CI/CD bien diseñado)
- Monitoring CodeCommit events in Amazon EventBridge and Amazon CloudWatch Events (Supervisión de eventos de CodeCommit en Amazon EventBridge y Amazon CloudWatch Events)
- Secrets detection in Amazon CodeGuru Review (Revisión de la detección de secretos en Amazon CodeGuru)
- Accelerate deployments on AWS with effective governance (Acelerar los despliegues en AWS con una gobernanza eficaz)
- How AWS approaches automating safe, hands-off deployments (Cómo AWS aborda la automatización de despliegues seguros y sin intervención)

Vídeos relacionados:

- Hands-off: Automating continuous delivery pipelines at Amazon (Sin intervención: automatización de canalizaciones de entrega continua en Amazon)
- Automating cross-account CI/CD pipelines (Automatización de canalizaciones CI/CD entre cuentas)

Ejemplos relacionados:

- Industry awareness for developers (Concienciación del sector para desarrolladores)
- AWS CodePipeline Governance (Gobernanza de AWS CodePipeline) (GitHub)
- Security for Developers workshop (Taller de seguridad para desarrolladores)
- AWS CI/CD Workshop (Taller de CI/CD de AWS)

SEC11-BP03 Realizar pruebas de penetración periódicas

Realice pruebas de penetración periódicas de su software. Este mecanismo ayuda a identificar posibles problemas de software que no pueden detectarse mediante pruebas automatizadas o una revisión manual del código. También puede ayudarle a comprender la eficacia de sus controles de detección. Las pruebas de penetración deben tratar de determinar si se puede hacer que el software realice operaciones inesperadas, como exponer datos que deberían estar protegidos o conceder permisos más amplios de lo esperado.

Resultado deseado: Las pruebas de penetración se utilizan para detectar, remediar y validar las propiedades de seguridad de la aplicación. Las pruebas de penetración periódicas y programadas deben formar parte del ciclo de vida de desarrollo de software (SDLC). Los hallazgos de las pruebas de penetración deben resolverse antes del lanzamiento del software. Debe analizar los resultados de las pruebas de penetración para identificar si hay problemas que podrían detectarse mediante la automatización. El uso de un proceso de pruebas de penetración periódicas y repetibles que incluya un mecanismo de retroalimentación activo ayuda a orientar a los creadores y mejora la calidad del software.

Patrones comunes de uso no recomendados:

- Hacer pruebas de penetración solo para problemas de seguridad conocidos o frecuentes.
- Hacer pruebas de penetración de aplicaciones sin herramientas ni bibliotecas de terceros dependientes.
- Hacer pruebas de penetración solo para problemas de seguridad de paquete, sin evaluar la lógica empresarial implementada.

Beneficios de establecer esta práctica recomendada:

- Mayor confianza en las propiedades de seguridad del software antes de su lanzamiento.
- Oportunidad de identificar los patrones de aplicación preferidos, lo que conduce a una mayor calidad del software.
- Un ciclo de retroalimentación que identifica en una fase más temprana del ciclo de desarrollo dónde la automatización o la formación adicional podrían mejorar las propiedades de seguridad del software.

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

Guía para la implementación

Las pruebas de penetración son un ejercicio estructurado de pruebas de seguridad en el que se ejecutan escenarios planificados de violación de la seguridad para detectar, remediar y validar los controles de seguridad. Las pruebas de penetración comienzan con el reconocimiento, durante el cual se recopilan datos basados en el diseño actual de la aplicación y sus dependencias. Luego, se elabora y ejecuta una lista seleccionada de escenarios de pruebas de seguridad. El objetivo principal de estas pruebas es descubrir problemas de seguridad en la aplicación, que podrían aprovecharse para obtener acceso no deseado a su entorno o acceso no autorizado a los datos. Debe llevar a cabo pruebas de penetración cuando lance nuevas características, o siempre que la aplicación haya sufrido cambios importantes en su funcionamiento o implementación técnica.

Debe identificar la etapa más apropiada del ciclo de vida de desarrollo en el que realizar las pruebas de penetración. Estas pruebas deben hacerse lo bastante tarde como para que la funcionalidad del sistema se aproxime al estado de lanzamiento previsto, pero con tiempo suficiente para solucionar cualquier problema.

Pasos para la aplicación

- Tenga un proceso estructurado para determinar el alcance de las pruebas de penetración. Basar este proceso en el modelo de amenazas es una buena forma de mantener el contexto.
- Identifique la etapa más apropiada del ciclo de desarrollo en el que realizar las pruebas de penetración. Debería ser cuando se espera un cambio mínimo en la aplicación, pero con tiempo suficiente para llevar a cabo la corrección.
- Forme a sus creadores sobre qué esperar de los resultados de las pruebas de penetración y cómo obtener información sobre la corrección.
- Utilice herramientas para acelerar el proceso de las pruebas de penetración mediante la automatización de pruebas habituales o repetibles.
- Analice los resultados de las pruebas de penetración con vistas a identificar problemas de seguridad sistémicos y utilice estos datos para efectuar pruebas automatizadas adicionales y para la formación continua de los creadores.

Recursos

Prácticas recomendadas relacionadas:

- SEC11-BP01 Formar en seguridad de las aplicaciones
- SEC11-BP02 Automatizar las pruebas a lo largo del ciclo de vida de desarrollo y lanzamiento

Documentos relacionados:

- Las pruebas de penetración de AWS ofrecen orientación detallada sobre las pruebas de penetración en AWS
- Accelerate deployments on AWS with effective governance (Acelerar los despliegues en AWS con una gobernanza eficaz)
- Socios con competencia en seguridad de AWS
- Modernize your penetration testing architecture on AWS Fargate (Modernice su arquitectura de pruebas de penetración en AWS Fargate)
- AWS Fault injection Simulator

Ejemplos relacionados:

- Automate API testing with AWS CodePipeline (Automatización de las pruebas de API con AWS CodePipeline) (GitHub)
- Automated security helper (Ayudante de seguridad automatizado) (GitHub)

SEC11-BP04 Revisiones manuales del código

Realice una revisión manual del código del software que produce. Este proceso ayuda a verificar que la persona que ha escrito el código no es la única que comprueba su calidad.

Resultado deseado: La inclusión de un paso de revisión manual del código durante el desarrollo aumenta la calidad del software que se está escribiendo, ayuda a mejorar las competencias de los miembros con menos experiencia del equipo y da la oportunidad de identificar los puntos en los que se puede utilizar la automatización. Las revisiones manuales del código pueden apoyarse en herramientas y pruebas automatizadas.

Patrones comunes de uso no recomendados:

- No revisar el código antes del despliegue.
- Tener una misma persona que escriba y revise el código.
- No utilizar la automatización para ayudar u organizar las revisiones del código.
- No formar a los creadores sobre la seguridad de las aplicaciones antes de que revisen el código.

Beneficios de establecer esta práctica recomendada:

- Mayor calidad del código.
- Mayor coherencia en el desarrollo del código gracias a la reutilización de estrategias comunes.
- Reducción del número de problemas revelados durante las pruebas de penetración y etapas posteriores.
- Mejora de la transferencia de conocimientos dentro del equipo.

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

Guía para la implementación

La etapa de revisión debe implementarse como parte del flujo general de gestión del código. Los pormenores dependen del planteamiento utilizado para la bifurcación, las solicitudes de incorporación de cambios y la fusión. Utilice AWS CodeCommit o soluciones de terceros como GitHub, GitLab o Bitbucket. Sea cual sea el método que utilice, es importante verificar que sus procesos requieren la revisión del código antes de desplegarlo en un entorno de producción. El uso de herramientas como Amazon CodeGuru Reviewer puede facilitar la organización del proceso de revisión del código.

Pasos para la aplicación

- Implemente un paso de revisión manual como parte del flujo de administración de código y realice esta revisión antes de continuar.
- Considere Amazon CodeGuru Reviewer para administrar y ayudar en las revisiones de código.
- Implemente un flujo de aprobación que exija que se complete una revisión del código antes de que este pueda pasar a la siguiente etapa.
- Compruebe que existe un proceso para identificar los problemas encontrados durante las revisiones manuales del código que podrían detectarse automáticamente.
- Integre el paso de revisión manual del código de forma que se ajuste a sus prácticas de desarrollo de código.

Recursos

Prácticas recomendadas relacionadas:

- SEC11-BP02 Automatizar las pruebas a lo largo del ciclo de vida de desarrollo y lanzamiento

Documentos relacionados:

- Working with pull requests in AWS CodeCommit repositories (Trabajo con solicitudes de incorporación de cambios en repositorios de AWS CodeCommit)
- Working with approval rule templates in AWS CodeCommit (Trabajar con plantillas de reglas de aprobación en AWS CodeCommit)
- About pull requests in GitHub (Acerca de las solicitudes de incorporación de cambios en GitHub)
- Automate code reviews with Amazon CodeGuru Reviewer (Revisiones automáticas de código con Amazon CodeGuru Reviewer)
- Automating detection of security vulnerabilities and bugs in CI/CD pipelines using Amazon CodeGuru Reviewer CLI (Automatización de la detección de vulnerabilidades y errores de seguridad en los procesos CI/CD mediante la CLI de Amazon CodeGuru Reviewer)

Vídeos relacionados:

- Continuous improvement of code quality with Amazon CodeGuru (Mejora continua de la calidad del código con Amazon CodeGuru)

Ejemplos relacionados:

- Security for Developers workshop (Taller de seguridad para desarrolladores)

SEC11-BP05 Centralizar los servicios para paquetes y dependencias

Proporcione servicios centralizados para que los equipos de creadores obtengan paquetes de software y otras dependencias. De este modo, se podrán validar los paquetes antes de incluirlos en el software que escriba y se dispondrá de un origen de datos para el análisis del software que se utiliza en su organización.

Resultado deseado: El software se compone de un conjunto de otros paquetes de software además del código que se escribe. Esto facilita el consumo de implementaciones de funcionalidades que se utilizan repetidamente, como un analizador JSON o una biblioteca de cifrado. La centralización lógica de los orígenes de estos paquetes y dependencias proporciona un mecanismo para que los equipos de seguridad validen las propiedades de los paquetes antes de utilizarlos. Este planteamiento también reduce el riesgo de que se produzca un problema inesperado debido a un cambio en un paquete existente o a la inclusión por equipos de creadores de paquetes arbitrarios directamente desde Internet. Utilice este planteamiento junto con los flujos de pruebas manuales y automatizadas para aumentar la confianza en la calidad del software que desarrolla.

Patrones comunes de uso no recomendados:

- Obtener paquetes de repositorios arbitrarios de Internet.
- No probar nuevos paquetes antes de ponerlos a disposición de los desarrolladores.

Beneficios de establecer esta práctica recomendada:

- Mejor comprensión de los paquetes que se utilizan en el software que se crea.
- Poder notificar a los equipos de carga de trabajo cuándo es necesario actualizar un paquete basándose en la comprensión de quién utiliza qué.
- Reducción del riesgo de que se incluya en el software un paquete con problemas.

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

Guía para la implementación

Proporcione servicios centralizados para paquetes y dependencias de una manera que resulte sencilla de consumir a los creadores. Los servicios centralizados pueden ser lógicamente centrales en lugar de implementarse como un sistema monolítico. Este método le permite proporcionar servicios de una manera que satisfaga las necesidades de los creadores. Debe implementar una forma eficaz de añadir paquetes al repositorio cuando se produzcan actualizaciones o surjan nuevos requisitos. Los servicios de AWS como AWS CodeArtifact o soluciones similares de socios de AWS son una forma de ofrecer esta capacidad.

Pasos para la aplicación:

- Implemente un servicio de repositorio lógicamente centralizado que esté disponible en todos los entornos en los que se desarrolla software.
- Incluya el acceso al repositorio como parte del proceso de aprovisionamiento de cuentas de Cuenta de AWS.
- Consolide la automatización para probar paquetes antes de que se publiquen en un repositorio.
- Mantenga métricas de los paquetes, lenguajes y equipos más utilizados y con mayor cantidad de cambios.
- Proporcione un mecanismo automatizado para que los equipos de creación soliciten nuevos paquetes y proporcionen comentarios.
- Analice periódicamente los paquetes del repositorio para identificar la posible repercusión de los problemas que se acaban de detectar.

Recursos

Prácticas recomendadas relacionadas:

- SEC11-BP02 Automatizar las pruebas a lo largo del ciclo de vida de desarrollo y lanzamiento

Documentos relacionados:

- Accelerate deployments on AWS with effective governance (Acelerar los despliegues en AWS con una gobernanza eficaz)
- Tighten your package security with CodeArtifact Package Origin Control toolkit (Refuerce la seguridad de sus paquetes con el kit de herramientas de control de origen de paquetes de CodeArtifact)
- Detecting security issues in logging with Amazon CodeGuru Reviewer (Detección de problemas de seguridad en el registro con Amazon CodeGuru Reviewer)
- Supply chain Levels for Software Artifacts (SLSA) (Niveles de la cadena de suministro de artefactos de software [SLSA])

Vídeos relacionados:

- Proactive security: Considerations and approaches (Seguridad proactiva: consideraciones y estrategias)
- The AWS Philosophy of Security (re:Invent 2017) (La filosofía de seguridad de AWS [re:Invent 2017])
- When security, safety, and urgency all matter: Handling Log4Shell (Cuando la seguridad, la protección y la urgencia son importantes: gestión de Log4Shell)

Ejemplos relacionados:

- Multi Region Package Publishing Pipeline (Canalización de publicación de paquetes multirregión) [GitHub])
- Publishing Node.js Modules on AWS CodeArtifact using AWS CodePipeline (Publicación de módulos Node.js en AWS CodeArtifact con AWS CodePipeline) (GitHub)
- AWS CDK Java CodeArtifact Pipeline Sample (Ejemplo de canalización de CodeArtifact en Java en AWS CDK) (GitHub)

- Distribute private .NET NuGet packages with AWS CodeArtifact (Distribuir paquetes NuGet .NET privados con AWS CodeArtifact) (GitHub)

SEC11-BP06 Desplegar software mediante programación

Siempre que sea posible, realice los despliegues de software mediante programación. Con este enfoque se reduce la probabilidad de que se produzca un error en el despliegue o de que surja un problema inesperado debido a un error humano.

Resultado deseado: Mantener a las personas alejadas de los datos es un principio clave para crear de forma segura en la Nube de AWS. Este principio incluye la forma de desplegar el software.

La ventaja de no depender de personas para desplegar el software es que tendrá mayor confianza en que se ha probado lo que se despliega, y que el despliegue se realice siempre de forma coherente. No tendrá que modificar el software para que funcione en distintos entornos. El uso de los principios del desarrollo de aplicaciones de doce factores, en concreto la externalización de la configuración, le permite desplegar el mismo código en varios entornos sin necesidad de realizar cambios. La firma criptográfica de los paquetes de software es una buena forma de verificar que no ha cambiado nada entre entornos. El resultado general de este método es que se reduce el riesgo en el proceso de cambio y mejorar la coherencia de las versiones de software.

Patrones comunes de uso no recomendados:

- Despliegue manual del software en producción.
- Realización manual de cambios en el software para adaptarlo a distintos entornos.

Beneficios de establecer esta práctica recomendada:

- Mayor confianza en el proceso de lanzamiento de software.
- Reducción del riesgo de que un cambio erróneo afecte a las funciones de la empresa.
- Aumento de la cadencia de lanzamiento debido al menor riesgo del cambio.
- Capacidad de reversión automática en caso de imprevistos durante el despliegue.
- Capacidad para demostrar criptográficamente que el software probado es el software desplegado.

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

Guía para la implementación

Cree su estructura de cuenta de Cuenta de AWS de forma que elimine el acceso humano persistente desde los entornos y utilice herramientas de CI/CD para realizar los despliegues. Diseñe las aplicaciones de manera que los datos de configuración específicos del entorno se obtengan de un origen externo, como el Parameter Store de AWS Systems Manager. Firme los paquetes después de probarlos y valide estas firmas durante el despliegue. Configure las canalizaciones de CI/CD para que envíen el código de la aplicación y utilice valores controlados para confirmar que el despliegue ha tenido lugar como corresponde. Utilice herramientas como AWS CloudFormation o AWS CDK para definir su infraestructura y, a continuación, use AWS CodeBuild y AWS CodePipeline para realizar las operaciones de CI/CD.

Pasos para la aplicación

- Cree canalizaciones de CI/CD bien definidas para agilizar el proceso de despliegue.
- Proporcione capacidad de CI/CD para simplificar la integración de las pruebas de seguridad en las canalizaciones con AWS CodeBuild y AWS Code Pipeline.
- Siga las directrices sobre separación de entornos del documento técnico Organizing Your AWS Environment Using Multiple Accounts (Organización del entorno de AWS con varias cuentas).
- Verifique que no haya acceso humano persistente a los entornos donde se ejecutan las cargas de trabajo de producción.
- Diseñe las aplicaciones de modo que admitan la externalización de datos de configuración.
- Piense en la posibilidad de llevar a cabo el despliegue mediante un modelo de despliegue azul-verde.
- Implemente valores controlados para validar el despliegue correcto del software.
- Utilice herramientas criptográficas como AWS Signer o AWS Key Management Service (AWS KMS) para firmar y verificar los paquetes de software que está desplegando.

Recursos

Prácticas recomendadas relacionadas:

- SEC11-BP02 Automatizar las pruebas a lo largo del ciclo de vida de desarrollo y lanzamiento

Documentos relacionados:

- AWS CI/CD Workshop (Taller de CI/CD de AWS)

- Accelerate deployments on AWS with effective governance (Acelerar los despliegues en AWS con una gobernanza eficaz)
- Automatización de implementaciones seguras y sin intervención
- Code signing using AWS Certificate Manager Private CA and AWS Key Management Service asymmetric keys (Firma de código mediante CA privada de AWS Certificate Manager y claves asimétricas deAWS Key Management Service)
- Code Signing, a Trust and Integrity Control for AWS Lambda (Firma de código, un control de confianza e integridad para AWS Lambda)

Vídeos relacionados:

- Hands-off: Automating continuous delivery pipelines at Amazon (Sin intervención: automatización de canalizaciones de entrega continua en Amazon)

Ejemplos relacionados:

- Blue/Green deployments with AWS Fargate Despliegues azul-verde AWS Fargate)

SEC11-BP07 Evaluar periódicamente las propiedades de seguridad de las canalizaciones

Aplique los principios del pilar de seguridad de Well-Architected a sus canalizaciones, prestando especial atención a la separación de permisos. Evalúe periódicamente las propiedades de seguridad de su infraestructura de canalización. La administración eficaz de la seguridad de las canalizaciones le permite garantizar la seguridad del software que pasa por ellas.

Resultado deseado: Las canalizaciones utilizadas para crear y desplegar el software deben seguir las mismas prácticas recomendadas que cualquier otra carga de trabajo en su entorno. Los desarrolladores no deben poder editar las pruebas que se implementan en las canalizaciones que utilizan. Las canalizaciones solo deben tener los permisos necesarios para los despliegues que están realizando y debe implementar salvaguardas para evitar que se desplieguen en los entornos equivocados. Las canalizaciones no deben depender de credenciales a largo plazo; además, deben estar configuradas para emitir estado de forma que se pueda validar la integridad de los entornos de compilación.

Patrones comunes de uso no recomendados:

- Pruebas de seguridad que los creadores pueden omitir.

- Permisos demasiado amplios para las canalizaciones de despliegue.
- Canalizaciones no configuradas para validar entradas.
- No revisar periódicamente los permisos asociados a la infraestructura de CI/CD.
- Uso de credenciales a largo plazo o codificadas.

Beneficios de establecer esta práctica recomendada:

- Mayor confianza en la integridad del software que se construye y despliega a través de las canalizaciones.
- Capacidad de detener un despliegue cuando hay actividades sospechosas.

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

Guía para la implementación

Comience con servicios de CI/CD administrados que admiten roles de IAM para reducir el riesgo de fuga de credenciales. La aplicación de los principios del pilar de seguridad a la infraestructura de canalización de CI/CD puede ayudarle a determinar dónde es posible realizar mejoras de seguridad. Siga la arquitectura de referencia de canalizaciones de despliegue de AWS. Es un buen punto de partida para crear entornos de CI/CD. Revise a intervalos regulares la implementación de la canalización y analice los registros para detectar comportamientos inesperados; esto puede ayudarle a comprender los patrones de uso de las canalizaciones que se utilizan para desplegar software.

Pasos para la aplicación

- Empiece con la arquitectura de referencia de canalizaciones de despliegue de AWS.
- Plantéese la posibilidad de utilizar AWS IAM Access Analyzer para generar mediante programación políticas de IAM de privilegios mínimos para las canalizaciones.
- Integre las canalizaciones con monitorización y alertas para que recibir notificaciones de actividad inesperada o anómala. Para los servicios administrados de AWS, Amazon EventBridge le permite enrutar datos a destinos como AWS Lambda o Amazon Simple Notification Service (Amazon SNS).

Recursos

Documentos relacionados:

- AWS Deployment Pipelines Reference Architecture (Arquitectura de referencia de canalizaciones de despliegue de AWS)
- Monitoring AWS CodePipeline (Monitorización de AWS CodePipeline)
- Security best practices for AWS CodePipeline (Prácticas recomendadas de seguridad de AWS CodePipeline)

Ejemplos relacionados:

- DevOps monitoring dashboard (Panel de monitorización de DevOps) (GitHub)

SEC11-BP08 Crear un programa que integre la propiedad de la seguridad en los equipos de la carga de trabajo

Elabore un programa o un mecanismo que permita a los equipos de creadores tomar decisiones de seguridad sobre el software que crean. Aún así, su equipo de seguridad debe validar estas decisiones durante una revisión, pero integrar la propiedad de la seguridad en los equipos de creadores permite crear cargas de trabajo más rápidas y seguras. Este mecanismo también fomenta una cultura de propiedad que repercute positivamente en el funcionamiento de los sistemas que se crean.

Resultado deseado: Para integrar la propiedad de la seguridad y la toma de decisiones en los equipos de creación, puede formar a los creadores sobre cómo pensar en la seguridad o puede mejorar su formación con personal de seguridad integrado o asociado a los equipos de creación. Cualquiera de las dos estrategias es válida y permite al equipo tomar decisiones de seguridad de mayor calidad en una fase más temprana del ciclo de desarrollo. Este modelo de propiedad se basa en la formación para lograr la seguridad de las aplicaciones. Empiece con el modelo de amenazas para la carga de trabajo concreta, lo que ayudará a dirigir el enfoque de diseño al contexto apropiado. Otra ventaja de contar con una comunidad de desarrolladores centrados en la seguridad, o con un grupo de ingenieros de seguridad que trabajen con equipos de creadores, es que es posible comprender más a fondo cómo se escribe el software. Esta comprensión le ayuda a determinar las próximas áreas de mejora en su capacidad de automatización.

Patrones comunes de uso no recomendados:

- Dejar todas las decisiones del diseño de la seguridad en manos del equipo de seguridad.
- No hacer frente a los requisitos de seguridad con suficiente antelación en el proceso de desarrollo.

- No obtener comentarios de los creadores y del personal de seguridad sobre el funcionamiento del programa.

Beneficios de establecer esta práctica recomendada:

- Reducción del tiempo necesario para completar las revisiones de seguridad.
- Reducción de los problemas de seguridad que solo se detectan en la fase de revisión de la seguridad.
- Mejora de la calidad general del software que se escribe.
- Oportunidad de identificar y comprender problemas sistémicos o áreas de mejora de alto valor.
- Reducción de la cantidad de tareas que es necesario repetir debido a los hallazgos de la revisión de seguridad.
- Mejora de la percepción de la función de seguridad.

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

Guía para la implementación

Empiece con la orientación de SEC11-BP01 Formar en seguridad de las aplicaciones. A continuación, identifique el modelo operativo para el programa que crea que puede funcionar mejor para su organización. Los dos modelos principales son formar a los desarrolladores o integrar al personal de seguridad en los equipos de creadores. Una vez que haya decidido el abordaje inicial, deberá realizar una prueba piloto con uno o un pequeño grupo de equipos de carga de trabajo para comprobar que el modelo funciona en su organización. El apoyo de los líderes de los departamentos de creación y seguridad de la organización contribuye a la implantación y al éxito del programa. A medida que cree este programa, es importante elegir métricas que sirvan para mostrar el valor del programa. Aprender de cómo AWS ha tratado este problema es una buena experiencia de aprendizaje. Esta práctica recomendada se centra en gran medida en la cultura y el cambio organizativo. Las herramientas que emplee deben apoyar la colaboración entre las comunidades de creadores y de seguridad.

Pasos para la aplicación

- Empiece por formar a los desarrolladores en la seguridad para las aplicaciones.
- Cree una comunidad y un programa de incorporación para educar a los creadores.

- Elija un nombre para el programa. Los más utilizados son «Guardians», «Champions» o «Advocates».
- Identifique el modelo a utilizar: formar a los desarrolladores, incorporar ingenieros de seguridad o tener roles de seguridad afines.
- Identifique a los patrocinadores del proyecto entre los encargados de la seguridad, los creadores y, quizá, otros grupos pertinentes.
- Haga un seguimiento del número de personas que participan en el programa, el tiempo necesario para las revisiones y los comentarios de los creadores y el personal de seguridad. Utilice estas métricas para acometer mejoras.

Recursos

Prácticas recomendadas relacionadas:

- SEC11-BP01 Formar en seguridad de las aplicaciones
- SEC11-BP02 Automatizar las pruebas a lo largo del ciclo de vida de desarrollo y lanzamiento

Documentos relacionados:

- How to approach threat modeling (Cómo abordar el modelado de amenazas)
- How to think about cloud security governance (Cómo concebir la gobernanza de la seguridad en la nube)

Vídeos relacionados:

- Proactive security: Considerations and approaches (Seguridad proactiva: consideraciones y estrategias)

Fiabilidad

El pilar de fiabilidad abarca la capacidad de una carga de trabajo para realizar su función prevista de forma correcta y coherente cuando se espera que lo haga. Encontrará recomendaciones de implementación en el documento técnico Pilar de fiabilidad.

Áreas de prácticas recomendadas

- Fundamentos

- Arquitectura de la carga de trabajo
- Administración de cambios
- Administración de errores

Fundamentos

Preguntas

- FIABILIDAD 1. ¿Cómo administra las Service Quotas y las restricciones?
- FIABILIDAD 2. ¿Cómo planifica la topología de la red?

FIABILIDAD 1. ¿Cómo administra las Service Quotas y las restricciones?

Para las arquitecturas de carga de trabajo basadas en la nube, existen Service Quotas, también denominadas límites de servicio. Estas cuotas existen para evitar aprovisionar por accidente más recursos de los necesarios y para limitar las tasas de solicitud en las operaciones de la API, de modo que los servicios queden protegidos ante posibles abusos. También existen restricciones de recursos, por ejemplo, la velocidad a la que se pueden introducir bits en un cable de fibra óptica o la cantidad de almacenamiento de un disco físico.

Prácticas recomendadas

- REL01-BP01 Conocimiento de las cuotas y restricciones del servicio
- REL01-BP02 Administrar cuotas de servicio en cuentas y regiones
- REL01-BP03 Adaptar las cuotas de servicio fijas y las restricciones a través de la arquitectura
- REL01-BP04 Supervisar y administrar cuotas
- REL01-BP05 Automatizar la administración de cuotas
- REL01-BP06 Garantizar que exista una diferencia suficiente entre las cuotas actuales y el uso máximo para permitir la conmutación por error

REL01-BP01 Conocimiento de las cuotas y restricciones del servicio

Conozca las cuotas predeterminadas y administre las solicitudes de aumento de cuota para su arquitectura de carga de trabajo. Sepa qué restricciones de recursos en la nube, como el disco o la red, pueden causar impacto.

Resultado deseado: los clientes pueden evitar el deterioro o la interrupción del servicio en sus Cuentas de AWS mediante la implementación de directrices adecuadas para la supervisión de las métricas clave, las revisiones de la infraestructura y la automatización de los pasos de corrección, a fin de verificar que no se alcanzan las cuotas y restricciones de los servicios que podrían provocar el deterioro o la interrupción del servicio.

Antipatrones usuales:

- Desplegar una carga de trabajo sin conocer las cuotas estrictas o flexibles y sus límites para los servicios utilizados.
- Desplegar una carga de trabajo de reemplazo sin analizar ni volver a configurar las cuotas necesarias o sin contactar previamente con el servicio de asistencia.
- Suponer que los servicios en la nube no tienen límites y que los servicios pueden utilizarse sin tener en cuenta tarifas, límites, recuentos o cantidades.
- Suponer que las cuotas se incrementarán automáticamente.
- Desconocer el proceso y la cronología de las solicitudes de cuotas.
- Suponer que la cuota de servicio predeterminada en la nube es la misma para todos los servicios en diferentes regiones.
- Suponer que se pueden incumplir las restricciones del servicio y que los sistemas se escalarán automáticamente o añadirán un aumento del límite más allá de las restricciones del recurso.
- No probar la aplicación en picos de tráfico para estresar el uso de sus recursos.
- Aprovisionar el recurso sin analizar el tamaño de recurso requerido.
- Sobreaprovisionar la capacidad mediante la elección de tipos de recursos que van mucho más allá de las necesidades reales o de los picos previstos.
- No evaluar las necesidades de capacidad para nuevos niveles de tráfico antes de que se produzca un nuevo evento con un cliente o de desplegar una nueva tecnología.

Beneficios de establecer esta práctica recomendada: la supervisión y la administración automatizada de las cuotas de servicio y las restricciones de recursos pueden reducir los errores de forma proactiva. Los cambios en los patrones de tráfico para el servicio de un cliente pueden provocar una interrupción o un deterioro si no se siguen las prácticas recomendadas. Con la supervisión y la administración de estos valores en todas las regiones y en todas las cuentas, las aplicaciones pueden tener una mayor resiliencia ante acontecimientos adversos o imprevistos.

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

Guía para la implementación

Service Quotas es un servicio de AWS que le ayuda a administrar sus cuotas para más de 250 servicios de AWS desde una sola ubicación. Además de consultar los valores de las cuotas, también puede solicitar y realizar un seguimiento de los aumentos de las cuotas desde la consola de Service Quotas o mediante el SDK de AWS. AWS Trusted Advisor ofrece una comprobación de las cuotas de servicio que muestra su uso y las cuotas para ciertos aspectos de algunos servicios. Las cuotas de servicio predeterminadas por servicio también están en la documentación de AWS del servicio respectivo (por ejemplo, consulte Cuotas de Amazon VPC).

Algunos límites de servicio, como los límites de velocidad en las API limitadas se establecen en Amazon API Gateway mediante la configuración de un plan de uso. Algunos límites que se establecen como configuración en sus respectivos servicios son las IOPS aprovisionadas, el almacenamiento de Amazon RDS asignado y las asignaciones de volumen Amazon EBS. Amazon Elastic Compute Cloud tiene su propio panel de límites de servicio que puede ayudarle a administrar sus límites de instancia, Amazon Elastic Block Store y de dirección IP elástica. Si tiene un caso de uso en el que las cuotas de servicio repercuten en el rendimiento de su aplicación y no se ajustan a sus necesidades, contacte con Soporte para ver si existen mitigaciones.

Las cuotas de servicio pueden ser específicas de una región o también pueden tener carácter global. El uso de un servicio de AWS que alcance su cuota no actuará del modo previsto en un uso normal y puede provocar la interrupción o el deterioro del servicio. Por ejemplo, una cuota de servicio limita el número de DL Amazon EC2 que pueden utilizarse en una región y ese límite puede alcanzarse durante un evento de escalamiento de tráfico mediante grupos de Auto Scaling (ASG).

Las cuotas de servicio de cada cuenta deben evaluarse periódicamente para determinar cuáles podrían ser los límites de servicio adecuados para esa cuenta. Estas cuotas de servicio existen como barreras de protección operativas para evitar aprovisionar por accidente más recursos de los necesarios. También sirven para limitar las tasas de solicitudes en las operaciones de API para proteger los servicios del abuso.

Las restricciones de servicio son diferentes de las cuotas de servicio. Las restricciones de servicio representan los límites de un recurso concreto, tal y como los define ese tipo de recurso. Pueden ser la capacidad de almacenamiento (por ejemplo, gp2 tiene un límite de tamaño de 1 GB - 16 TB) o el rendimiento del disco (10 0000 iops). Es esencial que las restricciones de un tipo de recurso se diseñen y evalúen constantemente para detectar un uso que pueda alcanzar su límite. Si se alcanza una restricción de forma inesperada, las aplicaciones o servicios de la cuenta pueden deteriorarse o interrumpirse.

Si existe un caso de uso en el que las cuotas de servicio repercuten en el rendimiento de una aplicación y no pueden ajustarse a las necesidades requeridas, contacte con Soporte para ver si existen mitigaciones. Para obtener más detalles sobre el ajuste de cuotas fijas, consulte REL01-BP03 Adaptar las cuotas de servicio fijas y las restricciones a través de la arquitectura.

Hay una serie de servicios y herramientas de AWS con las que se puede supervisar y administrar Service Quotas. El servicio y las herramientas se deben utilizar para proporcionar comprobaciones automatizadas o manuales de los niveles de cuota.

- AWS Trusted Advisor ofrece una comprobación de cuotas de servicio que muestra su uso y las cuotas para ciertos aspectos de algunos servicios. Puede ayudar a identificar los servicios que están cerca de la cuota.
- Consola de administración de AWS proporciona métodos para mostrar los valores de las cuotas de los servicios, administrar, solicitar nuevas cuotas, supervisar el estado de las solicitudes de cuotas y mostrar el historial de cuotas.
- AWS CLI y los CDK ofrecen métodos programáticos para administrar y supervisar automáticamente los niveles de cuota de servicio y el uso.

Pasos para la implementación

Para Service Quotas:

- Revise AWS Service Quotas.
- Para conocer sus cuotas de servicio existentes, determine los servicios (como IAM Access Analyzer) que se utilizan. Hay aproximadamente 250 servicios de AWS controlados por cuotas de servicio. A continuación, determine el nombre específico de la cuota de servicio que podría utilizarse en cada cuenta y región. Hay aproximadamente 3000 nombres de cuotas de servicio por región.
- Aumente este análisis de cuotas con AWS Config para encontrar todos los recursos de AWS utilizados en sus Cuentas de AWS.
- Use los datos de AWS CloudFormation para determinar los recursos de AWS utilizados. Examine los recursos que se han creado en la Consola de administración de AWS o con el comando `list-stack-resources` de la AWS CLI. También puede ver los recursos configurados para desplegarse en la propia plantilla.
- Consulte el código de despliegue para determinar todos los servicios que necesita su carga de trabajo.

- Determine las cuotas de servicio que se aplican. Use la información accesible mediante programación de Trusted Advisor y Service Quotas.
- Establezca un método de supervisión automatizado (consulte REL01-BP02 Administrar cuotas de servicio en cuentas y regiones y REL01-BP04 Supervisar y administrar cuotas) para alertar e informar si las cuotas de servicio están cerca de su límite o lo han alcanzado.
- Establezca un método automatizado y programático para comprobar si se ha modificado una cuota de servicio en una región pero no en otras regiones de la misma cuenta (consulte REL01-BP02 Administrar cuotas de servicio en cuentas y regiones y REL01-BP04 Supervisar y administrar cuotas).
- Automatice el análisis de los registros y las métricas de las aplicaciones para determinar si existen errores de cuotas o restricciones de servicio. Si se producen estos errores, envíe alertas al sistema de supervisión.
- Establezca procedimientos de ingeniería para calcular el cambio necesario en la cuota (consulte REL01-BP05 Automatizar la administración de cuotas) una vez que se haya identificado que se necesitan cuotas mayores para servicios específicos.
- Cree un flujo de trabajo de aprovisionamiento y aprobación para solicitar cambios en la cuota de servicio. Debería incluir un flujo de trabajo de excepciones en caso de denegación de la solicitud o de aprobación parcial.
- Cree un método de ingeniería para efectuar una revisión de las cuotas de servicio previa al aprovisionamiento y el uso de nuevos servicios de AWS antes de desplegarlos en entornos de producción o de carga (por ejemplo, una cuenta de pruebas de carga).

Para las restricciones de servicio:

- Establezca métodos de supervisión y medición para alertar de la lectura de los recursos próximos a sus restricciones. Use CloudWatch según proceda para las métricas o la supervisión de registros.
- Establezca umbrales de alerta para cada recurso con una restricción significativa para la aplicación o el sistema.
- Cree procedimientos de administración de flujos de trabajo e infraestructuras para cambiar el tipo de recurso si la restricción está próxima a su uso. Como práctica recomendada, este flujo de trabajo debe incluir pruebas de carga para verificar que el nuevo tipo sea el tipo de recurso correcto con las nuevas restricciones.
- Migre el recurso identificado al nuevo tipo de recurso recomendado, mediante los procedimientos y los procesos existentes.

Recursos

Prácticas recomendadas relacionadas:

- REL01-BP02 Administrar cuotas de servicio en cuentas y regiones
- REL01-BP03 Adaptar las cuotas de servicio fijas y las restricciones a través de la arquitectura
- REL01-BP04 Supervisar y administrar cuotas
- REL01-BP05 Automatizar la administración de cuotas
- REL01-BP06 Garantizar que exista una diferencia suficiente entre las cuotas actuales y el uso máximo para permitir la conmutación por error
- REL03-BP01 Elegir cómo segmentar su carga de trabajo
- REL10-BP01 Implementar la carga de trabajo en varias ubicaciones
- REL11-BP01 Supervisar todos los componentes de la carga de trabajo para detectar errores
- REL11-BP03 Automatizar la reparación en todas las capas
- REL12-BP05 Probar la resiliencia mediante la ingeniería del caos

Documentos relacionados:

- Pilar de fiabilidad de AWS Well-Architected Framework: disponibilidad
- AWS Service Quotas (denominados anteriormente límites de servicio)
- Comprobaciones de prácticas recomendadas de AWS Trusted Advisor (consulte la sección Límites de servicio)
- AWS Limit Monitor en AWS Answers
- Límites de servicio de Amazon EC2
- ¿Qué es Service Quotas?
- Cómo solicitar un aumento de cuota
- Puntos de conexión y cuotas de servicio
- Guía del usuario de Service Quotas
- Supervisor de cuotas para AWS
- Límites de aislamiento de errores de AWS
- Disponibilidad con redundancia
- AWS para datos

- ¿Qué es la integración continua?
- ¿Qué es la entrega continua?
- Socio de APN: socios que pueden ayudar con la administración de la configuración
- Managing the account lifecycle in account-per-tenant SaaS environments on AWS(Administración del ciclo de vida de las cuentas en entornos SaaS de cuenta por inquilino en AWS)
- Managing and monitoring API throttling in your workloads (Administrar y supervisar la limitación de las API en sus cargas de trabajo)
- View AWS Trusted Advisor recommendations at scale with AWS Organizations(Ver recomendaciones de AWS Trusted Advisor a escala con AWS Organizations)
- Automating Service Limit Increases and Enterprise Support with AWS Control Tower(Automatización de los aumentos del límite de servicio y asistencia a empresas con AWS Control Tower)

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas
- View and Manage Quotas for AWS Services Using Service Quotas (Ver y administrar cuotas para AWS Services con Service Quotas)
- AWS IAM Quotas Demo (Demostración de las cuotas de AWS IAM)

Herramientas relacionadas:

- Amazon CodeGuru Reviewer
- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager

- AWS Marketplace

REL01-BP02 Administrar cuotas de servicio en cuentas y regiones

Si utiliza múltiples cuentas o regiones, solicite las cuotas pertinentes en todos los entornos en los que se ejecutan sus cargas de trabajo de producción.

Resultado deseado: los servicios y las aplicaciones no deberían verse afectados si se agota la cuota de servicio en las configuraciones que abarcan cuentas o regiones, o que tienen diseños de resiliencia mediante la conmutación por error de zonas, regiones o cuentas.

Antipatrones usuales:

- Permitir que aumente el uso de recursos en una región aislada sin ningún mecanismo para mantener la capacidad en las demás.
- Configurar manualmente todas las cuotas de forma independiente en regiones aisladas.
- No considerar el efecto de las arquitecturas de resiliencia (activa o pasiva) en las futuras necesidades de cuota durante un deterioro de la región no principal.
- No evaluar las cuotas periódicamente ni realizar los cambios necesarios en cada región y cuenta donde se ejecuta la carga de trabajo.
- No utilizar las plantillas de solicitud de cuota para solicitar incrementos en varias regiones y cuentas.
- No actualizar las cuotas de servicio por pensar erróneamente que el aumento de las cuotas tiene implicaciones de coste, como las solicitudes de reserva de computación.

Beneficios de establecer esta práctica recomendada: verificar que puede gestionar su carga actual en regiones o cuentas secundarias si los servicios regionales dejan de estar disponibles. Esto puede reducir el número de errores o los niveles de deterioro que se producen durante la pérdida de una región.

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

Guía para la implementación

El seguimiento de las cuotas de servicio se realiza por cuenta. A no ser que se especifique lo contrario, cada cuota es específica de una Región de AWS. Además de los entornos de producción, administre también las cuotas en todos los entornos que no sean de producción aplicables, de modo que las pruebas y el desarrollo no se vean limitados. El mantenimiento de un elevado nivel

de resiliencia requiere que las cuotas de servicio se evalúen continuamente (ya sea de forma automatizada o manual).

Al haber más cargas de trabajo que abarcan regiones debido a la implementación de diseños que utilizan los enfoques Activo/Activo, Activo/Pasivo - En caliente, Activo/Pasivo - En frío y Activo/Pasivo - Luz piloto, es esencial comprender todos los niveles de cuotas de regiones y cuentas. Los patrones de tráfico anteriores no siempre son un buen indicador de si la cuota de servicio está configurada correctamente.

Igualmente importante es el hecho de que el límite de nombres de cuota de servicio no es siempre el mismo para todas las regiones. En una región, el valor podría ser cinco, y en otra, diez. La administración de estas cuotas debe abarcar los mismos servicios, cuentas y regiones para proporcionar una resiliencia coherente bajo carga.

Concilie todas las diferencias de cuota de servicio entre las distintas regiones (región activa o región pasiva) y cree procesos para conciliar continuamente estas diferencias. Los planes de prueba de las conmutaciones por error pasivas de las regiones en muy pocas ocasiones se escalan a la capacidad activa máxima, lo que significa que los ejercicios del día de juego o de mesa pueden no encontrar diferencias en las cuotas de servicio entre las regiones y tampoco mantener los límites correctos.

Es muy importante controlar y evaluar la desviación de cuota de servicio, la situación en la que los límites de la cuota de servicio para una determinada cuota con nombre se modifican en una región y no en todas las regiones. Debe considerarse la posibilidad de cambiar la cuota en las regiones con tráfico o con posibilidad de tener tráfico.

- Seleccione las cuentas y regiones que correspondan según sus requisitos de servicio, latencia, normativos y de recuperación de desastres (DR).
- Identifique las cuotas de servicio en todas las cuentas, regiones y zonas de disponibilidad pertinentes. Los límites se determinan por cuenta y región. Estos valores deben compararse para detectar diferencias.

Pasos para la implementación

- Revise los valores de Service Quotas que podrían haber superado el nivel de riesgo de uso. AWS Trusted Advisor proporciona alertas si superan los umbrales del 80 % y el 90 %.
- Revise los valores de las cuotas de servicio en cualquier región pasiva (en un diseño activo/pasivo). Verifique que la carga se ejecutará correctamente en las regiones secundarias si se produce un error en la región principal.

- Automatice la evaluación de si se ha producido alguna desviación de la cuota de servicio entre regiones de la misma cuenta y actúe en consecuencia para modificar los límites.
- Si las unidades organizativas (UO) del cliente están estructuradas de la forma admitida, las plantillas de cuotas de servicio deberán actualizarse para reflejar los cambios en las cuotas que deban aplicarse a varias regiones y cuentas.
  - Cree una plantilla y asocie regiones al cambio de cuota.
  - Revise todas las plantillas de cuota de servicio existentes por si fuera necesario realizar algún cambio (región, límites y cuentas).

Recursos

Prácticas recomendadas relacionadas:

- REL01-BP01 Conocimiento de las cuotas y restricciones del servicio
- REL01-BP03 Adaptar las cuotas de servicio fijas y las restricciones a través de la arquitectura
- REL01-BP04 Supervisar y administrar cuotas
- REL01-BP05 Automatizar la administración de cuotas
- REL01-BP06 Garantizar que exista una diferencia suficiente entre las cuotas actuales y el uso máximo para permitir la conmutación por error
- REL03-BP01 Elegir cómo segmentar su carga de trabajo
- REL10-BP01 Implementar la carga de trabajo en varias ubicaciones
- REL11-BP01 Supervisar todos los componentes de la carga de trabajo para detectar errores
- REL11-BP03 Automatizar la reparación en todas las capas
- REL12-BP05 Probar la resiliencia mediante la ingeniería del caos

Documentos relacionados:

- Pilar de fiabilidad de AWS Well-Architected Framework: disponibilidad
- AWS Service Quotas (denominados anteriormente límites de servicio)
- Comprobaciones de prácticas recomendadas de AWS Trusted Advisor (consulte la sección Límites de servicio)
- AWS Limit Monitor en AWS Answers
- Límites de servicio de Amazon EC2

- ¿Qué es Service Quotas?
- Cómo solicitar un aumento de cuota
- Puntos de conexión y cuotas de servicio
- Guía del usuario de Service Quotas
- Supervisor de cuotas para AWS
- Límites de aislamiento de errores de AWS
- Disponibilidad con redundancia
- AWS para datos
- ¿Qué es la integración continua?
- ¿Qué es la entrega continua?
- Socio de APN: socios que pueden ayudar con la administración de la configuración
- Managing the account lifecycle in account-per-tenant SaaS environments on AWS(Administración del ciclo de vida de las cuentas en entornos SaaS de cuenta por inquilino en AWS)
- Managing and monitoring API throttling in your workloads (Administrar y supervisar la limitación de las API en sus cargas de trabajo)
- View AWS Trusted Advisor recommendations at scale with AWS Organizations(Ver recomendaciones de AWS Trusted Advisor a escala con AWS Organizations)
- Automating Service Limit Increases and Enterprise Support with AWS Control Tower(Automatización de los aumentos del límite de servicio y asistencia a empresas con AWS Control Tower)

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas
- View and Manage Quotas for AWS Services Using Service Quotas (Ver y administrar cuotas para AWS Services con Service Quotas)
- AWS IAM Quotas Demo (Demostración de las cuotas de AWS IAM)

Servicios relacionados:

- Amazon CodeGuru Reviewer
- AWS CodeDeploy

- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP03 Adaptar las cuotas de servicio fijas y las restricciones a través de la arquitectura

Sea consciente de las cuotas de servicio inalterables, las restricciones de servicio y los límites de recursos físicos. Diseñe arquitecturas para aplicaciones y servicios que eviten que estos límites afecten a la fiabilidad.

Algunos ejemplos son el ancho de banda de la red, el tamaño de la carga útil de la invocación de funciones sin servidor, la tasa de ráfagas de aceleración para una puerta de enlace de API y las conexiones simultáneas de usuarios a una base de datos.

Resultado deseado: la aplicación o servicio funciona como se espera en condiciones normales y de alto tráfico. Se han diseñado para funcionar dentro de las limitaciones fijadas para ese recurso o cuotas de servicio.

Antipatrones usuales:

- Elegir un diseño que utilice un recurso de un servicio, sin saber que existen restricciones de diseño que provocarán que este diseño producirá un error en el escalado.
- Realizar una evaluación comparativa que no es realista y que alcanzará las cuotas fijas del servicio durante la evaluación. Por ejemplo, ejecutar pruebas con un límite de ráfagas, pero durante un período prolongado.
- Elegir un diseño que no pueda escalarse ni modificarse si se van a superar las cuotas de servicio fijas. Por ejemplo, un tamaño de carga útil SQS de 256 KB.
- No diseña ni implementa la observabilidad para supervisar y avisar sobre umbrales de cuotas de servicio que podrían estar en riesgo durante eventos de alto tráfico.

Beneficios de establecer esta práctica recomendada: verificación de que la aplicación funcionará bajo todos los niveles de carga de servicios previstos sin interrupciones ni degradaciones.

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

Guía para la implementación

A diferencia de las cuotas de servicio blandas o los recursos que pueden sustituirse por unidades de mayor capacidad, las cuotas fijas de los servicios de AWS no pueden modificarse. Esto significa que todos estos tipos de servicios de AWS deben evaluarse para detectar posibles límites estrictos de capacidad cuando se utilizan en el diseño de una aplicación.

Los límites estrictos se muestran en la consola de Service Quotas. Si las columnas muestran `AJUSTABLE = No`, el servicio tiene un límite estricto. Los límites estrictos también se muestran en algunas páginas de configuración de recursos. Por ejemplo, Lambda tiene límites estrictos específicos que no se pueden ajustar.

Por ejemplo, cuando se diseña una aplicación python para que se ejecute en una función Lambda, es necesario evaluar la aplicación para determinar si existe alguna posibilidad de que Lambda se ejecute durante más de 15 minutos. Si el código puede ejecutarse durante más tiempo de este límite de cuota de servicio, deben considerarse tecnologías o diseños alternativos. Si se alcanza el límite después del despliegue en producción, la aplicación sufrirá degradación e interrupciones hasta que pueda remediarse. A diferencia de las cuotas flexibles, no existe ningún método para cambiar estos límites, ni siquiera en caso de eventos de emergencia de gravedad 1.

Una vez que la aplicación se ha desplegado en un entorno de pruebas, se deben utilizar estrategias para averiguar si se puede alcanzar algún límite estricto. Las pruebas de estrés, las pruebas de carga y las pruebas de caos deben formar parte del plan de pruebas de introducción.

Pasos para la implementación

- Revise la lista completa de servicios de AWS que podrían utilizarse en la fase de diseño de la aplicación.
- Revise los límites de cuotas flexibles y estrictos para todos estos servicios. En la consola de Service Quotas no se muestran todos los límites. Algunos servicios describen estos límites en ubicaciones alternativas.
- Al diseñar su aplicación, revise los impulsores empresariales y tecnológicos de la carga de trabajo, como los resultados empresariales, el caso de uso, los sistemas dependientes, los objetivos de disponibilidad y los objetos de recuperación de desastres. Permita que sus impulsores

empresariales y tecnológicos guíen el proceso para identificar el sistema distribuido adecuado para su carga de trabajo.

- Analice la carga de servicio en todas las regiones y cuentas. Muchos límites estrictos se basan en la región para los servicios. Sin embargo, algunos límites se basan en las cuentas.
- Analice el uso de recursos de las arquitecturas de resistencia durante un error zonal y un error regional. En la progresión de los diseños multirregión que utilizan enfoques activo/activo, activo/pasivo: en caliente, activo/pasivo: en frío y activo/pasivo: luz piloto, estos casos de error provocarán un mayor uso. Esto crea un caso de uso potencial para alcanzar límites estrictos.

Recursos

Prácticas recomendadas relacionadas:

- REL01-BP01 Conocimiento de las cuotas y restricciones del servicio
- REL01-BP02 Administrar cuotas de servicio en cuentas y regiones
- REL01-BP04 Supervisar y administrar cuotas
- REL01-BP05 Automatizar la administración de cuotas
- REL01-BP06 Garantizar que exista una diferencia suficiente entre las cuotas actuales y el uso máximo para permitir la conmutación por error
- REL03-BP01 Elegir cómo segmentar su carga de trabajo
- REL10-BP01 Implementar la carga de trabajo en varias ubicaciones
- REL11-BP01 Supervisar todos los componentes de la carga de trabajo para detectar errores
- REL11-BP03 Automatizar la reparación en todas las capas
- REL12-BP05 Probar la resiliencia mediante la ingeniería del caos

Documentos relacionados:

- Pilar de fiabilidad de AWS Well-Architected Framework: disponibilidad
- AWS Service Quotas (denominados anteriormente límites de servicio)
- Comprobaciones de prácticas recomendadas de AWS Trusted Advisor (consulte la sección Límites de servicio)
- AWS Limit Monitor en AWS Answers
- Límites de servicio de Amazon EC2
- ¿Qué es Service Quotas?

- Cómo solicitar un aumento de cuota
- Puntos de conexión y cuotas de servicio
- Guía del usuario de Service Quotas
- Supervisor de cuotas para AWS
- Límites de aislamiento de errores de AWS
- Disponibilidad con redundancia
- AWS para datos
- ¿Qué es la integración continua?
- ¿Qué es la entrega continua?
- Socio de APN: socios que pueden ayudar con la administración de la configuración
- Managing the account lifecycle in account-per-tenant SaaS environments on AWS(Administración del ciclo de vida de las cuentas en entornos SaaS de cuenta por inquilino en AWS)
- Managing and monitoring API throttling in your workloads (Administrar y supervisar la limitación de las API en sus cargas de trabajo)
- View AWS Trusted Advisor recommendations at scale with AWS Organizations(Ver recomendaciones de AWS Trusted Advisor a escala con AWS Organizations)
- Automating Service Limit Increases and Enterprise Support with AWS Control Tower(Automatización de los aumentos del límite de servicio y asistencia a empresas con AWS Control Tower)
- Acciones, recursos y claves de condición de los servicios de Service Quotas

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas
- View and Manage Quotas for AWS Services Using Service Quotas (Ver y administrar cuotas para AWS Services con Service Quotas)
- AWS IAM Quotas Demo (Demostración de las cuotas de AWS IAM)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small (AWS re:Invent 2018: Cerrar los bucles y abrir las mentes: cómo asumir el control de los sistemas grandes y pequeños)

Herramientas relacionadas:

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP04 Supervisar y administrar cuotas

Evalúe el uso potencial y aumente las cuotas pertinentemente, lo que permitirá un crecimiento planificado del uso.

Resultado deseado: se despliegan sistemas activos y automatizados que administran y supervisan. Estas soluciones operativas garantizan que los umbrales de uso de las cuotas están a punto de alcanzarse. Esto se solucionaría de forma proactiva solicitando cambios en las cuotas.

Antipatrones usuales:

- No se configura la supervisión para comprobar el umbral de cuota de servicio.
- No se configura la supervisión de los límites estrictos, aunque esos valores no puedan modificarse.
- Se supone que el tiempo necesario para solicitar y asegurar un cambio de cuota flexible es inmediato o un de corta duración.
- Se configuran alarmas de aproximación para cuotas de servicio sin contar con ningún proceso para responder a una alerta.
- Solo se configuran alarmas para servicios compatibles con AWS Service Quotas y no se supervisan otros servicios de AWS.
- No se considera la administración de cuotas para diseños de resiliencia multirregional, como los enfoques activo/activo, activo/pasivo: en caliente, activo/pasivo: en frío y activo/pasivo: luz piloto.
- No se evalúan las diferencias de cuota entre regiones.
- No se evalúan las necesidades de cada región para una solicitud específica de aumento de cuota.

- No se aprovechan las plantillas de administración de cuotas multirregión.

Beneficios de establecer esta práctica recomendada: el seguimiento automático de las cuotas de servicio de AWSService Quotas y la supervisión del uso en comparación con dichas cuotas le permitirá comprobar cuándo se acerca al límite de una cuota. También puede utilizar estos datos de supervisión para ayudar a limitar cualquier degradación debida al agotamiento de cuotas.

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

Guía para la implementación

Para los servicios admitidos, puede supervisar las cuotas por medio de la configuración de varios servicios diferentes que pueden evaluar y luego enviar alertas o alarmas. Esto puede ayudar a supervisar el uso y alertarle de que se aproxima a las cuotas. Estas alarmas se pueden desencadenar con AWS Config, las funciones de Lambda, Amazon CloudWatch o con AWS Trusted Advisor. También puede usar filtros de métricas en CloudWatch Logs para buscar y extraer patrones en registros para determinar si el uso se aproxima a los umbrales de las cuotas.

Pasos para la implementación

Para la supervisión:

- Capture el consumo de recursos actual (por ejemplo, buckets o instancias). Utilice operaciones de API de servicio, como la API `DescribeInstances` de Amazon EC2 para recopilar el consumo actual de recursos.
- Capture las cuotas actuales que son esenciales y aplicables a los servicios que utiliza:
  - AWS Service Quotas
  - AWS Trusted Advisor
  - Documentación de AWS
  - Páginas específicas de los servicios de AWS
  - AWS Command Line Interface (AWS CLI)
  - Kit de desarrollo de la nube de AWS (AWS CDK)
- Utilice AWS Service Quotas, un servicio de AWS que le ayuda a administrar sus cuotas para más de 250 servicios de AWS desde una ubicación.
- Utilice los límites de servicio de Trusted Advisor para supervisar sus límites de servicio actuales en diversos umbrales.

- Utilice el historial de cuotas de servicio (consola o AWS CLI) para comprobar los aumentos regionales.
- Compare los cambios de cuota de servicio en cada región y cada cuenta para crear equivalencias, si es necesario.

Para la administración:

- Automatizada: configure una regla personalizada de AWS Config para analizar las cuotas de servicio en todas las regiones y comparar las diferencias.
- Automatizada: configure una función programada de Lambda para analizar las cuotas de servicio en todas las regiones y comparar las diferencias.
- Manual: analice la cuota de servicios a través de la AWS CLI, la API o la consola de AWS para analizar las cuotas de servicio en todas las regiones y comparar las diferencias. Informe de las diferencias.
- Si se identifican diferencias en las cuotas entre las regiones, solicite un cambio de cuota, si es necesario.
- Revise el resultado de todas las solicitudes.

Recursos

Prácticas recomendadas relacionadas:

- REL01-BP01 Conocimiento de las cuotas y restricciones del servicio
- REL01-BP02 Administrar cuotas de servicio en cuentas y regiones
- REL01-BP03 Adaptar las cuotas de servicio fijas y las restricciones a través de la arquitectura
- REL01-BP05 Automatizar la administración de cuotas
- REL01-BP06 Garantizar que exista una diferencia suficiente entre las cuotas actuales y el uso máximo para permitir la conmutación por error
- REL03-BP01 Elegir cómo segmentar su carga de trabajo
- REL10-BP01 Implementar la carga de trabajo en varias ubicaciones
- REL11-BP01 Supervisar todos los componentes de la carga de trabajo para detectar errores
- REL11-BP03 Automatizar la reparación en todas las capas
- REL12-BP05 Probar la resiliencia mediante la ingeniería del caos

Documentos relacionados:

- Pilar de fiabilidad de AWS Well-Architected Framework: disponibilidad
- AWS Service Quotas (denominados anteriormente límites de servicio)
- Comprobaciones de prácticas recomendadas de AWS Trusted Advisor (consulte la sección Límites de servicio)
- AWS Limit Monitor en AWS Answers
- Límites de servicio de Amazon EC2
- ¿Qué es Service Quotas?
- Cómo solicitar un aumento de cuota
- Puntos de conexión y cuotas de servicio
- Guía del usuario de Service Quotas
- Supervisor de cuotas para AWS
- Límites de aislamiento de errores de AWS
- Disponibilidad con redundancia
- AWS para datos
- ¿Qué es la integración continua?
- ¿Qué es la entrega continua?
- Socio de APN: socios que pueden ayudar con la administración de la configuración
- Managing the account lifecycle in account-per-tenant SaaS environments on AWS(Administración del ciclo de vida de las cuentas en entornos SaaS de cuenta por inquilino en AWS)
- Managing and monitoring API throttling in your workloads (Administrar y supervisar la limitación de las API en sus cargas de trabajo)
- View AWS Trusted Advisor recommendations at scale with AWS Organizations(Ver recomendaciones de AWS Trusted Advisor a escala con AWS Organizations)
- Automating Service Limit Increases and Enterprise Support with AWS Control Tower(Automatización de los aumentos del límite de servicio y asistencia a empresas con AWS Control Tower)
- Acciones, recursos y claves de condición de los servicios de Service Quotas

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas

- View and Manage Quotas for AWS Services Using Service Quotas (Ver y administrar cuotas para AWS Services con Service Quotas)
- AWS IAM Quotas Demo (Demostración de las cuotas de AWS IAM)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small (AWS re:Invent 2018: Cerrar los bucles y abrir las mentes: cómo asumir el control de los sistemas grandes y pequeños)

Herramientas relacionadas:

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP05 Automatizar la administración de cuotas

Implemente herramientas para alertarle cuando se acerque a los límites. Puede automatizar las solicitudes de incremento de cuotas utilizando las API de AWS Service Quotas y automatizar las solicitudes de incremento de cuotas.

Si integra su base de datos de administración de configuraciones (CMDB) o su sistema de emisión de tiques con Service Quotas, puede automatizar el seguimiento de las solicitudes de aumento de cuotas y las cuotas actuales. Además del SDK de AWS, Service Quotas ofrece automatización utilizando AWS Command Line Interface (AWS CLI).

Patrones de uso no recomendados comunes:

- Realizar el seguimiento de las cuotas y el uso en hojas de cálculo.
- Ejecutar informes de uso cada día, semana o mes y después comparar el uso con las cuotas.

Beneficios de establecer esta práctica recomendada: El control automático de las cuotas de servicio de AWS y la supervisión del uso en comparación con dichas cuotas le permite comprobar cuándo se acerca a una cuota. Puede configurar la automatización para que le ayude a solicitar un aumento de cuota cuando resulte necesario. Es posible que quiera plantearse la reducción de algunas cuotas cuando su uso adopte una tendencia opuesta para materializar los beneficios de un menor riesgo (en caso de que sus credenciales se hayan visto comprometidas) y el ahorro de costes.

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

Guía para la implementación

- Configure una supervisión automatizada. Implemente herramientas con SDK para alertarle cuando se acerque a los límites.
  - Utilice Service Quotas y potencie el servicio con una solución de supervisión de cuotas automatizada, como AWS Limit Monitor o una oferta de AWS Marketplace.
    - ¿Qué es Service Quotas?
    - Supervisor de cuotas en AWS: solución de AWS
  - Configure respuestas desencadenadas en función de umbrales de cuotas con las API de Amazon SNS y AWS Service Quotas.
  - Automatización de pruebas.
    - Configure umbrales de límites.
    - Integre con eventos de cambio de AWS Config, canalizaciones de despliegue, Amazon EventBridge o terceros.
    - Defina de forma artificial umbrales de cuota bajos para probar las respuestas.
    - Configure desencadenadores para realizar acciones pertinentes en las notificaciones y póngase en contacto con AWS Support cuando sea necesario.
    - Desencadene manualmente eventos de cambio.
    - Ejecute un día de juego para probar el proceso de cambio de aumento de cuotas.

Recursos

Documentos relacionados:

- Socio de APN: socios que pueden ayudar con la administración de la configuración
- AWS Marketplace: productos de CMDB que ayudan a hacer un seguimiento de los límites
- AWS Service Quotas (denominadas anteriormente límites de servicio)

- Comprobaciones de prácticas recomendadas de AWS Trusted Advisor (consulte la sección Límites de servicio)
- Supervisor de cuotas en AWS: solución de AWS
- Límites de servicio de Amazon EC2
- ¿Qué es Service Quotas?

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas

REL01-BP06 Garantizar que exista una diferencia suficiente entre las cuotas actuales y el uso máximo para permitir la conmutación por error

Cuando un recurso falla o es inaccesible, ese recurso puede seguir computando para una cuota dada hasta que se finalice correctamente. Compruebe que sus cuotas cubran el solapamiento de los recursos averiados o inaccesibles y sus sustitutos. A la hora de calcular esta brecha debe tener en cuenta casos de uso como errores de red, errores de la zona de disponibilidad o errores regionales.

Resultado deseado: los errores pequeños o grandes en los recursos o en la accesibilidad de los recursos pueden cubrirse dentro de los umbrales de servicio actuales. En la planificación de recursos se tienen en cuenta los errores de zona, de red o, incluso, regionales.

Antipatrones usuales:

- Se establecen cuotas de servicio sobre la base de las necesidades actuales sin tener en cuenta los casos de conmutación por error.
- No se tienen en cuenta los principios de estabilidad estática al calcular la cuota máxima de un servicio.
- No se tiene en cuenta el potencial de recursos inaccesibles al calcular la cuota total necesaria para cada región.
- No se tienen en cuenta los límites de aislamiento de errores del servicio de AWS para algunos servicios y sus posibles patrones de uso anómalos.

Beneficios de establecer esta práctica recomendada: cuando un evento de interrupción del servicio afecta a la disponibilidad de la aplicación, la nube le permite implementar estrategias para mitigar o recuperarse de estos eventos. Estas estrategias suelen incluir la creación de recursos adicionales

para sustituir aquellos que han experimentado algún error o a los que no se puede acceder. La estrategia de cuotas se adaptaría a estas condiciones de conmutación por error y no introduciría degradaciones adicionales debidas al agotamiento de los límites de servicio.

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

Guía para la implementación

Al evaluar los límites de cuota, considere los casos de conmutación por error que podrían producirse debido a alguna degradación. Deben tenerse en cuenta los siguientes tipos de casos de conmutación por error:

- Una VPC interrumpida o inaccesible.
- Una subred inaccesible.
- Una zona de disponibilidad que se ha degradado lo suficiente como para afectar a la accesibilidad de muchos recursos.
- Varias rutas de red o puntos de entrada y salida bloqueados o modificados.
- Una región que se ha degradado lo suficiente como para afectar a la accesibilidad de muchos recursos.
- Hay numerosos recursos, pero no todos se ven afectados por un error en una región o zona de disponibilidad.

Los errores como los de la lista anterior podrían ser el detonante del inicio de un evento de conmutación por error. La decisión de conmutar por error es única para cada situación y cliente, ya que el efecto empresarial puede variar drásticamente. Sin embargo, cuando operacionalmente se decide conmutar por error aplicaciones o servicios, la planificación de la capacidad de los recursos en la ubicación de la conmutación por error y sus cuotas correspondientes deben abordarse antes del evento.

Revise las cuotas de servicio para cada servicio teniendo en cuenta los picos más altos de lo normal que puedan producirse. Estos picos pueden estar relacionados con recursos a los que no se puede acceder debido a la red o a los permisos, pero que siguen activos. Los recursos activos no finalizados seguirán contando para el límite de cuota de servicio.

Pasos para la implementación

- Asegúrese de que haya una diferencia suficiente entre la cuota de servicio y el uso máximo para permitir la conmutación por error o una pérdida de accesibilidad.

- Determine sus cuotas de servicio, teniendo en cuenta sus patrones de despliegue, los requisitos de disponibilidad y el crecimiento del consumo.
- Solicite aumentos de la cuota si fuera necesario. Planifique el tiempo necesario para que se cumplan las solicitudes de aumentos de cuotas.
- Determine sus requisitos de fiabilidad (también conocidos como «número de nueves»).
- Establezca sus escenarios de error (por ejemplo, la pérdida de componentes, una zona de disponibilidad o una región).
- Establezca su metodología de despliegue (por ejemplo, valor controlado, azul-verde, rojo-negro o continua).
- Incluya un búfer adecuado (por ejemplo, del 15 %) en el límite actual.
- Incluya cálculos de estabilidad estática (zonal y regional) cuando proceda.
- Planifique el crecimiento de consumo (por ejemplo, supervise sus tendencias de consumo).
- Considere la repercusión de la estabilidad estática para las cargas de trabajo más fundamentales. Evalúe los recursos conforme a un sistema estáticamente estable en todas las regiones y zonas de disponibilidad.
- Considere el uso de reservas de capacidad bajo demanda para programar la capacidad antes de que se produzca una conmutación por error. Puede ser una estrategia útil durante las programaciones comerciales más cruciales para reducir los riesgos potenciales de obtener la cantidad y el tipo correctos de recursos durante la conmutación por error.

Recursos

Prácticas recomendadas relacionadas:

- REL01-BP01 Conocimiento de las cuotas y restricciones del servicio
- REL01-BP02 Administrar cuotas de servicio en cuentas y regiones
- REL01-BP03 Adaptar las cuotas de servicio fijas y las restricciones a través de la arquitectura
- REL01-BP04 Supervisar y administrar cuotas
- REL01-BP05 Automatizar la administración de cuotas
- REL03-BP01 Elegir cómo segmentar su carga de trabajo
- REL10-BP01 Implementar la carga de trabajo en varias ubicaciones
- REL11-BP01 Supervisar todos los componentes de la carga de trabajo para detectar errores

- REL11-BP03 Automatizar la reparación en todas las capas
- REL12-BP05 Probar la resiliencia mediante la ingeniería del caos

Documentos relacionados:

- Pilar de fiabilidad de AWS Well-Architected Framework: disponibilidad
- AWS Service Quotas (denominados anteriormente límites de servicio)
- Comprobaciones de prácticas recomendadas de AWS Trusted Advisor (consulte la sección Límites de servicio)
- AWS Limit Monitor en AWS Answers
- Límites de servicio de Amazon EC2
- ¿Qué es Service Quotas?
- Cómo solicitar un aumento de cuota
- Puntos de conexión y cuotas de servicio
- Guía del usuario de Service Quotas
- Supervisor de cuotas para AWS
- Límites de aislamiento de errores de AWS
- Disponibilidad con redundancia
- AWS para datos
- ¿Qué es la integración continua?
- ¿Qué es la entrega continua?
- Socio de APN: socios que pueden ayudar con la administración de la configuración
- Managing the account lifecycle in account-per-tenant SaaS environments on AWS(Administración del ciclo de vida de las cuentas en entornos SaaS de cuenta por inquilino en AWS)
- Managing and monitoring API throttling in your workloads (Administrar y supervisar la limitación de las API en sus cargas de trabajo)
- View AWS Trusted Advisor recommendations at scale with AWS Organizations(Ver recomendaciones de AWS Trusted Advisor a escala con AWS Organizations)
- Automating Service Limit Increases and Enterprise Support with AWS Control Tower(Automatización de los aumentos del límite de servicio y asistencia a empresas con AWS Control Tower)

- Acciones, recursos y claves de condición de los servicios de Service Quotas

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas
- View and Manage Quotas for AWS Services Using Service Quotas (Ver y administrar cuotas para AWS Services con Service Quotas)
- AWS IAM Quotas Demo (Demostración de las cuotas de AWS IAM)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small (AWS re:Invent 2018: Cerrar los bucles y abrir las mentes: cómo asumir el control de los sistemas grandes y pequeños)

Herramientas relacionadas:

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

FIABILIDAD 2. ¿Cómo planifica la topología de la red?

Suele haber cargas de trabajo en distintos entornos. Entre estos se incluyen los entornos de la nube (tanto públicamente accesibles como privados), y posiblemente la infraestructura del centro de datos existente. Los planes deben incluir consideraciones de la red, como la conectividad dentro de los sistemas y entre ellos, la administración de las direcciones IP públicas, la administración de las direcciones IP privadas y la resolución de nombres de dominio.

Prácticas recomendadas

- REL02-BP01 Usar conectividad de red de alta disponibilidad para los puntos de conexión públicos de la carga de trabajo
- REL02-BP02 Aprovisionar conectividad redundante entre las redes privadas en la nube y los entornos locales
- REL02-BP03 Garantizar que la asignación de subredes IP tenga en cuenta la expansión y la disponibilidad
- REL02-BP04 Preferir topologías radiales (hub-and-spoke) a una conexión en malla de varios a varios
- REL02-BP05 Emplear intervalos no superpuestos de direcciones IP privadas en todos los espacios de direcciones privadas que estén conectados

REL02-BP01 Usar conectividad de red de alta disponibilidad para los puntos de conexión públicos de la carga de trabajo

La creación de una conectividad de red de alta disponibilidad para los puntos de conexión públicos de las cargas de trabajo puede ayudarle a reducir el tiempo de inactividad debido a la pérdida de conectividad y mejorar la disponibilidad y el SLA de su carga de trabajo. Para conseguirlo, use DNS, redes de entrega de contenido (CDN), puertas de enlace de API, un equilibrador de carga o proxies inversos altamente disponibles.

Resultado deseado: es fundamental planificar, construir y poner en funcionamiento una conectividad de red de alta disponibilidad para sus puntos de conexión públicos. Si la carga de trabajo resulta inaccesible debido a una pérdida de conectividad, incluso si la carga de trabajo está en funcionamiento y disponible, los clientes verán su sistema como caído. Al combinar una conectividad de red de alta disponibilidad y resistente para los puntos de conexión públicos de la carga de trabajo, junto con una arquitectura resistente para la propia carga de trabajo, puede proporcionar la mejor disponibilidad y nivel de servicio posibles a sus clientes.

AWS Global Accelerator, Amazon CloudFront, Amazon API Gateway, las URL de función de AWS Lambda, las API de AWS AppSync y Elastic Load Balancing (ELB) ofrecen puntos de conexión públicos de alta disponibilidad. Amazon Route 53 proporciona un servicio DNS de alta disponibilidad para la resolución de nombres de dominio con el fin de verificar que las direcciones de los puntos de conexión públicos se puedan resolver.

También puede evaluar las aplicaciones de software de AWS Marketplace que proporcionen equilibrio de carga o uso de proxies.

Antipatrones usuales:

- Diseñar una carga de trabajo de alta disponibilidad sin planificar el DNS y la conectividad de red para alta disponibilidad.
- Usar direcciones de internet públicas en instancias o contenedores individuales y administrar la conectividad a ellas a con DNS.
- Usar direcciones IP en lugar de nombres de dominio para localizar los servicios.
- No hacer pruebas de escenarios en que se pierda la conectividad con sus puntos de conexión públicos.
- No analizar las necesidades de rendimiento de la red y los patrones de distribución.
- No hacer pruebas ni planificar escenarios en los que la conectividad de la red de internet a sus puntos de conexión públicos de la carga de trabajo pueda verse interrumpida.
- Proporcionar contenido (como páginas web, activos estáticos o archivos multimedia) a una gran área geográfica y no usar una red de entrega de contenido.
- No planificar en caso de que se produzcan ataques de denegación de servicio distribuido (DDoS). Los ataques DDoS corren el riesgo de cerrar el tráfico legítimo y reducir la disponibilidad para sus usuarios.

Beneficios de establecer esta práctica recomendada: se diseña una conectividad de red resistente y de alta disponibilidad que garantiza que la carga de trabajo esté accesible y disponible para los usuarios.

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

Guía para la implementación

El enrutamiento del tráfico es el núcleo de la creación de una conectividad de red de alta disponibilidad para sus puntos de conexión públicos. Para verificar que el tráfico puede llegar a los puntos de conexión, el DNS debe ser capaz de resolver los nombres de dominio en sus direcciones IP correspondientes. Utilice un sistema de nombres de dominio (DNS) escalable y de alta disponibilidad como Amazon Route 53 para administrar los registros DNS de su dominio. También puede utilizar las comprobaciones de estado proporcionadas por Amazon Route 53. Las comprobaciones de estado verifican que la aplicación sea accesible, esté disponible y funcione; se pueden configurar de manera que imiten el comportamiento de su usuario, como la solicitud de una página web o una URL concreta. En caso de error, Amazon Route 53 responde a las solicitudes de resolución de DNS y dirige el tráfico únicamente a los puntos de conexión de estado. También puede plantearse el uso de las capacidades de DNS geográfico y enrutamiento basado en la latencia que ofrece Amazon Route 53.

Para comprobar que la carga de trabajo en sí sea de alta disponibilidad, utilice Elastic Load Balancing (ELB). Amazon Route 53 se puede utilizar para dirigir el tráfico a ELB, que distribuye el tráfico a las instancias de computación de destino. También puede utilizar Amazon API Gateway junto con AWS Lambda para una solución sin servidor. Los clientes también pueden ejecutar cargas de trabajo en varias Regiones de AWS. Con el patrón activo/activo multisitio, la carga de trabajo puede atender tráfico de varias regiones. Con un patrón activo/pasivo multisitio, la carga de trabajo atiende tráfico desde la región activa, mientras que los datos se replican en la región secundaria y se activan en caso de error en la región principal. Las comprobaciones de estado de Route 53 se pueden utilizar para controlar la conmutación por error de DNS desde cualquier punto de conexión en una región principal y a un punto de conexión en una región secundaria, lo que verifica que la carga de trabajo esté accesible y disponible para los usuarios.

Amazon CloudFront proporciona una API sencilla para distribuir contenido con baja latencia y altas velocidades de transferencia de datos atendiendo las solicitudes mediante una red de ubicaciones periféricas en todo el mundo. Las redes de entrega de contenido (CDN) atienden a los clientes proporcionándoles contenido ubicado o almacenado en caché en una ubicación cercana al usuario. Esto también mejora la disponibilidad de su aplicación, ya que la carga de contenido se desplaza de sus servidores a las ubicaciones periféricas de CloudFront. Las ubicaciones periféricas y las cachés periféricas regionales mantienen copias en caché de su contenido cerca de sus usuarios, lo que permite una recuperación rápida y aumenta la accesibilidad y la disponibilidad de su carga de trabajo.

Para cargas de trabajo con usuarios distribuidos geográficamente, AWS Global Accelerator ayuda a mejorar la disponibilidad y el rendimiento de las aplicaciones. AWS Global Accelerator proporciona direcciones IP estáticas de difusión por proximidad que sirven como punto de entrada fijo a su aplicación alojada en una o más Regiones de AWS. Esto permite que el tráfico entre en la red global de AWS lo más cerca posible de sus usuarios, lo que mejora la accesibilidad y disponibilidad de su carga de trabajo. AWS Global Accelerator también supervisa el estado de los puntos de conexión de su aplicación mediante comprobaciones de estado de TCP, HTTP y HTTPS. Cualquier cambio en el estado o la configuración de sus puntos de conexión activa el redireccionamiento del tráfico de usuario a puntos de conexión en buen estado que ofrezcan el mejor rendimiento y disponibilidad a los usuarios. Además, AWS Global Accelerator cuenta con un diseño de aislamiento de errores que utiliza dos direcciones IPv4 estáticas atendidas por zonas de red independientes que aumentan la disponibilidad de las aplicaciones.

Para ayudar a proteger a los clientes de ataques DDoS, AWS proporciona AWS Shield Standard. Shield Standard se activa automáticamente y protege de los ataques habituales a la infraestructura (capas 3 y 4), como las inundaciones de SYN/UDP y los ataques de reflexión, para respaldar la alta

disponibilidad de sus aplicaciones en AWS. Para obtener protecciones adicionales contra ataques más sofisticados y grandes (como inundaciones de UDP) y ataques de agotamiento de estado (como inundaciones de TCP SYN), y para ayudar a proteger sus aplicaciones que se ejecutan en Amazon Elastic Compute Cloud (Amazon EC2), Elastic Load Balancing (ELB), Amazon CloudFront, AWS Global Accelerator y Route 53, puede considerar el uso de AWS Shield Advanced. Para la protección contra ataques en la capa de aplicación como HTTP POST o inundaciones de GET, utilice AWS WAF. AWS WAF puede utilizar condiciones de direcciones IP, encabezados HTTP, cuerpos HTTP, cadenas de URI, inyección de código SQL y scripting entre sitios para determinar si una solicitud debe bloquearse o permitirse.

Pasos para la implementación

1. Configure DNS de alta disponibilidad: Amazon Route 53 es un servicio web de sistema de nombres de dominio (DNS) (DNS) altamente disponible y escalable. Route 53 conecta las solicitudes de los usuarios con las aplicaciones de Internet que se ejecutan en AWS o localmente. Para obtener más información, consulte Configuración de Amazon Route 53 como servicio DNS.
2. Configure comprobaciones de estado: cuando utilice Route 53, verifique que solo se puedan resolver los destinos en buen estado. Empiece por Creación de comprobaciones de estado de Route 53 y configuración de la conmutación por error a nivel de DNS. Es importante tener en cuenta los siguientes aspectos a la hora de configurar las comprobaciones de estado:
   a. Cómo determina Amazon Route 53 si la comprobación de estado es correcta
   b. Creación, actualización y eliminación de comprobaciones de estado
   c. Supervisar el estado de la comprobación de estado y recibir notificaciones
   d. Prácticas recomendadas de DNS de Amazon Route 53
3. Conecte su servicio DNS a sus puntos de conexión.
   a. Al utilizar Elastic Load Balancing como destino de su tráfico, cree un registro de alias mediante Amazon Route 53 que apunte al punto de conexión regional de su equilibrador de carga. Durante la creación del registro de alias, establezca la opción de evaluación de estado del destino a Sí.
   b. Para cargas de trabajo sin servidor o API privadas cuando se utilice API Gateway, utilice Route 53 para dirigir el tráfico a API Gateway.
4. Decida la red de entrega de contenido.
   a. A la hora de entregar contenido mediante las ubicaciones periféricas más cercanas al usuario, comience por comprender cómo CloudFront entrega el contenido.

   b. Empiece con una distribución sencilla de CloudFront. CloudFront sabrá entonces desde dónde desea que se entregue el contenido, así como los detalles sobre cómo realizar el seguimiento y administrar la entrega de contenido. Es importante comprender y tener en cuenta los siguientes aspectos al configurar la distribución de CloudFront:
      i. Cómo funciona el almacenamiento en caché con ubicaciones periféricas de CloudFront
      ii. Incrementar la proporción de solicitudes que se atienden directamente desde las cachés de CloudFront (tasa de aciertos de caché)
      iii. Uso de Amazon CloudFront Origin Shield
      iv. Optimización de alta disponibilidad con conmutación por error de origen de CloudFront
5. Configure la protección de la capa de aplicación: AWS WAF le ayuda a protegerse contra ataques web y bots habituales que pueden afectar a la disponibilidad, comprometer la seguridad o consumir demasiados recursos. Para obtener una comprensión más profunda, revise cómo funciona AWS WAF y, cuando esté listo para implementar protecciones contra inundaciones de HTTP POST Y GET en la capa de aplicación, revise Introducción a AWS WAF. También puede utilizar AWS WAF con CloudFront. Consulte la documentación sobre cómo funciona AWS WAF con las características de Amazon CloudFront.
6. Configure protección DDoS adicional: de forma predeterminada, todos los clientes de AWS reciben protección frente a los ataques DDoS más habituales y frecuentes de la capa de red y transporte dirigidos a su sitio web o aplicación con AWS Shield Standard y sin ningún cargo adicional. Para obtener protección adicional de las aplicaciones orientadas a Internet que se ejecutan en Amazon EC2, Elastic Load Balancing, Amazon CloudFront, AWS Global Accelerator y Amazon Route 53 considere AWS Shield Advanced y revise los ejemplos de arquitecturas resistentes a DDoS. Para proteger su carga de trabajo y sus puntos de conexión públicos de ataques DDoS, consulte Introducción a AWS Shield Advanced.

Recursos

Prácticas recomendadas relacionadas:

- REL10-BP01 Implementar la carga de trabajo en varias ubicaciones
- REL10-BP02 Seleccionar las ubicaciones adecuadas para el despliegue en varias ubicaciones
- REL11-BP04 Confiar en el plano de datos y no en el plano de control durante la recuperación
- REL11-BP06 Enviar notificaciones cuando los eventos afecten a la disponibilidad

Documentos relacionados:

- Socio de APN: socios que pueden ayudarle a planificar sus redes
- AWS Marketplace for Network Infrastructure (AWS Marketplace para la infraestructura de red)
- ¿Qué es AWS Global Accelerator?
- ¿Qué es Amazon CloudFront?
- ¿Qué es Amazon Route 53?
- ¿Qué es Elastic Load Balancing?
- Network Connectivity capability - Establishing Your Cloud Foundations (Capacidad de conectividad de red: establecimiento de las bases de su nube)
- What is Amazon API Gateway? (¿Qué es Amazon API Gateway?)
- What are AWS WAF, AWS Shield, and AWS Firewall Manager? (¿Qué son AWS WAF, AWS Shield y AWS Firewall Manager?)
- What is Amazon Route 53 Application Recovery Controller? (¿Qué es el Controlador de recuperación de aplicaciones de Amazon Route 53?)
- Configurar las comprobaciones de estado personalizadas para la conmutación por error de DNS

Vídeos relacionados:

- AWS re:Invent 2022 - Improve performance and availability with AWS Global Accelerator (AWS re:Invent 2022: Mejorar el rendimiento y la disponibilidad con AWS Global Accelerator)
- AWS re:Invent 2020: Global traffic management with Amazon Route 53 (AWS re:Invent 2020: Administración de tráfico global con Amazon Route 53)
- AWS re:Invent 2022 - Operating highly available Multi-AZ applications (AWS re:Invent 2022: Funcionamiento de aplicaciones multi-AZ de alta disponibilidad)
- AWS re:Invent 2022 - Dive deep on AWS networking infrastructure (AWS re:Invent 2022: Profundización en la infraestructura de red de AWS)
- AWS re:Invent 2022 - Building resilient networks (AWS re:Invent 2022: Creación de redes resistentes)

Ejemplos relacionados:

- Disaster Recovery with Amazon Route 53 Application Recovery Controller (ARC) (Recuperación de desastres con el controlador de recuperación de aplicaciones [ARC] de Amazon Route 53)
- Reliability Workshops (Talleres de fiabilidad)

- AWS Global Accelerator Workshop (Taller de AWS Global Accelerator)

REL02-BP02 Aprovisionar conectividad redundante entre las redes privadas en la nube y los entornos locales

Use varias conexiones de AWS Direct Connect o túneles VPN entre redes privadas desplegadas por separado. Use varias ubicaciones de Direct Connect para tener alta disponibilidad. Si utiliza varias Regiones de AWS, garantice la redundancia en al menos dos de ellas. Puede interesarle evaluar dispositivos de AWS Marketplace que terminen las VPN. Si utiliza dispositivos de AWS Marketplace, implemente instancias redundantes para obtener alta disponibilidad en diferentes zonas de disponibilidad.

AWS Direct Connect es una solución de servicios en la nube que facilita el establecimiento de una conexión de red dedicada desde su entorno local a AWS. Gracias a Direct Connect Gateway, su centro de datos local puede conectarse a varias VPC de AWS repartidas por varias Regiones de AWS.

Esta redundancia soluciona los posibles errores que afectan a la resiliencia de la conectividad:

- ¿Cómo puede resistir los errores su topología?
- ¿Qué pasa si no configuro correctamente algo y elimino la conectividad?
- ¿Podrá gestionar un aumento inesperado del tráfico o del uso de sus servicios?
- ¿Podrá absorber un intento de ataque de denegación de servicio distribuido (DDoS)?

Cuando conecte su VPC a su centro de datos local a través de una VPN, deberá tener en cuenta los requisitos de resiliencia y ancho de banda que necesita cuando seleccione el proveedor y el tamaño de la instancia en la que necesita ejecutar el dispositivo. Si usa un dispositivo VPN que no es resistente en su implementación, debe tener una conexión redundante mediante un segundo dispositivo. Para todas estas situaciones, debe definir un tiempo aceptable para la recuperación y hacer pruebas para asegurarse de que puede cumplir esos requisitos.

Si decide conectar su VPC a su centro de datos mediante una conexión de Direct Connect y necesita que esta conexión tenga una alta disponibilidad, disponga de conexiones de Direct Connect redundantes desde cada centro de datos. La conexión redundante debe utilizar una segunda conexión de Direct Connect desde una ubicación diferente a la primera. Si tiene varios centros de datos, asegúrese de que las conexiones terminen en diferentes ubicaciones. Use el kit de herramientas de resiliencia de Direct Connect como ayuda para su configuración.

Si decide conmutar por error a una VPN por Internet mediante Site-to-Site VPN, es importante entender que admite hasta 1,25 Gbps de rendimiento por túnel VPN, pero no admite rutas múltiples de igual coste (ECMP) para el tráfico de salida en el caso de múltiples túneles de AWS Managed VPN que terminen en el mismo VGW. No le recomendamos que utilice AWS Managed VPN como respaldo de las conexiones de Direct Connect, a menos que pueda tolerar velocidades inferiores a 1 Gbps durante la conmutación por error.

También puede utilizar los puntos de conexión de VPC para conectar de forma privada la VPC a los servicios de AWS admitidos y a los servicios de punto de conexión de VPC con tecnología de AWS PrivateLink sin atravesar la Internet pública. Los puntos de conexión son dispositivos virtuales. Son componentes de VPC escalados horizontalmente, redundantes y de alta disponibilidad. Permiten la comunicación entre las instancias de la VPC y los servicios sin que ello suponga riesgos de disponibilidad o restricciones de ancho de banda en el tráfico de su red.

Patrones de uso no recomendados comunes:

- Tener un solo proveedor de conectividad entre la red local y AWS.
- Usar las funciones de conectividad de la conexión de AWS Direct Connect, pero tener una sola conexión.
- Tener una sola ruta para su conectividad de VPN

Beneficios de establecer esta práctica recomendada: al implementar la conectividad redundante entre el entorno en la nube y su entorno corporativo o local, puede garantizar que los servicios dependientes entre los entornos se puedan comunicar sin problemas.

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

Guía para la implementación

- Asegúrese de que tiene conectividad de alta disponibilidad entre AWS y el entorno local. Use varias conexiones de AWS Direct Connect o túneles VPN entre redes privadas desplegadas por separado. Use varias ubicaciones de Direct Connect para tener alta disponibilidad. Si utiliza varias Regiones de AWS, garantice la redundancia en al menos dos de ellas. Puede interesarle evaluar dispositivos de AWS Marketplace que terminen las VPN. Si utiliza dispositivos de AWS Marketplace, implemente instancias redundantes para obtener alta disponibilidad en diferentes zonas de disponibilidad.

- Asegúrese de que dispone de una conexión redundante a su entorno local. Es posible que necesite conexiones redundantes a múltiples Regiones de AWS para lograr sus necesidades de disponibilidad.
  - Recomendaciones sobre resiliencia de AWS Direct Connect
  - Usar conexiones de VPN de sitio a sitio para proporcionar conmutación por error
    - Use operaciones de la API de servicio para identificar el uso correcto de los circuitos de AWS Direct Connect.
      - DescribeConnections
      - DescribeConnectionsOnInterconnect
      - DescribeDirectConnectGatewayAssociations
      - DescribeDirectConnectGatewayAttachments
      - DescribeDirectConnectGateways
      - DescribeHostedConnections
      - DescribeInterconnects
    - Si solo existe una conexión de Direct Connect o no existe ninguna, configure túneles VPN redundantes hacia sus puertas de enlace privadas virtuales.
      - ¿Qué es AWS Site-to-Site VPN?
- Capture su conectividad actual (por ejemplo, Direct Connect, puertas de enlace privadas virtuales, dispositivos de AWS Marketplace).
  - Use operaciones de la API de servicio para consultar la configuración de las conexiones de Direct Connect.
    - DescribeConnections
    - DescribeConnectionsOnInterconnect
    - DescribeDirectConnectGatewayAssociations
    - DescribeDirectConnectGatewayAttachments
    - DescribeDirectConnectGateways
    - DescribeHostedConnections
    - DescribeInterconnects
  - Utilice las operaciones de la API de servicios para recopilar las puertas de enlace privadas virtuales cuando las tablas de enrutamiento las utilicen.
    - DescribeVpnGateways

    - DescribeRouteTables
- Utilice las operaciones de la API de servicios para recopilar las aplicaciones de AWS Marketplace en las que las tablas de enrutamiento las usan.
    - DescribeRouteTables

Recursos

Documentos relacionados:

- Socio de APN: socios que pueden ayudarle a planificar sus redes
- Recomendaciones sobre resiliencia de AWS Direct Connect
- AWS Marketplace para la infraestructura de red
- Documento técnico sobre las opciones de conectividad de Amazon Virtual Private Cloud
- Conectividad de red de alta disponibilidad en varios centros de datos
- Usar conexiones de VPN de sitio a sitio para proporcionar conmutación por error
- Usar el kit de herramientas de resiliencia de Direct Connect para empezar
- Puntos de conexión de VPC y servicios de punto de conexión de VPC (AWS PrivateLink)
- ¿Qué es Amazon VPC?
- ¿Qué es una puerta de enlace de tránsito?
- ¿Qué es AWS Site-to-Site VPN?
- Trabajar con puertas de enlace de Direct Connect

Vídeos relacionados:

- AWS re:Invent 2018: Advanced VPC Design and New Capabilities for Amazon VPC (Diseño avanzado de VPC y funciones nuevas de Amazon VPC) (NET303)
- AWS re:Invent 2019: AWS Transit Gateway reference architectures for many VPCs (Arquitecturas de referencia de AWS Transit Gateway para muchas VPC) (NET406-R1)

REL02-BP03 Garantizar que la asignación de subredes IP tenga en cuenta la expansión y la disponibilidad

Los intervalos de direcciones IP de Amazon VPC deben ser lo suficientemente amplios como para dar cabida a los requisitos de las cargas de trabajo, como la posible expansión futura y la asignación

de direcciones IP a las subredes de las zonas de disponibilidad. Esto incluye equilibradores de carga, instancias de EC2 y aplicaciones basadas en contenedores.

Cuando planifica la topología de su red, el primer paso es definir el espacio de la dirección IP. Se deben asignar rangos de direcciones IP privadas para cada VPC (siguiendo las directrices de la RFC 1918). Facilite los siguientes requisitos como parte de este proceso:

- Permita los espacios de direcciones IP para más de una VPC por región.
- En una VPC, deje espacio para múltiples subredes que abarquen varias zonas de disponibilidad.
- Deje siempre un espacio de bloque de CIDR sin usar en una VPC para posibles expansiones futuras..
- Asegúrese de que haya espacio de direcciones IP suficiente como para satisfacer las necesidades de flotas transitorias de instancias EC2 que podría usar, como flotas de spot para el machine learning, clústeres de Amazon EMR o clústeres de Amazon Redshift.
- Tenga en cuenta que las primeras cuatro direcciones IP y las últimas direcciones IP de cada bloque CIDR de subred están reservadas y no están disponibles para usar.
- Debería planear la implementación de grandes bloques de CIDR de VPC. Tenga en cuenta que el bloque de CIDR de la VPC inicial asignado a su VPC no debe cambiar ni eliminarse, pero puede añadir bloques de CIDR adicionales que no se solapen a la VPC. Los CIDR IPv4 de subred no se pueden cambiar; sin embargo, los CIDR IPv6 sí. Tenga en cuenta que la implementación de la VPC más grande posible (/16) supone más de 65 000 direcciones IP. Solo en el espacio de la dirección IP 10.x.x.x base, puede aprovisionar 255 de estas VPC. Por tanto, de equivocarse, debería hacerlo por exceso y no por defecto, para que administrar sus VPC resulte más sencillo.

Antipatrones usuales:

- Crear VPC pequeñas
- Crear subredes pequeñas y tener que añadir subredes a las configuraciones conforme crezca
- Calcular incorrectamente cuántas direcciones IP puede usar un equilibrador de carga elástico
- Implementar muchos equilibradores carga de tráfico en las mismas subredes

Beneficios de establecer esta práctica recomendada: De esta forma, se asegurará de dar cabida al crecimiento de sus cargas de trabajo y seguir proporcionando disponibilidad cuando aumente la capacidad.

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

Guía para la implementación

- Planificar su red para que se adapte al crecimiento, la conformidad normativa y la integración con otros. El crecimiento se puede subestimar, la conformidad normativa puede variar y las adquisiciones y conexiones de redes privadas pueden ser difíciles de realizar sin una planificación adecuada.
  - Seleccione las regiones y Cuentas de AWS que correspondan según sus requisitos de servicio, latencia, normativos y de recuperación de desastres (DR).
  - Identifique sus necesidades para implementaciones regionales de VPC.
  - Identifique el tamaño de las VPC.
    - Determine si va a implementar la conectividad de varias VPC.
      - ¿Qué es una puerta de enlace de tránsito?
      - Conectividad de varias VPC en una sola región
    - Determine si necesita una red segregada para los requisitos normativos.
    - Haga las VPC tan grandes como sea posible. El bloque de CIDR de la VPC inicial asignado a su VPC no debe cambiar ni eliminarse, pero puede añadir bloques de CIDR adicionales que no se solapen a la VPC. Sin embargo, esto podría fragmentar sus intervalos de direcciones.
    - Haga las VPC tan grandes como sea posible. El bloque de CIDR de la VPC inicial asignado a su VPC no debe cambiar ni eliminarse, pero puede añadir bloques de CIDR adicionales que no se solapen a la VPC. Sin embargo, esto podría fragmentar sus intervalos de direcciones.

Recursos

Documentos relacionados:

- Socio de APN: socios que pueden ayudarle a planificar sus redes
- AWS Marketplace para la infraestructura de red
- Documento técnico sobre las opciones de conectividad de Amazon Virtual Private Cloud
- Conectividad de red de alta disponibilidad en varios centros de datos
- Conectividad de varias VPC en una sola región
- ¿Qué es Amazon VPC?

Videos relacionados:

- AWS re:Invent 2018: Diseño avanzado de VPC y funciones nuevas de Amazon VPC (NET303)
- AWS re:Invent 2019: Arquitecturas de referencia de AWS Transit Gateway para muchas VPC (NET406-R1)

REL02-BP04 Preferir topologías radiales (hub-and-spoke) a una conexión en malla de varios a varios

Si se conectan más de dos espacios de direcciones de red (por ejemplo, VPC y redes locales) a través del emparejamiento de VPC, AWS Direct Connect o VPN, utilice un modelo radial hub-and-spoke como el proporcionado por AWS Transit Gateway.

Si tiene solo dos de esas subredes, puede conectarlas entre sí, pero si aumenta el número de redes, la complejidad de esas conexiones en malla se hace insostenible. AWS Transit Gateway proporciona una forma sencilla de mantener un modelo radial que permita el enrutamiento del tráfico entre sus distintas redes.

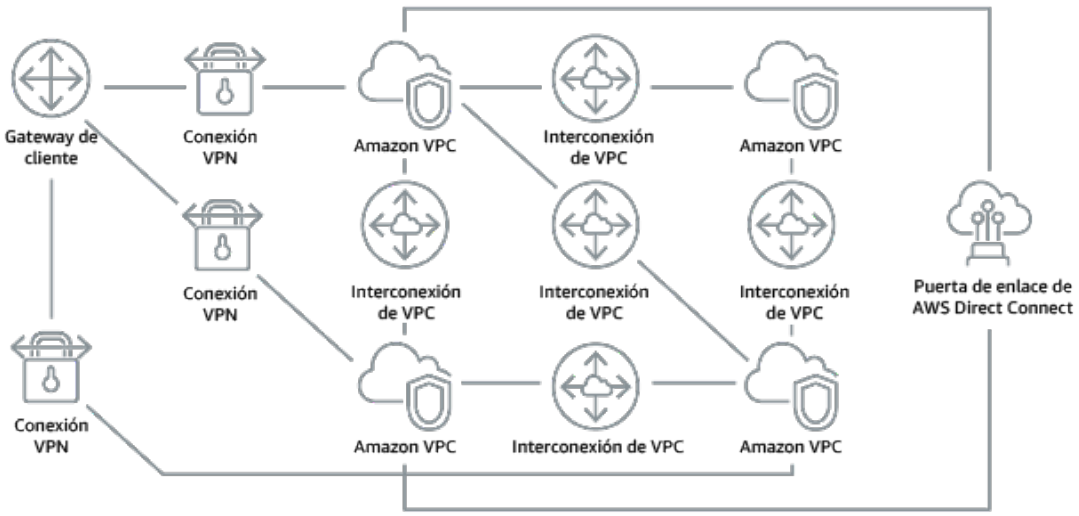


Figura 1: Sin AWS Transit Gateway, necesita emparejar cada Amazon VPC entre sí y con cada ubicación onsite mediante una conexión VPN, lo que puede volverse cada vez más completo a medida que escala el sistema.

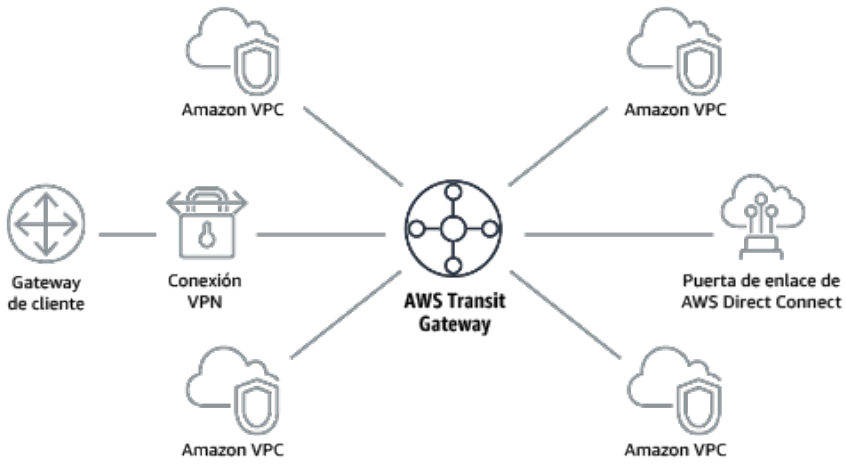


Figura 2: Con AWS Transit Gateway, simplemente tiene que conectar cada Amazon VPC o VPN a AWS Transit Gateway y esta enrutará el tráfico hacia y desde cada VPC o VPN.

Patrones de uso no recomendados comunes:

- Usar el emparejamiento de VPC para conectarse a más de dos VPC.
- Establecer varias sesiones de BGP para cada VPC para establecer conectividad que abarque las nubes virtuales privadas (VPC) repartidas entre las distintas Regiones de AWS.

Beneficios de establecer esta práctica recomendada: A medida que aumenta el número de redes, la complejidad de estas conexiones en malla se vuelve insostenible. AWS Transit Gateway proporciona una forma sencilla de mantener un modelo radial que permita el enrutamiento del tráfico entre sus distintas redes.

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

Guía para la implementación

- Priorice las topologías radiales (hub-and-spoke) sobre las conexiones en malla de varios a varios. Si se conectan más de dos espacios de direcciones de red (VPC y redes locales) a través del emparejamiento de VPC, AWS Direct Connect o VPN, utilice un modelo radial hub-and-spoke como el proporcionado por AWS Transit Gateway.
  - En el caso de haber solo dos de esas subredes, puede conectarlas entre sí, pero si aumenta el número de redes, la complejidad de esas conexiones en malla se hace insostenible. AWS

Transit Gateway proporciona una forma sencilla de mantener un modelo radial que permita el enrutamiento del tráfico entre sus distintas redes.

- ¿Qué es una puerta de enlace de tránsito?

Recursos

Documentos relacionados:

- Socio de APN: socios que pueden ayudarle a planificar sus redes
- AWS Marketplace para la infraestructura de red
- Conectividad de red de alta disponibilidad en varios centros de datos
- Puntos de conexión de VPC y servicios de punto de conexión de VPC (AWS PrivateLink)
- ¿Qué es Amazon VPC?
- ¿Qué es una puerta de enlace de tránsito?

Vídeos relacionados:

- AWS re:Invent 2018: diseño avanzado de VPC y funciones nuevas de Amazon VPC (NET303)
- AWS re:Invent 2019: arquitecturas de referencia de AWS Transit Gateway para muchas VPC (NET406-R1)

REL02-BP05 Emplear intervalos no superpuestos de direcciones IP privadas en todos los espacios de direcciones privadas que estén conectados

Los intervalos de direcciones IP de cada VPC no deben solaparse si se emparejan o conectan mediante VPN. Asimismo, debe evitar conflictos de direcciones IP entre una VPC y los entornos locales o con otros proveedores de servicios en la nube que utilice. También debe tener una forma de asignar intervalos de direcciones IP privadas cuando sea necesario.

Un sistema de administración de direcciones IP (IPAM) puede ayudar en este sentido. En AWS Marketplace hay disponibles varias IPAM.

Patrones de uso no recomendados comunes:

- Usar el mismo intervalo de direcciones IP en la VPC local o en la red corporativa
- No controlar los intervalos de direcciones IP de las VPC usadas para implementar sus cargas de trabajo

Beneficios de establecer esta práctica recomendada: La planificación activa de la red garantizará que no tenga varias instancias de la misma dirección IP en las redes interconectadas. Con esto evitará que se produzcan problemas de enrutamiento en las partes de la carga de trabajo que usan las diferentes aplicaciones.

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

Guía para la implementación

- Supervise y administre el uso de CIDR. Evalúe su potencial de uso en AWS, añada intervalos de CIDR a las VPC existentes y cree VPC para permitir un crecimiento planificado del uso.
  - Capture el consumo actual de CIDR (por ejemplo, VPC o subredes).
    - Use operaciones de la API de servicio para recopilar el consumo actual de CIDR.
  - Capture su uso actual de las subredes.
    - Use operaciones de la API de servicio para recopilar subredes por VPC en cada región.
      - DescribeSubnets
    - Registre el uso actual.
    - Determine si ha creado intervalos de IP superpuestos.
    - Calcule la capacidad de reserva.
    - Identifique los intervalos de IP superpuestos. Puede migrar a un intervalo de direcciones nuevo o usar dispositivos de traducción de redes y puertos (NAT) de AWS Marketplace si necesita conectar los intervalos superpuestos.

Recursos

Documentos relacionados:

- Socio de APN: socios que pueden ayudarle a planificar sus redes
- AWS Marketplace para la infraestructura de red
- Documento técnico sobre las opciones de conectividad de Amazon Virtual Private Cloud
- Conectividad de red de alta disponibilidad en varios centros de datos
- ¿Qué es Amazon VPC?
- ¿Qué es la IPAM?

Vídeos relacionados:

- AWS re:Invent 2018: Diseño avanzado de VPC y funciones nuevas de Amazon VPC (NET303)
- AWS re:Invent 2019: Arquitecturas de referencia de AWS Transit Gateway para muchas VPC (NET406-R1)

Arquitectura de la carga de trabajo

Preguntas

- FIABILIDAD 3. ¿Cómo diseña la arquitectura de servicio de su carga de trabajo?
- FIABILIDAD 4. ¿Cómo diseña las interacciones en un sistema distribuido para evitar los errores?
- FIABILIDAD 5. ¿Cómo diseña las interacciones en un sistema distribuido para mitigar o tolerar los errores?

FIABILIDAD 3. ¿Cómo diseña la arquitectura de servicio de su carga de trabajo?

Desarrolle cargas de trabajo escalables y fiables utilizando una arquitectura orientada a servicios (SOA) o una arquitectura de microservicios. La arquitectura orientada a servicios (SOA) es hacer que los componentes de software se puedan reutilizar mediante interfaces de servicio. La arquitectura de microservicios va más allá, para hacer que los componentes sean más pequeños y sencillos.

Prácticas recomendadas

- REL03-BP01 Elegir cómo segmentar su carga de trabajo
- REL03-BP02 Desarrollar servicios centrados en funcionalidades y dominios empresariales específicos
- REL03-BP03 Facilitar contratos de servicio por cada API

REL03-BP01 Elegir cómo segmentar su carga de trabajo

La segmentación de la carga de trabajo es importante a la hora de determinar los requisitos de resiliencia de su aplicación. La arquitectura monolítica debe evitarse siempre que sea posible. En su lugar, considere detenidamente qué componentes de la aplicación pueden dividirse en microservicios. Según los requisitos de su aplicación, esto puede terminar siendo una combinación de una arquitectura orientada a servicios (SOA) con microservicios cuando sea posible. Las cargas de trabajo que son capaces de no tener estado son más capaces desplegarse como microservicios.

Resultado deseado: Las cargas de trabajo deben ser soportables, escalables y estar tan poco acopladas como sea posible.

A la hora de elegir cómo segmentar la carga de trabajo, hay que sopesar las ventajas frente a las complejidades. Lo que puede ser adecuado para un nuevo producto encaminado a su primer lanzamiento es diferente a lo que necesita una carga de trabajo creada para escalarse desde el principio. Al refactorizar un monolito existente, tendrá que considerar en qué medida soportará la aplicación una descomposición hacia la falta de estado. Dividir los servicios en partes más pequeñas permite que equipos pequeños y bien definidos los desarrollen y administren. No obstante, los servicios más pequeños pueden introducir complejidades que incluyen un aumento de la latencia, una depuración más compleja y un mayor lastre operativo.

Patrones comunes de uso no recomendados:

- El microservicio Death Star es una situación en la que los componentes atómicos son tan interdependientes que el error de uno de ellos provoca un error mucho mayor, haciendo que los componentes sean tan rígidos y frágiles como un monolito.

Beneficios de establecer esta práctica:

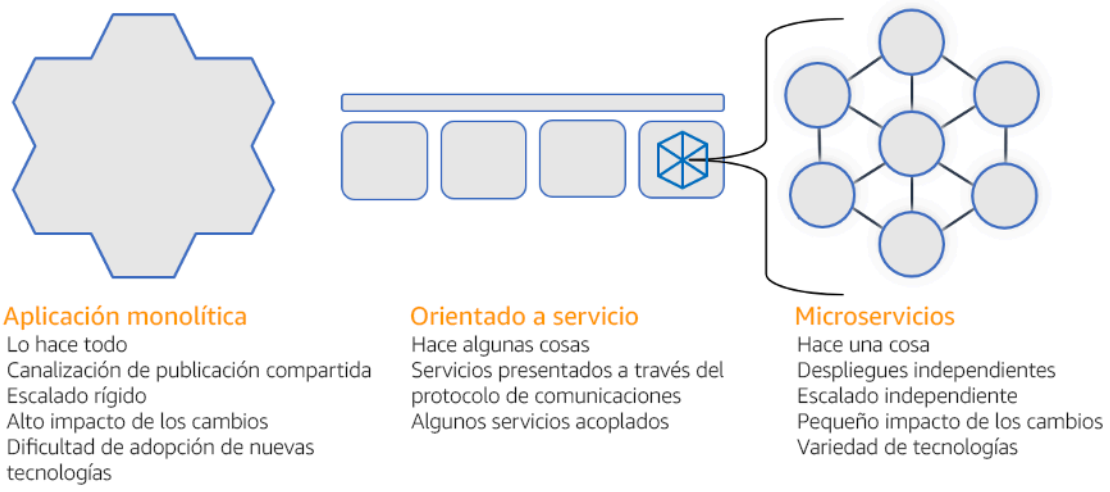
- Los segmentos más específicos conducen a una mayor agilidad, flexibilidad organizativa y escalabilidad.
- Reducción del impacto de las interrupciones del servicio.
- Los componentes de la aplicación pueden tener diferentes requisitos de disponibilidad, que pueden soportarse mediante una segmentación más atómica.
- Responsabilidades bien definidas para los equipos que apoyan la carga de trabajo.

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

Guía para la implementación

Seleccione el tipo de arquitectura en función de cómo va a segmentar su carga de trabajo. Seleccione una SOA o una arquitectura de microservicios (o, en algunos casos raros, una arquitectura monolítica). Incluso si decide empezar con una arquitectura monolítica, debe asegurarse de que sea modular y de que pueda evolucionar hacia SOA o microservicios de forma definitiva, a medida que su producto escala con la adopción por parte de los usuarios. La SOA y los microservicios ofrecen respectivamente una segmentación más pequeña, lo que resulta preferible como arquitectura moderna escalable y confiable, pero existen compensaciones a tener en cuenta, especialmente al desplegar una arquitectura de microservicios.

Una compensación principal es que se dispone de una arquitectura de informática distribuida que puede dificultar el cumplimiento de los requisitos de latencia del usuario y existe una complejidad adicional en la depuración y el rastreo de las interacciones del usuario. Puede utilizar AWS X-Ray para ayudarle a resolver este problema. Otro efecto que hay que tener en cuenta es el aumento de la complejidad operativa a medida que aumenta el número de aplicaciones que se administran, lo que requiere el despliegue de componentes con varias independencias.



Arquitecturas monolíticas, orientadas al servicio y de microservicios

Pasos para la aplicación

- Determine la arquitectura adecuada para refactorizar o desarrollar su aplicación. La SOA y los microservicios ofrecen respectivamente una segmentación más pequeña, lo que resulta preferible como arquitectura moderna escalable y confiable. La SOA puede ofrecer un término intermedio ideal para conseguir una segmentación más pequeña y, a la vez, evitar algunas de las complejidades de los microservicios. Para obtener más información, consulte Microservice Trade-Offs (Compensaciones de microservicios).
- Si su carga de trabajo lo admite y su organización puede permitírselo, debería usar una arquitectura de microservicios para conseguir la mejor agilidad y fiabilidad. Para obtener más información, consulte Implementación de microservicios en AWS
- Tenga en cuenta seguir el patrón de Strangler Fig para refactorizar un monolito en componentes más pequeños. Se trata de sustituir gradualmente componentes específicos de la aplicación por nuevas aplicaciones y servicios. AWS Migration Hub Refactor Spaces actúa como punto de partida para la refactorización incremental. Para obtener más información, consulte Seamlessly migrate

on-premises legacy workloads using a strangler pattern (Migrar sin problemas las cargas de trabajo heredadas localmente utilizando un patrón estrangulador).

- La implementación de microservicios puede requerir un mecanismo de detección de servicios para permitir que estos servicios distribuidos se comuniquen entre sí. AWS App Mesh se puede usar con arquitecturas orientadas a servicios para ofrecer una detección y un acceso confiable a los servicios. AWS Cloud Map también puede utilizarse para la detección dinámica de servicios basada en DNS.
- Si está migrando de un monolito a SOA, Amazon MQ puede ayudar a salvar la brecha como un bus de servicio cuando se rediseñan las aplicaciones heredadas en la nube.
- Para los monolitos existentes con una única base de datos compartida, elija cómo reorganizar los datos en segmentos más pequeños. Puede ser por unidad de negocio, patrón de acceso o estructura de datos. En este punto del proceso de refactorización, debe elegir entre una base de datos de tipo relacional o no relacional (NoSQL). Para obtener más información, consulte From SQL to NoSQL (De SQL a NoSQL).

Nivel de esfuerzo para el plan de implementación: Alto

Recursos

Prácticas recomendadas relacionadas:

- REL03-BP02 Desarrollar servicios centrados en funcionalidades y dominios empresariales específicos

Documentos relacionados:

- Amazon API Gateway: Configuración de una API de REST con OpenAPI
- ¿Qué es la arquitectura orientada a servicios?
- Contexto limitado (un patrón central en un diseño basado en dominios)
- Implementación de microservicios en AWS
- Microservice Trade-Offs (Compensaciones de microservicios)
- Microservicios: definición de este nuevo término de arquitectura
- Microservicios en AWS
- ¿Qué es AWS App Mesh?

Ejemplos relacionados:

- Taller de modernización de aplicaciones iterativas

Vídeos relacionados:

- Delivering Excellence with Microservices on AWS (Proporcionar excelencia con microservicios en AWS)

REL03-BP02 Desarrollar servicios centrados en funcionalidades y dominios empresariales específicos

La arquitectura orientada a servicios (SOA) define servicios con funciones bien delineadas que están determinadas por necesidades empresariales. Los microservicios utilizan modelos de dominio y contextos delimitados para trazar los límites de los servicios en los límites del contexto empresarial. Centrarse en los dominios y las funcionalidades empresariales ayuda a los equipos a definir requisitos de fiabilidad independientes para sus servicios. Los contextos delimitados aíslan y encapsulan la lógica empresarial, lo que permite a los equipos razonar mejor la forma de gestionar los errores.

Resultado deseado: los ingenieros y las partes interesadas de la empresa definen conjuntamente los contextos delimitados y los utilizan para diseñar sistemas como servicios que cumplan funciones empresariales específicas. Estos equipos utilizan prácticas establecidas, como las tormentas de eventos, para definir los requisitos. Las nuevas aplicaciones se diseñan como límites bien definidos de servicios y con acoplamiento flexible. Los monolitos existentes se descomponen en contextos delimitados y los diseños de sistemas se mueven a arquitecturas SOA o microservicios. Cuando los monolitos se refactorizan, se aplican enfoques establecidos, como contextos de burbujas y patrones de descomposición de monolitos.

Los servicios orientados al dominio se ejecutan como uno o más procesos que no comparten el estado. Responden de forma independiente a las fluctuaciones de la demanda y gestionan los escenarios de error teniendo en cuenta los requisitos específicos del dominio.

Patrones comunes de uso no recomendados:

- Se forman equipos en torno a dominios técnicos específicos, como la interfaz de usuario y la experiencia de usuario, el middleware o la base de datos, en lugar de formarse en torno a dominios empresariales específicos.

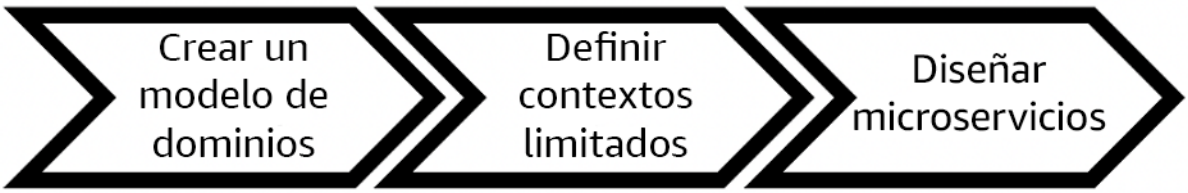
- Las aplicaciones abarcan las responsabilidades del dominio. Los servicios que abarcan contextos delimitados pueden ser más difíciles de mantener, exigen más pruebas y requieren la participación de equipos de varios dominios en las actualizaciones del software.
- Las dependencias de dominio, como las bibliotecas de entidades de dominio, se comparten entre los servicios, de modo que los cambios en un dominio de servicio requieren cambios en otros dominios de servicio.
- Los contratos de servicio y la lógica empresarial no expresan las entidades en un lenguaje de dominio común y coherente, lo que genera capas de traducción que complican los sistemas e incrementan los esfuerzos de depuración.

Beneficios de establecer esta práctica recomendada: las aplicaciones se diseñan como servicios independientes delimitados por dominios empresariales y utilizan un lenguaje empresarial común. Los servicios se pueden probar y desplegar de forma independiente. Los servicios cumplen los requisitos de resiliencia específicos del dominio implementado.

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

Guía para la implementación

El enfoque de decisiones impulsadas por dominio (DDD) es el enfoque fundamental para diseñar y crear software en torno a los dominios empresariales. Resulta útil trabajar con un marco existente a la hora de crear servicios centrados en dominios empresariales. Si trabaja con aplicaciones monolíticas existentes, puede utilizar patrones de descomposición que proporcionan técnicas establecidas para modernizar las aplicaciones y convertirlas en servicios.



Decisión impulsada por dominio

Pasos para la implementación

- Los equipos pueden realizar talleres de tormentas de eventos para identificar rápidamente eventos, comandos, agregados y dominios en un formato de notas adhesivas más ligero.

- Una vez que se hayan elaborado las entidades y funciones de dominio en el contexto de un dominio, puede dividir su dominio en servicios mediante un contexto delimitado, en el que se agrupan las entidades que comparten características y atributos similares. Si el modelo está dividido en contextos, tendrá una plantilla para limitar los microservicios.
  - Por ejemplo, las entidades del sitio web de Amazon.com podrían incluir el empaquetado, la entrega, la programación, el precio, el descuento y la divisa.
  - El empaquetado, la entrega y la programación se agrupan en el contexto del envío, mientras que el precio, el descuento y la divisa se agrupan en el contexto de los precios.
- En Decomposing monoliths into microservices (Descomposición de monolitos en microservicios), se describen patrones para refactorizar microservicios. El uso de patrones de descomposición por capacidad empresarial, subdominio o transacción se ajusta bien a los enfoques basados en dominios.
- Existen técnicas estratégicas, como el contexto de burbuja , que permiten introducir DDD en aplicaciones existentes o heredadas sin necesidad de reescrituras iniciales ni confirmaciones completas de las DDD. En un enfoque de contexto de burbujas, se establece un contexto pequeño y delimitado mediante la asignación y coordinación de servicios, o una capa anticorrupción, que protege el modelo de dominio recién definido de las influencias externas.

Una vez que los equipos hayan analizado el dominio y hayan definido las entidades y los contratos de servicio, pueden utilizar los servicios de AWS para implementar su diseño basado en dominio como servicios basados en la nube.

- Para comenzar el desarrollo, defina pruebas en las que se utilicen las reglas empresariales de su dominio. El desarrollo basado en pruebas (TDD) y el desarrollo basado en comportamiento (BDD) ayudan a los equipos a mantener los servicios centrados en resolver problemas empresariales.
- Seleccione los servicios de AWS que mejor se ajusten a los requisitos de su dominio empresarial y arquitectura de microservicios:
  - AWS sin servidor permite a su equipo centrarse en una lógica de dominio específica en lugar de administrar servidores e infraestructuras.
  - Los contenedores de AWS simplifican la administración de su infraestructura para que pueda centrarse en los requisitos de su dominio.
  - Las bases de datos personalizadas le ayudan a adaptar los requisitos de su dominio al tipo de base de datos más adecuado.
- En Building hexagonal architectures on AWS (Desarrollo de arquitecturas hexagonales en AWS), se describe un marco para integrar la lógica empresarial en los servicios que funcionan de manera

inversa desde un dominio empresarial para cumplir los requisitos funcionales y, a continuación, asociar los adaptadores de integración. Los patrones que separan los detalles de la interfaz de la lógica empresarial con los servicios de AWS ayudan a los equipos a centrarse en la funcionalidad del dominio y a mejorar la calidad del software.

Recursos

Prácticas recomendadas relacionadas:

- REL03-BP01 Elegir cómo segmentar su carga de trabajo
- REL03-BP03 Facilitar contratos de servicio por cada API

Documentos relacionados:

- AWS Microservicios
- Implementing Microservices on AWS (Implementación de microservicios en AWS)
- How to break a Monolith into Microservices (Cómo descomponer un sistema monolítico en microservicios)
- Getting Started with DDD when Surrounded by Legacy Systems (Introducción al DDD en un ambiente lleno de sistemas heredados)
- Domain-Driven Design: Tackling Complexity in the Heart of Software (Diseño basado en dominios: abordar la complejidad en el corazón del software)
- En Building hexagonal architectures on AWS (Desarrollo de arquitecturas hexagonales en AWS),
- Decomposing monoliths into microservices (Descomposición de monolitos en microservicios),
- Event Storming (Tormentas de eventos)
- Messages Between Bounded Contexts (Mensajes entre contextos delimitados)
- Microservicios
- Test-driven development (Desarrollo basado en pruebas)
- Behavior-driven development (Desarrollo basado en comportamiento)

Ejemplos relacionados:

- Enterprise Cloud Native Workshop (Taller sobre entornos nativos de la nube empresarial)

- Designing Cloud Native Microservices on AWS (from DDD/EventStormingWorkshop) (Diseño de microservicios nativos en la nube en AWS [de DDD/EventStormingWorkshop])

Herramientas relacionadas:

- Bases de datos en la Nube de AWS
- Sin servidor en AWS
- Contenedores de AWS

REL03-BP03 Facilitar contratos de servicio por cada API

Los contratos de servicio son acuerdos documentados entre los productores y los consumidores de las API que se encuentran en una definición de API legible por máquina. Una estrategia de control de versiones permite a los clientes seguir usando la API existente y migrar sus aplicaciones a la nueva API cuando estén listas. El despliegue del productor puede realizarse en cualquier momento, siempre y cuando se cumpla el contrato. Los equipos del servicio pueden usar la pila tecnológica que prefieran para cumplir el contrato de la API.

Resultado deseado:

Patrones comunes de uso no recomendados: las aplicaciones creadas con arquitecturas orientadas a servicios o de microservicios pueden funcionar de forma independiente y, al mismo tiempo, tener integrada una dependencia de la versión ejecutable. Los cambios desplegados en un consumidor o productor de API no interrumpen la estabilidad del sistema general cuando ambas partes utilizan el mismo contrato de API. Los componentes que se comunican a través de las API de servicio pueden realizar lanzamientos funcionales independientes, actualizar las dependencias de versiones ejecutables o realizar conmutaciones por error a un sitio de recuperación de desastres (DR) con poco o ningún impacto entre sí. Además, los servicios discretos pueden escalarse de forma independiente y absorber la demanda de recursos sin que sea necesario que otros servicios se escalen al unísono.

- Crear API de servicio sin esquemas estrictamente asignados. Como consecuencia, las API no se pueden usar para generar enlaces de API y las cargas útiles no se pueden validar mediante programación.
- No adoptar una estrategia de control de versiones, lo que obliga a los usuarios de la API a actualizarla y lanzarla; de lo contrario, fallará cuando los contratos de servicio evolucionen.
- Mensajes de error que filtran detalles de la implementación del servicio subyacente en lugar de describir los errores de integración en el contexto y el lenguaje del dominio.

- No utilizar contratos de API para desarrollar casos de prueba ni simulaciones de implementaciones de API para probar de forma independiente los componentes del servicio.

Beneficios de establecer esta práctica recomendada: los sistemas distribuidos que constan de componentes que se comunican a través de contratos de servicio de API pueden mejorar la fiabilidad. Los desarrolladores pueden detectar posibles problemas al principio del proceso de desarrollo mediante la comprobación de tipos durante la compilación para comprobar que las solicitudes y las respuestas cumplen el contrato de la API y que los campos obligatorios están presentes. Los contratos de la API proporcionan una interfaz clara y autodocumentada para las API y mejoran la interoperabilidad entre diferentes sistemas y lenguajes de programación.

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

Guía para la implementación

Una vez que hayan identificado los dominios empresariales y determinado la segmentación de la carga de trabajo, podrá desarrollar las API de sus servicios. Primero, defina contratos de servicio legibles por máquina para las API y, a continuación, implemente una estrategia de control de versiones de API. Cuando esté preparado para integrar servicios a través de protocolos comunes, como REST, GraphQL o eventos asíncronos, podrá incorporar servicios de AWS a su arquitectura para integrar sus componentes con contratos de API estrictamente asignados.

Servicios de AWS para contratos de API de servicios

Incorpore servicios de AWS, como Amazon API Gateway, AWS AppSyncy Amazon EventBridge, en su arquitectura para usar contratos de servicio de API en su aplicación. Amazon API Gateway le ayuda a integrarse directamente con servicios de AWS nativos y otros servicios web. API Gateway admite la especificación de OpenAPI y el control de versiones. AWS AppSync es un punto de conexión de GraphQL administrado que se configura definiendo un esquema de GraphQL para definir una interfaz de servicio para consultas, mutaciones y suscripciones. Amazon EventBridge usa esquemas de eventos para definir eventos y generar enlaces de código para sus eventos.

Pasos para la implementación

- Primero, defina un contrato para su API. En un contrato, se expresan las capacidades de una API y se definen objetos y campos de datos estrictamente asignados para la entrada y la salida de la API.
- Cuando configure las API en API Gateway, puede importar y exportar las especificaciones de OpenAPI para sus puntos de conexión.

  - La importación de una definición de OpenAPI simplifica la creación de su API y se puede integrar con la infraestructura de AWS, como las herramientas de código AWS Serverless Application Model y Kit de desarrollo de la nube de AWS (AWS CDK).
  - La exportación de una definición de API simplifica la integración con las herramientas de prueba de API y proporciona a los consumidores de servicios una especificación de la integración.
- Puede definir y administrar las API de GraphQL con AWS AppSync mediante la definición de un archivo de esquema de GraphQL para generar la interfaz del contrato y simplificar la interacción con modelos REST complejos, múltiples tablas de bases de datos o servicios heredados.
- Los proyectos de AWS Amplify que están integrados con AWS AppSync generan archivos de consulta de JavaScript estrictamente asignados para usarlos en su aplicación, así como una biblioteca de clientes de AWS AppSync GraphQL para tablas de Amazon DynamoDB .
- Cuando se consumen eventos de servicio de Amazon EventBridge, los eventos se ajustan a esquemas que ya existen en el registro de esquemas o que usted define con la especificación de OpenAPI. Si tiene un esquema definido en el registro, también puede generar enlaces de clientes desde el contrato de esquema para integrar el código con los eventos.
- Amplíe o realice un control de versiones de la API. Ampliar una API es la opción más sencilla cuando se añaden campos que se pueden configurar con campos opcionales o valores predeterminados para los campos obligatorios.
  - Los contratos basados en JSON para protocolos como REST y GraphQL pueden ser una buena opción para la ampliación del contrato.
  - Los contratos basados en XML para protocolos como SOAP deben probarse con los consumidores de servicios para determinar la viabilidad de la ampliación del contrato.
- Al realizar el control de versiones de una API, considere la posibilidad de implementar un control de versiones por proxy en el que se utilice una fachada para admitir las versiones, de modo que la lógica se pueda mantener en una única base de código.
  - Con API Gateway, puede usar mapeos de solicitudes y respuestas para simplificar la absorción de los cambios en los contratos mediante el establecimiento de una fachada que proporcione valores predeterminados para los campos nuevos o para quitar los campos eliminados de una solicitud o respuesta. Con este enfoque, el servicio subyacente puede mantener una única base de código.

Recursos

Prácticas recomendadas relacionadas:

- REL03-BP01 Elegir cómo segmentar su carga de trabajo
- REL03-BP02 Desarrollar servicios centrados en funcionalidades y dominios empresariales específicos
- REL04-BP02 Implementar dependencias con acoplamiento flexible
- REL05-BP03 Controlar y limitar las llamadas de reintento
- REL05-BP05 Definir tiempos de espera del cliente

Documentos relacionados:

- ¿Qué es una API?
- Implementing Microservices on AWS (Implementación de microservicios en AWS)
- Microservice Trade-Offs (Compensaciones de microservicios)
- Microservicios: definición de este nuevo término de arquitectura
- Microservicios en AWS
- Trabajar con extensiones de API Gateway para OpenAPI
- OpenAPI-Specification (Especificación de OpenAPI)
- GraphQL: Schemas and Types (GraphQL: esquemas y tipos)
- Amazon EventBridge code bindings (Enlaces de código de EventBridge)

Ejemplos relacionados:

- Amazon API Gateway: Configuración de una API de REST con OpenAPI
- Amazon API Gateway to Amazon DynamoDB CRUD application using OpenAPI (Aplicación CRUD de Amazon API Gateway en Amazon DynamoDB mediante OpenAPI)
- Modern application integration patterns in a serverless age: API Gateway Service Integration (Patrones de integración de aplicaciones modernos en una era sin servidores: integración de servicios de API Gateway)
- Implementing header-based API Gateway versioning with Amazon CloudFront (Implementación del control de versiones de API Gateway basado en encabezados con Amazon CloudFront)
- AWS AppSync: Building a client application (Creación de una aplicación cliente)

Vídeos relacionados:

- Using OpenAPI in AWS SAM to manage API Gateway (Uso de OpenAPI en AWS SAM para administrar API Gateway)

Herramientas relacionadas:

- Amazon API Gateway
- AWS AppSync
- Amazon EventBridge,

FIABILIDAD 4. ¿Cómo diseña las interacciones en un sistema distribuido para evitar los errores?

Los sistemas distribuidos dependen de las redes de comunicaciones para interconectar componentes, como servidores o servicios. Su carga de trabajo debe funcionar de manera fiable aunque se pierdan datos o haya latencia en estas redes. Los componentes del sistema distribuido deben funcionar de forma que no repercutan negativamente en otros componentes ni en la carga de trabajo. Estas prácticas recomendadas evitan que se produzcan errores y mejoran el tiempo medio entre errores (MTBF).

Prácticas recomendadas

- REL04-BP01 Identificar qué tipo de sistema distribuido se necesita
- REL04-BP02 Implementar dependencias con acoplamiento flexible
- REL04-BP03 Realizar un trabajo constante
- REL04-BP04 Hacer que todas las respuestas sean idempotentes

REL04-BP01 Identificar qué tipo de sistema distribuido se necesita

Los sistemas distribuidos en tiempo real estrictos requieren que las respuestas se proporcionen de forma sincrónica y rápidamente, mientras que los sistemas en tiempo real laxos cuentan con un plazo de tiempo más generoso, que se mide en minutos o más, para proporcionar una respuesta. Los sistemas sin conexión gestionan las respuestas a través del procesamiento por lotes o asincrónico. Los sistemas distribuidos en tiempo real estrictos tienen los requisitos de fiabilidad más exigentes.

Los desafíos más complicados relacionados con los sistemas distribuidos giran en torno a los sistemas distribuidos en tiempo real, conocidos también como servicios de solicitud/respuesta. Lo

que hace que sean tan difíciles es que las solicitudes llegan de forma impredecible y las respuestas deben emitirse rápidamente (por ejemplo, si el cliente está esperando una respuesta de forma activa). Entre algunos ejemplos se incluyen los servidores web de front-end, la canalización de pedidos, las transacciones con tarjetas de crédito, todas las API de AWS y la telefonía.

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

Guía para la implementación

- Identifique qué tipo de sistema distribuido se necesita. Entre los problemas de los sistemas distribuidos se incluyen la latencia, el escalado, conocer las API de red, la serialización y deserialización de los datos, y la complejidad de algoritmos como Paxos. A medida que los sistemas crecen y se hacen más distribuidos, los casos extremos teóricos se convierten en sucesos habituales.
  - La Amazon Builders' Library: Desafíos de los sistemas distribuidos
    - Los sistemas distribuidos en tiempo real estrictos requieren que las respuestas se proporcionen de forma sincrónica y rápidamente.
    - Los sistemas en tiempo real laxos cuentan con un plazo más generoso, que se mide en minutos o más, para proporcionar una respuesta.
    - Los sistemas sin conexión gestionan las respuestas a través del procesamiento por lotes o asincrónico.
    - Los sistemas distribuidos en tiempo real estrictos tienen los requisitos de fiabilidad más exigentes.

Recursos

Documentos relacionados:

- Amazon EC2: garantizar la idempotencia
- La Amazon Builders' Library: Desafíos de los sistemas distribuidos
- La Amazon Builders' Library: Fiabilidad, trabajo constante y una buena taza de café
- ¿Qué es Amazon EventBridge?
- ¿Qué es Amazon Simple Queue Service?

Videos relacionados:

- Cumbre de AWS en Nueva York 2019: Introducción a las arquitecturas basadas en eventos y Amazon EventBridge (MAD205)
- AWS re:Invent 2018: Bucles cerrados y mentes abiertas: cómo asumir el control de los sistemas grandes y pequeños ARC337 (incluye acoplamiento flexible, trabajo constante y estabilidad estática)
- AWS re:Invent 2019: Optar por arquitecturas basadas en eventos (SVS308)

REL04-BP02 Implementar dependencias con acoplamiento flexible

Las dependencias, como los sistemas de colas, los sistemas de transmisión, los flujos de trabajo y los equilibradores de carga, tienen un acoplamiento flexible. El acoplamiento flexible ayuda a aislar el comportamiento de un componente de otros componentes que dependen de él, lo que aumenta la resiliencia y la agilidad.

En sistemas de acoplamiento ajustado, los cambios en un componente pueden requerir cambios en otros componentes que dependen de él, lo que reduce el rendimiento de todos los componentes. El acoplamiento flexible elimina esta dependencia, de forma que los componentes dependientes solo necesitan conocer la interfaz publicada y versionada. La implementación de un acoplamiento flexible entre las dependencias aísla un fallo en una de ellas para que no afecte a otra.

El acoplamiento flexible permite modificar el código o añadir características a un componente y, al mismo tiempo, minimizar el riesgo para otros componentes que dependen de él. También permite una resiliencia granular a nivel de componente, lo que permite escalar horizontalmente o incluso cambiar la implementación subyacente de la dependencia.

Para mejorar aún más la resiliencia mediante el acoplamiento flexible, haga que las interacciones entre componentes sean asincrónicas siempre que sea posible. Este modelo es adecuado para cualquier interacción que no necesite una respuesta inmediata y en la que baste con el reconocimiento de que una solicitud se ha registrado. Consta de un componente que genera eventos y de otro que los consume. Ambos componentes no se integran mediante una interacción directa de punto a punto, sino que normalmente emplean una capa de almacenamiento duradera intermedia, como una cola Amazon SQS o una plataforma de restringa de datos como Amazon Kinesis o AWS Step Functions.

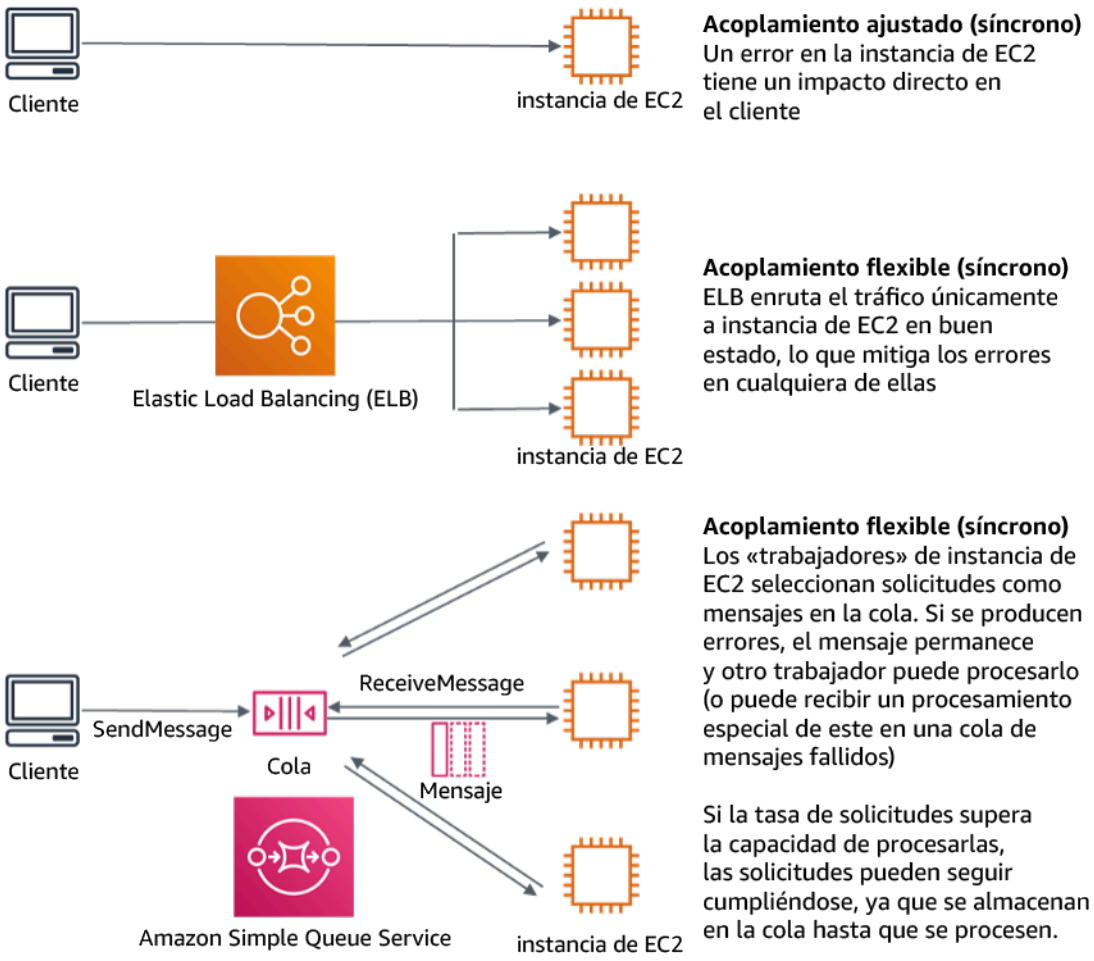


Figura 4: Las dependencias, como sistemas de colas y equilibradores de carga, tienen acoplamiento flexible

Las colas de Amazon SQS y los equilibradores de carga elásticos son solo dos formas de añadir una capa intermedia para el acoplamiento flexible. Las arquitecturas basadas en eventos también pueden integrarse en Nube de AWS utilizando Amazon EventBridge, lo que puede abstraer a los clientes (productores de eventos) de los servicios en los que confían (consumidores de eventos). Amazon Simple Notification Service (Amazon SNS) es una solución eficaz cuando se necesita una mensajería de alto rendimiento, basada en push y de varios a varios. Utilizando temas de Amazon SNS, los sistemas de su editor pueden repartir mensajes por una gran cantidad de puntos de conexión de suscriptores para procesarlos en paralelo.

Aunque las colas ofrecen varias ventajas, en la mayoría de sistemas inflexibles en tiempo real, las solicitudes que superan un umbral temporal (que suele ser de segundos) se consideran obsoletas (el cliente ha desistido y ya no espera una respuesta), por lo que no se procesan. De esta manera, se pueden procesar las solicitudes más recientes (y probablemente aún válidas) en su lugar.

Resultado deseado: la implementación de dependencias de acoplamiento flexible permite minimizar el área de superficie en caso de fallo a nivel de componente, lo que ayuda a diagnosticar y resolver problemas. También simplifica los ciclos de desarrollo, lo que permite a los equipos implementar cambios a nivel modular sin que eso afecte al rendimiento de otros componentes que dependen de él. Este enfoque ofrece la capacidad de escalar horizontalmente a nivel de componente en función de los recursos que sean necesarios, así como de utilizar un componente que contribuye a ahorrar costes.

Antipatrones usuales:

- Desplegar una carga de trabajo monolítica.
- Invocar directamente las API entre capas de la carga de trabajo sin la capacidad de conmutar por error ni procesar asincrónicamente la solicitud
- Utilizar un acoplamiento ajustado con datos compartidos. Los sistemas de acoplamiento flexible no deben compartir datos a través de bases de datos compartidas u otras formas de almacenamiento de datos de acoplamiento ajustado, que pueden reintroducir el acoplamiento ajustado y dificultar la escalabilidad.
- Ignorar la contrapresión. La carga de trabajo debe tener la capacidad de ralentizar o detener los datos entrantes cuando un componente no pueda procesarlos al mismo ritmo.

Ventajas de establecer esta práctica recomendada: el acoplamiento flexible ayuda a aislar el comportamiento de un componente de otros componentes que dependen de él, lo que aumenta la resiliencia y la agilidad. Un error en un componente está aislado de los demás componentes.

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

Guía para la implementación

Implemente dependencias con acoplamiento flexible. Existen varias soluciones que permiten crear aplicaciones con un acoplamiento flexible. Entre ellas, se incluyen servicios para implementar colas totalmente administradas, flujos de trabajo automatizados, reacción a eventos y API, entre otras, que pueden ayudar a aislar el comportamiento de los componentes de otros componentes y, por lo tanto, aumentar la resiliencia y la agilidad.

- Cree arquitecturas basadas en eventos: Amazon EventBridge le ayuda a crear arquitecturas basadas en eventos distribuidas y de acoplamiento flexible.
- Implemente colas en sistemas distribuidos: puede usar Amazon Simple Queue Service (Amazon SQS) para integrar y desacoplar sistemas distribuidos.

- Contenedorice los componentes como microservicios: los microservicios permiten a los equipos crear aplicaciones compuestas por pequeños componentes independientes que se comunican a través de API bien definidas. Amazon Elastic Container Service(Amazon ECS) y Amazon Elastic Kubernetes Service (Amazon EKS) pueden ayudarle a empezar a utilizar los contenedores con mayor rapidez.
- Administre los flujos de trabajo con Step Functions: Step Functions le ayudan a coordinar varios servicios de AWS para convertirlos en flujos de trabajo flexibles.
- Utilice las arquitecturas de mensajería de publicación y suscripción (pub/sub): Amazon Simple Notification Service (Amazon SNS) permite entregar mensajes de los editores a los suscriptores (también conocidos como productores y consumidores).

Pasos para la implementación

- Los componentes de una arquitectura basada en eventos se inician mediante eventos. Los eventos son acciones que ocurren en un sistema, como cuando un usuario añade un artículo a una cesta. Cuando una acción se realiza correctamente, se genera un evento que activa el siguiente componente del sistema.
  - «Building Event-driven Applications with Amazon EventBridge»
  - AWS «re:Invent 2022 - Designing Event-Driven Integrations using Amazon EventBridge»
- Los sistemas de mensajería distribuida tienen tres partes principales que deben implementarse para una arquitectura basada en colas. Incluyen los componentes del sistema distribuido, la cola que se usa para el desacoplamiento (distribuida en servidores de Amazon SQS) y los mensajes de la cola. Un sistema típico tiene productores que inician el mensaje en la cola y el consumidor que recibe el mensaje de la cola. La cola almacena los mensajes en varios servidores de Amazon SQS para garantizar la redundancia.
  - «Basic Amazon SQS architecture»
  - «Send Messages Between Distributed Applications with Amazon Simple Queue Service»
- Los microservicios, cuando se utilizan bien, facilitan el mantenimiento y aumentan la escalabilidad, ya que los componentes de acoplamiento flexible los administran equipos independientes. También permiten aislar los comportamientos en un solo componente en caso de que se realicen cambios.
  - «Implementing Microservices on AWS»
  - «Let's Architect! Architecting microservices with containers»

- Con AWS Step Functions, puede crear aplicaciones distribuidas, automatizar procesos y orquestar microservicios, entre otras cosas. La orquestación de varios componentes en un flujo de trabajo automatizado le permite desacoplar las dependencias de su aplicación.
  - «Create a Serverless Workflow with AWS Step Functions and AWS Lambda»
  - «Introducción a AWS Step Functions»

Recursos

Documentos relacionados:

- «Amazon EC2: Ensuring Idempotency»
- La Amazon Builders' Library: Desafíos de los sistemas distribuidos
- La Amazon Builders' Library: Fiabilidad, trabajo constante y una buena taza de café
- ¿Qué es Amazon EventBridge?
- «¿Qué es Amazon Simple Queue Service?»
- «Break up with your monolith»
- «Orchestrate Queue-based Microservices with AWS Step Functions and Amazon SQS»
- «Basic Amazon SQS architecture»
- «Queue-Based Architecture»

Vídeos relacionados:

- «AWS New York Summit 2019: Intro to Event-driven Architectures and Amazon EventBridge (MAD205)»
- «AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small ARC337» (incluye acoplamiento flexible, trabajo constante y estabilidad estática)
- «AWS re:Invent 2019: Moving to event-driven architectures (SVS308)»
- «AWS re:Invent 2019: Scalable serverless event-driven applications using Amazon SQS and Lambda (API304)»
- «AWS re:Invent 2019: Scalable serverless event-driven applications using Amazon SQS and Lambda«
- «AWS re:Invent 2022 - Designing event-driven integrations using Amazon EventBridge«
- «AWS re:Invent 2017: Elastic Load Balancing Deep Dive and Best Practices»

REL04-BP03 Realizar un trabajo constante

Los sistemas pueden producir error cuando hay cambios rápidos grandes en la carga. Por ejemplo, si la carga de trabajo está realizando una comprobación de estado que supervisa el estado de miles de servidores, debería enviar siempre una carga del mismo tamaño (una instantánea completa del estado actual). Si no hay errores en ningún servidor, o hay errores en todos ellos, el sistema de comprobación de estado estará haciendo un trabajo constante sin rápidos cambios de gran tamaño.

Por ejemplo, si el sistema de comprobación de estado supervisa 100 000 servidores, la carga contenida en él es nominal con un porcentaje de errores del servidor normalmente bajo. Sin embargo, si un evento importante deja a la mitad de esos servidores en mal estado, el sistema de comprobación de estado se sobrecargaría intentando actualizar los sistemas de notificación y comunicar el estado a sus clientes. Por ello, el sistema de comprobación de estado debería enviar cada vez la instantánea completa del estado actual. 100 000 estados de servidor, cada uno representado por un bit, solo constituiría una carga de 12,5 KB. Si no hay errores en ningún servidor, o hay errores en todos ellos, el sistema de comprobación de estado estará haciendo un trabajo constante y los cambios rápidos de gran tamaño no pondrán en peligro la estabilidad del sistema. En realidad, así es como Amazon Route 53 gestiona las comprobaciones de estado de los puntos de conexión (como las direcciones IP) para determinar cómo se enruta a los usuarios finales.

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

Guía para la implementación

- Realice un trabajo para que los sistemas no tengan errores cuando hay cambios grandes y rápidos en la carga.
- Implemente dependencias con acoplamiento flexible. Las dependencias, como los sistemas de colas, los sistemas de transmisión, los flujos de trabajo y los equilibradores de carga, tienen un acoplamiento flexible. El acoplamiento flexible ayuda a aislar el comportamiento de un componente de otros componentes que dependen de él, lo que aumenta la resiliencia y la agilidad.
  - La Amazon Builders' Library: Fiabilidad, trabajo constante y una buena taza de café
  - AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small ARC337 (includes constant work) (Cerrar los bucles y abrir las mentes: cómo tomar el control de los sistemas, grandes y pequeños ARC337 [incluye trabajo constante])
    - En el ejemplo de una sistema de comprobación de estado que supervisa 100 000 servidores, diseñe las cargas de trabajo de forma que los tamaños de la carga útil sean iguales independientemente del número de éxitos o fracasos.

Recursos

Documentos relacionados:

- Amazon EC2: garantizar la idempotencia
- La Amazon Builders' Library: Desafíos de los sistemas distribuidos
- La Amazon Builders' Library: Fiabilidad, trabajo constante y una buena taza de café

Vídeos relacionados:

- AWS New York Summit 2019: Intro to Event-driven Architectures and Amazon EventBridge (Introducción a las arquitecturas basadas en eventos y Amazon EventBridge) (MAD205)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small ARC337 (includes constant work) (Cerrar los bucles y abrir las mentes: cómo asumir el control de los sistemas grandes y pequeños ARC337 [incluye trabajo constante])
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small ARC337 (includes loose coupling, constant work, static stability) (Cerrar los bucles y abrir las mentes: cómo asumir el control de los sistemas grandes y pequeños ARC337 [incluye acoplamiento flexible, trabajo constante y estabilidad estática])
- AWS re:Invent 2019: Moving to event-driven architectures (Optar por arquitecturas basadas en eventos) (SVS308)

REL04-BP04 Hacer que todas las respuestas sean idempotentes

Un servicio idempotente promete que cada solicitud se completará una y solo una vez, de tal forma que realizar varias solicitudes idénticas tiene el mismo efecto que realizar una sola solicitud. Un servicio idempotente permite que un cliente implemente fácilmente los reintentos sin el temor de que una solicitud se procese erróneamente varias veces. Para ello, los clientes pueden usar solicitudes de API con un token de idempotencia: se utiliza el mismo token siempre que se repite la solicitud. Una API de servicio idempotente usa el token para devolver una respuesta idéntica a la que se devolvió por primera vez cuando se completó la solicitud.

En un sistema distribuido, es fácil llevar a cabo una acción una vez como máximo (el cliente realiza solo una solicitud) o al menos una vez (sigue realizando la solicitud hasta que el cliente obtiene una confirmación del éxito). Sin embargo, es difícil garantizar que una acción es idempotente, lo que significa que se lleva a cabo exactamente una vez, de modo que realizar varias solicitudes idénticas

tiene el mismo efecto que realizar una sola solicitud. Al usar tokens de idempotencia en las API, los servicios pueden recibir una solicitud de migración una o más veces sin crear registros duplicados ni efectos secundarios.

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

Guía para la implementación

- Haga que todas las respuestas sean idempotentes. Un servicio idempotente promete que cada solicitud se completará una y solo una vez, de tal forma que realizar varias solicitudes idénticas tiene el mismo efecto que realizar una sola solicitud.
  - Los clientes pueden usar solicitudes de API con un token de idempotencia: se utiliza el mismo token siempre que se repite la solicitud. Una API de servicio idempotente usa el token para devolver una respuesta idéntica a la que se devolvió por primera vez cuando se completó la solicitud.
    - Amazon EC2: garantizar la idempotencia

Recursos

Documentos relacionados:

- Amazon EC2: garantizar la idempotencia
- La Amazon Builders' Library: Desafíos de los sistemas distribuidos
- La Amazon Builders' Library: Fiabilidad, trabajo constante y una buena taza de café

Videos relacionados:

- Cumbre de AWS en Nueva York 2019: Introducción a las arquitecturas basadas en eventos y Amazon EventBridge (MAD205)
- AWS re:Invent 2018: Bucles cerrados y mentes abiertas: cómo asumir el control de los sistemas grandes y pequeños ARC337 (incluye acoplamiento flexible, trabajo constante y estabilidad estática)
- AWS re:Invent 2019: Optar por arquitecturas basadas en eventos (SVS308)

FIABILIDAD 5. ¿Cómo diseña las interacciones en un sistema distribuido para mitigar o tolerar los errores?

Los sistemas distribuidos dependen de las redes de comunicaciones para interconectar componentes, como servidores o servicios. Su carga de trabajo debe funcionar de manera fiable aunque se pierdan datos o haya latencia en estas redes. Los componentes del sistema distribuido deben funcionar de forma que no repercutan negativamente en otros componentes ni en la carga de trabajo. Estas prácticas recomendadas permiten que las cargas de trabajo toleren el estrés o los errores, se recuperen más rápidamente de ellos y mitiguen el impacto de dichos errores. El resultado es un tiempo medio de recuperación (MTTR) mejor.

Prácticas recomendadas

- REL05-BP01 Implementar una degradación estable para transformar las dependencias estrictas en flexibles
- REL05-BP02 Limitar las solicitudes
- REL05-BP03 Controlar y limitar las llamadas de reintento
- REL05-BP04 Responder rápido a los errores y limitar las colas
- REL05-BP05 Definir tiempos de espera del cliente
- REL05-BP06 Crear servicios sin estado cuando sea posible
- REL05-BP07 Implementar recursos de emergencia

REL05-BP01 Implementar una degradación estable para transformar las dependencias estrictas en flexibles

Los componentes de la aplicación deben seguir desempeñando su función principal incluso si las dependencias dejan de estar disponibles. Es posible que proporcionen datos ligeramente obsoletos, datos alternativos o incluso ningún dato. Esto garantiza que los errores localizados solo impidan lo mínimo del funcionamiento general del sistema y, al mismo tiempo, se obtenga el valor empresarial central.

Resultado deseado: Cuando las dependencias de un componente no están en buen estado, el propio componente puede seguir funcionando, aunque con la capacidad mermada. Los modos de errores de los componentes deben considerarse parte del funcionamiento normal. Los flujos de trabajo deben diseñarse de tal manera que dichos errores no produzcan un fallo total o, al menos, lleven a estados predecibles y recuperables.

Patrones comunes de uso no recomendados:

- No identificar la funcionalidad empresarial principal necesaria. No probar que los componentes funcionen, incluso durante los errores de dependencia.
- No proporcionar datos en caso de error o cuando solo una de las múltiples dependencias no está disponible y aún se pueden devolver resultados parciales.
- Crear un estado incoherente cuando una transacción falla parcialmente.
- No tener una forma alternativa de acceder a un almacén de parámetros central.
- Invalidar o vaciar un estado local como resultado de un fallo de actualización sin tener en cuenta las consecuencias.

Beneficios de establecer esta práctica recomendada: la degradación gradual mejora la disponibilidad del sistema en su conjunto y mantiene la funcionalidad de las funciones más importantes incluso cuando hay errores.

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

Guía para la implementación

La implementación de una degradación gradual ayuda a minimizar el impacto de los errores de dependencia en la función de los componentes. Lo ideal sería que un componente detectara los errores de dependencia y siguiese funcionando de una forma que afectara lo mínimo a otros componentes o clientes.

Diseñar una arquitectura que permita una degradación gradual implica considerar los posibles modos de errores durante el diseño de las dependencias. Para cada modo de error, disponga de una forma de ofrecer la mayoría o, al menos la funcionalidad más crítica del componente, a las personas que llaman o a los clientes. Estos factores pueden convertirse en requisitos adicionales que se pueden probar y verificar. Lo ideal es que un componente pueda realizar su función principal de manera aceptable incluso cuando falla una o varias dependencias.

Se trata tanto de un tema empresarial como técnico. Todos los requisitos empresariales son importantes y deben cumplirse si es posible. Sin embargo, es lógico preguntarse qué debe suceder cuando no se puedan cumplir todos. Se puede diseñar un sistema para que esté disponible y sea coherente, pero en circunstancias en las que haya que eliminar un requisito, ¿cuál es más importante? En el caso del procesamiento de pagos, puede ser la coherencia. En una aplicación en tiempo real, puede ser la disponibilidad. En el caso de un sitio web orientado al cliente, la respuesta dependería de las expectativas del cliente.

Lo que esto significa depende de los requisitos del componente y de lo que deba considerarse su función principal. Por ejemplo:

- Un sitio web de comercio electrónico podría mostrar en su página de inicio los datos de varios sistemas diferentes, como las recomendaciones personalizadas, los productos mejor clasificados y el estado de los pedidos de los clientes. Cuando un sistema anterior falla, sigue siendo lógico mostrar todo lo demás en lugar de mostrar una página de error al cliente.
- Un componente que realiza escrituras por lotes puede seguir procesando un lote si se produce un error en una de las operaciones individuales. Implementar un mecanismo de reintento debería ser sencillo. Se puede hacer devolviendo a la persona que llama información sobre qué operaciones se han realizado correctamente, cuáles han fallado y por qué han fallado, o colocando las solicitudes que han fallado en una cola de mensajes fallidos para implementar reintentos asíncronos. También se debe registrar la información sobre las operaciones que han fallado.
- Un sistema que procese las transacciones debe verificar que se ejecuten todas o ninguna de las actualizaciones individuales. En el caso de las transacciones distribuidas, se puede usar el patrón Saga para revertir operaciones anteriores en caso de que falle una operación posterior de la misma transacción. En este caso, la función principal es mantener la coherencia.
- Los sistemas en los que el tiempo es crítico deberían poder gestionar de la manera oportuna las dependencias que no respondan. En estos casos, se puede utilizar el patrón del disyuntor. Cuando se agota el tiempo de espera de las respuestas de una dependencia, el sistema puede cambiar a un estado cerrado en el que no se realizan llamadas adicionales.
- Una aplicación puede leer parámetros de un almacén de parámetros. Puede resultar útil crear imágenes de contenedores con un conjunto predeterminado de parámetros y utilizarlos en caso de que ese almacén de parámetros no esté disponible.

Tenga en cuenta que las soluciones que se adopten en caso de fallo de un componente deben probarse y deben ser significativamente más sencillas que la solución principal. En general, deben evitarse estrategias alternativas.

Pasos para la implementación

Identifique las dependencias externas e internas. Considere qué tipos de errores pueden producirse en ellas. Piense en formas de minimizar el impacto negativo en los sistemas anteriores y posteriores y en los clientes durante esos errores.

A continuación, tenemos una lista de dependencias y cómo degradar correctamente cuando fallan:

1. Fallo parcial de las dependencias: un componente puede realizar varias solicitudes a los sistemas posteriores, ya sean varias solicitudes a un sistema o una sola solicitud destinada a varios sistemas. Dependiendo del contexto empresarial, es posible que haya diferentes formas apropiadas de gestionar este problema (para obtener más información, consulte los ejemplos anteriores en la Guía de implementación).
2. Un sistema posterior no puede procesar las solicitudes debido a la alta carga: si las solicitudes a un sistema posterior fallan constantemente, no tiene sentido seguir intentándolo. Esto puede suponer una carga adicional para un sistema ya sobrecargado y dificultar la recuperación. Aquí se puede utilizar el patrón de disyuntor, que monitoriza las llamadas que han fallado al enviarlas a un sistema posterior. Si falla un gran número de llamadas, dejará de enviar más solicitudes al sistema posterior y solo permitirá ocasionalmente el paso de las llamadas para comprobar si el sistema posterior vuelve a estar disponible.
3. El almacén de parámetros no está disponible: para transformar un almacén de parámetros, se puede utilizar el almacenamiento en caché de dependencia flexible o los valores predeterminados en buen estado que se incluyen en las imágenes de contenedores o máquinas. Tenga en cuenta que estos valores predeterminados deben mantenerse actualizados e incluirse en los conjuntos de pruebas.
4. No hay disponible un servicio de monitorización u otra dependencia no funcional: si un componente no puede enviar registros, métricas o rastros de forma intermitente a un servicio de monitorización central, suele ser mejor seguir ejecutando las funciones empresariales como de costumbre. No registrar ni subir métricas de forma silenciosa durante mucho tiempo no suele ser aceptable. Además, algunos casos de uso pueden requerir entradas de auditoría completas para satisfacer los requisitos de cumplimiento.
5. Es posible que una instancia principal de una base de datos relacional no esté disponible: Amazon Relational Database Service, como casi todas las bases de datos relacionales, solo puede tener una instancia de escritor principal. Esto crea un único punto de error para las cargas de trabajo de escritura y dificulta el escalamiento. Este problema se puede mitigar parcialmente mediante el uso de una configuración multi-AZ para alta disponibilidad o Amazon Aurora sin servidor para mejorar el escalamiento. Cuando los requisitos de disponibilidad son muy altos, podría ser conveniente no utiliza en absoluto el escritor principal. Para consultas que solo leen, se pueden utilizar réplicas de lectura, que proporcionan redundancia y capacidad de escalamiento horizontal, no solo vertical. Las escrituras se pueden almacenar en búfer, por ejemplo, en una cola de Amazon Simple Queue Service, de modo que las solicitudes de escritura de los clientes puedan seguir aceptándose incluso si la principal no está disponible temporalmente.

Recursos

Documentos relacionados:

- Amazon API Gateway: Limitar las solicitudes de la API para mejorar el rendimiento
- CircuitBreaker (resumen del patrón de interruptor del libro «Release It!»)
- Error Retries and Exponential Backoff in AWS (Reintentos en caso de error y retroceso exponencial en AWS)
- Michael Nygard «Release It! Design and Deploy Production-Ready Software»
- La Amazon Builders' Library: Evitar el retroceso en sistemas distribuidos
- La Amazon Builders' Library: Evitar trabajos pendientes en colas insalvables
- La Amazon Builders' Library: Desafíos y estrategias del almacenamiento en caché
- La Amazon Builders' Library: Tiempos de espera, reintentos y retroceso con alteración

Vídeos relacionados:

- Retry, backoff, and jitter: AWS re:Invent 2019: Introducing The Amazon Builders’ Library (DOP328) (Reintento, retroceso y fluctuación: AWS re:Invent 2019: Presentación de Amazon Builders’ Library [DOP328])

Ejemplos relacionados:

- Laboratorio de Well-Architected: Nivel 300: Implementación de comprobaciones de estado y administración de dependencias para mejorar la fiabilidad

REL05-BP02 Limitar las solicitudes

Limite las solicitudes para mitigar el agotamiento de los recursos debido a aumentos inesperados de la demanda. Las solicitudes por debajo de los índices de limitación se procesan, pero las que superan el límite definido se rechazan y se envía un mensaje que indica que la solicitud no se ha procesado a causa de la limitación.

Resultado deseado: la limitación de las solicitudes mitiga los grandes picos de volumen, ya sea debido a un aumento repentino del tráfico de clientes, a ataques por desbordamiento o a tormentas de reintentos, lo que permite que las cargas de trabajo sigan procesando de manera normal el volumen de solicitudes admitido.

Patrones comunes de uso no recomendados:

- Las limitaciones de puntos de conexión de la API no se implementan o se mantienen en los valores predeterminados sin tener en cuenta los volúmenes esperados.
- Los puntos de conexión de la API no se someten a pruebas de carga ni se prueban las limitaciones.
- Los índices de solicitudes se reducen sin tener en cuenta el tamaño o la complejidad de las solicitudes.
- Los índices o el tamaño máximos de las solicitudes se prueban, pero por separado.
- Los recursos no se aprovisionan con los mismos límites establecidos en las pruebas.
- No se han configurado ni considerado planes de uso para los consumidores de API de aplicación a aplicación (A2A).
- Los consumidores de cola que escalan horizontalmente no tienen configurado un valor máximo de simultaneidad.
- No se ha implementado la limitación de índices por dirección IP.

Beneficios de establecer esta práctica recomendada: las cargas de trabajo que establecen límites pueden funcionar con normalidad y procesar correctamente la carga de solicitudes aceptada en caso de que se produzcan picos de volumen inesperados. Los picos repentinos o sostenidos de solicitudes a las API y las colas se limitan y no agotan los recursos de procesamiento de solicitudes. Hay límites de índices que limitan a solicitantes individuales para que un gran volumen de tráfico desde una sola dirección IP o un único consumidor de API no agote los recursos y afecte a otros consumidores.

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

Guía para la implementación

Los servicios deben diseñarse para procesar una capacidad de solicitudes conocida; esta capacidad se puede establecer mediante pruebas de carga. Si los índices de llegada de solicitudes superan los límites, se emite la respuesta correspondiente que indica que la solicitud no se ha procesado a causa de las limitaciones. Esto permite al consumidor gestionar el error y volver a intentarlo más tarde.

Cuando su servicio requiera la implementación de limitaciones, considere la posibilidad de implementar el algoritmo del bucket de tokens, en el que un token se refiere a una solicitud. Los tokens se recargan a un índice de limitación por segundo y se vacían de forma asíncrona a un ritmo de un token por solicitud.

El algoritmo del bucket de tokens.

Amazon API Gateway implementa el algoritmo del bucket de tokens de acuerdo con los límites de la cuenta y la región, y se puede configurar por cliente con planes de uso. Además, Amazon Simple Queue Service (Amazon SQS) y Amazon Kinesis pueden almacenar en búfer las solicitudes para aliviar el índice de solicitudes y permitir índices de limitaciones más altos para las solicitudes que se pueden atender. Por último, puede implementar la limitación de índices con AWS WAF para limitar a consumidores de API específicos que generan una carga inusualmente alta.

Pasos para la implementación

Puede configurar API Gateway con limitaciones para sus API y devolver errores `429 Demasiadas solicitudes` cuando se superan los límites. Puede utilizar AWS WAF con sus puntos de conexión AWS AppSync y API Gateway para habilitar la limitación de índices por dirección IP. Además, si su sistema tolera el procesamiento asíncrono, puede colocar los mensajes en una cola o secuencia para acelerar las respuestas a los clientes del servicio, lo que le permite ampliar los índices de limitación más altos.

Con el procesamiento asíncrono, cuando se haya configurado Amazon SQS como origen de eventos para AWS Lambda, puede configurar la simultaneidad máxima para evitar que los altos índices de eventos consuman la cuota de ejecución simultánea de la cuenta disponible que necesitan otros servicios de su carga de trabajo o cuenta.

Si bien API Gateway proporciona una implementación administrada del bucket de tokens, en los casos en que no pueda usar API Gateway, puede utilizar las implementaciones de código abierto

específicas de cada lenguaje (consulte los ejemplos relacionados en Recursos) del bucket de tokens para sus servicios.

- Comprenda y configure limitaciones de API Gateway a nivel de cuenta por región, API por etapa y clave de API por niveles de planes de uso.
- Utilice reglas de limitación de índices de AWS WAF para puntos de conexión de API Gateway y AWS AppSync para protegerse contra ataques de desbordamiento y bloquear las IP malintencionadas. Las reglas de limitación de índices también se pueden configurar en las claves de API de AWS AppSync para los consumidores de A2A.
- Analice si necesita un control de limitación superior a la limitación de índices para las API de AWS AppSync y, de ser así, configure una API Gateway enfrente de su punto de conexión de AWS AppSync.
- Si las colas de Amazon SQS están configuradas como disparadores para los consumidores de colas de Lambda, defina la simultaneidad máxima en un valor que procese lo suficiente para cumplir los objetivos de nivel de servicio, pero que no consuma límites de simultaneidad que afecten a otras funciones Lambda. Considere la posibilidad de configurar la simultaneidad reservada en otras funciones Lambda de la misma cuenta y región cuando consuma colas con Lambda.
- Utilice API Gateway con integraciones de servicios nativos para Amazon SQS o Kinesis para almacenar en búfer las solicitudes.
- Si no puede utilizar API Gateway, consulte las bibliotecas específicas del lenguaje para implementar el algoritmo del bucket de tokens para su carga de trabajo. Consulte la sección de ejemplos e investigue por su cuenta para encontrar una biblioteca adecuada.
- Pruebe los límites que tiene pensado establecer o que va a permitir que se aumenten, y documente los límites probados.
- No aumente los límites por encima de lo que establezca en las pruebas. Cuando aumente un límite, antes de aplicar ese aumento, compruebe que los recursos aprovisionados son equivalentes o superiores a los de los escenarios de prueba.

Recursos

Prácticas recomendadas relacionadas:

- REL04-BP03 Realizar un trabajo constante
- REL05-BP03 Controlar y limitar las llamadas de reintento

Documentos relacionados:

- Amazon API Gateway: Limitar las solicitudes de la API para mejorar el rendimiento
- AWS WAF: Rate-based rule statement (Declaración de la regla basada en índices)
- Introducing maximum concurrency of AWS Lambda when using Amazon SQS as an event source (Introducción a la simultaneidad máxima de funciones AWS Lambda cuando Amazon SQS se utiliza como origen de los eventos)
- AWS Lambda: Simultaneidad máxima

Ejemplos relacionados:

- The three most important AWS WAF rate-based rules (Las tres reglas más importantes basadas en índices de AWS WAF)
- Java Bucket4j
- Python token-bucket (Bucket de tokens de Python)
- Node token-bucket (Bucket de tokens de nodos)
- .NET System Threading Rate Limiting (Limitación de índices de subprocesos del sistema .NET)

Vídeos relacionados:

- Implementing GraphQL API security best practices with AWS AppSync (Implementación de prácticas recomendadas de seguridad de la API GraphQL con AWS AppSync)

Herramientas relacionadas:

- Amazon API Gateway
- AWS AppSync
- Amazon SQS
- Amazon Kinesis
- AWS WAF

REL05-BP03 Controlar y limitar las llamadas de reintento

Utilice un retroceso exponencial para reintentar las solicitudes a intervalos progresivamente más largos entre cada reintento. Introduzca una fluctuación entre reintentos para aleatorizar los intervalos de reintentos. Limite el número máximo de reintentos.

Resultado deseado: entre los componentes típicos de un sistema de software distribuido se incluyen servidores, equilibradores de carga, bases de datos y servidores DNS. Durante el funcionamiento normal, estos componentes pueden responder a las solicitudes con errores temporales o limitados, y también con errores que serían persistentes independientemente de los reintentos. Cuando los clientes realizan solicitudes a los servicios, esas solicitudes consumen recursos, como memoria, subprocesos, conexiones, puertos o cualquier otro recurso limitado. Controlar y limitar los reintentos es una estrategia para liberar y minimizar el consumo de recursos, de modo que los componentes del sistema sometidos a presión no se sobrecarguen.

Cuando se agota el tiempo de espera de las solicitudes del cliente o se reciben respuestas de error, deben determinar si deben volver a intentarlo o no. Si lo vuelven a intentar, lo hacen con un retroceso exponencial con fluctuaciones y un valor de reintento máximo. Como resultado, los servicios y procesos de backend tienen menos carga y más tiempo para recuperarse automáticamente, lo que se traduce en una recuperación más rápida y una tramitación satisfactoria de las solicitudes.

Patrones comunes de uso no recomendados:

- Implementar los reintentos sin añadir valores de retroceso exponencial, fluctuación y reintentos máximos. El retroceso y la fluctuación ayudan a evitar picos de tráfico artificiales debidos a reintentos coordinados involuntariamente a intervalos comunes.
- Implementar reintentos sin probar sus efectos o asumir que los reintentos ya están integrados en un SDK sin probar los escenarios de reintento.
- No entender los códigos de error publicados de las dependencias, lo que lleva a volver a intentar todos los errores, incluidos los que tienen una causa clara que indica una falta de permisos, un error de configuración u otro problema que es de esperar que no se pueda resolver sin una intervención manual.
- No utilizar prácticas de observabilidad, como monitorización y alertas en caso de errores de servicio repetidos, para conocer problemas subyacentes y poder solucionarlos.
- Desarrollar mecanismos de reintento personalizados cuando son suficientes las capacidades de reintento integradas o de terceros.
- Realizar reintentos en varias capas de la pila de aplicaciones de una forma que se acumulen, lo que consume aún más recursos en una tormenta de reintentos. Asegúrese de entender cómo

afectan estos errores a las dependencias en las que se basa y, a continuación, implemente los reintentos en un solo nivel.

- Reintentar llamadas de servicio que no son idempotentes, lo que provoca efectos secundarios inesperados, como resultados duplicados.

Beneficios de establecer esta práctica recomendada: los reintentos ayudan a los clientes a obtener los resultados deseados cuando las solicitudes fallan, pero también consumen más tiempo del servidor para obtener las respuestas satisfactorias que desean. Cuando los errores son poco frecuentes o transitorios, los reintentos funcionan bien. Cuando los errores se deben a una sobrecarga de recursos, los reintentos pueden empeorar las cosas. Añadir un retroceso exponencial con fluctuaciones para los reintentos de los clientes permite que los servidores se recuperen cuando los errores se deben a una sobrecarga de recursos. La fluctuación evita que haya picos de solicitudes y el retroceso disminuye el escalamiento de la carga provocado por la adición de reintentos a la carga normal de solicitudes. Por último, es importante configurar un número de reintentos máximo o un tiempo transcurrido máximo para evitar que se acumulen tareas pendientes que generen errores metaestables.

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

Guía para la implementación

Controle y limite las llamadas de reintento. Use el retroceso exponencial para los reintentos tras intervalos cada vez más largos. Introduzca una fluctuación para aleatorizar los intervalos de reintento y limite el número máximo de reintentos.

Algunos SDK de AWS implementan los reintentos y el retroceso exponencial de forma predeterminada. Utilice estas implementaciones de AWS integradas cuando corresponda en su carga de trabajo. Implemente una lógica similar en su carga de trabajo cuando llame a servicios que sean idempotentes y en los que los reintentos mejoren la disponibilidad de sus clientes. Decida cuáles son los tiempos de espera y cuándo dejar de reintentar según su caso de uso. Cree y realice escenarios de prueba para esos casos de uso de reintentos.

Pasos para la implementación

- Determine la capa óptima de la pila de aplicaciones para implementar los reintentos de los servicios de los que depende su aplicación.

- Tenga en cuenta que los SDK existentes implementan estrategias de reintento probadas con retroceso exponencial y fluctuaciones para el lenguaje que elija, y dé preferencia a estas estrategias en lugar de escribir sus propias implementaciones de reintentos.
- Verifique que los servicios sean idempotentes antes de implementar los reintentos. Una vez implementados, asegúrese de que se prueben y se utilicen regularmente en producción.
- Al llamar a las API del servicio de AWS, utilice los SDK de AWS y AWS CLI y comprenda las opciones de configuración de reintentos. Determine si los valores predeterminados funcionan para su caso de uso, pruébelos y ajústelos según sea necesario.

Recursos

Prácticas recomendadas relacionadas:

- REL04-BP04 Hacer que todas las respuestas sean idempotentes
- REL05-BP02 Limitar las solicitudes
- REL05-BP04 Responder rápido a los errores y limitar las colas
- REL05-BP05 Definir tiempos de espera del cliente
- REL11-BP01 Supervisar todos los componentes de la carga de trabajo para detectar errores

Documentos relacionados:

- Error Retries and Exponential Backoff in AWS (Reintentos en caso de error y retroceso exponencial en AWS)
- La Amazon Builders' Library: Tiempos de espera, reintentos y retroceso con alteración
- Exponential backoff and jitter (Retroceso exponencial y fluctuación)
- Making retries safe with idempotent APIs (Hacer que los reintentos sean seguros con API idempotentes)

Ejemplos relacionados:

- Spring Retry (Reintento de Spring)
- Resilience4j Retry (Reintento de Resilience4j)

Vídeos relacionados:

- Retry, backoff, and jitter: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328) (Reintento, retroceso y fluctuación: AWS re:Invent 2019: Presentación de Amazon Builders' Library [DOP328])

Herramientas relacionadas:

- AWS SDKs and Tools: Retry behavior (SDK y herramientas de AWS: comportamiento de reintento)
- AWS Command Line Interface: Reintentos de AWS CLI

REL05-BP04 Responder rápido a los errores y limitar las colas

Cuando un servicio no pueda responder correctamente a una solicitud, responda rápido a los errores. Esto permite que se liberen los recursos asociados a una solicitud y que un servicio se recupere cuando se le agotan los recursos. La respuesta rápida a los errores es un patrón de diseño de software bien establecido que se puede utilizar para conseguir cargas de trabajo enormemente fiables en la nube. Las colas también son un patrón de integración empresarial bien establecido que puede suavizar la carga y permitir a los clientes liberar recursos cuando se pueda tolerar el procesamiento asíncrono. Cuando un servicio puede responder correctamente en condiciones normales, pero falla cuando el índice de solicitudes es demasiado alto, utilice una cola para almacenar en búfer las solicitudes. Sin embargo, no permita que se acumulen largas colas de tareas pendientes, ya que eso podría hacer que se procesaran solicitudes obsoletas a las que un cliente ya ha renunciado.

Resultado deseado: cuando los sistemas sufren contención de recursos, tiempos de espera, excepciones o errores grises que hacen que los objetivos de nivel de servicio sean inalcanzables, las estrategias de respuesta rápida a los errores permiten recuperar el sistema más rápido. Los sistemas que deben absorber los picos de tráfico y pueden adaptarse al procesamiento asíncrono pueden mejorar la fiabilidad al permitir a los clientes liberar rápidamente las solicitudes mediante el uso de colas para almacenar en búfer las solicitudes a los servicios de backend. Cuando las solicitudes a las colas se almacenan en búfer, se implementan estrategias de administración de colas para evitar retrasos insuperables.

Patrones comunes de uso no recomendados:

- Implementar colas de mensajes, pero no configurar colas de mensajes fallidos (DLQ) ni alarmas en los volúmenes de DLQ para detectar cuándo está fallando un sistema.

- No medir la antigüedad de los mensajes de una cola, que es una medida de la latencia para saber cuándo los usuarios de la cola sufren retrasos o producen errores que dan lugar a reintentos.
- No borrar los mensajes pendientes de una cola cuando no sirve de nada procesar esos mensajes si la empresa ya no necesita hacerlo.
- Configurar colas de primero en entrar/primero en salir (FIFO) cuando las colas de último en entrar, primero en salir (LIFO) responderían mejor a las necesidades de los clientes, por ejemplo, cuando no se requieren pedidos estrictos y el procesamiento pendiente retrasa todas las solicitudes nuevas y urgentes, lo que hace que se infrinjan los niveles de servicio de todos los clientes.
- Exponer las colas internas a los clientes en lugar de exponer las API que administran la entrada de trabajo y colocan las solicitudes en colas internas.
- Combinar demasiados tipos de solicitudes de trabajo en una sola cola puede agravar las condiciones de las tareas pendientes al distribuir la demanda de recursos entre los tipos de solicitudes.
- Procesar solicitudes complejas y simples en la misma cola, a pesar de necesitar diferentes niveles de monitorización, tiempos de espera y asignaciones de recursos.
- No validar las entradas ni utilizar afirmaciones para implementar mecanismos de respuesta rápida a los errores en el software que envíen las excepciones a componentes de nivel superior que puedan gestionar los errores con facilidad.
- No eliminar los recursos que fallan del enrutamiento de solicitudes, especialmente cuando los errores grises emiten tanto éxitos como errores debido a bloqueos y reinicios, errores de dependencia intermitentes, una reducción de la capacidad o la pérdida de paquetes de red.

Beneficios de establecer esta práctica recomendada: los sistemas que responden rápido a los errores son más fáciles de depurar y corregir y, a menudo, revelan problemas de codificación y configuración antes de que las versiones se publiquen en producción. Los sistemas que incorporan estrategias de puesta en cola eficaces tienen una mayor resiliencia y fiabilidad a los picos de tráfico y a las condiciones de errores intermitentes del sistema.

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

Guía para la implementación

Las estrategias de respuesta rápida a los errores pueden codificarse en soluciones de software y también configurarse en la infraestructura. Además de la respuesta rápida a los errores, las colas son una técnica arquitectónica sencilla pero potente para desacoplar los componentes del sistema sin problemas de carga. Amazon CloudWatch proporciona capacidades para monitorizar los errores

y alertar en caso de que existan. Una vez que se sabe que un sistema está fallando, se pueden invocar estrategias de mitigación, como el alejamiento de los recursos deteriorados. Cuando los sistemas implementan colas con Amazon SQS y otras tecnologías de cola para facilitar la carga, deben considerar cómo administrar los atrasos en las colas, así como los errores en el consumo de mensajes.

Pasos para la implementación

- Implemente afirmaciones programáticas o métricas específicas en su software y utilícelas para alertar explícitamente sobre problemas del sistema. Amazon CloudWatch le ayuda a crear métricas y alarmas basadas en el patrón de registro de la aplicación y la instrumentación del SDK.
- Utilice métricas y alarmas de CloudWatch para alejarse de los recursos deteriorados que aumentan la latencia del procesamiento o que no procesan las solicitudes de forma reiterada.
- Utilice el procesamiento asíncrono diseñando API que acepten solicitudes y las anexen a las colas internas mediante Amazon SQS y luego respondan al cliente que produce los mensajes con un mensaje de éxito, de modo que el cliente pueda liberar recursos y continuar con otras tareas mientras los consumidores de la cola del backend procesan las solicitudes.
- Mida y monitorice la latencia de procesamiento de las colas generando una métrica de CloudWatch cada vez que se retire un mensaje de una cola comparándolo en ese momento con la marca de tiempo del mensaje.
- Cuando los errores impidan procesar correctamente los mensajes o los picos de tráfico en los volúmenes que no se pueden procesar dentro de los acuerdos de nivel de servicio, aparte el tráfico antiguo o excesivo y colóquelo en una cola secundaria. Esto permite procesar de forma prioritaria los trabajos nuevos y dejar los antiguos para cuando haya capacidad disponible. Esta técnica es una aproximación al procesamiento LIFO y permite que el sistema procese normalmente todos los trabajos nuevos.
- Utilice colas de mensajes fallidos o redireccione las colas para sacar de la lista de espera los mensajes que no se puedan procesar y colocarlos en una ubicación que pueda investigarse y resolverse más adelante
- Vuelva a intentarlo o, cuando sea tolerable, elimine los mensajes antiguos comparándolos en ese momento con la marca de tiempo del mensaje y descartando los mensajes que ya no sean relevantes para el cliente que los ha solicitado.

Recursos

Prácticas recomendadas relacionadas:

- REL04-BP02 Implementar dependencias con acoplamiento flexible
- REL05-BP02 Limitar las solicitudes
- REL05-BP03 Controlar y limitar las llamadas de reintento
- REL06-BP02 Definir y calcular métricas (agregación)
- REL06-BP07 Supervisar el seguimiento de las solicitudes de principio a fin en todo el sistema

Documentos relacionados:

- Cómo evitar demoras de colas insuperables
- Respuesta rápida a errores
- How can I prevent an increasing backlog of messages in my Amazon SQS queue? (¿Cómo puedo evitar que se acumulen mensajes en mi cola de Amazon SQS?
- Elastic Load Balancing: Zonal Shift (Cambio de zona)
- Amazon Route 53 Application Recovery Controller: Routing control for traffic failover (Amazon Route 53 Application Recovery Controller: control de enrutamiento para conmutación por error del tráfico)

Ejemplos relacionados:

- Enterprise Integration Patterns: Dead Letter Channel (Patrones de integración empresarial: canal de mensajes fallidos)

Vídeos relacionados:

- AWS re:Invent 2022 - Operating highly available Multi-AZ applications (AWS re:Invent 2022: Funcionamiento de aplicaciones multi-AZ de alta disponibilidad)

Herramientas relacionadas:

- Amazon SQS
- Amazon MQ
- AWS IoT Core
- Amazon CloudWatch

REL05-BP05 Definir tiempos de espera del cliente

Defina tiempos de espera adecuados para las conexiones y las solicitudes, verifíquelos sistemáticamente y no use los valores predeterminados, ya que no tienen en cuenta las características específicas de la carga de trabajo.

Resultado deseado: en los tiempos de espera de los clientes, se debe tener en cuenta el coste para el cliente, el servidor y la carga de trabajo asociados a la espera de las solicitudes que tardan un tiempo anormal en completarse. Dado que no es posible conocer la causa exacta de ningún tiempo de espera, los clientes deben utilizar el conocimiento de los servicios para fijar expectativas sobre las causas probables y los tiempos de espera adecuados.

El tiempo de espera de las conexiones del cliente se agota en función de los valores configurados. Cuando el tiempo de espera se agota, los clientes toman la decisión de dar marcha atrás y volver a intentarlo o abrir un disyuntor. Estos patrones evitan que se emitan solicitudes que puedan agravar una condición de error subyacente.

Patrones comunes de uso no recomendados:

- No estar al tanto de los tiempos de espera del sistema o de los tiempos de espera predeterminados.
- No estar al tanto del tiempo normal de finalización de las solicitudes.
- No conocer las posibles causas por las que las solicitudes tardan un tiempo anormalmente largo en completarse ni los costes para el rendimiento del cliente, el servicio o la carga de trabajo asociados a la espera a que se completen.
- No conocer la probabilidad de que la red deteriorada haga que una solicitud falle solo una vez que se haya agotado el tiempo de espera, ni de los costes que supone para el rendimiento del cliente y la carga de trabajo no utilizar un tiempo de espera más corto.
- No probar escenarios de tiempo de espera tanto para las conexiones como para las solicitudes.
- Definir tiempos de espera demasiado altos, lo que puede provocar tiempos de espera prolongados y aumentar la utilización de los recursos.
- Definir tiempos de espera demasiado bajos, lo que provoca errores artificiales.
- Pasar por alto los patrones para solucionar los errores de tiempo de espera de las llamadas remotas, como disyuntores y reintentos.
- No considerar la posibilidad de monitorizar los índices de errores de las llamadas de servicio, los objetivos de nivel de servicio referentes a la latencia y los valores atípicos de latencia. Estas métricas pueden proporcionar información sobre tiempos de espera agresivos o permisivos.

Beneficios de establecer esta práctica recomendada: los tiempos de espera de las llamadas remotas están configurados y los sistemas están diseñados para gestionar los tiempos de espera correctamente, de modo que los recursos se conserven cuando las llamadas remotas responden con una lentitud anormal y los clientes del servicio gestionan correctamente los errores de tiempo de espera.

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

Guía para la implementación

Defina un tiempo de espera de conexión y un tiempo de espera de solicitud en cualquier llamada de dependencia del servicio y, normalmente, en todas las llamadas de los procesos. Muchos marcos integran capacidades de tiempo de espera, pero tenga cuidado, ya que algunos tienen valores predeterminados que son infinitos o superiores a lo aceptable para sus objetivos de servicio. Un valor demasiado alto reduce la utilidad del tiempo de espera porque se siguen consumiendo recursos mientras el cliente espera a que transcurra el tiempo de espera. Un valor demasiado bajo puede generar un aumento del tráfico en el backend y un aumento de la latencia debido a que las solicitudes realizan demasiados reintentos. En algunos casos, esto puede producir una interrupción completa si se reintentan todas las solicitudes.

Tenga en cuenta lo siguiente al determinar las estrategias de tiempo de espera:

- Las solicitudes pueden tardar más de lo normal en procesarse debido a su contenido, a deficiencias en un servicio de destino o a un error en la partición de la red.
- Las solicitudes con contenido anormalmente caro podrían consumir recursos innecesarios del servidor y del cliente. En este caso, si se agota el tiempo de espera de estas solicitudes y no se vuelven a intentar, se pueden conservar los recursos. Los servicios también deberían protegerse del contenido anormalmente caro con restricciones y tiempos de espera del lado del servidor.
- Se puede agotar el tiempo de espera y volver a intentar las solicitudes que tarden un tiempo anormalmente largo debido a una interrupción del servicio. Se deben tener en cuenta los costes del servicio de la solicitud y el reintento, pero si la causa es una deficiencia localizada, es probable que el reintento no sea caro y reduzca el consumo de recursos del cliente. El tiempo de espera también puede liberar recursos del servidor según la naturaleza de la deficiencia.
- Se puede agotar el tiempo de espera y volver a intentar las solicitudes que tarden mucho en completarse porque la red no ha podido entregar la solicitud o la respuesta. Como la solicitud o la respuesta no se han entregado, el resultado habría sido un error independientemente del tiempo de espera. En este caso, el tiempo de espera no liberará los recursos del servidor, pero sí liberará los recursos del cliente y mejorará el rendimiento de la carga de trabajo.

Aproveche patrones de diseño bien establecidos, como los reintentos y los disyuntores, para gestionar los tiempos de espera correctamente y ofrecer enfoques de respuesta rápida a los errores. Los SDK de AWS y AWS CLI permiten configurar los tiempos de espera de las conexiones y las solicitudes y los reintentos con un retroceso exponencial y fluctuaciones. Las funciones de AWS Lambda admiten la configuración de tiempos de espera, y con AWS Step Functions, puede crear disyuntores de poco código que utilicen integraciones predefinidas con los servicios y los SDK de AWS. AWS App Mesh Envoy incluye capacidades de tiempo de espera y de disyuntor.

Pasos para la implementación

- Configure tiempos de espera en las llamadas de servicio remotas y aproveche las características integradas de tiempo de espera del lenguaje o las bibliotecas de tiempo de espera de código abierto.
- Cuando su carga de trabajo realice llamadas con un SDK de AWS, consulte la documentación para ver la configuración del tiempo de espera específica de cada lenguaje.
  - Python
  - PHP
  - .NET
  - Ruby
  - Java
  - Go
  - Node.js
  - C++
- Cuando utilice SDK de AWS o comandos de la AWS CLI en su carga de trabajo, defina los valores predeterminados de AWS al configurar los valores de tiempo de espera predeterminados para `connectTimeoutInMillis` y `tlsNegotiationTimeoutInMillis`.
- Utilice las opciones de línea de comandos `cli-connect-timeout` y `cli-read-timeout` para controlar comandos de la AWS CLI puntuales de los servicios de AWS.
- Monitorice las llamadas de servicio remotas para comprobar si hay tiempos de espera y configure alarmas en caso de errores persistentes para poder gestionar los escenarios de error de forma proactiva.
- Implemente Métricas de CloudWatch y detección de anomalías de CloudWatch en los índices de error de las llamadas, los objetivos de nivel de servicio en lo que se refiere a la latencia y los valores atípicos de latencia para proporcionar información sobre la administración de tiempos de espera demasiado agresivos o permisivos.

- Configure tiempos de espera en funciones de Lambda.
- Los clientes de API Gateway deben implementar sus propios reintentos al gestionar los tiempos de espera. API Gateway admite un tiempo de espera de integración de 50 milisegundos a 29 segundos para integraciones posteriores y no vuelve a intentarlo cuando se agota el tiempo de espera de las solicitudes de integración.
- Implemente el patrón de disyuntor para que no se realicen llamadas remotas cuando se agote el tiempo de espera. Abra el circuito para evitar llamadas fallidas y ciérrelo cuando las llamadas respondan normalmente.
- Para cargas de trabajo basadas en contenedores, consulte las características de App Mesh Envoy para utilizar los tiempos de espera y los disyuntores integrados.
- Utilice AWS Step Functions para crear disyuntores de poco código para las llamadas de servicio remotas, especialmente cuando se utilizan SDK nativos de AWS e integraciones de Step Functions compatibles para simplificar la carga de trabajo.

Recursos

Prácticas recomendadas relacionadas:

- REL05-BP03 Controlar y limitar las llamadas de reintento
- REL05-BP04 Responder rápido a los errores y limitar las colas
- REL06-BP07 Supervisar el seguimiento de las solicitudes de principio a fin en todo el sistema

Documentos relacionados:

- AWS SDK: Retries and Timeouts (SDK de AWS: reintentos y tiempos de espera)
- La Amazon Builders' Library: Tiempos de espera, reintentos y retroceso con alteración
- Cuotas de Amazon API Gateway y notas importantes
- AWS Command Line Interface: Command line options (Opciones de línea de comandos)
- AWS SDK for Java 2.x: Configure API Timeouts (Configurar los tiempos de espera de la API)
- AWS Botocore using the config object and Config Reference (AWS Botocore: Uso del objeto config y referencia de configuración)
- AWS SDK para .NET: Retries and Timeouts (Reintentos y tiempos de espera)
- AWS Lambda: Configuring Lambda function options (Configuración de las opciones de la función Lambda)

Ejemplos relacionados:

- Using the circuit breaker pattern with AWS Step Functions and Amazon DynamoDB (Uso del patrón del disyuntor con AWS Step Functions y Amazon DynamoDB)
- Martin Fowler: CircuitBreaker (Disyuntor)

Herramientas relacionadas:

- SDK de AWS
- AWS Lambda
- Amazon SQS
- AWS Step Functions
- AWS Command Line Interface

REL05-BP06 Crear servicios sin estado cuando sea posible

Los servicios deben o bien no requerir estado o bien descargar el estado, de forma que entre solicitudes de clientes distintos no haya dependencia en los datos almacenados localmente en disco y en memoria. Esto permite reemplazar los servidores a voluntad sin que la disponibilidad resulte afectada. Amazon ElastiCache y Amazon DynamoDB son buenos destinos para el estado descargado.

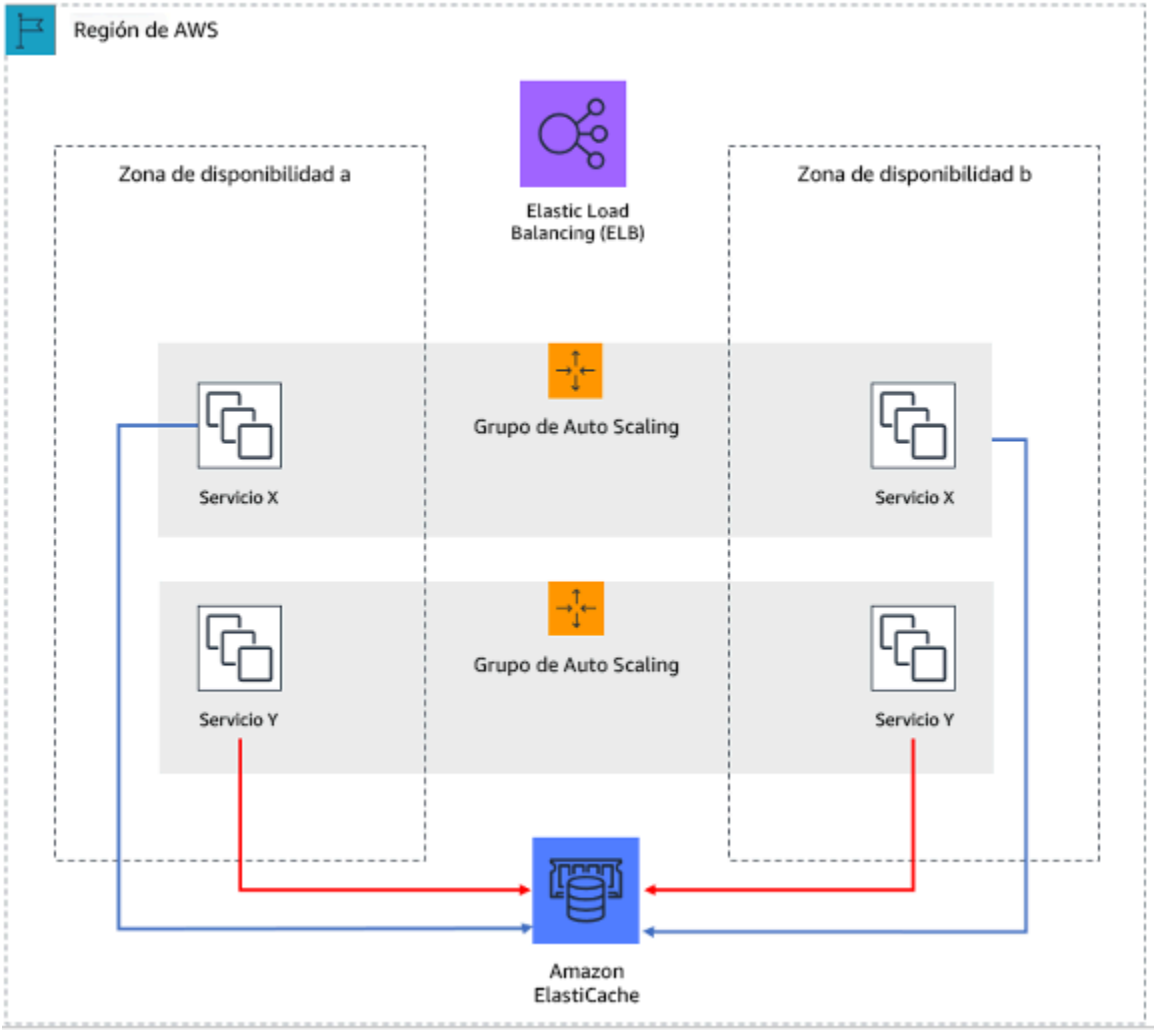


Figura 7: En esta aplicación web sin estado, el estado de la sesión se descarga en Amazon ElastiCache.

Cuando los usuarios o los servicios interactúan con una aplicación, suelen realizar una serie de interacciones que constituyen una sesión. Una sesión es un dato único para los usuarios que persiste entre las solicitudes mientras utilizan la aplicación. Una aplicación sin estado es aquella que no necesita conocer las interacciones anteriores y no almacena la información de la sesión.

Una vez se ha diseñado para no tener estado, puede utilizar servicios de computación sin servidor, como AWS Lambda o AWS Fargate.

Además del reemplazo del servidor, otro beneficio de las aplicaciones sin estado es que pueden escalar horizontalmente porque cualquiera de los recursos de computación disponibles (como las instancias EC2 y funciones AWS Lambda) puede dar servicio a cualquier solicitud.

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

Guía para la implementación

- Haga que sus aplicaciones no tengan estado. Las aplicaciones sin estado permiten el escalado horizontal y toleran el error de un nodo individual.
  - Elimine el estado que realmente podría almacenarse en parámetros de solicitud.
  - Tras examinar si el estado es realmente necesario, mueva cualquier seguimiento de estado a una caché o un almacén de datos resilientes de varias zonas como Amazon ElastiCache, Amazon RDS, Amazon DynamoDB o una solución de datos distribuidos de terceros. Almacene el estado que no se pudo mover a almacenes de datos resilientes.
    - Algunos datos (como las cookies) se pueden pasar a encabezados o parámetros de consulta.
    - Refactorice para eliminar el estado que se puede pasar rápidamente a las solicitudes.
    - Algunos datos pueden no resultar realmente necesarios para la solicitud y pueden recuperarse bajo demanda.
    - Elimine los datos que se puedan recuperar asincrónicamente.
    - Elija un almacén de datos que cumpla los requisitos para el estado requerido.
    - Considere la posibilidad de usar una base de datos NoSQL para datos no relacionales.

Recursos

Documentos relacionados:

- La Amazon Builders' Library: Evitar el retroceso en sistemas distribuidos
- La Amazon Builders' Library: Evitar trabajos pendientes en colas insalvables
- La Amazon Builders' Library: Desafíos y estrategias del almacenamiento en caché

REL05-BP07 Implementar recursos de emergencia

Los recursos de emergencia son procesos rápidos que pueden mitigar el impacto en la disponibilidad de la carga de trabajo.

Los recursos de emergencia desactivan, limitan o cambian el comportamiento de componentes o dependencias mediante mecanismos conocidos y probados. Esto puede aliviar las deficiencias de la carga de trabajo causadas por el agotamiento de los recursos debido a los aumentos inesperados de la demanda y reducir el impacto de los fallos en los componentes no críticos de la carga de trabajo.

Resultado deseado: al implementar recursos de emergencia, puede establecer procesos que se sabe que son buenos para mantener la disponibilidad de los componentes críticos de su carga de trabajo. La carga de trabajo debe degradarse de forma estable y seguir realizando sus funciones críticas para la empresa durante la activación de un recurso de emergencia. Para obtener más información sobre la degradación estable, consulte «REL05-BP01 Implementar una degradación estable para transformar las dependencias estrictas en flexibles».

Antipatrones usuales:

- El fallo de las dependencias no críticas repercute en la disponibilidad de su carga de trabajo principal.
- No probar o verificar el comportamiento de los componentes críticos durante el deterioro de los componentes no críticos.
- No definir criterios claros y deterministas para la activación o desactivación de un recurso de emergencia.

Beneficios de establecer esta práctica recomendada: la implementación de recursos de emergencia puede mejorar la disponibilidad de los componentes críticos de su carga de trabajo al proporcionar a sus solucionadores procesos establecidos para responder a picos inesperados de demanda o fallos de dependencias no críticas.

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

Guía para la implementación

- Identifique los componentes críticos de su carga de trabajo.
- Diseñe y cree los componentes críticos de su carga de trabajo para que resistan los fallos de los componentes no críticos.
- Realice pruebas para validar el comportamiento de sus componentes críticos durante el fallo de los componentes no críticos.
- Defina y supervise las métricas o los factores desencadenantes relevantes para iniciar los procedimientos de recursos de emergencia.
- Defina los procedimientos (manuales o automáticos) que componen el recurso de emergencia.

Pasos para la implementación

- Identifique los componentes críticos para la empresa en su carga de trabajo.

    - Cada componente técnico de su carga de trabajo debe asignarse a su función empresarial relevante y clasificarse como crítico o no crítico. Para ver ejemplos de funciones críticas y no críticas de Amazon, lea «Any Day Can Be Prime Day: How Amazon.com Search Uses Chaos Engineering to Handle Over 84K Requests Per Second».
    - Se trata de una decisión tanto técnica como empresarial, y varía según la organización y la carga de trabajo.
- Diseñe y cree los componentes críticos de su carga de trabajo para que resistan los fallos de los componentes no críticos.
    - Durante el análisis de dependencias, tenga en cuenta todos los modos de fallo potenciales y verifique que sus mecanismos de recursos de emergencia proporcionan la funcionalidad crítica a los componentes downstream.
- Realice pruebas para validar el comportamiento de sus componentes críticos durante la activación de sus recursos de emergencia.
    - Evite el comportamiento bimodal. Para obtener más información, consulte «REL11-BP05 Usar la estabilidad estática para evitar el comportamiento bimodal».
- Defina, supervise y alerte sobre las métricas relevantes para iniciar el procedimiento del recurso de emergencia.
    - Encontrar las métricas adecuadas para supervisar depende de su carga de trabajo. Algunos ejemplos de métricas son la latencia o el número de solicitudes fallidas a una dependencia.
- Defina los procedimientos (manuales o automáticos) que componen el recurso de emergencia.
    - Esto puede incluir mecanismos como el desbordamiento de carga, la limitación de solicitudes o la implementación de una degradación estable.

Recursos

Prácticas recomendadas relacionadas:

- REL05-BP01 Implementar una degradación estable para transformar las dependencias estrictas en flexibles
- REL05-BP02 Limitar las solicitudes
- REL11-BP05 Usar la estabilidad estática para evitar el comportamiento bimodal

Documentos relacionados:

- Automatización de implementaciones seguras y sin intervención

- «Any Day Can Be Prime Day: How Amazon.com Search Uses Chaos Engineering to Handle Over 84K Requests Per Second»

Vídeos relacionados:

- «AWS re:Invent 2020: Reliability, consistency, and confidence through immutability»

Administración de cambios

Preguntas

- FIABILIDAD 6. ¿Cómo supervisa los recursos de las cargas de trabajo?
- FIABILIDAD 7. ¿Cómo diseña su carga de trabajo para que se adapte a los cambios en la demanda?
- FIABILIDAD 8. ¿Cómo implementa los cambios?

FIABILIDAD 6. ¿Cómo supervisa los recursos de las cargas de trabajo?

Los registros y las métricas son una potente herramienta para obtener información sobre el estado de sus cargas de trabajo. Puede configurar su carga de trabajo de forma que supervise registros y métricas, y envíe notificaciones cuando se crucen ciertos umbrales o se produzcan eventos importantes. La supervisión permite que su carga de trabajo reconozca cuándo se cruzan umbrales de bajo rendimiento o cuándo se producen errores, para que pueda recuperarse de los errores de forma automática una vez recibida una respuesta.

Prácticas recomendadas

- REL06-BP01 Supervisar todos los componentes de la carga de trabajo (generación)
- REL06-BP02 Definir y calcular métricas (agregación)
- REL06-BP03 Enviar notificaciones (procesamiento y alarmas en tiempo real)
- REL06-BP04 Automatizar las respuestas (procesamiento y alarmas en tiempo real)
- REL06-BP05 Análisis
- REL06-BP06 Realizar revisiones con frecuencia
- REL06-BP07 Supervisar el seguimiento de las solicitudes de principio a fin en todo el sistema

REL06-BP01 Supervisar todos los componentes de la carga de trabajo (generación)

Supervise los componentes de la carga de trabajo con Amazon CloudWatch o herramientas de terceros. Supervise los servicios de AWS con el panel de AWS Health.

Debería supervisar todos los componentes de su carga de trabajo, incluidos los niveles del front-end, la lógica empresarial y el almacenamiento. Defina métricas claves, describa cómo extraerlas de los registros (si fuera necesario) y establezca umbrales para desencadenar los eventos de alarma correspondientes. Asegúrese de que las métricas sean pertinentes para los indicadores clave de rendimiento (KPI) de su carga de trabajo, y utilice métricas y registros para identificar signos de advertencia tempranos de degradación del servicio. Por ejemplo, una métrica relacionada con los resultados empresariales como el número de pedidos procesado satisfactoriamente por minuto, puede indicar problemas con la carga de trabajo más rápido que una métrica técnica, como el uso de la CPU. Utilice el panel de AWS Health para obtener una vista personalizada sobre el rendimiento y la disponibilidad de los servicios de AWS subyacentes a sus recursos de AWS.

La supervisión en la nube ofrece nuevas oportunidades. La mayoría de proveedores en la nube han desarrollado enlaces personalizables y pueden proporcionar conocimientos para ayudarle a supervisar varias capas de su carga de trabajo. Los servicios de AWS como Amazon CloudWatch aplican algoritmos estadísticos y de machine learning para analizar continuamente las métricas de los sistemas y aplicaciones, determinar las bases de referencia normales y hacer aflorar anomalías con una intervención mínima del usuario. Los algoritmos de detección de anomalías tienen en cuenta la estacionalidad y los cambios en las tendencias de las métricas.

AWS pone a disposición una gran cantidad de información de supervisión y registro para el consumo que se puede usar para definir métricas específicas de la carga de trabajo, procesos de cambio en la demanda y adoptar técnicas de machine learning independientemente de los conocimientos sobre ML.

Además, puede supervisar todos sus puntos de conexión externos para asegurarse de que sean independientes de su implementación base. Esta supervisión activa se puede llevar a cabo con transacciones sintéticas (a las que a veces se denomina «canaries» de usuario, y que no deben confundirse con los despliegues de valores controlados o «canary»), que ejecutan periódicamente varias tareas comunes que se ajustan a las acciones realizadas por los clientes de la carga de trabajo. Mantenga una duración breve para estas tareas y asegúrese de no sobrecargar sus cargas de trabajo durante las pruebas. Amazon CloudWatch Synthetics le permite: crear pruebas de transacciones o «canaries» sintéticas para supervisar sus puntos de conexión y API. También puede combinar los nodos de cliente de la «canary» sintética con la consola de AWS X-Ray para detectar

qué «canaries» sintéticas están teniendo problemas de errores, fallos o limitaciones para el periodo de tiempo seleccionado.

Resultado deseado:

Recopilar y usar métricas esenciales de todos los componentes de la carga de trabajo para garantizar la fiabilidad de la carga de trabajo y una experiencia de usuario óptima. Detectar que una carga de trabajo no consigue los resultados empresariales le permite declarar rápidamente una situación de desastre y recuperarse de un incidente.

Patrones de uso no recomendados comunes:

- Supervisar solamente las interfaces externas con su carga de trabajo
- No generar métricas específicas de una carga de trabajo y basarse solamente en las métricas que proporcionan los servicios de AWS que usa su carga de trabajo.
- Usar exclusivamente métricas técnicas en su carga de trabajo y no supervisar las métricas relacionadas con KPI no técnicos a los que contribuye la carga de trabajo.
- Basarse en el tráfico de producción y las comprobaciones de estado sencillas para supervisar y evaluar el estado de las cargas de trabajo.

Beneficios de establecer esta práctica recomendada: La supervisión de todos los niveles de la carga de trabajo le permite prever y resolver los problemas rápidamente en los componentes de la carga de trabajo.

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

Guía para la implementación

1. Habilite el registro cuando esté disponible. La supervisión de los datos debe obtenerse a partir de todos los componentes de las cargas de trabajo. Active métodos de registro adicionales, como los registros de acceso de S3, y permita que su carga de trabajo registre datos específicos de la carga de trabajo. Recopile métricas para los promedios de CPU, E/S de red y E/S de disco de servicios como Amazon ECS, Amazon EKS, Amazon EC2, Elastic Load Balancing, AWS Auto Scaling y Amazon EMR. Consulte Servicios de AWS que publican métricas de CloudWatch para consultar una lista de servicios de AWS que publican métricas en CloudWatch.
2. Revise todas las métricas predeterminadas y explore las carencias en cuanto a recopilación de datos. Todos los servicios generan métricas predeterminadas. La recopilación de métricas predeterminadas le permite comprender mejor las dependencias entre los componentes de la

carga de trabajo, y cómo la fiabilidad y el rendimiento de los componentes afectan a la carga de trabajo. También puede crear y publicar sus propias métricas en CloudWatch utilizando la AWS CLI o una API. Esto

3. Evalúe todas las métricas para decidir sobre cuáles alertar en cada servicio de AWS en su carga de trabajo. Puede decidir seleccionar un subconjunto de métricas que tenga un impacto importante en la fiabilidad de la carga de trabajo. Al centrarse en las métricas y umbrales críticos, podrá refinar el número de alertas de emergencia y contribuir a reducir al mínimo los falsos positivos.
4. Defina las alertas y los procesos de recuperación para su carga de trabajo una vez que se active la alerta. La definición de alertas le permite notificar, escalar y seguir los pasos necesarios rápidamente para recuperarse de un incidente y cumplir el objetivo de tiempo de recuperación (RTO) prescrito. Puede usar alarmas de Amazon CloudWatch para invocar flujos de trabajo automatizados e iniciar procedimientos de recuperación basados en los umbrales definidos.
5. Explore el uso de transacciones sintéticas para recopilar datos relevantes sobre el estado de las cargas de trabajo. La supervisión sintética sigue las mismas rutas y lleva a cabo las mismas acciones que un cliente, lo que le permite verificar continuamente su experiencia de usuario incluso si no tiene tráfico de cliente en sus cargas de trabajo. Al usar transacciones sintéticas, puede detectar los problemas antes de que lo hagan los clientes.

Recursos

Prácticas recomendadas relacionadas:

- REL11-BP03 Automatizar la reparación en todas las capas

Documentos relacionados:

- Introducción al panel de AWS Health: estado de su cuenta
- Servicios de AWS que publican métricas de CloudWatch
- Registros de acceso para su Network Load Balancer
- Registros de acceso para su Application Load Balancer
- Acceso a Amazon CloudWatch Logs para AWS Lambda
- Registro de acceso al servidor de Amazon S3
- Habilitar los registros de acceso para su Classic Load Balancer
- Exportación de datos de registro a Amazon S3

- Instalar el agente de CloudWatch en una instancia Amazon EC2
- Publicar métricas personalizadas
- Uso de paneles de Amazon CloudWatch
- Uso de métricas de Amazon CloudWatch
- Uso de «canaries» (Amazon CloudWatch Synthetics)
- ¿Qué son Amazon CloudWatch Logs?

Guías del usuario:

- Cree un registro de seguimiento
- Supervisión de memoria y métricas del disco para las instancias Linux de Amazon EC2
- Uso de CloudWatch Logs con instancias de contenedor
- Registros de flujo de VPC
- ¿Qué es Amazon DevOps Guru?
- ¿Qué es AWS X-Ray?

Blogs relacionados:

- Depuración con Amazon CloudWatch Synthetics y AWS X-Ray

Ejemplos relacionados y talleres:

- Laboratorios de AWS Well-Architected: excelencia operativa - supervisión de dependencias
- La Amazon Builders' Library: Instrumentación de los sistemas distribuidos para la visibilidad de las operaciones
- Taller sobre observabilidad

REL06-BP02 Definir y calcular métricas (agregación)

Almacene los datos de registro y aplique filtros cuando sea necesario para calcular métricas, como las veces que se produce un evento de registro específico o la latencia calculada a partir de las marcas temporales del evento de registro.

Amazon CloudWatch y Amazon S3 sirven como las capas principales de agregación y almacenamiento. En algunos servicios, como AWS Auto Scaling y Elastic Load Balancing, las

métricas predeterminadas se proporcionan listas para usar para la carga de CPU o la latencia promedio de solicitudes en un clúster o instancia. En servicios de streaming, como VPC Flow Logs o AWS CloudTrail, los datos del evento se envían a CloudWatch Logs y debe definir y aplicar filtros para extraer las métricas de los datos del evento. Esto le presenta datos sobre las series temporales, que pueden servir como entradas para las alarmas de CloudWatch que defina para activar las alertas.

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

Guía para la implementación

- Defina y calcule métricas (agregación). Almacene los datos de registro y aplique filtros cuando sea necesario para calcular métricas, como las veces que se produce un evento de registro específico o la latencia calculada de las marcas temporales del evento de registro.
  - Los filtros de métricas definen los términos y patrones que analizar en los datos de registro a medida que se envían a CloudWatch Logs. CloudWatch Logs usa estos filtros para convertir los datos de registro en métricas numéricas de CloudWatch que puede representar en gráficas o a partir de las cuales puede establecer alarmas.
    - Buscar y filtrar datos de registro
  - Use un tercero de confianza para agregar registros.
    - Siga las instrucciones de la solución externa. La mayoría de los productos de terceros se integran con CloudWatch y Amazon S3.
  - Algunos servicios de AWS pueden publicar registros directamente en Amazon S3. Si su requisito principal para los registros es almacenarlos en Amazon S3, puede hacer que el servidor que crea los registros los envíe directamente a Amazon S3 sin instalar infraestructura adicional.
    - Enviar registros directamente a Amazon S3

Recursos

Documentos relacionados:

- Consultas de ejemplo de Amazon CloudWatch Logs Insights
- Depuración con Amazon CloudWatch Synthetics y AWS X-Ray
- Taller sobre observabilidad
- Buscar y filtrar datos de registro
- Enviar registros directamente a Amazon S3

- La Amazon Builders' Library: Instrumentación de los sistemas distribuidos para la visibilidad de las operaciones

REL06-BP03 Enviar notificaciones (procesamiento y alarmas en tiempo real)

Cuando las organizaciones detectan posibles problemas, envían notificaciones y alertas en tiempo real al personal y los sistemas correspondientes para poder responder de manera rápida y eficaz a estos problemas.

Resultado deseado: es posible responder rápidamente a los eventos operativos con la configuración de las alarmas correspondientes en función de las métricas del servicio y la aplicación. Cuando se superan los umbrales de alarma, se avisa al personal y a los sistemas adecuados para que puedan abordar los problemas subyacentes.

Patrones comunes de uso no recomendados:

- Las alarmas están configuradas con un umbral excesivamente alto, lo que impide que se envíen notificaciones vitales.
- Las alarmas están configuradas con un umbral demasiado bajo, lo que provoca inacción en las alertas importantes por el ruido que genera el exceso de notificaciones.
- Las alarmas y los umbrales no se actualizan cuando hay cambios de uso.
- En el caso de las alarmas que se abordan mejor con acciones automatizadas, en lugar de generar la acción automatizada, se envían notificaciones al personal, lo que provoca un exceso de notificaciones.

Beneficios de establecer esta práctica recomendada: enviar notificaciones y alertas en tiempo real al personal y a los sistemas adecuados permite detectar problemas de forma temprana y responder rápidamente a los incidentes operativos.

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

Guía para la implementación

Las cargas de trabajo deben estar equipadas con sistemas de procesamiento y generación de alarmas en tiempo real que permitan mejorar la capacidad de detección de problemas que podrían afectar a la disponibilidad de la aplicación y actúen como desencadenantes de una respuesta automatizada. Las organizaciones pueden realizar el procesamiento y generar alarmas en tiempo

real creando alertas con métricas definidas para recibir notificaciones siempre que ocurran eventos importantes o una métrica supere un umbral.

Amazon CloudWatch le permite crear alarmas de métricas y alarmas compuestas mediante las alarmas de CloudWatch basadas en un umbral estático, la detección de anomalías y otros criterios. Para obtener más información sobre los tipos de alarmas que puede configurar con CloudWatch, consulte la sección de alarmas de la documentación de CloudWatch.

Puede crear vistas personalizadas de las métricas y alertas de los recursos de AWS para sus equipos mediante los paneles de CloudWatch. Las páginas de inicio personalizables de la consola de CloudWatch le permiten supervisar los recursos a través de una única vista de las diferentes regiones.

Las alarmas pueden realizar una o varias acciones, como enviar una notificación a un tema de Amazon SNS, realizar una acción de Amazon EC2 o de Amazon EC2 Auto Scaling o crear un elemento OpsItem o bien un incidente en AWS Systems Manager.

Amazon CloudWatch usa Amazon SNS para enviar notificaciones cuando la alarma cambia de estado, lo que permite que los editores (productores) envíen mensajes a los suscriptores (consumidores). Para obtener más detalles sobre la configuración de las notificaciones de Amazon SNS, consulte Configuración de Amazon SNS.

CloudWatch envía eventos EventBridge siempre que se crea, se actualiza, se elimina o cambia de estado un alarma de CloudWatch. Puede usar EventBridge con estos eventos para crear reglas que realicen acciones, como avisarle cada vez que cambie el estado de una alarma o que activen eventos en la cuenta de forma automática mediante la automatización de Systems Manager.

¿Cuándo debe usar EventBridge o Amazon SNS?

Tanto EventBridge como Amazon SNS se pueden utilizar para desarrollar aplicaciones basadas en eventos, así que la elección de uno u otro dependerá de sus necesidades específicas.

Se recomienda usar Amazon EventBridge si desea crear una aplicación que reaccione a los eventos de sus propias aplicaciones, de aplicaciones SaaS y de servicios de AWS. EventBridge es el único servicio basado en eventos que se integra directamente con socios de SaaS de terceros. EventBridge también ingiere automáticamente eventos de más de 200 servicios de AWS sin necesidad de que los desarrolladores tengan que crear ningún recurso en la cuenta.

EventBridge utiliza una estructura definida basada en JSON para los eventos y le ayuda a crear reglas que se aplican a todo el cuerpo del evento para seleccionar los eventos que se van a reenviar

a un destino. Actualmente, EventBridge admite más de 20 servicios de AWS como destino, entre los que se incluyen AWS Lambda, Amazon SQS, Amazon SNS, Amazon Kinesis Data Streamsy Amazon Data Firehose.

Se recomienda usar Amazon SNS con aplicaciones que necesiten una gran distribución (miles o millones de puntos de conexión). Un patrón habitual que vemos con frecuencia es que los clientes usan Amazon SNS como destino de la regla para filtrar los eventos que necesitan y distribuirlos a diversos puntos de conexión.

Los mensajes no están estructurados y pueden tener cualquier formato. Amazon SNS permite reenviar mensajes a seis tipos diferentes de destinos, como Lambda, Amazon SQS, puntos de conexión HTTP/S, SMS, notificaciones push móviles y correo electrónico. La latencia normal de Amazon SNS es inferior a 30 milisegundos. Hay un gran número de servicios de AWS que envían mensajes de Amazon SNS si se configuran para ello (hay más de 30, incluidos Amazon EC2, Amazon S3y Amazon RDS).

Pasos para la implementación

1. Cree una alarma con las alarmas de Amazon CloudWatch.
    a. Las alarmas de métricas supervisan una única métrica de CloudWatch o una expresión que depende de las métricas de CloudWatch. La alarma activa una o varias acciones en función del valor de la métrica o de la expresión en comparación con un umbral durante varios intervalos de tiempo. La acción puede consistir en enviar una notificación a un tema de Amazon SNS, realizar una acción de Amazon EC2 o de Amazon EC2 Auto Scaling o crear un elemento OpsItem o bien un incidente en AWS Systems Manager.
    b. Una alarma compuesta es una expresión de regla que tiene en cuenta las condiciones de otras alarmas que se han creado. La alarma compuesta solo entra en estado de alarma si se cumplen todas las condiciones de la regla. Las alarmas especificadas en la expresión de la regla de una alarma compuesta pueden ser alarmas de métricas y otras alarmas compuestas. Las alarmas compuestas pueden enviar notificaciones de Amazon SNS cuando su estado cambia y pueden crear elementos OpsItem de Systems Manager o bien incidentes cuando entran en estado de alarma, pero no pueden realizar ninguna acción de Amazon EC2 o Auto Scaling.
2. Configure Notificaciones de Amazon SNS. Al crear una alarma de CloudWatch, puede incluir un tema de Amazon SNS para enviar una notificación cuando la alarma cambie de estado.
3. Cree reglas en EventBridge que coincidan con las alarmas de CloudWatch especificadas. Cada regla admite varios destinos, incluidas las funciones de Lambda. Por ejemplo, puede

definir una alarma que se inicie cuando el espacio disponible en disco se esté agotando, lo que desencadenará una función de Lambda mediante una regla de EventBridge para limpiar el espacio. Para obtener más información sobre los destinos de EventBridge, consulte EventBridge targets.

Recursos

Prácticas recomendadas por Well-Architected:

- REL06-BP01 Supervisar todos los componentes de la carga de trabajo (generación)
- REL06-BP02 Definir y calcular métricas (agregación)
- REL12-BP01 Usar guías de estrategias para investigar los errores

Documentos relacionados:

- Amazon CloudWatch
- CloudWatch Logs insights
- Using Amazon CloudWatch alarms
- Using Amazon CloudWatch dashboards
- Using Amazon CloudWatch metrics
- Setting up Amazon SNS notifications
- detección de anomalías de CloudWatch
- Protección de datos en CloudWatch Logs
- Amazon EventBridge
- Amazon Simple Notification Service

Vídeos relacionados:

- Vídeos sobre observabilidad de reinvent 2022
- AWS re:Invent 2022 - Observability best practices at Amazon

Ejemplos relacionados:

- Taller sobre observabilidad

- Amazon EventBridge to AWS Lambda with feedback control by Amazon CloudWatch Alarms

REL06-BP04 Automatizar las respuestas (procesamiento y alarmas en tiempo real)

Use la automatización para actuar cuando se detecte un evento, por ejemplo, para sustituir componentes defectuosos.

El procesamiento automatizado de las alarmas en tiempo real se implementa para que los sistemas puedan tomar medidas correctivas rápidas e intentar evitar fallos o que el servicio se degrade cuando se activan las alarmas. Entre las respuestas automatizadas a las alarmas, se podría incluir la sustitución de los componentes que fallan, el ajuste de la capacidad de computación, el redireccionamiento del tráfico a hosts, zonas de disponibilidad u otras regiones en buen estado y la notificación a los operadores.

Resultado deseado: se identifican las alarmas en tiempo real y se configura el procesamiento automatizado de las alarmas para invocar las acciones apropiadas que se necesitan para mantener los objetivos de nivel de servicio y los acuerdos de nivel de servicio (SLA). La automatización puede abarcar desde actividades de autorreparación de componentes individuales hasta la conmutación por error de todo el sitio.

Antipatrones usuales:

- No tener un inventario o catálogo claros de las principales alarmas en tiempo real.
- No tener respuestas automatizadas en las alarmas críticas (por ejemplo, cuando los recursos de computación están a punto de agotarse, se produce un escalamiento automático).
- Acciones de respuesta a alarmas contradictorias.
- No tener procedimientos operativos estándar (SOP) que los operadores puedan seguir cuando reciben notificaciones de alerta.
- No supervisar los cambios de configuración, ya que los cambios de configuración no detectados pueden provocar un tiempo de inactividad en las cargas de trabajo.
- No tener una estrategia para deshacer los cambios de configuración no deseados.

Ventajas de establecer esta práctica recomendada: la automatización del procesamiento de alarmas puede mejorar la resiliencia del sistema. El sistema aplica las medidas correctivas automáticamente, lo que reduce las actividades manuales que dan lugar a intervenciones humanas que son más susceptibles a errores. Las operaciones de carga de trabajo cumplen los objetivos de disponibilidad y reducen la interrupción del servicio.

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

Guía para la implementación

Para administrar eficazmente las alertas y automatizar su respuesta, clasifique las alertas en función de su importancia y repercusión, documente los procedimientos de respuesta y planifique las respuestas antes de clasificar las tareas.

Identifique las tareas que requieren medidas específicas (suelen detallarse en los runbooks) y examine todos los runbooks y guías de estrategias para determinar qué tareas se pueden automatizar. Si se pueden definir acciones, estas suelen poderse automatizar. Si las acciones no se pueden automatizar, documente los pasos manuales en un SOP y forme a los operadores sobre ellos. Analice continuamente los procesos manuales en busca de oportunidades de automatización en las que pueda establecer y mantener un plan para automatizar las respuestas a las alertas.

Pasos para la implementación

1. Cree un inventario de alarmas: para obtener una lista de todas las alarmas, puede utilizar la AWS CLI con el comando de Amazon CloudWatch `describe-alarms`. Según el número de alarmas que haya configurado, es posible que tenga que usar la paginación para recuperar un subconjunto de alarmas para cada llamada o, alternativamente, puede usar el SDK de AWS para obtener las alarmas mediante una llamada a la API.
2. Documente todas las acciones de las alarmas: actualice un runbook con todas las alarmas y sus acciones, independientemente de si son manuales o automatizadas. AWS Systems Manager proporciona runbooks predefinidos. Para obtener más información sobre los runbooks, consulte «Creación de sus propios manuales de procedimientos». Para obtener más información sobre cómo ver el contenido del runbook, consulte «View runbook content».
3. Configure y administre las acciones de las alarmas: para cualquiera de las alarmas que requieran una acción, especifique la acción automatizada mediante el SDK de CloudWatch. Por ejemplo, puede cambiar el estado de sus instancias de Amazon EC2 automáticamente en función de una alarma de CloudWatch. Para ello, cree y habilite acciones en una alarma o deshabilite acciones en una alarma.

   También se puede utilizar Amazon EventBridge para responder automáticamente a los eventos del sistema, como los problemas de disponibilidad de las aplicaciones o los cambios en los recursos. Puede crear reglas para indicar qué eventos le interesan y las acciones que se deben realizar cuando un evento coincide con una regla. Entre las acciones que se pueden iniciar automáticamente, se incluye invocar una función AWS Lambda, invocar el comando de ejecución

de Amazon EC2, transmitir el evento a Amazon Kinesis Data Streams y ver cómo se automatiza Amazon EC2 con EventBridge.

4. Procedimientos operativos estándar (SOP): en función de los componentes que tenga su aplicación, AWS Resilience Hub recomienda varias plantillas de SOP. Puede utilizar estos SOP para documentar todos los procesos que debe seguir un operador en caso de que se genere una alerta. También puede crear un SOP basado en recomendaciones de Resilience Hub cuando necesite una aplicación Resilience Hub con una política de resiliencia asociada, así como una evaluación de resiliencia histórica en relación con esa aplicación. Las recomendaciones para su SOP provienen de la evaluación de resiliencia.

   Resilience Hub funciona con Systems Manager para automatizar los pasos de sus SOP al proporcionar una serie de SSMdocumentos que puede utilizar como base para esos SOP. Por ejemplo, Resilience Hub puede recomendar un SOP para añadir espacio en disco basándose en un documento de automatización de SSM existente.

5. Realice acciones automatizadas con Amazon DevOps Guru: puede utilizar Amazon DevOps Guru para supervisar automáticamente los recursos de la aplicación en busca de un comportamiento anómalo y ofrecer recomendaciones específicas para reducir el tiempo de identificación y resolución de problemas. ConDevOps Guru, puede supervisar secuencias de datos operativos casi en tiempo real desde múltiples orígenes, como métricas de Amazon CloudWatch, AWS Config, AWS CloudFormation y AWS X-Ray. También puede utilizar DevOps Guru para crear automáticamente OpsItems en OpsCenter y enviar eventos a EventBridge para una automatización adicional.

Recursos

Prácticas recomendadas relacionadas:

- REL06-BP01 Supervisar todos los componentes de la carga de trabajo (generación)
- REL06-BP02 Definir y calcular métricas (agregación)
- REL06-BP03 Enviar notificaciones (procesamiento y alarmas en tiempo real)
- REL08-BP01 Usar runbooks para actividades estándares como la implementación

Documentos relacionados:

- «AWS Systems Manager Automation»
- «Creating an EventBridge Rule That Triggers on an Event from an AWS Resource»

- Taller sobre observabilidad
- La Amazon Builders' Library: Instrumentación de los sistemas distribuidos para la visibilidad de las operaciones
- «What IsAmazon DevOps Guru?»
- Trabajar con documentos de automatización (guías de estrategias)

Vídeos relacionados:

- AWS re:Invent 2022 - Observability best practices at Amazon (AWS re:Invent 2022: Prácticas recomendadas de observabilidad en Amazon)
- «AWS re:Invent 2020: Automate anything with AWS Systems Manager»
- «Introduction to AWS Resilience Hub»
- «Create Custom Ticket Systems for Amazon DevOps Guru Notifications»
- «Enable Multi-Account Insight Aggregation with Amazon DevOps Guru»

Ejemplos relacionados:

- Reliability Workshops (Talleres de fiabilidad)
- «Amazon CloudWatch and Systems Manager Workshop»

REL06-BP05 Análisis

Recopile archivos de registros e historiales de métricas y analícelos para identificar tendencias e información sobre las cargas de trabajo.

Amazon CloudWatch Logs Insights es compatible con un lenguaje de consultas sencillo pero potente que puede usar para analizar datos de registro. Amazon CloudWatch Logs también admite suscripciones que permiten a los datos dirigirse de forma fluida hacia Amazon S3, donde podrá usar Amazon Athena para consultar los datos. También es compatible con consultas en una gran variedad de formatos. Consulte Formatos de SerDes y datos compatibles en la Guía del usuario de Amazon Athena para obtener más información. Para los análisis de conjuntos de archivos de registro enormes, puede ejecutar un clúster de Amazon EMR para ejecutar análisis en la escala de los petabytes.

Hay una serie de herramientas proporcionadas por socios de AWS y terceros que permiten la agregación, procesamiento, almacenamiento y análisis. Entre estas herramientas se incluyen New

Relic, Splunk, Loggly, Logstash, CloudHealth y Nagios. Sin embargo, la generación fuera de los registros del sistema y las aplicaciones es exclusiva de cada proveedor de la nube y, a menudo, exclusiva de cada servicio.

Una parte del proceso de monitoreo que a menudo se pasa por alto es la gestión de datos. Necesita determinar los requisitos de retención para supervisar los datos y, luego, aplicar las políticas del ciclo de vida correspondientemente. Amazon S3 permite la administración del ciclo de vida en el nivel del bucket de S3. Esta gestión del ciclo de vida se puede aplicar de manera diferente a diferentes rutas en el bucket. Hacia el final del ciclo de vida, puede realizar la transición de datos a Amazon Glacier para el almacenamiento a largo plazo y vencimiento, una vez alcanzado el final del periodo de retención. La clase de almacenamiento de S3 Intelligent-Tiering está diseñado para optimizar los costos trasladando automáticamente los datos al nivel de acceso más eficiente, sin que se vea afectado el rendimiento ni los gastos generales operativos.

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

Guía para la implementación

- CloudWatch Logs Insights le permite buscar y analizar de forma interactiva sus datos de registro en Amazon CloudWatch Logs.
  - Análisis de los datos de registro con CloudWatch Logs Insights
  - Consultas de ejemplo de Amazon CloudWatch Logs Insights
- Use Amazon CloudWatch Logs para enviar registros a Amazon S3, donde puede usar Amazon Athena para consultar los datos.
  - ¿Cómo analizo mis registros de acceso al servidor de Amazon S3 mediante Athena?
    - Cree una política de ciclo de vida de S3 para su bucket de registros de acceso al servidor. Configure la política de ciclo de vida para que se eliminen periódicamente los archivos de registros. De esta forma, reducirá la cantidad de datos que analiza Athena en cada consulta.
      - ¿Cómo creo una política de ciclo de vida para un bucket de S3?

Recursos

Documentos relacionados:

- Consultas de ejemplo de Amazon CloudWatch Logs Insights
- Análisis de los datos de registro con CloudWatch Logs Insights
- Depuración con Amazon CloudWatch Synthetics y AWS X-Ray

- ¿Cómo creo una política de ciclo de vida para un bucket de S3?
- ¿Cómo analizo mis registros de acceso al servidor de Amazon S3 mediante Athena?
- Taller sobre observabilidad
- La Amazon Builders' Library: Instrumentación de los sistemas distribuidos para la visibilidad de las operaciones

REL06-BP06 Realizar revisiones con frecuencia

Revise frecuentemente cómo está implementada la supervisión de cargas de trabajo y actualícela en función de eventos y cambios importantes.

La supervisión efectiva se basa en métricas empresariales claves. Asegúrese de que estas métricas tengan cabida en su carga de trabajo a medida que cambien las prioridades empresariales.

La auditoría de su supervisión le permite asegurarse de que sabrá cuándo cumple una aplicación con sus objetivos de disponibilidad. El análisis de las causas raíces requiere la capacidad de descubrir qué ha ocurrido cuando se produce un error. AWS facilita servicios que le permiten realizar un seguimiento del estado de sus servicios durante un incidente:

- Amazon CloudWatch Logs: puede almacenar sus registros en este servicio e inspeccionar sus contenidos.
- Amazon CloudWatch Logs Insights: es un servicio totalmente administrado que le permite analizar registros inmensos en segundos. Le ofrece consultas y visualizaciones rápidas e interactivas.
- AWS Config: puede ver qué infraestructura de AWS se ha estado utilizando en diferentes momentos.
- AWS CloudTrail: puede ver qué API de AWS se invocaron en qué momento y desde qué entidad principal.

En AWS, realizamos una reunión semanal para revisar el rendimiento operativo y compartir lo que hemos aprendido entre los equipos. Como hay tantos equipos en AWS, creamos La rueda para elegir al azar una carga de trabajo que revisar. El establecimiento de una cadencia regular para las revisiones de rendimiento operativo y el intercambio de conocimientos mejorará su capacidad para lograr un mayor rendimiento de sus equipos operativos.

Patrones de uso no recomendados comunes:

- Recopilar solo métricas predeterminadas

- Establecer una estrategia de supervisión y no revisarla nunca
- No considerar la supervisión cuando se implementan cambios importantes

Beneficios de establecer esta práctica recomendada: la revisión periódica de la supervisión le permite anticiparse a los posibles problemas en lugar de reaccionar a las notificaciones cuando se produzca un problema previsto.

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

Guía para la implementación

- Cree varios paneles para la carga de trabajo. Debe tener un panel general que contenga las principales métricas del negocio, así como las métricas técnicas que ha identificado como más relevantes para el estado previsto de la carga de trabajo conforme cambie su uso. También debe tener paneles para los distintos niveles y dependencias de la aplicación que puedan inspeccionarse.
  - Uso de paneles de Amazon CloudWatch
- Programe y realice revisiones periódicas de los paneles de cargas de trabajo. Realice una inspección periódica de los paneles. Puede tener diferentes cadencias para el alcance de la inspección.
  - Inspeccione las tendencias en las métricas. Compare los valores de las métricas con los valores históricos para saber si hay tendencias que puedan indicar que algo necesita ser investigado. Algunos ejemplos son un aumento de la latencia, una reducción de la función empresarial principal y un aumento de las respuestas a los errores.
  - Inspeccione valores atípicos o anomalías en las métricas. Los promedios o las medianas pueden ocultar valores atípicos y anomalías. Examine los valores más altos y más bajos durante el período de tiempo e investigue las causas de los valores extremos. Mientras elimina estas causas, la relajación de la definición de «extremo» le permitirá seguir mejorando la sistematicidad del rendimiento de sus cargas de trabajo.
  - Busque cambios bruscos en el comportamiento. Un cambio inmediato en la cantidad o en la dirección de una métrica podría indicar que se ha producido un cambio en la aplicación o factores externos que podrían necesitar la inclusión de métricas adicionales para su seguimiento.

Recursos

Documentos relacionados:

- Consultas de ejemplo de Amazon CloudWatch Logs Insights
- Depuración con Amazon CloudWatch Synthetics y AWS X-Ray
- Taller sobre observabilidad
- La Amazon Builders' Library: Instrumentación de los sistemas distribuidos para la visibilidad de las operaciones
- Uso de paneles de Amazon CloudWatch

REL06-BP07 Supervisar el seguimiento de las solicitudes de principio a fin en todo el sistema

Realice un seguimiento de las solicitudes a medida que se procesan a través de los componentes del servicio para que los equipos de producto puedan analizar y depurar los problemas con mayor facilidad y mejorar el rendimiento.

Resultado deseado: las cargas de trabajo con un rastreo exhaustivo en todos los componentes son fáciles de depurar, lo que mejora el tiempo medio de resolución (MTTR) de los errores y la latencia al simplificar la detección de la causa raíz. El rastreo integral reduce el tiempo necesario para descubrir los componentes afectados y analizar detalladamente las causas raíz de los errores o la latencia.

Patrones comunes de uso no recomendados:

- El rastreo se utiliza para algunos componentes, pero no para todos. Por ejemplo, si no se rastreara AWS Lambda, es posible que los equipos no entendieran con claridad la latencia que producen los arranques en frío en una carga de trabajo con picos.
- Los valores controlados sintéticos o la monitorización de usuarios reales (RUM) no tienen configurado el rastreo. Sin valores controlados ni RUM, la telemetría de interacción con el cliente se omite del análisis del rastreo, lo que da lugar a un perfil de rendimiento incompleto.
- Las cargas de trabajo híbridas incluyen herramientas de rastreo nativas en la nube y de terceros, pero no se han tomado medidas para integrar por completo una única solución de rastreo. En función de la solución de rastreo elegida, se deben utilizar SDK de rastreo nativos en la nube para instrumentar componentes que no sean nativos en la nube o se deben configurar herramientas de terceros para ingerir la telemetría de rastreo nativa en la nube.

Beneficios de establecer esta práctica recomendada: Cuando los equipos de desarrollo reciben alertas sobre los problemas, ven una imagen completa de las interacciones entre los componentes del sistema, incluida la correlación componente por componente con el registro, el rendimiento y los errores. Dado que el rastreo facilita la identificación visual de las causas raíz, se dedica menos tiempo a investigar estas causas. Los equipos que conocen bien las interacciones de los componentes toman decisiones mejores y más rápidas a la hora de resolver problemas. Las decisiones, como cuándo invocar una conmutación por error de recuperación de desastres (DR) o cuál es la mejor forma de implementar las estrategias de autorreparación, se pueden mejorar analizando los rastros de los sistemas y, en última instancia, puede mejorar la satisfacción del cliente con sus servicios.

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

Guía para la implementación

Los equipos que utilizan aplicaciones distribuidas pueden utilizar herramientas de rastreo para establecer un identificador de correlación, recopilar rastros de las solicitudes y crear mapas de servicio de los componentes conectados. Todos los componentes de la aplicación deben incluirse en los rastros de solicitudes, como las puertas de enlace de middleware, los buses de eventos y los clientes del servicio, los componentes de computación y el almacenamiento, incluidos los almacenes de valores clave y las bases de datos. Incluya valores controlados sintéticos y la monitorización de usuarios reales en su configuración de rastreo integral para medir las interacciones y la latencia de los clientes remotos, de modo que pueda evaluar con precisión el rendimiento de sus sistemas en función de sus acuerdos y objetivos de nivel de servicio.

Puede usar los servicios de instrumentación de monitorización de aplicaciones AWS X-Ray y Amazon CloudWatch para ofrecer una visión completa de las solicitudes a medida que pasan por su aplicación. X-Ray recopila la telemetría de las aplicaciones y le permite visualizarla y filtrarla en cargas útiles, funciones, rastros, servicios y API, y se puede activar para los componentes del sistema sin código o con poco código. La monitorización de aplicaciones de CloudWatch incluye ServiceLens para integrar sus rastros con métricas, registros y alarmas. La monitorización de aplicaciones de CloudWatch también incluye elementos sintéticos para monitorizar los puntos de conexión y las API, así como la monitorización de usuarios reales para instrumentar los clientes de sus aplicaciones web.

Pasos para la implementación

- Use AWS X-Ray en todos los servicios nativos admitidos, como Amazon S3, AWS Lambda y Amazon API Gateway. Estos servicios de AWS habilitan X-Ray con conmutadores de

configuración que utilizan la infraestructura como código, SDK de AWS o la Consola de administración de AWS.

- Aplicaciones de instrumentos AWS Distro for Open Telemetry y X-Ray o agentes de recopilación de terceros.
- Revise en la Guía para desarrolladores de AWS X-Ray la implementación específica del lenguaje de programación. En estas secciones de la documentación, se detalla cómo instrumentar las solicitudes HTTP, las consultas SQL y otros procesos específicos del lenguaje de programación de su aplicación.
- Utilice el rastreo de X-Ray para los valores controlados de Amazon CloudWatch Synthetics y Amazon CloudWatch RUM para analizar la ruta de la solicitud desde el cliente de su usuario final a través de su infraestructura posterior de AWS.
- Configure métricas y alarmas de CloudWatch en función del estado de los recursos y la telemetría de valores controlados para que los equipos reciban alertas de los problemas rápidamente y, a continuación, puedan analizar en profundidad los rastros y los mapas de servicio con ServiceLens.
- Habilite la integración de X-Ray con herramientas de rastreo de terceros, como Datadog, New Relico Dynatrace si utiliza herramientas de terceros para su solución de rastreo principal.

Recursos

Prácticas recomendadas relacionadas:

- REL06-BP01 Supervisar todos los componentes de la carga de trabajo (generación)
- REL11-BP01 Supervisar todos los componentes de la carga de trabajo para detectar errores

Documentos relacionados:

- ¿Qué es AWS X-Ray?
- Amazon CloudWatch: Application Monitoring (Monitorización de aplicaciones)
- Debugging with Amazon CloudWatch Synthetics and AWS X-Ray (Depuración con Amazon CloudWatch Synthetics y AWS X-Ray)
- La Amazon Builders' Library: Instrumentación de los sistemas distribuidos para la visibilidad de las operaciones
- Integrating AWS X-Ray with other AWS services (Integración de AWS X-Ray con otros servicios de AWS)

- AWS Distro for OpenTelemetry and AWS X-Ray (AWS Distro for OpenTelemetry y AWS X-Ray)
- Amazon CloudWatch: Using synthetic monitoring (Uso de la monitorización sintética)
- Amazon CloudWatch: Use CloudWatch RUM (Uso de CloudWatch RUM)
- Set up Amazon CloudWatch synthetics canary and Amazon CloudWatch alarm (Configuración del valor controlado en Amazon CloudWatch Synthetics y la alarma de Amazon CloudWatch)
- Availability and Beyond: Understanding and Improving the Resilience of Distributed Systems on AWS (Disponibilidad y más allá: comprender y mejorar la resiliencia de sistemas distribuidos en AWS)

Ejemplos relacionados:

- Taller sobre observabilidad

Vídeos relacionados:

- AWS re:Invent 2022 - How to monitor applications across multiple accounts (Cómo monitorizar aplicaciones en varias cuentas)
- How to Monitor your AWS Applications (Cómo monitorizar sus aplicaciones de AWS)

Herramientas relacionadas:

- AWS X-Ray
- Amazon CloudWatch
- Amazon Route 53

FIABILIDAD 7. ¿Cómo diseña su carga de trabajo para que se adapte a los cambios en la demanda?

Una carga de trabajo escalable proporciona elasticidad para agregar y eliminar recursos de forma automática a fin de que coincidan estrechamente con la demanda actual en cualquier momento dado.

Prácticas recomendadas

- REL07-BP01 Usar la automatización al obtener o escalar recursos
- REL07-BP02 Obtener recursos tras detectar un impedimento en una carga de trabajo

- REL07-BP03 Obtener recursos tras detectar que se necesitan más recursos para una carga de trabajo
- REL07-BP04 Realizar pruebas de la carga de trabajo

REL07-BP01 Usar la automatización al obtener o escalar recursos

Cuando reemplace recursos deteriorados o escale su carga de trabajo, automatice el proceso mediante el uso de servicios de AWS administrados, como Amazon S3 y AWS Auto Scaling. También puede utilizar herramientas de terceros y los SDK de AWS para automatizar el escalado.

Los servicios de AWS administrados incluyen Amazon S3, Amazon CloudFront, AWS Auto Scaling, AWS Lambda, Amazon DynamoDB, AWS Fargate y Amazon Route 53.

AWS Auto Scaling le permite detectar y reemplazar las instancias deterioradas. También le permite crear planes de escalado para los recursos, como instancias Amazon EC2 y flotas de spot, tareas de Amazon ECS , tablas e índices de Amazon DynamoDB y réplicas de Amazon Aurora .

Al escalar las instancias EC2, asegúrese de que utiliza varias zonas de disponibilidad (preferiblemente tres como mínimo) y agregue o elimine capacidad para mantener el equilibrio entre estas zonas de disponibilidad. Las tareas de ECS o los pods de Kubernetes (cuando se utiliza Amazon Elastic Kubernetes Service) también se deben distribuir en varias zonas de disponibilidad.

Al utilizar AWS Lambda, las instancias se escalan automáticamente. Cada vez que se recibe una notificación de evento para su función, AWS Lambda localiza rápidamente la capacidad libre en su flota de computación y ejecuta su código hasta la simultaneidad asignada. Debe asegurarse de que la simultaneidad necesaria está configurada en Lambda específico y en su Service Quotas.

Amazon S3 se escala automáticamente para gestionar las altas tasas de solicitudes. Por ejemplo, su aplicación puede alcanzar al menos 3500 solicitudes PUT/COPY/POST/DELETE o 5500 solicitudes GET/HEAD por segundo y prefijo en un bucket. No hay límites en el número de prefijos de un bucket. Puede aumentar su rendimiento de lectura o escritura si ejecuta en paralelo las lecturas. Por ejemplo, si crea diez prefijos en un bucket de Amazon S3 para ejecutar en paralelo las lecturas, podría escalar su rendimiento de lectura a 55 000 solicitudes de lectura por segundo.

Configure y use Amazon CloudFront o una red de entrega de contenido (CDN) de confianza. Una CDN puede proporcionar tiempos de respuesta más rápidos al usuario final y puede servir solicitudes de contenido desde la caché, lo que reduce la necesidad de escalar su carga de trabajo.

Patrones de uso no recomendados comunes:

- Implementar grupos de escalado automático para la reparación automatizada, pero no implementar la elasticidad.
- Utilizar el escalado automático para responder a los grandes aumentos de tráfico.
- Desplegar aplicaciones con un alto nivel de estado, lo que elimina la opción de la elasticidad.

Beneficios de establecer esta práctica recomendada: la automatización elimina la posibilidad de cometer errores manuales al desplegar y retirar los recursos. La automatización elimina el riesgo de sobrecostes y de denegación de servicio debido a la lentitud de respuesta en las necesidades de despliegue o retirada.

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

Guía para la implementación

- Configure y use AWS Auto Scaling. Esto supervisa sus aplicaciones y ajusta automáticamente la capacidad para mantener un rendimiento constante y predecible al menor coste posible. Con AWS Auto Scaling, puede configurar el escalado de aplicaciones para múltiples recursos en varios servicios.
  - ¿Qué es AWS Auto Scaling?
    - Configure el escalado automático en sus instancias y flotas de spot de Amazon EC2, tareas de Amazon ECS, tablas e índices de Amazon DynamoDB, réplicas de Amazon Aurora y dispositivos de AWS Marketplace según corresponda.
      - Administración automática de la capacidad de rendimiento con el escalado automático de DynamoDB
        - Use las operaciones de la API de servicio para especificar las alarmas, las políticas de escalado, los tiempos de calentamiento y los tiempos de enfriamiento.
- Use Elastic Load Balancing. Los equilibradores de carga pueden distribuir la carga por ruta o por conectividad de red.
  - ¿Qué es Elastic Load Balancing?
    - Application Load Balancers puede distribuir la carga por ruta.
      - ¿Qué es un Application Load Balancer?
        - Configure un Application Load Balancer para distribuir el tráfico entre diferentes cargas de trabajo en función de la ruta que corresponde al nombre de dominio.
        - Los Application Load Balancers se pueden utilizar para distribuir las cargas de manera que se integren con AWS Auto Scaling para administrar la demanda.

        - Usar un equilibrador de carga con un grupo de escalado automático
    - Los equilibradores de carga de red pueden distribuir la carga por conexión.
        - ¿Qué es un equilibrador de carga de red?
            - Configure un equilibrador de carga de red para distribuir el tráfico entre diferentes cargas de trabajo mediante TCP o para tener un conjunto constante de direcciones IP para la carga de trabajo.
            - Los equilibradores de carga de red se pueden utilizar para distribuir la carga de manera que se integre con AWS Auto Scaling para administrar la demanda.
- Use un proveedor de DNS de alta disponibilidad. Los nombres DNS permiten a sus usuarios acceder a sus cargas de trabajo con nombres en lugar de direcciones IP. Esta información se distribuye en un ámbito definido, normalmente de forma global para los usuarios de la carga de trabajo.
    - Utilice Amazon Route 53 o un proveedor de DNS de confianza.
        - ¿Qué es Amazon Route 53?
    - Use Route 53 para administrar las distribuciones de CloudFront y los equilibradores de carga.
        - Determine los dominios y subdominios que va a administrar.
        - Cree conjuntos de registros apropiados mediante registros ALIAS o CNAME.
            - Trabajar con los registros
- Utilice la red global de AWS para optimizar la ruta desde sus usuarios a sus aplicaciones. AWS Global Accelerator supervisa continuamente el estado de los puntos de conexión de sus aplicaciones y redirige el tráfico a los puntos en estado correcto en menos de 30 segundos.
    - AWS Global Accelerator es un servicio que mejora la disponibilidad y el rendimiento de sus aplicaciones con usuarios locales o globales. Proporciona direcciones IP estáticas que actúan como punto de entrada fijo a los puntos de conexión de aplicaciones en una o varias Regiones de AWS, como sus Application Load Balancers, equilibradores de carga de red o instancias Amazon EC2.
        - ¿Qué es AWS Global Accelerator?
- Configure y use Amazon CloudFront o una red de entrega de contenido (CDN) de confianza. Una red de entrega de contenido puede ofrecer tiempos de respuesta más rápidos a los usuarios finales y atender solicitudes de contenido que pueden provocar un escalado innecesario de las cargas de trabajo.
    - ¿Qué es Amazon CloudFront?

- Configure distribuciones de Amazon CloudFront para sus cargas de trabajo o utilice una CDN de terceros.
  - Puede limitar el acceso a sus cargas de trabajo para que solo sean accesibles desde CloudFront mediante los intervalos de IP para CloudFront en sus grupos de seguridad de puntos de conexión o políticas de acceso.

Recursos

Documentos relacionados:

- Socio de APN: socios que pueden ayudarle a crear soluciones informáticas automatizadas
- AWS Auto Scaling: cómo funcionan los planes de escalado
- AWS Marketplace: productos que pueden usarse con escalo automático
- Administración automática de la capacidad de rendimiento con el escalado automático de DynamoDB
- Usar un equilibrador de carga con un grupo de escalado automático
- ¿Qué es AWS Global Accelerator?
- ¿Qué es Amazon EC2 Auto Scaling?
- ¿Qué es AWS Auto Scaling?
- ¿Qué es Amazon CloudFront?
- ¿Qué es Amazon Route 53?
- ¿Qué es Elastic Load Balancing?
- ¿Qué es un equilibrador de carga de red?
- ¿Qué es un Application Load Balancer?
- Trabajar con los registros

REL07-BP02 Obtener recursos tras detectar un impedimento en una carga de trabajo

Escale recursos de forma retroactiva cuando sea necesario si la disponibilidad se ve afectada para restaurar la disponibilidad de la carga de trabajo.

Primero debe configurar las comprobaciones de estado y los criterios de dichas comprobaciones para indicar cuándo se ve afectada la disponibilidad por falta de recursos. A continuación, notifique al personal pertinente para que escale manualmente el recurso o inicie la automatización, a fin de que el escalamiento se realice de forma automática.

La escala puede ajustarse manualmente para su carga de trabajo (por ejemplo, se puede cambiar el número de instancias de EC2 en un grupo de Auto Scaling o se puede modificar el rendimiento de una tabla de DynamoDB mediante la Consola de administración de AWS o la AWS CLI). Sin embargo, la automatización debe usarse siempre que sea posible (consulte «Usar la automatización al obtener o escalar recursos»).

Resultado deseado: se inician las actividades de escalamiento (de forma automática o manual) para restablecer la disponibilidad al detectar un error o una experiencia del cliente degradada.

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

Guía para la implementación

Implemente la observabilidad y la supervisión en todos los componentes de su carga de trabajo para supervisar la experiencia del cliente y detectar errores. Defina los procedimientos, manuales o automatizados, que escalan los recursos necesarios. Para obtener más información, consulte «REL11-BP01 Supervisar todos los componentes de la carga de trabajo para detectar errores».

Pasos para la implementación

- Defina los procedimientos, manuales o automatizados, que escalan los recursos requeridos.
  - Los procedimientos de escalamiento dependen de cómo estén diseñados los distintos componentes de la carga de trabajo.
  - Estos procedimientos también varían según la tecnología subyacente que se utilice.
    - Los componentes que utilizan AWS Auto Scaling pueden usar planes de escalamiento para configurar un conjunto de instrucciones para escalar los recursos. Si trabaja con AWS CloudFormation o añade etiquetas a recursos de AWS, puede configurar planes de escalamiento para diferentes conjuntos de recursos por aplicación. Auto Scaling proporciona recomendaciones de estrategias de escalamiento personalizadas para cada recurso. Tras crear su plan de ajuste de escalamiento, Auto Scaling combina el escalamiento dinámico y los métodos de escalamiento predictivos para facilitar su estrategia de escalamiento. Para obtener más información, consulte «Cómo funcionan los planes de escalado».
    - Amazon EC2 Auto Scaling verifica que tiene el número correcto de instancias de Amazon EC2 disponibles para gestionar la carga de la aplicación. Debe crear colecciones de instancias de EC2, denominadas grupos de Auto Scaling. Puede especificar el número mínimo y máximo de instancias en cada grupo de Auto Scaling y Amazon EC2 Auto Scaling garantiza que su grupo nunca tenga un tamaño por encima o por debajo de este límite. Para obtener más información, consulte «¿Qué es Amazon EC2 Auto Scaling?»

- El escalamiento automático de Amazon DynamoDB usa el servicio Application Auto Scaling para ajustar dinámicamente la capacidad de rendimiento aprovisionada en su nombre como respuesta a los patrones de tráfico reales. Esto permite que una tabla o un índice secundario global aumente su capacidad de lectura y escritura aprovisionada para afrontar los picos repentinos de tráfico sin limitación. Para obtener más información, consulte «Administración automática de la capacidad de rendimiento con la función Auto Scaling de DynamoDB».

Recursos

Prácticas recomendadas relacionadas:

- «REL07-BP01 Usar la automatización al obtener o escalar recursos»
- REL11-BP01 Supervisar todos los componentes de la carga de trabajo para detectar errores

Documentos relacionados:

- «AWS Auto Scaling: Cómo funcionan los planes de escalado»
- «Administración automática de la capacidad de rendimiento con la función Auto Scaling de DynamoDB»
- What Is Amazon EC2 Auto Scaling? (¿Qué es Amazon EC2 Auto Scaling?)

REL07-BP03 Obtener recursos tras detectar que se necesitan más recursos para una carga de trabajo

Escale recursos de forma proactiva para satisfacer la demanda y evitar que la disponibilidad se vea impactada.

Muchos servicios de AWS se escalan automáticamente para satisfacer la demanda. Si usa instancias de Amazon EC2 o clústeres de Amazon ECS, puede configurar su escalado automático para que se lleve a cabo en función de métricas de uso que se correspondan con la demanda para su carga de trabajo. Para Amazon EC2, el uso medio de la CPU, el recuento de solicitudes al equilibrador de carga o el ancho de banda de la red se pueden usar para escalar (o desescalar) horizontalmente instancias de EC2. Para Amazon ECS, el uso medio de la CPU, el recuento de solicitudes al equilibrador de carga o el uso de memoria se pueden usar para escalar (o desescalar) horizontalmente tareas de ECS. Al utilizar el escalado automático por objetivos en AWS, el escalador automático actúa como un termostato doméstico y agrega o retira recursos para mantener el valor objetivo (por ejemplo, un uso de la CPU del 70 %) que haya especificado.

AWS Auto Scaling también puede llevar a cabo escalado automático predictivo, que utiliza machine learning para analizar la carga de trabajo histórica de cada recurso y predice regularmente la carga futura para los próximos dos días.

La ley de Little ayuda a calcular cuántas instancias de computación (instancias de EC2, funciones Lambda simultáneas, etc.) necesitará.

$R = \lambda W$

L = número de instancias (o simultaneidad media en el sistema)

λ = promedio de la tasa de llegada de solicitudes (solicitudes/s)

W = promedio del tiempo que pasa cada solicitud en el sistema (s)

Por ejemplo, a 100 sps, si cada solicitud tarda 0,5 segundos en procesarse, necesitará 50 instancias para satisfacer la demanda.

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

Guía para la implementación

- Obtenga recursos tras detectar que se necesitan más recursos para una carga de trabajo. Escale recursos de forma proactiva para satisfacer la demanda y evitar que la disponibilidad se vea impactada.
  - Calcule cuántos recursos de computación necesitará (simultaneidad de computación) para afrontar una tasa de solicitudes dada.
    - Presentación de historias sobre la ley de Little
  - Cuando tenga un patrón de uso histórico, configure el escalado programado para el escalado automático de Amazon EC2.
    - Escalado programado para Amazon EC2 Auto Scaling
  - Use el escalado predictivo de AWS.
    - Predictive Scaling for EC2, Powered by Machine Learning

Recursos

Documentos relacionados:

- AWS Auto Scaling: cómo funcionan los planes de escalado
- AWS Marketplace: productos que pueden usarse con Auto Scaling

- Administrar automáticamente la capacidad de rendimiento con Auto Scaling de DynamoDB
- Predictive Scaling for EC2, Powered by Machine Learning
- Escalado programado para Amazon EC2 Auto Scaling
- Presentación de historias sobre la ley de Little
- ¿Qué es Amazon EC2 Auto Scaling?

REL07-BP04 Realizar pruebas de la carga de trabajo

Adopte una metodología de prueba de carga para medir si la actividad de escalado satisface los requisitos de la carga de trabajo.

Es importante realizar pruebas de carga sostenidas. Las pruebas de carga deben descubrir el punto de ruptura y probar el rendimiento de su carga de trabajo. AWS facilita la creación de entornos de prueba temporales que modelan la escala de su carga de trabajo de producción. En la nube, puede crear un entorno de prueba a escala de producción, completar sus pruebas y desmantelar los recursos. Debido a que solo paga por el entorno de prueba cuando se ejecuta, puede simular su entorno en directo por una fracción del coste de las pruebas en las instalaciones.

Las pruebas de carga en producción también deben considerarse como parte de los días de juego en los que se estresa el sistema de producción, durante las horas de menor uso por parte de los clientes, con todo el personal a mano para interpretar los resultados y abordar cualquier problema que surja.

Patrones de uso no recomendados comunes:

- Realizar pruebas de carga en despliegues que no tienen la misma configuración que su producción.
- Realizar pruebas de carga solo en elementos individuales de su carga de trabajo y no en toda ella.
- Realizar pruebas de carga con un subconjunto de solicitudes y no con un conjunto representativo de solicitudes reales.
- Realizar pruebas de carga con un pequeño factor de seguridad por encima de la carga prevista.

Beneficios de establecer esta práctica recomendada: sabrá qué componentes de su arquitectura presentan errores bajo carga y podrá identificar qué métricas se deben vigilar para indicar que se está acercando a esa carga a tiempo para solucionar el problema, con lo que se evitará el impacto de ese error.

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

Guía para la implementación

- Realice pruebas de carga para identificar qué aspecto de su carga de trabajo indica que debe agregar o eliminar capacidad. Las pruebas de carga deben tener un tráfico representativo similar al que se recibe en producción. Aumente la carga mientras vigila las métricas que ha instrumentado para determinar qué métrica indica cuándo debe agregar o eliminar recursos.
  - Pruebas de carga distribuida en AWS: simular miles de usuarios conectados
    - Identifique la combinación de solicitudes. Es posible que tenga una combinación variada de solicitudes, por lo que deberá tener en cuenta diversos periodos de tiempo a la hora de identificar la combinación de tráfico.
    - Implemente un controlador de carga. Puede utilizar software de código abierto o comercial para implementar un controlador de carga.
    - Realice la prueba de carga inicialmente con una capacidad pequeña. Se ven algunos efectos inmediatos al pasar la carga a una capacidad menor, posiblemente tan pequeña como una instancia o un contenedor.
    - Realice una prueba de carga con una capacidad mayor. Los efectos serán diferentes en una carga distribuida, por lo que debe realizar las pruebas en un entorno lo más parecido posible al del producto.

Recursos

Documentos relacionados:

- Pruebas de carga distribuida en AWS: simular miles de usuarios conectados

FIABILIDAD 8. ¿Cómo implementa los cambios?

Los cambios controlados son necesarios para desplegar nuevas funcionalidades y comprobar que las cargas de trabajo y el entorno operativo ejecuten software conocido y que puedan recibir parches o reemplazos de manera predecible. Si estos cambios no se controlan, puede ser difícil prever su efecto o abordar los problemas que surjan a raíz de ellos.

Prácticas recomendadas

- REL08-BP01 Usar runbooks para actividades estándares como la implementación
- REL08-BP02 Integrar las pruebas funcionales como parte de su despliegue

- REL08-BP03 Integrar las pruebas de resiliencia como parte de su despliegue
- REL08-BP04 Desplegar mediante una infraestructura inmutable
- REL08-BP05 Desplegar cambios con automatización

REL08-BP01 Usar runbooks para actividades estándares como la implementación

Los runbooks son procedimientos predefinidos para obtener resultados concretos. Use runbooks para realizar actividades estándar manuales o automáticas. Algunos ejemplos incluyen implementar una carga de trabajo, aplicarle un parche a dicha carga de trabajo o realizar modificaciones de DNS.

Por ejemplo, se pueden implementar procesos para garantizar la seguridad de la restauración durante los despliegues. Tener la garantía de poder dar marcha atrás en un despliegue sin disrupciones para sus clientes es esencial a la hora de hacer que un servicio sea fiable.

Para los procedimientos de runbooks, empiece por un proceso manual efectivo válido, impleméntelo en el código y desencadene la ejecución automatizada cuando sea oportuno.

Incluso en el caso de cargas de trabajo sofisticadas con un alto nivel de automatización los runbooks siguen siendo útiles para ejecutar días de juego o ajustarse a los exhaustivos requisitos de presentación de informes y auditoría.

Tenga en cuenta que las guías de estrategias se usan en respuesta a incidentes específicos y los runbooks se usan para conseguir resultados determinados. A menudo, los runbooks se usan para actividades rutinarias, mientras que las guías de estrategias se utilizan para responder a eventos no rutinarios.

Patrones de uso no recomendados comunes:

- Realizar cambios imprevistos en la configuración en producción
- Omitir pasos del plan para realizar una implementación más rápida, lo que da lugar a una implementación errónea.
- Realizar cambios sin probar la revocación del cambio

Beneficios de establecer esta práctica recomendada: La planificación eficaz de los cambios aumenta su capacidad de realizar correctamente el cambio, ya que sabrá qué sistemas resultarán afectados. Validar el cambio en los entornos de prueba aumenta la confianza.

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

Guía para la implementación

- Puede obtener respuestas sistemáticas e inmediatas a eventos conocidos si documenta los procedimientos en runbooks.
  - Conceptos del AWS Well-Architected Framework: runbook
- Use el principio de la infraestructura como código para definir la infraestructura. Si usa AWS CloudFormation (o un tercero de confianza) para definir su infraestructura, puede utilizar software de control de versiones para crear versiones y hacer seguimiento de los cambios.
  - Use AWS CloudFormation (o un proveedor tercero de confianza) para definir su infraestructura.
    - ¿Qué es AWS CloudFormation?
  - Cree plantillas singulares y desacopladas, usando buenos principios de diseño de software.
    - Determine los permisos, las plantillas y las partes responsables de su implementación.
      - Control de acceso con AWS Identity and Access Management
    - Use herramientas de control de código, como AWS CodeCommit o las de terceros de confianza, para llevar un control de las versiones.
      - ¿Qué es AWS CodeCommit?

Recursos

Documentos relacionados:

- Socio de APN: socios que pueden ayudarle a crear soluciones de implementación automatizadas
- AWS Marketplace: productos que pueden usarse para automatizar sus implementaciones
- Conceptos del AWS Well-Architected Framework: runbook
- ¿Qué es AWS CloudFormation?
- ¿Qué es AWS CodeCommit?

Ejemplos relacionados:

- Automatización de operaciones con guías de estrategias y runbooks

REL08-BP02 Integrar las pruebas funcionales como parte de su despliegue

Las pruebas funcionales se ejecutan como parte de la implementación automatizada. Si no se satisfacen los criterios de éxito, la canalización se detiene o se revierte.

Estas pruebas se llevan a cabo en un entorno de preproducción, que se lleva a cabo antes de la producción en la canalización. Idealmente, esto se realiza como parte de una canalización de despliegue.

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

Guía para la implementación

- Integre las pruebas funcionales como parte de su despliegue. Las pruebas funcionales se ejecutan como parte de la implementación automatizada. Si no se satisfacen los criterios de éxito, la canalización se detiene o se revierte.
  - Invoque AWS CodeBuild durante la «acción de prueba» de sus canalizaciones de lanzamiento de software modeladas en AWS CodePipeline. Esta función le permite ejecutar fácilmente una gran variedad de pruebas en el código, como pruebas unitarias, análisis de código estático y pruebas de integración.
    - En AWS CodePipeline ahora se pueden hacer pruebas unitarias y de integración personalizadas con AWS CodeBuild
  - Use soluciones de AWS Marketplace para ejecutar pruebas automatizadas como parte de su canalización de entrega de software.
    - Automatización de pruebas de software

Recursos

Documentos relacionados:

- En AWS CodePipeline ahora se pueden hacer pruebas unitarias y de integración personalizadas con AWS CodeBuild
- Automatización de pruebas de software
- ¿Qué es AWS CodePipeline?

REL08-BP03 Integrar las pruebas de resiliencia como parte de su despliegue

Las pruebas de resiliencia (mediante los principios de la ingeniería del caos) se ejecutan como parte de la canalización de despliegue automatizada en un entorno previo a producción.

Estas pruebas se realizan por fases y se ejecutan en la canalización en un entorno previo a la producción. También deben ejecutarse en producción como parte de los días de juego.

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

Guía para la implementación

- Integre las pruebas de resiliencia como parte de su despliegue. Use la ingeniería del caos, la disciplina de poner a prueba una carga de trabajo para generar confianza en su capacidad de resistir condiciones adversas en producción.
  - Las pruebas de resiliencia introducen errores o degradación de los recursos para saber si la carga de trabajo responde con la resiliencia diseñada.
    - Laboratorio de Well-Architected: Nivel 300: pruebas de resiliencia de EC2 RDS y S3
  - Estas pruebas se pueden ejecutar periódicamente en entornos previos a producción en canalizaciones de implementaciones automatizadas.
  - También deben ejecutarse en producción como parte de los días de juego programados.
  - Usando los principios de ingeniería del caos, proponga hipótesis sobre cómo funcionará la carga de trabajo con distintos errores y después pruebe sus hipótesis mediante pruebas de resiliencia.
    - Principios de la ingeniería del caos

Recursos

Documentos relacionados:

- Principios de la ingeniería del caos
- ¿Qué es AWS Fault Injection Simulator?

Ejemplos relacionados:

- Laboratorio de Well-Architected: Nivel 300: pruebas de resiliencia de EC2 RDS y S3

REL08-BP04 Desplegar mediante una infraestructura inmutable

La infraestructura inmutable es un modelo que exige que no haya actualizaciones, parches de seguridad ni cambios de configuración en las cargas de trabajo de producción. Cuando es necesario realizar un cambio, la arquitectura se integra en una nueva infraestructura y se implementa en producción.

Utilice una estrategia de despliegue de infraestructura inmutable para aumentar la fiabilidad, la coherencia y la reproducibilidad de los despliegues de sus cargas de trabajo.

Resultado deseado: con una infraestructura inmutable, no está permitido realizar modificaciones in situ para ejecutar los recursos de infraestructura dentro de una carga de trabajo. En su lugar, cuando es necesario realizar un cambio, se despliega en paralelo un nuevo conjunto de recursos de infraestructura actualizados que contienen todos los cambios que es necesario realizar en los recursos existentes. Este despliegue se valida automáticamente y, si se realiza correctamente, el tráfico se desplaza gradualmente al nuevo conjunto de recursos.

Esta estrategia de despliegue se aplica a las actualizaciones de software, los parches de seguridad, los cambios de infraestructura, las actualizaciones de la configuración y las actualizaciones de las aplicaciones, entre otros.

Antipatrones usuales:

- Implementar cambios in situ en los recursos de infraestructura en ejecución.

Beneficios de establecer esta práctica recomendada:

- Mayor coherencia entre todos los entornos: dado que no hay diferencias en los recursos de infraestructura entre los entornos, se aumenta la coherencia y se simplifican las pruebas.
- Reducción de las desviaciones de la configuración: al reemplazar los recursos de la infraestructura por una configuración conocida y controlada por versiones, la infraestructura se establece en un estado conocido, probado y fiable, lo que evita desviaciones de la configuración.
- Despliegues atómicos fiables: los despliegues se realizan correctamente o no cambia nada, lo que aumenta la coherencia y la fiabilidad en el proceso de despliegue.
- Despliegues simplificados: los despliegues se simplifican porque no tienen que ser compatibles con las mejoras. Las mejoras son simplemente nuevos despliegues.
- Despliegues más seguros con procesos de recuperación y restauración rápidos: los despliegues son más seguros porque la versión activa anterior no se cambia. Puede restaurarla si se detecta algún error.
- Postura de seguridad mejorada: al no permitir cambios en la infraestructura, es posible deshabilitar los mecanismos de acceso remoto (como SSH). Esto reduce el vector de ataque y mejora la postura de seguridad de su organización.

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

Guía para la implementación

Automatización

A la hora de definir una estrategia de despliegue de infraestructura inmutable, se recomienda utilizar la automatización en la medida de lo posible para aumentar la reproducibilidad y minimizar la posibilidad de que se cometan errores humanos. Para obtener más información, consulte «REL08-BP05 Desplegar cambios con automatización» y «Automatización de implementaciones seguras y sin intervención».

Con la infraestructura como código (IaC), los pasos de aprovisionamiento, orquestación y despliegue de la infraestructura se definen de forma programática, descriptiva y declarativa, y se almacenan en un sistema de control del código fuente. El uso de la infraestructura como código simplifica la automatización del despliegue de la infraestructura y ayuda a lograr la inmutabilidad de la infraestructura.

Patrones de despliegue

Cuando es necesario realizar un cambio en la carga de trabajo, la estrategia inmutable de despliegue de la infraestructura requiere el despliegue de un nuevo conjunto de recursos de infraestructura que incluya todos los cambios necesarios. Es importante que este nuevo conjunto de recursos siga un patrón de despliegue que minimice la repercusión en los usuarios. Hay dos estrategias principales para este despliegue:

Despliegue de valores controlados: es la práctica de dirigir a una pequeña cantidad de los clientes a la nueva versión, que normalmente se ejecuta en una instancia de servicio único (la de valor controlado). A continuación, puede analizar en profundidad los errores o los cambios en el comportamiento que se hayan generado. Puede eliminar el tráfico del valor controlado si encuentra problemas críticos y enviar a los usuarios de vuelta a la versión anterior. Si el despliegue se realiza correctamente, puede seguir desplegando a la velocidad que desee, mientras supervisa los cambios en busca de errores, hasta completar el despliegue. AWS CodeDeploy puede configurarse con unos ajustes de despliegue que permitan un despliegue de valores controlados.

Despliegues azul-verde: son similares a los despliegues de valores controlados, excepto que una flota completa de la aplicación se despliega en paralelo. Alterne sus implementaciones en las dos pilas (azul y verde). Una vez más, puede enviar tráfico a la nueva versión y volver a la versión anterior si observa problemas con el despliegue. Normalmente, todo el tráfico se conmuta a la vez. Sin embargo, también puede usar fracciones de tráfico para cada versión para acelerar la adopción de la nueva versión utilizando las capacidades de enrutamiento ponderado por DNS de Amazon

Route 53. AWS CodeDeploy y AWS Elastic Beanstalk se pueden configurar con unos ajustes de despliegue que permitan un despliegue azul-verde.

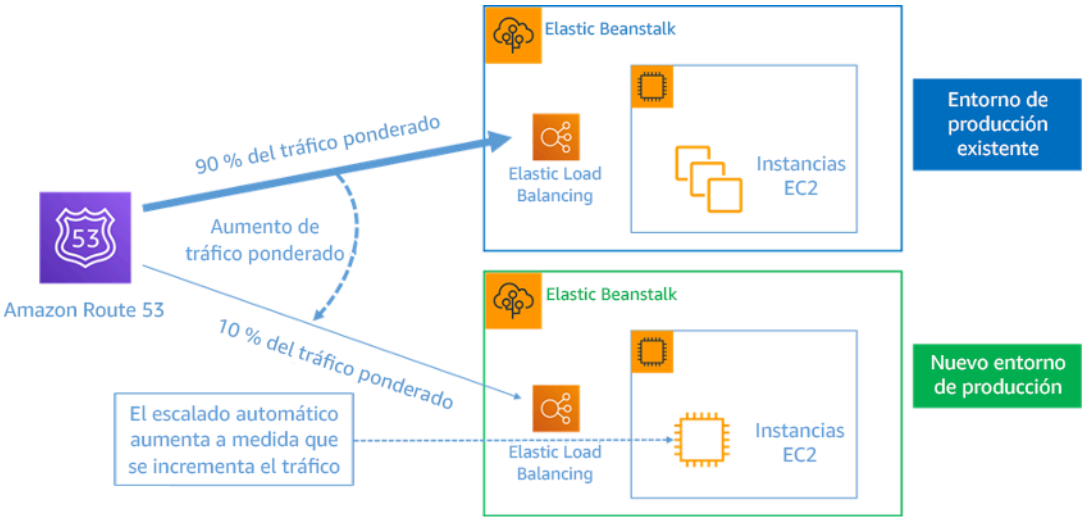


Figura 8: Despliegue azul-verde con AWS Elastic Beanstalk y Amazon Route 53

Detección de desviaciones

Una desviación es cualquier cambio que provoca que un recurso de la infraestructura tenga un estado o una configuración diferente a la esperada. Cualquier tipo de cambio de configuración no administrado va en contra de la noción de infraestructura inmutable y debe detectarse y remediarse para que la infraestructura inmutable se implemente correctamente.

Pasos para la implementación

- No permita que se realicen modificaciones in situ de los recursos de la infraestructura en ejecución.
  - Puede usar AWS Identity and Access Management (IAM) para especificar quién o qué puede acceder a los servicios y recursos de AWS, administrar de forma centralizada los permisos detallados y analizar el acceso para refinar los permisos en AWS.
- Automatice el despliegue de los recursos de la infraestructura para aumentar la reproducibilidad y minimizar la posibilidad de que se cometan errores humanos.
  - Como se describe en el documento técnico «Introduction to DevOps on AWS», la automatización es una piedra angular de los servicios de AWS y es compatible internamente con todos los servicios, características y ofertas.
  - La preparación previade su imagen de máquina de Amazon (AMI) puede acelerar el tiempo de lanzamiento. EC2 Image Builder es un servicio de AWS totalmente administrado que le ayuda

a automatizar la creación, el mantenimiento, la validación, el uso compartido y el despliegue de AMI para Linux o Windows personalizadas, seguras y actualizadas.

- Esto son algunos de los servicios que permiten la automatización:
  - AWS Elastic Beanstalk es un servicio para desplegar y escalar rápidamente aplicaciones web desarrolladas con Java, .NET, PHP, Node.js, Python, Ruby, Go y Docker en servidores conocidos como Apache, NGINX, Passenger e IIS.
  - AWS Proton ayuda a los equipos de plataforma a conectar y coordinar todas las herramientas que sus equipos de desarrollo necesitan para el aprovisionamiento de la infraestructura, los despliegues de código, la supervisión y las actualizaciones. AWS Proton permite una infraestructura automatizada, como el aprovisionamiento de código y el despliegue de aplicaciones basadas en contenedores y sin servidor.
- El uso de la infraestructura como código facilita la automatización del despliegue de la infraestructura y ayuda a lograr la inmutabilidad de la misma. AWS proporciona servicios que permiten la creación, el despliegue y el mantenimiento de la infraestructura de forma programática, descriptiva y declarativa.
  - AWS CloudFormation ayuda a los desarrolladores a crear recursos de AWS de una manera ordenada y predecible. Los recursos se escriben en archivos de texto en formato JSON o YAML. Las plantillas requieren una sintaxis y una estructura específicas que dependen de los tipos de recursos que se crean y administran. Los recursos se crean en JSON o YAML con cualquier editor de código, como AWS Cloud9, se registra en un sistema de control de versiones y, a continuación, CloudFormation crea los servicios especificados de una forma segura y repetible.
  - AWS Serverless Application Model (AWS SAM) es un marco de código abierto que puede utilizar para crear aplicaciones sin servidor. enAWS. AWS SAM se integra con otros servicios de AWS y es una extensión de CloudFormation.
  - Kit de desarrollo de la nube de AWS (AWS CDK) es un marco de desarrollo de software de código abierto para modelar y aprovisionar los recursos de sus aplicaciones en la nube mediante lenguajes de programación conocidos. Puede usar AWS CDK para modelar la infraestructura de las aplicaciones mediante TypeScript, Python, Java y .NET. AWS CDK utiliza CloudFormation en segundo plano para aprovisionar recursos de una forma segura y repetible.
  - API de control de nube de AWS presenta un conjunto común de API de creación, lectura, actualización, eliminación y enumeración (CRUDL) para ayudar a los desarrolladores a administrar su infraestructura en la nube de una forma sencilla y coherente. Las API comunes

de Cloud Control API permiten a los desarrolladores administrar de manera uniforme el ciclo de vida de los servicios de AWS y de terceros.

- Implemente patrones de despliegue que tengan la mínima repercusión en los usuarios.
  - Despliegue de valores controlados:
    - «Configuración de un despliegue de un lanzamiento canary de API Gateway»
    - «Create a pipeline with canary deployments for Amazon ECS using AWS App Mesh»
  - Despliegues azul-verde: en el documento técnico «Blue/Green Deployments on AWS», se describen ejemplos de técnicas para implementar estrategias de despliegue azul-verde.
- Detecte desviaciones de la configuración o el estado. Para obtener más información, consulte «Detección de cambios de configuración no administrados en pilas y recursos».

Recursos

Prácticas recomendadas relacionadas:

- «REL08-BP05 Desplegar cambios con automatización»

Documentos relacionados:

- Automatización de implementaciones seguras y sin intervención
- «Leveraging AWS CloudFormation to create an immutable infrastructure at Nubank»
- Infraestructura como código
- «Implementing an alarm to automatically detect drift in AWS CloudFormation stacks»

Vídeos relacionados:

- «AWS re:Invent 2020: Reliability, consistency, and confidence through immutability»

REL08-BP05 Desplegar cambios con automatización

Las implementaciones y la aplicación de parches se automatizan para eliminar su impacto negativo.

Los cambios en los sistemas de producción son una de las mayores áreas de riesgo para muchas organizaciones. Consideramos que los despliegues son un problema de primera clase que se debe resolver junto con los problemas comerciales que nuestro software aborda. Hoy en día, significa

el uso de la automatización siempre que sea práctico en las operaciones, incluidas las pruebas y el despliegue de cambios, la adición o eliminación de capacidad y la migración de datos. AWS CodePipeline le permite administrar los pasos necesarios para lanzar su carga de trabajo. Esto incluye un estado de despliegue utilizando AWS CodeDeploy para automatizar el despliegue del código de la aplicación en instancias de Amazon EC2, instancias locales, funciones de Lambda sin servidor o servicios de Amazon ECS.

> **Recomendación**
>
> Aunque la sabiduría convencional sugiere que mantenga a los humanos informados sobre los procedimientos operativos más difíciles, le sugerimos que automatice los procedimientos más difíciles por esa misma razón.

Patrones de uso no recomendados comunes:

- Realizar los cambios manualmente
- Omitir los pasos de la automatización a través de flujos de trabajo de emergencia
- No seguir los planes

Beneficios de establecer esta práctica recomendada: El uso de la automatización para implementar todos los cambios elimina la posibilidad de que se introduzcan errores humanos y proporciona la capacidad de probar los cambios antes de modificarlos en producción para garantizar que se cumplan los planes.

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

Guía para la implementación

- Automatice su proceso de despliegue. Las canalizaciones de implementación le permiten invocar pruebas automatizadas, detectar anomalías y detener la canalización en un paso determinado antes de la implementación en producción o revertir automáticamente un cambio.
  - La Amazon Builders' Library: Garantizar la seguridad de las reversiones durante las implementaciones
  - La Amazon Builders' Library: Agilizar el proceso con la entrega continua
    - Use AWS CodePipeline o un producto de terceros de confianza para definir y ejecutar los procesos.

- Configure la canalización para que se inicie cuando se confirme un cambio en su repositorio de código.
  - ¿Qué es AWS CodePipeline?
- Use Amazon Simple Notification Service (Amazon SNS) y Amazon Simple Email Service (Amazon SES) para enviar notificaciones sobre problemas en la canalización o integrar una herramienta de chat de equipo como Amazon Chime.
  - ¿Qué es Amazon Simple Notification Service?
  - ¿Qué es Amazon SES?
  - ¿Qué es Amazon Chime?
  - Automatice los mensajes de chat con webhooks.

Recursos

Documentos relacionados:

- Socio de APN: socios que pueden ayudarle a crear soluciones de implementación automatizadas
- AWS Marketplace: productos que pueden usarse para automatizar sus despliegues
- Automatice los mensajes de chat con webhooks.
- La Amazon Builders' Library: Garantizar la seguridad de las reversiones durante las implementaciones
- La Amazon Builders' Library: Agilizar el proceso con la entrega continua
- ¿Qué es AWS CodePipeline?
- ¿Qué es CodeDeploy?
- AWS Systems Manager Patch Manager
- ¿Qué es Amazon SES?
- ¿Qué es Amazon Simple Notification Service?

Vídeos relacionados:

- AWS Summit 2019: entrega e integración continuas en AWS

Administración de errores

Preguntas

- FIABILIDAD 9. ¿Cómo realiza una copia de seguridad de los datos?
- FIABILIDAD 10. ¿Cómo usa el aislamiento de errores para proteger su carga de trabajo?
- FIABILIDAD 11. ¿Cómo diseña su carga de trabajo para que soporte los errores de los componentes?
- FIABILIDAD 12. ¿Cómo pone a prueba la fiabilidad?
- FIABILIDAD 13. ¿Cómo planifica la recuperación de desastres (DR)?

FIABILIDAD 9. ¿Cómo realiza una copia de seguridad de los datos?

Realice una copia de seguridad de los datos, las aplicaciones y la configuración para satisfacer sus requisitos de objetivos de tiempo de recuperación (RTO) y objetivos de punto de recuperación (RPO).

Prácticas recomendadas

- REL09-BP01 Identificar todos los datos de los que se debe hacer una copia de seguridad y crearla o reproducir los datos a partir de los orígenes
- REL09-BP02 Proteger y cifrar copias de seguridad
- REL09-BP03 Realizar copias de seguridad de los datos automáticamente
- REL09-BP04 Realizar una recuperación periódica de los datos para verificar la integridad de la copia de seguridad y los procesos

REL09-BP01 Identificar todos los datos de los que se debe hacer una copia de seguridad y crearla o reproducir los datos a partir de los orígenes

Conozca y use las funciones de copia de seguridad de los servicios y recursos de datos usados por su carga de trabajo. La mayoría de los servicios ofrecen capacidades para realizar copias de seguridad de los datos de la carga de trabajo.

Resultado deseado: los orígenes de datos se han identificado y clasificado en función del nivel de criticidad. A continuación, establece una estrategia de recuperación de datos basada en el RPO. Esta estrategia supone crear una copia de seguridad de estos orígenes de datos o tener la capacidad de reproducir datos desde otros orígenes. En el caso de la pérdida de datos, la estrategia implementada permite la recuperación o reproducción de datos dentro de los RPO y RTO definidos.

Fase de madurez de la nube: básica

Antipatrones usuales:

- No ser consciente de todos los orígenes de datos para la carga de trabajo y su nivel de criticidad.
- No realizar copias de seguridad de orígenes de datos críticos.
- Realizar copias de seguridad solamente de algunos orígenes de datos sin usar la criticidad como criterio.
- RPO sin definir, o una frecuencia de copias de seguridad que no puede ajustarse al RPO.
- No evaluar si una copia de seguridad es necesaria o si se pueden reproducir datos desde otros orígenes.

Beneficios de establecer esta práctica recomendada: identificar los lugares en los que las copias de seguridad son necesarias e implementar un mecanismo para crear copias de seguridad, o ser capaz de reproducir los datos desde una fuente externa mejora la capacidad de restaurar y recuperar datos durante una interrupción.

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

Guía para la implementación

Todos los almacenes de datos de AWS ofrecen capacidades de copia de seguridad. En los servicios como Amazon RDS y Amazon DynamoDB también se pueden hacer copias de seguridad automatizadas, lo que facilita la recuperación a un momento dado (PITR). De este modo, podrá restaurar una copia de seguridad a cualquier momento hasta cinco minutos (o menos) antes del momento actual. Muchos servicios de AWS ofrecen la capacidad de copiar copias de seguridad en otra Región de AWS. AWS Backup es una herramienta que permite centralizar y automatizar la protección de datos entre servicios de AWS. AWS Elastic Disaster Recovery le permite copiar cargas de trabajo de servidor completas y mantener una protección de datos continua localmente o entre zonas de disponibilidad o regiones, con un objetivo de punto de recuperación (RPO) medido en segundos.

Amazon S3 puede usarse como destino de copias de seguridad para los orígenes de datos autoadministrados y administrados por AWS. Los servicios de AWS como Amazon EBS, Amazon RDS y Amazon DynamoDB tienen capacidades integradas para crear copias de seguridad. También se puede usar software de copias de seguridad de terceros.

Se pueden realizar copias de seguridad de los datos locales en Nube de AWS con AWS Storage Gateway o AWS DataSync. Los buckets de Amazon S3 se pueden usar para almacenar estos datos en AWS. Amazon S3 ofrece varios niveles de almacenamiento, como Amazon Glacier o Amazon Glacier Deep Archive para reducir el coste del almacenamiento de datos.

Es posible que pueda satisfacer las necesidades de recuperación de datos reproduciendo los datos desde otros orígenes. Por ejemplo, los nodos de réplicas de Amazon ElastiCache o bien las réplicas de lectura de Amazon RDS podrían usarse para reproducir datos si se pierde la principal. En casos en los que orígenes como este puedan usarse para cumplir su objetivo de punto de recuperación (RPO) y su objetivo de tiempo de recuperación (RTO), puede que no necesite una copia de seguridad. Otro ejemplo: si trabaja con Amazon EMR, puede que no sea necesario crear copias de seguridad de sus almacenes de datos HDFS, en la medida en que puede reproducir los datos enAmazon EMR desde Amazon S3.

Al seleccionar una estrategia de copia de seguridad, piense en el tiempo que se necesita para recuperar los datos. El tiempo necesario para recuperar datos depende del tipo de copia de seguridad (en el caso de una estrategia de copia de seguridad) o de la complejidad del mecanismo de reproducción de datos. Este tiempo debería ajustarse al RTO de la carga de trabajo.

Pasos para la implementación

1. Identifique todos los orígenes de datos para la carga de trabajo. Los datos se pueden almacenar en diversos recursos, como bases de datos, volúmenes, sistemas de archivos, sistemas de registro y almacenamiento de objetos. Consulte la sección Recursos para encontrar Documentos relacionados sobre distintos servicios deAWS en los que se almacenan los datos y la capacidad de copia de seguridad que proporcionan estos servicios.
2. Clasifique los orígenes de datos en función de su criticidad. Los distintos conjuntos de datos tendrán diferentes niveles de criticidad para una carga de trabajo y, por tanto, distintos requisitos de resiliencia. Por ejemplo, algunos datos podrían ser críticos y requerir un RPO cercano a cero, mientras que otros datos podrían ser menos críticos y tolerar un RPO más alto y cierta pérdida de datos. Del mismo modo, los distintos conjuntos de datos podrían tener también diferentes requisitos en cuanto al RTO.
3. Utilice AWS o servicios de terceros para crear copias de seguridad de los datos. AWS Backup es un servicio administrado que permite la creación de copias de seguridad de diferentes orígenes de datos en AWS. AWS Elastic Disaster Recovery administra la replicación automatizada de datos en menos de un segundo a una Región de AWS. La mayoría de los servicios de AWS también disponen de capacidades nativas para crear copias de seguridad. AWS Marketplace tiene muchas soluciones que ofrecen también estas capacidades. Consulte la sección Recursos que aparece a

continuación para ver información sobre cómo crear copias de seguridad de datos desde distintos servicios de AWS.

4. En el caso de los datos que no tengan copia de seguridad, establezca un mecanismo de reproducción de datos. Puede decidir no crear una copia de seguridad de datos que puedan reproducirse desde otros orígenes y por distintos motivos. Podría darse una situación en la que sea más barato reproducir datos de orígenes cuando sea necesario en lugar de crear una copia de seguridad, ya que podría existir un coste asociado con el almacenamiento de copias de seguridad. Otro ejemplo es cuando la restauración desde una copia de seguridad tarda más tiempo que la reproducción de los datos desde el origen, lo que implica un incumplimiento del RTO. En tales situaciones, sopese los pros y los contras y establezca un proceso bien definido sobre cómo se pueden reproducir los datos desde estos orígenes cuando sea necesaria una recuperación de los datos. Por ejemplo, si ha cargado datos desde Amazon S3 en un almacenamiento de datos (como Amazon Redshift) o un clúster de MapReduce (como Amazon EMR) para analizar dichos datos, esto podría ser un ejemplo de datos que se pueden reproducir desde otros orígenes. Siempre y cuando los resultados de estos análisis se almacenen en algún lugar o sean reproducibles, no sufriría una pérdida de datos por un error en el almacenamiento de datos o el clúster de MapReduce. Otros ejemplos que se pueden reproducir desde el origen son las cachés (como Amazon ElastiCache) o las réplicas de lectura de RDS.
5. Establezca una cadencia de copia de seguridad de los datos. La creación de copias de seguridad de orígenes de datos es un proceso periódico y la frecuencia debería depender del RPO.

Nivel de esfuerzo para el plan de implementación: moderado.

Recursos

Prácticas recomendadas relacionadas:

REL13-BP01 Definir objetivos de recuperación para la inactividad y la pérdida de datos

REL13-BP02 Usar estrategias de recuperación definidas para cumplir los objetivos de recuperación

Documentos relacionados:

- What Is AWS Backup? (¿Qué es AWS Backup?)
- What is AWS DataSync? (¿Qué es AWS DataSync?)
- What is Volume Gateway? (¿Qué es una puerta de enlace de volumen?)
- Socio de APN: socios que pueden ayudar con la copia de seguridad

- AWS Marketplace: products that can be used for backup (AWS Marketplace: productos que pueden usarse para la copia de seguridad)
- Instantáneas de Amazon EBS
- Backing Up Amazon EFS (Copia de seguridad de Amazon EFS)
- Backing up Amazon FSx for Windows File Server (Copia de seguridad de Amazon FSx para Windows File Server)
- Copia de seguridad y restauración de ElastiCache for Redis
- Creating a DB Cluster Snapshot in Neptune (Creación de una instantánea de base de datos en Neptune)
- Crear una instantánea de base de datos
- Creating an EventBridge Rule That Triggers on a Schedule (Creación de una regla de EventBridge que se ejecuta según una programación)
- Replicación entre regiones con Amazon S3
- AWS Backup de EFS a EFS
- Exportación de datos de registro a Amazon S3
- Administración del ciclo de vida de los objetos
- On-Demand Backup and Restore for DynamoDB (Copia de seguridad y restauración bajo demanda para DynamoDB)
- Recuperación a un momento dado en DynamoDB
- Trabajo con instantáneas de índice en Amazon OpenSearch Service
- What is AWS Elastic Disaster Recovery? (¿Qué es AWS Elastic Disaster Recovery?

Vídeos relacionados:

- AWS re:Invent 2021 - Backup, disaster recovery, and ransomware protection with AWS (AWS re:Invent 2021: Copia de seguridad, recuperación de desastres y protección contra ransomware con AWS)
- AWS Backup Demo: Cross-Account and Cross-Region Backup (Demostración de AWS Backup: copia de seguridad entre cuentas y entre regiones)
- AWS re:Invent 2019: Deep dive on AWS Backup, ft. Rackspace (STG341) (AWS re:Invent 2019: Análisis en profundidad en AWS Backup, ft. Rackspace)

Ejemplos relacionados:

- Well-Architected Lab - Implementing Bi-Directional Cross-Region Replication (CRR) for Amazon S3 (Laboratorio de Well-Architected: Implementación de la replicación bidireccional entre regiones (CRR) para Amazon S3)
- Well-Architected Lab - Testing Backup and Restore of Data (Laboratorio de Well-Architected: Probar la copia de seguridad y restauración de los datos)
- Well-Architected Lab - Backup and Restore with Failback for Analytics Workload (Laboratorio de Well-Architected: Copia de seguridad y restauración con conmutación por recuperación para cargas de trabajo de análisis)
- Well-Architected Lab - Disaster Recovery - Backup and Restore (Laboratorio de Well-Architected: Recuperación de desastres, copia de seguridad y restauración)

REL09-BP02 Proteger y cifrar copias de seguridad

Controle y detecte el acceso a las copias de seguridad con autenticación y autorización. Evite que la integridad de los datos de las copias de seguridad se vea comprometida (y detecte los casos en los que así sea) mediante el cifrado.

Antipatrones usuales:

- Tener el mismo acceso a las automatizaciones de las copias de seguridad y restauración que a los datos
- No cifrar las copias de seguridad

Beneficios de establecer esta práctica recomendada: proteger las copias de seguridad impide que se manipulen los datos y el cifrado de los datos impide el acceso a esos datos si se exponen por error.

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

Guía para la implementación

Controle y detecte el acceso a las copias de seguridad con autenticación y autorización, como AWS Identity and Access Management (IAM). Evite que la integridad de los datos de las copias de seguridad se vea comprometida (y detecte los casos en los que así sea) mediante el cifrado.

Amazon S3 admite varios métodos de cifrado de los datos en reposo. Con el cifrado del lado del servidor, Amazon S3 acepta sus objetos como datos sin cifrar y después los cifra a medida que se almacenan. Utilizando el cifrado del cliente, su aplicación de carga de trabajo es la responsable de cifrar los datos antes de que se envíen a Amazon S3. Ambos métodos le permiten utilizar AWS Key

Management Service (AWS KMS) para crear y almacenar la clave de los datos, o puede facilitar la suya propia, de la que será responsable. Con AWS KMS, puede establecer políticas utilizando IAM sobre quién puede acceder a sus claves de datos y datos descifrados y quién no.

Para Amazon RDS, si ha decidido cifrar las bases de datos, sus copias de seguridad estarán cifradas también. Las copias de seguridad de DynamoDB siempre están cifradas. Al usar AWS Elastic Disaster Recovery, todos los datos en tránsito y en reposo están cifrados. Con Elastic Disaster Recovery, los datos en reposo se pueden cifrar utilizando la clave de cifrado de volumen predeterminada de Amazon EBS o una clave personalizada administrada por el cliente.

Pasos para la implementación

1. Usar el cifrado en cada uno de los almacenes de datos. Si los datos de origen están cifrados, la copia de seguridad también estará cifrada.
   - Utilice el cifrado en Amazon RDS. Puede configurar el cifrado en reposo mediante AWS Key Management Service al crear una instancia de RDS.
   - Utilice el cifrado en volúmenes de Amazon EBS. Puede configurar el cifrado predeterminado o especificar una clave única al crear los volúmenes.
   - Utilice el cifrado de Amazon DynamoDB requerido. DynamoDB cifra todos los datos en reposo. Puede utilizar una clave AWS propiedad de AWS KMS o una clave KMS administrada por AWS, especificando una clave que esté almacenada en su cuenta.
   - Cifre los datos almacenados en Amazon EFS. Configure el cifrado cuando cree el sistema de archivos.
   - Configure el cifrado en las regiones de origen y destino. Puede configurar el cifrado en reposo en Amazon S3 con las claves almacenadas en KMS, pero las claves son específicas de la región. Puede especificar las claves de destino cuando configure la replicación.
   - Elija si desea utilizar el cifrado de Amazon EBS para Elastic Disaster Recovery predeterminado o personalizado. Esta opción cifrará sus datos replicados en reposo en los discos de la subred de la zona de preparación y en los discos replicados.
2. Implemente permisos de privilegios mínimos para acceder a las copias de seguridad. Siga las prácticas recomendadas para limitar el acceso a las copias de seguridad, instantáneas y réplicas de acuerdo con las prácticas recomendadas de seguridad.

Recursos

Documentos relacionados:

- AWS Marketplace: products that can be used for backup (AWS Marketplace: productos que pueden usarse para la copia de seguridad)
- Cifrado de Amazon EBS
- Amazon S3: Protección de datos mediante cifrado
- Configuración adicional de CRR: replicación de objetos creados con el cifrado del servidor (SSE) usando las claves de cifrado almacenadas en AWS KMS
- Cifrado en reposo en DynamoDB
- Cifrado de recursos de Amazon RDS
- Cifrado de datos y metadatos en Amazon EFS
- Cifrado para copias de seguridad en AWS
- Administrar las tablas cifradas
- Pilar de seguridad: AWS Well-Architected Framework
- What is AWS Elastic Disaster Recovery? (¿Qué es AWS Elastic Disaster Recovery?

Ejemplos relacionados:

- Well-Architected Lab - Implementing Bi-Directional Cross-Region Replication (CRR) for Amazon S3 (Laboratorio de Well-Architected: Implementación de la replicación bidireccional entre regiones (CRR) para Amazon S3)

REL09-BP03 Realizar copias de seguridad de los datos automáticamente

Configure las copias de seguridad para que se realicen automáticamente mediante el uso de un calendario periódico determinado por el objetivo de punto de recuperación (RPO) o cuando se produzcan cambios en el conjunto de datos. En el caso de los conjuntos de datos críticos con requisitos de pérdida de datos bajos, es necesario realizar una copia de seguridad automática con frecuencia, mientras que en el de los datos menos críticos para los que resultan aceptables ciertas pérdidas, las copias de seguridad pueden ser menos frecuentes.

Resultado deseado: un proceso automatizado que crea copias de seguridad de los orígenes de datos a un ritmo establecido.

Antipatrones usuales:

- Realizar las copias de seguridad manualmente

- Usar recursos que tengan la función de copia de seguridad, pero no incluir la copia de seguridad en la automatización

Beneficios de establecer esta práctica recomendada: la automatización de las copias de seguridad verifica que se realicen con regularidad en función del RPO, y emite una alerta en caso contrario.

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

Guía para la implementación

AWS Backup se puede usar para crear copias de seguridad de los datos automatizadas para varios orígenes de datos de AWS. Es posible realizar copias de seguridad de las instancias de Amazon RDS casi de forma continua cada cinco minutos, y de los objetos de Amazon S3 cada quince minutos, lo que facilita una recuperación a un momento dado (PITR) en un punto específico del historial de copias de seguridad. Para otros orígenes de datos de AWS, como los volúmenes Amazon EBS, las tablas de Amazon DynamoDB o los sistemas de archivos de Amazon FSx, AWS Backup puede ejecutar una copia de seguridad automatizada con una frecuencia que puede llegar a ser de una hora. Estos servicios ofrecen también capacidades de copia de seguridad nativas. Entre los servicios de AWS que ofrecen copia de seguridad automatizada con recuperación a un momento dado se incluyen Amazon DynamoDB, Amazon RDS y Amazon Keyspaces (para Apache Cassandra) —la restauración se puede realizar a un punto temporal específico dentro del historial de copias de seguridad—. La mayoría del resto de servicios de almacenamiento de datos de AWS ofrecen la capacidad de programar copias de seguridad periódicas, con una frecuencia que puede llegar a ser de una hora.

Amazon RDS y Amazon DynamoDB ofrecen copias de seguridad continuas con recuperación a un momento dado. El control de versiones de Amazon S3, una vez activado, es automático. Amazon Data Lifecycle Manager se puede utilizar para automatizar la creación, copia y eliminación de instantáneas de Amazon EBS. También puede automatizar la creación, copia, desuso y anulación de registro de imágenes de máquina de Amazon (AMI) basadas en Amazon EBS y sus instantáneas de Amazon EBS subyacentes.

AWS Elastic Disaster Recovery proporciona replicación continua a nivel de bloque desde el entorno de origen (local o AWS) a la región de recuperación de destino. El servicio crea y administra automáticamente instantáneas de Amazon EBS de un momento dado.

Para obtener una vista centralizada de la automatización y el historial de sus copias de seguridad, AWS Backup proporciona una solución de copia de seguridad totalmente administrada y basada en

políticas. Centraliza y automatiza la copia de seguridad de datos entre varios servicios de AWS en la nube y en el entorno local utilizando AWS Storage Gateway.

De forma adicional al control de versiones, Amazon S3 incluye también replicación. Todo el bucket de S3 se puede replicar automáticamente en otro bucket de la misma Región de AWS o una diferente.

Pasos para la implementación

1. Identifique los orígenes de datos de los que, actualmente, se están haciendo copias de seguridad de forma manual. Para obtener más detalles, consulte REL09-BP01 Identificar todos los datos de los que se debe hacer una copia de seguridad y crearla o reproducir los datos a partir de los orígenes.
2. Determine el RPO de la carga de trabajo. Para obtener más detalles, consulte REL13-BP01 Definir objetivos de recuperación para la inactividad y la pérdida de datos.
3. Utilice una solución de copia de seguridad o un servicio administrado automatizados. AWS Backup es un servicio totalmente administrado que facilita la centralización y automatización de la protección de datos entre servicios de AWS, en la nube y en el entorno local. Mediante el uso de planes de copia de seguridad en AWS Backup, cree reglas que definan de qué recursos se debe hacer copia de seguridad y con qué frecuencia deben crearse. Esta frecuencia debe determinarla el RPO establecido en el paso 2. Para obtener orientación práctica sobre cómo crear copias de seguridad automatizadas con AWS Backup, consulte Testing Backup and Restore of Data (Pruebas de copia de seguridad y restauración de datos). La mayoría de servicios de AWS que almacenan datos ofrecen capacidades de copia de seguridad nativas. Por ejemplo, se puede utilizar RDS para realizar copias de seguridad automatizadas con recuperación a un momento dado (PITR).
4. Para los orígenes de datos no compatibles con un servicio administrado o solución de copia de seguridad automatizada, como los orígenes de datos o las colas de mensajes locales, considere el uso de una solución de terceros de confianza para crear copias de seguridad automatizadas. Como alternativa, puede crear una automatización que se encargue de esto con la AWS CLI o algún SDK. Puede usar AWS Lambda Functions o AWS Step Functions para definir la lógica implicada en la creación de una copia de seguridad de datos y utilizar Amazon EventBridge para ejecutarla a una frecuencia basada en su RPO.

Nivel de esfuerzo para el plan de implementación: bajo.

Recursos

Documentos relacionados:

- Socio de APN: socios que pueden ayudar con la copia de seguridad
- AWS Marketplace: products that can be used for backup (AWS Marketplace: productos que pueden usarse para la copia de seguridad)
- Creating an EventBridge Rule That Triggers on a Schedule (Creación de una regla de EventBridge que se ejecuta según una programación)
- What Is AWS Backup? (¿Qué es AWS Backup?)
- ¿Qué es AWS Step Functions?
- What is AWS Elastic Disaster Recovery? (¿Qué es AWS Elastic Disaster Recovery?

Vídeos relacionados:

- AWS re:Invent 2019: Deep dive on AWS Backup, ft. Rackspace (STG341) (AWS re:Invent 2019: Análisis en profundidad en AWS Backup, ft. Rackspace)

Ejemplos relacionados:

- Well-Architected Lab - Testing Backup and Restore of Data (Laboratorio de Well-Architected: Probar la copia de seguridad y restauración de los datos)

REL09-BP04 Realizar una recuperación periódica de los datos para verificar la integridad de la copia de seguridad y los procesos

Valide que su implementación del proceso de copia de seguridad cumpla con los objetivos de tiempo de recuperación (RTO) y los objetivos de punto de recuperación (RPO) mediante una prueba de recuperación.

Resultado deseado: los datos de las copias de seguridad se recuperan periódicamente utilizando mecanismos bien definidos para verificar que la recuperación sea posible dentro del objetivo de tiempo de recuperación (RTO) determinado para la carga de trabajo. Verifique que la restauración a partir de una copia de seguridad dé como resultado un recurso que contenga los datos originales sin que ninguno de ellos resulte dañado o inaccesible, y una pérdida de datos coherente con el objetivo de punto de recuperación (RPO).

Antipatrones usuales:

- Restaurar una copia de seguridad, pero no consultar ni recuperar ningún dato para comprobar que la restauración sea posible.
- Suponer que existe una copia de seguridad.
- Suponer que la copia de seguridad de un sistema está plenamente operativa y que es posible recuperar datos de ella.
- Suponer que el tiempo de restauración o recuperación de datos de una copia de seguridad entra dentro del RTO para la carga de trabajo.
- Suponer que los datos que contiene la copia de seguridad están dentro del RPO para la carga de trabajo.
- Restaurar cuando sea necesario, sin usar un runbook, o fuera de un procedimiento automatizado.

Beneficios de establecer esta práctica recomendada: comprobar la recuperación de las copias de seguridad verifica que los datos puedan restaurarse cuando sea necesario sin tener que preocuparse por si los datos faltan o están dañados, por si la restauración y la recuperación son o no posibles dentro del RTO para la carga de trabajo y por si la pérdida de datos se ajusta al RPO de la carga de trabajo.

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

Guía para la implementación

La comprobación de la capacidad de copia de seguridad y restauración aumenta la confianza en la capacidad de llevar a cabo estas acciones durante una interrupción. Restaure periódicamente las copias de seguridad en una nueva ubicación y lleve a cabo pruebas para verificar la integridad de los datos. Algunas de las pruebas habituales que deberían realizarse son la comprobación de si todos los datos están disponibles, no están dañados, son accesibles y si la pérdida de datos (si la hay) se ajusta al RPO de la carga de trabajo. Estas pruebas también pueden ayudar a determinar si los mecanismos de recuperación son lo suficientemente rápidos como para tener capacidad para el RTO de la carga de trabajo.

Con AWS, puede crear un entorno de prueba y restaurar sus copias de seguridad para evaluar a las capacidades en cuanto al RTO y al RPO y llevar a cabo pruebas sobre el contenido y la integridad de los datos.

Además, Amazon RDS y Amazon DynamoDB permiten la recuperación a un momento dado (PITR). Mediante la copia de seguridad continua, puede restaurar su conjunto de datos al estado en el que se encontrara en una fecha y hora específicas.

Si todos los datos están disponibles, no están dañados, son accesibles y si la pérdida de datos (si la hay) se ajusta al RPO de la carga de trabajo. Estas pruebas también pueden ayudar a determinar si los mecanismos de recuperación son lo suficientemente rápidos como para tener capacidad para el RTO de la carga de trabajo.

AWS Elastic Disaster Recovery ofrece instantáneas de recuperación a un momento dado continuas de volúmenes de Amazon EBS. A medida que se replican los servidores de origen, los estados de un momento dado se cronifican en el tiempo en función de la política configurada. Elastic Disaster Recovery ayuda a verificar la integridad de estas instantáneas lanzando instancias con fines de prueba y simulacro sin redirigir el tráfico.

Pasos para la implementación

1. Identifique los orígenes de datos de los que se estén haciendo copias de seguridad y dónde se están almacenando dichas copias. Guía para la implementación REL09-BP01 Identificar todos los datos de los que se debe hacer una copia de seguridad y crearla o reproducir los datos a partir de los orígenes.
2. Establezca criterios de validación de datos para cada origen de datos. Los diferentes tipos de datos tendrán distintas propiedades, lo que podría requerir diferentes mecanismos de validación. Considere cómo se podrían validar estos datos antes de contar con la confianza suficiente para usarlos en producción. Algunas formas habituales de validar los datos son usar las propiedades de datos y copias de seguridad como el tipo de datos, el formato, la suma de comprobación, el tamaño o una combinación de ellas con lógica de validación personalizada. Por ejemplo, podría tratarse de una comparación de los valores de las sumas de comprobación entre el recurso restaurado y el origen de datos en el momento en que se creó la copia de seguridad.
3. Establezca RTO y RPO para restaurar los datos sobre la base de la importancia crítica de los datos. Guía para la implementación REL13-BP01 Definir objetivos de recuperación para la inactividad y la pérdida de datos.
4. Evalúe su capacidad de recuperación. Revise su estrategia de copia de seguridad y restauración para comprender si se ajusta a su RTO y RPO, y ajuste la estrategia según sea necesario. Con AWSResilience Hub, puede llevar a cabo una evaluación de su carga de trabajo. La evaluación compara la configuración de su aplicación con la política de resiliencia y notifica si se pueden cumplir los objetivos de RTO y RPO.

5. Realice una restauración de prueba con los procesos establecidos actualmente utilizados en producción para la restauración de datos. Estos procesos dependen de cómo se haya realizado la copia de seguridad del origen de datos, el formato y la ubicación del almacenamiento de la copia de seguridad, o de si los datos se reproducen desde otros orígenes. Por ejemplo, si utiliza un servicio administrado como AWS Backup, podría ser tan sencillo como restaurar la copia de seguridad en un nuevo recurso. Si utilizó AWS Elastic Disaster Recovery puede lanzar un simulacro de recuperación.
6. Valide la recuperación de datos desde el recurso restaurado en función de los criterios que estableciera anteriormente para la validación de datos. ¿Los datos restaurados y recuperados contienen el registro o elemento más reciente en el momento de la copia de seguridad? ¿Estos datos se ajustan al RPO de la carga de trabajo?
7. Mida el tiempo necesario para la restauración y la recuperación y compárelo con el RTO establecido. ¿Este proceso se ajusta al RTO para la carga de trabajo? Por ejemplo, compare las marcas de tiempo del momento en que se inició el proceso de restauración y de cuando se completó la validación de la recuperación para calcular cuánto tarda este proceso. Todas las llamadas a la API de AWS llevan una marca de tiempo y esta información está disponible en AWS CloudTrail. Aunque esta información puede proporcionar detalles sobre cuándo se inició el proceso de restauración, la marca de tiempo final para el momento de finalización de la validación debería quedar registrada mediante su lógica de validación. Si se utiliza un proceso automatizado, se pueden usar servicios como Amazon DynamoDB para almacenar esta información. Además, muchos servicios de AWS proporcionan un historial de eventos que facilita información con marcas de tiempo cuando ocurren determinadas acciones. En AWS Backup, las acciones de copia de seguridad y restauración se denominan trabajos y estos trabajos contienen información con marca de tiempo como parte de estos metadatos, que se pueden utilizar para medir el tiempo necesario para la restauración y la recuperación.
8. Notifique a las partes interesadas si falla la validación de datos o si el tiempo necesario para la restauración y la recuperación supera el RTO establecido para la carga de trabajo. Al implementar la automatización para que haga esto, como en este laboratorio, se pueden usar servicios como Amazon Simple Notification Service (Amazon SNS) para enviar notificaciones push, por ejemplo, por correo electrónico o SMS, a los interesados. Estos mensajes también se pueden publicar en aplicaciones de mensajería, como Amazon Chime, Slack o Microsoft Teams, o usarse para crear tareas como OpsItems con AWS Systems Manager OpsCenter.
9. Automatice este proceso para que se ejecute periódicamente. Por ejemplo, servicios como AWS Lambda o una máquina de estados en AWS Step Functions se pueden usar para automatizar los procesos de restauración y recuperación, y Amazon EventBridge se puede usar para desencadenar este flujo de trabajo de automatización periódicamente como se muestra en el

siguiente diagrama de arquitectura. Descubra cómo automatizar la validación de recuperación de datos con AWS Backup. Además, este laboratorio de Well-Architected contiene una experiencia práctica sobre una forma de llevar a cabo la automatización de varios de los pasos que aparecen aquí.

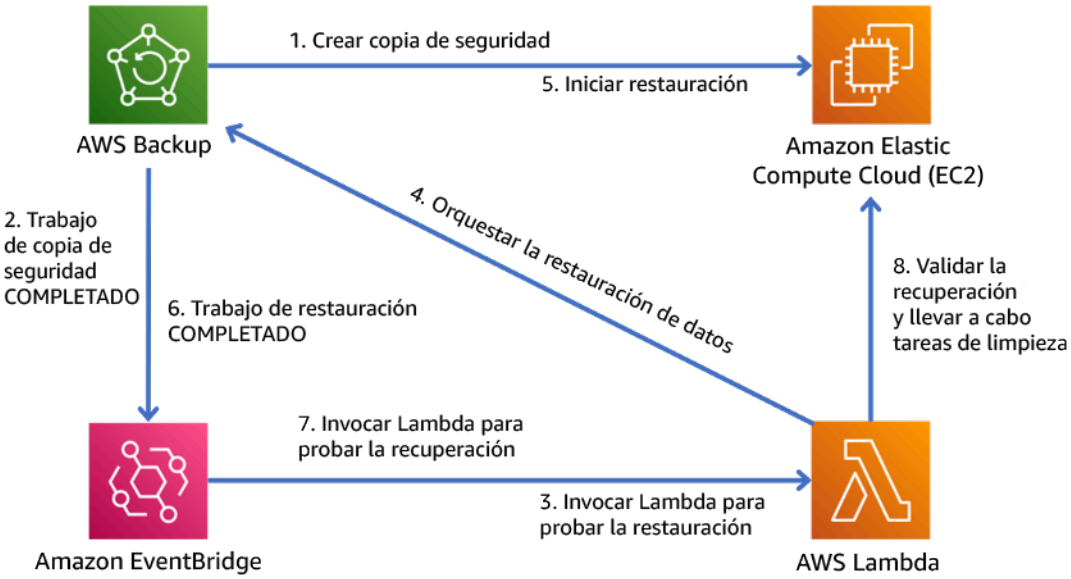


Figura 9. Un proceso de copia de seguridad y restauración automatizado

Nivel de esfuerzo para el plan de implementación: de moderado a alto, en función de la complejidad de los criterios de validación.

Recursos

Documentos relacionados:

- Automate data recovery validation with AWS Backup (Automatizar la validación de recuperación de datos con AWS Backup)
- Socio de APN: socios que pueden ayudar con la copia de seguridad
- AWS Marketplace: products that can be used for backup (AWS Marketplace: productos que pueden usarse para la copia de seguridad)
- Creating an EventBridge Rule That Triggers on a Schedule (Creación de una regla de EventBridge que se ejecuta según una programación)
- On-demand backup and restore for DynamoDB (Copia de seguridad y restauración bajo demanda para DynamoDB)
- What Is AWS Backup? (¿Qué es AWS Backup?)

- ¿Qué es AWS Step Functions?
- What is AWS Elastic Disaster Recovery (¿Qué es AWS Elastic Disaster Recovery?)
- AWS Elastic Disaster Recovery

Ejemplos relacionados:

- Laboratorio de Well-Architected: probar la copia de seguridad y restauración de los datos

FIABILIDAD 10. ¿Cómo usa el aislamiento de errores para proteger su carga de trabajo?

Los límites aislados de los errores acotan el efecto de un error en una carga de trabajo a un número limitado de componentes. Los componentes fuera del límite no resultan afectados por el error. Mediante el uso de varios límites aislados de error, puede acotar el impacto en su carga de trabajo.

Prácticas recomendadas

- REL10-BP01 Implementar la carga de trabajo en varias ubicaciones
- REL10-BP02 Seleccionar las ubicaciones adecuadas para el despliegue en varias ubicaciones
- REL10-BP03 Automatizar la recuperación de los componentes restringidos a una sola ubicación
- REL10-BP04 Usar arquitecturas herméticas para limitar el alcance del impacto

REL10-BP01 Implementar la carga de trabajo en varias ubicaciones

Distribuya los datos y los recursos de la carga de trabajo entre varias zonas de disponibilidad o, si es necesario, entre varias Regiones de AWS. Estas ubicaciones pueden ser tan diversas como sea necesario.

Uno de los principios fundamentales para el diseño de servicios en AWS es evitar puntos únicos de error en la infraestructura física subyacente. Esto nos motiva a desarrollar software y sistemas que utilizan múltiples zonas de disponibilidad y son resistentes a errores de una sola zona. Del mismo modo, los sistemas están diseñados para resistir los errores de un solo nodo de informática, un solo volumen de almacenamiento o una sola instancia de una base de datos. Cuando se desarrolla un sistema que depende de componentes redundantes, es importante asegurarse de que estos componentes funcionen de forma independiente y, en el caso de las Regiones de AWS, de forma autónoma. Los beneficios que se obtienen de los cálculos teóricos de disponibilidad con componentes redundantes solo son válidos si esto se cumple.

Zonas de disponibilidad (AZ)

Las Regiones de AWS tienen varias zonas de disponibilidad diseñadas para ser independientes entre sí. Cada zona de disponibilidad está separada por una distancia física significativa de otras zonas para evitar situaciones de error correlacionadas debido a peligros ambientales como incendios, inundaciones o tornados. Cada zona de disponibilidad también tiene una infraestructura física independiente: conexiones exclusivas para el suministro eléctrico, fuentes de energía de reserva independientes, servicios mecánicos independientes y conectividad de red independiente dentro y fuera de la zona de disponibilidad. Este diseño limita los errores en cualquiera de estos sistemas a la única AZ afectada. Pese a estar separadas geográficamente, las zonas de disponibilidad están situadas en la misma zona regional, lo que permite las redes de alto rendimiento y baja latencia. La totalidad de la Región de AWS (a través de todas las zonas de disponibilidad, que se componen de múltiples centros de datos físicamente independientes) se puede tratar como un único objetivo de despliegue lógico para su carga de trabajo, incluida la capacidad de replicar datos de forma síncrona (por ejemplo, entre bases de datos). De este modo, puede utilizar las zonas de disponibilidad en una configuración activa/activa o activa/en espera.

Las zonas de disponibilidad son independientes y, por lo tanto, la disponibilidad de la carga de trabajo se incrementa cuando esta se diseña para utilizar varias zonas. Algunos servicios de AWS (incluido el plano de datos de instancia Amazon EC2) se despliegan como servicios estrictamente zonales en los que tienen un destino compartido con la zona de disponibilidad en la que se encuentran. Las instancias Amazon EC2 de las demás AZ no se verán afectadas y seguirán funcionando. Del mismo modo, si un error en una zona de disponibilidad provoca el error de una base de datos de Amazon Aurora, es posible que una instancia de Aurora de réplica de lectura en una zona de disponibilidad no afectada se promueva automáticamente a principal. Los servicios de AWS regionales, como Amazon DynamoDB, utilizan internamente varias zonas de disponibilidad en una configuración activa/activa para alcanzar los objetivos de diseño de disponibilidad para ese servicio, sin que sea necesario configurar la ubicación AZ.

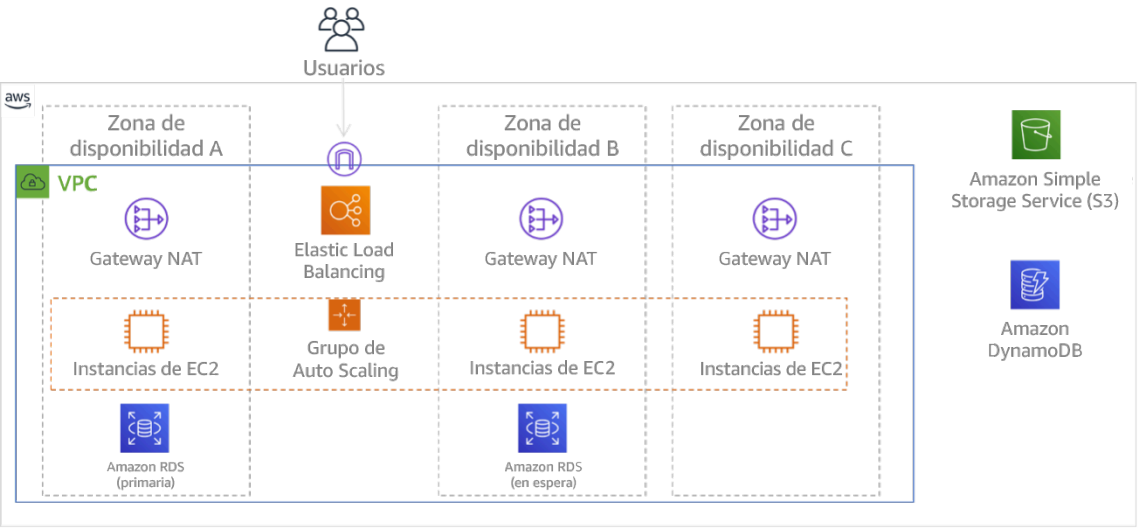


Figura 9: Diagrama que muestra la arquitectura de varios niveles desplegada en tres zonas de disponibilidad. Tenga en cuenta que Amazon S3 y Amazon DynamoDB son siempre Multi-AZ automáticamente. El ELB también se despliega en las tres zonas.

Si bien los planos de control de AWS generalmente ofrecen la capacidad de administrar recursos en toda la región (múltiples zonas de disponibilidad), ciertos planos de control (incluidos Amazon EC2 y Amazon EBS) pueden filtrar los resultados en una única zona de disponibilidad. Al hacerlo, la solicitud se procesa solo en la zona de disponibilidad indicada, lo que reduce la exposición a interrupciones en otras zonas de disponibilidad. Este ejemplo de AWS CLI ilustra la obtención de información de instancias Amazon EC2 solo de la zona de disponibilidad us-east-2c:

```
AWS ec2 describe-instances --filters Name=availability-zone,Values=us-east-2c
```

Zonas locales de AWS

Las zonas locales de AWS actúan de forma similar a las zonas de disponibilidad en su Región de AWS correspondiente en el sentido de que pueden seleccionarse como ubicación de recursos de AWS zonales, como subredes e instancias EC2. Lo que las hace especiales es que no están situadas en la Región de AWS asociada, sino cerca de los grandes centros de población, de la industria y de TI, donde no hay ninguna Región de AWS en la actualidad. Sin embargo, siguen reteniendo un gran ancho de banda y una conexión segura entre las cargas de trabajo de la zona local y las que se ejecutan en la Región de AWS. Debe utilizar las zonas locales de AWS para desplegar las cargas de trabajo más cerca de sus usuarios para cumplir los requisitos de baja latencia.

Red periférica global de Amazon

La red periférica global de Amazon consta de ubicaciones periféricas en ciudades de todo el mundo. Amazon CloudFront utiliza esta red para entregar contenido a los usuarios finales con una latencia menor. AWS Global Accelerator le permite crear sus puntos de conexión de la carga de trabajo en estas ubicaciones periféricas para proporcionar la incorporación a la red global de AWS cerca de sus usuarios. Amazon API Gateway permite que los puntos de conexión de la API optimizados para la periferia utilicen una distribución de CloudFront para facilitar el acceso de los clientes a través de la ubicación periférica más cercana.

Regiones de AWS

Las Regiones de AWS se han diseñado para ser autónomas, por lo que, para utilizar un enfoque multirregión, habría que desplegar copias dedicadas de los servicios en cada región.

Un enfoque multirregión es habitual en las estrategias de recuperación de desastres para cumplir los objetivos de recuperación cuando se producen eventos puntuales a gran escala. Consulte Planificar para la recuperación de desastres (DR) para obtener más información sobre estas estrategias. Sin embargo, aquí nos centramos en la disponibilidad, cuya finalidad es entregar un objetivo de tiempo de actividad medio a lo largo del tiempo. En el caso de los objetivos de alta disponibilidad, una arquitectura multirregión se diseñará generalmente para ser activa/activa, donde cada copia de servicio (en sus respectivas regiones) está activa (sirviendo solicitudes).

> **Recomendación**
>
> Los objetivos de disponibilidad para la mayoría de las cargas de trabajo pueden satisfacerse mediante una estrategia Multi-AZ en una sola Región de AWS. Considere la posibilidad de usar arquitecturas multirregión solo cuando las cargas de trabajo tengan requisitos extremos de disponibilidad, u otros objetivos empresariales, que requieran una arquitectura multirregión.

AWS le proporciona las capacidades para utilizar los servicios entre regiones. Por ejemplo, AWS proporciona una replicación continua y asíncrona de datos mediante la replicación de Amazon Simple Storage Service (Amazon S3), réplicas de lectura de Amazon RDS (incluidas las réplicas de lectura de Aurora) y las tablas globales de Amazon DynamoDB. Con la replicación continua, las versiones de sus datos están disponibles para usarse casi inmediatamente en cada una de sus regiones activas.

Con AWS CloudFormation, puede definir su infraestructura y desplegarla de forma coherente en varias Cuentas de AWS y Regiones de AWS. Y AWS CloudFormation StackSets amplía esta funcionalidad al permitirle crear, actualizar o eliminar pilas de AWS CloudFormation en varias cuentas y regiones con una sola operación. En el caso de los despliegues de instancias Amazon EC2, se utiliza una AMI (imagen de máquina de Amazon) para suministrar información como la configuración del hardware y el software instalado. Puede implementar una canalización del generador de imágenes de Amazon EC2 que cree las ANU que necesita y copiarlas en sus regiones activas. Esto garantiza que estas AMI doradas tiene todo lo que necesita para desplegar y escalar su carga de trabajo en cada nueva región.

Para enrutar el tráfico, tanto Amazon Route 53 como AWS Global Accelerator permiten la definición de políticas que determinen qué usuarios van a cada punto de conexión regional activo. Con Global Accelerator se establece un regulador de tráfico para controlar el porcentaje de tráfico que se dirige a cada punto de conexión de la aplicación. Route 53 es compatible con este enfoque porcentual y también con varias políticas disponibles, incluidas las basadas en la geoproximidad y la latencia. Global Accelerator aprovecha automáticamente la amplia red de servidores periféricos de AWS, para integrar el tráfico en la estructura de red de AWS de lo antes posible, lo que se traduce en menores latencias de solicitudes.

El funcionamiento de todas estas capacidades permite preservar la autonomía de cada región. Existen muy pocas excepciones a este enfoque, incluidos nuestros servicios que proporcionan entrega periférica global (como Amazon CloudFront y Amazon Route 53), junto con el plano de control para el servicio AWS Identity and Access Management (IAM). La gran mayoría de los servicios funcionan completamente en una sola región.

Centro de datos local

En el caso de las cargas de trabajo que se ejecutan en un centro de datos local, diseñe una experiencia híbrida cuando sea posible. AWS Direct Connect proporciona una conexión de red dedicada desde su entorno local a AWS, lo que le permite la ejecución en ambos.

Otra opción es ejecutar la infraestructura y los servicios de AWS localmente mediante AWS Outposts. AWS Outposts es un servicio completamente administrado que extiende la infraestructura de AWS, los servicios de AWS, las API y las herramientas a su centro de datos. La misma infraestructura de hardware que se utiliza en la Nube de AWS se instala en su centro de datos. AWS Outposts se conectan a las Región de AWS. A continuación, puede usar AWS Outposts para respaldar las cargas de trabajo que tengan requisitos de baja latencia o de procesamiento local de datos.

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

Guía para la implementación

- Use varias zonas de disponibilidad y Regiones de AWS. Distribuya los datos y los recursos de la carga de trabajo entre varias zonas de disponibilidad o, si es necesario, entre varias Regiones de AWS. Estas ubicaciones pueden ser tan diversas como sea necesario.
  - Los servicios regionales se implementan en las zonas de disponibilidad.
    - Esto incluye Amazon S3, Amazon DynamoDB y AWS Lambda (cuando no está conectado a una VPC)
  - Implemente su contenedor, instancia y cargas de trabajo basadas en funciones en múltiples zonas de disponibilidad. Use los almacenes de datos multizona, incluidas las memorias caché. Use las características de EC2 Auto Scaling, ubicación de tareas de ECS, configuración de la función AWS Lambda cuando se ejecute en su VPC y clústeres ElastiCache.
    - Utilice subredes en zonas de disponibilidad distintas cuando implemente grupos de Auto Scaling.
      - Ejemplo: distribución de instancias en zonas de disponibilidad
      - Estrategias de asignación de tareas de Amazon ECS
      - Configurar una función AWS Lambda para obtener acceso a los recursos en una Amazon VPC
      - Elección de regiones y zonas de disponibilidad
    - Utilice subredes en zonas de disponibilidad distintas cuando implemente grupos de Auto Scaling.
      - Ejemplo: distribución de instancias en zonas de disponibilidad
    - Utilice los parámetros de colocación de tareas de ECS, especificando grupos de subred de base de datos
      - Estrategias de asignación de tareas de Amazon ECS
    - Utilice subredes en múltiples zonas de disponibilidad cuando configure una función para que se ejecute en su VPC.
      - Configurar una función AWS Lambda para obtener acceso a los recursos en una Amazon VPC
    - Utilice múltiples zonas de disponibilidad con clústeres ElastiCache.
      - Elección de regiones y zonas de disponibilidad

- Si la carga de trabajo se debe desplegar en varias regiones, elija una estrategia multirregión. La mayoría de los requisitos de fiabilidad se pueden satisfacer con una sola Región de AWS que use una estrategia de varias zonas de disponibilidad. Use una estrategia multirregión cuando sea necesario para satisfacer las necesidades del negocio.
  - AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (Patrones de arquitectura para aplicaciones activas-activas en varias regiones) (ARC209-R2)
    - Contar con otra Región de AWS puede añadir otra capa de seguridad en cuanto a la disponibilidad de los datos.
    - Algunas cargas de trabajo tienen requisitos normativos que exigen una estrategia multirregión.
- Evalúe AWS Outposts para su carga de trabajo. Si su carga de trabajo requiere baja latencia en el centro de datos local o si tiene requisitos de procesamiento de datos locales, A continuación, ejecute la infraestructura de AWS y los servicios locales con AWS Outposts.
  - ¿Qué es AWS Outposts?
- Determine si las zonas locales de AWS le ayudan a prestar servicio a sus usuarios. Si tiene requisitos de baja latencia, compruebe si las zonas locales de AWS están cerca de sus usuarios. En caso afirmativo, úselas para implementar las cargas de trabajo más cerca de esos usuarios.
  - Preguntas frecuentes sobre las zonas locales de AWS

Recursos

Documentos relacionados:

- Infraestructura global de AWS
- Preguntas frecuentes sobre las zonas locales de AWS
- Estrategias de asignación de tareas de Amazon ECS
- Elección de regiones y zonas de disponibilidad
- Ejemplo: distribución de instancias en zonas de disponibilidad
- Tablas globales: replicación multirregión con DynamoDB
- Uso de las bases de datos globales de Amazon Aurora
- Serie de blog Creating a Multi-Region Application with AWS Services blog series (Creación de una aplicación multirregión con servicios de AWS)
- ¿Qué es AWS Outposts?

Vídeos relacionados:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (Patrones de arquitectura para aplicaciones activas-activas en varias regiones) (ARC209-R2)
- AWS re:Invent 2019: Innovation and operation of the AWS global network infrastructure (Innovación y funcionamiento de la infraestructura de red global de AWS) (NET339)

REL10-BP02 Seleccionar las ubicaciones adecuadas para el despliegue en varias ubicaciones

Resultado deseado

Para obtener una alta disponibilidad, despliegue siempre (cuando sea posible) sus componentes de carga de trabajo en varias zonas de disponibilidad (AZ), como se muestra en la figura 10. En el caso de cargas de trabajo con requisitos de resiliencia extremos, evalúe cuidadosamente las opciones de una arquitectura multirregión.

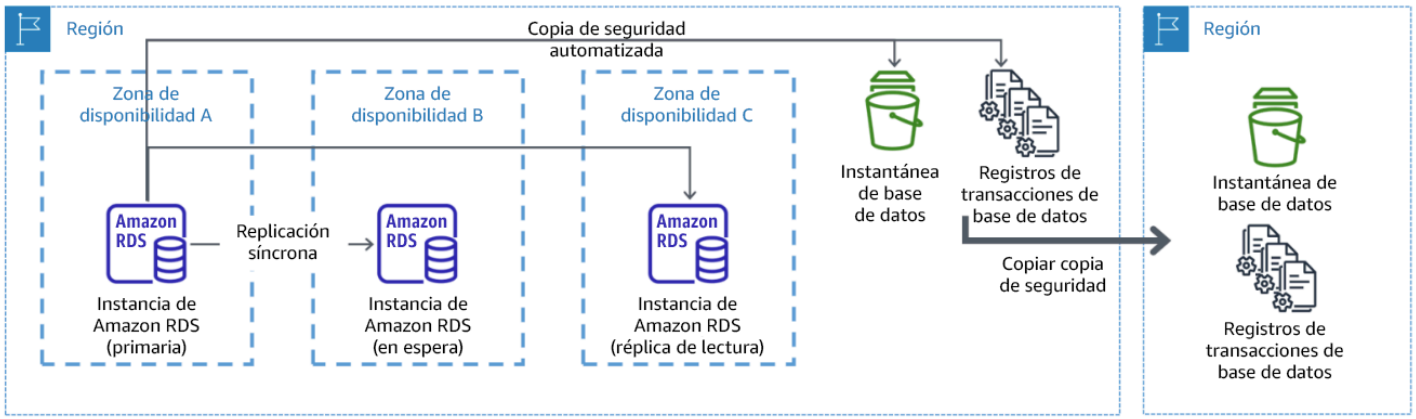


Figura 10: Un despliegue de base de datos multi-AZ resiliente con copia de seguridad en otra región de AWS

Patrones de uso no recomendados comunes

- Elegir el diseño de una arquitectura multirregión cuando una arquitectura multi-AZ satisfaría los requisitos.
- No tener en cuenta las dependencias entre los componentes de la aplicación si los requisitos de resiliencia y multiubicación difieren entre esos componentes.

Beneficios de establecer esta práctica recomendada

Para obtener resiliencia, debe utilizar un enfoque que cree capas de defensa. Una capa protege de las interrupciones más pequeñas y comunes mediante la creación de una arquitectura de alta

disponibilidad con múltiples AZ. Otra capa de defensa está pensada para proteger de eventos poco frecuentes como los desastres naturales generalizados y las interrupciones en el nivel de la región. Esta segunda capa implica la arquitectura de su aplicación para que abarque múltiples Regiones de AWS.

- La diferencia entre una disponibilidad del 99,5 % y del 99,99 % es de más de 3,5 horas al mes. La disponibilidad prevista de una carga de trabajo solo puede alcanzar los «cuatro nueves» si se encuentra en varias AZ.
- Al ejecutar su carga de trabajo en varias AZ, puede aislar las interrupciones de energía eléctrica, refrigeración y redes, y la mayoría de los desastres naturales como incendios e inundaciones.
- La implementación de una estrategia multirregión para su carga de trabajo le ayuda a protegerla de desastres naturales generalizados que afecten a una región geográfica amplia de un país o de errores técnicos de alcance regional. Tenga en cuenta que implementar una arquitectura multirregión puede ser significativamente complejo y no suele ser necesario para la mayoría de las cargas de trabajo.

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

Guía para la implementación

En el caso de un evento de desastre provocado por la interrupción o pérdida parcial de una zona de disponibilidad, la implementación de una carga de trabajo con alta disponibilidad en varias zonas de disponibilidad en una sola Región de AWS contribuye a mitigar los desastres naturales o técnicos. Cada Región de AWS consta de varias zonas de disponibilidad, cada una aislada de los errores de las demás zonas y separadas por una distancia significativa. Sin embargo, en el caso de un evento de desastre que implique el riesgo de perder varios componentes de zona de disponibilidad que están alejados entre sí, debe implementar opciones de recuperación de desastres para mitigar los errores de alcance regional. Para las cargas de trabajo que requieren una resiliencia extrema (infraestructuras críticas, aplicaciones relacionadas con la sanidad, infraestructuras de sistemas financieros, etc.), puede ser necesaria una estrategia multirregión.

Pasos de implementación

1. Evalúe su carga de trabajo y determine si las necesidades de resiliencia se pueden satisface con un enfoque multi-AZ (una sola Región de AWS) o si requieren un enfoque multirregión. La implementación de una arquitectura de multirregión para satisfacer estos requisitos supondrá una complejidad adicional, por lo que deberá considerar detenidamente su caso de uso y sus requisitos. Los requisitos de resiliencia se pueden cumplir casi siempre con una sola Región

de AWS. Tenga en cuenta los siguientes requisitos posibles a la hora de determinar si necesita utilizar varias regiones:

a. Recuperación de desastres (DR): en el caso de un evento de desastre provocado por la interrupción o pérdida parcial de una zona de disponibilidad, la implementación de una carga de trabajo con alta disponibilidad en varias zonas de disponibilidad en una sola Región de AWS contribuye a mitigar los desastres naturales o técnicos. En el caso de un evento de desastre que implique el riesgo de perder varios componentes de zona de disponibilidad que están alejados entre sí, debe implementar opciones de recuperación de desastres en varias regiones para mitigar los desastres naturales o los errores técnicos de alcance regional.

b. Alta disponibilidad: se puede utilizar una arquitectura de multirregión (mediante varias AZ en cada región) para lograr una disponibilidad superior a cuatro nueves (> 99,99 %).

c. Localización de pilas: al desplegar una carga de trabajo para una audiencia global, puede desplegar pilas localizadas en diferentes Regiones de AWS para atender a las audiencias de esas regiones. La localización puede incluir el idioma, la moneda y los tipos de datos almacenados.

d. Proximidad a los usuarios: al desplegar una carga de trabajo para una audiencia global, puede reducir la latencia si despliega las pilas en Regiones de AWS cerca de donde están los usuarios finales.

e. Residencia de los datos: algunas cargas de trabajo están sujetas a requisitos de residencia de datos, en los que los datos de ciertos usuarios deben permanecer dentro de las fronteras de un país específico. En función de la normativa en cuestión, puede optar por desplegar una pila completa, o solo los datos, en la Región de AWS en esas fronteras.

2. A continuación, se presentan algunos ejemplos de la funcionalidad multi-AZ proporcionada por los servicios de AWS:

   a. Para proteger las cargas de trabajo que utilizan EC2 o ECS, despliegue un equilibrador de carga elástico ante los recursos de computación. Elastic Load Balancing proporciona la solución para detectar las instancias en zonas con estado incorrecto y enrutar el tráfico a las que lo tienen correcto.

      i. Introducción a Application Load Balancers

      ii. Introducción a los equilibradores de carga de red

   b. En el caso de las instancias EC2 que ejecutan software estándar comercial y que no admiten el equilibrio de carga, puede conseguir una forma de tolerancia a errores mediante la implementación de una metodología de recuperación de desastres multi-AZ.

   i. the section called “REL13-BP02 Usar estrategias de recuperación definidas para cumplir los objetivos de recuperación”

c. Para las tareas de Amazon ECS, despliegue su servicio de forma homogénea en tres zonas de disponibilidad para lograr un equilibrio entre la disponibilidad y el coste.

   i. Amazon ECS availability best practices | Containers (Prácticas recomendadas de disponibilidad de Amazon ECS | Contenedores

d. En el caso de Aurora Amazon RDS, puede elegir Multi-AZ como una opción de configuración. En caso de error de la instancia de la base de datos principal, Amazon RDS promociona automáticamente una base de datos en espera para recibir el tráfico en otra zona de disponibilidad. También se pueden crear réplicas de lectura multirregión para mejorar la resiliencia.

   i. Despliegues multi-AZ de Amazon RDS

   ii. Creación de una réplica de lectura en una Región de AWS diferente

3. A continuación, se presentan algunos ejemplos de la funcionalidad multirregión proporcionada por los servicios de AWS:

a. Para las cargas de trabajo de Amazon S3, en las que la disponibilidad multi-AZ la proporciona automáticamente el servicio, considere la posibilidad de utilizar puntos de acceso multirregión si se necesita un despliegue multirregión.

   i. Puntos de acceso multirregión en Amazon S3

b. En el caso de las tablas de DynamoDB, en las que el servicio proporciona automáticamente la disponibilidad multi-AZ, puede convertir fácilmente las tablas existentes en tablas globales para aprovechar las ventajas de múltiples regiones.

   i. Convert Your Single-Region Amazon DynamoDB Tables to Global Tables (Convierta sus tablas de Amazon DynamoDB de una sola región en tablas globales)

c. Si su carga de trabajo está encabezada por Application Load Balancers o equilibradores de carga de red, use AWS Global Accelerator para mejorar la disponibilidad de su aplicación mediante el direccionamiento del tráfico a varias regiones que contengan puntos de conexión con el estado correcto.

   i. Endpoints for standard accelerators in AWS Global Accelerator - AWS Global Accelerator (Puntos de conexión para aceleradores estándar en AWS Global Accelerator - AWS Global Accelerator) (amazon.com)

d. En el caso de las aplicaciones que utilizan AWS EventBridge, considere la posibilidad de utilizar buses entre regiones para reenviar los eventos a otras Regiones que seleccione.

   i. Sending and receiving Amazon EventBridge events between Regiones de AWS (Envío y recepción de eventos de Amazon EventBridge entre Regiones de AWS)

e. En el caso de las bases de datos de Amazon Aurora, considere de usar bases de datos globales de Aurora, que abarcan varias regiones de AWS. Los clústeres existentes pueden modificarse para agregar también nuevas regiones.

   i. Introducción a las bases de datos globales de Amazon Aurora

f. Si su carga de trabajo incluye claves de cifrado de AWS Key Management Service (AWS KMS), considere si las claves multirregión son adecuadas para su aplicación.

   i. Claves multirregión en AWS KMS

g. Para conocer otras características de servicios de AWS, consulte esta serie de blog en Serie Creating a Multi-Region Application with AWS Services blog series (Creación de una aplicación multirregión con servicios de AWS)

Nivel de esfuerzo para el plan de implementación: De moderado a alto

Recursos

Documentos relacionados:

- Serie Creating a Multi-Region Application with AWS Services blog series (Creación de una aplicación multirregión con servicios de AWS)
- Disaster Recovery (DR) Architecture on AWS, Part IV: Multi-site Active/Active (Arquitectura de recuperación de desastres (DR) en AWS, parte IV: activa-activa multisitio)
- Infraestructura global de AWS
- Preguntas frecuentes sobre las zonas locales de AWS
- Disaster Recovery (DR) Architecture on AWS, Part I: Strategies for Recovery in the Cloud (Arquitectura de recuperación de desastres (DR) en AWS, parte I: estrategias de recuperación en la nube)
- La recuperación de desastres es diferente en la nube
- Tablas globales: replicación multirregión con DynamoDB

Vídeos relacionados:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (Patrones de arquitectura para aplicaciones activas-activas en varias regiones) (ARC209-R2)

- Auth0: arquitectura de alta disponibilidad en varias regiones que se amplía a más de 1500 millones de inicios de sesión en un mes con conmutación por error automatizada

Ejemplos relacionados:

- Disaster Recovery (DR) Architecture on AWS, Part I: Strategies for Recovery in the Cloud (Arquitectura de recuperación de desastres (DR) en AWS, parte I: estrategias de recuperación en la nube)
- DTCC consigue un nivel de resiliencia mayor del que podría obtener localmente
- Expedia Group usa una arquitectura de varias regiones y varias zonas de disponibilidad con un servicio DNS propio para agregar resiliencia a las aplicaciones
- Uber: recuperación de desastres para Kafka en varias regiones
- Netflix: estrategia activa-activa para la resiliencia multirregional
- Cómo creamos Data Residency for Atlassian Cloud
- Intuit TurboTax se ejecuta en dos regiones

REL10-BP03 Automatizar la recuperación de los componentes restringidos a una sola ubicación

Si los componentes de la carga de trabajo solo se pueden ejecutar en una zona de disponibilidad o en el centro de datos local, implemente la capacidad de volver a crear la carga de trabajo de acuerdo con los objetivos de recuperación definidos.

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

Guía para la implementación

Si la práctica recomendada de desplegar la carga de trabajo en varias ubicaciones no es posible por limitaciones tecnológicas, debe implementar una ruta alternativa hacia la resiliencia. Debe automatizar la capacidad de recrear la infraestructura necesaria, reimplementar las aplicaciones y recrear los datos necesarios para estos casos.

Por ejemplo, Amazon EMR lanza todos los nodos para un clúster determinado en la misma zona de disponibilidad, porque la ejecución de un clúster en la misma zona mejora el rendimiento de los flujos de trabajo, ya que ofrece una velocidad de acceso a los datos más alta. Si este componente resulta necesario para la resiliencia de la carga de trabajo, debe tener una forma de volver a desplegar el clúster y sus datos. Además, para Amazon EMR, debería aprovisionar la redundancia de formas diferentes al uso de multi-AZ. Puede aprovisionar varios nodos. Con el sistema de archivos EMR

(EMRFS), los datos en EMR se pueden almacenar en Amazon S3, lo que a su vez puede replicarse entre varias zonas de disponibilidad o Regiones de AWS.

De modo similar, en el caso de Amazon Redshift, aprovisiona de forma predeterminada el clúster en una zona de disponibilidad seleccionada al azar dentro de la Región de AWS que haya seleccionado. Todos los nodos del clúster se aprovisionan en la misma zona.

Para cargas de trabajo basadas en servidores con estado implementadas en un centro de datos local, puede utilizar AWS Elastic Disaster Recovery para proteger sus cargas de trabajo en AWS. Si ya está alojado en AWS, puede utilizar Elastic Disaster Recovery para proteger su carga de trabajo en una zona o región de disponibilidad alternativa. Elastic Disaster Recovery utiliza la replicación continua a nivel de bloque en un área de preparación ligera para proporcionar una recuperación rápida y fiable de las aplicaciones locales y basadas en la nube.

Pasos para la implementación

1. Implemente la autorrecuperación. Implemente sus instancias o contenedores con escalado automático siempre que sea posible. Si no puede usar el escalado automático, utilice la recuperación automática para instancias EC2 o implemente la automatización de autorrecuperación basada en eventos de ciclo de vida del contenedor de Amazon EC2 o ECS.
   - Utilice los grupos de Amazon EC2 Auto Scaling para instancias y cargas de trabajo de contenedor que no tienen requisitos para una sola dirección IP de instancia, dirección IP privada, dirección IP elástica y metadatos de instancia.
     - Los datos de usuario de la plantilla de lanzamiento se pueden usar para implementar una automatización que pueda solucionar la mayoría de las cargas de trabajo.
   - Utilice la recuperación de instancias Amazon EC2 automática para cargas de trabajo que requieren una única dirección ID de instancia, dirección IP privada, dirección IP elástica y metadatos de instancia.
     - La recuperación automática enviará alertas de estado de recuperación a un tema de SNS cuando se detecte un error en la instancia.
   - Utilice los eventos del ciclo de vida de la instancia Amazon EC2 o los eventos de Amazon ECS para automatizar la autorrecuperación cuando no se pueda utilizar el escalado automático ni la recuperación EC2.
     - Utilice los eventos para invocar la automatización que reparará su componente de acuerdo con la lógica de proceso que necesita.
   - Proteja las cargas de trabajo con estado que se limitan a una única ubicación con AWS Elastic Disaster Recovery.

Recursos

Documentos relacionados:

- Amazon ECS events (Eventos de Amazon ECS)
- Amazon EC2 Auto Scaling lifecycle hooks (Enlaces de ciclo de vida de Amazon EC2 Auto Scaling)
- Recupere la instancia.
- Escalado automático del servicio
- What Is Amazon EC2 Auto Scaling? (¿Qué es Amazon EC2 Auto Scaling?)
- AWS Elastic Disaster Recovery

REL10-BP04 Usar arquitecturas herméticas para limitar el alcance del impacto

La implementación de arquitecturas herméticas (también conocidas como arquitecturas basadas en celdas) restringe el efecto del fallo dentro de una carga de trabajo a un número limitado de componentes.

Resultado deseado: una arquitectura basada en celdas utiliza numerosas instancias aisladas de una carga de trabajo, donde cada instancia se conoce como celda. Cada celda es independiente, no comparte estado con otras celdas y gestiona un subconjunto de las solicitudes de la carga de trabajo global. Esto reduce la posible repercusión de un error, como una actualización de software incorrecta, en una celda individual y en las solicitudes que está procesando. Si una carga de trabajo utiliza 10 celdas para atender 100 peticiones cuando se produce un error, el 90 % del total de las solicitudes no se verá afectado por el error.

Antipatrones usuales:

- Permitir que las celdas crezcan sin límites.
- Aplicar actualizaciones o despliegues de código a todas las celdas al mismo tiempo.
- Compartir estado o componentes entre celdas (a excepción de la capa de enrutador).
- Añadir lógica compleja de negocio o de enrutamiento a la capa de enrutador.
- No minimizar las interacciones entre celdas.

Beneficios de establecer esta práctica recomendada: con las arquitecturas basadas en celdas, muchos tipos habituales de errores están contenidos dentro de la propia celda, lo que proporciona un aislamiento adicional de los errores. Estos límites de errores pueden proporcionar resiliencia

contra tipos de errores que, de otro modo, serían difíciles de contener, como despliegues de código infructuosos o solicitudes que se corrompen o activan un modo de error concreto (también conocidas como solicitudes de píldora envenenada).

Guía para la implementación

En un barco, los mamparos garantizan que una brecha en el casco quede contenida en una sola sección del casco. En los sistemas complejos, este modelo de contención suele imitarse para facilitar el aislamiento de errores. Los límites aislados de los errores restringen el efecto de un error en una carga de trabajo a un número limitado de componentes. Los componentes fuera del límite no resultan afectados por el error. Mediante el uso de varios límites aislados de error, puede acotar el impacto en su carga de trabajo. En AWS, los clientes pueden utilizar varias zonas y regiones de disponibilidad para proporcionar aislamiento de errores, pero el concepto de aislamiento de errores también puede extenderse a la arquitectura de su carga de trabajo.

La carga de trabajo global se divide en celdas mediante una clave de partición. Esta clave tiene que alinearse con la corriente del servicio, o la forma natural en que la carga de trabajo de un servicio puede subdividirse con mínimas interacciones entre celdas. Algunos ejemplos de claves de partición son el ID de cliente, el ID de recurso o cualquier otro parámetro fácilmente accesible en la mayoría de las llamadas a la API. Una capa de enrutador de celdas distribuye las solicitudes a celdas individuales en función de la clave de partición, y presenta un único punto de conexión a los clientes.

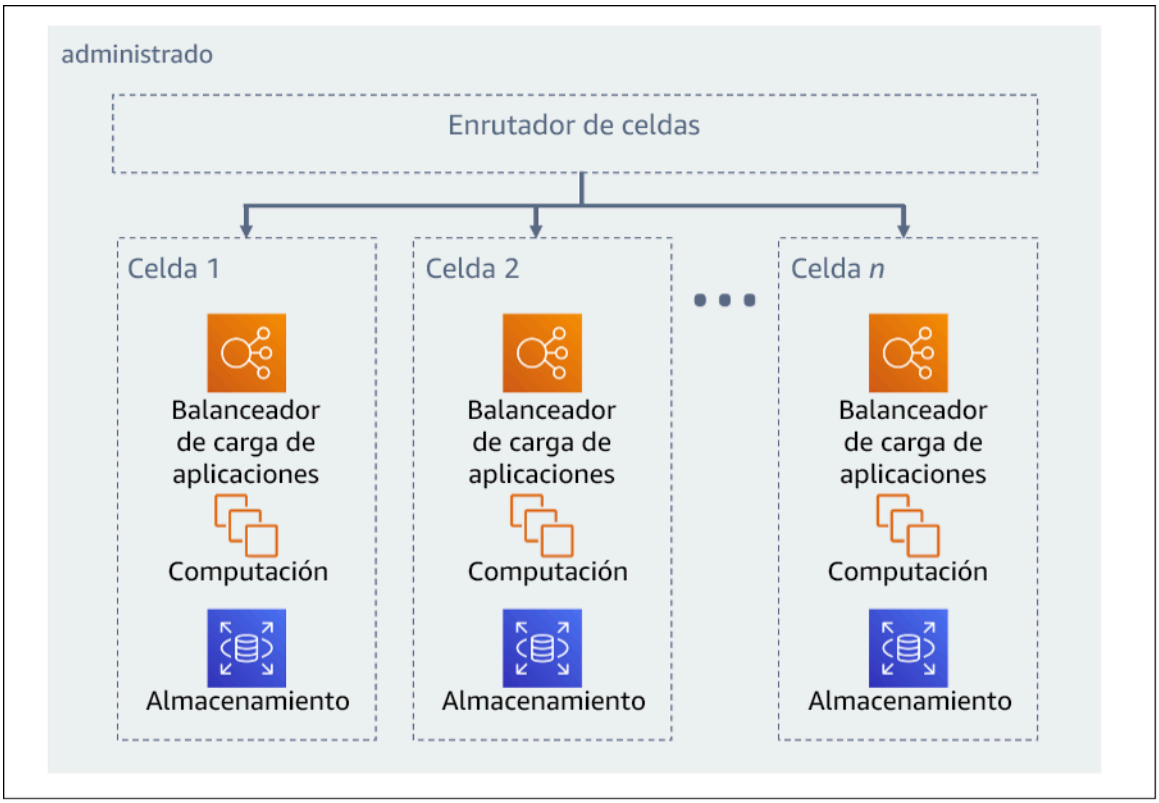


Figura 11: Arquitectura basada en celdas

Pasos para la implementación

Al diseñar una arquitectura basada en celdas, hay que tener en cuenta varias consideraciones de diseño.

1. Clave de partición: debe prestarse especial atención a la hora de elegir la clave de partición.
   - Debe alinearse con la corriente del servicio o con la forma natural en que la carga de trabajo de un servicio puede subdividirse con mínimas interacciones entre celdas. Algunos ejemplos son: `ID de cliente` o bien `ID de recurso`.
   - La clave de partición debe estar disponible en todas las solicitudes, ya sea de modo directo o de una manera que se pueda inferir con facilidad de forma determinista por otros parámetros.
2. Asignación persistente de celdas: los servicios ascendentes solo deben interactuar con una única celda durante el ciclo de vida de sus recursos.
   - Según la carga de trabajo, puede ser necesaria una estrategia de migración de celda para migrar datos de una celda a otra. Un posible escenario en el que puede ser precisa una migración de celda es si un usuario o recurso concreto de la carga de trabajo crece demasiado y requiere una celda dedicada.
   - Las celdas no deben compartir estados ni componentes entre ellas.
   - En consecuencia, las interacciones entre celdas deben evitarse o mantenerse al mínimo, ya que dichas interacciones crean dependencias entre las celdas y, por lo tanto, disminuyen las ventajas en el aislamiento de errores.
3. Capa de enrutador: la capa de enrutador es un componente compartido entre celdas, lo que significa que no puede seguir la misma estrategia de compartimentación que las celdas.
   - Se recomienda que la capa de enrutador distribuya las solicitudes a las celdas individuales mediante un algoritmo de asignación de particiones de una manera eficiente a nivel computacional, como la combinación de funciones hash criptográficas y aritmética modular para asignar claves de partición a las celdas.
   - Para evitar impactos multicelda, la capa de enrutador debe ser lo más simple y escalable horizontalmente posible, lo que requiere evitar una lógica de negocio compleja dentro de esta capa. Esto tiene la ventaja añadida de facilitar la comprensión de su comportamiento esperado en todo momento, lo que permite una comprobabilidad exhaustiva. Como explica Colm MacCárthaigh en Reliability, constant work, and a good cup of coffee (Fiabilidad, trabajo constante y una buena taza de café), los diseños sencillos y los patrones de trabajo constantes producen sistemas fiables y reducen la antifragilidad.

4. Tamaño de la celda: las celdas deben tener un tamaño máximo y no debe permitirse que lo superen.
    - Para determinar el tamaño máximo, se deben llevar a cabo pruebas exhaustivas hasta que se alcancen puntos de ruptura y se establezcan márgenes de funcionamiento seguros. Para obtener más detalles sobre cómo implementar prácticas de prueba, consulte REL07-BP04 Realizar pruebas de la carga de trabajo
    - La carga de trabajo global crecerá a medida que se añadan celdas adicionales, lo que permite escalar la carga de trabajo con los aumentos de la demanda.
5. Estrategias multi-AZ o en varias regiones: se deben aprovechar numerosas capas de resiliencia para ofrecer protección contra diferentes dominios de error.
    - Para obtener resiliencia, debe utilizar un enfoque que cree capas de defensa. Una capa protege de las interrupciones más pequeñas y frecuentes mediante la creación de una arquitectura de alta disponibilidad con múltiples AZ. Otra capa de defensa está pensada para proteger de eventos poco frecuentes, como las catástrofes naturales generalizadas y las interrupciones a nivel regional. Esta segunda capa implica la arquitectura de su aplicación para que abarque múltiples Regiones de AWS. La implementación de una estrategia multirregión para su carga de trabajo le ayuda a protegerla de catástrofes naturales generalizadas que afecten a una región geográfica amplia de un país o de errores técnicos de alcance regional. Tenga en cuenta que implementar una arquitectura multirregión puede ser significativamente complejo y no suele ser necesario para la mayoría de las cargas de trabajo. Para obtener más detalles, consulte REL10-BP02 Seleccionar las ubicaciones adecuadas para el despliegue en varias ubicaciones.
6. Despliegue de código: se prefiere una estrategia de despliegue de código escalonado en lugar de desplegar los cambios de código en todas las celdas al mismo tiempo.
    - Esto ayudará a minimizar posibles errores en numerosas celdas provocados por un despliegue incorrecto o a un error humano. Para obtener más detalles, consulte Automatización de implementaciones seguras y sin intervención.

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

Recursos

Prácticas recomendadas relacionadas:

- REL07-BP04 Realizar pruebas de la carga de trabajo
- REL10-BP02 Seleccionar las ubicaciones adecuadas para el despliegue en varias ubicaciones

Documentos relacionados:

- Reliability, constant work, and a good cup of coffee (Fiabilidad, trabajo constante y una buena taza de café)
- AWS and Compartmentalization (AWS y compartimentalización)
- Aislamiento de las cargas de trabajo a través de la fragmentación aleatoria
- Automatización de implementaciones seguras y sin intervención

Vídeos relacionados:

- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small (AWS re:Invent 2018: Cerrar los bucles y abrir las mentes: cómo asumir el control de los sistemas grandes y pequeños)
- AWS re:Invent 2018: How AWS Minimizes the Blast Radius of Failures (ARC338) (AWS re:Invent 2018: Cómo AWS minimiza el radio de efecto de los errores)
- Shuffle-sharding: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328) (Fragmentación aleatoria: AWS re:Invent 2019: Presentación de Amazon Builders' Library)
- AWS Summit ANZ 2021 - Everything fails, all the time: Designing for resilience (AWS Summit ANZ 2021: Todo falla todo el tiempo: diseñar para la resiliencia)

Ejemplos relacionados:

- Well-Architected Lab - Fault isolation with shuffle sharding (Laboratorio de Well-Architected: Aislamiento de errores con fragmentación aleatoria)

FIABILIDAD 11. ¿Cómo diseña su carga de trabajo para que soporte los errores de los componentes?

Las cargas de trabajo con un requisito de alta disponibilidad y un tiempo de recuperación (MTTR) bajo deben diseñarse para que sean resilientes.

Prácticas recomendadas

- REL11-BP01 Supervisar todos los componentes de la carga de trabajo para detectar errores
- REL11-BP02 Conmutación por error a recursos en buen estado
- REL11-BP03 Automatizar la reparación en todas las capas

- REL11-BP04 Confiar en el plano de datos y no en el plano de control durante la recuperación
- REL11-BP05 Usar la estabilidad estática para evitar el comportamiento bimodal
- REL11-BP06 Enviar notificaciones cuando los eventos afecten a la disponibilidad
- REL11-BP07 Diseñar su producto para cumplir objetivos de disponibilidad y acuerdos de nivel de servicio (SLA) de tiempo de actividad

REL11-BP01 Supervisar todos los componentes de la carga de trabajo para detectar errores

Supervise continuamente el estado de las cargas de trabajo para que usted y los sistemas automatizados sepan cuándo se produce degradaciones o errores en cuanto ocurran. Supervise los indicadores clave de rendimiento (KPI) en función del valor empresarial.

Todos los mecanismos de recuperación y corrección deben comenzar por la capacidad de detectar problemas rápidamente. Los fallos técnicos deberían detectarse en primer lugar para poder resolverse. Sin embargo, la disponibilidad se basa en la capacidad de su carga de trabajo para ofrecer valor empresarial, de modo que los indicadores clave de rendimiento (KPI) que midan esto tengan que formar parte de su estrategia de detección y corrección.

Resultado deseado: los componentes esenciales de una carga de trabajo se supervisan de forma independiente para detectar y alertar sobre los errores en el momento y el lugar en que se producen.

Patrones comunes de uso no recomendados:

- No se han configurado alarmas, por lo que las interrupciones se producen sin notificación.
- Existen alarmas, pero en umbrales que no proporcionan el tiempo necesario para reaccionar.
- No se recopilan métricas con la suficiente regularidad para satisfacer el objetivo de tiempo de recuperación (RTO).
- Solo se supervisan activamente las interfaces de la carga de trabajo orientadas a los clientes.
- Solo se recopilan métricas técnicas, no métricas de funciones empresariales.
- No hay métricas que midan la experiencia del usuario con la carga de trabajo.
- Se crean demasiadas supervisiones.

Beneficios de establecer esta práctica recomendada: Una supervisión adecuada de todas las capas le permite reducir el tiempo de recuperación al reducirse el tiempo de detección.

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

Guía para la implementación

Identifique todas las cargas de trabajo que se revisarán para su supervisión. Una vez que haya identificado todos los componentes de la carga de trabajo que deberán supervisarse, tendrá que determinar el intervalo de supervisión. El intervalo de supervisión tendrá un impacto directo en la rapidez con la que se puede iniciar la recuperación en función del tiempo que se tarde en detectar un error. El tiempo medio de detección (MTTD) es el tiempo transcurrido entre la aparición de un error y el inicio de las operaciones de reparación. La lista de servicios debe ser amplia y completa.

La supervisión debe cubrir todas las capas de la pila de aplicaciones, incluidas la aplicación, la plataforma, la infraestructura y la red.

Su estrategia de supervisión debe considerar el impacto de los errores grises. Para obtener más información sobre los errores grises, consulte la sección de errores grises en el documento técnico Advanced Multi-AZ Resilience Patterns.

Pasos para la implementación

- El intervalo de supervisión depende de la rapidez con la que deba recuperarse. El tiempo de recuperación depende del tiempo que tarde la recuperación, por lo que debe determinar la frecuencia de recopilación teniendo en cuenta este tiempo y el objetivo de tiempo de recuperación (RTO).
- Configure la supervisión detallada de los componentes y los servicios administrados.
  - Determine si supervisión detallada de instancias de EC2 y Auto Scaling es necesaria. La supervisión detallada proporciona métricas en intervalos de un minuto y la supervisión predeterminada proporciona métricas en intervalos de cinco minutos.
  - Determine si supervisión mejorada para RDS es necesaria. La supervisión mejorada usa un agente en las instancias de RDS para obtener información útil sobre los diferentes procesos o subprocesos.
  - Determine los requisitos de supervisión de los componentes sin servidor cruciales para Lambda, API Gateway, Amazon EKS, Amazon ECSy todos los tipos de equilibradores de carga.
  - Determine los requisitos de supervisión de los componentes de almacenamiento para Amazon S3, Amazon FSx, Amazon EFSy Amazon EBS.
- Cree métricas personalizadas para medir los indicadores clave de rendimiento (KPI) de la empresa. Las cargas de trabajo implementan funciones empresariales clave, que deben usarse como KPI para ayudar a identificar cuándo se produce un problema indirecto.

- Supervise la experiencia del usuario para detectar errores mediante valores controlados del usuario. Las pruebas de transacciones sintéticas (también denominadas pruebas de valores controlados, que no deben confundirse con los despliegues de valores controlados) que puedan ejecutar y simular el comportamiento de los clientes son uno de los procesos de prueba más importantes. Ejecute estas pruebas constantemente en los puntos de conexión de las cargas de trabajo desde distintas ubicaciones remotas.
- Cree métricas personalizadas que siguen la experiencia del usuario. Si puede instrumentar la experiencia del cliente, puede determinar cuándo se degrada la experiencia del cliente.
- Configure alarmas para detectar cuándo alguna parte de la carga de trabajo no funciona correctamente y para indicar cuándo escalar automáticamente los recursos. Las alarmas pueden mostrarse visualmente en paneles, enviar alertas a través de Amazon SNS o por correo electrónico y trabajar con Auto Scaling para escalar o desescalar verticalmente los recursos de la carga de trabajo.
- Cree paneles para visualizar las métricas. Se pueden usar paneles para visualizar las tendencias, los valores atípicos y otros indicadores de problemas potenciales, o para proporcionar una indicación de problemas que tal vez le convenga investigar.
- Cree supervisión de rastreo distribuido para sus servicios. Con la supervisión distribuida, podrá saber cómo se comporta su aplicación y sus servicios subyacentes para identificar y resolver la causa raíz de los problemas y errores de rendimiento.
- Cree paneles de sistemas de supervisión (mediante CloudWatch o bien X-Ray) y recopilación de datos en una región y una cuenta independientes.
- Cree una integración para la supervisión de Amazon Health Aware para poder supervisar la visibilidad de los recursos de AWS que podrían estar degradados. Para las cargas de trabajo empresariales esenciales, esta solución proporciona acceso a alertas proactivas y en tiempo real para los servicios de AWS.

Recursos

Prácticas recomendadas relacionadas:

- Definición de disponibilidad
- REL11-BP06 Enviar notificaciones cuando los eventos afecten a la disponibilidad

Documentos relacionados:

- Amazon CloudWatch Synthetics enables you to create user canaries

- Habilitar o deshabilitar la supervisión detallada de su instancia
- Monitoreo mejorado
- Monitoring Your Auto Scaling Groups and Instances Using Amazon CloudWatch
- Publicar métricas personalizadas
- Using Amazon CloudWatch Alarms
- Using CloudWatch Dashboards
- Using Cross Region Cross Account CloudWatch Dashboards
- Using Cross Region Cross Account X-Ray Tracing
- Understanding availability
- Implementing Amazon Health Aware (AHA)

Vídeos relacionados:

- Mitigating gray failures

Ejemplos relacionados:

- Laboratorio de Well-Architected: Nivel 300: Implementación de comprobaciones de estado y administración de dependencias para mejorar la fiabilidad
- One Observability Workshop: Explore X-Ray

Herramientas relacionadas:

- CloudWatch
- CloudWatch X-Ray

REL11-BP02 Conmutación por error a recursos en buen estado

Si un recurso fallara, los recursos en buen estado deberían seguir atendiendo las solicitudes. Para problemas de ubicación (como zonas de disponibilidad o Región de AWS), asegúrese de que dispone de sistemas para conmutar por error a recursos en buen estado en ubicaciones sin problemas.

Al diseñar un servicio, distribuya la carga entre los recursos, las zonas de disponibilidad o las regiones. De esta manera, el error de un recurso individual o el deterioro puede mitigarse al

desplazar el tráfico a los recursos restantes en buen estado. Tenga en cuenta cómo se descubren los servicios y cómo se enruta a ellos en caso de que se produzca un error.

Tenga en cuenta la recuperación de errores al diseñar sus servicios. En AWS, diseñamos servicios para minimizar el tiempo de recuperación de los errores y el impacto en los datos. Nuestros servicios utilizan principalmente almacenes de datos que confirman las solicitudes solo después de que se almacenan de forma duradera en varias réplicas en una región. Se han diseñado para utilizar el aislamiento basado en celdas y el aislamiento de errores que proporcionan las zonas de disponibilidad. Utilizamos ampliamente la automatización en nuestros procedimientos operativos. También optimizamos nuestra funcionalidad de reemplazo y reinicio para recuperarnos rápidamente de las interrupciones.

Los patrones y diseños que permiten la conmutación por error varían para cada servicio de plataforma de AWS. Muchos servicios administrados nativos de AWS son zonas de disponibilidad múltiples de forma nativa (como Lambda o API Gateway). Otros servicios de AWS (como EC2 y EKS) requieren diseños específicos de las prácticas recomendadas para admitir la conmutación por error de los recursos o el almacenamiento de datos en las AZ.

La supervisión debe configurarse para que compruebe que el recurso de conmutación por error esté en buen estado, realizar un seguimiento del progreso de los recursos de conmutación por error y supervisar la recuperación de los procesos empresariales.

Resultado deseado: los sistemas son capaces de utilizar nuevos recursos de forma automática o manual para recuperarse de la degradación.

Patrones comunes de uso no recomendados:

- La planificación para errores no forma parte de la fase de planificación y diseño.
- No se establecen el RTO y el RPO.
- Supervisión insuficiente para detectar recursos defectuosos.
- Aislamiento adecuado de los dominios de error.
- No se considera la conmutación por error multirregional.
- La detección de errores es demasiado sensible o agresiva a la hora de decidir realizar una conmutación por error.
- No se prueba ni se valida el diseño de la conmutación por error.
- Realizar la automatización de la autorreparación, pero no notificar que se necesita una reparación
- No hay un período de amortiguación para evitar que la conmutación por error se lleve a cabo demasiado pronto.

Beneficios de establecer esta práctica recomendada: con una degradación uniforme y una recuperación rápida, puede crear sistemas más resilientes que mantengan la fiabilidad cuando se producen errores.

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

Guía para la implementación

Los servicios de AWS, como Elastic Load Balancing y Amazon EC2 Auto Scaling, ayudan a distribuir la carga entre los recursos y las zonas de disponibilidad. Por lo tanto, el error de un recurso individual (como una instancia de EC2) o el deterioro de una zona de disponibilidad puede mitigarse si se desplaza el tráfico a los recursos restantes en buen estado.

Para las cargas de trabajo multirregión, los diseños son más complicados. Por ejemplo, las réplicas de lectura entre regiones le permiten desplegar sus datos en varias Regiones de AWS. Sin embargo, la conmutación por error sigue siendo necesaria para convertir la réplica de lectura en principal y, a continuación, dirigir el tráfico al nuevo punto de conexión. Amazon Route 53, Route 53 ARC, CloudFront y AWS Global Accelerator pueden ayudar a dirigir el tráfico a través de las Regiones de AWS.

Los servicios de AWS, como Amazon S3, Lambda, API Gateway, Amazon SQS, Amazon SNS, Amazon SES, Amazon Pinpoint, Amazon ECR, AWS Gestor de certificados, EventBridge o Amazon DynamoDB, se despliegan automáticamente en varias zonas de disponibilidad mediante AWS. En caso de error, estos servicios de AWS dirigen automáticamente el tráfico a ubicaciones en buen estado. Los datos se almacenan de forma redundante en varias zonas de disponibilidad y siguen estando disponibles.

Para Amazon RDS, Amazon Aurora, Amazon Redshift, Amazon EKS o Amazon ECS, Multi-AZ es una opción de configuración. AWS puede dirigir el tráfico a la instancia en buen estado si se inicia la conmutación por error. Esta acción de conmutación por error puede ser realizada por AWS o según lo requiera el cliente.

Para las instancias de Amazon EC2, Amazon Redshift, tareas de Amazon ECS o pods de Amazon EKS, usted elige en qué zonas de disponibilidad desea realizar el despliegue. En algunos diseños, Elastic Load Balancing proporciona la solución para detectar las instancias en las zonas que no tienen un estado correcto y enrutar el tráfico a las que sí lo tienen. Elastic Load Balancing también puede enrutar el tráfico a los componentes de su centro de datos local.

En cuanto a la conmutación por error del tráfico multirregional, el reenrutamiento puede utilizar Amazon Route 53, ARC, AWS Global Accelerator, Route 53 Private DNS for VPCs o CloudFront

para proporcionar una forma de definir dominios de Internet y asignar políticas de enrutamiento, incluidas comprobaciones de estado, para enrutar el tráfico a regiones en buen estado. AWS Global Accelerator proporciona direcciones IP estáticas que actúan como punto de entrada fijo a su aplicación; a continuación, se enrutan a los puntos de conexión de las Regiones de AWS que elija, mediante la red global de AWS en lugar de Internet para mejorar el rendimiento y la fiabilidad.

Pasos para la implementación

- Cree diseños de conmutación por error para todas las aplicaciones y servicios pertinentes. Aísle cada componente de la arquitectura y cree diseños de conmutación por error que satisfagan el RTO y el RPO de cada componente.
- Configure entornos inferiores (como los de desarrollo o prueba) con todos los servicios que sean necesarios para tener un plan de conmutación por error. Despliegue las soluciones mediante la infraestructura como código (IaC) para garantizar la repetibilidad.
- Configure un sitio de recuperación, como una segunda región, para implementar y probar los diseños de conmutación por error. Si fuera necesario, los recursos para las pruebas se pueden configurar temporalmente para limitar los costes adicionales.
- Determine qué planes de conmutación por error se automatizan mediante AWS, cuáles pueden automatizarse mediante un proceso de DevOps y cuáles pueden ser manuales. Documente y mida el RTO y el RPO de cada servicio.
- Cree una guía de estrategias de conmutación por error e incluya todos los pasos de la conmutación por error de cada recurso, aplicación y servicio.
- Cree una guía de estrategias de conmutación por recuperación e incluya todos los pasos de la conmutación por recuperación (con plazos) de cada recurso, aplicación y servicio.
- Cree un plan para iniciar y ensayar la guía de estrategias. Utilice simulaciones y pruebas de caos para poner a prueba los pasos de la guía de estrategias y la automatización.
- Para problemas de ubicación (como zonas de disponibilidad o Región de AWS), asegúrese de que dispone de sistemas para conmutar por error a recursos en buen estado en ubicaciones sin problemas. Compruebe la cuota, los niveles de escalado automático y los recursos en ejecución antes de realizar la prueba de conmutación por error.

Recursos

Prácticas recomendadas por Well-Architected:

- REL13: Plan para DR

- REL10: Uso del aislamiento de errores para proteger la carga de trabajo

Documentos relacionados:

- Setting RTO and RPO Targets
- Set up ARC with application loadbalancers
- Failover using Route 53 Weighted routing
- DR with ARC
- EC2 with autoscaling
- EC2 Deployments - Multi-AZ
- ECS Deployments - Multi-AZ
- Switch traffic using ARC
- Lambda with an Application Load Balancer and Failover
- ACM Replication and Failover
- Parameter Store Replication and Failover
- ECR cross region replication and Failover
- Secrets manager cross region replication configuration
- Enable cross region replication for EFS and Failover
- EFS Cross Region Replication and Failover
- Networking Failover
- S3 Endpoint failover using MRAP
- Create cross region replication for S3
- Failover Regional API Gateway with ARC
- Failover using multi-region global accelerator
- Failover with DRS
- Creating Disaster Recovery Mechanisms Using Amazon Route 53

Ejemplos relacionados:

- Disaster Recovery on AWS

- Elastic Disaster Recovery on AWS

REL11-BP03 Automatizar la reparación en todas las capas

Cuando se detecte un error, utilice las funciones automatizadas para tomar medidas correctivas. Las degradaciones pueden repararse automáticamente a través de mecanismos de servicio interno o requerir que los recursos se reinicien o eliminen a través de medidas de corrección.

Para las aplicaciones autoadministradas y la reparación entre regiones, los diseños de recuperación y los procesos de reparación automatizados se pueden extraer de las prácticas recomendadas existentes.

La capacidad de reiniciar o eliminar un recurso es una herramienta importante para corregir los errores. Una práctica recomendada es convertir los servicios en servicios sin estado siempre que sea posible. Esto evita la pérdida de datos o disponibilidad tras el reinicio del recurso. En la nube, puede (y generalmente debería) sustituir todo el recurso (por ejemplo, la instancia de computación o la función sin servidor) como parte del reinicio. El reinicio en sí es una forma sencilla y fiable de recuperarse de un error. En las cargas de trabajo ocurren muchos tipos de errores diferentes. Los errores pueden ocurrir en el hardware, el software, las comunicaciones y el funcionamiento.

El reinicio o el reintento también se aplican a las solicitudes de red. Se aplica el mismo enfoque de recuperación tanto a un tiempo de espera de la red como a un error en la dependencia, en el que la dependencia devuelve un error. Ambos eventos tienen un efecto similar en el sistema, por lo que en lugar de intentar convertir cada uno en un caso especial, se aplicaría una estrategia similar de reintento con retroceso exponencial y fluctuación. La capacidad de reiniciar es un mecanismo de recuperación que aparece en la computación orientada a la recuperación y en las arquitecturas de clústeres de alta disponibilidad.

Resultado deseado: se llevan a cabo medidas automatizadas para corregir la detección de un error.

Patrones comunes de uso no recomendados:

- Aprovisionar recursos sin escalado automático.
- Desplegar las aplicaciones en instancias o contenedores individualmente.
- Implementar aplicaciones que no se pueden implementar en varias ubicaciones sin usar la recuperación automática
- Reparar manualmente las aplicaciones que el escalamiento automático y la recuperación automática no pueden reparar.

- No hay automatización de las bases de datos de conmutación por error.
- Carencia de métodos automatizados para redirigir el tráfico a nuevos puntos de conexión.
- No hay replicación del almacenamiento.

Beneficios de establecer esta práctica recomendada: la reparación automática puede reducir el tiempo medio de recuperación y mejorar la disponibilidad.

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

Guía para la implementación

Los diseños de Amazon EKS u otros servicios de Kubernetes deben incluir conjuntos de réplicas o con estado mínimo y máximo y el tamaño mínimo del clúster y los grupos de nodos. Estos mecanismos proporcionan una cantidad mínima de recursos de procesamiento disponibles de forma continua y, al mismo tiempo, corrigen automáticamente cualquier error mediante el plano de control de Kubernetes.

Los patrones de diseño a los que se accede a través de un equilibrador de carga mediante clústeres de computación deben utilizar los grupos de Auto Scaling. Elastic Load Balancing (ELB) distribuye automáticamente el tráfico de aplicaciones entrante entre varios destinos y dispositivos virtuales en una o más zonas de disponibilidad (AZ).

Los diseños basados en computación en clúster que no utilizan el equilibrio de carga deben tener un diseño de tamaño que de cabida a la pérdida de al menos un nodo. Esto permitirá que el servicio siga funcionando con una capacidad potencialmente reducida mientras recupera un nuevo nodo. Algunos servicios son Mongo, DynamoDB Accelerator, Amazon Redshift, Amazon EMR, Cassandra, Kafka, MSK-EC2, Couchbase, ELK y Amazon OpenSearch Service. Muchos de estos servicios se pueden diseñar con características adicionales de autorreparación. Algunas tecnologías de clústeres deben generar una alerta ante la pérdida de un nodo, lo que desencadena un flujo de trabajo automático o manual para recrear un nuevo nodo. Este flujo de trabajo se puede automatizar con AWS Systems Manager para corregir los problemas rápidamente.

Se puede usar Amazon EventBridge para supervisar y filtrar los eventos, como las alarmas de Amazon EC2 Auto Scaling o cambios en el estado en otros servicios de AWS. En función de la información del evento, se puede invocar a AWS Lambda, la automatización de Systems Manager u otros destinos para ejecutar una lógica de corrección personalizada en su carga de trabajo. Amazon EC2 Auto Scaling se puede configurar para comprobar el estado de la instancia de EC2. Si la instancia está en un estado que no sea el de ejecución, o si el estado del sistema se ve

deteriorado, Amazon EC2 Auto Scaling considera que la instancia no está en buen estado y lanza una instancia de sustitución. Para sustituciones a gran escala (como la pérdida de toda una zona de disponibilidad), se prefiere la estabilidad estática para la alta disponibilidad.

Pasos para la implementación

- Use grupos de Auto Scaling para desplegar niveles en una carga de trabajo. Auto Scaling puede realizar una autorreparación de aplicaciones sin estado, y añadir y eliminar capacidad.
- Para las instancias de computación indicadas anteriormente, utilice el equilibrio de carga y elija el tipo de equilibrador de carga adecuado.
- Considere la posibilidad de reparación para Amazon RDS. Con las instancias en espera, configure la conmutación por error automática a la instancia en espera. Para la réplica de lectura de Amazon RDS, se requiere un flujo de trabajo automatizado para convertir una réplica de lectura en principal.
- Implemente la recuperación automática en instancias de EC2 que tengan aplicaciones desplegadas que no se puedan desplegar en varias ubicaciones y puedan tolerar el reinicio tras un error. La recuperación automática se puede usar para reemplazar hardware defectuoso y reiniciar la instancia cuando la aplicación no se puede implementar en varias ubicaciones. Se conservan los metadatos de la instancia y las direcciones IP asociadas, así como los volúmenes de EBS y los puntos de montaje en Amazon Elastic File System o bien sistemas de archivos para Lustre y Windows. Con AWS OpsWorks, puede configurar la autorreparación de las instancias de EC2 en el nivel de capa.
- Implemente la recuperación automatizada mediante AWS Step Functions y AWS Lambda cuando no pueda usar el escalamiento automático ni la recuperación automática, o cuando la recuperación automática produzca un error. Cuando no pueda usar el escalamiento automático ni la recuperación automática, o esta produzca un error, puede automatizar la reparación con AWS Step Functions y AWS Lambda.
- Amazon EventBridge se puede usar para supervisar y filtrar los eventos, como las alarmas de CloudWatch o los cambios en el estado en otros servicios de AWS. En función de la información del evento, se puede invocar a AWS Lambda (u otros destinos) para ejecutar una lógica de corrección personalizada en su carga de trabajo.

Recursos

Prácticas recomendadas relacionadas:

- Definición de disponibilidad

- REL11-BP01 Supervisar todos los componentes de la carga de trabajo para detectar errores

Documentos relacionados:

- How AWS Auto Scaling Works
- Amazon EC2 Automatic Recovery
- Amazon Elastic Block Store (Amazon EBS)
- Amazon Elastic File System (Amazon EFS)
- What is Amazon FSx for Lustre?
- What is Amazon FSx for Windows File Server?
- AWS OpsWorks: Using Auto Healing to Replace Failed Instances
- What isAWS Step Functions?
- What is AWS Lambda?
- What is Amazon EventBridge?
- Using Amazon CloudWatch Alarms
- Amazon RDS Failover
- SSM - Systems Manager Automation
- Resilient Architecture Best Practices

Vídeos relacionados:

- Automatically Provision and Scale OpenSearch Service
- Amazon RDS Failover Automatically

Ejemplos relacionados:

- Workshop on Auto Scaling
- Amazon RDS Failover Workshop

Herramientas relacionadas:

- CloudWatch

- CloudWatch X-Ray

REL11-BP04 Confiar en el plano de datos y no en el plano de control durante la recuperación

Los planos de control proporcionan las API administrativas que se utilizan para crear, leer y describir, actualizar, eliminar y enumerar los recursos (CRUDL), mientras que los planos de datos gestionan el tráfico de servicio diario. Al implementar respuestas de recuperación o mitigación a eventos que puedan afectar a la resiliencia, concéntrese en utilizar un número mínimo de operaciones del plano de control para recuperar, reescalar, restaurar, reparar o conmutar por error el servicio. La acción del plano de datos debe reemplazar cualquier actividad durante estos eventos de degradación.

Por ejemplo, las siguientes son todas acciones del plano de control: lanzar una nueva instancia de computación, crear almacenamiento en bloques y describir los servicios de colas. Al lanzar instancias de computación, el plano de control debe realizar varias tareas, como encontrar un host físico con capacidad, asignar interfaces de red, preparar los volúmenes de almacenamiento en bloques locales, generar credenciales y añadir reglas de seguridad. Los planos de control suelen tener una orquestación complicada.

Resultado deseado: cuando un recurso entra en un estado deteriorado, el sistema es capaz de recuperarse automática o manualmente al cambiar el tráfico de recursos deteriorados a recursos en buen estado.

Patrones comunes de uso no recomendados:

- Dependencia de cambiar los registros de DNS para redirigir el tráfico.
- Dependencia de las operaciones de escalado del plano de control para reemplazar los componentes dañados debido a que no se han aprovisionado suficientes recursos.
- Confiar en amplias acciones del plano de control, multiservicio y multiAPI para corregir cualquier categoría de deterioro.

Beneficios de establecer esta práctica recomendada: el aumento de la tasa de éxito de la corrección automatizada puede reducir el tiempo medio de recuperación y mejorar la disponibilidad de la carga de trabajo.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Medio. En ciertos tipos de degradaciones del servicio, los planos de control se ven afectados. La dependencia del uso extensivo del plano de control para la corrección puede aumentar el tiempo de recuperación (RTO) y el tiempo medio de recuperación (MTTR).

Guía para la implementación

Para limitar las acciones del plano de datos, evalúe cada servicio para determinar qué acciones son necesarias para restablecer el servicio.

Utilice Amazon Application Recovery Controller (ARC) para cambiar el tráfico de DNS. Estas características supervisan continuamente la capacidad de la aplicación de recuperarse de los errores y le permiten controlar la recuperación de la aplicación en las distintas Regiones de AWS, zonas de disponibilidad y localmente.

Las políticas de enrutamiento de Route 53 utilizan el plano de control, por lo que no debe confiar en él para la recuperación. Los planos de datos de Route 53 responden a consultas de DNS y llevan a cabo y evalúan comprobaciones de estado. Se distribuyen a nivel mundial y están diseñados para un acuerdo de nivel de servicio (SLA) del 100 % de disponibilidad..

Las API de administración de Route 53 y las consolas en las que se crean, actualizan y eliminan recursos de Route 53 se ejecutan en planos de control diseñados para dar prioridad a la sólida coherencia y durabilidad que necesita al administrar DNS. Para conseguirlo, los planos de control se encuentran en una única región: Este de EE. UU. (Norte de Virginia). Aunque ambos sistemas se han diseñado para ser muy fiables, los planos de control no están incluidos en el SLA. Podría haber eventos poco frecuentes en los que el diseño resiliente del plano de datos permita mantener la disponibilidad mientras que los planos de control no lo permitan. Con los mecanismos de recuperación de desastres y conmutación por error, utilice las funciones del plano de datos para proporcionar la mejor fiabilidad posible.

Para Amazon EC2, utilice diseños de estabilidad estática para limitar las acciones del plano de control. Las acciones del plano de control incluyen la ampliación de los recursos de forma individual o mediante grupos de Auto Scaling (ASG). Para obtener los niveles más altos de resiliencia, aprovisione suficiente capacidad en el clúster utilizado para la conmutación por error. Si este umbral de capacidad debe limitarse, establezca reguladores en todo el sistema de principio a fin para limitar de forma segura el tráfico total que llega al conjunto limitado de recursos.

Para servicios como Amazon DynamoDB, Amazon API Gateway, los equilibradores de carga y los servicios de AWS Lambda sin servidor, el uso de esos servicios utiliza el plano de datos. Sin embargo, la creación de nuevas funciones, equilibradores de carga, puertas de enlace de API o tablas de DynamoDB es una acción del plano de control y debe completarse antes de la degradación como preparación para un evento y ensayo de las acciones de conmutación por error. En el caso de Amazon RDS, las acciones del plano de datos permiten el acceso a los datos.

Para obtener más información sobre los planos de datos, los planos de control y cómo AWS crea servicios para cumplir los objetivos de alta disponibilidad, consulte Estabilidad estática con zonas de disponibilidad.

Comprenda qué operaciones están en el plano de datos y cuáles están en el plano de control.

Pasos para la implementación

Para cada carga de trabajo que deba restaurarse después de un evento de degradación, evalúe el runbook de conmutación por error, el diseño de alta disponibilidad, el diseño de reparación automática o el plan de restauración de recursos de alta disponibilidad. Identifique cada acción que pueda considerarse una acción del plano de control.

Considere cambiar la acción de control por una acción del plano de datos:

- Auto Scaling (plano de control) en comparación con los recursos de Amazon EC2 preescalados (plano de datos).
- Migre a Lambda y sus métodos de escalado (plano de datos) o a Amazon EC2 y ASG (plano de control).
- Evalúe cualquier diseño con Kubernetes y la naturaleza de las acciones del plano de control. Añadir pods es una acción del plano de datos en Kubernetes. Las acciones deben limitarse a añadir pods y no a añadir nodos. Con nodos sobreaprovisionados es el método preferido para limitar las acciones del plano de control.

Considere enfoques alternativos que permitan que las acciones del plano de datos afecten a la misma corrección.

- Cambio de registro de Route 53 (plano de control) o ARC (plano de datos).
- Comprobaciones de estado de Route 53 para obtener actualizaciones más automatizadas.

Considere algunos servicios en una región secundaria, si el servicio es crucial para la misión, para permitir más acciones del plano de control y el plano de datos en una región no afectada.

- Amazon EC2 Auto Scaling o Amazon EKS en una región principal en comparación con Amazon EC2 Auto Scaling o Amazon EKS en una región secundaria y enrutamiento del tráfico a la región secundaria (acción del plano de control).
- Hacer réplicas de lectura en la región principal secundaria o intentar la misma acción en la región principal (acción del plano de control).

Recursos

Prácticas recomendadas relacionadas:

- Definición de disponibilidad
- REL11-BP01 Supervisar todos los componentes de la carga de trabajo para detectar errores

Documentos relacionados:

- APN Partner: socios que pueden ayudar con la automatización de su tolerancia a errores
- AWS Marketplace: productos que pueden usarse para tolerancia a errores
- La Amazon Builders' Library: Evitar la sobrecarga de los sistemas distribuidos asumiendo el control del servicio más pequeño
- API de Amazon DynamoDB (plano de control y plano de datos)
- AWS Lambda Executions (divididas entre el plano de control y el plano de datos)
- AWS Elemental MediaStore Data Plane
- Building highly resilient applications using Amazon Route 53 Application Recovery Controller, Part 1: Single-Region stack
- Building highly resilient applications using Amazon Route 53 Application Recovery Controller, Part 2: Multi-Region stack
- Creating Disaster Recovery Mechanisms Using Amazon Route 53
- What is Route 53 Application Recovery Controller
- Kubernetes Control Plane and data plane

Vídeos relacionados:

- Back to Basics - Using Static Stability
- Building resilient multi-site workloads using AWS global services

Ejemplos relacionados:

- Introducing Amazon Route 53 Application Recovery Controller
- La Amazon Builders' Library: Evitar la sobrecarga de los sistemas distribuidos asumiendo el control del servicio más pequeño

- Building highly resilient applications using Amazon Route 53 Application Recovery Controller, Part 1: Single-Region stack
- Building highly resilient applications using Amazon Route 53 Application Recovery Controller, Part 2: Multi-Region stack
- Estabilidad estática con zonas de disponibilidad

Herramientas relacionadas:

- Amazon CloudWatch
- AWS X-Ray

REL11-BP05 Usar la estabilidad estática para evitar el comportamiento bimodal

Las cargas de trabajo deben ser estáticamente estables y funcionar solo en un único modo normal. El comportamiento bimodal se produce cuando la carga de trabajo presenta un comportamiento diferente en los modos normal y de error.

Por ejemplo, puede intentar recuperarse de un error en una zona de disponibilidad lanzando nuevas instancias en una zona de disponibilidad diferente. Esto puede dar como resultado una respuesta bimodal durante un modo de error. En lugar de ello, debe crear cargas de trabajo que sean estables estáticamente y operen dentro de un solo modo. En este ejemplo, esas instancias deberían haberse aprovisionado en la segunda zona de disponibilidad antes del error. Este diseño de estabilidad estática verifica que la carga de trabajo solo funcione en un solo modo.

Resultado deseado: las cargas de trabajo no muestran un comportamiento bimodal durante los modos normal y de error.

Patrones comunes de uso no recomendados:

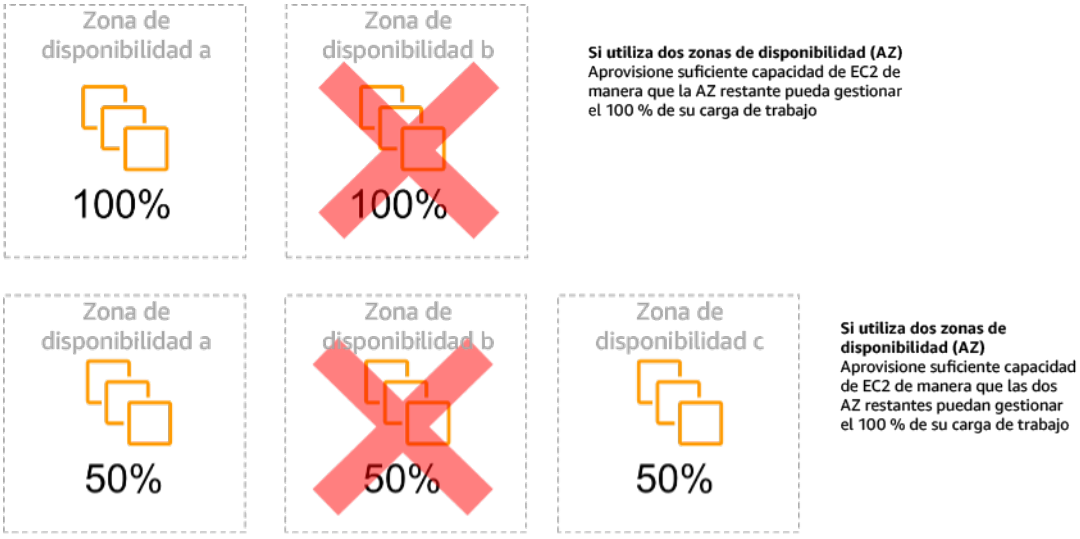
- Suponer que los recursos siempre se pueden aprovisionar independientemente del alcance del error.
- Intentar adquirir recursos de forma dinámica durante un error.
- No aprovisionar los recursos adecuados en todas las zonas o regiones hasta que se produzca un error.
- Considerar diseños estáticos estables solo para recursos de computación.

Beneficios de establecer esta práctica recomendada: las cargas de trabajo que se ejecutan con diseños estáticamente estables pueden tener resultados predecibles durante eventos normales y de error.

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

Guía para la implementación

El comportamiento bimodal ocurre cuando la carga de trabajo exhibe diferentes comportamientos en los modos normal y de error (como confiar en el lanzamiento de nuevas instancias si se produce un error en una zona de disponibilidad). Un ejemplo de comportamiento bimodal ocurre cuando los diseños de Amazon EC2 estables aprovisionan suficientes instancias en cada zona de disponibilidad para gestionar la carga de trabajo si se eliminara una de estas zonas. Se comprobaría el estado de Elastic Load Balancing o Amazon Route 53 para retirar una carga de las instancias dañadas. Una vez que el tráfico ha cambiado, use AWS Auto Scaling para sustituir de forma asíncrona las instancias de la zona con errores y lanzarlas en las zonas en buen estado. La estabilidad estática para despliegues de computación (como instancias EC2 o contenedores) da como resultado la máxima fiabilidad.



Estabilidad estática de las instancias EC2 entre zonas de disponibilidad

Esto debe sopesarse en relación al coste de este modelo y el valor empresarial de mantener la carga de trabajo en todos los casos de resiliencia. Es menos caro aprovisionar menos capacidad de computación y confiar en el lanzamiento de nuevas instancias en caso de error, pero en el caso de errores a gran escala (como una deterioro regional o de zona de disponibilidad), este enfoque es menos eficaz porque se basa tanto en un plano operativo como en la disponibilidad de recursos suficientes en las zonas o regiones no afectadas.

Su solución también debe sopesar la fiabilidad en comparación con los costes necesarios para su carga de trabajo. Las arquitecturas de estabilidad estática se aplican a una variedad de arquitecturas, incluidas las instancias de computación distribuidas en las zonas de disponibilidad, los diseños de réplicas de lectura de bases de datos, los diseños de clústeres de Kubernetes (Amazon EKS) y las arquitecturas de conmutación por error multirregional.

También es posible implementar un diseño más estable desde el punto de vista estático mediante el uso de más recursos en cada zona. Al agregar más zonas, reduce la cantidad de procesamiento adicional que necesita para la estabilidad estática.

Un ejemplo de comportamiento bimodal sería un tiempo de espera de la red que podría provocar que un sistema intente actualizar el estado de configuración de todo el sistema. Se añadiría una carga inesperada a otro componente, lo que podría hacer que falle y desencadene otras consecuencias inesperadas. Este bucle de retroalimentación negativa afecta a la disponibilidad de su carga de trabajo. En lugar de ello, puede crear cargas de trabajo que sean estables estáticamente y operen en un solo modo. Un diseño estáticamente estable haría un trabajo constante y actualizaría continuamente el estado de configuración a una cadencia establecida. Cuando una llamada genera un error, la carga de trabajo utiliza el valor previamente almacenado en caché e inicia una alarma.

Otro ejemplo de comportamiento bimodal es permitir que los clientes eludan la caché de la carga de trabajo si se produce un error. Esto podría parecer una solución para satisfacer las necesidades del cliente, pero puede cambiar notablemente la demanda de la carga de trabajo y es probable que produzca un error.

Evalúe las cargas de trabajo críticas para determinar cuáles requieren este tipo de diseño de resiliencia. Se debe revisar cada componente de la aplicación en las cargas que se consideren cruciales. Algunos tipos de servicios que requieren evaluaciones de estabilidad estática son:

- Computación: Amazon EC2, EKS-EC2, ECS-EC2, EMR-EC2
- Bases de datos: Amazon Redshift, Amazon RDS, Amazon Aurora
- Storage (Almacenamiento): Amazon S3 (zona única), Amazon EFS (montajes), Amazon FSx (montajes)
- Equilibradores de carga: en ciertos diseños

Pasos para la implementación

- Cree cargas de trabajo que sean estables estáticamente y operen en un solo modo. En este caso, aprovisione suficientes instancias en cada región o zona de disponibilidad para gestionar la

capacidad de la carga de trabajo si se eliminara una región o zona de disponibilidad. Puede usar una variedad de servicios para el enrutamiento a recursos en buen estado, como:

  - Enrutamiento de DNS entre regiones
  - Enrutamiento de punto de acceso de varias regiones de Amazon S3
  - AWS Global Accelerator
  - Amazon Application Recovery Controller (ARC)
- Configure las réplicas de lectura de base de datos de modo que tengan en cuenta la pérdida de una única instancia principal o una réplica de lectura. Si las réplicas de lectura atienden el tráfico, la cantidad en cada zona de disponibilidad y cada región debe ser igual a la necesidad general en caso de que se produzca un error en la zona o región.
- Configure los datos esenciales en el almacenamiento Amazon S3 que está diseñado para ser estáticamente estable para los datos almacenados en caso de que se produzca un error en la zona de disponibilidad. Si se usa la clase de almacenamiento de acceso poco frecuente en una única zona de Amazon S3, no debe considerarse estable desde el punto de vista estático, ya que la pérdida de esa zona minimiza el acceso a los datos almacenados.
- Los equilibradores de carga a veces están configurados incorrectamente o por diseño para prestar servicio a una zona de disponibilidad específica. En este caso, el diseño estáticamente estable podría consistir en distribuir una carga de trabajo entre varias zonas de disponibilidad en un diseño más complejo. El diseño original se puede utilizar para reducir el tráfico entre zonas por motivos de seguridad, latencia o coste.

Recursos

Prácticas recomendadas por Well-Architected:

- Definición de disponibilidad
- REL11-BP01 Supervisar todos los componentes de la carga de trabajo para detectar errores
- REL11-BP04 Confiar en el plano de datos y no en el plano de control durante la recuperación

Documentos relacionados:

- Minimizar las dependencias en un plan de recuperación de desastres
- La Amazon Builders' Library: Estabilidad estática con zonas de disponibilidad
- Límites de aislamiento de errores
- Estabilidad estática con zonas de disponibilidad

- RDS multizona
- Minimizar las dependencias en un plan de recuperación de desastres
- Enrutamiento de DNS entre regiones
- Enrutamiento de punto de acceso de varias regiones de Amazon S3
- AWS Global Accelerator
- ARC
- Zona única de Amazon S3
- Equilibrio de carga entre zonas

Vídeos relacionados:

- Static stability in AWS: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328)

Ejemplos relacionados:

- La Amazon Builders' Library: Estabilidad estática con zonas de disponibilidad

REL11-BP06 Enviar notificaciones cuando los eventos afecten a la disponibilidad

Se envían notificaciones cuando se detecta que se han superado los umbrales, incluso si el evento que causó el problema se ha resuelto automáticamente.

La corrección automática permite que la carga de trabajo sea fiable. Sin embargo, también puede ocultar problemas subyacentes que deberían abordarse. Implemente una supervisión y unos eventos apropiados para poder detectar patrones de problemas, incluidos los que pueden abordarse mediante corrección automática, para que pueda resolver los problemas de la causa principal.

Los sistemas resilientes están diseñados para que los eventos de degradación se comuniquen inmediatamente a los equipos correspondientes. Estas notificaciones deben enviarse a través de uno o varios canales de comunicación.

Resultado deseado: las alertas se envían inmediatamente a los equipos de operaciones cuando se superan los umbrales, como las tasas de error, la latencia u otras métricas cruciales de los indicadores clave de rendimiento (KPI), para que estos problemas se resuelvan lo antes posible y se evite o minimice el impacto en los usuarios.

Patrones comunes de uso no recomendados:

- Enviar demasiadas alarmas.
- Enviar alarmas que no son procesables.
- Establecer umbrales de alarma demasiado altos (muy sensibles) o demasiado bajos (poco sensibles).
- No enviar alarmas para dependencias externas.
- No tener en cuenta los errores grises al diseñar la supervisión y las alarmas.
- Realizar la automatización de la reparación, pero sin notificar al equipo adecuado que se necesita una reparación.

Beneficios de establecer esta práctica recomendada: las notificaciones de recuperación permiten que los equipos operativos y empresariales estén al tanto de las degradaciones del servicio para que puedan reaccionar de inmediato y minimizar tanto el tiempo medio de detección (MTTD) como el tiempo medio de reparación (MTTR). Las notificaciones de los eventos de recuperación también garantizan que no se ignoren problemas que ocurren con poca frecuencia.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: Medio. Si no se implementan los mecanismos adecuados de supervisión y notificación de eventos, es posible que no se detecten patrones de problemas, incluidos los que se pueden abordarse mediante la corrección automática. El equipo solo se descubrirá la degradación del sistema cuando los usuarios contacten con el servicio de atención al cliente o por casualidad.

Guía para la implementación

Al definir una estrategia de supervisión, la activación de una alarma es un evento frecuente. Es probable que este evento contenga un identificador de la alarma, el estado de la alarma (como `En alarma` o bien `Aceptar`) y detalles sobre qué la desencadenó. En muchos casos, se debe detectar el evento de alarma y enviar una notificación por correo electrónico. Este es un ejemplo de una acción en una alarma. La notificación de alarmas es fundamental en la observabilidad, ya que informa a las personas adecuadas de que existe un problema. Sin embargo, cuando la acción sobre los eventos madura en su solución de observabilidad, puede solucionar el problema automáticamente sin necesidad de intervención humana.

Una vez que se hayan establecido las alarmas de supervisión de los KPI, se deben enviar alertas a los equipos correspondientes cuando se superen los umbrales. Esas alertas también se pueden usar para activar procesos automatizados que intentarán corregir la degradación.

Para una supervisión de umbrales más compleja, se deben considerar las alarmas compuestas. Las alarmas compuestas utilizan una serie de alarmas de supervisión de KPI para crear una alerta basada en la lógica empresarial operativa. Las alarmas de CloudWatch se pueden configurar para enviar correos electrónicos o para registrar incidentes en sistemas de seguimiento de incidentes de terceros mediante la integración con Amazon SNS o Amazon EventBridge.

Pasos para la implementación

Cree varios tipos de alarmas en función de la forma en que se supervisan las cargas de trabajo, como por ejemplo:

- Las alarmas de las aplicaciones se utilizan para detectar cuando alguna parte de la carga de trabajo no funciona correctamente.
- Las alarmas de infraestructura indican cuándo escalar los recursos. Las alarmas se pueden mostrar visualmente en paneles, enviar alertas a través de Amazon SNS o por correo electrónico y trabajar con Auto Scaling para aumentar o reducir los recursos de la carga de trabajo.
- Se pueden crear alarmas estáticas simples para supervisar cuando una métrica supera un umbral estático durante un número específico de períodos de evaluación.
- Las alarmas compuestas pueden abarcar alarmas complejas de numerosas fuentes.
- Una vez creada la alarma, cree los eventos de notificación adecuados. Puede invocar directamente una API de Amazon SNS para enviar notificaciones y vincular cualquier automatización para su corrección o comunicación.
- Integrar Amazon Health Aware para poder supervisar la visibilidad de los recursos de AWS que podrían estar degradados. Para las cargas de trabajo empresariales esenciales, esta solución proporciona acceso a alertas proactivas y en tiempo real para los servicios de AWS.

Recursos

Prácticas recomendadas por Well-Architected:

- Definición de disponibilidad

Documentos relacionados:

- Cree una alarma de CloudWatch basada en un umbral estático
- What is Amazon EventBridge?
- ¿Qué es Amazon Simple Notification Service?

- Publicar métricas personalizadas
- Using Amazon CloudWatch Alarms
- Amazon Health Aware (AHA)
- Setup CloudWatch Composite alarms
- What's new in AWS Observability at re:Invent 2022

Herramientas relacionadas:

- CloudWatch
- CloudWatch X-Ray

REL11-BP07 Diseñar su producto para cumplir objetivos de disponibilidad y acuerdos de nivel de servicio (SLA) de tiempo de actividad

Diseñe su producto para que cumpla objetivos de disponibilidad y acuerdos de nivel de servicio (SLA) de tiempo de actividad. Si publica o acuerda en privado objetivos de disponibilidad o SLA de tiempo de actividad, verifique que su arquitectura y procesos operativos están diseñados para darles cabida.

Resultado deseado: cada aplicación tiene un objetivo definido de disponibilidad y un SLA para las métricas de rendimiento, que pueden supervisarse y mantenerse para alcanzar los resultados empresariales.

Antipatrones usuales:

- Diseño y despliegue de cargas de trabajo sin establecer acuerdos de nivel de servicio.
- Las métricas de los SLA se fijan en niveles altos sin justificación ni requisitos empresariales.
- Establecimiento de SLA sin tener en cuenta las dependencias y sus SLA subyacentes.
- Los diseños de aplicaciones se crean sin tener en cuenta el modelo de responsabilidad compartida para la resiliencia.

Beneficios de establecer esta práctica recomendada: el diseño de aplicaciones basado en objetivos clave de resistencia ayuda a cumplir los objetivos empresariales y las expectativas de los clientes. Estos objetivos contribuyen a impulsar el proceso de diseño de aplicaciones que evalúa diferentes tecnologías y tiene en cuenta varios compromisos.

Guía para la implementación

El diseño de aplicaciones debe tener en cuenta una serie de requisitos derivados de objetivos empresariales, operativos y financieros. Dentro de los requisitos operativos, las cargas de trabajo deben tener objetivos concretos de métricas de resistencia para que se puedan supervisar y respaldar adecuadamente. Las métricas de resistencia no deben establecerse ni derivarse después de desplegar la carga de trabajo. En cambio, deben definirse durante la fase de diseño y ayudar a orientar diversas decisiones y compromisos.

- Cada carga de trabajo debe tener su propio conjunto de métricas de resistencia. Esas métricas pueden ser diferentes de las de otras aplicaciones empresariales.
- Reducir las dependencias puede tener un efecto positivo en la disponibilidad. Cada carga de trabajo debe considerar sus dependencias y sus SLA. En general, seleccione dependencias con objetivos de disponibilidad iguales o superiores a los objetivos de su carga de trabajo.
- Siempre que sea posible, examine diseños de acoplamiento flexible para que la carga de trabajo pueda funcionar correctamente a pesar del deterioro de las dependencias.
- Reduzca las dependencias del plano de control, especialmente durante la recuperación o una degradación. Evalúe diseños que sean estáticamente estables para las cargas de trabajo cruciales para la misión. Utilice el ahorro de recursos para aumentar la disponibilidad de esas dependencias en una carga de trabajo.
- La observabilidad y la instrumentación son fundamentales para alcanzar los SLA al reducir el tiempo medio de detección (MTTD) y el tiempo medio de reparación (MTTR).
- Los tres factores que se utilizan para mejorar la disponibilidad en los sistemas distribuidos son menos errores frecuentes (MTBF más largo), tiempos de detección de errores más cortos (MTTD más corto) y tiempos de reparación más cortos (MTTR más corto).
- Establecer y cumplir las métricas de resistencia para una carga de trabajo es un elemento imprescindible en todo diseño eficaz. Estos diseños deben tener en cuenta los compromisos de la complejidad del diseño, las dependencias de los servicios, el rendimiento, la escalabilidad y los costes.

Pasos para la implementación

- Revise y documente el diseño de la carga de trabajo teniendo presente las siguientes preguntas:
  - ¿Dónde se utilizan los planos de control en la carga de trabajo?
  - ¿Cómo implementa la carga de trabajo la tolerancia a errores?

  - ¿Cuáles son los patrones de diseño para el escalado, el escalado automático, la redundancia y los componentes de alta disponibilidad?
  - ¿Cuáles son los requisitos de coherencia y disponibilidad de los datos?
  - ¿Se tiene en cuenta el ahorro de recursos o la estabilidad estática de los recursos?
  - ¿Cuáles son las dependencias de los servicios?
- Defina las métricas de los SLA basándose en la arquitectura de la carga de trabajo mientras trabaja con las partes interesadas. Considere los SLA de todas las dependencias utilizadas por la carga de trabajo.
- Una vez establecido el objetivo del SLA, optimice la arquitectura para que cumpla el SLA.
- Una vez establecido el diseño que cumplirá el SLA, implemente cambios operativos, automatización de procesos y runbooks que también se centren en reducir el MTTD y el MTTR.
- Una vez desplegado, supervise y cree informes del SLA.

Recursos

Prácticas recomendadas relacionadas:

- REL03-BP01 Elegir cómo segmentar su carga de trabajo
- REL10-BP01 Implementar la carga de trabajo en varias ubicaciones
- REL11-BP01 Supervisar todos los componentes de la carga de trabajo para detectar errores
- REL11-BP03 Automatizar la reparación en todas las capas
- REL12-BP05 Probar la resiliencia mediante la ingeniería del caos
- REL13-BP01 Definir objetivos de recuperación para la inactividad y la pérdida de datos
- Comprender el estado de las cargas de trabajo

Documentos relacionados:

- Disponibilidad con redundancia
- Pilar de fiabilidad: disponibilidad
- Measuring availability (Medición de la disponibilidad)
- Límites de aislamiento de errores de AWS
- Shared Responsibility Model for Resiliency (Modelo de responsabilidad compartida para la resiliencia)

- Estabilidad estática con zonas de disponibilidad
- Acuerdos de nivel de servicios (SLA) de AWS
- Guidance for Cell-based Architecture on AWS (Guía para la arquitectura basada en celdas en AWS)
- AWS infrastructure (Infraestructura de AWS)
- Advanced Multi-AZ Resiliance Patterns whitepaper (Documento técnico sobre patrones avanzados de resistencia multi-AZ)

Servicios relacionados:

- Amazon CloudWatch
- AWS Config
- AWS Trusted Advisor

FIABILIDAD 12. ¿Cómo pone a prueba la fiabilidad?

Una vez diseñada la carga de trabajo para que sea resiliente al estrés de producción, las pruebas son la única forma de comprobar que funcionará según lo previsto y proporcionará la resiliencia esperada.

Prácticas recomendadas

- REL12-BP01 Usar guías de estrategias para investigar los errores
- REL12-BP02 Realizar un análisis después del incidente
- REL12-BP03 Comprobar los requisitos funcionales
- REL12-BP04 Requisitos de escalado y rendimiento de las pruebas
- REL12-BP05 Probar la resiliencia mediante la ingeniería del caos
- REL12-BP06 Planificación regular de días de juego

REL12-BP01 Usar guías de estrategias para investigar los errores

Puede obtener respuestas sistemáticas e inmediatas a escenarios de error que no se entiendan bien documentando el proceso de investigación en guías de estrategias. Las guías de estrategias son pasos predefinidos realizados para identificar los factores que contribuyen a un escenario de error. Los resultados de cualquier paso del proceso se utilizan para determinar los siguientes pasos, hasta que el problema se haya identificado o deba derivarse.

Las guías de estrategias implican una planificación proactiva que debe llevar a cabo para poder emprender acciones reactivas de forma eficaz. Cuando se encuentran en producción casos de error que no están contemplados en la guía de estrategias, primero debe solucionar el problema (apagar el fuego). Luego, deberá volver y analizar los pasos que ha seguido para abordar el problema y, sobre ellos, añadir una nueva entrada en la guía.

Tenga en cuenta que las guías de estrategias se usan en respuesta a incidentes específicos y los runbooks se usan para conseguir resultados determinados. A menudo, los runbooks se usan para actividades rutinarias, mientras que las guías de estrategias se utilizan para responder a eventos no rutinarios.

Antipatrones usuales:

- Planificar la implementación de una carga de trabajo sin conocer los procesos para diagnosticar los problemas o responder a los incidentes
- Decisiones no planificadas sobre de qué sistemas se recopilan registros y métricas cuando se investiga un evento
- No conservar las métricas y los eventos el tiempo suficiente para poder recuperar los datos

Beneficios de establecer esta práctica recomendada: La captura de esta información en guías de estrategias garantiza que el proceso pueda seguirse sistemáticamente. La creación de guías de estrategias limita la introducción de errores de la actividad manual. La automatización de guías de estrategias reduce el tiempo para responder a un evento al eliminar el requisito de intervención de un miembro del equipo o al disponer de información adicional al inicio de su intervención.

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

Guía para la implementación

- Use guías de estrategias para identificar problemas. Las guías de estrategias son procesos documentados para investigar problemas. Permita las respuestas sistemáticas e inmediatas a escenarios de error documentando los procesos en guías de estrategias. Las guías de estrategias deben contener la información y las instrucciones necesarias para que alguien con la formación adecuada reúna la información correspondiente, identifique las posibles fuentes de error, aísle los errores y determine los factores que han contribuido al problema (realizar un análisis después del incidente).
  - Implemente en código las guías de estrategias. Realice sus operaciones como código creando scripts de sus guías de estrategias para garantizar la sistematicidad y reducir los errores

causados por los procesos manuales. Las guías de estrategias pueden constar de varios scripts que representen los diferentes pasos que podrían ser necesarios para identificar los factores que contribuyen a un problema. Se pueden programar o realizar actividades de runbook como parte de las actividades de una guía de estrategias, o se puede solicitar la ejecución de una guía de estrategias en respuesta a eventos identificados.

- Automatizar las guías de estrategias operativas con AWS Systems Manager
- AWS Systems Manager Run Command
- AWS Systems Manager Automation
- ¿Qué es AWS Lambda?
- ¿Qué es Amazon EventBridge?
- Uso de alarmas de Amazon CloudWatch

Recursos

Documentos relacionados:

- AWS Systems Manager Automation
- AWS Systems Manager Run Command
- Automatizar las guías de estrategias operativas con AWS Systems Manager
- Uso de alarmas de Amazon CloudWatch
- Uso de valores controlados (Amazon CloudWatch Synthetics)
- ¿Qué es Amazon EventBridge?
- ¿Qué es AWS Lambda?

Ejemplos relacionados:

- Automatización de operaciones con guías de estrategias y runbooks

REL12-BP02 Realizar un análisis después del incidente

Revise los eventos que afectan a los clientes e identifique los factores que contribuyen al evento y las medidas preventivas. Use esta información para desarrollar un plan de mitigación que limite o evite la reaparición del problema. Desarrolle procedimientos para proporcionar respuestas rápidas y eficaces. Comunique los factores que han contribuido al problema y las medidas correctivas según

corresponda, adaptados al público de destino. Disponga de un método para comunicar estas causas a otros usuarios según sea necesario.

Evalúe por qué las pruebas existentes no han detectado el problema. Añada pruebas para este caso si no hay pruebas ya establecidas.

Resultado deseado: sus equipos tienen un enfoque uniforme y consensuado para gestionar el análisis posterior a los incidentes. Uno de los mecanismos es el proceso de corrección de errores (COE). El proceso COE ayuda a sus equipos a identificar, comprender y abordar las causas fundamentales de los incidentes, a la vez que crea mecanismos y barreras de protección para limitar la probabilidad de que se repita el mismo incidente.

Antipatrones usuales:

- Buscar los factores que han contribuido al problema, pero no seguir investigando si existen otros problemas potenciales o enfoques que mitigar
- Identificar solo los errores humanos y no proporcionar ninguna formación o automatización que pueda evitar estos errores
- Concentrarse en determinar la culpa en lugar de en conocer la causa raíz, lo que da lugar a una cultura de miedo y obstaculiza la comunicación abierta
- Falta de intercambio de ideas, lo que hace que los resultados del análisis de incidentes los conozca solo un grupo pequeño e impide que otros se beneficien de las lecciones aprendidas
- No tener ningún mecanismo para capturar el conocimiento institucional, por lo que se pierde información valiosa al no preservar las lecciones aprendidas en forma de actualizaciones de las prácticas recomendadas y, por lo tanto, se repiten incidentes con la misma causa raíz o una similar

Ventajas de establecer esta práctica recomendada: realizar análisis después de un incidente y compartir los resultados permite que el riesgo se mitigue en otras cargas de trabajo si estas tienen implementados los mismos factores que han contribuido al problema, y permite también implementar la mitigación o la recuperación automatizada antes de que se produzca un incidente.

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

Guía para la implementación

Un buen análisis posterior a un incidente ofrece oportunidades de proponer soluciones comunes para problemas con patrones arquitectónicos que se utilizan en otros lugares de los sistemas.

Una piedra angular del proceso COE es documentar y abordar los problemas. Es recomendable definir una forma estandarizada de documentar las causas raíz críticas y asegurarse de que estas se revisan y solucionan. Asigne una propiedad clara al proceso de análisis posterior al incidente. Designe a un equipo o persona responsable que supervise las investigaciones y el seguimiento de los incidentes.

Fomente una cultura que se centre en el aprendizaje y la mejora en lugar de en la culpa. Haga hincapié en que el objetivo es prevenir futuros incidentes, no penalizar a las personas.

Desarrolle procedimientos bien definidos para realizar análisis posteriores a los incidentes. En estos procedimientos, se deben describir los pasos que se deben seguir, la información que se va a recopilar y las preguntas clave que se abordarán durante el análisis. Investigue los incidentes a fondo y vaya más allá de las causas inmediatas para identificar las causas raíz y los factores que contribuyen a ellos. Use técnicas como los cinco porqués para profundizar en los problemas subyacentes.

Mantenga un repositorio de las lecciones aprendidas de los análisis de incidentes. Este conocimiento institucional puede servir como referencia para futuros incidentes y esfuerzos de prevención. Comparta las conclusiones y los conocimientos de los análisis posteriores a los incidentes y considere la posibilidad de celebrar reuniones de revisión de invitación abierta después de los incidentes para analizar las lecciones aprendidas.

Pasos para la implementación

- Al realizar un análisis posterior al incidente, asegúrese de que en el proceso no se culpe a nadie. Esto permite que las personas involucradas en el incidente se muestren imparciales con respecto a las medidas correctivas propuestas, además de fomentar una autoevaluación honesta y la colaboración entre los equipos.
- Defina una forma estandarizada de documentar los problemas críticos. Un ejemplo de estructura para dicho documento es el siguiente:
  - ¿Qué ha ocurrido?
  - ¿Cómo ha afectado a los clientes y a la empresa?
  - ¿Cuál ha sido la causa raíz?
  - ¿Qué datos tiene para corroborarlo?
    - Por ejemplo, métricas y gráficos
  - ¿Qué pilares básicos estuvieron implicados, con especial atención a la seguridad?

    - Al diseñar cargas de trabajo, se hacen concesiones entre pilares según el contexto del negocio. Estas decisiones de negocios pueden impulsar sus prioridades de ingeniería. Podría dar preferencia a reducir el costo a expensas de la fiabilidad en el desarrollo de entornos o, para soluciones de misión crítica, podría optimizar la fiabilidad con costos aumentados. La seguridad siempre es la tarea primordial, ya que sus clientes deben estar protegidos.
  - ¿Qué lecciones aprendió?
  - ¿Qué medidas correctivas está tomando?
    - Medidas
    - Artículos relacionados
- Cree procedimientos operativos estándar bien definidos para realizar análisis posteriores a los incidentes.
- Configure un proceso estandarizado de notificación de incidentes. Documente todos los incidentes de manera exhaustiva, incluido el informe inicial del incidente, los registros, las comunicaciones y las medidas tomadas durante el incidente.
- Recuerde que un incidente no requiere una interrupción. Podría tratarse de un cuasi incidente o de un sistema que funciona de una forma inesperada, pero que cumple su función.
- Mejore continuamente su proceso de análisis posterior a un incidente en función de los comentarios y las lecciones aprendidas.
- Registre los resultados clave en un sistema de administración del conocimiento y considere cualquier patrón que deba añadirse a las guías para desarrolladores o a las listas de verificación previas al despliegue.

Recursos

Documentos relacionados:

- «Why you should develop a correction of error (COE)»

Vídeos relacionados:

- «Amazon’s approach to failing successfully»
- «AWS re:Invent 2021 - Amazon Builders’ Library: Operational Excellence at Amazon»

REL12-BP03 Comprobar los requisitos funcionales

Use técnicas como pruebas unitarias y pruebas de integración que validen la funcionalidad necesaria.

Conseguirá los mejores resultados cuando estas pruebas se lleven a cabo automáticamente como parte de las acciones de compilación y despliegue. Por ejemplo, al utilizar AWS CodePipeline, los desarrolladores confirman los cambios en un repositorio de origen en el que CodePipeline detecta los cambios automáticamente. Esos cambios se incorporan y se realizan pruebas. Una vez completadas las pruebas, el código compilado se implementa en los servidores provisionales para comprobarlo. Desde el servidor provisional, CodePipeline ejecuta más pruebas, como pruebas de integración o carga. Una vez completadas correctamente esas pruebas, CodePipeline implementa el código comprobado y aprobado en instancias de producción.

Además, la experiencia demuestra que las pruebas de transacciones sintéticas (denominadas también pruebas de valores controlados, que no deben confundirse con las implementaciones de valores controlados) que puedan ejecutar y simular el comportamiento de los clientes son uno de los procesos de prueba más importantes. Ejecute estas pruebas constantemente en los puntos de conexión de las cargas de trabajo desde distintas ubicaciones remotas. Amazon CloudWatch Synthetics le permite crear valores controlados para supervisar sus puntos de conexión y API.

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

Guía para la implementación

- Compruebe los requisitos funcionales. Entre estas se incluyen las pruebas unitarias y las pruebas de integración que validan la funcionalidad necesaria.
  - Use CodePipeline y AWS CodeBuild para probar el código y ejecutar compilaciones
  - AWS CodePipeline añade compatibilidad a las pruebas unitarias y de integración personalizadas con AWS CodeBuild
  - Entrega continua e integración continua
  - Uso de valores controlados (Amazon CloudWatch Synthetics)
  - Automatización de pruebas de software

Recursos

Documentos relacionados:

- Socio de APN: socios que pueden ayudar con la implementación de una canalización de integración continua
- AWS CodePipeline añade compatibilidad a las pruebas unitarias y de integración personalizadas con AWS CodeBuild
- AWS Marketplace: productos que pueden usarse para la integración continua
- Entrega continua e integración continua
- Automatización de pruebas de software
- Use CodePipeline y AWS CodeBuild para probar el código y ejecutar compilaciones
- Uso de valores controlados (Amazon CloudWatch Synthetics)

REL12-BP04 Requisitos de escalado y rendimiento de las pruebas

Use técnicas como las pruebas de carga para validar que la carga de trabajo satisface los requisitos de escalado y rendimiento.

En la nube, puede crear un entorno de pruebas a escala de producción bajo demanda para su carga de trabajo. Si ejecuta estas pruebas en una infraestructura desescalada verticalmente, debe escalar los resultados observados a lo que cree que ocurrirá en producción. Las pruebas de carga y rendimiento también pueden realizarse en producción si se tiene cuidado de no afectar a los usuarios reales y se etiquetan los datos de prueba para que no se mezclen con los datos de usuarios reales y alteren las estadísticas de uso o los informes de producción.

Con las pruebas, asegúrese de que sus recursos base, la configuración de escalado, las cuotas de servicio y el diseño de resiliencia funcionan del modo esperado bajo carga.

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

Guía para la implementación

- Pruebe los requisitos de escalado y de rendimiento. Realice pruebas de carga para validar que la carga de trabajo satisface los requisitos de escalado y rendimiento.
  - Pruebas de carga distribuida en AWS: simular miles de usuarios conectados
  - Apache JMeter
    - Implemente la aplicación en un entorno idéntico al de producción y ejecute una prueba de carga.
      - Utilice conceptos de infraestructura como código para crear un entorno tan similar al entorno de producción como sea posible.

Recursos

Documentos relacionados:

- Pruebas de carga distribuida en AWS: simular miles de usuarios conectados
- Apache JMeter

REL12-BP05 Probar la resiliencia mediante la ingeniería del caos

Realice experimentos de caos con regularidad en entornos que estén en producción o lo más cerca posible de ella para entender cómo responde su sistema a condiciones adversas.

Resultado deseado:

La resiliencia de la carga de trabajo se verifica regularmente aplicando la ingeniería del caos en forma de experimentos de inyección de errores o inyección de carga inesperada, además de las pruebas de resiliencia que validan el comportamiento esperado conocido de su carga de trabajo durante un evento. Combine la ingeniería del caos y las pruebas de resiliencia para tener la seguridad de que su carga de trabajo puede sobrevivir a los errores de los componentes y puede recuperarse de las interrupciones inesperadas con un impacto mínimo o nulo.

Patrones comunes de uso no recomendados:

- Diseñar para lograr la resiliencia, pero no verificar cómo funciona la carga de trabajo en su conjunto cuando se producen errores.
- No experimentar nunca en condiciones reales y con la carga prevista.
- No tratar los experimentos como código ni mantenerlos durante el ciclo de desarrollo.
- No ejecutar experimentos de caos tanto como parte de su canalización de CI/CD, así como fuera de los despliegues.
- No utilizar los análisis posteriores a los incidentes a la hora de determinar los errores con los que experimentar.

Beneficios de establecer esta práctica recomendada: Inyectar errores para verificar la resiliencia de la carga de trabajo permite ganar confianza sobre el hecho de que los procedimientos de recuperación de su diseño resiliente funcionarán en caso de un error real.

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

Guía para la implementación

La ingeniería del caos proporciona a sus equipos capacidades para inyectar continuamente interrupciones del mundo real (simulaciones) de forma controlada a nivel de proveedor de servicios, infraestructura, carga de trabajo y componentes, con un impacto mínimo o nulo para sus clientes. Permite que sus equipos aprendan de los fallos y observen, midan y mejoren la resiliencia de sus cargas de trabajo, además de validar que las alertas se disparen y que los equipos reciban notificaciones en caso de algún evento.

Cuando se realiza de forma continua, la ingeniería del caos puede poner de manifiesto deficiencias en sus cargas de trabajo que, si no se abordan, podrían afectar negativamente la disponibilidad y al funcionamiento.

> **Note**
>
> La ingeniería del caos es la disciplina que consiste en experimentar en un sistema para generar confianza en la capacidad del sistema de resistir condiciones adversas en producción. Principios de la ingeniería del caos

Si un sistema es capaz de soportar estas interrupciones, el experimento del caos debería mantenerse como una prueba de regresión automatizada. De este modo, los experimentos de caos deben realizarse como parte de su ciclo de vida de desarrollo de sistemas (SDLC) y como parte de su canalización de CI/CD.

Para asegurarse de que su carga de trabajo puede sobrevivir a los errores de los componentes, inyecte eventos del mundo real como parte de sus experimentos. Por ejemplo, experimente con la pérdida de instancias de Amazon EC2 o la conmutación por error de la instancia primaria de la base de datos de Amazon RDS y verifique que su carga de trabajo no se ve afectada (o solo mínimamente). Utilice una combinación de errores de componentes para simular los eventos que puede causar una interrupción en una zona de disponibilidad.

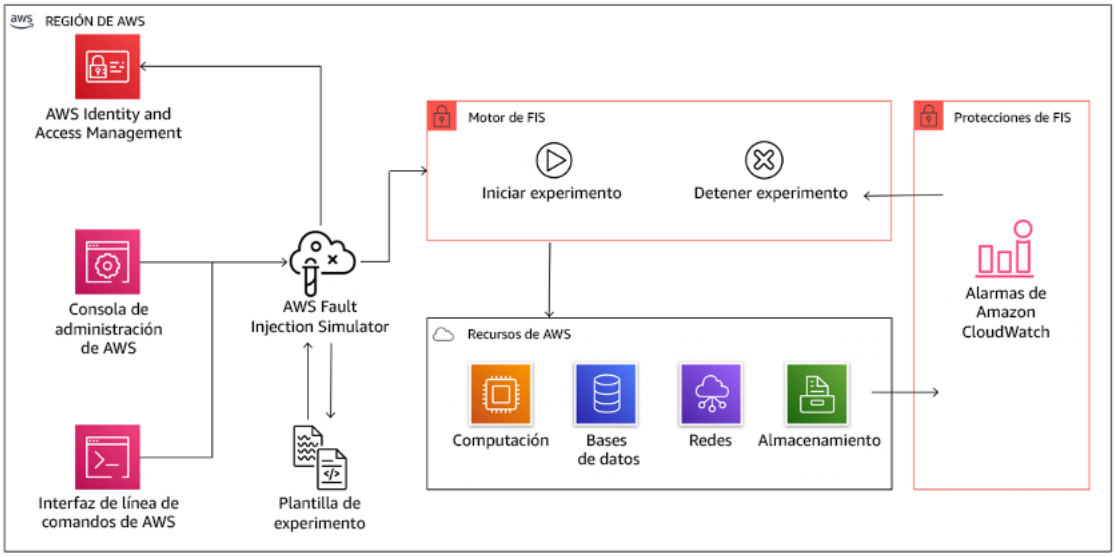
Para los errores a nivel de aplicación (como las caídas), se puede empezar con factores de estrés como el agotamiento de la memoria y la CPU.

Para validar los mecanismos de recuperación o de conmutación por error para las dependencias externas debido a interrupciones intermitentes de la red, sus componentes deben simular un evento de este tipo bloqueando el acceso a los proveedores de terceros durante una duración especificada que puede durar desde segundos hasta horas.

Otros modos de degradación podrían provocar una funcionalidad reducida y respuestas lentas, lo que a menudo da como resultado una interrupción de sus servicios. Las fuentes comunes de esta degradación son una mayor latencia en los servicios críticos y una comunicación de red poco fiable (paquetes omitidos). Los experimentos con estos errores, que incluyen efectos de red como la latencia, los mensajes perdidos y los errores de DNS, podrían incluir la incapacidad de resolver un nombre, alcanzar el servicio DNS o establecer conexiones con servicios dependientes.

Herramientas de ingeniería del caos:

AWS Fault Injection Service (AWS FIS) es un servicio completamente administrado para realizar experimentos de inserción de errores que puede utilizarse como parte de su canalización de CD. AWS FIS es una buena opción para usar durante los días de juego de ingeniería del caos. Admite la introducción simultánea de errores en diferentes tipos de recursos, como Amazon EC2, Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS) y Amazon RDS. Estos errores incluyen la terminación de los recursos, forzado de conmutación por error, estrés de CPU o memoria, limitación, latencia y pérdida de paquetes. Al estar integrado con Amazon CloudWatch Alarms, puede configurar las condiciones de parada como barreras de protección para revertir un experimento si provoca un impacto inesperado.



AWS Fault Injection Service se integra con recursos de AWS para permitirle ejecutar experimentos de inserción de errores para sus cargas de trabajo.

También hay varias opciones de terceros para los experimentos de inserción de errores. Incluyen herramientas de código abierto como Chaos Toolkit, Chaos Mesh y Litmus Chaos, además de opciones comerciales como Gremlin. Para ampliar el alcance de los errores que se pueden inyectar en AWS, AWS FIS se integra con Chaos Mesh y Litmus Chaos, lo que le permite coordinar los

flujos de trabajo de inyección de errores entre varias herramientas. Por ejemplo, puede ejecutar una prueba de estrés en la CPU de un pod utilizando errores de Chaos Mesh o Litmus mientras termina un porcentaje seleccionado al azar de nodos del clúster utilizando acciones de error de AWS FIS.

Pasos para la aplicación

- Determine qué errores se van a utilizar en los experimentos.

  Evalúe el diseño de su carga de trabajo para la resiliencia. Estos diseños (creados utilizando las prácticas recomendadas del Marco de buena arquitectura) tienen en cuenta los riesgos basados en las dependencias críticas, los eventos pasados, los problemas conocidos y los requisitos de cumplimiento. Enumere cada elemento del diseño destinado a mantener la resiliencia y los errores que pretende mitigar. Para obtener más información sobre la creación de estas listas, consulte el documento técnico Revisión de la preparación operativa que le orienta sobre cómo crear un proceso para evitar que se repitan incidentes anteriores. El proceso de análisis de modos de error y efectos (AMFE) le proporciona un marco para realizar un análisis de los fallos a nivel de componente y cómo afectan a su carga de trabajo. Adrian Cockcroft describe con más detalle el AMFE en Failure Modes and Continuous Resilience (Modos de error y resiliencia continua).
- Asigne una prioridad a cada error.

  Comience con una categorización gruesa como alta, media o baja. Para evaluar la prioridad, hay que tener en cuenta la frecuencia del error y el impacto del mismo en la carga de trabajo global.

  Al considerar la frecuencia de un determinado error, analice los datos anteriores de esta carga de trabajo cuando estén disponibles. Si no están disponibles, utilice datos de otras cargas de trabajo que se ejecuten en un entorno similar.

  Cuando se considera el impacto de un error determinado, cuanto mayor sea el alcance del error, generalmente mayor será el impacto. También hay que tener en cuenta el diseño y la finalidad de la carga de trabajo. Por ejemplo, la capacidad de acceder a los almacenes de datos de origen es fundamental para una carga de trabajo que realice transformaciones y análisis de datos. En este caso, se daría prioridad a los experimentos de errores de acceso, así como al acceso limitado y a la inserción de latencia.

  Los análisis posteriores a los incidentes son un buena fuente de datos para comprender tanto la frecuencia como el impacto de los modos de error.

  Utilice la prioridad asignada para determinar con qué fallos experimentar primero y el orden con el que desarrollar nuevos experimentos de inyección de errores.

- En cada experimento que realice, siga la ingeniería del caos y el volante de resiliencia continua.

Ingeniería del caos y volante de resiliencia continua, utilizando el método científico de Adrian Hornsby.

- Defina el estado estable como un resultado medible de una carga de trabajo que indica un comportamiento normal.

  Su carga de trabajo exhibe un estado estable si está operando de manera fiable y como se espera. Por tanto, valide que su carga de trabajo tenga un buen estado antes de definir el estado estable. El estado estable no significa necesariamente que no haya impacto en la carga de trabajo cuando se produce un error, ya que un cierto porcentaje en los errores podría estar dentro de los límites aceptables. El estado estable es la línea de base que observará durante el experimento, que pondrá de manifiesto las anomalías si su hipótesis definida en el siguiente paso no resulta de la forma esperada.

Por ejemplo, un estado estable de un sistema de pagos puede definirse como el procesamiento de 300 TPS con una tasa de éxito del 99 % y un tiempo de ida y vuelta de 500 ms.

- Formule una hipótesis sobre cómo reaccionará la carga de trabajo ante el error.

  Una buena hipótesis se basa en cómo se espera que la carga de trabajo mitigue el error para mantener el estado estable. La hipótesis establece que dado el error de un tipo específico, el sistema o la carga de trabajo continuará en estado estable, porque la carga de trabajo fue diseñada con mitigaciones específicas. En la hipótesis deben especificarse el tipo específico de error y las mitigaciones.

  Se puede utilizar la siguiente plantilla para la hipótesis (pero también se acepta otra redacción):

  > **Note**
  >
  > Si se produce el *`error específico`* , la carga de trabajo *`nombre de carga de trabajo`* , *`describirá los controles mitigantes`* para mantener el *`impacto de las métricas empresariales o técnicas`*.

  Por ejemplo:

  - Si el 20 % de los nodos del grupo de nodos de Amazon EKS se caen, la API de creación de transacciones sigue sirviendo el percentil 99 de peticiones en menos de 100 ms (estado estable). Los nodos de Amazon EKS se recuperarán en cinco minutos, y los pods se programarán y procesarán el tráfico en ocho minutos tras el inicio del experimento. Las alertas se disparan en tres minutos.
  - Si se produce un error de instancia de Amazon EC2, la comprobación de estado de Elastic Load Balancing del sistema de pedidos hará que Elastic Load Balancing solo envíe solicitudes a las instancias en buen estado restantes mientras Amazon EC2 Auto Scaling sustituye la instancia con error, manteniendo un aumento inferior al 0,01 % en los errores del lado del servidor (5xx) (estado estable).
  - Si la instancia de la base de datos primaria de Amazon RDS falla, la carga de trabajo de recopilación de datos de la cadena de suministro se conmutará por error y se conectará a la instancia de la base de datos de Amazon RDS en espera para mantener menos de 1 minuto de errores de lectura o escritura en la base de datos (estado estable).

- Realiza el experimento inyectando el error.

Un experimento debería ser por defecto a prueba de errores y tolerado por la carga de trabajo. Si sabe que la carga de trabajo va a fallar, no realice el experimento. Debe utilizarse la ingeniería del caos para encontrar conocidos-desconocidos o desconocidos-desconocidos. Conocidos-desconocidos son cosas de las que es consciente pero no comprende del todo, y desconocidos-desconocidos son cosas de las que no es consciente ni comprende del todo. Experimentar con una carga de trabajo que sabe que está rota no proporcionará nuevas ideas. Su experimento debe estar cuidadosamente planificado, tener un alcance claro de impacto y proporcionar un mecanismo de retroceso que pueda aplicarse en caso de turbulencias inesperadas. Si su diligencia demuestra que su carga de trabajo debería sobrevivir al experimento, siga adelante con el mismo. Hay varias opciones para inyectar los errores. Para cargas de trabajo en AWS, AWS FIS proporciona muchas simulaciones de errores predefinidas llamadas acciones. También puede definir acciones personalizadas que se ejecuten en AWS FIS utilizando documentos de AWS Systems Manager.

Desaconsejamos el uso de scripts personalizados para los experimentos de caos, a menos que los scripts tengan la capacidad de entender el estado actual de la carga de trabajo, sean capaces de emitir registros y proporcionen mecanismos para retrocesos y condiciones de parada cuando sea posible.

Un marco o conjunto de herramientas eficaz que apoye la ingeniería del caos debe hacer el seguimiento del estado actual de un experimento, emitir registros y proporcionar mecanismos de reversión para apoyar la ejecución controlada de un experimento. Comience con un servicio establecido como AWS FIS que permite realizar experimentos con un alcance claramente definido y mecanismos de seguridad que reviertan el experimento en el caso de que introduzca turbulencias inesperadas. Para conocer una mayor variedad de experimentos con AWS FIS, consulte también el Laboratorio de Aplicaciones resilientes y bien diseñadas con ingeniería del caos. Además, AWS Resilience Hub analizará su carga de trabajo y creará experimentos que puede elegir para implementar y ejecutar en AWS FIS.

> **Note**
>
> Para cada experimento, comprenda claramente el alcance y su impacto. Recomendamos que los fallos se simulen primero en un entorno no productivo antes de ejecutarlos en producción.

Los experimentos deben realizarse en producción bajo carga real utilizando despliegue de valores controlados que acelera tanto el despliegue de un sistema de control como el experimental, cuando es factible. Ejecutar los experimentos durante las horas de menor actividad es una buena práctica para mitigar el impacto potencial cuando se experimenta por primera vez en producción. Además, si utilizar el tráfico real del cliente supone demasiado riesgo, puede realizar experimentos utilizando tráfico sintético en la infraestructura de producción contra los despliegues de control y experimentales. Cuando no sea posible utilizar la producción, ejecute los experimentos en entornos de preproducción que sean lo más parecidos posible a la producción.

Debe establecer y supervisar las barreras de seguridad para garantizar que el experimento no afecte al tráfico de producción o a otros sistemas más allá de los límites aceptables. Establezca condiciones de parada para detener un experimento si alcanza un umbral en una métrica de barrera que defina. Esto debería incluir las métricas para el estado estable de la carga de trabajo, así como la métrica contra los componentes en los que está inyectando el error. Una monitorización sintética (también conocida como valor controlado) es una métrica que normalmente debería incluir como proxy de usuario. Las condiciones de parada para AWS FIS se admiten como parte de la plantilla del experimento, permitiendo hasta cinco condiciones de parada por plantilla.

Uno de los principios del caos es minimizar el alcance del experimento y su impacto:

Aunque hay que tener en cuenta algún impacto negativo a corto plazo, es responsabilidad y obligación del ingeniero del caos garantizar que las consecuencias de los experimentos se minimicen y contengan.

Un método para verificar el alcance y el impacto potencial es realizar el experimento primero en un entorno de no producción, verificando que los umbrales para las condiciones de parada se activan como se espera durante un experimento y la observabilidad está implantada para detectar una excepción, en lugar de experimentar directamente en la producción.

Cuando se realicen experimentos de inyección de errores, verifique que todas las partes responsables estén bien informadas. Comuníquese con los equipos adecuados, como los equipos de operaciones, los equipos de fiabilidad del servicio y el servicio de atención al cliente, para informarles de cuándo se llevarán a cabo los experimentos y qué pueden esperar. Proporcione a estos equipos herramientas de comunicación para que informen a los que dirigen el experimento si observan algún efecto adverso.

Debe restablecer la carga de trabajo y sus sistemas subyacentes al estado bueno conocido original. A menudo, el diseño resistente de la carga de trabajo se autorrepara. Pero algunos diseños de errores o experimentos fallidos pueden dejar su carga de trabajo en un estado de error inesperado. Al final del experimento, debe ser consciente de ello y restablecer la carga de trabajo y los sistemas. Con AWS FIS puede establecer una configuración de reversión (también llamada acción posterior) dentro de los parámetros de la acción. Una acción posterior devuelve el objetivo al estado en el que se encontraba antes de ejecutar la acción. Ya sean automatizadas (como el uso de AWS FIS) o manuales, estas acciones posteriores deben formar parte de una guía de estrategias que describa cómo detectar y gestionar los errores.

- Verifique la hipótesis.

  Principios de la ingeniería del caos ofrece estas directrices sobre cómo verificar el estado estable de su carga de trabajo:

  Céntrese en los resultados medibles de un sistema, más que en los atributos internos del mismo. Las mediciones de esa producción durante un corto periodo de tiempo constituyen una aproximación al estado estable del sistema. El rendimiento global del sistema, las tasas de error y los percentiles de latencia podrían ser métricas de interés que representen el comportamiento en estado estable. Al centrarse en los patrones de comportamiento sistémico durante los experimentos, la ingeniería del caos verifica que el sistema funcione, en lugar de intentar validar cómo funciona.

  En nuestros dos ejemplos anteriores, incluimos las métricas de estado estable de menos del 0,01 % de aumento de errores del lado del servidor (5xx) y menos de un minuto de errores de lectura y escritura en la base de datos.

  Los errores 5xx son una buena métrica porque son una consecuencia del modo de error que un cliente de la carga de trabajo experimentará directamente. La medición de los errores de la base de datos es buena como consecuencia directa del error, pero también debe complementarse con una medición del impacto en el cliente, como las solicitudes fallidas de los clientes o los errores que aparecen en el cliente. Además, incluya una monitorización sintética (también conocida como valor controlado) en cualquier API o URI al que acceda directamente el cliente de su carga de trabajo.

- Mejore el diseño de la carga de trabajo para la resiliencia.

  Si el estado estable no se mantuvo, entonces investigue cómo se puede mejorar el diseño de la carga de trabajo para mitigar el error, aplicando las mejores prácticas del Pilar de fiabilidad

de AWS Well-Architected. Se pueden encontrar orientaciones y recursos adicionales en la AWS Builder's Libraryque aloja artículos sobre cómo mejorar las comprobaciones de estado o bien emplear reintentos con retroceso en el código de su aplicación, entre otros.

Una vez aplicados estos cambios, vuelva a realizar el experimento (mostrado por la línea de puntos en el volante de ingeniería del caos) para determinar su eficacia. Si el paso de verificación indica que la hipótesis es cierta, entonces la carga de trabajo estará en estado estable, y el ciclo continúa.

- Realice experimentos con regularidad.

  Un experimento de caos es un ciclo, y los experimentos deben realizarse regularmente como parte de la ingeniería del caos. Después de que una carga de trabajo cumpla con la hipótesis del experimento, este debe automatizarse para ejecutarse continuamente como parte de la regresión de su canalización de CI/CD. Para saber cómo hacerlo, consulte este blog sobre cómo ejecutar experimentos de AWS FIS con AWS CodePipeline. Este laboratorio sobre experimentos de AWS FIS periódicos en una canalización de CI/CD le permite trabajar de forma práctica.

  Los experimentos de inyección de errores también forman parte de los días de juego (consulte REL12-BP06 Planificación regular de días de juego). En los días de juego se simula un error o un evento para verificar los sistemas, los procesos y las respuestas de los equipos. El objetivo es, de hecho, realizar las acciones que llevaría a cabo el equipo si se produjera un evento excepcional.

- Capture y almacene los resultados de los experimentos.

  Los resultados de los experimentos de inyección de errores deben capturarse y persistir. Incluya todos los datos necesarios (como el tiempo, la carga de trabajo y las condiciones) para poder analizar posteriormente los resultados y las tendencias del experimento. Algunos ejemplos de resultados pueden ser capturas de pantalla de paneles de control, volcados CSV de la base de datos de su métrica o un registro escrito a mano de los eventos y observaciones del experimento. Experimentar el registro con AWS FIS puede formar parte de esta captura de datos.

Recursos

Prácticas recomendadas relacionadas:

- REL08-BP03 Integrar las pruebas de resiliencia como parte de su despliegue
- REL13-BP03 Probar la implementación de recuperación de desastres para validarla

Documentos relacionados:

- ¿Qué es AWS Fault Injection Service?
- ¿Qué es AWS Resilience Hub?
- Principios de la ingeniería del caos
- Ingeniería del caos: Planificar su primer experimento
- Ingeniería de resiliencia: aprender a asumir los errores
- Historias de ingeniería del caos
- Evitar los planes alternativos en los sistemas distribuidos
- Despliegue de valores controlados para experimentos de caos

Vídeos relacionados:

- AWS re:Invent 2020: Pruebas de resistencia mediante la ingeniería del caos (ARC316)
- AWS re:Invent 2019: Mejorar la resiliencia con la ingeniería del caos (DOP309-R1)
- AWS re:Invent 2019: Realizar ingeniería del caos en un mundo sin servidores (CMY301)

Ejemplos relacionados:

- Laboratorio de Well-Architected: Nivel 300: pruebas de resiliencia de Amazon EC2, Amazon RDS y Amazon S3
- Laboratorio de ingeniería del caos en AWS
- Laboratorio de aplicaciones resilientes y bien diseñadas con ingeniería del caos
- Laboratorio de caos sin servidor
- Medir y mejorar la resiliencia de sus aplicaciones con laboratorio de AWS Resilience Hub

Herramientas relacionadas:

- AWS Fault Injection Service
- AWS Marketplace: Gremlin Chaos Engineering Platform (Plataforma de ingeniería del caos de Gremlin)
- Chaos Toolkit
- Chaos Mesh

- Litmus

REL12-BP06 Planificación regular de días de juego

Utilice días de juego para poner en práctica frecuentemente sus procedimientos para responder a los eventos y los errores lo más cerca de la fecha de lanzamiento a producción posible (incluidos los entornos de producción) con las personas que trabajarán en los escenarios de error reales. Los días de juego sirven para imponer medidas que garanticen que los eventos de producción no afecten a los usuarios.

En los días de juego se simula un error o un evento para probar los sistemas, los procesos y las respuestas de los equipos. El objetivo es, de hecho, realizar las acciones que llevaría a cabo el equipo si se produjera un evento excepcional. Esto ayudará a comprender dónde se pueden realizar mejoras y a desarrollar la experiencia organizacional en la gestión de eventos. Deberían hacerse habitualmente para que el equipo desarrolle “memoria muscular” sobre cómo responder.

Una vez que cuente con un diseño de resiliencia y que lo haya probado en entornos que no sean de producción, los días de juego son la fórmula ideal para garantizar que todo funcione según lo previsto en el entorno de producción. Un día de juego, especialmente el primero, es una actividad para todo el equipo en la que los ingenieros y el personal de operaciones están informados de cuándo ocurrirá y de qué pasará. Hay runbooks preparados. Se ejecutan eventos simulados, incluidos los posibles eventos de error, en los sistemas de producción y de la forma prescrita y, entonces, se evalúa el impacto. Si todos los sistemas funcionan según lo diseñado, la detección y la autocorrección se producirán con un impacto mínimo o inexistente. Sin embargo, si se observa algún impacto negativo, se da marcha atrás a la prueba y los problemas con la carga de trabajo se remedian, de forma manual si fuera necesario (utilizando el runbook). Dado que los días de juego suelen desarrollarse en el entorno de producción, deben tomarse todas las precauciones necesarias para garantizar que no haya ningún impacto sobre la disponibilidad para los clientes.

Antipatrones usuales:

- Documentar los procedimientos, pero no ponerlos nunca en práctica
- No incluir a los responsables de la toma de decisiones del negocio en los ejercicios de prueba

Beneficios de establecer esta práctica recomendada: Realizar días de juego periódicamente garantiza que todos los empleados sigan las políticas y los procedimientos cuando se produzca un incidente real y valida que esas políticas y procedimientos son apropiados.

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

Guía para la implementación

- Programe días de juego para practicar con las guías de estrategias y runbooks. Todo el mundo que pueda verse involucrado en un evento de producción debe participar en los días de juego: el propietario de la empresa, el personal de desarrollo, el personal de operaciones y los equipos de respuesta a incidentes.
  - Ejecute sus pruebas de carga o rendimiento y después su inyección de errores.
  - Busque anomalías en los runbooks y oportunidades para poner en práctica las guías de estrategias.
    - Si se desvía de los runbooks, mejórelos o corrija este comportamiento. Si utiliza la guía de estrategias, identifique el runbook que debería haberse usado o cree uno nuevo.

Recursos

Documentos relacionados:

- ¿Qué es AWS GameDay?

Videos relacionados:

- AWS re:Invent 2019: Mejorar la resiliencia con la ingeniería del caos (DOP309-R1)

  Ejemplos relacionados:

- Laboratorios de AWS Well-Architected: comprobación de resiliencia

FIABILIDAD 13. ¿Cómo planifica la recuperación de desastres (DR)?

Disponer de copias de seguridad y de componentes de cargas de trabajo redundantes es el principio de su estrategia de DR. El RTO y el RPO son sus objetivos para la restauración de su carga de trabajo. Estos se definen en función de las necesidades del negocio. Implemente una estrategia para satisfacer estos objetivos teniendo en cuenta las ubicaciones y la función de los recursos de las cargas de trabajo y los datos. La probabilidad de una interrupción y el coste de recuperación son también factores clave que ayudan a conocer el valor empresarial de proporcionar recuperación de desastres para una carga de trabajo.

Prácticas recomendadas

- REL13-BP01 Definir objetivos de recuperación para la inactividad y la pérdida de datos
- REL13-BP02 Usar estrategias de recuperación definidas para cumplir los objetivos de recuperación
- REL13-BP03 Probar la implementación de recuperación de desastres para validarla
- REL13-BP04 Administrar la desviación de la configuración en el sitio de o en la región de recuperación de desastres
- REL13-BP05 Automatizar la recuperación

REL13-BP01 Definir objetivos de recuperación para la inactividad y la pérdida de datos

La carga de trabajo tiene un objetivo de tiempo de recuperación (RTO) y un objetivo de punto de recuperación (RPO).

Objetivo de tiempo de recuperación (RTO) es el retraso máximo aceptable entre la interrupción del servicio y su restablecimiento. Esto determina lo que se considera un intervalo de tiempo aceptable cuando el servicio no está disponible.

Objetivo de punto de recuperación (RPO) es el periodo de tiempo máximo aceptable desde el último punto de recuperación de datos. Determina lo que se considera una pérdida aceptable de datos entre el último punto de recuperación y la interrupción del servicio.

Los valores de RTO y RPO son consideraciones importantes a la hora de seleccionar una estrategia de recuperación de desastres (DR) adecuada para su carga de trabajo. Estos objetivos los determina la empresa y los utilizan los equipos técnicos para seleccionar e implementar una estrategia de recuperación de desastres.

Resultado deseado:

Cada carga de trabajo tiene un RTO y un RPO asignados, definidos en función del impacto empresarial. La carga de trabajo se asigna en un nivel predefinido, lo que define la disponibilidad del servicio y la pérdida de datos aceptable, con un RTO y un RPO asociados. Si no es posible esta jerarquización, se puede asignar de forma personalizada por carga de trabajo, con la intención de crear niveles más adelante. RTO y RPO se utilizan como una de las principales consideraciones para la selección de la implementación de una estrategia de recuperación de desastres para la carga de trabajo. Otras consideraciones a la hora de elegir una estrategia de recuperación de desastres son las restricciones de costes, las dependencias de la carga de trabajo y los requisitos operativos.

Para RTO, entienda el impacto basado en la duración de una interrupción. ¿Es lineal o hay implicaciones no lineales? (por ejemplo, después de cuatro horas, se cierra una línea de fabricación hasta el comienzo del siguiente turno).

Una matriz de recuperación de desastres, como la siguiente, puede ayudarle a entender cómo se relaciona la criticidad de la carga de trabajo con los objetivos de recuperación. (Tenga en cuenta que los valores reales de los ejes X e Y deben adaptarse a las necesidades de su organización).

Matriz de recuperación de desastres						
		Objetivo de punto de recuperación				
		< 1 minuto	< 1 hora	< 6 horas	< 1 día	Más de 1 día
Objetivo de tiempo de recuperación	< 10 minutos	Crítico	Crítico	Alto	Medio	Medio
	< 2 horas	Crítico	Alto	Medio	Medio	Bajo
	< 8 horas	Alto	Medio	Medio	Bajo	Bajo
	< 24 horas	Medio	Medio	Bajo	Bajo	Bajo
	Más de 24 horas	Medio	Bajo	Bajo	Bajo	Bajo

Figura 16: Matriz de recuperación de desastres

Patrones de uso no recomendados comunes:

- No hay objetivos de recuperación definidos.
- Seleccionar objetivos de recuperación arbitrarios.
- Seleccionar objetivos de recuperación demasiado permisivos y no satisfacer los objetivos empresariales
- No entender el impacto del tiempo de inactividad y la pérdida de datos.
- Seleccionar objetivos de recuperación poco realistas, como el tiempo de recuperación cero y la pérdida de datos cero, que pueden no ser alcanzables para la configuración de su carga de trabajo.
- Seleccionar objetivos de recuperación más estrictos que los objetivos empresariales reales Esto obliga a realizar implementaciones de recuperación de desastres más costosas y complejas de lo que necesita la carga de trabajo.
- Seleccionar objetivos de recuperación incompatibles con los de una carga de trabajo dependiente.
- Sus objetivos de recuperación no tienen en cuenta los requisitos de cumplimiento normativo.

- RTO y RPO definidos para una carga de trabajo, pero nunca se han probado.

Beneficios de establecer esta práctica recomendada: Los objetivos de recuperación de tiempo y pérdida de datos son necesarios para guiar su implementación de DR.

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

Guía para la implementación

Para la carga de trabajo dada, debe entender el impacto del tiempo de inactividad y la pérdida de datos en su empresa. Por lo general, el impacto aumenta con un mayor tiempo de inactividad o pérdida de datos, pero la forma de este crecimiento puede variar en función del tipo de carga de trabajo. Por ejemplo, puede ser capaz de tolerar un tiempo de inactividad de hasta una hora con poco impacto, pero después el impacto aumenta rápidamente. El impacto en la empresa se manifiesta de muchas formas, como el coste económico (por ejemplo, la pérdida de ingresos), la confianza de los clientes (y el impacto en la reputación), los problemas operativos (por ejemplo, la pérdida de nóminas o la disminución de la productividad) y el riesgo normativo. Siga estos pasos para entender estos impactos y establecer RTO y RPO para su carga de trabajo.

Pasos de implementación

1. Determine las partes interesadas de su empresa para esta carga de trabajo y colabore con ellas para implementar estos pasos. Los objetivos de recuperación de una carga de trabajo son una decisión empresarial. Después, los equipos técnicos trabajan con las partes interesadas de la empresa para utilizar estos objetivos para seleccionar una estrategia de recuperación de desastres.

   > **Note**
   >
   > Para los pasos 2 y 3, puede usar la the section called “Hoja de trabajo de implementación”.

2. Responda a las preguntas siguientes para reunir la información necesaria a fin de tomar una decisión.
3. ¿Tiene categorías o niveles de criticidad para el impacto de la carga de trabajo en su organización?

   a. En caso afirmativo, asigne esta carga de trabajo a una categoría.

b. En caso contrario, establezca estas categorías. Cree un máximo de cinco categorías y ajuste el intervalo de su objetivo de tiempo de recuperación para cada una. Algunos ejemplos de categorías son: crítica, alta, media, baja. Para entender cómo se asignan las cargas de trabajo a las categorías, considere si la carga de trabajo es de misión crítica, importante para la empresa o no lo es.

c. Establezca el RTO y el RPO de la carga de trabajo en función de la categoría. Acceda a este paso para elegir siempre una categoría más estricta (RTO y RPO más bajos) que los valores sin procesar calculados. Si esto da lugar a un cambio de valor inadecuado, considere la posibilidad de crear una nueva categoría.

4. Según estas respuestas, asigne los valores de RTO y RPO a la carga de trabajo. Se puede hacer directamente o mediante la asignación de la carga de trabajo a un nivel de servicio predefinido.
5. Documente el plan de recuperación de desastres (DRP) de esta carga de trabajo, que forma parte del plan de continuidad del negocio (BCP)de su organización, en una ubicación accesible al equipo de la carga de trabajo y las partes interesadas.

   a. Registre el RTO y el RPO, así como la información utilizada para determinar estos valores. Incluya la estrategia que se utiliza para evaluar el impacto de la carga de trabajo en la empresa.

   b. Registre otras métricas, además de RTO y RPO, de las que hace un seguimiento o planifica hacerlo para los objetivos de recuperación de desastres.

   c. Agregará los detalles de su estrategia de recuperación de desastres y su runbook a este plan cuando los cree.

6. Si busca la criticidad de la carga de trabajo en una matriz como la de la figura 15, puede empezar a establecer los niveles de servicio predefinidos para su organización.
7. Después de haber implementado una estrategia de recuperación de desastres (o una prueba de concepto para una estrategia de este tipo) según the section called “REL13-BP02 Usar estrategias de recuperación definidas para cumplir los objetivos de recuperación”, pruebe esta estrategia para determinar la capacidad de tiempo de recuperación (RTC) y de punto de recuperación (RPC) reales de la carga de trabajo. Si no cumplen los objetivos de recuperación previstos, es posible colaborar con las partes interesadas de su empresa para ajustar dichos objetivos o realizar cambios en la estrategia de RD para cumplir los objetivos previstos.

Preguntas principales

1. ¿Cuál es el tiempo máximo que la carga de trabajo puede estar inactiva antes de que se produzca un impacto grave en la empresa?

a. Determine el coste económico (impacto financiero directo) para la empresa por minuto si se interrumpe la carga de trabajo.

b. Considere que el impacto no siempre es lineal. El impacto puede ser limitado al principio e ir aumentando rápidamente a partir de un punto crítico.

2. ¿Cuál es la cantidad máxima de datos que puede perderse antes de que se produzca un impacto grave en la empresa?

   a. Considere este valor para su almacén de datos más crítico. Identifique la criticidad correspondiente de otros almacenes de datos.

   b. ¿Se pueden recrear los datos de la carga de trabajo si se pierden? Si esto es operativamente más fácil que la copia de seguridad y la restauración, elija el RPO en función de la criticidad de los orígenes de los datos que se utilizan para recrear los datos de la carga de trabajo.

3. ¿Cuáles son los objetivos de recuperación y las expectativas de disponibilidad de las cargas de trabajo de las que depende esta (descendente), o de las cargas de trabajo que dependen de esta (ascendente)?

   a. Elija objetivos de recuperación que permitan a esta carga de trabajo cumplir los requisitos de las dependencias ascendentes.

   b. Elija objetivos de recuperación que sean alcanzables teniendo en cuenta las capacidades de recuperación de las dependencias descendentes. Se pueden excluir las dependencias descendentes no críticas (aquellas que puede «resolver»). O bien, trabaje con las dependencias críticas posteriores para mejorar sus capacidades de recuperación cuando sea necesario.

Preguntas adicionales

Considere estas preguntas y cómo pueden aplicarse a esta carga de trabajo:

4. ¿Tiene diferentes RTO y RPO en función del tipo de interrupción región con respecto a AZ, etc.)?
5. ¿Hay algún momento específico (estacionalidad, eventos de ventas, lanzamientos de productos) en el que pueda cambiar su RTO/RPO? Si es así, ¿cuál es el límite de medida y tiempo diferente?
6. ¿Cuántos clientes se verán afectados si se interrumpe la carga de trabajo?
7. ¿Cuál es el impacto en la reputación si se interrumpe la carga de trabajo?
8. ¿Qué otros impactos operativos pueden producirse si se interrumpe la carga de trabajo? Por ejemplo, el impacto en la productividad de los empleados si los sistemas de correo electrónico no están disponibles o si los sistemas de nómina no pueden enviar las transacciones.

9. ¿Cómo se alinean el RTO y el RPO de la carga de trabajo con la línea de negocio y la estrategia organizativa de recuperación de desastres?
10. ¿Existen obligaciones contractuales internas para la prestación de un servicio? ¿Existen sanciones por incumplirlas?
11. ¿Cuáles son las restricciones normativas o de cumplimiento con los datos?

Hoja de trabajo de implementación

Puede utilizar esta hoja de trabajo para implementar los pasos 2 y 3. Puede ajustar esta hoja de trabajo para adaptarla a sus necesidades específicas, por ejemplo, puede agregar preguntas adicionales.

Paso 2: Preguntas principales	¿Se aplica a la carga de trabajo?	RTO de carga de trabajo	RPO de carga de trabajo	Ajuste de RTO	Ajuste de RPO	Instrucciones
[1] tiempo máximo que puede estar inoperativa la carga de trabajo						medido en tiempo desde el inicio de la interrupción hasta la recuperación
[2] cantidad máxima de datos que se pueden perder						medido en tiempo desde el último conjunto de datos restaurable correcto conocido
[3a] dependencias upstream						introduzca los objetivos de recuperación upstream más estrictos
[3b] dependencias downstream						introduzca los objetivos de recuperación downstream menos estrictos
[3a] dependencias upstream reconciliadas						Si el valor upstream es menor que los valores actuales y el valor downstream es mayor,
[3b] dependencias downstream reconciliadas						trabaje con las dependencias para reconciliarlas e introducir aquí los valores reconciliados
[3] dependencias						reduzca los valores para ajustarse a las dependencias upstream o auméntelos en función de las capacidades de dependencias downstream
Paso 2: Preguntas adicionales						Indique si se aplica la pregunta. Si no se aplica, omítala
RTO/RPO base						Traslade los valores de RTO y RPO de arriba aquí
[4] tipo de interrupción	[ ]S / [ ]N					Introduzca los objetivos de recuperación del tipo de evento con los requisitos más estrictos
[5] objetivos específicos basados en el tiempo	[ ]S / [ ]N					Introduzca los objetivos de recuperación de tiempo con los requisitos más estrictos
[6] clientes afectados	[ ]S / [ ]N					Realice un gráfico de los clientes afectados en función del tiempo de inactividad o de los datos perdidos. Utilícelo para introducir los RTO y RPO máximos permisibles en función del impacto sobre los clientes
[7] impacto reputacional	[ ]S / [ ]N					Trabaje con la empresa para determinar los RTO y RPO máximos en función del impacto sobre la reputación
[8] impacto operativo	[ ]S / [ ]N					Introduzca los RTO y RPO máximos en función del impacto operativo
[9] alineación organizativa	[ ]S / [ ]N					Introduzca los RTO y RPO máximos para cargas de trabajo de este tipo según los requisitos organizativos y de LOB
[10] obligaciones contractuales	[ ]S / [ ]N					Introduzca los RTO y RPO máximos en función de las obligaciones contractuales
[11] conformidad normativa	[ ]S / [ ]N					Introduzca los RTO y RPO máximos en función de la conformidad normativa pertinente
objetivo basado en preguntas adicionales						Tome el valor mínimo (valor más estricto) de entre Q 4-11 e introdúzcalo aquí
objetivo ajustado						Si no se puede dar cabida a los objetivos de la línea anterior, colabore con los interesados para transigir en los límites e introduzca aquí un nuevo valor mínimo
RTO/RPO ajustados						Introduzca los valores base de RPO/RTO o el objetivo ajustado, el menor de los valores
Paso 3						
Mapa a categoría o nivel predefinidos						Ajuste ambos valores a la baja (lo más estricto) para alinearlos con el nivel definido más cercano

Hoja de trabajo

Nivel de esfuerzo para el plan de implementación: Bajo

Recursos

Prácticas recomendadas relacionadas:

- the section called “REL09-BP04 Realizar una recuperación periódica de los datos para verificar la integridad de la copia de seguridad y los procesos”
- the section called “REL13-BP02 Usar estrategias de recuperación definidas para cumplir los objetivos de recuperación”
- the section called “REL13-BP03 Probar la implementación de recuperación de desastres para validarla”

Documentos relacionados:

- Blog de arquitectura de AWS: serie de recuperación de desastres
- Recuperación de desastres de las cargas de trabajo en AWS: recuperación en la nube (documento técnico de AWS)
- Managing resiliency policies with AWS Resilience Hub (Administración de las políticas de resiliencia con AWS Resilience Hub)
- Socio de APN: socios que pueden ayudar con la recuperación de desastres
- AWS Marketplace: productos que pueden usarse para la recuperación de desastres

Vídeos relacionados:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (Patrones de arquitectura para aplicaciones activas-activas en varias regiones) (ARC209-R2)
- Disaster Recovery of Workloads on AWS (Recuperación de desastres de cargas de trabajo en AWS)

REL13-BP02 Usar estrategias de recuperación definidas para cumplir los objetivos de recuperación

Defina una estrategia de recuperación de desastres (DR) que se ajuste a los objetivos de recuperación de su carga de trabajo. Elija una estrategia como copia de seguridad y restauración, estado de espera (activa/pasiva) o activa/activa.

Resultado deseado: para cada carga de trabajo, existe una estrategia de DR definida e implementada que permite que esa carga de trabajo alcance los objetivos de DR. Las estrategias de DR entre cargas de trabajo emplean patrones reutilizables (como las estrategias descritas anteriormente).

Antipatrones usuales:

- Implementar procedimientos de recuperación incoherentes para cargas de trabajo con objetivos de DR similares.
- Dejar la estrategia de DR para implementarla ad hoc cuando se produzca un desastre.
- No tener un plan de recuperación de desastres.
- Depender de las operaciones del plano de control durante la recuperación.

Beneficios de establecer esta práctica recomendada:

- El uso de estrategias de recuperación definidas le permite emplear herramientas y procedimientos de prueba comunes.
- El uso de estrategias de recuperación definidas mejora el intercambio de conocimiento entre equipos y la implementación de la DR en las cargas de trabajo que se encuentran bajo su responsabilidad.

Nivel de riesgo expuesto si no se establece esta práctica recomendada: alto Sin una estrategia de DR planificada, implementada y probada, es poco probable que consiga sus objetivos de recuperación en caso de desastre.

Guía para la implementación

Una estrategia de DR depende de la capacidad de poner en marcha su carga de trabajo en un sitio de recuperación si su ubicación principal deja de estar disponible para la ejecución de dicha carga de trabajo. Los objetivos de recuperación más comunes son el RTO y el RPO, como explicamos en REL13-BP01 Definir objetivos de recuperación para la inactividad y la pérdida de datos.

Una estrategia de DR en varias zonas de disponibilidad (AZ) dentro de una única Región de AWS puede ofrecer mitigación contra eventos de desastres como incendios, inundaciones y cortes de suministro eléctrico considerables. Si es necesario implementar medidas de protección contra un evento poco probable que evite que su carga de trabajo pueda ejecutarse en una Región de AWS determinada, puede seguir una estrategia de DR que abarque múltiples regiones.

A la hora de diseñar una estrategia de DR en varias regiones, debería elegir una de las siguientes estrategias. Se indican en orden ascendente de coste y complejidad y en orden descendente en cuanto al RTO y RPO. Región de recuperación se refiere a una Región de AWS diferente de la principal que se usa para su carga de trabajo.

Figura 17: Estrategias de recuperación de desastres (DR)

- Generación de copias de seguridad y restauración (RPO en horas, RTO en 24 horas o menos): cree una copia de seguridad de sus datos y aplicaciones en la región de recuperación. El uso de copias de seguridad automatizadas o continuas permitirá la recuperación a un momento dado, lo que puede reducir el RPO a hasta 5 minutos en algunos casos. En caso de desastre, desplegará su infraestructura (utilizando la infraestructura como código para reducir el RTO), desplegará su código y restaurará los datos desde la copia de seguridad para recuperarse del desastre en la región de recuperación.
- Luz piloto (RPO en minutos, RTO en decenas de minutos): aprovisione una copia de la infraestructura principal de su carga de trabajo en la región de recuperación. Replique sus datos en la región de recuperación y cree allí copias de seguridad de estos. Los recursos necesarios para permitir la replicación y copia de seguridad de los datos, como el almacenamiento de bases de datos y objetos, están siempre disponibles. Otros elementos, como los servidores de aplicaciones o la computación sin servidor, no se despliegan, pero pueden crearse cuando sea necesario con la configuración y el código de aplicación pertinentes.
- Espera semiactiva (RPO en segundos, RTO en minutos): mantenga una versión reducida pero totalmente funcional de su carga de trabajo que se ejecute continuamente en la región de recuperación. Los sistemas críticos se duplican en su totalidad y siempre están activos, pero con una flota reducida. Los datos se replican y están activos en la región de recuperación. Cuando llegue el momento de la recuperación, el sistema se amplía rápidamente para asumir la carga de producción. Cuanto mayor sea la escala de la espera semiactiva, menor será el RTO y la fiabilidad del plano de control. Cuando alcanza su plena escala, la espera pasa a denominarse espera activa.

- Activa-activa multirregión (multisitio) (RPO próximo a cero, RTO potencial de cero): la carga de trabajo está desplegada en varias Regiones de AWS y entrega tráfico de forma activa desde estas regiones. Esta estrategia requiere que sincronice datos entre regiones. Los posibles conflictos causados por escrituras en el mismo registro en dos diferentes réplicas regionales deben evitarse o gestionarse, lo que puede resultar complejo. La replicación de datos es útil para la sincronización de datos y le protegerá ante algunos tipos de desastres, pero no ante el daño o la destrucción de datos, a no ser que su solución incluya también opciones para una recuperación a un momento dado.

> **Note**
>
> La diferencia entre la luz piloto y la espera semiactiva a veces puede ser difícil de comprender. Ambos métodos incluyen un entorno en su región de recuperación con copias de los activos de su región principal. La distinción es que la luz piloto no puede procesar solicitudes sin tomar primero acciones adicionales, mientras que la espera semiactiva puede gestionar el tráfico (a niveles de capacidad reducidos) inmediatamente. La luz piloto exige que active servidores, posiblemente que despliegue infraestructura adicional (no principal) y que escale verticalmente, mientras que la espera semiactiva solo requiere que escale verticalmente (ya está todo desplegado y en ejecución). Elija una de estas opciones en función de sus necesidades de RTO y RPO.
>
> Cuando el coste sea una preocupación y desee alcanzar unos objetivos de RPO y RTO similares a los definidos en la estrategia de espera semiactiva, podría plantearse soluciones nativas de la nube, como AWS Elastic Disaster Recovery, que adoptan el enfoque de luz piloto y ofrecen objetivos de RPO y RTO mejorados.

Pasos para la implementación

1. Determine una estrategia de DR que satisfaga los requisitos de recuperación de esta carga de trabajo.

La selección de una estrategia de DR requiere alcanzar un punto de equilibrio entre la reducción del tiempo de inactividad y la pérdida de datos (RTO y RPO) y los costes y la complejidad de implementar la estrategia. Debería evitar implementar una estrategia que sea más exigente de lo necesario, ya que esto supone costes innecesarios.

Por ejemplo, en el siguiente diagrama, la empresa ha determinado su RTO máximo permisible y el límite de gasto en su estrategia de restauración del servicio. Dados los objetivos de la empresa, las estrategias de DR de luz piloto o espera semiactiva satisfarán tanto el RTO como los criterios de coste.

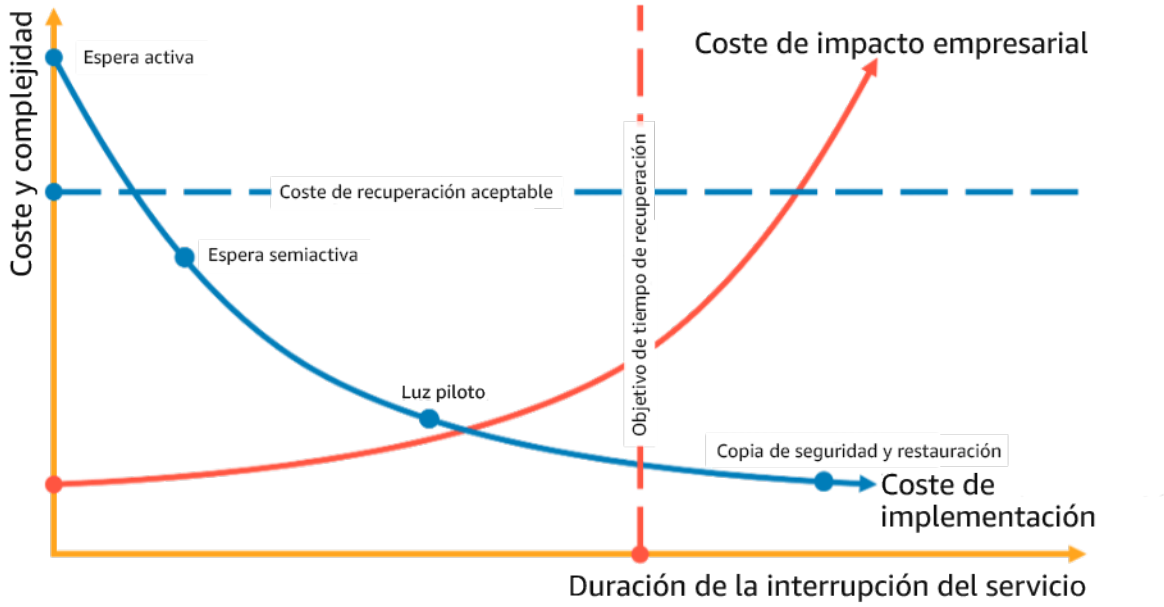


Figura 18: Selección de una estrategia de DR basada en el RTO y el coste

Para obtener más información, consulte Business Continuity Plan (BCP) (Plan de continuidad del negocio [BCP]).

2. Revise los patrones de implementación de la estrategia de DR seleccionada.

Este paso implica comprender cómo implementará la estrategia seleccionada. Las estrategias se explican utilizando Regiones de AWS para determinar un sitio principal y otro de recuperación. Sin embargo, también puede decidir utilizar zonas de disponibilidad en una única región como estrategia de DR, que utiliza los elementos de varias de estas estrategias.

En los siguientes pasos, puede aplicar la estrategia a su carga de trabajo específica.

Generación de copias de seguridad y restauración

Generación de copias de seguridad y restauración es la estrategia menos compleja de implementar, pero requiere más tiempo y esfuerzo para restaurar la carga de trabajo, lo que genera un mayor RTO y RPO. Se recomienda realizar siempre copias de seguridad de los datos y copiarlas en otro sitio (por ejemplo, otra Región de AWS).

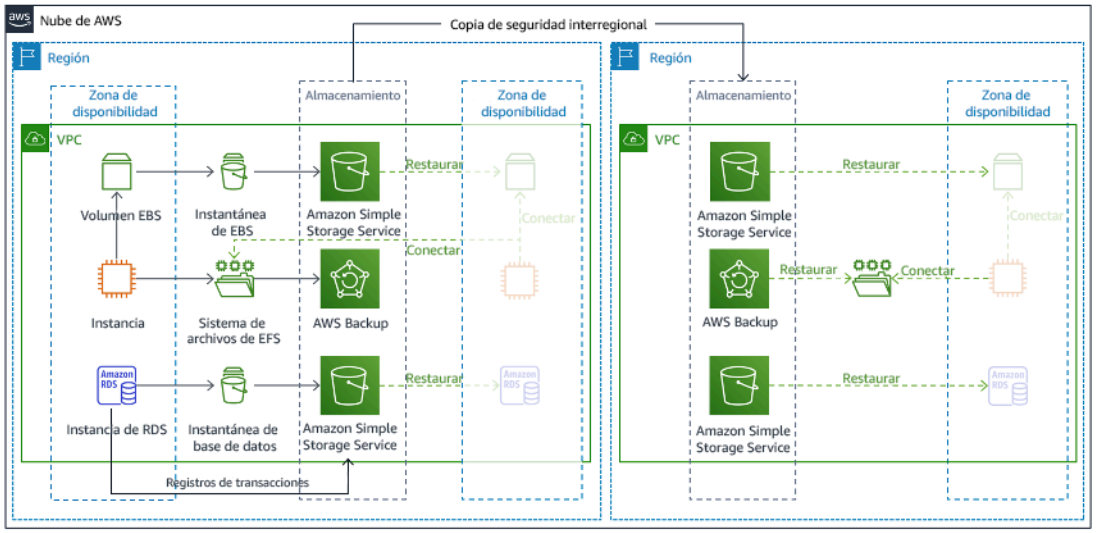


Figura 19: Arquitectura de copia de seguridad y restauración

Para obtener más detalles sobre esta estrategia, consulte Disaster Recovery (DR) Architecture on AWS, Part II: Backup and Restore with Rapid Recovery (Arquitectura de recuperación de desastres [DR] en AWS, parte II: copias de seguridad y restauración con Rapid Recovery).

Luz piloto

Con el enfoque luz piloto, replica sus datos de la región principal en la región de recuperación. Los recursos principales utilizados para la infraestructura de la carga de trabajo se despliegan en la región de recuperación; sin embargo, se siguen necesitando recursos adicionales y las dependencias pertinentes para que esta pila sea funcional. Por ejemplo, en la figura 20 no se despliegan instancias de computación.

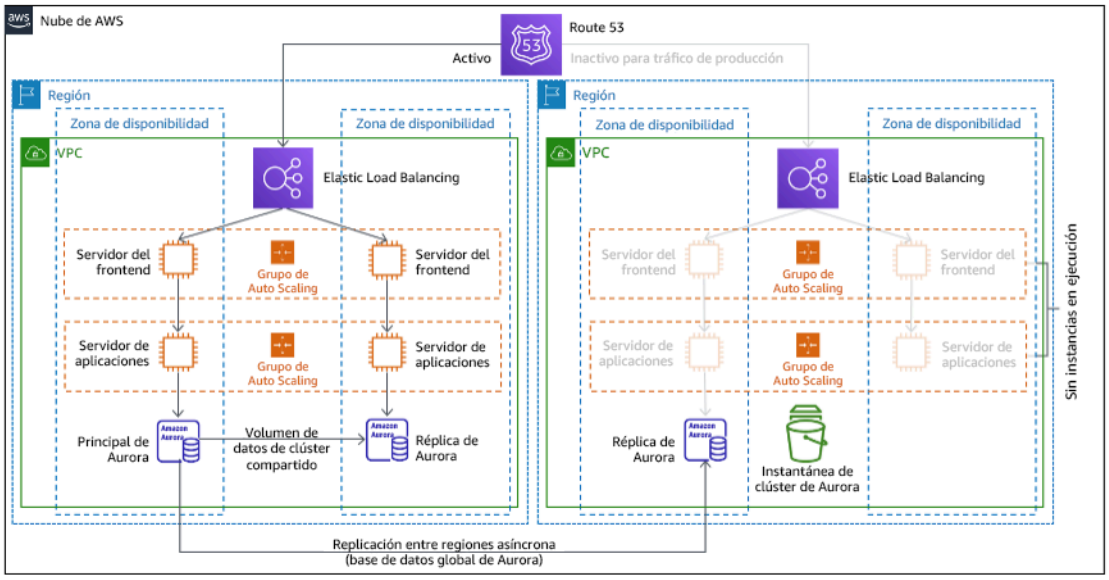


Figura 20: Arquitectura de luz piloto

Para obtener más detalles sobre esta estrategia, consulte Disaster Recovery (DR) Architecture on AWS, Part III: Pilot Light and Warm Standby (Arquitectura de recuperación de desastres [DR] en AWS, parte III: luz piloto y espera semiactiva).

Espera semiactiva

El enfoqueespera semiactiva supone garantizar que exista una copia con desescalada vertical pero con plena funcionalidad de su entorno de producción en otra región. Este enfoque extiende el concepto de luz piloto y reduce el tiempo de recuperación, ya que su carga de trabajo tiene disponibilidad permanente en otra región. Si la región de recuperación se despliega a plena capacidad, esto se denomina espera activa.

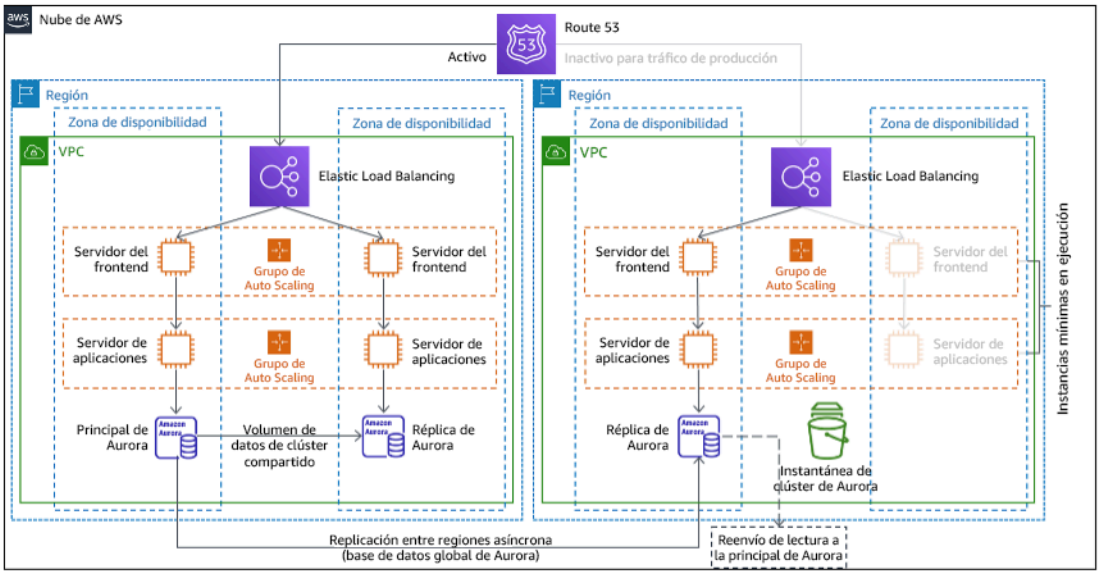


Figura 21: Arquitectura de espera semiactiva

El uso de las estrategias espera semiactiva o luz piloto requiere escalar verticalmente los recursos en la región de recuperación. Para verificar que la capacidad esté disponible cuando se necesita, considere el uso de reservas de capacidad para instancias EC2. Si usa AWS Lambda, entonces la simultaneidad aprovisionada puede proporcionar entornos de ejecución de forma que estén preparados para responder de inmediato a las invocaciones de la función.

Para obtener más detalles sobre esta estrategia, consulte Disaster Recovery (DR) Architecture on AWS, Part III: Pilot Light and Warm Standby (Arquitectura de recuperación de desastres [DR] en AWS, parte III: luz piloto y espera semiactiva).

Activa-activa multisitio

Puede ejecutar su carga de trabajo de forma simultánea en varias regiones como parte de una estrategia activa-activa multisitio. La estrategia activa-activa multisitio suministra tráfico desde todas las regiones en las que se despliega. Los clientes podrían seleccionar esta estrategia por motivos ajenos a la DR. Se puede usar para aumentar la disponibilidad o cuando se despliega una carga de trabajo para una audiencia global (para colocar el punto de conexión más cerca de los usuarios o para desplegar pilas localizadas para la audiencia de esa región). Como estrategia de DR, si la carga de trabajo no es compatible en una de las Regiones de AWS en la que esté desplegada, esa región se evacúa y las regiones restantes se usan para mantener la disponibilidad. La estrategia activa-activa multisitio es la más compleja de las estrategias de DR a nivel operativo, y solo debería seleccionarse cuando los requisitos empresariales lo exijan.

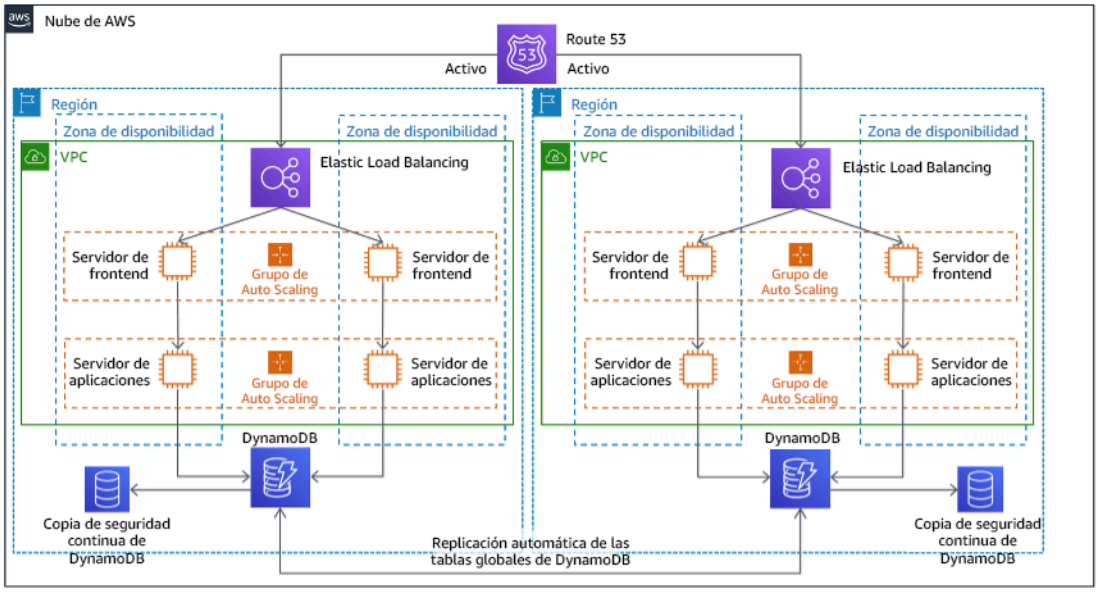


Figura 22: Arquitectura activa-activa multisitio

Para obtener más detalles sobre esta estrategia, consulte Disaster Recovery (DR) Architecture on AWS, Part IV: Multi-site Active/Active (Arquitectura de recuperación de desastres [DR] en AWS, parte IV: activa-activa multisitio).

AWS Elastic Disaster Recovery

Si está considerando la estrategia luz piloto o espera semiactiva para la recuperación de desastres, AWS Elastic Disaster Recovery podría ofrecer un enfoque alternativo con mayores ventajas. Elastic Disaster Recovery puede ofrecer un objetivo de RPO y RTO similar al de la estrategia espera semiactiva, pero manteniendo el enfoque de bajo coste de la estrategia luz piloto. Elastic Disaster Recovery replica los datos desde la región primaria a la región de recuperación, utilizando la protección continua de datos para lograr un RPO medido en segundos y un RTO que puede medirse en minutos. En la región de recuperación solo se despliegan los recursos necesarios para replicar los datos, lo que mantiene los costes bajos, de forma similar a la estrategia luz piloto. Cuando se utiliza Elastic Disaster Recovery, el servicio coordina y organiza la recuperación de los recursos de computación cuando se inicia como parte de una conmutación por error o un simulacro.

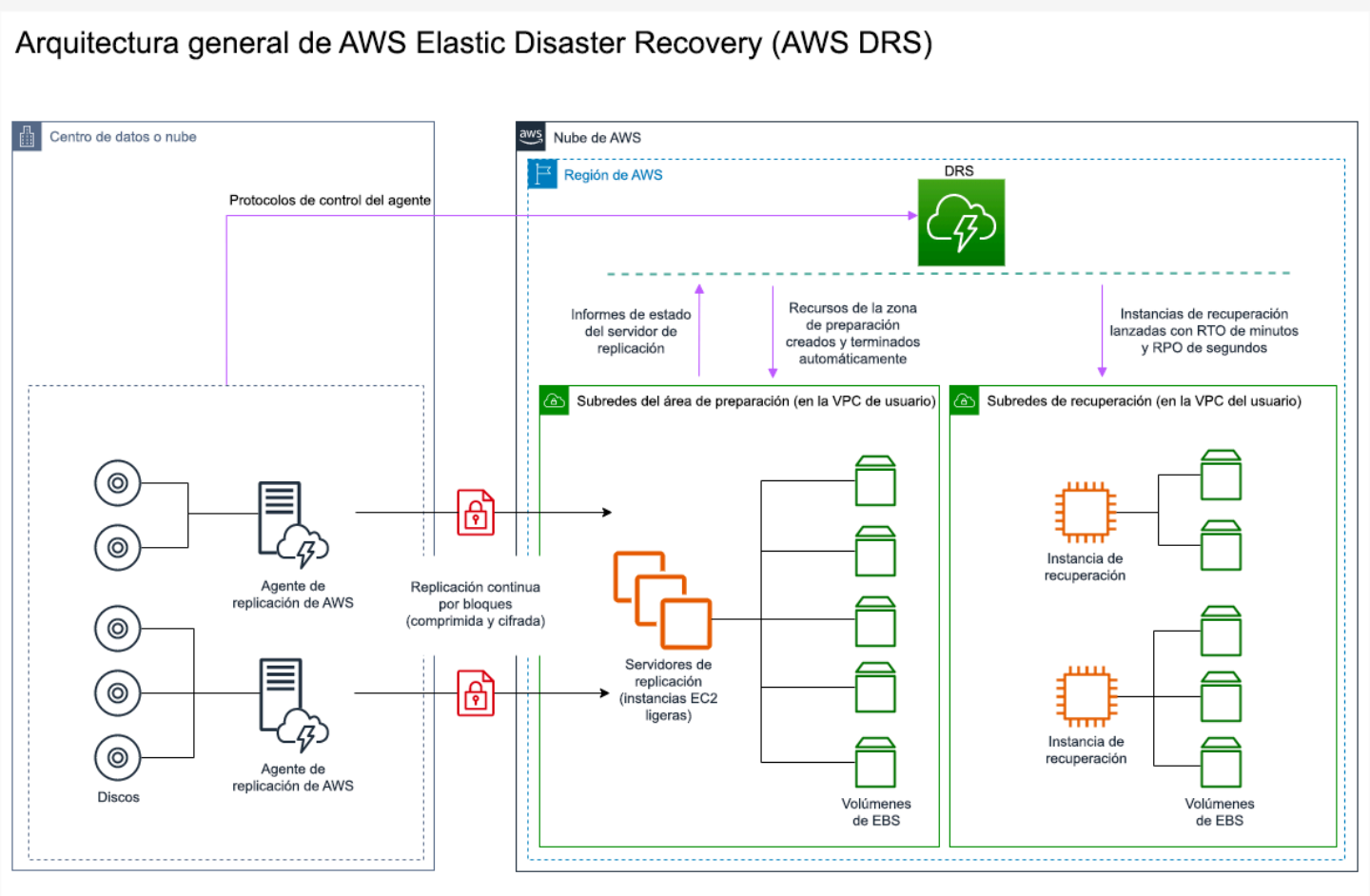


Figura 23: Arquitectura de AWS Elastic Disaster Recovery

Prácticas adicionales para la protección de datos

Con todas las estrategias, también debe mitigar los posibles desastres de datos. La replicación de datos continua le protegerá ante algunos tipos de desastres, pero no ante el daño o la destrucción de datos, a no ser que su estrategia incluya también control de versiones para una recuperación a un momento dado. También debe realizar una copia de seguridad de los datos replicados en el sitio de recuperación para crear copias de seguridad a un momento dado además de las réplicas.

Uso de varias zonas de disponibilidad (AZ) en una únicaRegión de AWS

Al usar varias AZ en una única región, su implementación de DR utiliza varios elementos de las estrategias anteriores. Primero, debe crear una arquitectura de alta disponibilidad utilizando varias AZ, como se muestra en la figura 23. Esta arquitectura hace uso de un enfoque activo-activo multisitio, ya que las instancias Amazon EC2 y el equilibrador de carga elástico tienen recursos

desplegados en varias AZ, que gestionan las solicitudes de forma activa. La arquitectura también dispone de espera activa, en la que si la instancia principal de Amazon RDS produce un error (o la propia AZ tiene un error), la instancia en espera pasa a ser la principal.

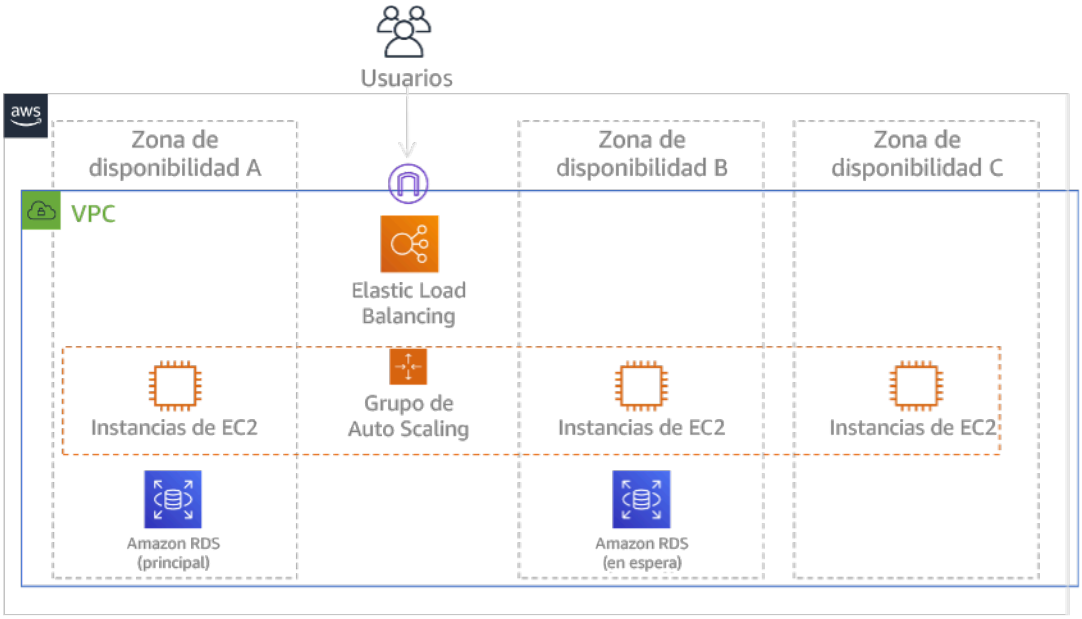


Figura 24: Arquitectura multi-AZ

Además de esta arquitectura de alta disponibilidad, necesita agregar copias de seguridad con todos los datos necesarios para ejecutar su carga de trabajo. Esto es especialmente importante para datos que estén limitados a una única zona, como los volúmenes de Amazon EBS o los clústeres de Amazon Redshift. Si falla una AZ, tendrá que restaurar estos datos en otra AZ. Siempre que sea posible, deberá copiar las copias de seguridad de los datos en otra Región de AWS como capa de protección adicional.

En la publicación de blog, Building highly resilient applications using Amazon Route 53 Application Recovery Controller, Part 1: Single-Region stack (Creación de aplicaciones de alta resiliencia con Amazon Route 53, parte 1: pila de una sola región) se ilustra un enfoque alternativo menos habitual a la RD multi-AZ de única región. Aquí, la estrategia es mantener la máxima cantidad posible de aislamiento entre las AZ, tal y como funcionan las regiones. Al utilizar esta estrategia alternativa, puede elegir un enfoque activo-activo o activo-pasivo.

> **Note**
>
> Algunas cargas de trabajo tienen requisitos normativos de residencia de los datos. Si esto se aplica a su carga de trabajo en una ubicación que actualmente tenga solo una Región

de AWS, el enfoque multirregión no se adaptará a sus necesidades empresariales. Las estrategias multi-AZ ofrecen una buena protección contra la mayor parte de los desastres.

3. Evalúe los recursos de su carga de trabajo y cuál será su configuración en la región de recuperación antes de la conmutación por error (durante la operación normal).

Para la infraestructura y los recursos de AWS, utilice infraestructura como código como AWS CloudFormation o herramientas de terceros como Hashicorp Terraform. Para un despliegue en varias cuentas y regiones con una única operación, puede utilizar AWS CloudFormation StackSets. En las estrategias activa-activa multisitio o espera activa, la infraestructura desplegada en su región de recuperación tiene los mismos recursos que su región principal. En las estrategias luz piloto y espera semiactiva, la infraestructura desplegada requerirá acciones adicionales para prepararse para la producción. Con parámetros y lógica condicional de CloudFormation, puede controlar si una pila desplegada está activa o en espera con una sola plantilla. Al utilizar Elastic Disaster Recovery, el servicio replicará y organizará la restauración de las configuraciones de las aplicaciones y los recursos de computación.

Todas las estrategias de DR requieren que se realicen copias de seguridad de los orígenes de datos en la Región de AWS, y que estas copias de seguridad se copien en la región de recuperación. AWS Backup proporciona una vista centralizada en la que puede configurar, programar y supervisar las copias de seguridad para estos recursos. En el caso de las opciones de luz piloto, espera semiactiva y activa-activa multisitio, también debería replicar los datos de la región principal en los recursos de datos de la región de recuperación, como las instancias de base de datos de Amazon Relational Database Service (Amazon RDS) o tablas de Amazon DynamoDB. De esta forma, estos recursos de datos estarán activos y preparados para responder a solicitudes en la región de recuperación.

Para obtener más información sobre cómo funcionan los servicios de AWS en todas las regiones, consulte esta serie de blogs sobre Creating a Multi-Region Application with AWS Services (Creación de una aplicación multirregión con servicios de AWS).

4. Determine e implemente cómo preparará su región de recuperación para la conmutación por error cuando sea necesario (durante un evento de desastre).

En el caso de la opción activa-activa multisitio, la conmutación por error implica evacuar una región y recurrir a las regiones activas restantes. En general, esas regiones están listas para aceptar tráfico. En las estrategias luz piloto y espera semiactiva, sus acciones de recuperación tendrán que

desplegar los recursos faltantes, como las instancias EC2 en la figura 20, además de otros recursos faltantes.

En todas las estrategias anteriores, es posible que tenga que promover instancias de solo lectura de bases de datos para que se conviertan en la instancia de lectura y escritura principal.

En copias de seguridad y restauración, la restauración de datos desde una copia de seguridad crea recursos para esos datos, como volúmenes EBS, instancias de bases de datos RDS y tablas de DynamoDB. También tiene que restaurar la infraestructura y desplegar el código. Puede usar AWS Backup para restaurar datos en la región de recuperación. Consulte REL09-BP01 Identificar todos los datos de los que se debe hacer una copia de seguridad y crearla o reproducir los datos a partir de los orígenes para obtener más información. La reconstrucción de la infraestructura incluye la creación de recursos como instancias EC2, además de Amazon Virtual Private Cloud (Amazon VPC), subredes y grupos de seguridad necesarios. Puede automatizar gran parte del proceso de restauración. Para saber cómo, consulte esta publicación de blog.

5. Determine e implemente cómo redirigirá su tráfico a la conmutación por error cuando sea necesario (durante un evento de desastre).

Esta operación de conmutación por error se puede iniciar automáticamente o manualmente. La conmutación por error iniciada automáticamente basada en comprobaciones de estado o alarmas se debe usar con cuidado, ya que una conmutación por error innecesaria (falsa alarma) supone ciertos inconvenientes, como la falta de disponibilidad y la pérdida de datos. Por tanto, la conmutación por error iniciada manualmente es la que se suele utilizar. En este caso, debe seguir automatizando los pasos de la conmutación por error, de modo que la iniciación manual sea como pulsar un botón.

Hay varias opciones de administración del tráfico que tener en cuenta al usar servicios de AWS. Una opción es utilizar Amazon Route 53. Al usar Amazon Route 53, puede asociar varios puntos de conexión de IP en una o varias Regiones de AWS con un nombre de dominio de Route 53. Para implementar la conmutación por error iniciada manualmente, puede utilizar el Controlador de recuperación de aplicaciones de Amazon Route 53, que proporciona una API de plano de datos de alta disponibilidad para redirigir el tráfico a la región de recuperación. Al implementar la conmutación por error, use las operaciones del plano de datos y evite las del plano de control, como se describe en REL11-BP04 Confiar en el plano de datos y no en el plano de control durante la recuperación.

Para obtener más información sobre esta y otras opciones, consulte esta sección del documento técnico de recuperación de desastres.

6. Diseñe un plan para la recuperación tras error de su carga de trabajo.

La conmutación por recuperación es cuando se devuelve la operación de una carga de trabajo a la región principal una vez que amaina un evento de desastre. El aprovisionamiento de la infraestructura y el código en la región principal generalmente sigue los mismos pasos que se utilizaron inicialmente, recurriendo a la infraestructura como código y las canalizaciones de despliegue del código. El reto que plantea la conmutación por recuperación es restaurar los almacenes de datos y asegurarse de que sean coherentes con la región de recuperación en funcionamiento.

En el estado de conmutación por error, las bases de datos en la región de recuperación están activas y tienen los datos actualizados. El objetivo es resincronizar desde la región de recuperación a la región principal, garantizando así que esté actualizada.

Algunos servicios de AWS harán esto automáticamente. Si utiliza las tablas globales de Amazon DynamoDB, incluso si la tabla en la región principal ha dejado de estar disponible, cuando vuelva a estar online, DynamoDB volverá a propagar las escrituras pendientes. Si utiliza la base de datos global de Amazon Aurora y utiliza la conmutación por error planificada administrada, se mantendrá la topología de replicación existente de la base de datos global de Aurora. Por tanto, la instancia de lectoescritura anterior en la región principal se convertirá en una réplica y recibirá actualizaciones desde la región de recuperación.

En los casos en los que esto no se haga automáticamente, tendrá que restablecer la base de datos en la región principal como una réplica de la base de datos en la región de recuperación. En muchos casos, esto supondrá eliminar la antigua base de datos principal y crear nuevas réplicas. Por ejemplo, para obtener instrucciones sobre cómo hacer esto con la base de datos global de Amazon Aurora en el caso de una conmutación por error no planificada, consulte este laboratorio: Fail Back a Global Database (Conmutación por recuperación a una base de datos global).

Tras una conmutación por error, si puede seguir operando en su región de recuperación, considere convertir esta región en la nueva región principal. Seguiría realizando los pasos anteriores para hacer que la antigua región principal fuera una región de recuperación. Algunas organizaciones llevan a cabo una rotación programada y cambian sus regiones principal y de recuperación periódicamente (por ejemplo, cada tres meses).

Todos los pasos necesarios para la conmutación por error y la restauración tras error deben mantenerse en una guía de estrategias que esté a disposición de todos los miembros del equipo y se revise periódicamente.

Si utiliza Elastic Disaster Recovery, el servicio ayudará a organizar y automatizar el proceso de conmutación por recuperación. Para obtener más información, consulte Performing a failback (Realizar una conmutación por recuperación).

Nivel de esfuerzo para el plan de implementación: alto

Recursos

Prácticas recomendadas relacionadas:

- the section called “REL09-BP01 Identificar todos los datos de los que se debe hacer una copia de seguridad y crearla o reproducir los datos a partir de los orígenes”
- the section called “REL11-BP04 Confiar en el plano de datos y no en el plano de control durante la recuperación”
- the section called “REL13-BP01 Definir objetivos de recuperación para la inactividad y la pérdida de datos”

Documentos relacionados:

- AWS Architecture Blog: Disaster Recovery Series (Blog de arquitectura de AWS: serie de recuperación de desastres)
- Disaster Recovery of Workloads on AWS: Recovery in the Cloud (Recuperación de cargas de trabajo en caso de desastre en AWS: recuperación en la nube) (Documento técnico de AWS)
- Opciones de recuperación de desastres en la nube
- Crear una solución backend activa-activa sin servidor en varias regiones en una hora
- Backend sin servidor en varias regiones: actualizado
- RDS: replicación de una réplica de lectura entre regiones
- Route 53: Configuración de la recuperación ante errores a nivel de DNS
- S3: replicación entre regiones
- What Is AWS Backup? (¿Qué es AWS Backup?)
- What is Route 53 Application Recovery Controller? (¿Qué es el Controlador de recuperación de aplicaciones de Route 53?)
- AWS Elastic Disaster Recovery
- HashiCorp Terraform: Get Started - AWS (HashiCorp Terraform: primeros pasos: AWS)
- Socio de APN: socios que pueden ayudar con la recuperación de desastres

- AWS Marketplace: products that can be used for disaster recovery (AWS Marketplace: productos que pueden usarse para la recuperación de desastres)

Vídeos relacionados:

- Disaster Recovery of Workloads on AWS (Recuperación de desastres de cargas de trabajo en AWS)
- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2) (AWS re:Invent 2018: Patrones de arquitectura para aplicaciones activas-activas en varias regiones)
- Get Started with AWS Elastic Disaster Recovery | Amazon Web Services (Introducción a AWS Elastic Disaster Recovery | Amazon Web Services)

Ejemplos relacionados:

- Well-Architected Lab - Disaster Recovery - Series of workshops illustrating DR strategies (Laboratorio de Well-Architected: Recuperación de desastres: serie de talleres que ilustran las estrategias de DR)

REL13-BP03 Probar la implementación de recuperación de desastres para validarla

Compruebe periódicamente la conmutación por error a su sitio de recuperación para verificar que funcione adecuadamente y que se cumplan el RTO y el RPO.

Antipatrones usuales:

- No llevar a cabo nunca conmutaciones por error en producción.

Beneficios de establecer esta práctica recomendada: las pruebas periódicas del plan de recuperación de desastres verifican que el plan funcione cuando llegue el momento y que su equipo sepa cómo ejecutar la estrategia.

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

Guía para la implementación

Un patrón que debe evitarse es el desarrollo de rutas de recuperación que se pongan en práctica pocas veces. Por ejemplo, puede tener un almacén de datos secundario que se utilice para consultas

de solo lectura. Cuando escribe en un almacén de datos y el almacén principal falla, es posible que quiera conmutar por error al almacén de datos secundario. Si no se prueba frecuentemente esta conmutación por error, es posible que sus suposiciones sobre las capacidades del almacén de datos secundario sean incorrectas. Es posible que la capacidad del almacén de datos secundario, que quizás fuera suficiente cuando se probó por última vez, ya no pueda tolerar la carga en esta situación. Nuestra experiencia ha demostrado que la única forma de recuperación de errores que funciona es aquella que prueba constantemente. Por ello, es mejor tener un número reducido de rutas de recuperación. Puede establecer patrones de recuperación y probarlos con frecuencia. Si tiene una ruta de recuperación compleja o crítica, todavía debe llevar a efecto ese error en producción periódicamente para asegurarse de que la ruta funcione. En el ejemplo que acabamos de comentar, se debe conmutar por error al modo de espera con regularidad, sin importar si es necesario.

Pasos para la implementación

1. Diseñe sus cargas de trabajo para que se puedan recuperar. Pruebe regularmente sus rutas de recuperación. La computación orientada a la recuperación identifica las características de los sistemas que mejoran la recuperación: aislamiento y redundancia, capacidad en todo el sistema para revertir los cambios, capacidad para supervisar y determinar el estado, capacidad para proporcionar diagnósticos, recuperación automatizada, diseño modular y capacidad para reiniciar. Ponga en práctica la ruta de recuperación para verificar que pueda cumplir la recuperación en el tiempo especificado para el estado especificado. Use sus runbooks durante esta recuperación para documentar los problemas y encontrar soluciones para ellos antes de la próxima prueba.
2. Para cargas de trabajo basadas en Amazon EC2, utilice AWS Elastic Disaster Recovery para implementar y lanzar instancias de simulacro para su estrategia de recuperación de desastres. AWS Elastic Disaster Recovery ofrece la posibilidad de ejecutar simulacros de manera eficiente, lo que le ayuda a prepararse para un evento de conmutación por error. También puede lanzar con frecuencia sus instancias mediante Elastic Disaster Recovery para realizar pruebas y simulacros sin redirigir el tráfico.

Recursos

Documentos relacionados:

- Socio de APN: socios que pueden ayudar con la recuperación de desastres
- AWS Architecture Blog: Disaster Recovery Series (Blog de arquitectura de AWS: serie de recuperación de desastres)

- AWS Marketplace: products that can be used for disaster recovery (AWS Marketplace: productos que pueden usarse para la recuperación de desastres)
- AWS Elastic Disaster Recovery
- Disaster Recovery of Workloads on AWS: Recovery in the Cloud (Recuperación de cargas de trabajo en caso de desastre en AWS: Recuperación en la nube) (documento técnico de AWS)
- AWS Elastic Disaster Recovery Preparing for Failover (AWS Elastic Disaster Recovery: Preparación para la conmutación por error)
- Proyecto de informática orientada a la recuperación de Berkeley/Stanford
- ¿Qué es AWS Fault Injection Simulator?

Vídeos relacionados:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (AWS re:Invent 2018: Patrones de arquitectura para aplicaciones activas-activas en varias regiones)
- AWS re:Invent 2019: Backup-and-restore and disaster-recovery solutions with AWS (AWS re:Invent 2019: Copia de seguridad y restauración y soluciones de recuperación de desastres con AWS)

Ejemplos relacionados:

- Well-Architected Lab - Testing for Resiliency (Laboratorio de Well-Architected: Prueba de resiliencia)

REL13-BP04 Administrar la desviación de la configuración en el sitio de o en la región de recuperación de desastres

Asegúrese de que la infraestructura, los datos y la configuración estén cuando se necesiten en el sitio o región de DR. Por ejemplo, compruebe que las AMI y las cuotas de servicio están actualizadas.

AWS Config supervisa y registra continuamente las configuraciones de sus recursos de AWS. Puede detectar la desviación y desencadenar Automatización de AWS Systems Manager para solucionarlo y generar alarmas. Además, AWS CloudFormation puede detectar la desviación en las pilas que ha desplegado.

Patrones de uso no recomendados comunes:

- No realizar actualizaciones en sus ubicaciones de recuperación, cuando realice cambios de configuración o de infraestructura en sus ubicaciones primarias.
- No considerar las posibles limitaciones (como las diferencias en los servicios) en las ubicaciones principales y de recuperación.

Beneficios de establecer esta práctica recomendada: Comprobar que su entorno de DR es coherente con el entorno existente garantiza una recuperación completa.

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

Guía para la implementación

- Asegúrese de que sus canalizaciones de entrega realizan la entrega tanto al sitio principal como al de copia de seguridad. Las canalizaciones de entrega para implementar aplicaciones en producción deben distribuir la entrega a todas las ubicaciones de la estrategia de recuperación de desastres especificadas, incluidos los entornos de desarrollo y pruebas.
- Habilite AWS Config para realizar un seguimiento de las posibles ubicaciones con desviaciones. Use reglas de AWS Config para crear sistemas que apliquen sus estrategias de recuperación de desastres y creen alertas si detectan divergencias.
  - Corrección de recursos de AWS disconformes con Reglas de AWS Config
  - Automatización de AWS Systems Manager
- Use AWS CloudFormation para desplegar su infraestructura. AWS CloudFormation puede detectar la desviación entre lo que especifican sus plantillas de CloudFormation y lo que realmente está desplegado.
  - AWS CloudFormation: detectar desviaciones en una pila completa de CloudFormation

Recursos

Documentos relacionados:

- Socio de APN: socios que pueden ayudar con la recuperación de desastres
- Blog de arquitectura de AWS: serie de recuperación de desastres
- AWS CloudFormation: detectar desviaciones en una pila completa de CloudFormation
- AWS Marketplace: productos que pueden usarse para la recuperación de desastres
- Automatización de AWS Systems Manager

- Recuperación de desastres de las cargas de trabajo en AWS: recuperación en la nube (documento técnico de AWS)
- ¿Cómo implemento una solución de administración de la configuración de la infraestructura en AWS?
- Corrección de recursos de AWS disconformes con Reglas de AWS Config

Vídeos relacionados:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (Patrones de arquitectura para aplicaciones activas-activas en varias regiones) (ARC209-R2)

REL13-BP05 Automatizar la recuperación

Use AWS o herramientas de terceros para automatizar la recuperación del sistema y dirigir el tráfico al sitio o región de DR.

En función de las comprobaciones de estado configuradas, los servicios de AWS, como Elastic Load Balancing y AWS Auto Scaling, pueden distribuir la carga a zonas de disponibilidad en buen estado mientras que los servicios, como Amazon Route 53 y AWS Global Accelerator, pueden dirigir la carga a Regiones de AWS en buen estado. Amazon Route 53 Application Recovery Controller le ayuda a administrar y coordinar la conmutación por error mediante comprobaciones de idoneidad y funciones de control de enrutamiento. Estas características supervisan continuamente la capacidad de la aplicación de recuperarse de los errores, de modo que pueda controlar la recuperación de la aplicación en las distintas Regiones de AWS, zonas de disponibilidad y localmente.

Para cargas de trabajo en centros de datos físicos o virtuales existentes o nubes privadas, AWS Elastic Disaster Recovery, disponible en AWS Marketplace, permite a las organizaciones configurar una estrategia de recuperación de desastres automatizada en AWS. CloudEndure también admite la recuperación de desastres entre regiones o AZ en AWS.

Antipatrones usuales:

- La implementación de técnicas de conmutación por error y de conmutación por recuperación idénticas puede producir una alteración cuando surge un error.

Beneficios de establecer esta práctica recomendada: La recuperación automatizada reduce el tiempo de recuperación al eliminar la posibilidad de que se produzcan errores manuales.

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

Guía para la implementación

- Automatice las rutas de recuperación. Para tiempos de recuperación cortos, las decisiones y las acciones humanas no pueden usarse para escenarios de alta disponibilidad. El sistema debe recuperarse automáticamente en cada situación.
  - Use la recuperación de desastres de Cloudendure para la conmutación por error y la restauración tras error automatizadas. La recuperación de desastres de CloudEndure replica continuamente las máquinas (incluido el sistema operativo, la configuración de estado del sistema, las bases de datos, las aplicaciones y los archivos) en un área de ensayo de bajo costo en su Cuenta de AWS de destino y región preferida. En caso de desastre, puede indicar a CloudEndure Disaster Recovery que lance automáticamente miles de máquinas en su estado aprovisionado completo en solo unos minutos.
    - Realizar la conmutación por error y la conmutación por recuperación de recuperación de desastres
    - CloudEndure Disaster Recovery

Recursos

Documentos relacionados:

- Socio de APN: socios que pueden ayudar con la recuperación de desastres
- Blog de arquitectura de AWS: serie de recuperación de desastres
- AWS Marketplace: productos que pueden usarse para la recuperación de desastres
- AWS Systems Manager Automation
- Recuperación de desastres de CloudEndure en AWS
- Recuperación de desastres de las cargas de trabajo en AWS: recuperación en la nube (documento técnico de AWS)

Videos relacionados:

- AWS re:Invent 2018: Patrones de arquitectura para aplicaciones activas-activas en varias regiones (ARC209-R2)

Eficiencia del rendimiento

El pilar de eficiencia del rendimiento incluye la capacidad de utilizar de forma eficaz los recursos de computación para satisfacer los requisitos del sistema, así como el mantenimiento de esta eficiencia a medida que la demanda cambia y las tecnologías evolucionan. Encontrará recomendaciones de implementación en el documento técnico Pilar de eficiencia del rendimiento.

Áreas de prácticas recomendadas

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

Selección de arquitectura

Preguntas

- RENDIMIENTO 1. ¿Cómo selecciona los recursos y la arquitectura de nube adecuados para su carga de trabajo?

RENDIMIENTO 1. ¿Cómo selecciona los recursos y la arquitectura de nube adecuados para su carga de trabajo?

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

Prácticas recomendadas

- PERF01-BP01 Descubrir y comprender los servicios y las características disponibles en la nube
- PERF01-BP02 Seguir las recomendaciones de su proveedor de servicios en la nube o de un socio adecuado para conocer los modelos arquitectónicos y las prácticas recomendadas
- PERF01-BP03 Tener en cuenta los costes en sus decisiones arquitectónicas
- PERF01-BP04 Analizar cómo sus decisiones afectan a los clientes y a la eficiencia de la arquitectura

- PERF01-BP05 Usar políticas y arquitecturas de referencia
- PERF01-BP06 Realizar pruebas comparativas para tomar decisiones arquitectónicas
- PERF01-BP07 Aplicar un enfoque basado en los datos en sus decisiones arquitectónicas

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

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

Patrones comunes de uso no recomendados:

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

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

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

Guía para la implementación

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

Pasos para la implementación

- Haga un inventario del software y la arquitectura de su carga de trabajo para los servicios relacionados. Decida la categoría de productos sobre la que desea obtener más información.

- Explore las ofertas de AWS para identificar y conocer los servicios y las opciones de configuración pertinentes que pueden ayudarle a mejorar el rendimiento y a reducir los costes y la complejidad operativa.
  - Novedades en AWS
  - Blog de AWS
  - AWS Skill Builder
  - Eventos y Webinars de AWS
  - Formación de AWS and Certifications
  - Canal de YouTube de AWS
  - Talleres de AWS
  - Comunidades de AWS
- Utilice entornos aislados (que no sean de producción) para aprender y experimentar con los nuevos servicios sin incurrir en costes extraordinarios.

Recursos

Documentos relacionados:

- Centro de arquitectura de AWS
- AWS Partner Network
- La Biblioteca de soluciones de AWS
- Centro de conocimiento de AWS
- Build modern applications on AWS

Vídeos relacionados:

- This is my Architecture

Ejemplos relacionados:

- Ejemplos de AWS
- Ejemplos de SDK de AWS

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

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

Patrones comunes de uso no recomendados:

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

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

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

Guía para la implementación

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

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

Pasos para la implementación

- Revise la documentación y las directrices de AWS y siga las prácticas recomendadas. Estos recursos pueden ayudarle a elegir y configurar los servicios de manera eficaz y a lograr un mejor rendimiento.
  - Documentación de AWS (como guías de usuario y documentos técnicos)
  - Blog de AWS

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

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- This is my Architecture

Ejemplos relacionados:

- Ejemplos de AWS
- Ejemplos de SDK de AWS

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

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

Patrones comunes de uso no recomendados:

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

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

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

Guía para la implementación

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

Pasos para la implementación

- Establezca objetivos de costes, como los límites presupuestarios de la carga de trabajo en la nube.
- Identifique los componentes clave (como las instancias y el almacenamiento) que influyen en los costes de su carga de trabajo. Puede usar el Calculadora de precios de AWS y AWS Cost Explorer para identificar los principales factores que influyen en los costes de su carga de trabajo.

- Utilice las prácticas recomendadas de optimización de costes de Well-Architected para optimizar los costes de estos componentes clave.
- Supervise y analice los costes de forma continua para identificar oportunidades que le permitan optimizar los gastos de su carga de trabajo.
  - Utilice AWS Budgets para recibir alertas sobre costes inaceptables.
  - Utilice AWS Compute Optimizer o bien AWS Trusted Advisor para obtener recomendaciones sobre la optimización de costes.
  - Utilice la detección de anomalías en los costes de AWS para detectar automáticamente las anomalías en los costes y analizar la causa raíz.

Recursos

Documentos relacionados:

- A Detailed Overview of the Cost Intelligence Dashboard
- Centro de arquitectura de AWS
- AWS Partner Network
- La Biblioteca de soluciones de AWS
- Centro de conocimiento de AWS

Vídeos relacionados:

- This is my Architecture
- Optimize performance and cost for your AWS compute

Ejemplos relacionados:

- Ejemplos de AWS
- Ejemplos de SDK de AWS
- Rightsizing with Compute Optimizer and Memory utilization enabled
- AWS Compute Optimizer Demo code

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

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

Patrones comunes de uso no recomendados:

- Da por hecho que habría que implementar todas las ventajas relacionadas con el rendimiento, aunque esta implementación tenga repercusiones.
- Solo evalúa los cambios en las cargas de trabajo cuando un problema de rendimiento ha alcanzado un punto crítico.

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

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

Guía para la implementación

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

Pasos para la implementación

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

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- Diseñe un plan de monitoreo
- Optimize applications through Amazon CloudWatch RUM
- Demostración de Amazon CloudWatch Synthetics

Ejemplos relacionados:

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

PERF01-BP05 Usar políticas y arquitecturas de referencia

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

Patrones comunes de uso no recomendados:

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

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

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

Guía para la implementación

Contar con políticas internas para seleccionar los recursos y la arquitectura proporciona estándares y pautas que pueden seguirse al tomar decisiones arquitectónicas. Estas directrices agilizan el proceso de toma de decisiones a la hora de elegir el servicio de nube correcto y pueden ayudar a mejorar la eficiencia del rendimiento. Despliegue la carga de trabajo utilizando políticas o arquitecturas de referencia. Integre los servicios en su despliegue en la nube y, a continuación, utilice las pruebas de rendimiento para asegurarse de que puede seguir cumpliendo los requisitos establecidos.

Pasos para la implementación

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

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- This is my Architecture

Ejemplos relacionados:

- Ejemplos de AWS
- Ejemplos de SDK de AWS

PERF01-BP06 Realizar pruebas comparativas para tomar decisiones arquitectónicas

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

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 realizar una prueba de carga.

Puede crear sus propias pruebas comparativas personalizadas, o bien usar un estándar industrial, 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 realizar en su arquitectura.

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

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

Pasos para la implementación

- Defina las métricas (como el uso de la CPU, la latencia o el rendimiento) para evaluar el rendimiento de su carga de trabajo.
- Identifique y configure una herramienta de pruebas comparativas que sea adecuada para su carga de trabajo. Puede utilizar servicios de AWS (como Amazon CloudWatch) o una herramienta de terceros que sea compatible con su carga de trabajo.
- Realice las pruebas comparativas y supervise las métricas durante la prueba.
- Analice y documente los resultados de las pruebas comparativas para identificar problemas y cuellos de botella.
- Utilice los resultados de las pruebas para tomar decisiones arquitectónicas y ajustar su carga de trabajo. Para ello, puede ser necesario cambiar los servicios o adoptar nuevas características.
- Tras realizar el ajuste, repita las pruebas de su carga de trabajo.

Recursos

Documentos relacionados:

- Centro de arquitectura de AWS
- AWS Partner Network
- La Biblioteca de soluciones de AWS
- Centro de conocimiento de AWS
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics

Vídeos relacionados:

- This is my Architecture

- Optimize applications through Amazon CloudWatch RUM
- Demostración de Amazon CloudWatch Synthetics

Ejemplos relacionados:

- Ejemplos de AWS
- Ejemplos de SDK de AWS
- Pruebas de carga distribuidas
- Medición del tiempo de carga de la página con Amazon CloudWatch Synthetics
- Cliente web de Amazon CloudWatch RUM

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

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

Patrones comunes de uso no recomendados:

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

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

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

Guía para la implementación

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

La lista de tareas pendientes para las cargas de trabajo críticas no solo debe incluir casos de usuario que brinden una funcionalidad relevante para la empresa y los usuarios, sino también casos

técnicos que conformen un plan arquitectónico para la carga de trabajo. Este plan se nutre de nuevos avances en tecnología y nuevos servicios, que se incorporan con arreglo a los datos y de forma justificada. Esto garantiza que la arquitectura siempre está preparada para el futuro y no se queda anquilosada.

Pasos para la implementación

- Hable con las principales partes interesadas para definir los requisitos de la carga de trabajo, incluidas las consideraciones de rendimiento, disponibilidad y costes. Tenga en cuenta factores como la cantidad de usuarios y el modo de uso de la carga de trabajo.
- Cree un plan arquitectónico o una lista de tareas pendientes relacionadas con la tecnología que tengan la misma prioridad que las tareas pendientes relacionadas con la funcionalidad.
- Evalúe los diferentes servicios en la nube (para obtener más información, consulte PERF01-BP01 Descubrir y comprender los servicios y las características disponibles en la nube).
- Analice diferentes patrones arquitectónicos, como los microservicios o la computación sin servidor, que se ajusten a sus requisitos de rendimiento (para obtener más información, consulte PERF01-BP02 Seguir las recomendaciones de su proveedor de servicios en la nube o de un socio adecuado para conocer los modelos arquitectónicos y las prácticas recomendadas).

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

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

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- This is my Architecture

Ejemplos relacionados:

- Ejemplos de AWS
- Ejemplos de SDK de AWS

Computación y hardware

RENDIMIENTO 2. ¿Cómo selecciona y utiliza los recursos de computación en su carga de trabajo?

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

Prácticas recomendadas

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

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

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

Patrones comunes de uso no recomendados:

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

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

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

Guía para la implementación

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

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

Pasos para la implementación

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

Servicio de AWS	Características clave	Casos de uso habituales
Amazon Elastic Compute Cloud (Amazon EC2)	Cuenta con una opción dedicada para hardware, requisitos de licencia, una amplia selección de distintas familias de instancias, tipos de procesadores y acelerado res de cómputo	Migraciones mediante lift-and-shift, aplicación monolític a, entornos híbridos, aplicacio nes empresariales
Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS)	Despliegue sencillo, entornos coherentes, escalable	Microservicios, entornos híbridos
AWS Lambda	Servicio de computación sin servidor que ejecuta código como respuesta a eventos y administra automáticamente los recursos de computación subyacentes.	Microservicios, aplicaciones basadas en eventos
AWS Batch	Aprovisiona y escala de forma eficiente y dinámica los recursos de computaci ón de Amazon Elastic Container Service (Amazon	HPC, entrenamiento de modelos de ML

Servicio de AWS	Características clave	Casos de uso habituales
	ECS), Amazon Elastic Kubernetes Service (Amazon EKS)y AWS Fargate con la opción para usar instancias de spot o bajo demanda en función de los requisitos de su trabajo.	
Amazon Lightsail	Aplicación de Linux y Windows preconfigurada para ejecutar cargas de trabajo pequeñas	Aplicaciones web simples, sitio web personalizado

3. Calcule el coste (por ejemplo, el coste por hora o la transferencia de datos) y los gastos generales de administración (como la aplicación de parches y el escalamiento) asociados a cada opción de computación.
4. Realice experimentos y pruebas comparativas en un entorno que no sea de producción para identificar qué opción de computación puede satisfacer mejor los requisitos de su carga de trabajo.
5. Una vez que haya probado e identificado su nueva solución de computación, planifique la migración y valide sus métricas de rendimiento.
6. Use herramientas de supervisión de AWS como Amazon CloudWatch y servicios de optimización como AWS Compute Optimizer para optimizar los recursos de computación de manera continua en función de patrones de uso reales.

Recursos

Documentos relacionados:

- Cloud Compute with AWS
- Amazon EC2 Instance Types
- Amazon EKS Containers: Amazon EKS Worker Nodes
- Amazon ECS Containers: Amazon ECS Container Instances
- Funciones: configuración de funciones de Lambda
- Prescriptive Guidance for Containers

- Prescriptive Guidance for Serverless

Vídeos relacionados:

- How to choose compute option for startups (Cómo elegir la opción de computación para las empresas emergentes)
- Optimize performance and cost for your AWS compute
- Amazon EC2 foundations
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Deploy ML models for inference at high performance and low cost
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2

Ejemplos relacionados:

- Migrating the Web application to containers
- Run a Serverless Hello World

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

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

Patrones comunes de uso no recomendados:

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

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

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

Guía para la implementación

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

Pasos para la implementación

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

Opción de configuración	Ejemplos
Tipo de instancia	• Las instancias optimizadas para la computación son ideales para las cargas de trabajo que requieren una relación entre vCPU y memoria más alta. • Las instancias optimizadas para la memoria ofrecen grandes cantidades de memoria para admitir cargas de trabajo que hacen un uso intensivo de la memoria. • Las instancias optimizadas para el almacenamiento están diseñadas para cargas de trabajo que requieren un alto acceso secuencial de lectura y escritura (IOPS) al almacenamiento local.

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

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

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- Amazon EC2 foundations
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Optimize performance and cost for your AWS compute

Ejemplos relacionados:

- Rightsizing with Compute Optimizer and Memory utilization enabled
- AWS Compute Optimizer Demo code

PERF02-BP03 Recopilar métricas relacionadas con la computación

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

Patrones comunes de uso no recomendados:

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

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

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

Guía para la implementación

Las cargas de trabajo en la nube pueden generar grandes volúmenes de datos, como métricas, registros y eventos. En Nube de AWS, la recopilación de métricas es un paso crucial para mejorar la seguridad, la rentabilidad, el rendimiento y la sostenibilidad. AWS ofrece una amplia variedad de métricas relacionadas con el rendimiento a través de servicios de supervisión como Amazon CloudWatch, que le proporcionan una información valiosa. Las métricas como la utilización de la CPU, la utilización de la memoria, las operaciones de E/S del disco y la entrada y salida de la red pueden proporcionar información sobre los niveles de utilización o los cuellos de botella del rendimiento. Utilice estas métricas como parte de un enfoque basado en datos para ajustar y optimizar activamente los recursos de su carga de trabajo. En un supuesto ideal, debería recopilar todas las métricas relacionadas con sus recursos de computación en una única plataforma que tuviera políticas de retención implementadas para satisfacer los objetivos operativos y financieros.

Pasos para la implementación

1. Identifique qué métricas relacionadas con el rendimiento son relevantes para su carga de trabajo. Debe recopilar métricas sobre la utilización de los recursos y la forma en que funciona su carga de trabajo en la nube (por ejemplo, el tiempo de respuesta y el rendimiento).
   a. Métricas predeterminadas de Amazon EC2

   b. Métricas predeterminadas de Amazon ECS
   c. Métricas predeterminadas de Amazon EKS
   d. Métricas predeterminadas de Lambda
   e. Métricas de memoria y disco de Amazon EC2
2. Elija y configure la solución de registro y supervisión adecuada para su carga de trabajo.
   a. Observabilidad nativa de AWS
   b. AWS Distro para OpenTelemetry
   c. Amazon Managed Service for Prometheus
3. Defina el filtro y la agregación que se necesitan para las métricas en función de los requisitos de su carga de trabajo.
   a. Cuantifique métricas de aplicación personalizadas con Amazon CloudWatch Logs y filtros de métrica
   b. Recopile métricas personalizadas con el etiquetado estratégico de Amazon CloudWatch
4. Configure políticas de retención de datos para que las métricas se ajusten a los objetivos operativos y de seguridad.
   a. Retención de datos predeterminada para métricas de CloudWatch
   b. Retención de datos predeterminada para CloudWatch Logs
5. Si es necesario, cree alarmas y notificaciones para sus métricas, lo que le ayudará a responder de manera proactiva a los problemas relacionados con el rendimiento.
   a. Cree alarmas para métricas personalizadas con la detección de anomalías de Amazon CloudWatch
   b. Cree métricas y alarmas para páginas web específicas con Amazon CloudWatch RUM
6. Utilice la automatización para desplegar los agentes de agregación de métricas y registros.
   a. Automatización de AWS Systems Manager
   b. Colector de OpenTelemetry

Recursos

Documentos relacionados:

- Amazon CloudWatch documentation
- Recopilación de métricas y registros de instancias Amazon EC2 y en los servidores en las instalaciones con el agente de CloudWatch

- Accessing Amazon CloudWatch Logs for AWS Lambda
- Using CloudWatch Logs with container instances
- Publique métricas personalizadas
- AWS Answers: Centralized Logging
- Servicios de AWS que publican métricas de CloudWatch
- Monitoring Amazon EKS on AWS Fargate

Vídeos relacionados:

- Application Performance Management on AWS

Ejemplos relacionados:

- Level 100: Monitoring with CloudWatch Dashboards
- Level 100: Monitoring Windows EC2 instance with CloudWatch Dashboards
- Level 100: Monitoring an Amazon Linux EC2 instance with CloudWatch Dashboards

PERF02-BP04 Configurar y dimensionar correctamente los recursos de computación

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

Patrones comunes de uso no recomendados:

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

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

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

Guía para la implementación

Un dimensionamiento correcto permite a las organizaciones gestionar la infraestructura en la nube de manera eficiente y rentable, al tiempo que abordan sus necesidades empresariales. Un aprovisionamiento excesivo de recursos en la nube puede generar costes adicionales, mientras que un aprovisionamiento insuficiente puede provocar un rendimiento deficiente y una experiencia de cliente negativa. AWS proporciona herramientas como AWS Compute Optimizer y AWS Trusted Advisor que utilizan datos históricos para ofrecer recomendaciones que permiten dimensionar correctamente los recursos informáticos.

Pasos para la implementación

- Elija el tipo de instancia que mejor se adapte a sus necesidades:
  - How do I choose the appropriate Amazon EC2 instance type for my workload? (¿Cómo elijo el tipo de instancia de EC2 apropiado para mi carga de trabajo?)
  - Selección de tipo de instancia basada en atributos para la flota de Amazon EC2
  - Crear un grupo de Auto Scaling con la selección de un tipo de instancia basada en atributos
  - Optimizar los costes computacionales de Kubernetes con la consolidación de Karpenter
- Analice las distintas características de rendimiento de su carga de trabajo y la relación que tienen con el uso de memoria, redes y CPU. Use estos datos para elegir recursos que encajen bien con el perfil de la carga de trabajo y los objetivos de rendimiento.
- Controle el uso de los recursos con las herramientas de supervisión de AWS, comoAmazon CloudWatch.
- Seleccione la configuración correcta para cada recurso informático.
  - Para las cargas de trabajo efímeras, evalúe las métricas de Amazon CloudWatch de instancias como `CPUUtilization` para identificar si la instancia está infrautilizada o sobreutilizada.

    - En las cargas de trabajo estables, consulte regularmente las herramientas de dimensionamiento de AWS, como AWS Compute OptimizerAWS Trusted Advisor, para identificar oportunidades de optimizar y dimensionar las instancias de forma correcta.
        - Well-Architected Lab - Rightsizing Recommendations (Laboratorio de Well-Architected: recomendaciones de redimensionamiento)
        - Well-Architected Lab - Rightsizing with Compute Optimizer (Laboratorio de Well-Architected: redimensionamiento con Compute Optimizer)
- Pruebe los cambios de configuración en un entorno que no sea de producción antes de implementarlos en un entorno activo.
- Revalúe continuamente las nuevas ofertas de computación y compárelas con las necesidades de la carga de trabajo.

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- Amazon EC2 foundations
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2
- Deploy ML models for inference at high performance and low cost
- Optimize performance and cost for your AWS compute
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Simplifying Data Processing to Enhance Innovation with Serverless Tools

Ejemplos relacionados:

- Rightsizing with Compute Optimizer and Memory utilization enabled
- AWS Compute Optimizer Demo code

PERF02-BP05 Escalar los recursos computacionales de forma dinámica

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

Patrones comunes de uso no recomendados:

- Reacciona a las alarmas aumentando manualmente la capacidad.
- Utiliza las mismas directrices de dimensionamiento (por lo general, una infraestructura estática) que en el entorno local.
- Deja la capacidad aumentada después de un evento de ajuste de escala en lugar de volver a desescalar verticalmente.

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

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

Guía para la implementación

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

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

- Enfoque de seguimiento de objetivos: supervise la métrica de escalamiento y aumente o reduzca de forma automática la capacidad en función de sus necesidades.
- Escalamiento predictivo: escale de antemano según las tendencias diarias y semanales previstas.
- Enfoque basado en programación: establezca su propia programación de escalamiento según los cambios de carga predecibles.

- Escalamiento de servicios: elija servicios (como los servicios sin servidor) diseñados para escalar automáticamente.

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

Pasos para la implementación

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

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

- Normalmente, se habla del escalamiento en relación con los servicios de computación, como las instancias de Amazon EC2 o las funciones de AWS Lambda. No olvide que también debe tener en cuenta la configuración de otros servicios no computacionales como AWS Glue para satisfacer la demanda.
- Asegúrese de que las métricas de escalamiento se ajustan a las características de la carga de trabajo que se está desplegando. Si está desplegando una aplicación de transcodificación de vídeo, se espera una utilización del 100 % de la CPU y no debería ser su métrica principal. En su lugar, utilice la profundidad de la cola de trabajos de transcodificación. Puede usar una

métrica personalizada para su política de escalamiento, si es necesario. Para elegir las métricas adecuadas, tenga en cuenta las siguientes directrices para Amazon EC2:

  - La métrica debe ser una métrica de utilización válida y describir el grado de ocupación de una instancia.
  - El valor de la métrica debe aumentar o disminuir proporcionalmente al número de instancias del grupo de Auto Scaling.
- Asegúrese de utilizar el escalado dinámico en vez del escalado manual para su grupo de Auto Scaling. También le recomendamos que utilice políticas de escalado de seguimiento de destino en su escalado dinámico.
- Compruebe que los despliegues de la carga de trabajo puedan gestionar ambos eventos de escalamiento (escalamiento y desescalamiento verticales). Como ejemplo, puede usar el historia de actividades para verificar una actividad de escalamiento para un grupo de Auto Scaling.
- Evalúe los patrones predecibles de su carga de trabajo y escale de forma proactiva al anticiparse a los cambios previstos y planeados en la demanda. Con el escalamiento predictivo, puede eliminar la necesidad de aprovisionar capacidad en exceso. Para obtener más detalles, consulte Predictive scaling with Amazon EC2 Auto Scaling (Escalamiento predictivo con Amazon EC2 Auto Scaling).

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- Amazon EC2 foundations
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2

- Optimize performance and cost for your AWS compute
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Build a cost-, energy-, and resource-efficient compute environment (Crear un entorno de computación rentable, eficiente en términos de costes, energía y recursos)

Ejemplos relacionados:

- Amazon EC2 Auto Scaling Group Examples
- Implement Autoscaling with Karpenter

PERF02-BP06 Utilización de aceleradores computacionales optimizados basados en hardware

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

Patrones comunes de uso no recomendados:

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

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 ejecutar determinadas funciones de procesamiento de manera más eficiente.

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

Guía para la implementación

Las instancias de computación acelerada proporcionan acceso a aceleradores de computación basados en hardware, como las GPU y las FPGA. Estos aceleradores de hardware realizan ciertas funciones, como el procesamiento gráfico o la concordancia de patrones de datos, de forma más eficiente que las alternativas basadas en CPU. Muchas cargas de trabajo aceleradas, como el renderizado, la transcodificación y el machine learning, son muy variables en cuanto al uso de

recursos. Ejecute este hardware únicamente durante el tiempo necesario y retírelo de forma automatizada cuando no sea necesario para mejorar la eficiencia general del rendimiento.

Pasos para la implementación

- Identifique qué instancias de computación acelerada pueden satisfacer sus necesidades.
- Para las cargas de trabajo de machine learning, utilice hardware personalizado específico para su carga de trabajo, como AWS Trainium, AWS Inferentiay Amazon EC2 DL1. Las instancias de AWS Inferentia, como las instancias Inf2, ofrecen hasta un 50 % más de rendimiento por vatio en comparación con instancias de Amazon EC2 comparables.
- Recopile las métricas de uso de sus instancias de computación acelerada. Por ejemplo, puede usar un agente de CloudWatch para recopilar métricas como `utilization_gpu` y `utilization_memory` para sus GPU, como se muestra en Collect NVIDIA GPU metrics with Amazon CloudWatch (Recopilación de métricas de CPU de 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.
  - Optimizar la configuración de GPU
  - GPU Monitoring and Optimization in the Deep Learning AMI (Supervisión y optimización de la GPU en la AMI de aprendizaje profundo)
  - Optimizing I/O for GPU performance tuning of deep learning training in Amazon SageMaker AI (Optimización de la E/S para el ajuste del rendimiento de la GPU en el entrenamiento del aprendizaje profundo en Amazon SageMaker)
- Utilice las bibliotecas de alto rendimiento y los controladores de GPU más recientes.
- Use la automatización para liberar instancias de GPU cuando no se estén usando.

Recursos

Documentos relacionados:

- GPU instances
- Instances with AWS Trainium
- Instances with AWS Inferentia
- Let’s Architect! Architecting with custom chips and accelerators (Arquitectura con chips y aceleradores personalizados)

- Computación acelerada
- Instancias VT1 de Amazon EC2
- How do I choose the appropriate Amazon EC2 instance type for my workload? (¿Cómo elijo el tipo de instancia de EC2 apropiado para mi carga de trabajo?)
- Choose the best AI accelerator and model compilation for computer vision inference with Amazon SageMaker AI (Elija el mejor acelerador de IA y compilación de modelos para la inferencia de visión artificial con Amazon SageMaker)

Vídeos relacionados:

- How to select Amazon EC2 GPU instances for deep learning (Cómo seleccionar las instancias de GPU de Amazon EC2 para el aprendizaje profundo)
- Deploying Cost-Effective Deep Learning Inference (Despliegue rentable de la inferencia del aprendizaje profundo)

Gestión de datos

RENDIMIENTO 3. ¿Cómo almacena, administra y accede a los datos de su carga de trabajo?

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, dinámico), y restricciones de disponibilidad y durabilidad. Las cargas de trabajo Well-Architected utilizan almacenes de datos diseñados específicamente que admiten diferentes características para mejorar el rendimiento.

Prácticas recomendadas

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

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

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

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 que es adecuada para las necesidades de su carga de trabajo.

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

Guía para la implementación

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

Pasos para la implementación

- Realice un inventario de los distintos tipos de datos que existen en su carga de trabajo.

- Estudie y documente las características y los requisitos de los datos, como:
  - Tipo de datos (no estructurados, semiestructurados o relacionales)
  - Volumen y crecimiento de los datos
  - Durabilidad de los datos: persistentes, efímeros o transitorios
  - Requisitos de ACID (atomicidad, consistencia, aislamiento, durabilidad)
  - Patrones de acceso a los datos (lectura o escritura intensivas)
  - Latencia
  - Rendimiento
  - IOPS (operaciones de entrada/salida por segundo)
  - Período de retención de los datos
- Obtenga información sobre los diferentes almacenes de datos 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 Descubrir y comprender los servicios y las características disponibles en la nube). Estos son algunos ejemplos de tecnologías de almacenamiento de AWS y sus principales características:

Tipo	Servicios de AWS	Características clave
Clases de almacenamiento	Amazon S3	Escalabilidad ilimitada, alta disponibilidad y múltiples opciones de accesibilidad. La transferencia y el acceso a los objetos dentro y fuera de Amazon S3 puede utilizar un servicio como Aceleración de transferencia o bien Puntos de accesopara respaldar su ubicación, necesidades de seguridad y patrones de acceso.
Almacenamiento de archivos	Amazon Glacier	Diseñado para archivar datos.
Almacenamiento en streaming	Amazon Kinesis	Ingesta y almacenamiento eficientes de datos de streaming.

Tipo	Servicios de AWS	Características clave
	Amazon Managed Streaming for Apache Kafka (Amazon MSK)	
Sistema de archivos compartidos	Amazon Elastic File System (Amazon EFS)	Sistema de archivos montable al que pueden acceder varios tipos de soluciones de computación.
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. En Amazon FSx, su latencia, rendimiento y E/S por segundo varían según el sistema de archivos y deben tenerse en cuenta a la hora de seleccionar el sistema de archivos adecuado para sus necesidades de carga de trabajo.

Tipo	Servicios de AWS	Características clave
Almacenamiento de bloques	Amazon Elastic Block Store (Amazon EBS)	Servicio de almacenamiento de bloques escalable y de alto rendimiento diseñado para Amazon Elastic Compute Cloud (Amazon EC2). Amazon EBS incluye almacenamiento respaldado por SSD para cargas de trabajo transaccionales y de IOPS intensivas, y almacenamiento respaldado por HDD para cargas de trabajo de rendimiento intensivo.
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 los 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.

Tipo	Servicios de AWS	Características clave
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.
Base de datos para 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.

Tipo	Servicios de AWS	Características clave
Bases 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.
Base de datos de grafos	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 grafos para detección de fraude, redes sociales y motores de recomendaciones.

Tipo	Servicios de AWS	Características clave
Base de datos de serie temporal	Amazon Timestream	Se usa para recopilar, sintetizar y obtener informaci ón de forma eficaz a partir de datos que cambian con el tiempo. Las aplicaciones de IoT, DevOps y telemetrí a industrial pueden utilizar bases de datos de serie temporal.
Columnas anchas	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.

Tipo	Servicios de AWS	Características clave
Libro mayor	Amazon Quantum Ledger Database (Amazon QLDB)	Proporcionan 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, utilice la arquitectura de datos moderna de AWS para integrar su lago de datos, almacenamiento de datos y almacenes de datos personalizados.
- Las principales preguntas que debe hacerse al elegir un almacén de datos para su carga de trabajo son las siguientes:

Pregunta	Aspectos que deben tenerse en cuenta
¿Cómo están estructurados los datos?	• Si los datos no están estructurados, considere la posibilidad de usar un almacén de objetos como Amazon S3 o una base de datos NoSQL como Amazon DocumentDB • Para los datos de clave-valor, considere la posibilidad de usar DynamoDB, Amazon ElastiCache (Redis OSS) o bien Amazon MemoryDB
¿Qué nivel de integridad referencial se requiere?	• Para las restricciones de clave externa, las bases de datos relacionales como Amazon RDS y Aurora pueden proporcionar este nivel de integridad.

Pregunta	Aspectos que deben tenerse en cuenta
	• 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 una base de datos NoSQL,puede utilizar lecturas con coherencia fuerte con DynamoDB.
¿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) 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 particionamiento.

Pregunta	Aspectos que deben tenerse en cuenta
¿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 con muchas lecturas pueden beneficiarse de una capa de caché, como ElastiCache o bien DAX si la base de datos es DynamoDB. • Las lecturas también pueden descargarse en réplicas de lectura con bases de datos relacionales como Amazon RDS.
¿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 realizar cambios, considere la posibilidad de usar una base de datos NoSQL, como. DynamoDB. • En el caso de los patrones de lectura complejos y de alto rendimiento (como una combinación) que tienen coherencia, use Amazon RDS. • Para las consultas analíticas, considere la posibilidad de usar una base de datos en columnas como Amazon Redshift o de exportar los datos a Amazon S3 y realizar análisis mediante Athena o bien Amazon Quick.

Pregunta	Aspectos que deben tenerse en cuenta
¿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 períodos de tiempo.
¿Existe el deseo de evitar los motores de bases de datos comerciales o los costes de licencia?	• Considere la posibilidad de utilizar motores de código abierto como PostgreSQL y MySQL en Amazon RDS o Aurora. • Utilice AWS Database Migration Service y AWS Schema Conversion Tool para realizar 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 costes operativos.

Pregunta	Aspectos que deben tenerse en cuenta
¿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 necesite mantener la compatibilidad con las bases de datos que se utilizan allí. Amazon RDS es totalmente compatible con las diferentes versiones de motores que admite, como Microsoft SQL Server, Oracle, MySQL y PostgreSQL.

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

Recursos

Documentos relacionados:

- Amazon EBS Volume Types (Tipos de volumen de Amazon EBS)
- Amazon EC2 Storage (Almacenamiento de Amazon EC2)
- Amazon EFS: Amazon EFS Performance (Amazon EFS: rendimiento de Amazon EFS)
- Amazon FSx for Lustre Performance
- Amazon FSx for Windows File Server Performance (Rendimiento de Amazon FSx for Windows File Server)
- Amazon Glacier: Amazon Glacier Documentation
- Amazon S3: Request Rate and Performance Considerations (Amazon S3: tasa de solicitud y consideraciones de rendimiento)
- Cloud Storage with AWS (Almacenamiento en la nube con AWS)
- Características de E/S de Amazon EBS
- Cloud Databases with AWS
- AWS Database Caching
- DynamoDB Accelerator
- Prácticas recomendadas para Amazon Aurora
- Desempeño de Amazon Redshift

- Amazon Athena top 10 performance tips
- Amazon Redshift Spectrum best practices
- Amazon DynamoDB best practices
- Choose between Amazon EC2 and Amazon RDS
- Best Practices for Implementing Amazon ElastiCache

Vídeos relacionados:

- Deep dive on Amazon EBS
- Optimize your storage performance with Amazon S3
- Modernize apps with purpose-built databases
- Amazon Aurora storage demystified: How it all works
- Amazon DynamoDB deep dive: Advanced design patterns

Ejemplos relacionados:

- Amazon EFS CSI Driver (Controlador CSI de Amazon EFS)
- Amazon EBS CSI Driver (Controlador CSI de Amazon EBS)
- Amazon EFS Utilities (Utilidades de Amazon EFS)
- Amazon EBS Autoscale (Escala automática de Amazon EBS)
- Amazon S3 Examples (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 Evaluar las opciones de configuración disponibles

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

Patrones comunes de uso no recomendados:

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

Beneficios de establecer esta práctica recomendada: si explora y experimenta con las configuraciones de almacenamiento de datos, puede reducir el coste de la infraestructura, mejorar el rendimiento y reducir el esfuerzo necesario para mantener sus cargas de trabajo.

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

Guía para la implementación

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

Pasos para la implementación

- Estudie las configuraciones actuales (como el tipo de instancia, el tamaño de almacenamiento o la versión del motor de base de datos) de su almacén de datos.
- Consulte la documentación y las prácticas recomendadas de AWS para obtener información sobre las opciones de configuración recomendadas que pueden ayudarle a mejorar el rendimiento de su almacén de datos. Las principales opciones de almacenamiento de datos que debe tener en cuenta son las siguientes:

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 utilizando 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 añadir réplicas de lectura para descargar las partes de lectura de la carga de trabajo. • Las bases de datos sin servidor, como DynamoDB, se escalarán automátic amente. Asegúrese de que tiene suficient es unidades de capacidad de lectura (RCU) aprovisionadas para gestionar la carga de trabajo.

Opción de configuración	Ejemplos
Escalamiento de escrituras (como la fragmentación de claves de partición o la introducción de una cola)	• En el caso de las bases de datos relaciona les, puede aumentar el tamaño de la instancia para acomodar una mayor carga de trabajo o aumentar las IOPS aprovisio nadas 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 desbordad a. • Si agrupa las solicitudes de escritura en lugar de crear muchas transacciones de corta duración, puede mejorar el rendimien to de las bases de datos relacionales con un gran volumen de operaciones de escritura. • Las bases de datos sin servidor como DynamoDB pueden escalar el rendimiento de escritura automáticamente o ajustando las unidades de capacidad de escritura (WCU) aprovisionadas en función del modo de capacidad. • Puede tener problemas con las particiones activas si alcanza los límites de rendimien to de una clave de partición determinada. Esto puede mitigarse eligiendo una clave de partición distribuida de manera más uniforme o particionando la escritura en función de la clave de partición.

Opción de configuración	Ejemplos
Políticas para administrar el ciclo de vida de los conjuntos de datos	• Puede usar el Ciclo de vida de Amazon S3 para administrar los objetos a lo largo de su ciclo de vida. Si los patrones de acceso no se conocen, experimentan cambios o son impredecibles, puede usar 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 más baratos. Puede utilizar las métricas de Amazon S3 Storage Lens para identific ar las oportunidades de optimización y las lagunas en la administración del ciclo de vida. • Administración del ciclo de vida de Amazon EFS administra automáticamente el almacenamiento en los 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 escalamie nto automático para hacer frente a los picos de carga.

- Realice experimentos y pruebas comparativas en un entorno que no sea de producción para identificar qué opción de computación se ajusta a los requisitos de la carga de trabajo.
- Una vez hecho esto, planifique la migración y valide las métricas de rendimiento.

- Use las herramientas de supervisión de AWS (como Amazon CloudWatch) y de optimización (como Amazon S3 Storage Lens) para optimizar continuamente el almacén de datos con patrones de uso reales.

Recursos

Documentos relacionados:

- Cloud Storage with AWS (Almacenamiento en la nube con AWS)
- Amazon EBS Volume Types (Tipos de volumen de Amazon EBS)
- Amazon EC2 Storage (Almacenamiento de Amazon EC2)
- Amazon EFS: Amazon EFS Performance (Amazon EFS: rendimiento de Amazon EFS)
- Amazon FSx for Lustre Performance
- Amazon FSx for Windows File Server Performance (Rendimiento de Amazon FSx for Windows File Server)
- Amazon Glacier: Amazon Glacier Documentation
- Amazon S3: Request Rate and Performance Considerations (Amazon S3: tasa de solicitud y consideraciones de rendimiento)
- Cloud Storage with AWS (Almacenamiento en la nube con AWS)
- Cloud Storage with AWS (Almacenamiento en la nube con AWS)
- Características de E/S de Amazon EBS
- Cloud Databases with AWS
- AWS Database Caching
- DynamoDB Accelerator
- Prácticas recomendadas para Amazon Aurora
- Desempeño de Amazon Redshift
- Amazon Athena top 10 performance tips
- Amazon Redshift Spectrum best practices
- Amazon DynamoDB best practices

Vídeos relacionados:

- Deep dive on Amazon EBS

- Optimize your storage performance with Amazon S3
- Modernize apps with purpose-built databases
- Amazon Aurora storage demystified: How it all works
- Amazon DynamoDB deep dive: Advanced design patterns

Ejemplos relacionados:

- Amazon EFS CSI Driver (Controlador CSI de Amazon EFS)
- Amazon EBS CSI Driver (Controlador CSI de Amazon EBS)
- Amazon EFS Utilities (Utilidades de Amazon EFS)
- Amazon EBS Autoscale (Escala automática de Amazon EBS)
- Amazon S3 Examples (Ejemplos de Amazon S3)
- Amazon DynamoDB Examples
- AWS Database migration samples
- Database Modernization Workshop
- Working with parameters on your Amazon RDS for Postgress DB

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

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

Patrones comunes de uso no recomendados:

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

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

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

Guía para la implementación

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

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

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

Pasos para la implementación

1. Identifique las métricas de rendimiento clave del almacén de datos que desee supervisar.
    a. Métricas y dimensiones de Amazon S3
    b. Supervisión de las métricas de una instancia de Amazon RDS
    c. Supervisión de la carga de bases de datos con Información sobre rendimiento en Amazon RDS
    d. Descripción general de la supervisión mejorada
    e. Métricas y dimensiones de DynamoDB
    f. Supervisión de DynamoDB Accelerator
    g. Supervisión de Amazon MemoryDB con Amazon CloudWatch

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

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- AWS purpose-built databases
- Amazon Aurora storage demystified: How it all works
- Amazon DynamoDB deep dive: Advanced design patterns
- Best Practices for Monitoring Redis Workloads on Amazon ElastiCache

Ejemplos relacionados:

- Level 100: Monitoring with CloudWatch Dashboards
- AWS Dataset Ingestion Metrics Collection Framework
- Amazon RDS Monitoring Workshop

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

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

Patrones comunes de uso no recomendados:

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

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

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

Guía para la implementación

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

uso de recursos y generar cuellos de botella, lo que reduce la eficiencia general de los almacenes de datos.

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

Pasos para la implementación

- Estudie y analice las consultas de datos críticos que se realizan en el almacén de datos.
- Identifique las consultas de ejecución lenta del almacén de datos y utilice planes de consulta para conocer su estado actual.
  - Análisis del plan de consulta en Amazon Redshift
  - Uso de EXPLAIN y EXPLAIN ANALYZE en Athena
- Implemente estrategias para mejorar el rendimiento de las consultas. Algunas de las estrategias clave son:
  - Uso de un formato de archivo en columnas (como Parquet u ORC).
  - Comprimir los datos en el almacén de datos para reducir el espacio de almacenamiento y la operación de E/S.
  - Crear particiones de datos para dividir la información en partes más pequeñas y reducir el tiempo de análisis de los datos.
    - Partición de datos en Athena
    - Particiones y distribución de datos
  - Indexar los datos de las columnas más frecuentes de la consulta.
  - Elegir la operación de unión correcta para la consulta. Cuando una dos tablas, especifique la tabla mayor en el lado izquierdo de la unión y la tabla menor en el lado derecho de la unión.
  - Usar una solución de almacenamiento en caché distribuida para mejorar la latencia y reducir la cantidad de operaciones de E/S de la base de datos.
  - Realizar un mantenimiento regular, como la ejecución de estadísticas.
- Experimente y pruebe estrategias en un entorno que no sea de producción.

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- Optimize Data Pattern using Amazon Redshift Data Sharing
- Optimize Amazon Athena Queries with New Query Analysis Tools

Ejemplos relacionados:

- Amazon EFS CSI Driver (Controlador CSI de Amazon EFS)

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

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

Patrones comunes de uso no recomendados:

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

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

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

Guía para la implementación

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

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

Almacenamiento en caché del lado del cliente

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

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

Almacenamiento remoto en caché

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

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

Pasos para la implementación

1. Identifique las bases de datos, las API y los servicios de red que podrían beneficiarse del almacenamiento en caché. Los servicios que tienen cargas de trabajo de lectura pesadas, tienen una alta relación de lectura y escritura o son caros de escalar son candidatos para el almacenamiento en caché.
   - Almacenamiento en caché de base de datos
   - Habilitación del almacenamiento en caché de la API para mejorar la capacidad de respuesta
2. Identifique el tipo de estrategia de almacenamiento en caché adecuada que mejor se adapte a su patrón de acceso.
   - Estrategias de almacenamiento en caché
   - Soluciones de almacenamiento en caché de AWS
3. Siga las prácticas recomendadas del almacenamiento en caché para su almacén de datos.
4. Configure una estrategia de invalidación de caché, como un tiempo de vida (TTL), para todos los datos que equilibre la actualización de los datos y reduzca la presión sobre el almacén de datos de backend.
5. Habilite características como reintentos de conexión automáticos, retroceso exponencial, tiempos de espera del lado del cliente y agrupación de conexiones en el cliente, si están disponibles, ya que pueden mejorar el rendimiento y la fiabilidad.
   - Prácticas recomendadas: clientes de Redis y Amazon ElastiCache (Redis OSS)
6. Supervise la tasa de aciertos de la caché con un objetivo del 80 % o superior. Los valores más bajos pueden indicar un tamaño de caché insuficiente o un patrón de acceso que no se beneficia del almacenamiento en caché.
   - ¿Qué métricas debo supervisar?
   - Best practices for monitoring Redis workloads on Amazon ElastiCache
   - Monitoring best practices with Amazon ElastiCache (Redis OSS) using Amazon CloudWatch
7. Implemente la replicación de datos para descargar las lecturas en varias instancias y mejorar el rendimiento y la disponibilidad de la lectura de datos.

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- Amazon ElastiCache Learning Path
- Design for success with Amazon ElastiCache best practices

Ejemplos relacionados:

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

Redes y entrega de contenido

RENDIMIENTO 4. ¿Cómo selecciona y configura los recursos de red en su carga de trabajo?

La solución de base de datos más eficaz para un sistema varía según los requisitos de disponibilidad, constancia, tolerancia de partición, latencia, durabilidad, escalabilidad y capacidad de consulta. Muchos sistemas utilizan diferentes soluciones de bases de datos para varios subsistemas y activan diferentes características para mejorar el rendimiento. Seleccionar la solución de base de datos y las características incorrectas para un sistema puede conducir a una menor eficiencia de rendimiento.

Prácticas recomendadas

- PERF04-BP01 Comprender cómo afectan las redes al rendimiento
- PERF04-BP02 Evaluar las características de las redes disponibles
- PERF04-BP03 Elegir la conectividad o VPN dedicadas adecuadas para la carga de trabajo
- PERF04-BP04 Utilizar el equilibrio de carga para distribuir el tráfico entre varios recursos

- PERF04-BP05 Elegir los protocolos de red para mejorar el rendimiento
- PERF04-BP06 Elegir la ubicación de la carga de trabajo en función de los requisitos de la red
- PERF04-BP07 Optimizar la configuración de red según las métricas

PERF04-BP01 Comprender cómo afectan las redes al rendimiento

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

Patrones comunes de uso no recomendados:

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

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

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

Guía para la implementación

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

Disponga de una lista documentada de los requisitos de redes de la carga de trabajo, incluida la latencia, el tamaño de los paquetes, las reglas de enrutamiento, los protocolos y los patrones

de tráfico que admiten. Examine las soluciones de red disponibles e identifique qué servicio se ajusta a las características de red de su carga de trabajo. Las redes basadas en la nube se pueden reconstruir rápidamente, de modo que hacer evolucionar su arquitectura de red con el tiempo resulta necesario para mantener la eficiencia del rendimiento.

Pasos para la aplicación:

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

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

Características	Herramientas y métricas
Características de rastreo	• AWS X-Ray • VPC Reachability Analyzer • Network Access Analyzer • Amazon Inspector • Amazon CloudWatch RUM

4. Compare y pruebe el rendimiento de la red:
   a. Compare el rendimiento de la red ya que algunos factores pueden afectar al rendimiento de red de Amazon EC2 cuando las instancias están en la misma VPC. Mida el ancho de banda de la red entre las instancias Linux de Amazon EC2 en la misma VPC.
   b. Realice pruebas de carga para experimentar con soluciones y opciones de redes.

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures

- Optimizing Network Performance for Amazon EC2 Instances
- Improve Global Network Performance for Applications
- EC2 Instances and Performance Optimization Best Practices
- Optimizing Network Performance for Amazon EC2 Instances
- Networking best practices and tips with the Well-Architected Framework (Prácticas recomendadas y consejos para la creación de redes con Well-Architected Framework)
- AWS networking best practices in large-scale migrations

Ejemplos relacionados:

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

PERF04-BP02 Evaluar las características de las redes disponibles

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

Patrones comunes de uso no recomendados:

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

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

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

Guía para la implementación

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

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

Pasos para la implementación

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

- Para los cuellos de botella de rendimiento que identifique, revise las opciones de configuración de sus soluciones para identificar las oportunidades de mejora del rendimiento. Eche un vistazo a las siguientes opciones y características de red clave:

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

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

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

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

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures
- Optimizing Network Performance for Amazon EC2 Instances
- AWS Global Accelerator

Ejemplos relacionados:

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

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

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

Patrones comunes de uso no recomendados:

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

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

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

Guía para la implementación

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

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

Pasos para la implementación

1. Calcule los requisitos de ancho de banda y de latencia de sus aplicaciones actuales.
   a. En el caso de cargas de trabajo existentes que se trasladan a AWS, utilice los datos de sus sistemas internos de supervisión de red.
   b. En el caso de cargas de trabajo nuevas o existentes para las que no disponga de datos de supervisión, consulte con los propietarios del producto para determinar las métricas de rendimiento adecuadas y ofrecer una buena experiencia de usuario.
2. Seleccione una conexión dedicada o VPN como opción de conectividad. En función de todos los requisitos de la carga de trabajo (necesidades de cifrado, ancho de banda y tráfico), puede elegir

AWS Direct Connect o Site-to-Site VPN (o ambas). El siguiente diagrama puede ayudarle a elegir el tipo de conexión adecuado.

a. AWS Direct Connect ofrece conectividad dedicada al entorno de AWS, desde 50 Mbps hasta 100 Gbps, mediante conexiones dedicadas o conexiones alojadas. Esto le ofrece un ancho de banda aprovisionado y una latencia administrada y controlada, a fin de que su carga de trabajo pueda conectarse de manera eficiente a otros entornos. Mediante el uso de socios de AWS Direct Connect, puede disponer de conectividad de extremo a extremo desde varios entornos, lo que proporciona una red ampliada con un rendimiento coherente. AWS ofrece un ancho de banda de conexión directa escalable mediante 100 Gbps nativos, un grupo de agregación de enlaces (LAG) o varias rutas de igual coste (ECMP) con BGP.

b. La AWS Site-to-Site VPN proporciona un servicio de VPN administrado compatible con la seguridad del protocolo de Internet (IPsec). Cuando se crea una conexión VPN, cada conexión VPN incluye dos túneles para ofrecer una alta disponibilidad.

3. Siga la documentación de AWS para elegir la opción de conectividad adecuada:

   a. Si decide usar Direct Connect, seleccione el ancho de banda adecuado para su conectividad.

   b. Si utiliza una AWS Site-to-Site VPN a través de numerosas ubicaciones para conectarse a una Región de AWS, use una conexión de Site-to-Site VPN acelerada para tener la oportunidad de mejorar el rendimiento de la red.

   c. Si el diseño de su red consiste en una conexión VPN IPsec a través de AWS Direct Connect, considere usar una VPN con IP privada para mejorar la seguridad y lograr la segmentación. La VPN con IP privada de AWS Site-to-Site se despliega sobre la interfaz virtual de tránsito (VIF).

   d. AWS Direct Connect SiteLink permite crear conexiones redundantes y de baja latencia entre sus centros de datos de todo el mundo mediante el envío de datos a través de la ruta más corta entre las ubicaciones de AWS Direct Connect, sin pasar por las Regiones de AWS.

4. Valide la configuración de la conectividad antes del despliegue en producción. Lleve a cabo pruebas de seguridad y rendimiento para asegurarse de que cumple los requisitos de ancho de banda, fiabilidad, latencia y cumplimiento.

5. Supervise periódicamente el rendimiento y el uso de la conectividad y optimícelo si es necesario.

Diagrama de flujo de rendimiento determinístico

Recursos

Documentos relacionados:

- Network Load Balancer
- Productos de redes con AWS
- AWS Transit Gateway
- Transitioning to latency-based Routing in Amazon Route 53
- Puntos de conexión de VPC
- Site-to-Site VPN
- Building a Scalable and Secure Multi-VPC AWS Network Infrastructure
- Direct Connect
- Client VPN (Cliente VPN)

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures
- Optimizing Network Performance for Amazon EC2 Instances
- AWS Global Accelerator
- Direct Connect
- AWS Transit Gateway Connect
- VPN Solutions (Soluciones de VPN)
- Security with VPN Solutions (Seguridad con soluciones de VPN)

Ejemplos relacionados:

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

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

Distribuya el tráfico entre varios recursos o servicios para que su carga de trabajo aproveche la elasticidad que ofrece la nube. También puede utilizar el equilibrio de carga para descargar la

terminación del cifrado con el objetivo de mejorar el rendimiento, la fiabilidad y administrar y enrutar el tráfico de manera eficaz.

Patrones comunes de uso no recomendados:

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

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

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

Guía para la implementación

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

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

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

Network Load Balancer es el más adecuado para el equilibrio de carga del tráfico de TCP en donde se necesite un rendimiento extremo. Es capaz de gestionar millones de solicitudes por segundo

manteniendo latencias ultrabajas, y está optimizado para manejar patrones de tráfico repentinos y volátiles.

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

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

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

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

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

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

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

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

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

Pasos para la implementación

- Defina sus requisitos de equilibrio de carga, incluidos un volumen, una disponibilidad y una escalabilidad de aplicaciones excelentes.
- Elija el tipo de equilibrador de carga adecuado para su aplicación.
  - Utilice Application Load Balancer para cargas de trabajo HTTP/HTTPS.
  - Utilice Network Load Balancer para cargas de trabajo distintas de HTTP que se ejecuten en TCP o UDP.
  - Use una combinación de ambos (ALB como destino de NLB) si desea utilizar las características de ambos productos. Por ejemplo, puede hacerlo si desea utilizar las IP estáticas de NLB junto con el enrutamiento basado en encabezado HTTP de ALB, o si desea exponer su carga de trabajo HTTP a una AWS PrivateLink.
  - Para obtener una comparación completa de los equilibradores de carga, consulte ELB product comparison(Comparación de productos ELB).
- Utilice la descarga SSL/TLS si es posible.
  - Configure los agentes de escucha HTTPS/TLS con Application Load Balancer y Network Load Balancer integrados con AWS Gestor de certificados.
  - Tenga en cuenta que algunas cargas de trabajo pueden requerir cifrado de extremo a extremo por motivos de conformidad. En este caso, es un requisito permitir el cifrado en los destinos.
  - Para conocer las prácticas recomendadas de seguridad, consulte SEC09-BP02 Aplicar el cifrado en tránsito.
- Seleccione el algoritmo de enrutamiento adecuado (solo ALB).
  - El algoritmo de enrutamiento puede marcar la diferencia en el grado de utilización de sus destinos de backend y, por lo tanto, en su repercusión en el rendimiento. Por ejemplo, ALB proporciona dos opciones para algoritmos de enrutamiento:
  - Solicitudes menos pendientes: utilícelo para lograr una mejor distribución de la carga a sus destinos de backend para los casos en que las solicitudes de la aplicación varíen en complejidad o los destinos varíen en capacidad de procesamiento.

    - Distribución: 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 zonal.
    - Desactive el aislamiento entre zonas (aislamiento zonal) para mejorar la latencia y los dominios de error zonal. Está desactivado de forma predeterminada en NLB y, en ALB, puede desactivarlo por grupo de destino..
    - Active el aislamiento entre zonas para aumentar la disponibilidad y flexibilidad. Está activado de forma predeterminada en ALB y, en NLB, puede activarlo por grupo de destino..
- Active la conexión persistente HTTP para sus cargas de trabajo HTTP (solo ALB). Con esta característica, el equilibrador de carga puede reutilizar las conexiones de backend hasta que expire el tiempo de espera activo, lo que mejora el tiempo de solicitud y respuesta HTTP, además de reducir la utilización de recursos en los destinos de backend. Para obtener información detallada sobre cómo hacer esto para Apache y Nginx, consulte ¿Cuál es la configuración óptima para utilizar Apache o NGINX como servidor backend para ELB?
- Active la supervisión de su equilibrador de carga.
    - Active los registros de acceso para su Application Load Balancer y Network Load Balancer.
    - Los principales campos a tener en cuenta para ALB son `request_processing_time`, `request_processing_time`y `response_processing_time`.
    - Los principales campos a tener en cuenta para NLB son `connection_time` y `tls_handshake_time`.
    - Esté preparado para consultar los registros cuando los necesite. Puede utilizar Amazon Athena para consultar tanto los registros de ALB y los registros de NLB.
    - Cree alarmas para las métricas relacionadas con el rendimiento, como `TargetResponseTime` para ALB..

Recursos

Documentos relacionados:

- ELB product comparison(Comparación de productos ELB)
- Infraestructura global de AWS
- Improving Performance and Reducing Cost Using Availability Zone Affinity (Mejora del rendimiento y reducción de costes mediante la afinidad de zonas de disponibilidad)
- Step by step for Log Analysis with Amazon Athena

- Querying Application Load Balancer logs
- Monitor your Application Load Balancers
- Monitor your Network Load Balancer
- Use Elastic Load Balancing to distribute traffic across the instances in your Auto Scaling group

Vídeos relacionados:

- AWS re:Invent 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:Inforce 2022 - How to use Elastic Load Balancing to enhance your security posture at scale
- AWS re:Invent 2019: Get the most from Elastic Load Balancing for different workloads

Ejemplos relacionados:

- CDK and CloudFormation samples for Log Analysis with Amazon Athena

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

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

Existe una relación entre la latencia y el ancho de banda para lograr el rendimiento. Si la transferencia de archivos utiliza el protocolo de control de transmisión (TCP), las latencias más altas probablemente reducirán el rendimiento general. Existen enfoques para solucionar esto con el ajuste de TCP y protocolos de transferencia optimizados, pero una solución es utilizar el protocolo de datagramas de usuario (UDP).

Patrones comunes de uso no recomendados:

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

Beneficios de establecer esta práctica recomendada: Verificar que se utiliza un protocolo adecuado para la comunicación entre los usuarios y los componentes de la carga de trabajo ayuda a mejorar la experiencia general del usuario para sus aplicaciones. Por ejemplo, UDP sin conexión permite

una alta velocidad, pero no ofrece retransmisión ni alta fiabilidad. TCP es un protocolo con todas las características, pero requiere una mayor sobrecarga para procesar los paquetes.

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

Guía para la implementación

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

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

Cuándo considerar el uso de TCP

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

Cuándo considerar el uso de UDP

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

añade una pequeña cantidad de sobrecarga a los paquetes UDP, ya que incluye campos adicionales para indicar los parámetros de seguridad y detectar manipulaciones.

Cuándo considerar el uso de SRD

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

Pasos para la implementación

1. Utilice la AWS Global Accelerator y AWS Transfer Family para mejorar el rendimiento de sus aplicaciones de transferencia de archivos en línea. El servicio AWS Global Accelerator le ayuda a conseguir una latencia menor entre sus dispositivos cliente y su carga de trabajo en AWS. Con AWS Transfer Family, puede utilizar protocolos basados en TCP como el protocolo de transferencia de archivos de shell seguro (SFTP) y el protocolo de transferencia de archivos sobre SSL (FTPS) para escalar y administrar de forma segura las transferencias de archivos a los servicios de almacenamiento de AWS.
2. Utilice la latencia de la red para determinar si TCP es adecuado para la comunicación entre los componentes de la carga de trabajo. Si la latencia de la red entre la aplicación cliente y el servidor es alta, la comunicación TCP de tres vías puede tardar un tiempo, lo que afectará a la capacidad de respuesta de la aplicación. Para medir la latencia de la red pueden utilizarse métricas, como el tiempo hasta el primer byte (TTFB) y el tiempo de ida y vuelta (RTT). Si su carga de trabajo ofrece contenido dinámico a los usuarios, considere la posibilidad de utilizar Amazon CloudFront, que establece una conexión persistente con cada origen para el contenido dinámico para eliminar el tiempo de configuración de la conexión que, de otro modo, ralentizaría cada solicitud del cliente.
3. El uso de TLS con TCP o UDP puede aumentar la latencia y reducir el rendimiento de la carga de trabajo debido al impacto del cifrado y el descifrado. Para este tipo de cargas de trabajo, considere la posibilidad de descargar SSL/TLS en Elastic Load Balancing para mejorar el rendimiento de la carga de trabajo al permitir que el equilibrador de carga gestione el proceso de cifrado y descifrado SSL/TLS, en lugar de que lo hagan las instancias de backend. Esto puede ayudar a reducir la utilización de la CPU en las instancias backend, lo que puede mejorar el rendimiento y aumentar la capacidad.

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

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures
- Optimizing Network Performance for Amazon EC2 Instances

Ejemplos relacionados:

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

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

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

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 Outpostso AWS Wavelength . La selección de esta ubicación influye en la latencia y el rendimiento de la red desde una ubicación de usuario. Los de periferia, como Amazon CloudFront y AWS Global Accelerator , también se pueden utilizar para mejorar el rendimiento de la red al almacenar contenido en caché en ubicaciones periféricas o proporcionar a los usuarios una ruta óptima a la carga de trabajo a través de la red global de AWS.

Amazon EC2 ofrece grupos de colocación para la creación de redes. Un grupo de registro es una agrupación lógica de instancias para reducir la latencia. El uso de grupos de colocación con tipos

de instancias compatibles y un Elastic Network Adapter (ENA) permite que las cargas de trabajo participen en una red de 25 Gbps de baja latencia y fluctuación reducida. Se recomiendan grupos de colocación para cargas de trabajo que aprovechan la baja latencia de red, el alto rendimiento de red o ambos.

Los servicios sensibles a la latencia se prestan en ubicaciones periféricas mediante una red global de AWS, 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 periféricos para reducir la latencia y permitir el almacenamiento en caché del contenido. Configure correctamente el control de caché para DNS y HTTP/HTTPS a fin de obtener el mayor beneficio de estos enfoques.

Pasos para la implementación

- Recoja información sobre el tráfico IP que entra y sale de las interfaces de red.
  - Registro del tráfico de IP con registros de flujo de la VPC
  - Cómo se conserva la dirección IP del cliente en AWS Global Accelerator
- Analice los patrones de acceso de la red en su carga de trabajo para identificar cómo utilizan los usuarios su aplicación.
  - Use herramientas de monitorización como Amazon CloudWatch y AWS CloudTrailpara recopilar datos sobre las actividades de la red.
  - Analice los datos para identificar el patrón de acceso de la red.
- Seleccione regiones para el despliegue de la carga de trabajo en función de los siguientes elementos clave:
  - Dónde se encuentran sus datos: en el caso de las aplicaciones con gran cantidad de datos (como macrodatos y machine learning), el código de la aplicación debe ejecutarse lo más cerca posible de los datos.
  - Dónde se encuentran sus 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 restricciones como el coste y el cumplimiento como se explica en Qué tener en cuenta al seleccionar una región para las cargas de trabajo.
- Utilice Zonas locales de AWS para ejecutar cargas de trabajo como la renderización de vídeo. Las zonas locales le permiten beneficiarse de tener recursos de computación y almacenamiento más cerca de los usuarios finales.
- Utilice AWS Outposts para cargas de trabajo que deban seguir siendo locales y en las que desee que esa carga de trabajo se ejecute sin problemas con el resto de sus demás cargas de trabajo en AWS.
- Aplicaciones como la transmisión de vídeo en directo de alta resolución, audio de alta fidelidad y realidad aumentada/realidad virtual (RA/RV) requieren una latencia ultrabaja para dispositivos 5G. Para este tipo de aplicaciones, considere AWS Wavelength. AWS Wavelength integra los servicios de computación y almacenamiento de AWS en las redes 5G, proporcionando una infraestructura de computación periférica móvil para desarrollar, desplegar y escalar aplicaciones de ultrabaja latencia.
- Utilice almacenamiento en caché local o Soluciones de almacenamiento en caché de AWS para los recursos de uso frecuente con el fin de mejorar el rendimiento, reducir el movimiento de datos y disminuir el impacto medioambiental.

Servicio	Cuándo usar
Amazon CloudFront	Se usa para almacenar en caché el contenido estático como imágenes, scripts y vídeos, así como el contenido dinámico como respuestas de API y aplicaciones web.
Amazon ElastiCache	Se usa para almacenar en caché el contenido de las aplicaciones web.
DynamoDB Accelerator	Se usa para añadir aceleración en memoria a sus tablas de DynamoDB.

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

Servicio	Cuándo 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.
Funciones de Amazon CloudFront	Se usan para casos de uso sencillos como las manipulaciones de solicitudes o respuesta s HTTP(s) que pueden iniciarse mediante funciones de corta duración.
AWS IoT Greengrass	Se usa para ejecutar la computación local, la mensajería y el almacenamiento en caché de datos para los dispositivos conectados.

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

Recursos

Prácticas recomendadas relacionadas:

- COST07-BP02 Implementar regiones según los costes
- COST08-BP03 Implementar servicios para reducir los costes de transferencia de datos

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

Documentos relacionados:

- Infraestructura global de AWS
- AWS Local Zones and AWS Outposts, choosing the right technology for your edge workload
- Grupos de ubicación
- Zonas locales de AWS
- AWS Outposts
- AWS Wavelength
- Amazon CloudFront
- AWS Global Accelerator
- AWS Direct Connect
- AWS Site-to-Site VPN
- Amazon Route 53

Vídeos relacionados:

- AWS Local Zones Explainer Video
- AWS Outposts: Overview and How it Works
- 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 Workshop
- Handling Rewrites and Redirects using Edge Functions (Gestión de reescrituras y redireccionamientos mediante funciones periféricas)

PERF04-BP07 Optimizar la configuración de red según las métricas

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

Patrones comunes de uso no recomendados:

- Supone que todos los problemas de rendimiento están relacionados con las aplicaciones.
- Solo hace pruebas del rendimiento de la red desde una ubicación cercana al punto de implementación de la carga de trabajo.
- Se utilizan configuraciones predeterminadas para todos los servicios de red.
- Se sobreaprovisionan los recursos de red para proporcionar capacidad suficiente.

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

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

Guía para la implementación

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

Pasos para la implementación

- Identifique las métricas clave de rendimiento, como la latencia o la pérdida de paquetes, que desee recopilar. AWS proporciona varias herramientas que pueden ayudarle a recopilar estas métricas.

Mediante las siguientes herramientas, puede inspeccionar más a fondo la información sobre el uso del tráfico, el acceso a la red y los registros:

Herramienta de AWS	Dónde se usa
Amazon VPC IP Address Manager.	Utilice IPAM para planificar, seguir y supervisar las direcciones IP para sus cargas de trabajo de AWS y locales. Esta es una práctica recomendada para optimizar el uso y la asignación de direcciones IP.
Registros de flujo de VPC	Utilice los registros de flujo de la VPC para capturar información detallada sobre el tráfico hacia y desde las interfaces de red en sus VPC. Con los registros de flujo de la VPC, puede diagnosticar reglas de grupos de seguridad excesivamente restrictivas o permisivas y determinar la dirección del tráfico hacia y desde las interfaces de red.
Registros de flujo de AWS Transit Gateway	Utilice los registros de AWS Transit Gateway flujo para recoger información sobre el tráfico IP que entra y sale de sus puertas de enlace de tránsito.
Registro de consultas de DNS	Registre información sobre las consultas de DNS públicas o privadas que recibe Route 53. Con los registros de DNS, puede optimizar las configuraciones de DNS al conocer el dominio o subdominio que se solicitó o las ubicaciones periféricas de Route 53 que respondieron a las consultas de DNS.

Herramienta de AWS	Dónde se usa
Reachability Analyzer	Reachability Analyzer le ayuda a analizar y depurar la accesibilidad de la red. Reachability Analyzer es una herramienta de análisis de configuración que le permite realizar pruebas de conectividad entre un recurso de origen y un recurso de destino en sus VPC. Esta herramienta le ayuda a verificar que la configuración de su red coincida con la conectividad prevista.
Network Access Analyzer	Network Access Analyzer le ayuda a comprender el acceso a la red a sus recursos. Puede utilizar el Network Access Analyzer para especificar los requisitos de acceso a la red e identificar posibles rutas de red que no cumplan los requisitos especificados. Al optimizar su configuración de red correspondiente, puede comprender y verificar el estado de su red y demostrar si su red en AWS cumple con sus requisitos de conformidad.
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 NAT de VPC, AWS Transit Gateway, túneles de VPN, AWS Network Firewall, Elastic Load Balancing y AWS Direct Connect. La supervisión continua de las métricas es una práctica recomendada para observar y comprender el estado y el uso de su red, lo que le ayuda a optimizar la configuración de la red basándose en sus observaciones.

Herramienta de 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. Network Manager proporciona una latencia de red agregada entre las Regiones de AWS y las zonas de disponibilidad y dentro de cada zona de disponibilidad, lo que le permite comprender mejor la relación entre el rendimiento de las aplicaciones y el rendimiento de la red de AWS subyacente.
Amazon CloudWatch RUM	Use Amazon CloudWatch RUM para recopilar las métricas que le proporcionan información que le ayuda a identificar, comprender y mejorar la experiencia del usuario.

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

Recursos

Documentos relacionados:

- Registros de flujo de VPC
- Habilite los registros de consultas de DNS públicos.
- What is IPAM?
- What is Reachability Analyzer?
- What is Network Access Analyzer?
- Métricas de CloudWatch para sus VPC

- Optimize performance and reduce costs for network analytics with VPC Flow Logs in Apache Parquet format (Optimice el rendimiento y reduzca los costes de los análisis de red con los registros de flujo de VPC en formato Apache Parquet)
- Monitoring your global and core networks with Amazon CloudWatch metrics
- Continuously monitor network traffic and resources (Supervisar de forma continua el tráfico y los recursos de red)

Vídeos relacionados:

- Networking best practices and tips with the AWS Well-Architected Framework
- Monitoring and troubleshooting network traffic (Supervisión y solución de problemas del tráfico de red)

Ejemplos relacionados:

- AWS Networking Workshops (Talleres de red de AWS)
- AWS Network Monitoring

Proceso y cultura

RENDIMIENTO 5. ¿Cómo contribuyen sus prácticas y cultura de la organización a la eficiencia del rendimiento en su carga de trabajo?

Al diseñar cargas de trabajo, hay principios y prácticas que puede adoptar para ayudarle a ejecutar mejor cargas de trabajo en la nube eficientes y de alto rendimiento. Para adoptar una cultura que fomente la eficiencia del rendimiento de las cargas de trabajo en la nube, tenga en cuenta estos principios y prácticas clave:

Prácticas recomendadas

- PERF05-BP01 Establecer indicadores clave de rendimiento (KPI) para medir el estado y el rendimiento de la carga de trabajo
- PERF05-BP02 Utilizar soluciones de supervisión para saber en qué áreas es más crítico el rendimiento
- PERF05-BP03 Definir un proceso para mejorar el rendimiento de la carga de trabajo
- PERF05-BP04 Realizar pruebas de la carga de trabajo

- PERF05-BP05 Utilizar la automatización para solucionar de forma proactiva los problemas relacionados con el rendimiento
- PERF05-BP06 Mantener la carga de trabajo y los servicios actualizados
- PERF05-BP07 Revisar las métricas a intervalos regulares

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

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

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 están publicando y compartiendo como datos de métricas estándar.
- No tiene definido un KPI cuantitativo y medible.
- Los KPI no se corresponden con los objetivos o estrategias empresariales.

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, y del modo en que estos factores se combinan para producir resultados empresariales. Por ejemplo, en una carga de trabajo de un sitio web, el tiempo de carga de la página podría usarse como indicativo del rendimiento general. Esta métrica sería uno de los múltiples puntos de datos que miden la experiencia del usuario. Además de identificar los umbrales de los tiempos de carga de la página, debería documentar el resultado previsto o el riesgo empresarial si no se cumple el ideal de rendimiento. Si una página tarda en cargarse, los usuarios

finales se ven directamente afectados, se reduce su valoración de la experiencia y se pueden perder clientes. Cuando defina los umbrales de KPI, combine tanto las referencias del sector como las expectativas de los usuarios finales. Por ejemplo, si la referencia sectorial actual es que una página web se cargue en dos segundos, pero los usuarios esperan que tarde solamente un segundo, debería tener en cuenta estos dos puntos de datos al establecer el KPI.

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

Pasos para la implementación

1. Identifique y documente las principales partes interesadas de la empresa.
2. Trabaje con estas partes interesadas para definir y documentar los objetivos de su carga de trabajo.
3. Revise las prácticas sectoriales recomendadas para identificar los KPI relevantes que se ajustan a los objetivos de su carga de trabajo.
4. Utilice las prácticas sectoriales recomendadas 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.
5. Identifique y documente el riesgo y el impacto del incumplimiento de los KPI.
6. Identifique y documente las métricas que pueden ayudarle a establecer los KPI.
7. Emplee herramientas de supervisión como Amazon CloudWatch o bien AWS Config para recopilar métricas y medir los KPI.
8. Utilice paneles de control para visualizar los KPI y comunicarlos a las partes interesadas.
9. Revise y analice periódicamente las métricas para identificar las áreas de la carga de trabajo que deben mejorarse.
10. Revise los KPI cuando cambien los objetivos empresariales o el rendimiento de la carga de trabajo.

Recursos

Documentos relacionados:

- CloudWatch documentation
- Monitoring, Logging, and Performance AWS Partners
- Documentación de X-Ray
- Uso de paneles de Amazon CloudWatch
- Quick KPIs

Vídeos relacionados:

- AWS re:Invent 2019: Scaling up to your first 10 million users
- Cut through the chaos: Gain operational visibility and insight
- Diseñe un plan de monitoreo

Ejemplos relacionados:

- Creating a dashboard with Quick

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

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

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.

Beneficios 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 deficientemente fragmentados y particionados. Identifique las áreas restringidas de la carga de trabajo mediante pruebas de carga o supervisión.

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

Pasos para la implementación

1. 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 realizar 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.

Servicio	Dónde se usa
Supervisión mejorada de Amazon RDS	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.

2. Lleve a cabo pruebas para generar métricas, identificar patrones de tráfico, cuellos de botella y áreas críticas de rendimiento. Estos son algunos ejemplos de cómo se realizan las pruebas:
    - Configure «canaries» 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.
    - Utilice la Pruebas de carga distribuidas en AWS para generar picos de tráfico o probar la carga de trabajo con la tasa de crecimiento prevista.
3. 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.
4. Experimente con las mejoras de rendimiento y mida los cambios con datos. Como ejemplo, puede usar CloudWatch Evidently para probar las nuevas mejoras y el impacto en el rendimiento de la carga de trabajo.

Recursos

Documentos relacionados:

- Amazon Builders' Library
- Documentación de X-Ray
- Amazon CloudWatch RUM
- Amazon DevOps Guru

Vídeos relacionados:

- The Amazon Builders’ Library: 25 years of Amazon operational excellence

- Visual Monitoring of Applications with Amazon CloudWatch Synthetics

Ejemplos relacionados:

- Medición del tiempo de carga de la página con Amazon CloudWatch Synthetics
- Cliente web de Amazon CloudWatch RUM
- SDK de X-Ray para Node.js
- SDK de X-Ray para Python
- SDK de X-Ray para Java
- SDK de X-Ray para .Net
- SDK de X-Ray para Ruby
- Daemon de X-Ray
- Pruebas de carga distribuidas en AWS

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

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

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 de métricas.

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

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

Guía para la implementación

El rendimiento de su carga de trabajo tiene algunas limitaciones clave. Documéntelos para que sepa qué tipos de innovación pueden mejorar el rendimiento de su carga de trabajo. Utilice esta

información cuando conozca nuevos servicios o tecnologías a medida que estén disponibles para identificar formas de mitigar las limitaciones o cuellos de botella.

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

Pasos para la implementación

- Identifique los KPI de rendimiento de la carga de trabajo tal y como se describe en PERF05-BP01 Establecer indicadores clave de rendimiento (KPI) para medir el estado y el rendimiento de la carga de trabajo para establecer los puntos de referencia de su carga de trabajo.
- Utilice las herramientas de observabilidad de AWS para recopilar métricas de rendimiento y medir los KPI.
- Realice un análisis exhaustivo para identificar las áreas de la carga de trabajo (como la configuración y el código de la aplicación) que tienen un rendimiento inferior, tal y como se describe en PERF05-BP02 Utilizar soluciones de supervisión para saber en qué áreas es más crítico el rendimiento.
- Utilice sus herramientas de análisis y rendimiento para identificar la estrategia de optimización del rendimiento.
- Utilice entornos aislados o de preproducción para validar la eficacia de la estrategia.
- Implemente los cambios en la producción y supervise continuamente el rendimiento de la carga de trabajo.
- Documente las mejoras y comuníqueselo a las partes interesadas.

Recursos

Documentos relacionados:

- Blog de AWS
- Novedades de AWS

Vídeos relacionados:

- Canal de YouTube de eventos de AWS
- Canal de YouTube de AWS Online Tech Talks

- Canal de YouTube de Amazon Web Services

Ejemplos relacionados:

- GitHub de AWS
- AWS Skill Builder

PERF05-BP04 Realizar pruebas de la carga de trabajo

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

Patrones comunes de uso no recomendados:

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

Beneficios de establecer esta práctica recomendada: La medición del rendimiento en una prueba de carga le mostrará dónde se verá afectado 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 es un proceso que permite medir el rendimiento de la carga de trabajo en la nube bajo condiciones realistas, con la carga de usuarios esperada. Este proceso implica el aprovisionamiento de un entorno de nube similar al de producción, el uso de herramientas de pruebas de carga para generar la carga y el análisis de métricas para evaluar la capacidad de la carga de trabajo a la hora de gestionar una carga realista. Las pruebas de carga deben ejecutarse

con versiones sintéticas o saneadas de los datos de producción (debe eliminarse la información confidencial o de identificación). Realice automáticamente pruebas de carga en la canalización de entrega y compare los resultados con los KPI y los umbrales predefinidos. Este proceso le ayudará a seguir alcanzando el rendimiento requerido.

Pasos para la implementación

- Configure el entorno de prueba con arreglo a su entorno de producción. Puede usar los servicios de AWS para ejecutar entornos a escala de producción y poner a prueba su arquitectura.
- Elija y configure la herramienta de prueba de carga que se ajuste a su carga de trabajo.
- Defina los escenarios y los parámetros de las pruebas de carga (como la duración de la prueba y el número de usuarios).
- Cree escenarios de prueba a escala. Utilice la Nube de AWS para probar la carga de trabajo y detectar las áreas en las que el escalamiento no se realiza correctamente o no se produce de forma lineal. Por ejemplo, utilice Spot Instances para generar cargas a bajo costo y descubrir obstáculos antes que se experimenten en la producción
- Supervise y registre las métricas de funcionamiento (como el rendimiento y el tiempo de respuesta). Amazon CloudWatch puede recopilar métricas en los diferentes recursos de la arquitectura. También puede recopilar y publicar métricas del cliente para negocios de superficie o métricas derivadas.
- Analice los resultados para identificar los cuellos de botella del rendimiento y las áreas en las que se pueden mejorar.
- Documente el proceso y los resultados de las pruebas de carga, y cree los informes pertinentes.

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- Solving with AWS Solutions: Distributed Load Testing

- Optimize applications through Amazon CloudWatch RUM
- Demostración de Amazon CloudWatch Synthetics

Ejemplos relacionados:

- Pruebas de carga distribuidas en AWS

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

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

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 concentrarse 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 alarmas críticas.

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

Guía para la implementación

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

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

Pasos para la implementación

- Identifique y estudie si el problema de rendimiento puede solucionarse automáticamente. Use soluciones de supervisión de AWS, como Amazon CloudWatch o AWS X-Ray, para ayudarle a comprender mejor la causa raíz del problema.
- Cree un plan y un proceso de corrección paso a paso que pueda utilizar para solucionar el problema automáticamente.
- Configure el activador que va a 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.
- 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.
- Pruebe el proceso de corrección automatizado en un entorno de preproducción.
- Una vez realizadas las pruebas, implemente el proceso de corrección en el entorno de producción y supervíselo continuamente para identificar posibles áreas de mejora.

Recursos

Documentos relacionados:

- CloudWatch Documentation
- Monitoring, Logging, and Performance AWS Partner Network Partners
- Documentación de X-Ray
- Using Alarms and Alarm Actions in CloudWatch

Vídeos relacionados:

- Intelligently automating cloud operations
- Setting up controls at scale in your AWS environment
- Automating patch management and compliance using AWS
- How Amazon uses better metrics for improved website performance

Ejemplos relacionados:

- CloudWatch Logs Customize Alarms

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

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

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 su 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, realice pruebas de rendimiento para medir el impacto que tienen en el rendimiento de la carga de trabajo.

Pasos para la implementación

- Inventariar el software y la arquitectura de su carga de trabajo e identificar los componentes que deben actualizarse.
- 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 los productos que coinciden con su componente de carga de trabajo. Puede suscribirse a la fuente RSS o administrar sus suscripciones de correo electrónico.

- Establezca un calendario para evaluar nuevos servicios y características con su carga de trabajo.
  - Puede usar el 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 ejecutando el software y las configuraciones requeridas por su política de software así como las instancias que deben actualizarse.
- 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 su carga de trabajo.
- Utilice la automatización del proceso de actualización para reducir el nivel de esfuerzo para desplegar nuevas funciones y limitar los errores causados por los procesos manuales.
  - Puede usar Entrega e integración continuas (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 con AWS Systems Manager Maintenance Windows.
- 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

Vídeos relacionados:

- Canal de YouTube de eventos de AWS
- Canal de YouTube de AWS Online Tech Talks
- Canal de YouTube de Amazon Web Services

Ejemplos relacionados:

- Well-Architected Labs - Inventory and Patch Management (Laboratorios de Well-Architected: administración de inventario y parches)
- Laboratorio: AWS Systems Manager

PERF05-BP07 Revisar las métricas a intervalos regulares

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

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 no puede accionar un sistema de automatización.

Beneficios de establecer esta práctica recomendada: revisar 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 parámetros fueron útiles para abordar el problema y qué parámetros podrían haber ayudado a los que no se están controlando actualmente. Utilice este método para mejorar la calidad de las métricas que recopila, de modo que pueda prevenir o resolver incidentes en el futuro con mayor rapidez.

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

Pasos para la implementación

1. Defina las métricas de rendimiento críticas para comprobar que estén alineadas con el objetivo de su carga de trabajo.
2. Establezca una base de referencia y el valor que desee para cada métrica.
3. Establezca una cadencia (como semanal o mensual) para revisar las métricas críticas.
4. 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.
5. Lleve a cabo un análisis exhaustivo de la causa raíz de los problemas identificados para conocer qué los provoca.
6. Documente sus resultados y utilice estrategias para hacer frente a los problemas y cuellos de botella identificados.
7. Evalúe y mejore continuamente el proceso de revisión de las métricas.

Recursos

Documentos relacionados:

- CloudWatch Documentation
- Recopilación de métricas y registros de instancias Amazon EC2 y en los servidores en las instalaciones con el agente de CloudWatch
- Monitoring, Logging, and Performance AWS Partner Network Partners
- Documentación de X-Ray

Vídeos relacionados:

- Setting up controls at scale in your AWS environment
- How Amazon uses better metrics for improved website performance

Ejemplos relacionados:

- Creating a dashboard with Quick
- Level 100: Monitoring with CloudWatch Dashboards

Optimización de costes

El pilar de optimización de costes incluye la capacidad de ejecutar sistemas para ofrecer valor empresarial al precio más bajo posible. Encontrará recomendaciones de implementación en el documento técnico Pilar de optimización de costes.

Áreas de prácticas recomendadas

- Práctica de administración financiera en la nube
- Conocimiento del gasto y del uso
- Recursos rentables
- Administración de la demanda y suministro de recursos
- Optimización a lo largo del tiempo

Práctica de administración financiera en la nube

Pregunta

- COSTE 1. ¿Cómo implementar la administración financiera en la nube?

COSTE 1. ¿Cómo implementar la administración financiera en la nube?

Implementar la administración financiera en la nube ayuda a las empresas a obtener valor empresarial y éxito financiero al optimizar su coste y uso, y al escalar en AWS.

Prácticas recomendadas

- COST01-BP01 Establecer la responsabilidad de la optimización de costes
- COST01-BP02 Establecer la colaboración entre los departamentos de Finanzas y Tecnología
- COST01-BP03 Establecer presupuestos y previsiones de la nube
- COST01-BP04 Implementar la conciencia de costes en los procesos organizativos
- COST01-BP05 Crear informes y notificar la optimización de costes
- COST01-BP06 Supervisar los costes de forma proactiva
- COST01-BP07 Estar al día sobre las nuevas versiones de los servicios
- COST01-BP08 Crear una cultura de conciencia de costes
- COST01-BP09 Cuantificar el valor empresarial a partir de la optimización de costes

COST01-BP01 Establecer la responsabilidad de la optimización de costes

Cree un equipo (Oficina de negocios en la nube, Centro de excelencia en la nube o equipo de FinOps) que se encargue de establecer y afianzar la concienciación sobre los costes en toda la organización. El responsable de la optimización de costes puede ser una persona o un equipo (requiere representantes de los equipos financieros, tecnológicos y empresariales) que comprenda toda la organización y las finanzas en la nube.

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

Guía para la implementación

Esta es la introducción a un equipo de Oficina de negocios en la nube (CBO) o Centro de excelencia en la nube (CCOE) que se encargue de establecer y afianzar una cultura de concienciación sobre los costes en la computación en la nube. Esta función puede ser una persona existente, un equipo de la organización o un nuevo equipo formado por representantes clave de los departamentos financiero, tecnológico y organizativo de la organización.

Esta función (la persona o el equipo) prioriza y dedica el porcentaje necesario de su tiempo a actividades de administración y optimización de costes. En una organización pequeña, es posible que esta función dedique menos tiempo a ello, si lo comparamos con una función a tiempo completo de una empresa grande.

Esta función (la persona o el equipo) prioriza y dedica el porcentaje necesario de su tiempo a actividades de administración y optimización de costes. En una organización pequeña, es posible que esta función dedique menos tiempo a actividades de administración y optimización de costes en comparación con una función a tiempo completo de una empresa grande.

Esta función debe tener carácter multidisciplinar, es decir, que debe tener experiencia en administración de proyectos, ciencia de datos, análisis financiero y desarrollo de software o infraestructura. Pueden mejorar la eficiencia de la carga de trabajo ejecutando optimizaciones de costes dentro de tres propiedades diferentes:

- Centralizada: a través de equipos designados, como el equipo de FinOps, el equipo de Administración financiera en la nube (CFM), la Oficina de negocios en la nube (CBO) o el Centro de excelencia en la nube (CCoE), los clientes pueden diseñar e implementar mecanismos de gobernanza e impulsar las prácticas recomendadas en toda la empresa.
- Descentralizada: se influye en los equipos tecnológicos para que realicen optimizaciones de costes.

- Híbrida: una combinación de equipos centralizados y descentralizados que pueden trabajar de forma conjunta para ejecutar optimizaciones de costes.

La función se evalúa según su capacidad de ejecutar y alcanzar los objetivos de optimización de costes (por ejemplo, las métricas de eficiencia de las cargas de trabajo).

Debe conseguir el patrocinio de los ejecutivos para esta función, lo cual es un factor clave para el éxito. El patrocinador es considerado el campeón del consumo rentable de la nube y proporciona apoyo al equipo para garantizar que las actividades de optimización de costes se traten según el nivel de prioridad definido por la organización. De lo contrario, se ignorarán las directrices y no se dará prioridad a las oportunidades de ahorro. De forma conjunta, el patrocinador y el equipo garantizan que su organización haga un consumo eficiente de la nube y ofrezca valor empresarial.

Si tiene un plan de asistencia Business, Enterprise-On-Ramp o Enterprise y necesita ayuda para crear este equipo o función, póngase en contacto con los expertos de Administración financiera en la nube (CFM) a través de su equipo de cuentas.

Pasos para la implementación

- Defina los miembros clave: todas las partes pertinentes de la organización deben contribuir y estar interesadas en la administración de costes. En general, los equipos de las organizaciones constan de equipos de finanzas, propietarios de aplicaciones o productos, administración y técnicos (DevOps). Algunos tienen dedicación completa (técnicos y financieros) mientras que otros participan periódicamente, según sea necesario. Las personas o los equipos encargadas de la CFM precisan el siguiente conjunto de habilidades:
  - Desarrollo de software: en el caso de que se creen scripts y automatizaciones.
  - Ingeniería de infraestructuras: para desplegar scripts, automatizar procesos y entender cómo se aprovisionan los servicios o recursos.
  - Perspicacia en las operaciones: la CFM consiste en operar en la nube de forma eficiente, para lo que se mide, supervisa, modifica, planifica y escala el uso eficiente de la nube.
- Establezca objetivos y métricas: Esta función debe proporcionar valor a la organización de distintas maneras. Estos objetivos se definen y evolucionan de forma continua a medida que evoluciona la organización. Estas son las actividades habituales: crear y ejecutar programas educativos sobre optimización de costes en la organización, desarrollar estándares para toda la organización, como la supervisión y la creación de informes de optimización de costes, y establecer objetivos de carga de trabajo sobre la optimización. Esta función también debe informar regularmente a la organización sobre la capacidad de optimizar costes de la organización.

Puede definir indicadores clave de rendimiento (KPI) basados en el valor o el coste. Cuando se definen los KPI, se puede calcular el coste previsto en términos de eficiencia y el resultado empresarial esperado. Los KPI basados en el valor vinculan las métricas de coste y uso a los impulsores del valor empresarial y ayudan a racionalizar los cambios en el gasto de AWS. El primer paso para derivar los KPI basados en el valor es trabajar juntos, entre organizaciones, para seleccionar y acordar un conjunto estándar de KPI.

- Establezca una cadencia regular: el grupo (equipos de finanzas, tecnología y negocios) debe reunirse de manera regular para revisar sus objetivos y métricas. Una cadencia típica implica revisar el estado de la organización, revisar los programas que se ejecutan actualmente y las métricas generales financieras y de optimización. Después, se debe informar sobre las cargas de trabajo clave con mayor detalle.

  Durante estas revisiones periódicas, se puede revisar la eficiencia de la carga de trabajo (coste) y los resultados empresariales. Por ejemplo, un aumento del 20 % en el coste de una carga de trabajo puede coincidir con un mayor uso por parte del cliente. En este caso, este aumento del 20 % de los costes puede interpretarse como una inversión. Estas reuniones de cadencia periódicas pueden ayudar a los equipos a identificar los KPI de valor que proporcionan significado a toda la organización.

Recursos

Documentos relacionados:

- Blog de CCOE de AWS
- Creating Cloud Business Office (Creación de la Oficina de negocios en la nube)
- CCOE - Cloud Center of Excellence

Vídeos relacionados:

- Historia de éxito de CCOE en Vanguard

Ejemplos relacionados:

- Using a Cloud Center of Excellence (CCOE) to Transform the Entire Enterprise (Uso del Centro de excelencia en la nube [CCOE] para transformar toda la empresa)

- Building a CCOE to transform the entire enterprise (Creación de un CCOE para transformar toda la empresa)
- 7 Pitfalls to Avoid When Building CCOE (7 obstáculos que evitar al crear el CCOE)

COST01-BP02 Establecer la colaboración entre los departamentos de Finanzas y Tecnología

Debe implicar a los equipos de finanzas y tecnología en las discusiones sobre costes y uso en todas las etapas del traspaso a la nube. Los equipos deben reunirse y tratar regularmente sobre temas como los objetivos organizativos, el estado actual de los costes y el uso, y las prácticas contables y financieras.

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

Guía para la implementación

Los equipos tecnológicos innovan más rápido en la nube gracias a que los ciclos de aprobación, adquisición y despliegue de la infraestructura son más cortos. Esto puede implicar un ajuste para las organizaciones financieras acostumbradas a tanto a ejecutar procesos que requieren mucho tiempo y consumen muchos recursos para obtener y desplegar capital en entornos de centros de datos y a nivel local, como a asignar los costes solo al aprobar el proyecto.

Desde el punto de vista de la organización financiera y de adquisiciones, el proceso de presupuestos de capital, las solicitudes de capital, las aprobaciones, las adquisiciones y la instalación de la infraestructura física es un proceso que se ha aprendido y estandarizado durante décadas:

- Los equipos de Ingeniería o TI suelen ser los solicitantes
- Varios equipos financieros actúan como aprobadores y compradores
- Los equipos de operaciones montan rack, apilan y entregan la infraestructura lista para usar

Con la adopción de la nube, la adquisición y el consumo de infraestructura dejan de estar a expensas de una cadena de dependencias. En el modelo de la nube, los equipos de tecnología y productos ya no son solo creadores, sino operadores y propietarios de sus productos, responsables de la mayoría de las actividades históricamente asociadas a los equipos de finanzas y operaciones, incluidas las adquisiciones y el despliegue.

Todo lo que se necesita para aprovisionar recursos en la nube es una cuenta de usuario y la serie de permisos adecuada. Esto es también lo que reduce el riesgo de la TI y de las finanzas, lo cual significa que, con unos pocos clics o llamadas a la API, los equipos pueden eliminar los recursos inactivos o innecesarios en la nube. Esto también permite a los equipos de tecnología innovar más rápidamente gracias a la agilidad y la capacidad de poner en marcha experimentos

y luego desmantelarlos. Aunque la naturaleza variable del consumo de la nube puede afectar a la previsibilidad desde el punto de vista de los presupuestos y las previsiones de capital, la nube ofrece a las organizaciones la posibilidad de reducir el coste del exceso de aprovisionamiento así como el coste de oportunidad asociado al subaprovisionamiento conservador.

Establezca la colaboración entre los partes interesadas clave de Finanzas y Tecnología para lograr un entendimiento común de los objetivos organizativos y desarrollar mecanismos para obtener éxito financiero en el modelo de gasto variable de la computación en la nube. Los equipos relevantes de su organización deben estar presentes en las discusiones sobre costes y uso en todas las etapas del traspaso a la nube, incluidos:

- Líderes en finanzas: Los directores financieros, controllers financieros, planificadores financieros, analistas empresariales, responsables de adquisición, de abastecimiento y de cuentas a pagar deben entender el modelo de consumo en la nube, las opciones de compra y el proceso de facturación mensual. El departamento financiero debe asociarse con los equipos de tecnología

para crear y compartir una historia de valor de TI, ayudando al equipo del departamento comercial a comprender cómo el gasto en tecnología está vinculado a los resultados empresariales. De esta manera, los gastos en tecnología no se consideran costes, sino inversiones. Dado que hay diferencias fundamentales entre la nube (por ejemplo, la velocidad del cambio en el uso, los precios del pago por uso, los precios por niveles, los modelos de precios y la información detallada sobre la facturación y el uso) y las operaciones locales, resulta esencial que el equipo de Finanzas comprenda de qué manera puede afectar el uso de la nube a aspectos empresariales como los procesos de adquisición, el seguimiento de incentivos, la asignación de costes y los estados financieros.

- Líderes en tecnología: Los líderes en tecnología (incluidos los propietarios de aplicaciones y productos) deben conocer los requisitos financieros (por ejemplo, las limitaciones presupuestarias), así como los requisitos empresariales (por ejemplo, los acuerdos de nivel de servicio). Esto permite que la carga de trabajo se implemente para lograr los objetivos empresariales deseados.

La colaboración entre los departamentos de Finanzas y Tecnología aporta los siguientes beneficios:

- Los equipos de finanzas y tecnología tienen visibilidad casi en tiempo real de los costes y el uso.
- Los equipos de finanzas y tecnología establecen un procedimiento operativo estándar para gestionar la variación del gasto en la nube.
- Las partes interesadas de finanzas actúan como asesores estratégicos en cuanto a cómo se utiliza el capital para comprar descuentos por compromiso de compra (por ejemplo, instancias reservadas o Savings Plans de AWS), y cómo se utiliza la nube para hacer crecer la organización.
- Las cuentas a pagar y los procesos de adquisición existentes también se usan en la nube.
- Los equipos de finanzas y tecnología colaboran a la hora de prever los costes y el uso de AWS en el futuro para adaptar y diseñar los presupuestos organizativos.
- Mejor comunicación dentro de la organización al compartir el mismo lenguaje y tener un conocimiento común de los conceptos financieros.

Otras partes interesadas de su organización que deberían estar implicadas en las discusiones sobre costes y uso son:

- Propietarios de unidades de negocio: Los propietarios de unidades de negocio deben comprender el modelo de negocio en la nube para poder establecer directrices para las unidades de negocio y toda la empresa. Este conocimiento de la nube resulta esencial para realizar previsiones de

crecimiento y de uso de las cargas de trabajo, pero también al valorar diferentes opciones de compra, por ejemplo, las instancias reservadas o los Savings Plans.

- Equipo de ingeniería: Establecer una asociación entre los equipos de finanzas y tecnología es esencial para crear una cultura sensibilizada con los costes que anime a los ingenieros a actuar en la administración financiera en la nube (CFM). Uno de los problemas habituales de los profesionales de la CFM o de las operaciones financieras y de los equipos de finanzas es conseguir que los ingenieros entiendan todo el negocio en la nube, sigan las prácticas recomendadas y adopten las medidas recomendadas.
- Terceros: Si en su organización participan terceros (por ejemplo, consultores o herramientas), asegúrese de que también sigan sus objetivos empresariales y que lo demuestren a través de sus modelos de compromiso y el retorno de la inversión (ROI). Por lo general, los terceros contribuyen a la generación de informes y al análisis de las cargas de trabajo que administren, y también aportan análisis de costes de cualquier carga de trabajo que diseñen.

Implementar la CFM y tener éxito requiere la colaboración entre los equipos de finanzas, tecnología y comercial, y un cambio en la forma en que se comunica y evalúa el gasto en la nube en toda la organización. Incluya a los equipos de ingeniería para que puedan formar parte de estos debates sobre costes y uso en todas las etapas, y anímelos a seguir las prácticas recomendadas y a adoptar las medidas acordadas de forma apropiada.

Pasos para la aplicación

- Defina los miembros clave: Compruebe que todos los miembros relevantes de sus equipos de finanzas y tecnología participen en la colaboración. Los miembros de Finanzas relevantes serán aquellos que interactúen con la factura de la nube. Suelen ser los directores financieros, los controllers financieros, los planificadores financieros, los analistas empresariales, los responsables de adquisiciones y los responsables de abastecimiento. Los miembros del equipo de Tecnología suelen ser los propietarios de las aplicaciones y de los productos, y los gerentes y representantes técnicos de todos los equipos que crean en la nube. Otros miembros pueden ser los propietarios de la unidad de negocio, como el departamento de Marketing, pues influyen en el uso de los productos, y terceros como consultores, para unirse a los objetivos y mecanismos, y para asistir en la gestión de informes.
- Defina los temas de discusión: Defina los temas comunes a todos los equipos o que requieran una compresión compartida. Haga un seguimiento del coste desde el momento en que se genera hasta que se paga la factura. Tome nota de todos los miembros implicados y de los procesos organizativos que deben aplicarse. Comprenda cada paso o proceso por el que pasa

y la información asociada, como los modelos de precios disponibles, los precios por niveles, los modelos de descuento, la creación de presupuestos y los requisitos financieros.

- Establezca una cadencia regular: Para crear una asociación entre finanzas y tecnología, establezca una cadencia de comunicación regular para crear y mantener la coherencia. El grupo debe reunirse de forma regular para tratar sobre sus objetivos y métricas. Una cadencia típica implica revisar el estado de la organización, revisar los programas que se ejecutan actualmente y las métricas generales financieras y de optimización. Después, se debe informar sobre las cargas de trabajo clave con mayor detalle.

Recursos

Documentos relacionados:

- Blog de noticias de AWS

COST01-BP03 Establecer presupuestos y previsiones de la nube

Ajuste los procesos de presupuestos y previsión organizativos para que sean compatibles con la naturaleza altamente variable de los costes y el uso de la nube. Los procesos deben ser dinámicos y usar algoritmos basados en tendencias o el motor principal de la empresa, o en una combinación de ambos.

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

Guía para la implementación

Los clientes usan la nube por su eficiencia, velocidad y agilidad, lo que genera una cantidad muy variable de costes y usos. Los costes pueden bajar (o a veces aumentar) a medida que las cargas de trabajo sean más eficientes o al desplegar cargas de trabajo y características nuevas. Las cargas de trabajo se pueden escalar para prestar servicio a más clientes, lo que aumenta el uso y los costes de la nube. Los recursos están ahora más accesibles que nunca. La elasticidad de la nube también conlleva una elasticidad de costes y previsiones. Los procesos de presupuestos existentes de la empresa deben modificarse en consecuencia para incorporar esta variabilidad.

Por lo general, el presupuesto se prepara para un solo año y permanece fijo, lo que requiere un estricto cumplimiento por parte de todas las partes involucradas. Por el contrario, las previsiones son más flexibles, ya que permiten reajustes a lo largo del año y proporcionan proyecciones dinámicas durante un período de uno, dos o tres años. Tanto la elaboración de presupuestos

como de previsiones desempeñan un papel crucial a la hora de establecer las expectativas financieras entre las diversas partes interesadas de la tecnología y la empresa. La previsión y la implementación veraces también exigen responsabilidades a las partes interesadas que son directamente responsables de los costes de aprovisionamiento en primer lugar, y también pueden aumentar su concienciación general de los costes.

Ajuste los procesos de elaboración de presupuestos y previsiones existentes para que sean más dinámicos, ya sea utilizando un algoritmo basado en las tendencias (que usa los costes históricos como entradas) o algoritmos basados en impulsores (por ejemplo, lanzamientos de productos nuevos, expansión regional o nuevos entornos para cargas de trabajo), lo que resulta ideal para un entorno de gasto dinámico y variable, o una combinación tanto de impulsores empresariales como de tendencias.

Puede utilizar AWS Cost Explorer para previsiones basadas en tendencias en un intervalo de tiempo futuro definido en función de su gasto anterior. El motor de pronóstico de AWS Cost Explorer segmenta los datos históricos en función de los tipos de cargo (por ejemplo, instancias reservadas) y utiliza una combinación de machine learning y modelos basados en reglas para predecir el gasto en todos los tipos de cargo individualmente.

Identifique los impulsores empresariales que pueden repercutir en sus costes de uso y haga previsiones para cada uno de ellos por separado para garantizar que el uso esperado se calcule con antelación. Algunos de los impulsores están relacionados con los equipos de TI y de productos de la organización. Los líderes de ventas, marketing y negocios conocen otros impulsores empresariales, como los eventos de marketing, las promociones, las fusiones y las adquisiciones, y es importante colaborar y tener en cuenta también todos esos impulsores de la demanda. Debe trabajar en estrecha colaboración con ellos para comprender la repercusión en los nuevos impulsores internos.

Una vez que haya determinado su previsión basada en tendencias con Cost Explorer o cualquier otra herramienta, utilice la Calculadora de precios de AWS para estimar su caso de uso de AWS y los costes futuros en función del uso previsto (tráfico, solicitudes por segundo o instancia de Amazon EC2 necesaria). También puede utilizarla para ayudarle a planificar sus gastos, encontrar oportunidades de ahorro y tomar decisiones informadas al utilizar AWS. Es importante hacer un seguimiento de la precisión de esa previsión, ya que los presupuestos deben establecerse en función de estos cálculos y estimaciones de previsión.

Utilice AWS Budgets para establecer presupuestos personalizados con gran nivel de detalle mediante la especificación del periodo de tiempo, la recurrencia o la cantidad (fija o variable) y la adición de filtros tales como servicio, Región de AWS y etiquetas. Para mantenerse informado sobre el rendimiento de sus presupuestos existentes, puede crear y programar informes de presupuesto

de AWS Budgets para que se envíen por correo electrónico tanto a usted y como a otras partes interesadas con regularidad. También puede crear alertas de AWS Budgets basadas en los costes reales (de naturaleza reactiva) o en los costes previstos, lo que proporciona tiempo para implementar mitigaciones contra posibles sobrecostes. Puede recibir una alerta cuando el coste o el uso superen, o se prevea que superen, el importe presupuestado.

Utilice AWS Cost Anomaly Detection para evitar o reducir las sorpresas en los costes y mejorar el control sin frenar la innovación. AWS Cost Anomaly Detection saca provecho del machine learning para identificar los gastos anómalos y las causas raíz, de modo que pueda adoptar medidas rápidamente. Con tres sencillos pasos, puede crear su propio monitor contextualizado y recibir alertas cuando se detecte cualquier gasto anómalo.

Como se menciona en la sección Colaboración entre los departamentos de finanzas y tecnologíadel pilar de optimización de costes de Well-Architected, es importante contar con asociaciones y cadencias entre los departamentos de TI, finanzas y otras partes interesadas para garantizar que todos utilicen las mismas herramientas o procesos en aras de la coherencia. En los casos en que los presupuestos deban cambiar, el aumento de los puntos de contacto de la cadencia puede ayudar a reaccionar a esos cambios más rápidamente.

Pasos para la implementación

- Analice las previsiones basadas en tendencias: utilice las herramientas preferidas de previsión basada en tendencias, como AWS Cost Explorer y Amazon Forecast. Analice su coste de uso en las diferentes dimensiones, como servicio, cuentas, etiquetas y categorías de costes. Si se requiere una previsión avanzada, importe sus datos de AWS Cost and Usage Report a Amazon Forecast (que aplica la regresión lineal como una forma de machine learning para realizar previsiones).
- Analice las previsiones basadas en impulsores: identifique el efecto de los impulsores empresariales en el uso de la nube y realice previsiones para cada uno de ellos por separado para calcular el coste de uso esperado con antelación. Trabaje en estrecha colaboración con los responsables de las unidades empresariales y las partes interesadas para comprender la repercusión en los nuevos impulsores y calcular los cambios de costes esperados para definir presupuestos veraces.
- Actualice los procesos de elaboración de presupuestos y previsiones existentes: defina sus procesos de elaboración de presupuestos y previsiones en función de los métodos de previsión adoptados, como los basados en tendencias, los basados en impulsores empresariales o una combinación. Los presupuestos deben ser calculados y realistas, en función de estos procesos de elaboración de previsiones.

- Configure alertas y notificaciones: Utilice las alertas de AWS Budgets y AWS Cost Anomaly Detection para recibir alertas y notificaciones.
- Realice revisiones periódicas con las principales partes interesadas: Por ejemplo, las partes interesadas en TI, finanzas, equipos de plataforma y otras áreas de la empresa deben alinearse con los cambios en la dirección y el uso de la empresa.

Recursos

Documentos relacionados:

- AWS Cost Explorer
- AWS Cost and Usage Report
- Quick Forecasting
- Amazon Forecast
- AWS Budgets
- Blog de noticias de AWS

Vídeos relacionados:

- How can I use AWS Budgets to track my spending and usage
- AWS Cost Optimization Series: AWS Budgets

Ejemplos relacionados:

- Comprender y crear previsiones basadas en impulsores
- Cómo establecer e impulsar una cultura de previsión
- Cómo mejorar sus previsiones de costes en la nube
- Uso de las herramientas adecuadas para la previsión de costes de la nube

COST01-BP04 Implementar la conciencia de costes en los procesos organizativos

Implemente la conciencia de costes, cree transparencia y responsabilidad de los costes en los procesos nuevos y existentes que afecten al uso, y aproveche los procesos existentes para tomar conciencia de los costes. Implemente la conciencia de costes en la formación del personal.

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

Guía para la implementación

La conciencia de los costes debe implementarse en los procesos organizativos nuevos y existentes. Es una de las capacidades fundamentales, requisito previo para otras prácticas recomendadas. Se recomienda reutilizar y modificar los procesos existentes cuando sea posible, lo que minimiza el impacto en la agilidad y la velocidad. Informe de los costes de la nube a los equipos de tecnología y a los responsables de la toma de decisiones en los equipos de negocio y finanzas para concienciar sobre los costes, y establecer indicadores clave de rendimiento (KPI) de eficiencia para las partes interesadas de finanzas y negocios. Las siguientes recomendaciones ayudarán a implementar la conciencia de los costes en su carga de trabajo:

- Verifique que la gestión de los cambios incluya la medición de los costes para cuantificar el efecto financiero de sus cambios. Esto ayuda a abordar de forma proactiva las preocupaciones relacionadas con los costes y destacar el ahorro de costes.
- Verifique que la optimización de costes sea un componente central de sus capacidades operativas. Por ejemplo, puede aprovechar los procesos de gestión de incidentes para investigar e identificar la causa raíz de las anomalías de los costes y el uso o costes excesivos.
- Acelere el ahorro de costes y la materialización del valor de negocio a través de la automatización o las herramientas. Al pensar en el coste de implementación, enmarque la conversación para que incluya un componente de rendimiento de la inversión (ROI) para justificar la inversión de tiempo o dinero.
- Asigne los costes de la nube mediante la aplicación de devoluciones o reembolsos de los gastos en la nube, incluidos los gastos en las opciones de compra basadas en el compromiso, los servicios compartidos y las compras en el mercado para impulsar el consumo de la nube teniendo siempre presentes los costes.
- Amplíe los programas de formación y desarrollo existentes para que incluyan la sensibilización con los costes en toda la organización. Se recomienda incluir formación y certificaciones continuas. Con ello logrará tener una organización capaz de autoadministrar los costes y el uso.
- Aproveche las herramientas nativas gratuitas de AWS como AWS Cost Anomaly Detection, AWS Budgets e informes de presupuesto de AWS Budgets.

cuando las organizaciones adoptan de forma sistemática prácticas de Administración financiera en la nube (CFM), esos comportamientos se arraigan en la forma de trabajar y tomar decisiones. El resultado es una cultura que tiene más en cuenta los costes, desde los desarrolladores que diseñan

una nueva aplicación nacida en la nube hasta los administradores financieros que analizan el retorno de estas nuevas inversiones en la nube.

Pasos para la aplicación

- Identifique los procesos organizativos relevantes: Cada unidad organizativa debe revisar sus procesos e identificar los procesos que afecten a los costes y el uso. Cualquier proceso que conlleve la creación o finalización de un recurso debe incluirse en la revisión. Debe buscar procesos que ayuden a tomar conciencia de los costes en su negocio, como la administración de incidentes y la formación.
- Establezca una cultura autosuficiente en materia de costes: asegúrese de que todas las partes interesadas pertinentes se alinean con la causa del cambio y el impacto como coste para que entiendan el coste de la nube. Esto permitirá a su organización establecer una cultura de innovación autosuficiente y sensibilizada con los costes.
- Actualice los procesos con conciencia de costes: debe cambiarse cada proceso para que incluya la toma de conciencia de costes. El proceso puede requerir controles previos adicionales, como valorar el efecto del coste, o controles posteriores que validen que se han producido los cambios esperados en el coste y el uso. Dar soporte a procesos tales como la formación y la gestión de incidentes puede ampliarse para incluir elementos de coste y uso.

Para obtener ayuda, póngase en contacto con los expertos de CFM a través de su equipo de cuentas, o explore los recursos y documentos relacionados a continuación.

Recursos

Documentos relacionados:

- Administración financiera en la nube de AWS

Ejemplos relacionados:

- Strategy for Efficient Cloud Cost Management (Estrategia para la eficiencia en la Administración financiera en la nube)
- Cost Control Blog Series #3: How to Handle Cost Shock (Serie de blog sobre control de costes n.º 3: cómo gestionar el choque de costes)
- A Beginner's Guide to AWS Cost Management (Guía para principiantes de AWS Cost Management)

COST01-BP05 Crear informes y notificar la optimización de costes

Establezca presupuestos en la nube y configure mecanismos para detectar anomalías en el uso. Configure las herramientas relacionadas para que proporcionen alertas de coste y uso respecto a objetivos predefinidos y reciba notificaciones cuando el uso supere esos objetivos. Organice reuniones periódicas para analizar la rentabilidad de las cargas de trabajo y promover la concienciación en cuanto a los costes.

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

Guía para la implementación

Debe informar regularmente sobre la optimización de costes y el uso en su organización. Puede implementar sesiones dedicadas a examinar el rendimiento de costes o incluir la optimización de costes en los ciclos de preparación de informes operativos habituales para sus cargas de trabajo. Utilice los servicios y las herramientas para supervisar el rendimiento de los costes con regularidad e implementar oportunidades de ahorro de costes.

Examine los costes y uso con diversos filtros y especificidad mediante el uso de AWS Cost Explorer, que proporciona paneles e informes, como costes por servicio o por cuenta, costes diarios o costes del marketplace. Realice un seguimiento de la evolución del coste y el uso según los presupuestos creados con informes de presupuesto de AWS Budgets.

Utilice AWS Budgets para establecer presupuestos personalizados para hacer un seguimiento de los costes y el uso, y responder rápidamente a las alertas recibidas por correo electrónico o mediante notificaciones de Amazon Simple Notification Service (Amazon SNS) si supera el umbral. Establezca su período de presupuesto preferido como diario, mensual, trimestral o anual, y cree límites presupuestarios específicos para mantenerse informado sobre el progreso de los costes reales o previstos y el uso hacia su umbral presupuestario. También puede configurar alertas y acciones contra esas alertas para que se ejecuten automáticamente, o mediante un proceso de aprobación cuando se supere un objetivo presupuestario.

Implemente notificaciones sobre costes y uso para garantizar que se pueda actuar rápidamente ante cambios en los costes y el uso en caso de que sean inesperados. AWS Cost Anomaly Detection permite reducir las sorpresas de costes y mejorar el control sin ralentizar la innovación. AWS Cost Anomaly Detection identifica los gastos anómalos y las causas que los originan, lo que ayuda a reducir el riesgo de sorpresas en la facturación. Con tres sencillos pasos, puede crear su propio monitor contextualizado y recibir alertas cuando se detecte cualquier gasto anómalo.

También puede utilizar Quick con datos de AWS Cost and Usage Report (CUR) para proporcionar informes altamente personalizados con datos más pormenorizados. Quick permite programar informes y recibir periódicamente correos electrónicos de informes de costes para conocer el historial de costes y uso, o las oportunidades de ahorro. Consulte nuestra solución Panel de inteligencia de costes (CID) basada en Quick, que le proporciona una visibilidad avanzada.

Utilice AWS Trusted Advisor, que proporciona orientación para verificar si los recursos aprovisionados están en consonancia con las prácticas recomendadas de AWS para la optimización de costes.

Compare sus recomendaciones de Savings Plans a través de gráficos visuales con sus costes y uso detallados. Los gráficos por hora muestran el gasto bajo demanda junto con el compromiso de Savings Plans recomendado, lo que proporciona información sobre los ahorros, la cobertura de los Savings Plans y la utilización de Savings Plans estimados. Esto ayuda a las organizaciones a comprender cómo sus Savings Plans se aplican a cada hora de gasto sin tener que invertir tiempo y recursos en la creación de modelos para analizar el gasto.

Cree periódicamente informes que contengan un resumen de los Savings Plans, las instancias reservadas y Amazon EC2 las recomendaciones de tamaño adecuado de AWS Cost Explorer para empezar a reducir el coste asociado a las cargas de trabajo en estado estable, los recursos ociosos y los infrautilizados. Identifique y recupere el gasto asociado a los residuos de la nube para los recursos que se despliegan. El desperdicio en la nube se produce cuando se crean recursos de tamaño incorrecto o se observan patrones de uso diferentes a los esperados. Siga las prácticas recomendadas de AWS para reducir el despilfarro o pida a su socio y equipo de cuentas que le ayuden a optimizar y ahorrar costes de la nube.

Genere informes con regularidad para mejorar las opciones de compra de sus recursos y reducir los costes unitarios de sus cargas de trabajo. Las opciones de compra, como los Savings Plans, las instancias reservadas o las instancias de spot de Amazon EC2, ofrecen el mayor ahorro de costes para las cargas de trabajo con tolerancia a errores y permiten a las partes interesadas (propietarios de la empresa, equipos financieros y técnicos) formar parte de estas conversaciones de compromiso.

Comparta los informes que contengan oportunidades o anuncios de nuevas versiones que puedan ayudarle a reducir el coste total de propiedad (TCO) de la nube. Adopte nuevos servicios, regiones, funciones, soluciones o nuevas formas de lograr una mayor reducción de costes.

Pasos para la implementación

- Configure AWS Budgets: Configure AWS Budgets en todas las cuentas de su carga de trabajo. Establezca un presupuesto para el gasto general de la cuenta y un presupuesto para la carga de trabajo con etiquetas.
  - Well-Architected Labs: coste y uso de la gobernabilidad
- Informe sobre optimización de costes: Defina un ciclo habitual para tratar y analizar la eficiencia de la carga de trabajo. Utilice las métricas establecidas, notifique las métricas alcanzadas y el coste para alcanzarlas. Identifique y corrija las tendencias negativas, así como las tendencias positivas que puede promover en su organización. La preparación de informes debe incluir a representantes de los equipos y propietarios de las aplicaciones, de las finanzas y los principales responsables de la toma de decisiones con respecto al gasto en la nube.

Recursos

Documentos relacionados:

- AWS Cost Explorer
- AWS Trusted Advisor
- AWS Budgets
- AWS Cost and Usage Report
- Prácticas recomendadas de AWS Budgets
- Análisis de Amazon S3

Ejemplos relacionados:

- Well-Architected Labs: coste y uso de la gobernabilidad
- Formas clave para empezar a optimizar los costes de la nube de AWS

COST01-BP06 Supervisar los costes de forma proactiva

Implemente herramientas y paneles para supervisar los costes de forma proactiva para la carga de trabajo. Revise periódicamente los costes con herramientas configuradas o listas para usar, no se limite a mirar los costes y las categorías cuando reciba las notificaciones. Supervisar y analizar los costes de forma proactiva ayuda a identificar las tendencias positivas y permite promoverlas en toda la organización.

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

Guía para la implementación

Se recomienda supervisar los costes y el uso de forma proactiva dentro de la organización, no solo cuando surjan anomalías o excepciones. Disponer de unos paneles muy visibles en la oficina o en el entorno de trabajo permite a las personas clave tener acceso a la información que necesitan y transmite la idea de que la organización se centra en la optimización de costes. Los paneles visibles permiten promover de forma activa los resultados de éxito e implementarlos en toda la organización.

Cree una rutina diaria o frecuente para utilizar AWS Cost Explorer o cualquier otro panel como Amazon Quick para ver los costes y analizarlos de forma proactiva. Analice el uso y los costes de los servicios de AWS al nivel de cuenta de AWS, a nivel de carga de trabajo o a nivel de servicio específico de AWS con agrupamientos y filtrado, y valide si son los esperados o no. Utilice la granularidad a nivel de hora y de recurso y las etiquetas para filtrar e identificar los costes incurridos para los principales recursos. También puede crear sus propios informes con el Panel de inteligencia de costes, una solución Amazon Quick desarrollada por arquitectos de soluciones de AWS, y comparar sus presupuestos con el coste y el uso reales.

Pasos para la aplicación

- Informe sobre optimización de costes: Defina un ciclo habitual para tratar y analizar la eficiencia de la carga de trabajo. Utilice las métricas establecidas, notifique las métricas alcanzadas y el coste para alcanzarlas. Identifique y corrija las tendencias negativas e identifique las tendencias positivas que quiere promover en su organización. La gestión de informes debe implicar a los representantes de los equipos de aplicaciones y de los propietarios, del departamento financiero y de dirección.
- Cree y active AWS Budgets de granularidad diaria del coste y el uso para adoptar las medidas oportunas para evitar cualquier posible sobrecoste: AWS Budgets le permiten configurar notificaciones de alerta, para que esté informado si alguno de sus tipos de presupuesto se sale de los umbrales preconfigurados. La mejor manera de aprovechar AWS Budgets es establecer los costes y el uso previstos como límites, de modo que todo lo que supere los presupuestos se considere un gasto excesivo.
- Cree AWS Cost Anomaly Detection para la supervisión de costes: AWS Cost Anomaly Detection utiliza tecnología avanzada de machine learning para identificar los gastos anómalos y las causas que los originan para que pueda adoptar medidas rápidamente. Le permite configurar monitores de costes que definen los segmentos de gastos que desea evaluar (por ejemplo, servicios individuales de AWS, cuentas de miembros, etiquetas de asignación de costes y categorías

de costes) y le permite establecer cuándo, dónde y cómo recibir sus notificaciones de alerta. Para cada monitor, adjunte varias suscripciones de alerta para los propietarios de negocios y los equipos de tecnología, que incluyan un nombre, un umbral de impacto de costes y la frecuencia de las alertas (alertas individuales, resumen diario, resumen semanal) para cada suscripción.

- Utilice AWS Cost Explorer o integre sus datos de AWS Cost and Usage Report (CUR) con paneles de Amazon Quick para visualizar los costes de su organización: AWS Cost Explorer tiene una interfaz sencilla que le ayuda a visualizar, comprender y administrar los costes y el uso de AWS a lo largo del tiempo. El Panel de inteligencia de costes es un panel personalizable y accesible para ayudar a crear la base de su propia herramienta de administración y optimización de costes.

Recursos

Documentos relacionados:

- AWS Budgets
- AWS Cost Explorer
- Daily Cost and Usage Budgets (Presupuestos de coste y uso diarios)
- AWS Cost Anomaly Detection

Ejemplos relacionados:

- Well-Architected Labs: visualización
- Well-Architected Labs: visualización avanzada
- Well-Architected Labs: paneles de inteligencia en la nube
- Well-Architected Labs: visualización de los costes
- AWS Cost Anomaly Detection Alert with Slack (Alerta de AWS Cost Anomaly Detection con Slack)

COST01-BP07 Estar al día sobre las nuevas versiones de los servicios

Consulte regularmente con expertos o socios de AWS qué servicios y características proporcionan un coste inferior. Revise los blogs de AWS y otras fuentes de información.

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

Guía para la implementación

AWS está añadiendo constantemente nuevas capacidades para que pueda aprovechar las últimas tecnologías para experimentar e innovar más rápidamente. Puede implementar nuevos servicios y características de AWS para incrementar la rentabilidad de su carga de trabajo. Revise periódicamente la Administración de costes de AWS, el Blog de noticias de AWS, el blog de Administración de costes de AWS y las novedades de AWS para obtener información sobre las nuevas versiones de los servicios y características. Las publicaciones sobre las novedades ofrecen un breve resumen de todos los anuncios de servicios, funciones y ampliación de regiones de AWS a medida que se publican.

Pasos para la aplicación

- Suscríbase a los blogs: Vaya a las páginas de los blogs de AWS y suscríbase al blog de novedades y a otros blogs relevantes. Puede inscribirse en la página de preferencias de comunicaciones con su dirección de correo electrónico.
- Suscríbase a las noticias de AWS: Revise periódicamente el Blog de noticias de AWS y las novedades de AWS para obtener información sobre las nuevas versiones de los servicios y características. Suscríbase al canal RSS o con su correo electrónico para seguir los anuncios y lanzamientos.
- Siga las Reducciones de precios de AWS: La reducción periódica de los precios de todos nuestros servicios ha sido una forma habitual para que AWS traslade a nuestros clientes las eficiencias económicas obtenidas gracias a nuestra escala. A partir de abril de 2022, AWS ha reducido los precios 115 veces desde su lanzamiento en 2006. Si tiene alguna decisión comercial pendiente por cuestiones de precio, puede volver a revisarla después de las reducciones de precio y las nuevas integraciones de servicios. Puede conocer los esfuerzos anteriores de reducción de precios, incluidas las instancias de Amazon Elastic Compute Cloud (Amazon EC2), en la categoría de reducción de precios del blog de noticias de AWS.
- Eventos y reuniones de AWS: Asista a la cumbre local deAWS y a cualquier reunión local con otras organizaciones de su zona. Si no puede asistir en persona, intente asistir a los eventos virtuales para conocer mejor a los expertos de AWS y los casos empresariales de otros clientes.
- Reúnase con el equipo de cuentas: Programe una cadencia regular con su equipo de cuentas, reúnase con él y trate sobre las tendencias del sector y los servicios de AWS. Hable con el gerente de cuentas, el arquitecto de soluciones y el equipo de soporte.

Recursos

Documentos relacionados:

- Administración de costes de AWS
- las novedades de AWS
- Blog de noticias de AWS

Ejemplos relacionados:

- Amazon EC2 – 15 Years of Optimizing and Saving Your IT Costs (Amazon EC2: 15 años de optimización y ahorro de costes de TI)
- Blog de noticias de AWS: reducción de precios

COST01-BP08 Crear una cultura de conciencia de costes

Implemente cambios o programas en la organización para crear una cultura de conciencia de costes. Se recomienda empezar discretamente, y a medida que crezcan las capacidades y el uso de la nube por parte de la empresa implementar programas grandes y de gran alcance.

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

Guía para la implementación

Afianzar una cultura de conciencia de los costes permite mejorar la optimización de costes y la administración financiera en la nube (operaciones financieras, centro de excelencia en la nube, equipos de operaciones en la nube, etc.) a través de las prácticas recomendadas llevadas a cabo de forma orgánica y descentralizada en toda la organización. La concienciación sobre los costes permite obtener grandes niveles de capacidad en la organización con un esfuerzo mínimo, en comparación con un enfoque centralizado y descendente.

La concienciación sobre los costes en la computación en la nube, especialmente para los principales impulsores de costes en la computación en la nube, permite a los equipos comprender los resultados esperados de cualquier cambio en la perspectiva de los costes. Los equipos que acceden a los entornos de la nube deben conocer los modelos de precios y la diferencia entre los centros de datos tradicionales locales y la computación en la nube.

La principal ventaja de una cultura consciente de los costes es que los equipos tecnológicos los optimizan de forma proactiva y continua (por ejemplo, se consideran un requisito no funcional a

la hora de diseñar nuevas cargas de trabajo o de realizar cambios en las existentes) en lugar de realizar optimizaciones de costes reactivas según sea necesario.

Aplicar unos pequeños cambios en la cultura puede tener un gran impacto en la eficiencia de las cargas de trabajo actuales y futuras. Ejemplos:

- Dar visibilidad y sensibilizar a los equipos de ingeniería para que comprendan lo que hacen y su impacto en términos de costes.
- Ludificar los costes y el uso en toda la organización. Esto se puede realizar con un panel visible para todo el personal o mediante un informe que compare los costes normalizados y el uso de los diferentes equipos (por ejemplo, coste por carga de trabajo, coste por transacción).
- Reconocer la rentabilidad. Premiar los logros voluntarios o espontáneos de optimización de costes de forma pública o privada, y aprender de los errores para no repetirlos en el futuro.
- Crear requisitos organizativos descendentes para que las cargas de trabajo se lleven a cabo con presupuestos predefinidos.
- Cuestionar los requisitos empresariales de los cambios y el impacto de los costes de los cambios solicitados en la infraestructura de la arquitectura o la configuración de la carga de trabajo para asegurarse de que se paga solo lo que se necesita.
- Asegurarse de que el planificador del cambio es consciente de los cambios previstos que tienen un impacto en los costes, y que estos son confirmados por las partes interesadas para obtener resultados empresariales de forma rentable.

Pasos para la aplicación

- Informe de los costes de la nube a los equipos tecnológicos: Para aumentar la concienciación sobre costes y establecer indicadores clave de eficiencia para las partes interesadas de las finanzas y la empresa.
- Informe a las partes interesadas o a los miembros del equipo sobre los cambios previstos: Cree un punto en el orden del día para debatir los cambios previstos y el impacto del coste-beneficio en la carga de trabajo durante las reuniones semanales sobre cambios.
- Reúnase con el equipo de cuentas: Establezca una cadencia de reuniones regular con su equipo de cuentas, y trate las tendencias del sector y los servicios de AWS. Hable con el gerente de cuentas, el arquitecto y el equipo de soporte.
- Comparta casos de éxito: Comparta historias de éxito sobre la reducción de costes para cualquier carga de trabajo, Cuenta de AWS u organización para crear una actitud positiva y un estímulo en torno a la optimización de costes.

- Entrenamiento: Asegúrese de que los equipos técnicos o los miembros del equipo reciban formación para conocer los costes de los recursos en Nube de AWS.
- Eventos y reuniones de AWS: Asista a las cumbres locales deAWS y a cualquier reunión local con otras organizaciones de su zona.
- Suscríbase a los blogs: Vaya a las páginas de blogs de AWS y suscríbase al blog de novedades y otros blogs relevantes para seguir las nuevas versiones, implementaciones, ejemplos y cambios compartidos por AWS.

Recursos

Documentos relacionados:

- Blog de AWS
- Administración de costes de AWS
- Blog de noticias de AWS

Ejemplos relacionados:

- Administración financiera en la nube de AWS
- AWS Well-Architected Labs: administración financiera en la nube

COST01-BP09 Cuantificar el valor empresarial a partir de la optimización de costes

Cuantificar el valor empresarial a partir de la optimización de costes le permite comprender todos los beneficios para su organización. Dado que la optimización de costes es una inversión necesaria, cuantificar el valor empresarial le permite explicar el retorno de la inversión a las partes interesadas. Cuantificar el valor empresarial le puede ayudar a lograr mayor aceptación de las partes interesadas para realizar inversiones futuras en optimización de costes y, además, le proporciona un marco para medir los resultados de las actividades de optimización de costes de la organización.

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

Guía para la implementación

Cuantificar el valor empresarial implica determinar los beneficios que las empresas obtienen de las acciones y decisiones que toman. El valor empresarial puede ser tangible (por ejemplo, una

reducción de los gastos o un aumento de los beneficios) o intangible (por ejemplo, la mejora de la reputación de la marca o el aumento de la satisfacción del cliente).

Para cuantificar el valor empresarial que se produce con la optimización de costes, hay que determinar cuánto valor o beneficio obtiene de sus esfuerzos por gastar de manera más eficiente. Por ejemplo, si una empresa gasta 100 000 USD en desplegar una carga de trabajo en AWS y, posteriormente, la optimiza, el nuevo coste pasa a ser de solo 80 000 USD sin sacrificar la calidad ni el rendimiento. En este escenario, el valor empresarial cuantificado de la optimización de costes sería un ahorro de 20 000 USD. Pero más allá del ahorro, la empresa también podría cuantificar el valor en términos de tiempos de entrega más rápidos, una mayor satisfacción del cliente u otras métricas que se deriven de los esfuerzos de optimización de costes. Las partes interesadas deben tomar decisiones sobre el valor potencial de la optimización de costes, el coste de optimizar la carga de trabajo y el valor de retorno.

Además de informar sobre los ahorros de la optimización de costes, se recomienda cuantificar el valor adicional conseguido. Los beneficios de la optimización de costes se suelen cuantificar en términos de menos costes por resultado empresarial. Por ejemplo, puede cuantificar los ahorros de costes de Amazon Elastic Compute Cloud (Amazon EC2) al comprarSavings Plans, lo que reduce los costes y mantiene los niveles de producción de la carga de trabajo. Puede cuantificar reducciones de costes en el gasto de AWS cuando se eliminan las instancias de Amazon EC2 inactivas o cuando se eliminan volúmenes de Amazon Elastic Block Store (Amazon EBS) no asociados.

Sin embargo, la optimización de costes tiene muchos más beneficios, aparte de reducir o evitar costes. Plantéese capturar más datos para medir las mejoras en la rentabilidad y el valor empresarial.

Pasos para la implementación

- Evalúe los beneficios empresariales: en este proceso, se analizan y ajustan los costes de Nube de AWS de manera que se maximice el beneficio obtenido por el dinero gastado. En lugar de centrarse en la reducción de costes sin valor empresarial, considere los beneficios empresariales y la rentabilidad de la optimización de los costes, lo que puede aportar más valor al dinero que gasta. Se trata de gastar con prudencia y realizar inversiones y gastos en las áreas que tengan el mejor rendimiento.
- Analice las previsiones de los costes de AWS: las previsiones permiten a las partes interesadas financieras establecer expectativas con otras partes interesadas internas y externas de la organización, además de mejorar las predicciones financieras de la organización. AWS Cost Explorer se puede usar para realizar previsiones de costes y uso.

Recursos

Documentos relacionados:

- «Economía»
- Blog de AWS
- «AWS Cost Management»
- Blog de noticias de AWS
- Documento técnico «Pilar de fiabilidad: AWS Well-Architected Framework»
- «Explorador de costes de AWS«

Vídeos relacionados:

- «Unlock Business Value with Windows on AWS»

Ejemplos relacionados:

- «Measuring and Maximizing the Business Value of Customer 360»
- «The Business Value of Adopting Amazon Web Services Managed Databases»
- «The Business Value of Amazon Web Services for Independent Software Vendors»
- «Business Value of Cloud Modernization»
- «The Business Value of Migration to Amazon Web Services»

Conocimiento del gasto y del uso

Preguntas

- COSTE 2. ¿Cómo controla el uso?
- COSTE 3. ¿Cómo supervisa sus costes y su uso?
- COSTE 4. ¿Cómo retira los recursos?

COSTE 2. ¿Cómo controla el uso?

Establezca políticas y mecanismos para comprobar que se incurre en costes apropiados mientras se alcanzan los objetivos. Cuando emplea un enfoque de evaluar la situación, puede innovar sin gastar de más.

Prácticas recomendadas

- COST02-BP01 Desarrollar políticas basadas en los requisitos de su organización
- COST02-BP02 Implementar objetivos y metas
- COST02-BP03 Implementar una estructura de cuentas
- COST02-BP04 Implementar grupos y roles
- COST02-BP05 Implementación de controles de costes
- COST02-BP06 Controlar el ciclo de vida de los proyectos

COST02-BP01 Desarrollar políticas basadas en los requisitos de su organización

Desarrolle políticas que definan la forma en que su organización administra los recursos e inspecciónelas periódicamente. Las políticas deben abarcar los aspectos de coste de los recursos y las cargas de trabajo, como su creación, modificación y retirada durante la vida útil del recurso.

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

Guía para la implementación

Es fundamental comprender los costes y lo que impulsa su organización para administrar eficazmente el coste y el uso, e identificar las oportunidades de reducción de costes. Las organizaciones suelen operar con múltiples cargas de trabajo que gestionan varios equipos. Estos equipos pueden estar en diferentes unidades organizativas, cada una con su propio flujo de ingresos. La capacidad de atribuir los costes de los recursos a las cargas de trabajo, a la organización individual o a los propietarios de los productos impulsa un comportamiento de uso eficiente y contribuye a reducir los residuos. La monitorización precisa de los costes y el uso le ayuda a comprender el nivel de optimización de una carga de trabajo, así como la rentabilidad de las unidades y los productos de la organización. Este conocimiento permite tomar decisiones más fundamentadas sobre dónde asignar los recursos dentro de la organización. El conocimiento del uso en todos los niveles de la organización es clave para impulsar el cambio, ya que el cambio en el uso impulsa los cambios en el coste. Considere la posibilidad de adoptar un enfoque multifacético para conocer su uso y sus gastos.

El primer paso para llevar a cabo la gobernanza es utilizar los requisitos de su organización para desarrollar políticas para el uso de la nube. Estas políticas definen cómo su organización utiliza la nube y cómo se administran los recursos. Las políticas deben tratar todos los aspectos de los recursos y las cargas de trabajo que tienen que ver con el coste o el uso, como su creación,

modificación y retirada durante la vida útil del recurso. Verifique que se siguen e implementan las políticas y los procedimientos ante cualquier cambio en un entorno en la nube. Durante sus reuniones de administración de cambios de TI, formule preguntas para averiguar el impacto en los costes de los cambios previstos, tanto si aumentan como si disminuyen, la justificación empresarial y el resultado esperado.

Las políticas deben ser sencillas para que se comprendan fácilmente y puedan implementarse con eficacia en toda la organización. Las políticas también deben ser fáciles de seguir e interpretar (para que se usen) y específicas (para que no haya malinterpretaciones entre los equipos). Además, deben inspeccionarse periódicamente (igual que nuestros mecanismos) y actualizarse a medida que cambien las condiciones empresariales o las prioridades de los clientes, ya que esto podría hacer que la política quedara obsoleta.

Empiece con políticas amplias y generales, como la región geográfica que se usará o las horas del día en las que deben funcionar los recursos. Mejore gradualmente las políticas para las distintas unidades organizativas y cargas de trabajo. Entre las políticas más comunes se incluyen los servicios y las características que pueden utilizarse (por ejemplo, el almacenamiento de menor rendimiento en los entornos de prueba o de desarrollo), los tipos de recursos que pueden utilizar los distintos grupos (por ejemplo, el mayor tamaño de recurso en una cuenta de desarrollo es el medio) y durante cuánto tiempo estarán en uso estos recursos (temporalmente, a corto plazo o durante un periodo de tiempo específico).

Ejemplo de política

A continuación, tenemos un ejemplo de política que puede utilizar para crear sus propias políticas de gobernanza de la nube que se centren en la optimización de costes. Asegúrese de ajustar la política en función de los requisitos de su organización y de las solicitudes de las partes interesadas.

- Nombre de la política: defina un nombre claro, como «Política de optimización de recursos y reducción de costes».
- Finalidad: explique por qué se debe utilizar esta política y cuál es el resultado esperado. El objetivo de esta política es verificar que se requiere un coste mínimo para desplegar y ejecutar la carga de trabajo deseada con el fin de cumplir los requisitos empresariales.
- Ámbito: defina claramente quién debe usar esta política y cuándo debe usarse; por ejemplo, podría indicar que el equipo X de DevOps X debe usar esta política en los clientes de la región us-east para el entorno X (de producción o no de producción).

Declaración de la política

1. Seleccione us-east-1 o varias regiones de us-east en función del entorno y los requisitos empresariales de su carga de trabajo (desarrollo, pruebas de aceptación de los usuarios, preproducción o producción).
2. Programe instancias de Amazon EC2 y Amazon RDS para que se ejecuten entre las seis de la mañana y las ocho de la tarde (hora estándar del este [EST]).
3. Detenga todas las instancias de Amazon EC2 no utilizadas después de ocho horas y las instancias de Amazon RDS no utilizadas después de 24 horas de inactividad.
4. Termine todas las instancias de Amazon EC2 no utilizadas después de 24 horas de inactividad en entornos que no sean de producción. Recuérdele al propietario de la instancia de Amazon EC2 (basándose en las etiquetas) que revise las instancias de Amazon EC2 detenidas en producción e infórmele de que sus instancias de Amazon EC2 se cancelarán en un plazo de 72 horas si no están en uso.
5. Utilice una familia y un tamaño de instancias genéricos, como m5.large, y luego cambie el tamaño de la instancia en función del uso de la CPU y la memoria mediante AWS Compute Optimizer.
6. Priorice el uso del escalamiento automático para ajustar dinámicamente la cantidad de instancias en ejecución en función del tráfico.
7. Utilice instancias de spot para cargas de trabajo no críticas.
8. Revise los requisitos de capacidad para confirmar Savings Plans o instancias reservadas para cargas de trabajo predecibles e informe al equipo de administración financiera en la nube.
9. Utilice políticas de ciclo de vida de Amazon S3 para mover los datos a los que se accede con poca frecuencia a niveles de almacenamiento más económicos. Si no se ha definido ninguna política de retención, utilice Amazon S3 Intelligent Tiering para mover los objetos al nivel de archivado automáticamente.
10. Monitorice el uso de los recursos y configure alarmas para activar eventos de escalamiento mediante Amazon CloudWatch.
11. Para cada Cuenta de AWS, utilice AWS Budgets para establecer presupuestos de costes y uso para su cuenta en función del centro de costes y las unidades empresariales.
12. Si usa AWS Budgets para establecer presupuestos de costes y uso para su cuenta, puede resultarle más fácil controlar sus gastos y evitar facturas inesperadas, lo que te permitirá controlar mejor sus costes.

Procedimiento: proporcione procedimientos detallados para implementar esta política o consulte otros documentos en los que se describe cómo implementar cada declaración de la política. En esta sección, se deben proporcionar instrucciones paso a paso para cumplir los requisitos de la política.

Para implementar esta política, puede utilizar diversas herramientas o reglas de AWS Config de terceros para comprobar si se cumple la declaración de la política y activar medidas de corrección automatizadas mediante funciones AWS Lambda. También puede utilizar AWS Organizations para hacer cumplir la política. Además, debe revisar periódicamente el uso de sus recursos y ajustar la política según sea necesario para comprobar que sigue satisfaciendo las necesidades de su empresa.

Pasos para la implementación

- Reúnase con las partes interesadas: para desarrollar políticas, pida a las partes interesadas (oficinas de la empresa en la nube, ingenieros o responsables de la toma de decisiones funcionales para la aplicación de las políticas) de su organización que especifiquen sus requisitos y los documenten. Adopte un enfoque iterativo; para ello, empiece con un enfoque amplio y vaya reduciendo hasta llegar a las unidades más pequeñas en cada paso. Entre los miembros del equipo se encuentran los que tienen un interés directo en la carga de trabajo, como las unidades organizativas o los propietarios de las aplicaciones, además de los grupos de asistencia, como los equipos de seguridad y finanzas.
- Obtenga la confirmación: asegúrese de que los equipos se ponen de acuerdo en las políticas sobre quién puede acceder y desplegar en la Nube de AWS. Asegúrese de que siguen las políticas de su organización y confirme que sus creaciones de recursos se ajustan a las políticas y procedimientos acordados.
- Cree sesiones de formación de incorporación: pida a los nuevos miembros de la organización que completen los cursos de formación de incorporación para crear concienciación sobre los costes y que conozcan los requisitos de la organización. Es posible que asuman políticas diferentes debido a su experiencia anterior o que no piensen en ellas en absoluto.
- Defina las ubicaciones de la carga de trabajo: defina dónde opera su carga de trabajo, incluido el país y la zona dentro del país. Esta información se utiliza para el mapeo de Regiones de AWS y las zonas de disponibilidad.
- Defina y agrupe los servicios y recursos: defina los servicios que requieren las cargas de trabajo. Para cada servicio, especifique los tipos, el tamaño y el número de recursos necesarios. Defina grupos para los recursos por función, como servidores de aplicaciones o almacenamiento de bases de datos. Los recursos pueden pertenecer a varios grupos.
- Defina y agrupe a los usuarios por función: defina a los usuarios que interactúan con la carga de trabajo; para ello, céntrese en lo que hacen y en cómo utilizan la carga de trabajo, no en quiénes son o en su posición en la organización. Agrupe usuarios o funciones similares. Puede utilizar las políticas administradas de AWS como guía.

- Defina las acciones: mediante las ubicaciones, los recursos y los usuarios identificados anteriormente, defina las acciones que requiere cada uno de ellos para lograr los resultados de la carga de trabajo a lo largo de su vida útil (desarrollo, funcionamiento y retirada). Identifique las acciones en función de los grupos, y no de los elementos individuales de los grupos, en cada ubicación. Empiece a grandes rasgos con la lectura o la escritura y, después, vaya reduciendo hasta llegar a las acciones específicas para cada servicio.
- Defina el periodo de revisión: las cargas de trabajo y los requisitos organizativos pueden cambiar con el tiempo. Defina el calendario de revisión de la carga de trabajo para asegurarse de que se mantiene alineado con las prioridades organizativas.
- Documente las políticas: verifique que las políticas que se han definido sean accesibles tal y como lo requiere su organización. Estas políticas se utilizan para implementar, mantener y auditar el acceso de sus entornos.

Recursos

Documentos relacionados:

- Administración de cambios en la nube
- Managed Policies de AWS para funciones de trabajo
- Estrategia de facturación de varias cuentas de AWS
- Acciones, recursos y claves de condición de los servicios de AWS
- Administración y gobernanza en AWS
- Control access to Regiones de AWS using IAM policies (Control del acceso a las regiones de AWS mediante políticas de IAM)
- Zonas de disponibilidad y regiones de infraestructuras globales

Vídeos relacionados:

- AWS Management and Governance at Scale (Administración y gobernanza en AWS a escala)

Ejemplos relacionados:

- VMware - What Are Cloud Policies? (VMware - ¿Qué son las políticas de la nube?)

COST02-BP02 Implementar objetivos y metas

Implemente objetivos de costes y uso para la carga de trabajo. Los objetivos son una guía de resultados previstos para la organización. Las metas proporcionan resultados medibles específicos que se deben alcanzar para las cargas de trabajo.

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

Guía para la implementación

Desarrolle objetivos y metas de costes y uso para su organización. Para una organización en crecimiento en AWS, es importante establecer objetivos de optimización de costes y realizar un seguimiento de ellos. Estos objetivos o indicadores clave de rendimiento (KPI) pueden incluir aspectos como el porcentaje del gasto bajo demanda o la adopción de ciertos servicios optimizados, como instancias de AWS Graviton o tipos de volúmenes gp3 de EBS. Establecer objetivos cuantificables y factibles puede ayudarle a seguir midiendo las mejoras de eficiencia, algo que es importante para las operaciones empresariales en curso. Los objetivos son una guía sobre los resultados esperados para su organización. Las metas proporcionan los resultados medibles que se deben alcanzar. En resumen, un objetivo es la dirección en la que quiere ir y la meta es hasta dónde ir en esa dirección y cuándo debe lograrse ese objetivo (mediante una orientación específica, medible, alcanzable, realista y oportuna, o SMART, por sus siglas en inglés). Un ejemplo de un objetivo es que el uso de la plataforma debería incrementarse de forma significativa con tan solo un ligero incremento (no lineal) del coste. Un ejemplo de meta es un incremento del 20 % del uso de la plataforma con un incremento de menos del 5 % de los costes. Otro objetivo común es que las cargas de trabajo deben ser más eficientes cada seis meses. La meta correspondiente sería que las métricas de coste por empresa disminuyan un 5 % cada seis meses.

Un objetivo de la optimización de costes es aumentar la eficiencia de la carga de trabajo, lo que significa reducir el coste por resultado empresarial de la carga de trabajo con el tiempo. Se recomienda implementar este objetivo para todas las cargas de trabajo y, además, establecer una meta como, por ejemplo, un aumento del 5 % en la eficiencia cada seis meses o un año. Esto se puede conseguir en la nube mediante la creación de capacidades de optimización de costes y el lanzamiento de nuevos servicios y características.

Es importante tener una visibilidad casi en tiempo real de los KPI y las oportunidades de ahorro relacionadas, y hacer un seguimiento del progreso a lo largo del tiempo. Para empezar a definir y hacer un seguimiento de los objetivos de los KPI, recomendamos utilizar el panel de KPI del marco de paneles de inteligencia en la nube (CID). En función de los datos de AWS Cost and Usage Report, el panel de KPI proporciona una serie de KPI de optimización de costes recomendados con

la capacidad de establecer objetivos personalizados y realizar un seguimiento de su progreso a lo largo del tiempo.

Si dispone de otra solución que le permita establecer los objetivos de los KPI y realizar un seguimiento de ellos, asegúrese de que la adopten todas las partes interesadas en la administración financiera de la nube de su organización.

Pasos para la implementación

- Defina los niveles de uso esperados: céntrese primero en los niveles de uso. Interactúe con los propietarios de aplicaciones, los equipos de marketing y otros equipos grandes de la empresa para entender los niveles de uso esperados de la carga de trabajo. ¿Cómo cambiará la demanda de los clientes con el tiempo y habrá cambios debido a los incrementos de temporada o a las campañas de marketing?
- Defina los recursos y los costes de las cargas de trabajo: una vez definidos los niveles de uso, se deben cuantificar los cambios en los recursos de las cargas de trabajo necesarios para ajustarse a dichos niveles de uso. Es posible que tenga que incrementar el tamaño o el número de recursos para un componente de carga de trabajo, incrementar la transferencia de datos o cambiar los componentes de las cargas de trabajo por un servicio distinto en un nivel determinado. Especifique los costes de estos aspectos principales y qué cambios sufrirán los costes si hay cambios en el uso.
- Defina los objetivos empresariales: debe combinar el resultado de los cambios previstos en el uso y los costes con los cambios previstos en la tecnología, o cualquier programa que esté ejecutando, y establecer objetivos para la carga de trabajo. Los objetivos deben tratar sobre el uso y los costes y la relación de ambos. Los objetivos deben ser sencillos y generales. Además, deben ayudar a otras personas a entender lo que espera la empresa en cuanto a los resultados (por ejemplo, asegurarse de que los recursos sin usar estén por debajo de un determinado nivel de coste). No tiene que definir objetivos para cada tipo de recurso no utilizado ni definir costes que provoquen pérdidas para los objetivos y las metas. Verifique que haya programas organizativos (por ejemplo, desarrollo de capacidades a través de cursos de formación) si se prevén cambios en los costes sin cambios en el uso.
- Defina metas: debe especificar una meta cuantificable para cada de uno de los objetivos definidos. Si el objetivo es incrementar la eficiencia de la carga de trabajo, la meta cuantificará la mejora (normalmente en base a los resultados empresariales por cada dólar gastado) y cuándo tendrá lugar. Por ejemplo, si establece el objetivo de minimizar el desperdicio debido al aprovisionamiento excesivo, su objetivo puede ser que dicho desperdicio de computación en el primer nivel de cargas de trabajo de producción no supere el 10 % del coste de computación del nivel y que en el

segundo nivel de cargas de trabajo de producción no sea superior al 5 % del coste de computación del nivel.

Recursos

Documentos relacionados:

- AWS managed policies for job functions (Políticas administradas por AWS para funciones de trabajo)
- AWS multi-account strategy for your AWS Control Tower landing zone (Estrategia multicuenta de AWS para su zona de almacenamiento de AWS Control Tower)
- Control access to Regiones de AWS using IAM policies (Control del acceso a las regiones de AWS mediante políticas de IAM)
- Objetivos SMART

Vídeos relacionados:

- Well-Architected Labs: Goals and Targets (Level 100) (Well-Architected Labs: objetivos y metas [nivel 100])

Ejemplos relacionados:

- Well-Architected Labs: retirada de recursos (objetivos y metas)
- Well-Architected Labs: tipo, tamaño y número de recursos (objetivos y metas)

COST02-BP03 Implementar una estructura de cuentas

Implante una estructura de cuentas adaptada a su organización. Esto ayuda a asignar y administrar los costes en toda la organización.

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

Guía para la implementación

AWS Organizations le permite crear varias Cuentas de AWS que pueden ayudarle a controlar de forma centralizada su entorno a medida que escala sus cargas de trabajo en AWS. Puede modelar su jerarquía organizativa si agrupa las Cuentas de AWS en una estructura de unidades organizativas (OU) y crea varias Cuentas de AWS cada OU. Para crear una estructura de cuentas, primero debe

decidir cuál de sus Cuentas de AWS será la de administración. Después, puede crear Cuentas de AWS nuevas o seleccionar las existentes como cuentas de miembros en función de la estructura de cuentas que haya diseñado según las prácticas recomendadas de cuentas de administración y de miembros.

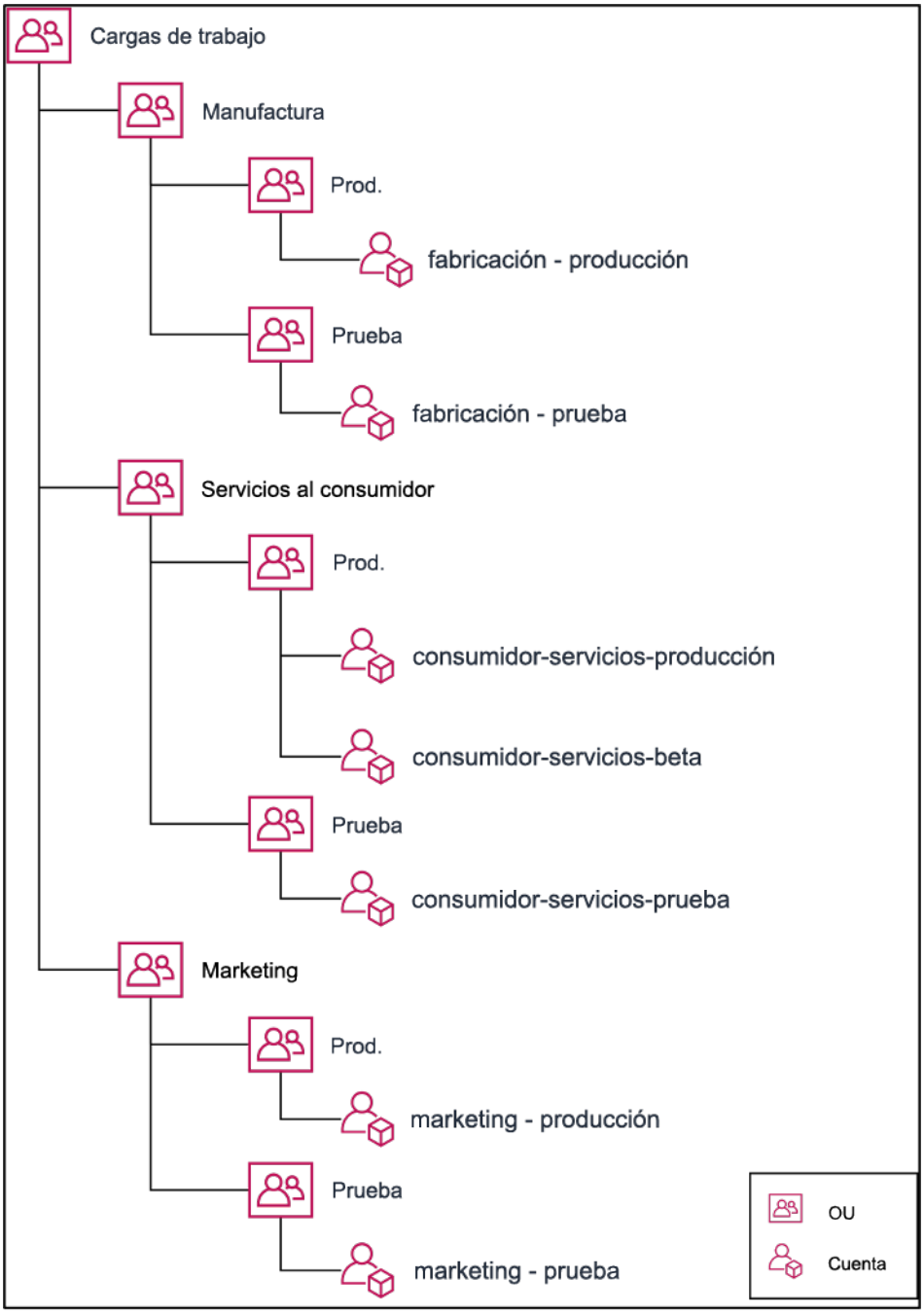
Se aconseja tener siempre al menos una cuenta de administración con una cuenta de miembro vinculada, sin importar el tamaño de la organización o su uso. Los recursos de las cargas de trabajo deberían estar solo en las cuentas de miembro y no se debería crear ningún recurso en la cuenta de administración. En cuanto a la pregunta sobre la cantidad de Cuentas de AWS que se debe tener, no existe una sola respuesta correcta para todas las situaciones. Primero debe evaluar sus modelos operativos y de costes, tanto actuales como futuros, para asegurarse de que la estructura de sus Cuentas de AWS refleje los de su organización. Algunas empresas crean varias cuentas de Cuentas de AWS por motivos empresariales, por ejemplo:

- Se requiere aislamiento administrativo o fiscal y de facturación entre unidades organizativas, centros de costes o cargas de trabajo específicas.
- Los límites de servicio de AWS están configurados para ser específicos para cargas de trabajo concretas.
- Existe un requisito de aislamiento y separación entre cargas de trabajo y recursos.

En AWS Organizations, la facturación unificada crea el componente entre una o más cuentas de miembro y la cuenta de administración. Las cuentas de miembro le permiten aislar y distinguir los costes y el uso por grupos. Una práctica común es tener cuentas de miembro independientes para cada unidad organizativa (como finanzas, marketing y ventas), para cada ciclo de vida del entorno (como desarrollo, prueba y producción) o para cada carga de trabajo (carga de trabajo a, b y c) y luego agregar dichas cuentas vinculadas mediante la facturación unificada.

La facturación unificada le permite unificar el pago de varias Cuentas de AWS de miembro en una sola cuenta de administración y proporcionar a la vez visibilidad de la actividad de cada cuenta vinculada. A medida que se añaden costes y uso a la cuenta de administración, puede maximizar los descuentos de volumen de servicio y el uso de los descuentos por compromiso (Savings Plans e instancias reservadas) para obtener los mayores descuentos.

En el siguiente diagrama se muestra cómo puede utilizar AWS Organizations con unidades organizativas (OU) para agrupar varias cuentas y colocar múltiples Cuentas de AWS en cada OU. Se recomienda utilizar OU para diversos casos de uso y cargas de trabajo, lo que proporciona patrones para organizar las cuentas.



Ejemplo de agrupación de varias Cuentas de AWS en unidades organizativas.

AWS Control Tower puede configurar rápidamente varias cuentas de AWS y garantizar que la gobernanza esté alineada con los requisitos de la organización.

Pasos para la aplicación

- Definir los requisitos de separación: los requisitos de separación son una combinación de múltiples factores, como la seguridad, la fiabilidad y los componentes financieros. Defina cada factor por

orden y especifique si la carga de trabajo o el entorno de la carga de trabajo debería separarse de otras cargas de trabajo. La seguridad promueve la adhesión a los requisitos de acceso y datos. La fiabilidad administra los límites de tal forma que los entornos y las cargas de trabajo no afecten a los demás. Consulte periódicamente los pilares de seguridad y fiabilidad del marco Well-Architected Framework y siga las prácticas recomendadas. Los componentes financieros crean una separación financiera estricta (centro de coste diferente, propietarios de la carga de trabajo y responsabilidad). Los ejemplos comunes de separación son que las cargas de trabajo de producción y prueba se ejecuten en cuentas separadas, o que se use una cuenta separada para que los datos de facturación y de las facturas se puedan proporcionar a las unidades de negocio o departamentos individuales de la organización o parte interesada propietaria de la cuenta.

- Definir requisitos de agrupación: los requisitos de agrupación no anulan los de separación, pero se utilizan para contribuir a la administración. Agrupe entornos o cargas de trabajo similares que no requieran separación. Un ejemplo es agrupar múltiples entornos de prueba o desarrollo de una o varias cargas de trabajo.
- Definir la estructura de cuentas: con estas separaciones y agrupaciones, especifique una cuenta para cada grupo y compruebe que se cumplan los requisitos de separación. Estas cuentas son sus cuentas de miembro o vinculadas. Al agrupar estas cuentas de miembro en una única cuenta de administración o de pagador, combina el uso, lo que le permite disfrutar de descuentos de mayor volumen en todas las cuentas y le proporciona una sola factura para todas las cuentas. No se pueden separar los datos de facturación y proporcionar a cada cuenta de miembro una vista individual de sus datos de facturación. Si una cuenta de miembro no debe tener los datos de facturación o de uso visibles para las demás cuentas, o si se requiere una factura distinta de AWS, defina múltiples cuentas de administración o de pagador. En este caso, cada cuenta de miembro tiene su propia cuenta de administración o de pagador. Los recursos deberían colocarse siempre en las cuentas de miembro o vinculadas. Las cuentas de administración o de pagador solo deben usarse para tareas de administración.

Recursos

Documentos relacionados:

- Uso de etiquetas de asignación de costes
- Políticas administradas de AWS para las funciones del trabajo
- Estrategia de facturación de varias cuentas de AWS
- Controlar el acceso a las Regiones de AWS mediante políticas de IAM
- AWS Control Tower

- AWS Organizations
- Prácticas recomendadas para las cuentas de administración y de miembro
- Organización de su entorno de AWS mediante varias cuentas
- Activación de instancias reservadas compartidas y descuentos de Savings Plans
- Facturación unificada
- Facturación unificada

Ejemplos relacionados:

- División del acceso compartido y CUR

Vídeos relacionados:

- Introducción a AWS Organizations
- Set Up a Multi-Account AWS Environment that Uses Best Practices for AWS Organizations (Configurar un entorno de AWS de varias cuentas que utilice las prácticas recomendadas para AWS Organizations)

Ejemplos relacionados:

- Well-Architected Labs: Create an AWS Organization (Level 100) (Laboratorios de Well-Architected: crear una organización de AWS [nivel 100])
- División del acceso compartido y AWS Cost and Usage Report
- Defining an AWS Multi-Account Strategy for telecommunications companies (Definición de una estrategia de varias cuentas de AWS para empresas de telecomunicaciones)
- Best Practices for Optimizing Cuentas de AWS (Prácticas recomendadas para la optimización de Cuentas de AWS)
- Best Practices for Organizational Units with AWS Organizations (Prácticas recomendadas para unidades organizativas con AWS Organizations)

COST02-BP04 Implementar grupos y roles

Implemente grupos y roles que se ajusten a sus políticas y controle quién puede crear, modificar o retirar instancias y recursos en cada grupo. Por ejemplo, implementar grupos de desarrollo, de pruebas y de producción. Esto se aplica a los servicios de AWS y a las soluciones de terceros.

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

Guía para la implementación

Los roles y los grupos de usuarios son elementos fundamentales en el diseño y la implementación de sistemas seguros y eficaces. Los roles y los grupos ayudan a las organizaciones a equilibrar la necesidad de control con el requisito de flexibilidad y productividad, lo que facilita, en última instancia, los objetivos de la organización y las necesidades de los usuarios. Tal y como se recomienda en la sección «Administración de identidades y accesos» del pilar de seguridad de AWS Well-Architected Framework, necesita tener una administración de identidades y permisos sólida para proporcionar acceso a los recursos correctos a las personas adecuadas en las condiciones adecuadas. Los usuarios reciben solo el acceso necesario para completar sus tareas. Esto minimiza el riesgo asociado con el acceso no autorizado o el uso indebido.

Después de desarrollar las políticas, puede crear grupos lógicos y roles de usuario dentro de la organización. Esto le permite asignar permisos, controlar el uso y ayudar a implementar mecanismos de control de acceso sólidos, lo que evita el acceso no autorizado a la información confidencial. Empiece con grupos de personas de alto nivel. Esto suele corresponderse con unidades organizativas y roles de trabajos (por ejemplo, el administrador de sistemas del departamento de TI, el controlador financiero o los analistas empresariales). Los grupos permiten clasificar a las personas que realizan tareas similares y necesitan accesos similares. Los roles definen lo que debe hacer un grupo. Es más fácil administrar permisos de grupos y roles que permisos de usuarios individuales. Los roles y los grupos asignan permisos de manera uniforme y sistemática a todos los usuarios, lo que evita errores e incoherencias.

Cuando cambia el rol de un usuario, los administradores pueden ajustar el acceso a nivel de rol o grupo, en lugar de volver a configurar cuentas de usuario individuales. Por ejemplo, un administrador de sistemas del departamento de TI requiere acceso para crear todos los recursos, pero un miembro del equipo de análisis solo lo necesita para crear recursos de análisis.

Pasos para la implementación

- Implemente grupos: use los grupos de usuarios definidos en sus políticas organizativas para implementar los grupos correspondientes, si es necesario. Para conocer las prácticas recomendadas sobre usuarios, grupos y autenticación, consulte «Pilar de seguridad: AWS Well-Architected Framework».
- Implemente roles y políticas: use las acciones definidas en sus políticas organizativas para crear los roles y las políticas de acceso necesarios. Para conocer las prácticas recomendadas sobre roles y políticas, consulte «Pilar de seguridad: AWS Well-Architected Framework».

Recursos

Documentos relacionados:

- «Managed Policies de AWS para funciones de trabajo»
- Estrategia de facturación de varias cuentas de AWS
- «Pilar de seguridad: AWS Well-Architected Framework»
- AWS Identity and Access Management (IAM)
- Políticas de AWS Identity and Access Management

Vídeos relacionados:

- «Why use Identity and Access Management»

Ejemplos relacionados:

- Well-Architected Lab Basic Identity and Access (Laboratorio de Well-Architechted: identidad y acceso básicos)
- «Control access to Regiones de AWS using IAM policies»
- «Starting your Cloud Financial Management journey: Cloud cost operations»

COST02-BP05 Implementación de controles de costes

Aplique controles basados en las políticas de la organización y en grupos y roles definidos. De este modo se certifica que los costes solo se producen según los requisitos de la organización, como controlar el acceso a regiones o tipos de recursos.

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

Guía para la implementación

Un primer paso común en la implementación de los controles de costes es establecer notificaciones cuando se producen eventos de costes o de uso fuera de las políticas. Puede actuar con rapidez y verificar si es necesaria una acción correctiva, sin restringir ni afectar negativamente a las cargas de trabajo o a la nueva actividad. Una vez que conozca los límites de la carga de trabajo y del entorno, podrá aplicar la gobernanza. AWS Budgets le permite establecer notificaciones

y definir presupuestos mensuales para sus costes, uso y descuentos por compromiso de AWS (Savings Plans e Instancias reservadas). Puede crear presupuestos en un nivel de coste agregado (por ejemplo, todos los costes), o en un nivel más detallado en el que incluya solo dimensiones específicas como, por ejemplo, cuentas vinculadas, servicios, etiquetas o zonas de disponibilidad.

Una vez que haya establecido los límites de su presupuesto con AWS Budgets, utilice AWS Cost Anomaly Detection para reducir el coste imprevisto. AWS Cost Anomaly Detection es un servicio de administración de costes que utiliza machine learning para supervisar continuamente su coste y uso, así como para detectar gastos inusuales. Le ayuda a identificar los gastos anómalos y las causas que los originan para que pueda adoptar medidas rápidamente. En primer lugar, cree un monitor de costes en AWS Cost Anomaly Detection, y, a continuación, elija su preferencia de alerta mediante el establecimiento de un umbral en dólares (como una alerta sobre anomalías con un impacto superior a 1000 USD). Una vez que reciba las alertas, podrá analizar la causa raíz que provoca la anomalía y el impacto en los costes. También puede supervisar y realizar sus propios análisis de anomalías en AWS Cost Explorer.

Aplique las políticas de gobernanza en AWS mediante AWS Identity and Access Management y las políticas de control de servicios (SCP) de AWS Organizations. IAM le permite administrar de forma segura el acceso a los servicios y recursos de AWS. Mediante IAM, puede controlar quién puede crear o administrar los recursos de AWS, el tipo de recursos que se pueden crear y dónde se pueden crear. Esto minimiza la posibilidad de que se creen recursos fuera de la política definida. Utilice los roles y grupos creados anteriormente y asigne las políticas de IAM para aplicar el uso correcto. La SCP ofrece un control centralizado de los permisos máximos disponibles para todas las cuentas de su organización, lo que mantiene sus cuentas según las directrices de control de acceso. Las SCP están disponibles solo en una organización que tenga todas las características activadas. Puede configurar las SCP para denegar o permitir acciones en las cuentas de los miembros de forma predeterminada. Para obtener más detalles sobre la implementación de la administración del acceso, consulte el documento técnico Pilar de seguridad de Well-Architected.

La gobernanza también puede implementarse a través de la administración de las cuotas de servicio de AWS. Si garantiza que las cuotas de servicio se configuran con los gastos generales mínimos y se mantienen correctamente, puede minimizar la creación de recursos que no necesite su organización. Para lograrlo, debe conocer la velocidad con la que pueden cambiar sus requisitos, comprender los proyectos en curso (tanto la creación como la retirada de recursos) y tener en cuenta la rapidez con la que se pueden implementar los cambios de cuota. Las cuotas de servicio se pueden usar para aumentar las cuotas cuando sea necesario.

Pasos para la aplicación

- Implementar notificaciones sobre el gasto: mediante el uso de las políticas definidas por su organización, cree AWS Budgets para recibir notificaciones cuando el gasto no cumpla las políticas. Configure varios presupuestos de costes, uno para cada cuenta, que le notifiquen el gasto global de la cuenta. Configure presupuestos de costes adicionales en cada cuenta para unidades más pequeñas en ella. Estas unidades varían en función de la estructura de la cuenta. Algunos ejemplos comunes son las Regiones de AWS, las cargas de trabajo (mediante etiquetas) o los servicios de AWS. Configure una lista de distribución de correo electrónico como destinatario de las notificaciones y no una cuenta de correo electrónico individual. Puede configurar un presupuesto real en caso de que se supere una cantidad o utilizar un presupuesto previsto para notificar el uso previsto. También puede preconfigurar acciones presupuestarias de AWS que pueden aplicar políticas de IAM o SCP específicas, o detener las instancias Amazon EC2 y Amazon RDS de destino. Las acciones presupuestarias se pueden ejecutar automáticamente o requerir la aprobación del flujo de trabajo.
- Implementar notificaciones sobre el gasto anómalo: use AWS Cost Anomaly Detection para reducir los costes sorpresa en su organización y analizar la causa raíz del posible gasto anómalo. Una vez que haya creado la supervisión de costes para identificar los gastos inusuales con el detalle que especifique y haya configurado las notificaciones en AWS Cost Anomaly Detection, le enviará una alerta cuando se detecten gastos inusuales. Esto le permitirá analizar el origen de la anomalía y comprender el impacto en el coste. Utilice las categorías de costes de AWS durante la configuración de AWS Cost Anomaly Detection para identificar qué equipo de proyecto o de unidad de negocio puede analizar la causa raíz del coste inesperado y tomar las medidas necesarias a tiempo.
- Implementar controles de uso: mediante las políticas de organización definidas, implemente políticas y roles de IAM para especificar qué acciones pueden realizar los usuarios y cuáles no. En una política de AWS pueden incluirse múltiples políticas organizativas. De la misma manera en que ha definido las políticas, empiece de manera amplia y, a continuación, aplique controles más detallados en cada paso. Los límites de servicio son también un control eficaz del uso. Implemente los límites de servicio correctos en todas las cuentas.

Recursos

Documentos relacionados:

- Políticas administradas de AWS para las funciones del trabajo
- Estrategia de facturación de varias cuentas de AWS
- Controlar el acceso a las Regiones de AWS mediante políticas de IAM

- AWS Budgets
- AWS Cost Anomaly Detection
- Controle los costes de AWS

Vídeos relacionados:

- How can I use AWS Budgets to track my spending and usage (Como puedo usar AWS Budgets para hacer un seguimiento de mis gastos y el uso)

Ejemplos relacionados:

- Políticas de administración de acceso de IAM de ejemplo
- Políticas de control de servicios de ejemplo
- AWS Budgets Actions (Acciones de AWS Budgets)
- Create IAM Policy to control access to Amazon EC2 resources using Tags (Crear una política de IAM para controlar el acceso a los recursos de Amazon EC2 mediante etiquetas)
- Restrict the access of IAM Identity to specific Amazon EC2 resources (Restringir el acceso de la identidad de IAM a recursos de Amazon EC2 específicos)
- Create an IAM Policy to restrict Amazon EC2 usage by family (Crear una política de IAM para restringir el uso de Amazon EC2 por familia)
- Well-Architected Labs: Cost and Usage Governance (Level 100) (Laboratorios de Well-Architected: gobernanza de coste y uso [nivel 100])
- Well-Architected Labs: Cost and Usage Governance (Level 200) (Laboratorios de Well-Architected: gobernanza de coste y uso [nivel 200])
- Slack integrations for Cost Anomaly Detection using Amazon Q Developer in chat applications (Integraciones de Slack para Cost Anomaly Detection mediante Amazon Q Developer in chat applications)

COST02-BP06 Controlar el ciclo de vida de los proyectos

Controle, mida y audite el ciclo de vida de los proyectos, equipos y entornos para evitar el uso y el pago de recursos innecesarios.

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

Guía para la implementación

Un seguimiento eficaz del ciclo de vida de los proyectos permite a las organizaciones controlar mejor los costes gracias a una mejor planificación, administración y optimización de los recursos, el tiempo y la calidad. Los conocimientos que se obtienen con el seguimiento son muy valiosos para tomar decisiones informadas que contribuyen a la rentabilidad y al éxito general del proyecto.

El seguimiento de todo el ciclo de vida de la carga de trabajo le ayuda a comprender cuándo las cargas de trabajo o sus componentes dejan de ser necesarios. Puede que las cargas de trabajo y los componentes existentes parezcan estar en uso, pero cuando AWS publica nuevos servicios o características, estos pueden retirarse o adoptarse. Compruebe las etapas anteriores de las cargas de trabajo. Cuando una carga de trabajo ya no está en producción, los entornos previos se pueden retirar o reducirse en gran medida hasta que se necesiten de nuevo.

AWS proporciona un número de servicios de administración y gobernanza que puede usar para controlar el ciclo de vida de la entidad. Puede utilizar AWS Config o AWS Systems Manager para proporcionar un inventario detallado de sus recursos y configuración de AWS. Se recomienda realizar una integración con sus sistemas de administración de proyectos o recursos existentes para realizar un seguimiento de los proyectos y productos activos en su organización. Mediante la combinación del sistema actual con el amplio conjunto de eventos y métricas que ofrece AWS, podrá crear una vista de eventos importantes del ciclo de vida y administrar de forma proactiva los recursos a fin de reducir costes innecesarios.

Al igual que en la administración del ciclo de vida de la aplicación (ALM), el seguimiento del ciclo de vida de los proyectos debe implicar que varios procesos, herramientas y equipos trabajen juntos (por ejemplo, diseño y desarrollo, pruebas, producción, soporte y redundancia de la carga de trabajo).

Al supervisar cuidadosamente cada fase del ciclo de vida de un proyecto, las organizaciones obtienen información crucial y mejoran el control, lo que facilita la planificación, implementación y finalización exitosas del proyecto. En esta cuidadosa supervisión, se verifica que los proyectos no solo cumplan los estándares de calidad, sino que se entreguen a tiempo y dentro del presupuesto, lo que fomenta el ahorro de costes.

Para obtener más información sobre la implementación del seguimiento del ciclo de vida de las entidades, consulte el documento técnico «Pilar de excelencia operativa: AWS Well-Architected».

Pasos para la implementación

- Establezca un proceso de supervisión del ciclo de vida del proyecto: el equipo del Centro de excelencia en la nube debe establecer un proceso de supervisión del ciclo de vida del proyecto.

Establezca un enfoque estructurado y sistemático para supervisar las cargas de trabajo con el objeto de mejorar el control, la visibilidad y el resultado de los proyectos. Haga que el proceso de supervisión sea transparente y colaborativo y esté centrado en la mejora continua para maximizar su eficacia y valor.

- Realice revisiones de la carga de trabajo: de acuerdo con lo que se haya definido en las políticas de la organización, establezca una regularidad para auditar los proyectos existentes y realizar revisiones de la carga de trabajo. El esfuerzo dedicado a la auditoría debería ser proporcional al riesgo, el valor o el coste aproximados de la organización. Las principales áreas que debería incluir en la auditoría son el riesgo de incidente o interrupción en la organización, el valor o la contribución a la organización (medida en ingresos o reputación de la marca), el coste de la carga de trabajo (medido como coste total de los recursos y costes operativos) y el uso de la carga de trabajo (medido en número de resultados organizativos por unidad de tiempo). Si estas áreas cambian durante el ciclo de vida, se deberá ajustar la carga de trabajo, por ejemplo, mediante una retirada total o parcial.

Recursos

Documentos relacionados:

- «Guidance for Tagging on AWS»
- «¿Qué es la administración del ciclo de vida de las aplicaciones (ALM)?»
- Políticas administradas de AWS para las funciones del trabajo

Ejemplos relacionados:

- Controlar el acceso a las Regiones de AWS mediante políticas de IAM

Herramientas relacionadas:

- «AWS Config»
- «AWS Systems Manager»
- «AWS Budgets»
- «AWS Organizations»
- «AWS CloudFormation»

COSTE 3. ¿Cómo supervisa sus costes y su uso?

Establezca políticas y procedimientos para monitorear y asignar adecuadamente sus costes. Esto le permite medir y mejorar la rentabilidad de esta carga de trabajo.

Prácticas recomendadas

- COST03-BP01 Configurar los orígenes de información detallados
- COST03-BP02 Añadir información de la organización a los costes y el uso
- COST03-BP03 Identificar las categorías de atribución de costes
- COST03-BP04 Establecer métricas de organización
- COST03-BP05 Configurar herramientas de facturación y administración de costes
- COST03-BP06 Asignar costes según las métricas de carga de trabajo

COST03-BP01 Configurar los orígenes de información detallados

Configure las herramientas de administración de costes y generación de informes para ofrecer una especificidad por horas a fin de proporcionar información detallada sobre los costes y el uso, lo que permitirá un análisis y una transparencia más exhaustivos. Configure la carga para generar o disponer de entradas de registro para cada resultado empresarial entregado.

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

Guía para la implementación

La información de facturación detallada, como la especificidad por horas en las herramientas de administración de costes, permite a las organizaciones realizar un seguimiento más detallado de su consumo y les ayuda a identificar algunos de los motivos del aumento de costes. Estos orígenes de datos proporcionan la visión más veraz del coste y uso en toda la organización.

AWS proporciona una especificidad de uso diario o por horas, tarifas, costes y atributos de uso de todos los servicios de AWS Cost and Usage Report de pago. Todas las dimensiones posibles están en el CUR, por ejemplo: etiquetado, ubicación, atributos de recursos e ID de cuentas.

Configure el CUR con las siguientes personalizaciones:

- Inclusión de los ID de recurso
- Actualización automática del CUR

- Especificidad por horas
- Control de versiones: sobrescritura del informe existente
- Integración de datos: Athena (formato y compresión Parquet)

Utilice AWS Glue para preparar los datos para el análisis y emplee Amazon Athena para analizar los datos y SQL para su consulta. También puede utilizar Quick para crear visualizaciones personalizadas y complejas, y distribuirlas por toda la organización.

Pasos para la implementación

- Configurar el informe de coste y uso: configure al menos un informe de coste y uso con la consola de facturación. Configure un informe detallado por horas que incluya todos los identificadores y los ID de recurso. También puede crear otros informes con distintos niveles de especificidad para proporcionar información resumida de nivel superior.
- Configurar la especificidad por horas en Cost Explorer: habilite Por hora y Datos del nivel de recurso para acceder a los datos sobre coste y uso con especificidad por horas de los últimos 14 días y especificidad a nivel de recursos.
- Configurar el registro de aplicaciones: verifique que su aplicación registra cada resultado empresarial que ofrece para que se pueda hacer el seguimiento y la medición. Asegúrese de que la especificidad de estos datos es, como mínimo, por horas, para que coincidan con los datos de coste y uso. Para obtener más información sobre el registro y la supervisión, consulte Pilar de excelencia operativa de Well-Architected.

Recursos

Documentos relacionados:

- AWS Cost and Usage Report
- AWS Glue
- Quick
- AWS Cost Management Pricing
- Tagging AWS resources (Etiquetado de recursos de AWS)
- Analyzing your costs with AWS Budgets
- Analyzing your costs with Cost Explorer (Análisis de los costes con Cost Explorer)
- Managing AWS Cost and Usage Reports

- Pilar de excelencia operativa de Well-Architected

Ejemplos relacionados:

- AWS Account Setup
- AWS Cost Explorer's New Look and Common Use Cases

COST03-BP02 Añadir información de la organización a los costes y el uso

Defina un esquema de etiquetado basado en su organización, los atributos de carga de trabajo y las categorías de asignación de costes para poder filtrar y buscar recursos o supervisar el coste y el uso en las herramientas de administración de costes. Implemente un etiquetado coherente en todos los recursos, siempre que sea posible, por finalidad, equipo, entorno u otros criterios relevantes para su empresa.

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

Guía para la implementación

Implemente el etiquetado en AWS para añadir información de la organización a los recursos que, a su vez, se añadirá a la información de costes y uso. Una etiqueta es un par clave-valor: la clave está definida y debe ser única en toda la organización, mientras que el valor es único para un grupo de recursos. Un ejemplo de par clave-valor es la clave `Entorno` con un valor de `Producción`. Todos los recursos del entorno de producción tendrán este par clave-valor. El etiquetado le permite categorizar y controlar los costes con información de la organización relevante y útil. Puede aplicar etiquetas que representen categorías de la organización (como centros de costes, nombres de aplicación, proyectos o propietarios) e identificar cargas de trabajo y características de cargas de trabajo (como de prueba o producción) para categorizar sus costes y uso en toda la organización.

Cuando aplica etiquetas a sus recursos de AWS (como instancias Amazon Elastic Compute Cloud o buckets de Amazon Simple Storage Service) y las activa, AWS añade esta información a los informes de uso y costes. Puede ejecutar informes y realizar análisis en recursos con etiquetas o sin ellas para permitir un mayor cumplimiento de las políticas de administración de costes internos y garantizar una atribución precisa.

Con la creación e implementación de un estándar de etiquetado de AWS en las cuentas de su organización, podrá administrar y controlar sus entornos de AWS de manera coherente y uniforme. Use políticas de etiquetado en AWS Organizations para definir reglas sobre cómo se pueden usar

los recursos de AWS en sus cuentas de AWS Organizations. Las políticas de etiquetado le permiten adoptar un enfoque estandarizado para los recursos de etiquetado de AWS.

El editor de etiquetas de AWS le permite añadir, eliminar y administrar etiquetas de múltiples recursos. Con el editor de etiquetas, puede buscar los recursos que desea etiquetar y, a continuación, administrar las etiquetas de los recursos que aparecen en los resultados de la búsqueda.

Las categorías de costes de AWS le permiten asignar un significado de organización a los costes, sin necesitar etiquetas en los recursos. Puede asignar la información de costes y uso a estructuras organizativas internas únicas. Debe definir reglas de categorías para asignar y categorizar los costes mediante dimensiones de facturación, como cuentas y etiquetas. Esto proporciona otro nivel de capacidad de administración, además del etiquetado. También puede asignar cuentas específicas y etiquetas a varios proyectos.

Pasos para la aplicación

- Definir un esquema de etiquetado: reúna a todas las partes interesadas de la empresa para definir un esquema. En general, son personas con roles técnicos, financieros o de administración. Defina una lista de etiquetas que deben tener todos los recursos, así como una lista de las etiquetas que deberían tener los recursos. Compruebe que los nombres y los valores de las etiquetas sean coherentes en toda la organización.
- Etiquetar recursos: con las categorías de atributos de costes definidas, coloque etiquetas en todos los recursos en sus cargas de trabajo según las categorías. Use herramientas como la CLI, el editor de etiquetas o AWS Systems Manager para incrementar la eficiencia.
- Implementar categorías de costes de AWS: puede crear categorías de costes sin implementar el etiquetado. Las categorías de costes usan las dimensiones de costes y uso existentes. Cree reglas de categorías a partir de su esquema e impleméntelas en las categorías de costes.
- Automatizar el etiquetado: para comprobar que mantiene altos niveles de etiquetado en todos los recursos, automatice el etiquetado para que los recursos reciban etiquetas automáticamente en cuanto se creen. Use servicios como AWS CloudFormation para verificar que los recursos se etiquetan al crearse. También puede crear una solución personalizada para etiquetar automáticamente con funciones de Lambda o use un microservicio personalizado que escanee la carga de trabajo periódicamente y elimine cualquier recurso que no tenga etiqueta, lo que es ideal para los entornos de prueba y desarrollo.
- Supervisar las etiquetas y elaborar informes de ellas: para comprobar que mantiene altos niveles de etiquetado en toda la organización, elabore informes de las etiquetas de sus cargas de trabajo

y supervise dichas etiquetas. Puede usar AWS Cost Explorer para ver el coste de los recursos etiquetados o no, o bien usar servicios tales como el editor de etiquetas. Revise periódicamente el número de recursos no etiquetados y añada etiquetas hasta alcanzar el nivel de etiquetado que desee.

Recursos

Documentos relacionados:

- Prácticas recomendadas sobre etiquetado
- Etiqueta de recurso de AWS CloudFormation
- Categorías de costes de AWS
- Etiquetado de recursos de AWS
- Análisis de los costes con AWS Budgets
- Análisis de los costes con Cost Explorer
- Administración de los informes de coste y uso de AWS

Vídeos relacionados:

- How can I tag my AWS resources to divide up my bill by cost center or project (Cómo puedo etiquetar mis recursos de AWS para dividir mi factura por centro de coste o proyecto)
- Tagging AWS Resources (Etiquetado de recursos de AWS)

Ejemplos relacionados:

- Automatically tag new AWS resources based on identity or role (Etiquetar automáticamente los nuevos recursos de AWS a partir de la identidad o el rol)

COST03-BP03 Identificar las categorías de atribución de costes

Identifique las categorías de la organización como las unidades empresariales, los departamentos o los proyectos que podrían utilizarse para asignar los costes dentro de su organización a las entidades consumidoras internas. Utilice esas categorías para imponer la responsabilidad del gasto, crear concienciación sobre los costes y fomentar comportamientos de consumo eficaces.

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

Guía para la implementación

El proceso de categorización de los costes es crucial en la elaboración de presupuestos, la contabilidad, los informes financieros, la toma de decisiones, las evaluaciones comparativas y la administración de proyectos. Al clasificar y categorizar los gastos, los equipos pueden comprender mejor los tipos de costes en los que incurrirán durante su traspaso a la nube, lo que les ayuda a tomar decisiones fundamentadas y a administrar los presupuestos de manera eficaz.

La responsabilidad de los gastos en la nube es un buen incentivo para conseguir una administración disciplinada de la demanda y los costes. Como resultado, las organizaciones que destinan la mayor parte de su gasto en la nube a unidades empresariales o equipos que consumen recursos ahorran mucho más en costes en la nube. Además, la asignación del gasto en la nube ayuda a las organizaciones a adoptar más prácticas recomendadas de gobernanza centralizada en la nube.

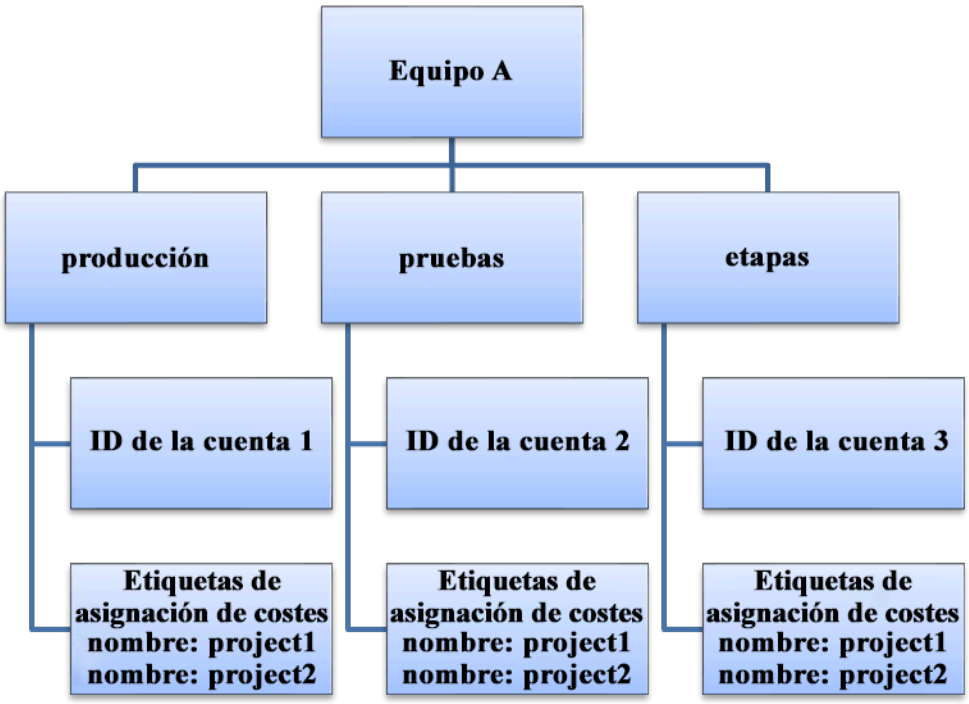
Colabore con el equipo financiero y otras partes interesadas pertinentes para comprender los requisitos sobre cómo deben asignarse los costes en la organización durante las llamadas de cadencia periódicas. Los costes de carga de trabajo deben asignarse a todo el ciclo de vida, como las fases de desarrollo, pruebas, producción y retirada. Debe saber qué costes de la organización proceden de la formación, el desarrollo del personal y la creación de ideas. Puede ser útil para asignar correctamente las cuentas que se usan para los presupuestos de formación y desarrollo, en lugar de presupuestos genéricos de costes de TI.

Tras definir las categorías de atribución de costes con las partes interesadas de la organización, utilice las AWS Cost Categories (Categorías de costes de AWS) para agrupar la información de costes y uso en categorías significativas en la Nube de AWS, como el coste de un proyecto específico o las Cuentas de AWS de departamentos o unidades empresariales. Puede crear categorías personalizadas y asignar su información de costes y uso a estas categorías en función de las reglas que defina mediante varias dimensiones, tales como: cuenta, etiqueta, servicio o tipo de cargo. Tras configurar las categorías de costes, puede ver la información de costes y uso por estas categorías, lo que permite a la organización tomar mejores decisiones estratégicas y de compra. Estas categorías también están visibles en AWS Cost Explorer, AWS Budgets y AWS Cost and Usage Report.

Por ejemplo, cree categorías de costes para sus unidades empresariales (equipo de DevOps), y en cada categoría cree varias reglas (para cada subcategoría) con múltiples dimensiones (Cuentas de AWS, etiquetas de asignación de costes, servicios o tipo de cargo) basadas en las agrupaciones definidas. Con las categorías de costes, puede organizar sus costes mediante un motor basado en reglas. Las reglas que configure organizan sus costes en categorías. En estas reglas, puede realizar el filtrado con varias dimensiones para cada categoría, como Cuentas de AWS, servicios de AWS

o tipos de cargos específicos. Después, podrá utilizar estas categorías en varios productos en la consola de Administración de facturación y costos de AWS and Cost Management .. Esto incluye AWS Cost Explorer, AWS Budgets, AWS Cost and Usage Report y AWS Cost Anomaly Detection.

En el siguiente diagrama se muestra, a modo de ejemplo, cómo agrupar la información de costes y uso de la organización si tiene varios equipos (categoría de costes) con varios entornos (reglas) y cada entorno tiene varios recursos o activos (dimensiones).



Organigrama de costes y uso

También puede crear agrupaciones de costes mediante categorías de costes. Después de crear las categorías de costes (deje que transcurran hasta 24 horas desde la creación de una categoría de costes para que sus registros de uso se actualicen con valores), estas aparecen en AWS Cost Explorer, AWS Budgets, AWS Cost and Usage Reporty AWS Cost Anomaly Detection. En AWS Cost Explorer y AWS Budgets, una categoría de coste aparece como una dimensión de facturación adicional. Puede utilizarla para filtrar por el valor específico de la categoría de costes o agrupar por dicha categoría.

Pasos para la implementación

- Defina las categorías de la organización: reúnase con las partes interesadas internas y las unidades de negocio para definir las categorías que reflejen la estructura y los requisitos de su organización. Estas categorías deben reflejar directamente la estructura de las categorías

financieras existentes, como unidad empresarial, presupuestaria, centro de costes o departamento. Consulte los resultados de la nube para su empresa, como la formación o la educación, pues también son categorías de la organización.

- Defina las categorías funcionales: reúnase con las partes interesadas internas y las unidades de negocio para definir las categorías que reflejen las funciones de su empresa. Pueden ser los nombres de las aplicaciones o las cargas de trabajo y el tipo de entorno, como producción, pruebas o desarrollo.
- Defina categorías de costes de AWS: cree categorías de costes para organizar la información de costes y uso mediante AWS Cost Categories (Categorías de costes de AWS) y asigne su coste y uso de AWS a categorías significativas. Se pueden asignar varias categorías a un recurso y un recurso puede estar en muchas categorías distintas, por lo que se recomienda definir tantas categorías como sea necesario para que pueda administrar sus costes en la estructura categorizada mediante categorías de costes de AWS.

Recursos

Documentos relacionados:

- Tagging AWS resources (Etiquetado de recursos de AWS)
- Uso de las etiquetas de asignación de costos
- Analyzing your costs with AWS Budgets
- Analyzing your costs with Cost Explorer (Análisis de los costes con Cost Explorer)
- Administración de instancias de AWS Cost and Usage Report
- AWS Cost Categories (Categorías de costes de AWS)
- Administración de costos con AWS Cost Categories
- Creación de categorías de costes
- Etiquetado de categorías de costes
- División de cargos en categorías de costes
- AWS Cost Categories Features (Características de las categorías de costes de AWS)

Ejemplos relacionados:

- Organize your cost and usage data with AWS Cost Categories (Organice sus datos de costes y uso con las categorías de costes de AWS)

- Administración de costos con AWS Cost Categories
- Well-Architected Labs: visualización de costes y uso
- Well-Architected Labs: categorías de costes

COST03-BP04 Establecer métricas de organización

Establezca las métricas de organización necesarias para esta carga de trabajo. Algunos ejemplos de métricas de cargas de trabajo son los informes de clientes producidos o las páginas web que se entregan a los clientes.

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

Guía para la implementación

Comprenda cómo se mide el rendimiento de su carga de trabajo en relación con el éxito empresarial. Cada carga de trabajo suele tener un pequeño conjunto de resultados principales que indican el rendimiento. Si tiene una carga de trabajo compleja con muchos componentes, puede priorizar la lista o definir y hacer un seguimiento de las métricas de cada componente. Colabore con sus equipos para entender qué métricas utilizar. Esta unidad se usará para comprender la eficiencia de la carga de trabajo o el coste de cada resultado empresarial.

Pasos para la aplicación

- Definir los resultados de la carga de trabajo: reúnase con las partes interesadas de la empresa y defina los resultados de la carga de trabajo. Son una medida principal del uso de los clientes y deben ser métricas empresariales y no técnicas. Debe haber un pequeño número de métricas generales (menos de cinco) por carga de trabajo. Si la carga de trabajo produce varios resultados para diferentes casos de uso, agrúpelos en una sola métrica.
- Definir los resultados de los componentes de la carga de trabajo: de manera opcional, si tiene una carga de trabajo grande y compleja, o puede dividir fácilmente su carga de trabajo en componentes (como microservicios) con entradas y salidas bien definidas, establezca métricas para cada componente. El esfuerzo debe reflejar el valor y el coste del componente. Empiece por los componentes más grandes y continúe con los más pequeños.

Recursos

Documentos relacionados:

- Etiquetado de recursos de AWS

- Análisis de los costes con AWS Budgets
- Análisis de los costes con Cost Explorer
- Administración de los informes de coste y uso de AWS

COST03-BP05 Configurar herramientas de facturación y administración de costes

Configure las herramientas de administración de costes de acuerdo con las políticas de su organización para administrar y optimizar el gasto en la nube. Esto incluye servicios, herramientas y recursos para organizar y hacer un seguimiento de los datos de costes y uso, mejorar el control mediante una facturación consolidada y permisos de acceso, mejorar la planificación mediante presupuestos y previsiones, recibir notificaciones o alertas y reducir aún más los costes con optimizaciones de recursos y precios.

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

Guía para la implementación

Para establecer una responsabilidad sólida, lo primero que hay que hacer es tener en cuenta la estrategia de cuentas en la estrategia de asignación de costes. Si lo hace bien, es posible que no necesite nada más. En caso contrario, puede haber desconocimiento y problemas adicionales.

Para fomentar la responsabilidad del gasto en la nube, los usuarios deben tener acceso a herramientas que proporcionen visibilidad sobre sus costes y su uso. Se recomienda que todas las cargas de trabajo y equipos tengan configuradas las herramientas para lo siguiente:

- Organización: establezca su base de referencia de asignación de costes y gobernanza con su propia estrategia de etiquetado y categorizaciones.
- Organización: establezca su base de referencia de asignación de costes y gobernanza con su propia estrategia de etiquetado y taxonomía. Etiquete los recursos que admite AWS y clasifíquelos de manera significativa en función de la estructura de su organización (unidades empresariales, departamentos o proyectos). Etiquete los nombres de las cuentas para centros de costes específicos y asígnelos a las categorías de costes de AWS para agrupar las cuentas de determinadas unidades empresariales con sus centros de costes, de modo que el propietario de la unidad empresarial pueda ver el consumo de varias cuentas en un solo lugar.
- Acceso: realice un seguimiento de la información de facturación de toda la organización en una facturación unificada y verifique que las partes interesadas y los propietarios empresariales adecuados tengan acceso a ella.

- Control: cree mecanismos de gobernanza efectivos con las barreras de protección adecuadas para evitar escenarios inesperados cuando se utilicen las políticas de control de servicios (SCP), las políticas de etiquetas y las alertas de presupuestos. Por ejemplo, con mecanismos de control eficaces puede permitir que los equipos creen recursos solo en regiones preferidas.
- Estado actual: configure un panel que muestre los niveles actuales de coste y uso. El panel debe estar disponible en un lugar muy visible en el entorno de trabajo, de forma similar a un panel de operaciones. Puede usar el panel de inteligencia en la nube (CID) o cualquier otro producto admitido para conseguir esta visibilidad.
- Notificaciones: proporcione notificaciones cuando el coste o el uso sobrepasen los límites definidos y cuando se produzcan anomalías con AWS Budgets o AWS Cost Anomaly Detection.
- Informes: resuma toda la información sobre costes y uso y aumente la concienciación y la responsabilidad de su gasto en la nube con datos de costes detallados y atribuibles. Los informes deben ser relevantes para el equipo que los consume y lo ideal es que contengan recomendaciones.
- Seguimiento: muestre el coste y uso actuales con respecto a los objetivos o las metas configurados.
- Análisis: permita que los miembros del equipo realicen análisis personalizados y exhaustivos con un nivel de detalle por hora, con todas las dimensiones posibles.
- Inspección: manténgase al día de sus oportunidades de despliegue de recursos y optimización de costes. Reciba notificaciones (con Amazon CloudWatch, Amazon SNS o Amazon SES) sobre los despliegues de recursos en el nivel de la organización y revise las recomendaciones de optimización de costes (por ejemplo, AWS Compute Optimizer o AWS Trusted Advisor).
- Tendencias: muestre la variabilidad del coste y uso durante el periodo de tiempo requerido con el nivel de detalle necesario.
- Previsiones: muestre los costes futuros estimados y calcule el uso de sus recursos y el gasto con paneles de previsión creados por usted.

Puede utilizar herramientas de AWS como AWS Cost Explorer, Administración de facturación y costos de AWSo AWS Budgets para lo esencial, o puede integrar datos de CUR con Amazon Athena y Quick para tener esta capacidad de obtener vistas más detalladas. Si no tiene las habilidades o el ancho de banda esenciales en su organización, puede trabajar con AWS ProServ, AWS Managed Services (AMS)o AWS Partners y usar sus herramientas. También puede utilizar herramientas de terceros, pero verifique primero que el coste aporta valor a su organización.

Pasos para la implementación

- Permita el acceso basado en equipos a las herramientas: configure sus cuentas y cree grupos que tengan acceso a los informes de costes y uso necesarios para sus consumos, y use AWS Identity and Access Management para controlar el acceso a herramientas como AWS Cost Explorer. Estos grupos deben incluir a representantes de todos los equipos que poseen o administran una aplicación. De este modo, se certifica que cada equipo tiene acceso a su información de costes y uso para realizar el seguimiento de su consumo.
- Configure AWS Budgets: Configure AWS Budgets en todas las cuentas de su carga de trabajo. Establezca presupuestos para el gasto general de la cuenta y presupuestos para la carga de trabajo con etiquetas. Configure las notificaciones en AWS Budgets para recibir alertas cuando supere los importes presupuestados o cuando los costes estimados superen sus presupuestos.
- Configure AWS Cost Explorer: Configure AWS Cost Explorer para su carga de trabajo y cuentas para visualizar los datos de costes y realizar un análisis posterior. Cree un panel para la carga de trabajo que realice un seguimiento del gasto general, las métricas clave de uso de la carga de trabajo y la previsión de los costes futuros a partir de sus datos históricos de costes.
- Configure AWS Cost Anomaly Detection: use AWS Cost Anomaly Detection para sus cuentas, servicios básicos o categorías de costes que haya creado para supervisar el coste y el uso, y detectar gastos fuera de lo habitual. Puede recibir las alertas individualmente en informes agregados y en un correo electrónico o un tema de Amazon SNS que le permita analizar y determinar la causa principal de la anomalía, e identificar el factor que está provocando el aumento de los costes.
- Configure herramientas avanzadas: de forma opcional, puede crear herramientas personalizadas para su organización que proporcionen información y detalles adicionales. Puede implementar la capacidad de análisis avanzado mediante Amazon Athenay paneles con Quick. Considere la posibilidad de utilizar la solución CID, que cuenta con paneles avanzados preconfigurados. También hay AWS Partners con los que puede trabajar y adoptar sus soluciones de administración de la nube para habilitar la monitorización y optimización de la facturación en la nube en una ubicación única y práctica.

Recursos

Documentos relacionados:

- Administración de costes de AWS
- Etiquetado Recursos de AWS

- Analyzing your costs with AWS Budgets
- Analyzing your costs with Cost Explorer (Análisis de los costes con Cost Explorer)
- Managing AWS Cost and Usage Report
- AWS Cost Categories (Categorías de costes de AWS)
- Administración financiera en la nube con AWS
- Políticas de control de servicios de ejemplo
- AWS APN Partners – Cost Management

Vídeos relacionados:

- Deploying Cloud Intelligence Dashboards (Despliegue de paneles de inteligencia en la nube)
- Get Alerts on any FinOps or Cost Optimization Metric or KPI (Recibir alertas sobre cualquier métrica o KPI de FinOps o de optimización de costes)

Ejemplos relacionados:

- Well-Architected Labs - AWS Account Setup (Configuración de la cuenta de AWS)
- Well-Architected Labs: visualización de facturación
- Well-Architected Labs: coste y uso de la gobernabilidad
- Well-Architected Labs: análisis de costes y uso
- Well-Architected Labs: visualización de costes y uso
- Well-Architected Labs: paneles de inteligencia en la nube
- Cómo utilizar las SCP para establecer barreras de protección en todas las cuentas

COST03-BP06 Asignar costes según las métricas de carga de trabajo

Asigne los costes de la carga de trabajo por métricas de uso o resultados empresariales para medir la eficiencia de los costes. Implemente un proceso para analizar los datos de costes y uso con servicios de análisis que pueden proporcionar información y capacidad de recuperación.

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

Guía para la implementación

La optimización de costes significa aportar resultados empresariales al menor precio, lo que solo se puede conseguir asignando los costes de la carga de trabajo por métricas de carga de trabajo (medidas por eficiencia de la carga de trabajo). Controle las métricas de carga de trabajo definidas mediante archivos de registro u otro tipo de monitorización de la aplicación. Combine estos datos con los costes de la carga de trabajo, que pueden obtenerse consultando los costes que tienen un valor de etiqueta o identificador de cuenta específicos. Se recomienda realizar este análisis a nivel de hora. Por lo general, su eficiencia cambiará si tiene algunos componentes de costes estáticos (por ejemplo, una base de datos backend que se ejecuta permanentemente) con un índice de solicitudes variable (por ejemplo, el uso alcanza su punto máximo entre las nueve de la mañana y las cinco de la tarde, pero hay pocas solicitudes por la noche). Comprender la relación entre los costes variables y fijos le ayudará a centrar sus actividades de optimización.

Crear métricas de carga de trabajo para los recursos compartidos puede resultar un desafío en comparación con recursos como las aplicaciones en contenedores en Amazon Elastic Container Service (Amazon ECS) y Amazon API Gateway. Sin embargo, hay ciertas formas de clasificar el uso y realizar un seguimiento de los costes. Si necesita realizar un seguimiento de los recursos compartidos de Amazon ECS y AWS Batch, puede habilitar los datos de asignación de costes divididos en AWS Cost Explorer. Al dividir los datos de asignación de costes, puede comprender y optimizar el coste y el uso de sus aplicaciones en contenedores y volver a asignar los costes de las aplicaciones a entidades empresariales individuales en función de cómo se consumen los recursos compartidos de computación y memoria. Si tiene un uso compartido de las funciones API Gateway y AWS Lambda, puede usar AWS Application Cost Profiler para clasificar su consumo en función de su `identificador de inquilino` o bien `ID de cliente`.

Pasos para la implementación

- Asigne costes a las métricas de carga de trabajo: use las métricas definidas y las etiquetas configuradas, y cree una métrica que combine el resultado de la carga de trabajo y el coste de la carga de trabajo. Use servicios de análisis como Amazon Athena y Amazon Quick para crear un panel de eficiencia para la carga de trabajo global y para cualquier otro componente.

Recursos

Documentos relacionados:

- Tagging AWS resources (Etiquetado de recursos de AWS)

- Analyzing your costs with AWS Budgets (Análisis de los costes con AWS Budgets)
- Analyzing your costs with Cost Explorer (Análisis de los costes con Cost Explorer)
- Managing AWS Cost and Usage Reports (Administración de los informes de coste y uso de AWS)

Ejemplos relacionados:

- Improve cost visibility of Amazon ECS and AWS Batch with AWS Split Cost Allocation Data (Mejore la visibilidad de los costes de Amazon ECS y AWS Batch con datos de asignación de costes divididos de AWS)

COSTE 4. ¿Cómo retira los recursos?

Implemente control de cambios y administración de recursos desde el inicio del proyecto hasta su finalización. Esto garantiza el cierre o la terminación de recursos no utilizados para reducir el desperdicio.

Prácticas recomendadas

- COST04-BP01 Seguimiento de los recursos a lo largo de su ciclo de vida
- COST04-BP02 Implementar un proceso de retirada
- COST04-BP03 Retirar recursos
- COST04-BP04 Retirar los recursos automáticamente
- COST04-BP05 Aplicación de políticas de retención de datos

COST04-BP01 Seguimiento de los recursos a lo largo de su ciclo de vida

Defina e implemente un método para hacer un seguimiento de los recursos y sus asociaciones con los sistemas a lo largo de su ciclo de vida. Puede usar etiquetas para identificar la carga de trabajo o la función del recurso.

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

Guía para la implementación

Retire los recursos de la carga de trabajo que ya no necesite. Por ejemplo, después de hacer pruebas, los recursos empleados ya no se necesitan y se pueden retirar. El seguimiento de los recursos con etiquetas (y ejecutar informes de dichas etiquetas) puede ayudarle a identificar los elementos que se deben retirar, ya que no estarán en uso o caducará su licencia. Usar etiquetas

es una forma efectiva de hacer un seguimiento de los recursos. Se puede etiquetar el recurso con su función o una fecha conocida en la que se puede retirar. Puede ejecutar informes de estas etiquetas. Un valor de ejemplo del etiquetado de características es `feature-X testing` para identificar el propósito del recurso en términos de ciclo de vida de la carga de trabajo. Otro ejemplo es usar `LifeSpan` o `TTL` para los recursos, como el nombre y el valor de la clave de etiqueta que se eliminará para definir el periodo de tiempo o el momento específico para la retirada.

Pasos para la aplicación

- Implementar un esquema de etiquetado: implemente un esquema de etiquetado que identifique la carga de trabajo a la que pertenece el recurso y compruebe que todos los recursos de la carga de trabajo estén etiquetados en consonancia. El etiquetado le ayuda a categorizar los recursos por finalidad, equipo, entorno u otros criterios pertinentes para su empresa. Para obtener más detalle sobre el etiquetado de casos de uso, estrategias y técnicas, consulte AWS Tagging Best Practices (Prácticas recomendadas de etiquetado de AWS).
- Implementar la supervisión del rendimiento de la carga de trabajo o de los resultados: implemente la supervisión del rendimiento de la carga de trabajo o las alarmas que desencadenen solicitudes de entrada o finalizaciones de salida. Configúrela para que proporcione notificaciones cuando las solicitudes de carga de trabajo o los resultados lleguen a cero, lo que significa que ya no se usan los recursos de la carga de trabajo. Incorpore un factor de tiempo si la carga de trabajo baja a cero de forma periódica en condiciones normales. Para obtener más detalles sobre los recursos no utilizados o infrautilizados, consulte AWS Trusted Advisor Cost Optimization checks (Comprobaciones de optimización de costes de AWS Trusted Advisor).
- Agrupar recursos de AWS: cree grupos para recursos de AWS. Puede utilizar Grupos de recursos de AWS para organizar y administrar sus recursos de AWS que se encuentran en la misma Región de AWS. Puede añadir etiquetas a la mayoría de sus recursos como ayuda para identificarlos y clasificarlos en su organización. Utilice Editor de etiquetas para añadir etiquetas a los recursos admitidos en bloque. Considere la posibilidad de utilizar AWS Service Catalog para crear, administrar y distribuir carteras de productos aprobados a los usuarios finales y administrar el ciclo de vida de los productos.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Trusted Advisor

- AWS Trusted Advisor Cost Optimization Checks (Comprobaciones de optimización de costes de AWS Trusted Advisor)
- Etiquetado de recursos de AWS
- Publicar métricas personalizadas

Vídeos relacionados:

- How to optimize costs using AWS Trusted Advisor (Cómo optimizar los costes mediante AWS Trusted Advisor)

Ejemplos relacionados:

- Organize AWS resources (Organizar recursos de AWS)
- Optimize cost using AWS Trusted Advisor (Optimizar el coste mediante AWS Trusted Advisor)

COST04-BP02 Implementar un proceso de retirada

Implemente un proceso para identificar y retirar los recursos sin usar.

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

Guía para la implementación

Implemente un proceso estandarizado en toda la organización para identificar y eliminar los recursos que no se utilizan. El proceso debería definir la frecuencia con que se realizan las búsquedas y los procesos para retirar el recurso para verificar que se cumplan todos los requisitos de la organización.

Pasos para la aplicación

- Crear e implementar un proceso de retirada: trabaje con los desarrolladores y propietarios de las cargas de trabajo para diseñar un proceso de retirada de la carga de trabajo y sus recursos. El proceso debería incluir un método para verificar si se usa la carga de trabajo y también si se usa cada recurso de la carga de trabajo. Detalle los pasos necesarios para retirar el recurso del servicio garantizando el cumplimiento de cualquier requisito normativo. Se debe incluir cualquier recurso asociado, como licencias o almacenamiento asociado. Notifique a los propietarios de las cargas de trabajo que se ha iniciado el proceso de retirada.

  Siga estos pasos de retirada como guía sobre lo que se debe comprobar como parte del proceso:

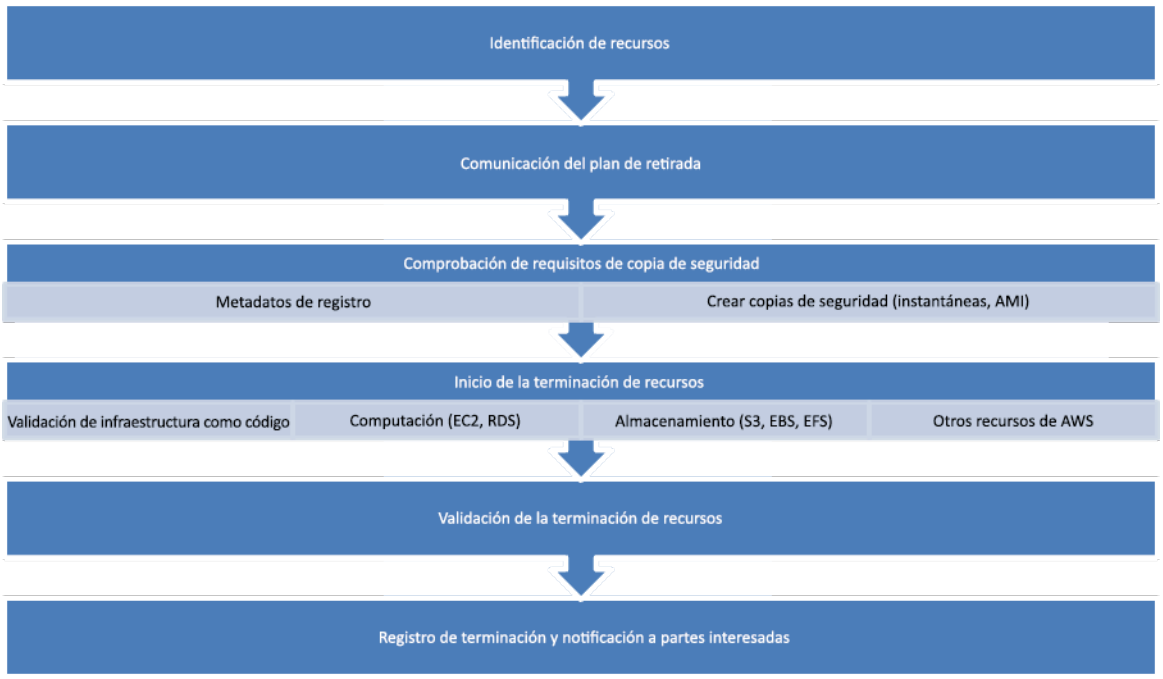
- Identificar los recursos que deben retirarse del servicio: identifique los recursos candidatos para retirarse en su Nube de AWS. Registre toda la información necesaria y programe la retirada. En su cronología, asegúrese de tener en cuenta si surgen (y cuándo surgen) problemas inesperados durante el proceso.
- Coordinar y comunicar: colabore con los propietarios de las cargas de trabajo para confirmar el recurso que se va a retirar.
- Registrar metadatos y crear copias de seguridad: registre metadatos (como IP públicas, región, AZ, VPC, subred y grupos de seguridad) y, si es necesario, cree copias de seguridad (como instantáneas de Amazon Elastic Block Store o realice AMI, exportación de claves y exportación de certificados) para los recursos del entorno de producción o si se trata de recursos críticos.
- Validar la infraestructura como código: determine si los recursos se han desplegado con CloudFormation, Terraform, Kit de desarrollo de la nube de AWS (AWS CDK) o cualquier otra herramienta de despliegue de infraestructura como código para poder volver a desplegarlos en caso necesario.
- Impedir el acceso: aplique controles restrictivos durante un periodo de tiempo para impedir el uso de recursos mientras determina si el recurso es necesario. Verifique que el entorno del recurso se puede revertir a su estado original si es necesario.
- Seguir su proceso de retirada interno: siga las tareas administrativas y el proceso de retirada de su organización, como eliminar el recurso del dominio de su organización, eliminar el registro DNS y eliminar el recurso de su herramienta de administración de configuración, herramienta de supervisión, herramienta de automatización y herramientas de seguridad.

Si el recurso es una instancia Amazon EC2, consulte la siguiente lista. Para obtener más detalles, consulte ¿Cómo eliminar o terminar recursos de Amazon EC2?

- Detenga o termine todas las instancias de Amazon EC2 y equilibradores de carga. Las instancias Amazon EC2 son visibles en la consola por poco tiempo después de su terminación. No se facturan las instancias que no están en estado de ejecución.
- Elimine su infraestructura de Auto Scaling.
- Libere todos los host dedicados.
- Elimine todos los volúmenes de Amazon EBS y las instantáneas de Amazon EBS.
- Libere todas las direcciones IP elásticas.
- Anule el registro de todas las imágenes de máquina de Amazon (AMI).
- Termine todos los entornos de AWS Elastic Beanstalk.

Si el recurso es un objeto en el almacenamiento de Amazon Glacier y si elimina un archivo antes de cumplir la duración de almacenamiento mínima, se le cobrará una tarifa prorrateada por eliminación anticipada. La duración de almacenamiento mínima de Amazon Glacier depende de la clase de almacenamiento utilizada. Para obtener un resumen de la duración de almacenamiento mínima de cada clase de almacenamiento, consulte Rendimiento de las clases de almacenamiento de Amazon S3. Para más detalles sobre cómo se calculan las tasas de eliminación anticipada, consulte Precios de Amazon S3.

En el sencillo diagrama de flujo del proceso de retirada que figura a continuación se describen las etapas de retirada. Antes de retirar los recursos, verifique que los que ha identificado para retirar no los usa la organización.



Flujo de retirada de recursos.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Trusted Advisor
- AWS CloudTrail

Vídeos relacionados:

- Delete CloudFormation stack but retain some resources (Eliminar la pila de pero retener algunos recursos CloudFormation)
- Find out which user launched Amazon EC2 instance (Averiguar qué usuario ha lanzado la instancia Amazon EC2)

Ejemplos relacionados:

- Delete or terminate Amazon EC2 resources (Eliminar o terminar recursos de Amazon EC2)
- Find out which user launched an Amazon EC2 instance (Averiguar qué usuario ha lanzado una instancia Amazon EC2)

COST04-BP03 Retirar recursos

Retire los recursos que algunos eventos generan, como las auditorías periódicas o los cambios en el uso. La retirada se suele realizar periódicamente y es manual o automática.

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

Guía para la implementación

La frecuencia y el esfuerzo dedicados a buscar recursos que no se utilizan deberían reflejar el ahorro potencial, de manera que una cuenta con pocos costes debería analizarse con menos frecuencia que una cuenta con costes mayores. Las búsquedas y los eventos de retirada pueden producirse por cambios de estado de la carga de trabajo, como el fin de la vida útil de un producto o su reemplazo. También pueden producirse por eventos externos, como cambios en las condiciones de mercado o la finalización de un producto.

Pasos para la aplicación

- Retirar recursos: se trata de la fase de amortización de los recursos de AWS que ya no se necesitan o de la finalización de un acuerdo de licencia. Complete todas las comprobaciones finales realizadas antes de pasar a la fase de eliminación y retirada de recursos para evitar interrupciones no deseadas, como la realización de instantáneas o copias de seguridad. Use el proceso de retirada para retirar los recursos identificados como no utilizados.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Trusted Advisor

Ejemplos relacionados:

- Well-Architected Labs: Decommission resources (Level 100) (Laboratorios de Well-Architected: retirada de recursos [nivel 100])

COST04-BP04 Retirar los recursos automáticamente

Diseñe su carga de trabajo para que gestione de manera sencilla la finalización de recursos a medida que identifica y retira recursos que no son críticos, recursos innecesarios o recursos con poco uso.

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

Guía para la implementación

Use la automatización para reducir o eliminar los costes asociados al proceso de retirada. El diseño de la carga de trabajo para que aplique procesos de retirada automáticos reducirá los costes generales de la carga de trabajo durante su vida. Puede usar AWS Auto Scaling para realizar el proceso de retirada. También puede implementar código personalizado con la API o el SDK para retirar los recursos de la carga de trabajo de forma automática.

Las aplicaciones modernas se crean primero sin servidor, una estrategia en la que se prioriza la adopción de servicios sin servidor. AWS ha desarrollado servicios sin servidor para las tres capas de su pila: computación, integración y almacenes de datos. El uso de la arquitectura sin servidor le permitirá ahorrar costes durante periodos de poco tráfico, con escalamiento y desescalamiento verticales de forma automática.

Pasos para la aplicación

- Implementar AWS Auto Scaling: en el caso de los recursos que se admitan, configúrelos con AWS Auto Scaling. AWS Auto Scaling puede ayudarle a optimizar la eficiencia de uso y costes al consumir servicios de AWS. Cuando baje la demanda, AWS Auto Scaling eliminará automáticamente cualquier exceso de capacidad de recursos para evitar un gasto excesivo.

- Configurar CloudWatch para terminar instancias: las instancias se pueden configurar para que finalicen con alarmas de CloudWatch. Use las métricas del proceso de retirada e implemente una alarma con una acción de Amazon Elastic Compute Cloud. Verifique la operación en un entorno no productivo antes de la implementación.
- Implementar código en la carga de trabajo: use el SDK de AWS CLI o la AWS para retirar los recursos de la carga de trabajo. Implemente código en la aplicación que se integre con AWS y finalice o elimine recursos que ya no se usan.
- Utilizar servicios sin servidor: dé prioridad a la creación de arquitecturas sin servidor y a la arquitectura basada en eventos en AWS para crear y ejecutar sus aplicaciones. AWS ofrece múltiples servicios de tecnología sin servidor que proporcionan de forma inherente una utilización de recursos optimizada automáticamente y una retirada automatizada (escalar y desescalar horizontalmente). Con las aplicaciones sin servidor, la utilización de los recursos se optimiza automáticamente y nunca pagará por un exceso de aprovisionamiento.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Trusted Advisor
- Sin servidor en AWS
- Crear alarmas para detener, terminar, reiniciar o recuperar una instancia
- Getting Started with Amazon EC2 Auto Scaling (Introducción a Amazon EC2 Auto Scaling)
- Agregar acciones de detención a las alarmas de Amazon CloudWatch

Ejemplos relacionados:

- Scheduling automatic deletion of AWS CloudFormation stacks (Programación de la eliminación automática de las pilas de AWS CloudFormation)
- Well-Architected Labs – Decommission resources automatically (Level 100) (Laboratorios de Well-Architected: retirar los recursos automáticamente [nivel 100])
- Servian AWS Auto Cleanup

COST04-BP05 Aplicación de políticas de retención de datos

Defina políticas de retención de datos en los recursos admitidos para gestionar la eliminación de objetos según los requisitos de su organización. Identifique y elimine los recursos y objetos innecesarios o huérfanos que ya no sean necesarios.

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

Utilice las políticas de retención de datos y de ciclo de vida para reducir los costes asociados al proceso de retirada y los costes de almacenamiento de los recursos identificados. Definir sus políticas de retención de datos y de ciclo de vida para realizar la migración y eliminación automatizadas de clases de almacenamiento reducirá los costes generales de almacenamiento durante su vida útil. Puede utilizar Amazon Data Lifecycle Manager para automatizar la creación y eliminación de instantáneas de Amazon Elastic Block Store e imágenes de máquina de Amazon (AMI) respaldadas por Amazon EBS, y utilizar Amazon S3 Intelligent-Tiering o una configuración del ciclo de vida de Amazon S3 para administrar el ciclo de vida de sus objetos de Amazon S3. También puede implementar código personalizado mediante el uso de la API o el SDK para crear políticas de ciclo de vida y reglas de política, a fin de que los objetos se eliminen automáticamente.

Pasos para la implementación

- Utilización Amazon Data Lifecycle Manager: utilice políticas de ciclo de vida en Amazon Data Lifecycle Manager para automatizar la eliminación de instantáneas de Amazon EBS y AMI respaldadas por Amazon EBS.
- Configuración el ciclo de vida de un bucket: utilice la configuración del ciclo de vida de Amazon S3 en un bucket para definir las acciones que realizará Amazon S3 durante el ciclo de vida de un objeto, así como su eliminación al final del ciclo de vida del objeto, en función de los requisitos de su empresa.

Recursos

Documentos relacionados:

- AWS Trusted Advisor
- Amazon Data Lifecycle Manager
- Cómo establecer la configuración del ciclo de vida en el bucket de Amazon S3

Vídeos relacionados:

- Automate Amazon EBS Snapshots with Amazon Data Lifecycle Manager (Automatizar las instantáneas de Amazon EBS con Amazon Data Lifecycle Manager)
- Empty an Amazon S3 bucket using a lifecycle configuration rule (Vaciar un bucket de Amazon S3 mediante una regla de configuración de ciclo de vida)

Ejemplos relacionados:

- Empty an Amazon S3 bucket using a lifecycle configuration rule (Vaciar un bucket de Amazon S3 mediante una regla de configuración de ciclo de vida)
- Well-Architected Lab: Decommission resources automatically (Level 100) (Laboratorio de Well-Architected: retirar los recursos automáticamente [nivel 100])

Recursos rentables

Preguntas

- COSTE 5. ¿Cómo evalúa el costo cuando selecciona servicios?
- COSTE 6. ¿Cómo cumple los objetivos de costes cuando selecciona el tipo, el tamaño y el número de recursos?
- COSTE 7. ¿Cómo utiliza los modelos de fijación de precios para reducir los costos?
- COSTE 8. ¿Cómo planifica los gastos de transferencia de datos?

COSTE 5. ¿Cómo evalúa el costo cuando selecciona servicios?

Amazon EC2, Amazon EBS y Amazon S3 son servicios de AWS básicos. Los servicios administrados, como Amazon RDS y Amazon DynamoDB, son servicios de AWS de nivel superior o de aplicación. Cuando selecciona los bloques de creación y los servicios administrados apropiados, puede optimizar esta carga de trabajo para el coste. Por ejemplo, cuando usa servicios administrados, puede reducir o eliminar gran parte de sus gastos administrativos y operativos, lo que le permite trabajar en aplicaciones y actividades relacionadas con el negocio.

Prácticas recomendadas

- COST05-BP01 Identificar los requisitos de la organización en relación con el coste
- COST05-BP02 Analizar todos los componentes de la carga de trabajo
- COST05-BP03 Análisis exhaustivo de cada componente

- COST05-BP04 Seleccionar software con licencias rentables
- COST05-BP05 Seleccionar los componentes de la carga de trabajo para optimizar los costes de acuerdo con las prioridades de la organización
- COST05-BP06 Analizar los costes para diferentes usos a lo largo del tiempo

COST05-BP01 Identificar los requisitos de la organización en relación con el coste

Trabaje con los miembros del equipo para definir el equilibrio entre la optimización de costos y otros pilares, como el rendimiento y la fiabilidad, de la carga de trabajo.

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

Guía para la implementación

En la mayoría de las organizaciones, el departamento de tecnología de la información (TI) está compuesto por varios equipos pequeños, cada uno con su propia agenda y área de enfoque, que refleja las especialidades y habilidades de los miembros de su equipo. Debe conocer los objetivos, prioridades y metas generales de su organización y cómo cada departamento o proyecto contribuye a estos objetivos. La clasificación de todos los recursos esenciales, incluidos el personal, el equipo, la tecnología, los materiales y los servicios externos, es crucial para lograr los objetivos de la organización y una planificación del presupuesto exhaustiva. La adopción de este enfoque sistemático para la identificación y comprensión de los costes es fundamental para establecer un plan de costes realista y sólido para la organización.

A la hora de seleccionar los servicios para su carga de trabajo, es fundamental que entienda las prioridades de su organización. Cree un equilibrio entre la optimización de costes y otros pilares de AWS Well-Architected Framework, como el rendimiento y la fiabilidad. Este proceso debe llevarse a cabo de manera sistemática y regular para reflejar los cambios en los objetivos de la organización, las condiciones del mercado y la dinámica operativa. Una carga de trabajo totalmente optimizada en cuanto a costes es la solución que más se ajusta a los requisitos de su organización, no necesariamente la de menor coste. Reúnase con todos los equipos de su organización (por ejemplo, de producto, empresarial, técnico y financiero) para recopilar información. Evalúe el impacto de las compensaciones que se realizan entre intereses opuestos o enfoques alternativos para ayudar a tomar decisiones fundamentadas a la hora de determinar dónde centrar los esfuerzos o elegir una vía de acción.

Por ejemplo, comercializar más rápido las nuevas características puede primar sobre la optimización de los costes, o se podría elegir una base de datos relacional para los datos no relacionales para

simplificar el esfuerzo de migración de un sistema en lugar de migrar a una base de datos optimizada para su tipo de datos y actualizar su aplicación.

Pasos para la implementación

- Identifique los requisitos de coste de la organización: reúnase con los miembros de los equipos de su organización, incluidos los de administración de productos, los propietarios de aplicaciones, los equipos de desarrollo y operativos, y roles de administración y financieros. Priorice los pilares de Well-Architected para esta carga de trabajo y sus componentes. El resultado debería ser una lista ordenada de los pilares. También puede añadir una ponderación a cada pilar para indicar cuánto enfoque adicional tiene, o las similitudes de un enfoque entre dos pilares.
- Aborde la deuda técnica y documéntela: durante la revisión de la carga de trabajo, aborde la deuda técnica. Documente una tarea pendiente para revisar la carga de trabajo en el futuro, con el objetivo de refactorizarlo o rediseñarlo para optimizarlo aún más. Es esencial comunicar claramente a otras partes interesadas las compensaciones que se han realizado.

Recursos

Prácticas recomendadas relacionadas:

- REL11-BP07 Diseñar su producto para cumplir objetivos de disponibilidad y acuerdos de nivel de servicio (SLA) de tiempo de actividad
- «OPS01-BP06 Evaluar compensaciones»

Documentos relacionados:

- Calculadora de coste total de propiedad (TCO) de AWS
- Clases de almacenamiento de Amazon S3
- Productos de la nube

COST05-BP02 Analizar todos los componentes de la carga de trabajo

Asegúrese de que se analice cada componente de la carga de trabajo, independientemente del tamaño o del coste actuales. El esfuerzo de revisión debería reflejar el beneficio potencial, como los costes actuales y previstos.

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

Guía para la implementación

Los componentes de la carga de trabajo, que están diseñados para ofrecer valor empresarial a la organización, pueden abarcar varios servicios. Para cada componente, se pueden elegir servicios específicos de Nube de AWS para abordar las necesidades empresariales. Esta selección podría estar influenciada por factores como la familiaridad con estos servicios o la experiencia previa con ellos.

Después de identificar los requisitos de su organización (tal y como se menciona en «COST05-BP01 Identificar los requisitos de la organización en relación con los costes»), lleve a cabo un análisis exhaustivo de todos los componentes de su carga de trabajo. Analice cada componente teniendo en cuenta los costes y tamaños actuales y previstos. Compare el coste del análisis con cualquier posible ahorro en la carga de trabajo a lo largo de su ciclo de vida. El esfuerzo dedicado a analizar todos los componentes de esta carga de trabajo debe corresponderse con los posibles ahorros o mejoras que se tiene previsto conseguir si se optimiza ese componente específico. Por ejemplo, si el coste del recurso propuesto es de 10 USD al mes y, según las cargas previstas, no superaría los 15 USD al mes, dedicar un día de esfuerzo a reducir los costes un 50 % (5 USD al mes) no debería superar el beneficio potencial durante la vida del sistema. Usar una estimación basada en datos más eficiente y rápida produciría el mejor resultado global para este componente.

Las cargas de trabajo pueden cambiar con el tiempo y el conjunto adecuado de servicios podría no ser óptimo si la arquitectura o el uso de la carga de trabajo cambia. En el análisis para seleccionar los servicios, se deben incluir estados de carga de trabajo actuales y futuros y niveles de uso. Implementar un servicio para un estado o uso de la carga de trabajo futura puede reducir los costes globales al reducir o eliminar el esfuerzo requerido para realizar cambios en el futuro. Por ejemplo, podría ser adecuado usar Amazon EMR Serverless al principio. Sin embargo, a medida que aumenta el consumo de ese servicio, la transición a Amazon EMR en Amazon EC2 podría reducir los costes de ese componente de la carga de trabajo.

La revisión estratégica de todos los componentes de la carga de trabajo, independientemente de sus atributos actuales, puede traducirse en mejoras y ahorros financieros importantes al cabo del tiempo. El esfuerzo invertido en este proceso de revisión debe ser deliberado y deben estudiarse cuidadosamente las ventajas que podrían conseguirse.

AWS Cost Explorer y AWS Cost and Usage Report (CUR) pueden analizar el coste de una prueba de concepto (PoC) o del entorno en ejecución. También se puede usar Calculadora de precios de AWS para calcular los costes de la carga de trabajo.

Pasos para la implementación

- Enumere los componentes de la carga de trabajo: cree una lista de los componentes de la carga de trabajo. Esta lista se usa para comprobar que se haya analizado cada componente. El esfuerzo dedicado debería reflejar la importancia de la carga de trabajo, tal como definen las prioridades de la organización. Agrupar los recursos mejora la eficiencia funcional (por ejemplo, el almacenamiento de datos de producción, si hay varias bases de datos).
- Priorice la lista de componentes: priorice la lista de componentes por esfuerzo. En general, esto se corresponde con el coste del componente, es decir, de más caro a menos caro, o según la importancia definida por las prioridades de la organización.
- Realice el análisis: revise las opciones y los servicios disponibles para cada componente de la lista y elija la opción que mejor se adapte a las prioridades de la organización.

Recursos

Documentos relacionados:

- «Calculadora de precios de AWS»
- AWS Cost Explorer
- Clases de almacenamiento de Amazon S3
- Productos de la nube

COST05-BP03 Análisis exhaustivo de cada componente

Consulte el coste total que supone a la organización cada componente. Calcule el coste total de propiedad teniendo en cuenta el coste de las operaciones y la administración, sobre todo cuando utilice servicios administrados por el proveedor de servicios en la nube. El esfuerzo de revisión debe reflejar los posibles beneficios (por ejemplo, el tiempo empleado en analizar es proporcional al coste de los componentes).

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

Guía para la implementación

Considere el ahorro de tiempo que permitirá a su equipo centrarse en la retirada de la deuda técnica, la innovación, las características que añaden valor y la creación de lo que diferencia a la empresa. Por ejemplo, puede que necesite migrar mediante lift-and-shift (también conocido como «volver a alojar») sus bases de datos de su entorno local a la nube lo más rápidamente posible y

optimizarlas más tarde. Merece la pena explorar el ahorro que puede suponer el uso de servicios administrados en AWS que puedan eliminar o reducir los costes de las licencias. Los servicios administrados en AWS eliminan la carga operativa y administrativa del mantenimiento de un servicio, como la aplicación de parches o la actualización del sistema operativo, y le permiten centrarse en la innovación y la empresa.

Dado que los servicios administrados operan a la escala de la nube, pueden ofrecer un coste menor por transacción o servicio. Puede realizar optimizaciones potenciales para obtener alguna ventaja tangible, sin cambiar la arquitectura principal de la aplicación. Por ejemplo, puede que quiera reducir el tiempo que dedica a administrar instancias de bases de datos mediante la migración a una plataforma de base de datos como servicio como Amazon Relational Database Service (Amazon RDS) o la migración de su aplicación a una plataforma completamente administrada como AWS Elastic Beanstalk.

Normalmente, los servicios administrados tienen atributos que puede configurar para garantizar una capacidad suficiente. Debe configurar y supervisar estos atributos para que su exceso de capacidad se mantenga al mínimo y el rendimiento se maximice. Puede modificar los atributos de AWS Managed Services mediante la Consola de administración de AWS o las API y los SDK de AWS para adaptar las necesidades de recursos a la demanda cambiante. Por ejemplo, puede aumentar o disminuir el número de nodos de un clúster de Amazon EMR (o un clúster de Amazon Redshift) para escalar o desescalar horizontalmente.

También puede empaquetar varias instancias en un recurso de AWS para conseguir un uso de mayor densidad. Por ejemplo, puede aprovisionar varias bases de datos pequeñas en una sola instancia de base de datos de Amazon Relational Database Service (Amazon RDS). A medida que aumenta el uso, puede migrar una de las bases de datos a una instancia de base de datos de Amazon RDS dedicada mediante un proceso de restauración y una instantánea.

Cuando aprovisione cargas de trabajo mediante servicios administrados, debe conocer los requisitos para ajustar la capacidad del servicio. Estos requisitos suelen ser tiempo, esfuerzo y cualquier impacto en el funcionamiento normal de la carga de trabajo. El recurso aprovisionado debe dejar tiempo para que se produzca cualquier cambio, por lo que debe aprovisionar la sobrecarga necesaria para permitirlo. El esfuerzo continuo requerido para modificar los servicios se puede reducir a prácticamente cero mediante el uso de API y SDK que se integran con el sistema y las herramientas de supervisión, como Amazon CloudWatch.

Amazon RDS, Amazon Redshift y Amazon ElastiCache proporcionan un servicio de base de datos administrada. Amazon Athena, Amazon EMR y Amazon OpenSearch Service proporcionan un servicio de análisis administrados.

AMS es un servicio que utiliza la infraestructura de AWS en nombre de los clientes y socios de la empresa. Proporciona un entorno seguro y conforme en el que puede desplegar sus cargas de trabajo. AMS utiliza modelos operativos de nube empresarial con automatización para permitirle cumplir con los requisitos de su organización, pasar a la nube más rápidamente y reducir los costes de administración continua.

Pasos para la implementación

- Realizar un análisis exhaustivo: mediante la lista de componentes, examine cada uno de ellos de mayor a menor prioridad. En el caso de los componentes con mayor prioridad y más costosos, realice un análisis adicional y evalúe todas las opciones disponibles y su impacto a largo plazo. En el caso de los componentes con menor prioridad, evalúe si los cambios en el uso modificarían la prioridad del componente y, a continuación, realice un análisis del esfuerzo adecuado.
- Comparar los recursos administrados y no administrados: considere el coste operativo de los recursos que administra y compárelos con los recursos administrados por AWS. Por ejemplo, revise sus bases de datos que se ejecutan en instancias de Amazon EC2 y compárelas con las opciones de Amazon RDS (un servicio administrado por AWS) o Amazon EMR en comparación con la ejecución de Apache Spark en Amazon EC2. Cuando cambie de una carga de trabajo autoadministrada a una completamente administrada por AWS, investigue cuidadosamente sus opciones. Los tres factores más importantes que se deben tener en cuenta son el tipo de servicio administrado que desea utilizar, el proceso que empleará para migrar sus datos y comprender el modelo de responsabilidad compartida de AWS.

Recursos

Documentos relacionados:

- Calculadora de coste total de propiedad (TCO) de AWS
- Clases de almacenamiento de Amazon S3
- Productos de Nube de AWS
- Modelo de responsabilidad compartida de AWS

Vídeos relacionados:

- Why move to a managed database? (¿Por qué cambiar a una base de datos administrada?)
- What is Amazon EMR and how can I use it for processing data? (¿Qué es Amazon EMR y cómo se puedo utilizar para el procesamiento de datos?)

Ejemplos relacionados:

- Why move to a managed database? (¿Por qué cambiar a una base de datos administrada?)
- Consolidate data from identical SQL Server databases into a single Amazon RDS for SQL Server database using AWS DMS (Consolidar los datos de bases de datos SQL Server idénticas en una única base de datos de Amazon RDS for SQL Server mediante AWS DMS)
- Deliver data at scale to Amazon Managed Streaming for Apache Kafka (Amazon MSK) (Entregar datos a escala a Amazon Managed Streaming for Apache Kafka [Amazon MSK])
- Migrate an ASP.NET web application to AWS Elastic Beanstalk (Migrar una aplicación web ASP.NET a AWS Elastic Beanstalk)

COST05-BP04 Seleccionar software con licencias rentables

El software de código abierto elimina los costes de licencias de software, lo que puede repercutir enormemente en los costes de las cargas de trabajo. Si se requiere software con licencia, evite licencias vinculadas a atributos arbitrarios como las CPU y busque licencias vinculadas a los resultados. El coste de estas licencias está más vinculado al beneficio que aportan.

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

Guía para la implementación

El código abierto se originó en el contexto del desarrollo de software e indica que el software cumple ciertos criterios para su distribución gratuita. El software de código abierto consta de código fuente que cualquiera puede inspeccionar, modificar y mejorar. En función de los requisitos empresariales, la habilidad de los ingenieros, el uso previsto u otras dependencias tecnológicas, las organizaciones pueden considerar la posibilidad de utilizar software de código abierto en AWS para minimizar los costes de sus licencias. Dicho de otro modo, el coste de las licencias de software se puede reducir utilizando software de código abierto. Esto puede repercutir de forma significativa en los costes de la carga de trabajo a medida que esta aumente.

Determine los beneficios del software con licencia teniendo en cuenta el coste total para optimizar su carga de trabajo. Haga simulaciones de los cambios en las licencias y estudie cómo afectaría a los costes de la carga de trabajo. Si un proveedor cambia el coste de la licencia de la base de datos, investigue cómo afecta eso a la eficiencia general de la carga de trabajo. Consulte los anuncios de precios históricos de sus proveedores para ver las tendencias en los cambios de las licencias en sus productos. Los costes de licencia también pueden variar sin tener en cuenta el rendimiento o el uso, como las licencias que varían según el hardware (licencias vinculadas a la CPU). Estas

licencias deberían evitarse porque sus costes pueden incrementarse rápidamente sin que haya unos resultados correspondientes.

Por ejemplo, utilizar una instancia de Amazon EC2 en us-east-1 con un sistema operativo Linux le permite reducir los costes en aproximadamente un 45 %, en comparación con la ejecución de otra instancia de Amazon EC2 que se ejecute en Windows.

La Calculadora de precios de AWS ofrece una forma completa de comparar los costes de varios recursos con diferentes opciones de licencias, como instancias de Amazon RDS y diferentes motores de bases de datos. Además, AWS Cost Explorer proporciona una perspectiva muy valiosa de los costes de las cargas de trabajo existentes, especialmente aquellas que vienen con diferentes licencias. Para la administración de licencias, AWS License Manager ofrece un método simplificado para supervisar y gestionar las licencias de software. Los clientes pueden desplegar y poner en funcionamiento su software de código abierto preferido en Nube de AWS.

Pasos para la implementación

- Analice las opciones de licencias: revise las condiciones de la licencia del software disponible. Busque versiones de código abierto que dispongan de las funciones requeridas y si los beneficios del software con licencia superan su coste. Si las condiciones son favorables, el coste del software se compensa con el beneficio que aporta.
- Analice al proveedor de software: revise cualquier cambio histórico de precios o licencias del proveedor. Busque cambios que no se alineen con los resultados, tales como términos punitivos si se ejecuta hardware o se trabaja con plataformas de proveedores específicos. Además, fíjese en cómo realizan las auditorías y las sanciones que se podrían aplicar.

Recursos

Documentos relacionados:

- «Open Source at AWS»
- Calculadora de coste total de propiedad (TCO) de AWS
- Clases de almacenamiento de Amazon S3
- Productos de la nube

Ejemplos relacionados:

- «Open Source Blogs»

- «AWS Open Source Blogs»
- Evaluación de optimización y licencias

COST05-BP05 Seleccionar los componentes de la carga de trabajo para optimizar los costes de acuerdo con las prioridades de la organización

Tenga en cuenta el coste al seleccionar los componentes de su carga de trabajo. Esto incluye el uso de servicios administrados y de nivel de aplicación o de una arquitectura sin servidor, de contenedores o basada en eventos para reducir el coste global. Minimice los costes de licencia con software de código abierto, software que no tenga costes de licencia o alternativas para reducir el coste.

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

Guía para la implementación

Tenga en cuenta el coste de los servicios y las opciones a la hora de seleccionar los componentes. Esto incluye el uso de servicios administrados y de nivel de aplicación, como Amazon Relational Database Service (Amazon RDS), Amazon DynamoDB, Amazon Simple Notification Service (Amazon SNS) y Amazon Simple Email Service (Amazon SES) para reducir el coste total de la organización.

Use aplicaciones sin servidor y contenedores para la computación, como AWS Lambda y Amazon Simple Storage Service (Amazon S3) para sitios web estáticos. Si es posible, incluya su aplicación en un contenedor y use los servicios de contenedores administrados de AWS, como Amazon Elastic Container Service (Amazon ECS) o Amazon Elastic Kubernetes Service (Amazon EKS).

Minimice los costes de licencia con software de código abierto o software que no tenga costes de licencia (por ejemplo, Amazon Linux para cargas de trabajo de computación o migre bases de datos a Amazon Aurora).

Puede usar servicios de nivel de aplicación o sin servidor, como Lambda, Amazon Simple Queue Service (Amazon SQS), Amazon SNSy Amazon SES. Estos servicios eliminan la necesidad de administrar un recurso y proporcionan la función de ejecución de código, servicios de colas y entrega de mensajes. La otra ventaja es que desescalan horizontalmente el rendimiento y el coste de acuerdo con el uso, por lo que permiten la asignación y atribución de costes de forma eficiente.

Usar una arquitectura basada en eventos también es posible con los servicios sin servidor. Las arquitecturas basadas en eventos se basan en la inserción, por lo que todo sucede bajo demanda

a medida que el evento se presenta en el enrutador. De esta forma, no pagará por un sondeo continuo para comprobar si hay algún evento. Esto se traduce en un menor consumo de ancho de banda de la red, una menor utilización de la CPU, una menor capacidad inactiva de la flota y menos establecimientos de protocolo de enlace SSL/TLS.

Para obtener más información sobre la tecnología sin servidor, consulte el documento técnico Serverless Applications Lens for the AWS Well-Architected Framework.

Pasos para la implementación

- Seleccionar cada servicio para optimizar costes: Use la lista de prioridades y el análisis para seleccionar la opción que se adapte mejor a las prioridades de la organización. En lugar de aumentar la capacidad para satisfacer la demanda, considere otras opciones que puedan ofrecerle un mejor rendimiento con un coste menor. Por ejemplo, si debe revisar el tráfico previsto para sus bases de datos en AWS, considere la posibilidad de aumentar el tamaño de la instancia o de utilizar servicios de Amazon ElastiCache (Redis o Memcached) a fin de proporcionar mecanismos de caché para sus bases de datos.
- Arquitectura basada en eventos: el uso de una arquitectura sin servidor también le permite crear una arquitectura basada en eventos para aplicaciones distribuidas basadas en microservicios, lo que le ayuda a crear soluciones escalables, resilientes, ágiles y rentables.

Recursos

Documentos relacionados:

- Calculadora de coste total de propiedad (TCO) de AWS
- AWS sin servidor
- Qué es la arquitectura basada en eventos
- Clases de almacenamiento de Amazon S3
- Productos de la nube
- Amazon ElastiCache (Redis OSS)

Ejemplos relacionados:

- Getting started with event-driven architecture (Introducción a la arquitectura basada en eventos)
- Arquitectura basada en eventos
- How Statsig runs 100x more cost-effectively using Amazon ElastiCache (Redis OSS)

- Prácticas recomendadas para trabajar con funciones de AWS Lambda

COST05-BP06 Analizar los costes para diferentes usos a lo largo del tiempo

Las cargas de trabajo pueden cambiar con el tiempo. Algunos servicios o características son más rentables en diferentes niveles de uso. Al analizar cada componente a lo largo del tiempo, así como el uso previsto, la carga de trabajo se mantiene rentable durante su vida útil.

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

Guía para la implementación

A medida que AWS lanza nuevos servicios y características, los servicios óptimos para su carga de trabajo también pueden cambiar. El esfuerzo necesario debe reflejar las ventajas potenciales. La frecuencia de revisión de la carga de trabajo depende de los requisitos de la organización. Si se trata de una carga de trabajo con un coste importante, implementar nuevos servicios antes maximizará el ahorro, por lo que realizar la revisión con mayor frecuencia puede ser de gran ayuda. Otro punto a revisar es el cambio en los patrones de uso. Unos cambios significativos en el uso pueden indicar que unos servicios alternativos serían óptimos.

Si necesita trasladar datos a Nube de AWS, puede seleccionar cualquier amplia variedad de servicios que ofrece AWS y herramientas de socios para ayudarle a migrar sus conjuntos de datos, ya sean archivos, bases de datos, imágenes de máquinas, volúmenes de bloques o, incluso, copias de seguridad en cinta. Por ejemplo, para trasladar una gran cantidad de datos con destino y origen en AWS o procesar datos en la periferia, puede utilizar uno de los dispositivos personalizados de AWS para trasladar de forma rentable petabytes de datos fuera de línea. Otro ejemplo: para tasas de transferencia de datos más elevadas, un servicio de conexión directa puede resultar más barato que una VPN que proporcione la coherencia de conectividad necesaria para su empresa.

Revise su actividad de escalamiento basándose en el análisis de costes para diferentes usos a lo largo del tiempo. Analice el resultado para ver si la política de escalamiento puede ajustarse para añadir instancias con varios tipos de instancia y opciones de compra. Revise la configuración para ver si es posible reducir el mínimo para atender las solicitudes de los usuarios, pero con una flota de menor tamaño, y añada más recursos para satisfacer la elevada demanda prevista.

Hable con las partes interesadas de su organización para realizar un análisis de los costes de los distintos usos a lo largo del tiempo y utilice la característica de previsión de AWS Cost Explorer para prever el impacto potencial de los cambios en el servicio. Supervise los desencadenadores del

nivel de uso mediante AWS Budgets, alarmas de facturación de CloudWatch y AWS Cost Anomaly Detection para identificar e implementar antes los servicios más rentables.

Pasos para la implementación

- Definir patrones de uso previstos: trabaje con los distintos equipos de la organización, como el departamento de marketing y los propietarios de producto, para documentar los patrones de uso previstos para la carga de trabajo. Hable con las partes interesadas de la empresa sobre el aumento de coste y uso, tanto históricos como previstos, y asegúrese de que el aumento se ajusta a los requisitos de la empresa. Identifique los días naturales, las semanas o los meses en los que espera que más usuarios utilicen sus recursos de AWS, lo que indica que debe aumentar la capacidad de los recursos existentes o adoptar servicios adicionales para reducir el coste y aumentar el rendimiento.
- Realizar análisis de costes con el uso previsto: aplique los patrones de uso definidos y analícelos en cada uno de estos puntos. El esfuerzo de análisis debería reflejar el resultado potencial. Por ejemplo, si el cambio de uso es grande, debería realizarse un análisis exhaustivo para verificar los costes y los cambios. En otras palabras, cuando el coste aumenta, el uso también debería aumentar para la empresa.

Recursos

Documentos relacionados:

- Calculadora de coste total de propiedad (TCO) de AWS
- Clases de almacenamiento de Amazon S3
- Productos de la nube
- Amazon EC2 Auto Scaling
- Migración de datos a la nube
- AWS Snow Family

Vídeos relacionados:

- AWS OpsHub for Snow Family

COSTE 6. ¿Cómo cumple los objetivos de costes cuando selecciona el tipo, el tamaño y el número de recursos?

Compruebe que elige el tamaño y el número de recurso apropiados para la tarea en cuestión. Al seleccionar el tipo, el tamaño y el número más rentables, minimiza el desperdicio.

Prácticas recomendadas

- COST06-BP01 Realizar modelado de costes
- COST06-BP02 Seleccionar el tipo, tamaño y número de recursos en función de los datos
- COST06-BP03 Seleccionar tipo, tamaño y número de recursos automáticamente en función de las métricas

COST06-BP01 Realizar modelado de costes

Identifique los requisitos de la organización (como las necesidades empresariales y los compromisos existentes) y realice un modelado de costes (costes generales) de la carga de trabajo y de cada uno de sus componentes. Realice actividades de referencia para la carga de trabajo bajo diferentes cargas previstas y compare los costes. El esfuerzo para realizar el modelado debería reflejar la ventaja potencial. Por ejemplo, el tiempo dedicado debe ser proporcional al coste del componente.

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

Guía para la implementación

Realice el modelado de costes para la carga de trabajo y cada uno de sus componentes para comprender el equilibrio entre los recursos. Busque el tamaño adecuado para cada recurso de la carga de trabajo según un determinado nivel de rendimiento. Comprender las consideraciones de costes puede servir de base al caso empresarial de su organización y al proceso de toma de decisiones cuando se evalúen los resultados de obtención de valor para el despliegue planificado de la carga de trabajo.

Realice actividades de referencia para la carga de trabajo bajo diferentes cargas previstas y compare los costes. El esfuerzo de modelado debe reflejar las posibles ventajas; por ejemplo, el tiempo empleado es proporcional al coste de los componentes o al ahorro previsto. Para conocer las prácticas recomendadas, consulte la sección Revisión del pilar Eficiencia del rendimiento del marco AWS Well-Architected Framework.

Por ejemplo, crear un modelado de costes para una carga de trabajo compuesta por recursos de computación, AWS Compute Optimizer puede ayudar a modelar los costes de las cargas de trabajo

en ejecución. Proporciona recomendaciones de tamaño ideal para los recursos de computación basándose en el uso histórico. Asegúrese de que se despliegan agentes de CloudWatch en las instancias de Amazon EC2 para recopilar métricas de memoria que le ayuden con recomendaciones más precisas en AWS Compute Optimizer. Se trata del origen de datos ideal para los recursos de computación porque es un servicio gratuito que usa el machine learning para realizar numerosas recomendaciones en función de los niveles de riesgo.

Existen varios servicios que puede utilizar con registros personalizados como orígenes de datos para las operaciones de dimensionamiento correcto para otros servicios y componentes de la carga de trabajo, como AWS Trusted Advisor, Amazon CloudWatch y Amazon CloudWatch Logs. AWS Trusted Advisor comprueba los recursos y marca los que tienen poco uso, lo que puede ayudarle a dimensionar correctamente sus recursos y a crear modelos de costes.

Estas son recomendaciones de datos y métricas de modelado de costes:

- La supervisión debe reflejar fielmente la experiencia del usuario. Seleccione la granularidad correcta del periodo y elija cuidadosamente el percentil 99 o el percentil máximo en lugar del promedio.
- Seleccione el nivel de detalle correcto para el periodo de análisis necesario a fin de cubrir cualquier ciclo de carga de trabajo. Por ejemplo, si se lleva a cabo un análisis de dos semanas, es posible que esté pasando por alto un ciclo mensual de alta utilización, lo que podría generar un aprovisionamiento insuficiente.
- Elija los servicios de AWS adecuados para la carga de trabajo prevista; para ello, tenga en cuenta sus compromisos existentes, los modelos de precios seleccionados para otras cargas de trabajo y la capacidad de innovar más rápidamente y centrarse en el valor empresarial principal.

Pasos para la implementación

- Realizar modelado de costes: despliegue la carga de trabajo o una prueba de concepto en una cuenta aparte con los tipos y tamaños de recurso específicos de la prueba. Ejecute la carga de trabajo con los datos de la prueba y registre los resultados de la salida, así como los datos de costes del momento en que se ejecutó la prueba. Después, vuelva a desplegar la carga de trabajo o cambie los tipos y tamaños de recurso y vuelva a ejecutar la prueba. Incluya las tarifas de licencia de cualquier producto que pueda utilizar con estos recursos y los costes estimados de las operaciones (mano de obra o ingenieros) para desplegar y administrar estos recursos durante la creación del modelado de costes. Considere el modelado de costes para un periodo (por hora, por día, por mes, por año o por trienio).

Recursos

Documentos relacionados:

- AWS Auto Scaling
- Identificar oportunidades para el tamaño correcto
- Características de Amazon CloudWatch
- Optimización de costes: redimensionamiento adecuado de Amazon EC2
- AWS Compute Optimizer
- Calculadora de precios de AWS

Ejemplos relacionados:

- Perform a Data-Driven Cost Modelling (Realizar un modelo de costes basado en datos)
- Estimate the cost of planned AWS resource configurations (Calcular el coste de las configuraciones de recursos de AWS previstas)
- Choose the right AWS tools (Elegir las herramientas de AWS adecuadas)

COST06-BP02 Seleccionar el tipo, tamaño y número de recursos en función de los datos

Seleccione el tamaño o el tipo de recurso en función de los datos sobre las características de la carga de trabajo y de los recursos. Por ejemplo, computación, memoria, rendimiento o uso intensivo de escritura. Para realizar esta selección, suele utilizarse una versión anterior (local) de la carga de trabajo, la documentación u otras fuentes de información sobre la carga de trabajo.

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

Guía para la implementación

Amazon EC2 ofrece una amplia selección de tipos de instancias con diferentes niveles de capacidad de CPU, memoria, almacenamiento y red para adaptarse a diferentes casos de uso. Estos tipos de instancias ofrecen diferentes combinaciones de capacidades de CPU, memoria, almacenamiento y red, lo que le proporciona versatilidad a la hora de seleccionar la combinación de recursos adecuada para sus proyectos. Cada tipo de instancia se ofrece en varios tamaños, de modo que puede ajustar sus recursos en función de las demandas de su carga de trabajo. Para determinar qué tipo de instancia necesita, recopile datos sobre los requisitos del sistema de la aplicación o el software que tiene pensado ejecutar en su instancia. Estos datos deben incluir lo siguiente:

- Sistema operativo
- Número de núcleos de CPU
- Núcleos de GPU
- Cantidad de memoria del sistema (RAM)
- Tipo y espacio de almacenamiento
- Requisitos de ancho de banda de la red

Identifique el propósito de los requisitos de computación y qué instancia se necesita y, a continuación, examine las distintas familias de instancias de Amazon EC2. Amazon ofrece las siguientes familias de tipos de instancias:

- Uso general
- Optimizadas para la computación
- Optimizadas para la memoria
- Optimizadas para el almacenamiento
- Computación acelerada
- Optimizadas para HPC

Para obtener una comprensión más profunda de los propósitos y casos de uso específicos que puede cumplir una familia de Amazon EC2 instancias en particular, consulte Tipos de AWS instancias.

La recopilación de requisitos los requisitos del sistema es fundamental para seleccionar la familia de instancias y el tipo de instancia específicos que mejor se ajusten a sus necesidades. Los nombres de los tipos de instancias están compuestos por el nombre de la familia y el tamaño de la instancia. Por ejemplo, la instancia t2.micro pertenece a la familia T2 y tiene el tamaño micro.

Seleccione el tamaño o el tipo de recurso en función de las características de la carga de trabajo y de los recursos. Por ejemplo: computación, memoria, rendimiento o uso intensivo de escritura. Para realizar esta selección, suele utilizarse el modelado de costes, una versión anterior de la carga de trabajo (por ejemplo, una versión local), documentación o u otras fuentes de información sobre la carga de trabajo (documentos técnicos o soluciones publicadas). El uso de calculadoras de precios o herramientas de administración de costes de AWS puede ayudar a tomar decisiones informadas sobre los tipos, tamaños y configuraciones de las instancias.

Pasos para la implementación

- Seleccione los recursos en función de los datos: utilice sus datos de modelado de costes para seleccionar el nivel de uso previsto de la carga de trabajo y elija el tipo y tamaño de recursos especificados. Basándose en los datos del modelado de costes, determine el número de CPU virtuales, la memoria total (GiB), el volumen del almacén de instancias local (GB), los volúmenes de Amazon EBS y el nivel de rendimiento de la red, teniendo en cuenta la velocidad de transferencia de datos necesaria para la instancia. Realice siempre selecciones basadas en análisis detallados y datos precisos para optimizar el rendimiento al tiempo que administra los costes de forma eficaz.

Recursos

Documentos relacionados:

- «Tipos de instancias de AWS»
- AWS Auto Scaling
- Características de Amazon CloudWatch
- Cost Optimization: EC2 Right Sizing

Vídeos relacionados:

- «Selecting the right Amazon EC2 instance for your workloads»
- «Right size your service»

Ejemplos relacionados:

- «It just got easier to discover and compare Amazon EC2 instance types»

COST06-BP03 Seleccionar tipo, tamaño y número de recursos automáticamente en función de las métricas

Use métricas de la carga de trabajo actual para seleccionar el tamaño y tipo correcto para optimizar el costo. Aprovisione de forma adecuada el rendimiento, el tamaño y el almacenamiento para los servicios de computación, almacenamiento, datos y redes. Esto puede hacerse con un bucle de retroalimentación, como el escalamiento automático, o mediante un código personalizado en la carga de trabajo.

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

Guía para la implementación

Cree un bucle de retroalimentación en la carga de trabajo que use métricas activas de la carga de trabajo en ejecución para realizar cambios en dicha carga de trabajo. Puede utilizar un servicio administrado, como AWS Auto Scaling, que se configura para llevar a cabo automáticamente las operaciones de tamaño adecuadas. AWS también proporciona API, SDK y características que permiten modificar los recursos con un mínimo esfuerzo. Puede programar una carga de trabajo para que detenga e inicie una instancia Amazon EC2 a fin de poder realizar un cambio en el tamaño o el tipo de instancia. Esto permite obtener el tamaño adecuado y, además, permite eliminar casi todo el coste operativo necesario para realizar el cambio.

Algunos servicios de AWS tienen una selección de tipo o tamaño automática, como Amazon Simple Storage Service Intelligent-Tiering. Amazon S3 Intelligent-Tiering mueve automáticamente los datos entre dos niveles de acceso (frecuente y poco frecuente) en función de sus patrones de uso.

Pasos para la aplicación

- Aumentar la observabilidad mediante la configuración de las métricas de la carga de trabajo: capture las métricas clave de la carga de trabajo. Estas métricas son indicativas de la experiencia del cliente, como el resultado de la carga de trabajo, y alinean las diferencias que hay entre los tipos y los tamaños de los recursos, como la CPU y el uso de memoria. En el caso del recurso de computación, analice los datos de rendimiento para determinar el tamaño adecuado de sus instancias Amazon EC2. Identifique las instancias inactivas y las infrautilizadas. Las métricas clave a tener en cuenta son el uso de la CPU y la utilización de la memoria (por ejemplo, un 40 % de utilización de la CPU el 90 % del tiempo, como se explica en Rightsizing with AWS Compute Optimizer and Memory Utilization Enabled [Redimensionamiento con AWS Compute Optimizer y uso de memoria activado]). Identifique las instancias con un uso de CPU y una utilización de memoria máximos inferiores al 40 % durante un periodo de cuatro semanas. Estas son las instancias que hay que dimensionar correctamente para reducir costes. En el caso de recursos de almacenamiento como Amazon S3, puede utilizar Amazon S3 Storage Lens, que le permite ver 28 métricas en varias categorías en el nivel de bucket y 14 días de datos históricos en el panel de forma predeterminada. Puede filtrar su panel de Amazon S3 Storage Lens por resumen y optimización de costes o eventos para analizar métricas específicas.
- Consultar las recomendaciones de redimensionamiento: utilice las recomendaciones de redimensionamiento en AWS Compute Optimizer y la herramienta de redimensionamiento Amazon EC2 en la consola de administración de costes, o revise el redimensionamiento de

AWS Trusted Advisor de sus recursos para realizar ajustes en la carga de trabajo. Es importante utilizar las herramientas adecuadas a la hora de redimensionar los distintos recursos y seguir las directrices de redimensionamiento, ya se trate de una instancia Amazon EC2, de clases de almacenamiento de AWS o de tipos de instancia Amazon RDS. En el caso de los recursos de almacenamiento, puede utilizar Amazon S3 Storage Lens, que le ofrece visibilidad sobre el uso del almacenamiento de objetos, las tendencias de actividad y le proporciona recomendaciones prácticas para optimizar los costes y aplicar las prácticas recomendadas de protección de datos. Gracias a las recomendaciones contextuales que Amazon S3 Storage Lens obtiene del análisis de las métricas de toda su organización, podrá tomar medidas inmediatas para optimizar el almacenamiento.

- Seleccionar el tipo y el tamaño de los recursos automáticamente según las métricas: use las métricas de la carga de trabajo y seleccione sus recursos de la carga de trabajo de forma manual o automática. En el caso de los recursos de computación, configurar AWS Auto Scaling o implementar el código en su aplicación puede reducir el esfuerzo necesario si deben realizarse cambios frecuentes, y así podrá implementar cambios potenciales antes que con el proceso manual. Puede lanzar y escalar automáticamente una flota de instancias bajo demanda e instancias de spot en un mismo grupo de Auto Scaling. Además de beneficiarse de descuentos por utilizar instancias de spot, puede usar las instancias reservadas o un Savings Plan para obtener descuentos en los precios habituales de las instancias bajo demanda. Todos estos factores combinados le ayudarán a optimizar el ahorro de costes de las instancias Amazon EC2 y a determinar la escala y el rendimiento que desea para su aplicación. También puede utilizar una estrategia de selección de tipo de instancia basada en atributos (ABS) en Auto Scaling Groups (ASG), que le permite expresar sus requisitos de instancia como un conjunto de atributos, por ejemplo, vCPU, memoria y almacenamiento. Puede utilizar automáticamente los tipos de instancia de nueva generación cuando se lancen y acceder a una gama más amplia de capacidad con las instancias de spot de Amazon EC2. Flota de Amazon EC2 y Amazon EC2 Auto Scaling seleccionan y lanzan instancias que se ajusten a los atributos especificados, por lo que no es necesario elegir manualmente los tipos de instancia. En el caos de los recursos de almacenamiento, puede utilizar las características Amazon S3 Intelligent Tiering y Amazon EFS Infrequent Access, que le permiten seleccionar automáticamente las clases de almacenamiento que ofrecen un ahorro automático de costes de almacenamiento cuando cambian los patrones de acceso a los datos, sin generar impacto en el rendimiento ni sobrecarga operativa.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- Tamaño correcto de AWS
- AWS Compute Optimizer
- Características de Amazon CloudWatch
- Configuración inicial de CloudWatch
- Publicar métricas personalizadas de CloudWatch
- Getting Started with Amazon EC2 Auto Scaling (Introducción a Amazon EC2 Auto Scaling)
- Amazon S3 Storage Lens
- Amazon S3 Intelligent-Tiering
- Amazon EFS Infrequent Access
- Lanzamiento de una instancia Amazon EC2 mediante el SDK

Vídeos relacionados:

- Right Size Your Services (Tamaño correcto de sus servicios)

Ejemplos relacionados:

- Attribute based Instance Type Selection for Auto Scaling for Amazon EC2 Fleet (Selección de tipo de instancia basada en atributos para Auto Scaling para Flota de Amazon EC2)
- Optimizing Amazon Elastic Container Service for cost using scheduled scaling (Optimización de Amazon Elastic Container Service para coste mediante el escalamiento programado)
- Predictive scaling with Amazon EC2 Auto Scaling (Escalamiento predictivo con Amazon EC2 Auto Scaling)
- Optimize Costs and Gain Visibility into Usage with Amazon S3 Storage Lens (Optimizar los costes y obtener visibilidad del uso con Amazon S3 Storage Lens)
- Well-Architected Labs: Rightsizing Recommendations (Level 100) (Laboratorios de Well-Architected: recomendaciones de redimensionamiento [nivel 100])
- Well-Architected Labs: Rightsizing with AWS Compute Optimizer and Memory Utilization Enabled (Level 200) (Laboratorios de Well-Architected: redimensionamiento con AWS Compute Optimizer y uso de memoria activado [nivel 200])

COSTE 7. ¿Cómo utiliza los modelos de fijación de precios para reducir los costos?

Use el modelo de fijación de precios más apropiado para sus recursos a fin de minimizar los gastos.

Prácticas recomendadas

- COST07-BP01 Analizar los modelos de precios
- COST07-BP02 Elegir regiones según el coste
- COST07-BP03 Seleccionar acuerdos de terceros con condiciones rentables
- COST07-BP04 Implementar modelos de precios para todos los componentes de la carga de trabajo
- COST07-BP05 Realizar análisis de modelos de precios en el nivel de la cuenta de administración

COST07-BP01 Analizar los modelos de precios

Analice cada componente de la carga de trabajo. Determine si el componente y los recursos se ejecutarán durante períodos extensos (por descuentos por compromiso) o períodos dinámicos y de corta ejecución (para spot o bajo demanda). Realice un análisis de la carga de trabajo mediante las recomendaciones de las herramientas de administración de costes y aplique las reglas empresariales a dichas recomendaciones para conseguir un alto rendimiento.

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

Guía para la implementación

AWS tiene varios modelos de precios que le permiten pagar por los recursos de la manera más rentable en función de las necesidades de su organización y según el producto. Determine con sus equipos el modelo de precios más apropiado. Su modelo de precios suele constar de una combinación de varias opciones, según lo determine su disponibilidad

Con las instancias bajo demanda puede pagar por capacidad de computación o de base de datos por hora o por segundo (60 segundos como mínimo) según las instancias que ejecute, sin compromisos a largo plazo o pagos por adelantados.

Los Savings Plans son un modelo de precios flexible que ofrece precios bajos por el uso de Amazon EC2, Lambda y AWS Fargate a cambio de un compromiso de uso constante (medido en USD por hora) durante un plazo de uno o tres años.

Las instancias de spot son un mecanismo de precios de Amazon EC2 que le permite solicitar capacidad de computación sobrante a una tarifa por hora con descuento (hasta un 90 % sobre el precio bajo demanda) sin compromiso inicial.

Con las instancias reservadas puede obtener hasta un 75 % de descuento si paga por adelantado la capacidad. Para obtener más detalles, consulte Optimizing costs with reservations (Optimización de los costes con las reservas).

Puede incluir un Savings Plans para los recursos asociados a los entornos de producción, calidad y desarrollo. Como alternativa, debido a que los recursos del entorno aislado solo se encienden cuando se necesitan, podría elegir un modelo bajo demanda para los recursos de ese entorno. Utilice las instancias de spot de Amazon para reducir los costes de Amazon EC2 o utilice Savings Plans para computación a fin de reducir el coste de Amazon EC2, Fargate y Lambda. La herramienta de recomendaciones de AWS Cost Explorer ofrece oportunidades de descuentos por compromiso con Savings Plans.

Si en el pasado ha estado comprando instancias reservadas por Amazon EC2 o ha establecido prácticas de asignación de costes en su organización, puede seguir usando las instancias reservadas de Amazon EC2 por el momento. Sin embargo, recomendamos elaborar una estrategia para usar Savings Plans en el futuro como mecanismo más flexible de ahorro de costes. Puede actualizar las recomendaciones de Savings Plans (SP) en AWS Cost Management para generar nuevas recomendaciones de Savings Plans en cualquier momento. Use las instancias reservadas (RI) para reducir los costes de Amazon RDS, Amazon Redshift, Amazon ElastiCache y Amazon OpenSearch Service. Los Savings Plans y las instancias reservadas están disponibles en tres modalidades de pago: puede abonarse el total por adelantado, abonarse parte por adelantado y no abonarse nada por adelantado. Utilice las recomendaciones de compra de RI y SP de AWS Cost Explorer.

Para buscar oportunidades para cargas de trabajo de spot, use una vista por hora del uso general y busque períodos regulares de uso cambiante o de elasticidad. Puede utilizar instancias de spot para diversas aplicaciones flexibles y tolerantes a errores. Algunos ejemplos son los servidores web sin estado, los puntos de conexión de API, las aplicaciones de macrodatos y análisis, las cargas de trabajo en contenedores, CI/CD y otras cargas de trabajo flexibles.

Analice sus instancias de Amazon EC2 y Amazon RDS si pueden desactivarse cuando no las utilice (fuera de horario laboral y en fines de semana). Este enfoque le permitirá reducir costes en un 70 % o más con respecto a su uso ininterrumpido. Si tiene clústeres de Amazon Redshift que solo deben estar disponibles en momentos concretos, puede pausar el clúster y reanudarlo más tarde.

Cuando se detiene el clúster de Amazon Redshift o la instancia de Amazon EC2 y Amazon RDS, la facturación de computación se detiene y solo se aplica el cargo por almacenamiento.

Tenga en cuenta que las reservas de capacidad bajo demanda (ODCR) no suponen un descuento en los precios. Las reservas de capacidad se cobran según la tarifa bajo demanda equivalente, tanto si ejecuta instancias con capacidad reservada como si no. Deben tenerse en cuenta cuando necesite proporcionar suficiente capacidad para los recursos que tiene previsto ejecutar. Las ODCR no tienen por qué estar vinculadas a compromisos a largo plazo, ya que pueden cancelarse cuando ya no las necesite, pero también pueden beneficiarse de los descuentos que ofrecen los Savings Plans o las instancias reservadas.

Pasos para la implementación

- Analizar la elasticidad de la carga de trabajo: mediante el detalle por horas en Cost Explorer o un panel personalizado, analice la elasticidad de la carga de trabajo. Busque cambios regulares en el número de instancias que se están ejecutando. Las instancias de corta duración son candidatas para las instancias o la flota de spot.
  - Laboratorio de Well-Architected: Cost Explorer
  - Laboratorio de Well-Architected: visualización de los costes
- Revisar los contratos de precios existentes: revise los contratos o los compromisos actuales de sus necesidades a largo plazo. Analice lo que tiene actualmente y en qué medida se utilizan esos compromisos. Aproveche los descuentos contractuales o los acuerdos empresariales preexistentes. Los acuerdos empresariales ofrecen a los clientes la opción de adaptar los acuerdos que mejor se adapten a sus necesidades. En el caso de compromisos a largo plazo, considere los descuentos por precios reservados, las instancias reservadas o Savings Plans para el tipo de instancia específico, la familia de instancias, Región de AWS y las zonas de disponibilidad.
- Realizar un análisis de descuentos por compromiso: use Savings Plans en su cuenta para revisar las recomendaciones de Cost Explorer y de instancias reservadas. Para comprobar que está implementando las recomendaciones correctas con los descuentos y el riesgo necesarios, siga los laboratorios de Well-Architected.

Recursos

Documentos relacionados:

- Acceso a recomendaciones de instancias reservadas

- Opciones de compra de instancias
- AWS Enterprise

Vídeos relacionados:

- Ahorre hasta un 90 % y ejecute cargas de trabajo de producción en spot

Ejemplos relacionados:

- Laboratorio de Well-Architected: Cost Explorer
- Laboratorio de Well-Architected: visualización de los costes
- Laboratorio de Well-Architected: modelos de precios

COST07-BP02 Elegir regiones según el coste

Los precios de los recursos pueden variar según la región. Identifique las diferencias regionales de costes y realice el despliegue solo en las regiones con costes más elevados para cumplir los requisitos de latencia, residencia de los datos y soberanía de los datos. Si tiene en cuenta el coste de la región, podrá pagar el precio global más bajo por esta carga de trabajo.

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

Guía para la implementación

La infraestructura de Nube de AWS es global, está alojada en múltiples ubicaciones en todo el mundoy se basa en Regiones de AWS, zonas de disponibilidad, zonas locales, AWS Outposts y zonas de longitud de onda. Una región es una ubicación física en el mundo y cada región es un área geográfica independiente en la que AWS tiene varias zonas de disponibilidad. Las zonas de disponibilidad, que son varias ubicaciones aisladas en cada región, constan de uno o varios centros de datos discretos, cada uno de ellos con alimentación, redes y conectividad redundantes.

Cada Región de AWS opera según las condiciones del mercado local, y el precio de los recursos es distinto en cada región debido a las diferencias del coste del terreno, la fibra, la electricidad y los impuestos, por ejemplo. Elija una región específica en la que desee aplicar un componente de su solución o la solución completa a fin de poder ejecutar al precio más bajo posible a nivel mundial. Utilice la calculadora de AWS para calcular los costes de su carga de trabajo en varias regiones mediante la búsqueda de servicios por tipo de ubicación (región, zona de longitud de onda y zona local) y región.

Al diseñar soluciones, una práctica recomendada es intentar colocar los recursos informáticos más cerca de los usuarios a fin de brindar una latencia más baja y una soberanía de datos sólida. Seleccione la ubicación geográfica en función de los requisitos de su empresa, privacidad de datos, rendimiento y seguridad. En el caso de aplicaciones con usuarios finales en todo el mundo, utilice varias ubicaciones.

Recurra a las regiones que ofrecen precios más bajos por los servicios de AWS para desplegar sus cargas de trabajo si no tiene obligaciones en materia de privacidad de datos, seguridad y requisitos de empresa. Por ejemplo, si su región predeterminada es ap-southeasth-2 (Sídney) y si no existen restricciones (privacidad de los datos o seguridad, por ejemplo) para utilizar otras regiones, desplegar instancias de Amazon EC2 no críticas (desarrollo y pruebas) en la región north-east-1 (N. Virginia) tendrá menos costes.

	Cumplimiento	*Latencia*	*Coste*	*Servicios/características*
Región 1	✓	15 ms	$$	✓
Región 2	✓	20 ms	$$$	X
Región 3	✓	80 ms	$	✓
Región 4	✓	15 ms	$$	✓
Región 5	✓	20 ms	$$$	X
Región 6	✓	15 ms	$	✓
Región 7	✓	80 ms	$	✓
Región 8	✓	15 ms	$	X

Tabla matricial de características de las regiones

La tabla matricial anterior nos muestra que la Región 4 es la mejor opción para este escenario específico, porque la latencia es baja en comparación con otras regiones, el servicio está disponible y es la región menos cara.

Pasos para la implementación

- Revise los precios de las Región de AWS: analice los costes de la carga de trabajo de la región actual. A partir de los costes más elevados por servicio y tipo de uso, calcule los costes en otras

regiones que estén disponibles. Si el ahorro previsto supera el coste de trasladar el componente o la carga de trabajo, migre a la nueva región.

- Revise los requisitos de los despliegues en varias regiones: analice los requisitos y las obligaciones de su empresa (privacidad de los datos, seguridad o rendimiento) para averiguar si existe alguna restricción que le impida utilizar varias regiones. Si no hay obligaciones que restrinjan el uso de una sola región, utilice varias.
- Analice la transferencia de datos requerida: tenga en cuenta los costes de transferencia de datos al seleccionar las regiones. Mantenga sus datos cerca de su cliente y de los recursos. Seleccione Regiones de AWS menos costosas donde fluyan los datos y donde la transferencia de datos sea mínima. Dependiendo de los requisitos de su empresa para la transferencia de datos, puede utilizar Amazon CloudFront, AWS PrivateLink, AWS Direct Connecty AWS Virtual Private Network para reducir los costes de red, mejorar el rendimiento y mejorar la seguridad.

Recursos

Documentos relacionados:

- Acceso a recomendaciones de instancias reservadas
- Precios de Amazon EC2
- Opciones de compra de instancias
- Tabla de regiones

Vídeos relacionados:

- Ahorre hasta un 90 % y ejecute cargas de trabajo de producción en spot

Ejemplos relacionados:

- Overview of Data Transfer Costs for Common Architectures (Información general de los costes de transferencia de datos para arquitecturas comunes)
- Cost Considerations for Global Deployments (Consideraciones sobre costes para despliegues globales)
- Qué tener en cuenta al seleccionar una región para las cargas de trabajo
- Well-Architected Labs: Restrict service usage by Region (Level 200) (Laboratorios de Well-Architected: restricción del uso del servicio por región [nivel 200])

COST07-BP03 Seleccionar acuerdos de terceros con condiciones rentables

Los acuerdos y condiciones rentables garantizan que el coste de estos servicios vaya a la par de los beneficios que proporcionan. Seleccione acuerdos y precios que se escalen cuando proporcionen beneficios adicionales a la organización.

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

Guía para la implementación

Hay varios productos en el mercado que pueden ayudarle a administrar los costes en sus entornos de nube. Puede que haya algunas diferencias en lo que se refiere a las características que dependen de los requisitos del cliente. Por ejemplo, puede que algunos se centren en la gobernanza de costes o la visibilidad de costes y otros en la optimización de costes. Un factor clave para que la optimización de los costes y la gobernanza sean efectivas es utilizar la herramienta adecuada con las características necesarias y el modelo de precios correcto. Estos productos tienen diferentes modelos de precios. Algunos cobran un porcentaje determinado de la factura mensual, mientras que otros cobran un porcentaje del ahorro que se consigue. Lo ideal es que pague solo lo que necesita.

Al utilizar soluciones o servicios de terceros en la nube, es importante que las estructuras de precios se ajusten a los resultados deseados. Los precios deben ir a la par de los resultados y el valor que aportan. Un ejemplo de ello es el software que se lleva una parte del ahorro que proporciona: cuanto más ahorra (resultado), más cobra. Los acuerdos de licencias en los que paga más a medida que aumentan sus gastos no siempre le convienen para optimizar costes. Sin embargo, si el proveedor ofrece ventajas claras en todas las partes de su factura, este aumento de tarifa podría estar justificado.

Por ejemplo, una solución que proporciona recomendaciones para Amazon EC2 y cobra un porcentaje de toda la factura podría ser más cara si usa otros servicios que no generan ningún beneficio. Otro ejemplo es un servicio administrado que se cobra a un porcentaje del coste de los recursos que se administran. Un mayor tamaño de la instancia no tiene por qué requerir un mayor esfuerzo de administración, aunque sí se podría cobrar más. A fin de impulsar la eficiencia, asegúrese de que en estos acuerdos de precios del servicio, se incluya un programa o características de optimización de costes en su servicio.

Los clientes podrían encontrar en el mercado estos productos más avanzados o fáciles de usar. Debe considerar el coste de estos productos y pensar en los posibles resultados de optimización de costes a largo plazo.

Pasos para la implementación

- Analice los acuerdos y condiciones de terceros: revise los precios de los acuerdos de terceros. Realice modelados de los diferentes niveles de uso y tenga en cuenta nuevos costes, como el uso de nuevos servicios o incrementos en los servicios actuales debido al crecimiento de la carga de trabajo. Decida si los costes adicionales proporcionan los beneficios necesarios para su empresa.

Recursos

Documentos relacionados:

- Acceso a recomendaciones de instancias reservadas
- Opciones de compra de instancias

Vídeos relacionados:

- Ahorre hasta un 90 % y ejecute cargas de trabajo de producción en spot

COST07-BP04 Implementar modelos de precios para todos los componentes de la carga de trabajo

Al ejecutar recursos de forma permanente, se debe utilizar la capacidad reservada, como los Savings Plans o las instancias reservadas. La capacidad a corto plazo se configura con instancias o una flota de spot. Las instancias bajo demanda solo se usan para cargas de trabajo a corto plazo que no se pueden interrumpir y que no se ejecutan lo suficiente como para tener capacidad reservada, es decir, de un 25 a un 75 % del período, según el tipo de recurso.

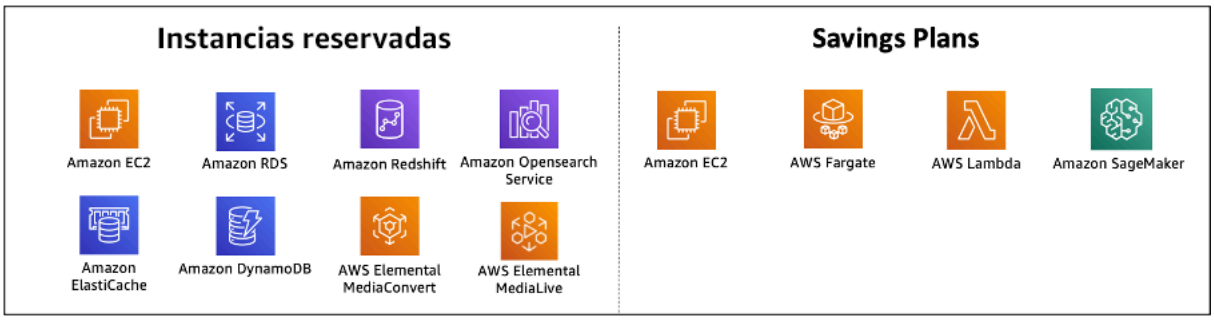
Nivel de riesgo expuesto si no se establece esta práctica recomendada: bajo

Guía para la implementación

Para mejorar la rentabilidad, AWS proporciona varias recomendaciones de compromiso basadas en el uso pasado. Estas recomendaciones pueden servirle para saber lo que puede ahorrar y cómo se utilizará el compromiso. Puede utilizar estos servicios como instancias bajo demanda, spot, o comprometerse por un período de tiempo determinado y reducir sus costes bajo demanda con instancias reservadas (RI) y Savings Plans (SP). Necesita conocer, no solo los componentes de cada carga de trabajo y los múltiples servicios de AWS, sino también los descuentos por compromiso, las opciones de compra y las instancias de spot de estos servicios para optimizar su carga de trabajo.

Tenga en cuenta los requisitos de los componentes de su carga de trabajo e infórmese de los diferentes modelos de precios de estos servicios. Defina el requisito de disponibilidad de estos componentes. Determine si hay varios recursos independientes que ejecuten la función en la carga de trabajo y cuáles son los requisitos de la carga de trabajo a lo largo del tiempo. Compare el coste de los recursos con el modelo de precios bajo demanda predeterminado y otros modelos aplicables. Tenga en cuenta cualquier cambio potencial en los recursos o en los componentes de la carga de trabajo.

Veamos, por ejemplo, esta arquitectura de aplicaciones web en AWS. Esta carga de trabajo de ejemplo se compone de varios servicios de AWS, como Amazon Route 53, AWS WAF, Amazon CloudFront, instancias de Amazon EC2, instancias de Amazon RDS, equilibradores de carga, almacenamiento Amazon S3 y Amazon Elastic File System (Amazon EFS). Debe revisar cada uno de estos servicios e identificar las posibles oportunidades de ahorro de costes con diferentes modelos de precios. Algunos de ellos podrían ser aptos para RI o SP, mientras que otros podrían estar disponibles solo bajo demanda. Como se muestra en la siguiente imagen, algunos de los servicios de AWS pueden comprometerse mediante RI o SP.



Servicios de AWS comprometidos mediante instancias reservadas y Savings Plans

Pasos para la implementación

- Implemente modelos de precios: utilice los resultados de sus análisis para comprarSavings Plans o instancias reservadas, o implementar instancias de spot. Si se trata de su primera compra de compromiso, elija las cinco o diez mejores recomendaciones de la lista y, a continuación, supervise y analice los resultados durante uno o dos meses. AWS Cost Management Console le guiará a lo largo del proceso. Revise las recomendaciones de RI o SP desde la consola, personalice las recomendaciones (tipo, pago y plazo), revise el compromiso por hora (por ejemplo, 20 dólares por hora) y, a continuación, añádalo a la cesta. Los descuentos se aplican automáticamente al uso elegible. Compre una pequeña cantidad de descuentos por compromiso en ciclos regulares (por ejemplo, cada 2 semanas o mensualmente). Implemente instancias de spot para las cargas

de trabajo que se puedan interrumpir o no tengan estado. Por último, seleccione instancias de Amazon EC2 bajo demanda y asigne recursos para los requisitos restantes.

- Ciclo de revisión de la carga de trabajo: implemente un ciclo de revisión de la carga de trabajo que analice específicamente la cobertura del modelo de precios. Cuando la carga de trabajo tenga la cobertura requerida, compre descuentos por compromiso adicionales parcialmente (cada pocos meses) o a medida que cambie el uso en la organización.

Recursos

Documentos relacionados:

- «Understanding your Savings Plans recommendations»
- Acceso a recomendaciones de instancias reservadas
- Cómo comprar instancias reservadas
- Opciones de compra de instancias
- Instancias de spot
- Modelos de reserva para otros servicios de AWS
- «Savings Plans Supported Services»

Vídeos relacionados:

- Ahorre hasta un 90 % y ejecute cargas de trabajo de producción en spot

Ejemplos relacionados:

- «What should you consider before purchasing Savings Plans?»
- «How can I use Cost Explorer to analyze my spending and usage?»

COST07-BP05 Realizar análisis de modelos de precios en el nivel de la cuenta de administración

Consulte las herramientas de facturación y administración de costes y vea los descuentos recomendados con compromisos y reservas para realizar análisis periódicos en el nivel de la cuenta de administración.

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

Guía para la implementación

La realización periódica de modelos de costes le ayuda a implementar oportunidades de optimización en múltiples cargas de trabajo. Por ejemplo, si varias cargas de trabajo usan instancias bajo demanda en un nivel agregado, el riesgo de cambio es menor e implementar un descuento basado en el compromiso puede tener un coste general inferior. Se recomienda realizar análisis en ciclos regulares de dos semanas a un mes. De este modo, podrá realizar compras de ajustes pequeños para que sus modelos de precios puedan seguir evolucionando a medida que cambien sus cargas de trabajo y sus componentes.

Utilice la herramienta de recomendaciones AWS Cost Explorer para buscar descuentos por compromiso en su cuenta de administración. Las recomendaciones en el nivel de cuenta de administración se calculan teniendo en cuenta el uso en todas las cuentas de su organización de AWS que tengan activadas las instancias de reserva (RI) o Savings Plans (SP). También se calculan cuando se activa la opción de compartir descuentos para recomendar un compromiso que maximice los ahorros en todas las cuentas.

Aunque la compra a nivel de cuenta de administración permite conseguir el máximo ahorro en muchos casos, puede haber situaciones en las que podría plantearse la posibilidad de comprar SP a nivel de la cuenta vinculada cuando, por ejemplo, desee que los descuentos se apliquen primero al uso de esa cuenta vinculada en particular. Las recomendaciones para las cuentas de los miembros se calculan en el nivel de cuenta individual, para maximizar el ahorro de cada cuenta aislada. Si su cuenta es propietaria de compromisos de RI y SP, se aplicarán en este orden:

1. RI de zona
2. RI estándar
3. RI convertible
4. SP para instancias
5. SP para computación

Si compra un SP a nivel de cuenta de administración, el ahorro se aplicará en función del porcentaje de descuento más alto al más bajo. Los SP a nivel de cuenta de administración examinan todas las cuentas vinculadas y aplican el ahorro allí donde el descuento sea más alto. Si desea restringir dónde se aplica el ahorro, puede comprar un Savings Plan a nivel de cuenta vinculada para que, cada vez que esa cuenta utilice servicios de computación que cumplan los requisitos, el descuento se aplique primero allí. Cuando la cuenta no utilice servicios de computación que cumplan los requisitos, el descuento se compartirá entre las demás cuentas vinculadas de la misma cuenta de

administración. La opción de compartir descuentos está activada de forma predeterminada, pero se puede desactivar si es necesario.

En una familia de facturación unificada, los Savings Plans se aplican primero al uso de la cuenta del propietario y, luego, al uso de otras cuentas. Esto solo ocurre si tiene activada la opción de compartir descuentos. Sus Savings Plans se aplican primero al porcentaje de ahorro más alto. Si hay varios usos que tienen porcentajes de ahorro iguales, los Savings Plans se aplican al primer uso con la tarifa de Savings Plans más baja. Los Savings Plans seguirán aplicándose hasta que no queden más usos o hasta que se agote su compromiso. El uso restante se cobrará a las tarifas bajo demanda. Puede actualizar las recomendaciones de Savings Plans en AWS Cost Management para generar nuevas recomendaciones de Savings Plans en cualquier momento.

Tras analizar la flexibilidad de las instancias, puede comprometerse siguiendo las recomendaciones. Cree modelos de costes al analizar los costes de la carga de trabajo a corto plazo con posibles opciones de recursos diferentes, además de analizar los modelos de precios de AWS y su alineación con los requisitos empresariales para averiguar el coste total de propiedad y las oportunidades de optimización de costes.

Pasos para la implementación

Realizar una análisis de descuentos por compromiso: use Cost Explorer en su cuenta para revisar las recomendaciones de Savings Plans y de instancias reservadas. Asegúrese de que entiende las recomendaciones del Saving Plan y calcule el gasto y el ahorro mensual. Revise las recomendaciones en el nivel de cuenta de administración que se calculan teniendo en cuenta el uso en todas las cuentas de miembros de su organización de AWS que tengan activadas las instancias reservadas o el reparto de descuentos de Savings Plans para obtener el máximo ahorro en todas las cuentas. Puede verificar que ha implementado las recomendaciones correctas con los descuentos y riesgos necesarios si sigue los laboratorios de Well-Architected.

Recursos

Documentos relacionados:

- ¿Cómo funcionan los precios de AWS?
- Opciones de compra de instancias
- Información general de Savings Plans
- Recomendaciones de Savings Plans
- Acceso a recomendaciones de instancias reservadas
- Cómo entender las recomendaciones de Savings Plans

- How Savings Plans apply to your AWS usage (Cómo se aplican los Savings Plans a su uso de AWS)
- Planes de ahorro con facturación unificada
- Activación de instancias reservadas compartidas y descuentos de Savings Plans

Vídeos relacionados:

- Ahorre hasta un 90 % y ejecute cargas de trabajo de producción en spot

Ejemplos relacionados:

- AWS Well-Architected Lab: Pricing Models (Level 200)
- AWS Well-Architected Labs: Pricing Model Analysis (Level 200)
- ¿Qué debo tener en cuenta antes de comprar un Savings Plan?
- How can I use rolling Savings Plans to reduce commitment risk?
- Cuándo usar las instancias de spot

COSTE 8. ¿Cómo planifica los gastos de transferencia de datos?

Compruebe que planifica y supervisa los cargos de transferencia de datos para que pueda tomar decisiones en cuanto al diseño y minimizar los costes. Un cambio de diseño pequeño, pero efectivo, puede reducir drásticamente sus costos operativos con el tiempo.

Prácticas recomendadas

- COST08-BP01 Realizar un modelado de transferencia de datos
- COST08-BP02 Seleccionar componentes para optimizar el coste de la transferencia de datos
- COST08-BP03 Implementar servicios para reducir los costes de transferencia de datos

COST08-BP01 Realizar un modelado de transferencia de datos

Reúna los requisitos de la organización y realice un modelado de transferencia de datos de la carga de trabajo y de cada uno de sus componentes. Se identifica el punto de costo más bajo para los requisitos de transferencia de datos actuales.

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

Guía para la implementación

Cuando se diseña una solución en la nube, las tarifas de transferencia de datos suelen olvidarse. Esto se debe a la costumbre de diseñar la arquitectura utilizando centros de datos locales o a la falta de conocimientos. Los cargos por transferencia de datos en AWS vienen determinados por el origen, el destino y el volumen del tráfico. Si se tienen en cuenta estas tarifas durante la fase de diseño, es posible ahorrar costes. Saber dónde se produce la transferencia de datos en su carga de trabajo, el coste de la transferencia y su beneficio asociado es muy importante para calcular con precisión el coste total de propiedad (TCO). De este modo, podrá tomar una decisión informada para modificar o aceptar la decisión arquitectónica. Por ejemplo, podría tener una configuración de la zona de disponibilidad múltiple donde replicar los datos entre las zonas de disponibilidad.

Después, modele los componentes de los servicios que transfieren los datos en su carga de trabajo y determine que se trata de un coste aceptable (similar al de pagar por la computación y el almacenamiento en la zona de disponibilidad) para lograr la fiabilidad y resiliencia requeridas. Debe modelar los costes según los distintos niveles de uso. El uso de la carga de trabajo puede cambiar con el tiempo y algunos servicios podrían ser más rentables en diferentes niveles.

Al modelar la transferencia de datos, piense en la cantidad de datos que se ingieren y de dónde provienen esos datos. Además, considere cuántos datos se procesan y cuánta capacidad de almacenamiento o computación se necesita. Durante el modelado, siga las prácticas recomendadas de redes para la arquitectura de su carga de trabajo con el fin de optimizar los costes potenciales de la transferencia de datos.

Calculadora de precios de AWS puede ayudarle a conocer los costes estimados de servicios de AWS específicos y la transferencia de datos esperada. Si ya tiene una carga de trabajo en ejecución (con fines de prueba o en un entorno de preproducción), utilice AWS Cost Explorer o AWS Cost and Usage Report (CUR) para conocer y modelar sus costes de transferencia de datos. Configure una prueba de concepto (POC) o pruebe su carga de trabajo y ejecute una prueba con una carga simulada realista. Puede modelar sus costes con distintas demandas de carga de trabajo.

Pasos para la implementación

- Identifique los requisitos: ¿cuál es el objetivo principal y los requisitos empresariales para la transferencia de datos planificada entre el origen y el destino? ¿Cuál es el resultado empresarial que se espera al final? Recopile los requisitos empresariales y defina los resultados esperados.
- Identifique el origen y el destino: ¿cuál es el origen y el destino de los datos para la transferencia de datos? Por ejemplo, dentro de Regiones de AWS, a servicios de AWS o a Internet.
  - Data transfer within an Región de AWS

  - Data transfer between Regiones de AWS
  - Data transfer out to the internet
- Identifique las clasificaciones de datos: ¿cuál es la clasificación de datos de esta transferencia de datos? ¿Qué tipo de datos son? ¿Qué tamaño tienen los datos? ¿Con qué frecuencia se deben transferir los datos? ¿Los datos son confidenciales?
- Identifique los servicios de AWS o herramientas que se van a utilizar: ¿qué servicios de AWS se utilizan para esta transferencia de datos? ¿Es posible utilizar un servicio ya aprovisionado para otra carga de trabajo?
- Calcule los costes de transferencia de datos: utilice los precios de AWS y el modelo de transferencia de datos que creó anteriormente para calcular los costes de transferencia de datos de la carga de trabajo. Calcule los costes de transferencia de datos en distintos niveles de uso para los incrementos y las reducciones del uso de la carga de trabajo. Si hay múltiples opciones para la arquitectura de la carga de trabajo, calcule el coste de cada opción para compararlas.
- Vincule costes y resultados: especifique el resultado obtenido por la carga de trabajo para cada coste de transferencia de datos incurrido. Si es una transferencia entre componentes, puede deberse a un desacoplamiento y, si es entre zonas de disponibilidad, puede deberse a la redundancia.
- Cree modelos de transferencia de datos: después de recopilar toda la información, cree un modelo de transferencia de datos de base conceptual para varios casos de uso y diferentes cargas de trabajo.

Recursos

Documentos relacionados:

- Soluciones de almacenamiento en caché de AWS
- Precios de AWS
- Precios de Amazon EC2
- Precios de Amazon VPC
- Understanding data transfer charges

Vídeos relacionados:

- Supervisión y optimización de los costes de transferencia de datos
- S3 Transfer Acceleration

Ejemplos relacionados:

- Overview of Data Transfer Costs for Common Architectures (Información general de los costes de transferencia de datos para arquitecturas comunes)
- «AWS Prescriptive Guidance for Networking»

COST08-BP02 Seleccionar componentes para optimizar el coste de la transferencia de datos

Se seleccionan todos los componentes y se diseña la arquitectura para reducir los costos de transferencia de datos. Incluye el uso de componentes como la optimización de la red de área extensa (WAN) y las configuraciones de varias zonas de disponibilidad (AZ).

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

Guía para la implementación

La arquitectura para la transferencia de datos minimiza los costes de transferencia de datos. Para ello, es posible que deba usar redes de entrega de contenido para colocar los datos más cerca de los usuarios, o bien enlaces de red dedicados entre sus instalaciones y AWS. También puede usar la optimización de WAN y de las aplicaciones para reducir la cantidad de datos que se transfieren entre los componentes.

Al transferir datos a Nube de AWS, o dentro de ella, es muy importante conocer el destino en función de los diversos casos de uso, la naturaleza de los datos y los recursos de red disponibles para seleccionar los servicios de AWS correctos para optimizar la transferencia de datos. AWS ofrece una amplia gama de servicios de transferencia de datos diseñados para satisfacer diversos requisitos de migración de datos. Seleccione las opciones adecuadas de almacenamiento y transferencia de datos en función de las necesidades empresariales de su organización.

Al planificar o revisar la arquitectura de la carga de trabajo, tenga en cuenta lo siguiente:

- Use puntos de conexión de VPC dentro de AWS: los puntos de conexión de VPC permiten conexiones privadas entre su VPC y los servicios de AWS compatibles. Esto le permite evitar el uso de la Internet pública, que puede dar lugar a costes de transferencia de datos.
- Use una puerta de enlace NAT: use una puerta de enlace NAT para que las instancias de una subred privada puedan conectarse a Internet o a los servicios fuera de su VPC. Compruebe si los recursos que hay detrás de la puerta de enlace NAT que envían la mayor cantidad de tráfico se encuentran en la misma zona de disponibilidad que la puerta de enlace NAT. Si no lo están, cree

nuevas puertas de enlace NAT en la misma zona de disponibilidad que el recurso para reducir los cargos por transferencia de datos entre AZ.

- El uso de AWS Direct Connect Direct Connect omite la Internet pública y establece una conexión directa y privada entre su red local y AWS. Esto puede resultar más rentable y coherente que la transferencia de grandes volúmenes de datos a través de Internet.
- Evite transferir datos a través de las fronteras regionales: las transferencias de datos entre Regiones de AWS (de una región a otra) suelen incurrir en cargos. La vía multirregional debería ser una decisión muy meditada. Para obtener más información, consulte «Situaciones de varias regiones».
- Supervise la transferencia de datos: utilice Amazon CloudWatch y registros de flujo de VPC para obtener detalles sobre la transferencia de datos y el uso de la red. Analice la información de tráfico de red de sus VPC, como la dirección IP o el rango, que va y viene de las interfaces de red.
- Analice el uso de la red: utilice herramientas de medición y generación de informes como AWS Cost Explorer, los paneles CUDOS o CloudWatch, para conocer el coste de transferencia de datos de su carga de trabajo.

Pasos para la implementación

- Seleccione los componentes para la transferencia de datos: use el modelado de transferencia de datos que se explica en COST08-BP01 Realizar un modelado de transferencia de datos para centrarse en dónde se encuentran los mayores costes de transferencia de datos o dónde estarían si cambia el uso de la carga de trabajo. Busque arquitecturas alternativas o componentes adicionales que eliminen o reduzcan la necesidad de transferir datos (o que reduzcan su coste).

Recursos

Prácticas recomendadas relacionadas:

- COST08-BP01 Realizar un modelado de transferencia de datos
- COST08-BP03 Implementar servicios para reducir los costes de transferencia de datos

Documentos relacionados:

- Migración de datos a la nube
- Soluciones de almacenamiento en caché de AWS
- «Deliver content faster with Amazon CloudFront»

Ejemplos relacionados:

- Overview of Data Transfer Costs for Common Architectures (Información general de los costes de transferencia de datos para arquitecturas comunes)
- «AWS Network Optimization Tips»
- Optimize performance and reduce costs for network analytics with VPC Flow Logs in Apache Parquet format (Optimice el rendimiento y reduzca los costes de los análisis de red con los registros de flujo de VPC en formato Apache Parquet)

COST08-BP03 Implementar servicios para reducir los costes de transferencia de datos

Implemente servicios para reducir la transferencia de datos. Por ejemplo, utilice ubicaciones periféricas o redes de entrega de contenido (CDN) para ofrecer contenido a los usuarios finales, cree capas de almacenamiento en caché delante de sus servidores de aplicaciones o bases de datos y utilice conexiones de red dedicadas en lugar de VPN para conectarse a la nube.

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

Guía para la implementación

Existen varios servicios de AWS que pueden ayudarle a optimizar el uso de la transferencia de datos de la red. Según los componentes de la carga de trabajo, el tipo y la arquitectura de la nube, estos servicios pueden ayudar a comprimir, almacenar en caché y compartir y distribuir su tráfico en la nube.

- Amazon CloudFront es una red global de entrega de contenido que brinda datos con baja latencia y altas velocidades de transferencia. Almacena en caché los datos en ubicaciones de borde en todo el mundo, lo que reduce la carga de sus recursos. Con CloudFront, puede reducir el esfuerzo administrativo que supone ofrecer contenido a una gran cantidad de usuarios a nivel mundial y hacerlo con una latencia mínima. El paquete de ahorro en seguridad puede ayudarle a ahorrar hasta un 30 % en el uso de CloudFront si planea un aumento de este con el tiempo.
- AWS Direct Connect permite establecer una conexión de red dedicada con AWS. Esto puede reducir los costos de red, aumentar el ancho de banda y brindar una experiencia de red más coherente que las conexiones basadas en Internet.
- Site-to-Site VPN le permite establecer una conexión segura y privada entre su red privada y la red global de AWS. Es ideal para oficinas pequeñas o socios empresariales porque proporciona una conectividad simplificada, además de ser un servicio elástico y completamente administrado.

- Puntos de conexión de VPC permiten la conexión entre los servicios de AWS a través de la red privada y se pueden usar para reducir los costes de la transferencia de datos pública y de la puerta de enlace de NAT. Los puntos de conexión de VPC de la puerta de enlace no se cobran por hora y son compatibles con Amazon S3 y Amazon DynamoDB. Los puntos de conexión de VPC de la interfaz los proporciona AWS PrivateLink y tienen un tarifa por hora y un coste de uso por GB.
- Las puertas de enlace de NAT ofrecen escalado y administración integrados para reducir los costes, a diferencia de una instancia de NAT independiente. Coloque las puertas de enlace de NAT en las mismas zonas de disponibilidad que las instancias de alto tráfico y plantéese usar puntos de conexión de VPC para las instancias que necesiten acceder a Amazon DynamoDB o Amazon S3 a fin de reducir los costes de transferencia y procesamiento de datos.
- Utilice dispositivos de AWS Snow Family que tengan recursos de computación para recopilar y procesar datos en la periferia. Los dispositivos de AWS Snow Family (Snowball Edge, Snowball Edge y Snowmobile) le permiten transferir petabytes de datos a la Nube de AWS de forma rentable y sin conexión.

Pasos para la implementación

- Implementar servicios: seleccione los servicios de red de AWS aplicables en función del tipo de carga de trabajo del servicio mediante el modelado de transferencia de datos y la revisión de registros de flujo de VPC. Observe dónde están los mayores costes y los mayores flujos de volumen. Revise los servicios de AWS y evalúe si existe un servicio que reduzca o elimine la transferencia, especialmente en relación con la entrega de contenido y las redes. Busque también servicios de almacenamiento en caché donde haya acceso repetido a los datos o grandes cantidades de datos.

Recursos

Documentos relacionados:

- AWS Direct Connect
- Explore nuestros productos de AWS
- AWS caching solutions
- Amazon CloudFront
- AWS Snow Family
- Paquete de ahorro de seguridad Amazon CloudFront

Vídeos relacionados:

- Supervisión y optimización de los costes de transferencia de datos
- AWS Cost Optimization Series: CloudFront
- ¿Cómo puedo reducir los cargos de transferencia de datos de mi puerta de enlace de NAT?

Ejemplos relacionados:

- How-to chargeback shared services: An AWS Transit Gateway example
- Understand AWS data transfer details in depth from cost and usage report using Athena query and QuickSight
- Overview of Data Transfer Costs for Common Architectures (Información general de los costes de transferencia de datos para arquitecturas comunes)
- Using AWS Cost Explorer to analyze data transfer costs
- Cost-Optimizing your AWS architectures by utilizing Amazon CloudFront features
- ¿Cómo puedo reducir los cargos de transferencia de datos de mi puerta de enlace de NAT?

Administración de la demanda y suministro de recursos

Pregunta

- COSTE 9. ¿Cómo administra la demanda y aprovisiona los recursos?

COSTE 9. ¿Cómo administra la demanda y aprovisiona los recursos?

Para una carga de trabajo que tenga un gasto y un rendimiento equilibrados, compruebe que en todo lo que invierte se utiliza y evite las instancias de infrautilización significativas. Una métrica de utilización sesgada en cualquier dirección tiene un efecto adverso en su organización, ya sea en los costes operativos (rendimiento degradado debido a la sobreutilización) o en los gastos desperdiciados de AWS (debido al sobreaprovisionamiento).

Prácticas recomendadas

- COST09-BP01 Realizar un análisis de la demanda de la carga de trabajo
- COST09-BP02 Despliegue un buffer o un acelerador para gestionar la demanda de la carga de trabajo
- COST09-BP03 Aprovisionar recursos de forma dinámica

COST09-BP01 Realizar un análisis de la demanda de la carga de trabajo

Analice la demanda de la carga de trabajo a lo largo del tiempo. Compruebe que el análisis cubra las tendencias estacionales y represente con precisión las condiciones de servicio durante toda la vida útil de la carga de trabajo. El análisis debe reflejar los posibles beneficios; por ejemplo, el tiempo empleado es proporcional al coste de la carga de trabajo.

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

Guía para la implementación

El análisis de la demanda de la carga de trabajo para la computación en la nube implica comprender los patrones y las características de las tareas de computación que se inician en el entorno de la nube. Este análisis ayuda a los usuarios a optimizar la asignación de recursos, administrar los costes y verificar que el rendimiento cumpla con los niveles requeridos.

Debe conocer los requisitos de la carga de trabajo. Los requisitos de su organización deben indicar los tiempos de respuesta de la carga de trabajo frente a las solicitudes. El tiempo de respuesta se puede usar para determinar si la demanda está administrada o si el suministro de recursos debe cambiar para adaptarse a la demanda.

El análisis debe incluir la previsibilidad y la repetibilidad de la demanda, el ritmo y la cantidad de cambio en la demanda. Realice el análisis a lo largo de un período suficiente de tiempo para que incorpore variantes estacionales, como un procesamiento de final de mes o los picos de las vacaciones.

La actividad de análisis debe reflejar los posibles beneficios de la implementación del escalado. Consulte el coste total previsto del componente y cualquier incremento o descenso del uso, así como el coste durante el período de vida de la carga de trabajo.

Estos son algunos aspectos clave que se deben tener en cuenta al realizar un análisis de la demanda de la carga de trabajo para la computación en la nube:

1. Métricas de uso y rendimiento de los recursos: analice cómo se utilizan los recursos de AWS a lo largo del tiempo. Determine los patrones de uso en las horas punta y fuera de las horas punta para optimizar la asignación de recursos y las estrategias de escalado. Supervise las métricas de rendimiento, como los tiempos de respuesta, la latencia, el rendimiento y las tasas de error. Estas métricas ayudan a evaluar el estado general y la eficiencia de la infraestructura de la nube.
2. Comportamiento de escalado de usuarios y aplicaciones: comprenda el comportamiento de los usuarios y cómo afecta a la demanda de la carga de trabajo. El examen de los patrones del

tráfico de usuarios ayuda a mejorar la entrega de contenido y la capacidad de respuesta de las aplicaciones. Analice cómo se escalan las cargas de trabajo a medida que aumenta la demanda. Determine si los parámetros de escalado automático están configurados de forma correcta y eficaz para gestionar las fluctuaciones de carga.

3. Tipos de carga de trabajo: identifique los diferentes tipos de cargas de trabajo que se ejecutan en la nube, como el procesamiento por lotes, el procesamiento de datos en tiempo real, las aplicaciones web, las bases de datos o el machine learning. Cada tipo de carga de trabajo puede tener requisitos de recursos y perfiles de rendimiento diferentes.
4. Acuerdos de nivel de servicio (SLA): compare el rendimiento real con los SLA para garantizar el cumplimiento e identificar las áreas que necesitan mejoras.

Puede usar el Amazon CloudWatch para recopilar y realizar un seguimiento de las métricas, supervisar los archivos de registro, configurar alarmas y reaccionar automáticamente a los cambios en los recursos de AWS. También puede usar Amazon CloudWatch para lograr visibilidad de la utilización de los recursos de todo el sistema, el rendimiento de las aplicaciones y el estado operativo.

Con AWS Trusted Advisor, puede aprovisionar sus recursos conforme a las prácticas recomendadas para mejorar el rendimiento y la fiabilidad del sistema, aumentar la seguridad y buscar oportunidades para ahorrar dinero. También puede desactivar las instancias que no son de producción y utilizar Amazon CloudWatch y Auto Scaling para adaptarlas a los aumentos o reducciones de la demanda.

Por último, puede usar AWS Cost Explorer o bien Quick con el archivo del AWS Cost and Usage Report (CUR) o los registros de la aplicación para realizar un análisis avanzado de la demanda de la carga de trabajo.

En general, un análisis integral de la demanda de la carga de trabajo permite a las organizaciones tomar decisiones informadas sobre el aprovisionamiento, el escalado y la optimización de los recursos, lo que se traduce en un mejor rendimiento, rentabilidad y satisfacción de los usuarios.

Pasos para la implementación

- Analizar los datos de la carga de trabajo existente: analice los datos de la carga de trabajo existente, las versiones anteriores de la carga de trabajo o los patrones de uso previstos. Utilice Amazon CloudWatch, los archivos de registro y los datos de supervisión para obtener información sobre cómo se utilizó la carga de trabajo. Analice un ciclo completo de la carga de trabajo y recopile datos de los cambios estacionales, como los eventos de final de mes o de final de año. El esfuerzo reflejado en este análisis debe mostrar las características de la carga de trabajo.

Debe ponerse mayor empeño en las cargas de trabajo de mayor valor con mayores cambios en la demanda. Debe ponerse menor empeño en las cargas de trabajo de menor valor con menores cambios en la demanda.

- Prever la influencia exterior: reúnase con miembros de equipos de toda la organización que puedan influir o cambiar la demanda de la carga de trabajo. Estos equipos suelen ser los de Ventas, Marketing o Desarrollo empresarial. Colabore con ellos para conocer los ciclos en los que operan y si hay eventos especiales que puedan cambiar la demanda de la carga de trabajo. Haga una previsión de la demanda de la carga de trabajo con estos datos.

Recursos

Documentos relacionados:

- Amazon CloudWatch
- AWS Trusted Advisor
- AWS X-Ray
- AWS Auto Scaling
- AWS Instance Scheduler
- Getting started with Amazon SQS (Introducción a Amazon SQS)
- AWS Cost Explorer
- Quick

Vídeos relacionados:

Ejemplos relacionados:

- Supervisión, seguimiento y análisis para optimizar los costes
- Searching and analyzing logs in CloudWatch

COST09-BP02 Despliegue un buffer o un acelerador para gestionar la demanda de la carga de trabajo

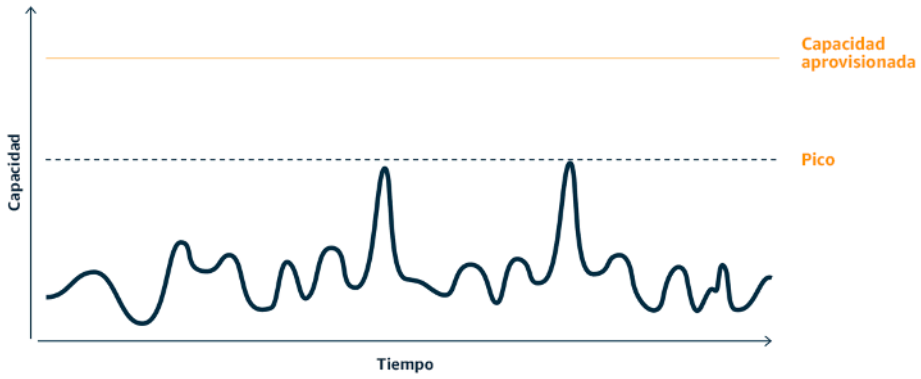
El almacenamiento en búfer y la limitación modifican la demanda de la carga de trabajo y suavizan los picos. Implemente limitaciones cuando sus clientes realicen reintentos. Implemente el almacenamiento en búfer para almacenar la solicitud y aplazar el procesamiento para más adelante.

Verifique que las limitaciones y los búferes se hayan diseñado de tal manera que los clientes reciban una respuesta en el tiempo requerido.

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

Guía para la implementación

La implementación de un búfer o limitador es crucial en la computación en la nube para administrar la demanda y reducir la capacidad aprovisionada necesaria para la carga de trabajo. Para conseguir un rendimiento óptimo, es esencial evaluar la demanda total, incluidos los picos, el ritmo de cambio en las solicitudes y el tiempo de respuesta necesario. Cuando los clientes tienen la posibilidad de reenviar sus solicitudes, es práctico aplicar la limitación. Por el contrario, para los clientes que carecen de funcionalidades de reintento, lo ideal es implementar una solución de búfer. Estos búferes agilizan la afluencia de solicitudes y optimizan la interacción de las aplicaciones con diferentes velocidades operativas.



Curva de demanda con dos picos diferenciados que requieren una alta capacidad aprovisionada

Supongamos que tenemos una carga de trabajo con la curva de demanda que se muestra en la imagen anterior. Esta carga de trabajo tiene dos picos y, para gestionarlos, se aprovisiona la capacidad de recursos que muestra la línea naranja. Los recursos y la energía utilizados para esta carga de trabajo no están indicados en el área situada debajo de la curva de demanda, sino en el área situada debajo de la línea de capacidad aprovisionada, ya que esta capacidad es la que se necesita para gestionar esos dos picos. El aplanamiento de la curva de demanda de la carga de trabajo puede ayudarle a reducir la capacidad aprovisionada para una carga de trabajo y a reducir su impacto medioambiental. Para suavizar el pico, considere la posibilidad de implementar una solución de limitación o almacenamiento en búfer.

Vamos a profundizar en las limitaciones y el almacenamiento en búfer para entenderlos mejor.

Limitación: si el origen de la demanda tiene capacidad de reintento, puede implementar una limitación. La limitación le dice al origen que, si no puede atender la solicitud en ese momento, debe intentarlo más tarde. El origen espera un tiempo y vuelve a intentar la solicitud. Implementar una limitación tiene la ventaja de que se limita la cantidad máxima de recursos y costes de la carga de trabajo. En AWS, puede usar Amazon API Gateway para implementar la limitación.

Basado en búfer: un enfoque basado en búfer utiliza productores (componentes que envían mensajes a la cola), consumidores (componentes que reciben mensajes de la cola) y una cola (que contiene mensajes) para almacenar los mensajes. De este modo, los consumidores pueden leer y procesar los mensajes, lo que permite que dichos mensajes se ejecuten a la velocidad que cumpla con los requisitos empresariales de los consumidores. Al utilizar una metodología centrada en los búferes, los mensajes de los productores se alojan en colas o secuencias, listos para que los consumidores accedan a ellos a un ritmo que se ajuste a sus demandas operativas.

EnAWS, puede elegir entre varios servicios para implementar un enfoque de almacenamiento en búfer. Amazon Simple Queue Service(Amazon SQS) es un servicio administrado que proporciona colas que permiten a un solo consumidor leer mensajes individuales. Amazon Kinesis proporciona una secuencia que permite que muchos consumidores lean los mismos mensajes.

El almacenamiento en búfer y las limitaciones pueden suavizar cualquier pico al modificar la demanda de la carga de trabajo. Utilice limitaciones cuando los clientes vuelvan a intentar realizar acciones y utilice el almacenamiento en búfer para retener la solicitud y procesarla más adelante. Al trabajar con un enfoque basado en búfer, diseñe su carga de trabajo para atender la solicitud en el tiempo requerido y verifique que pueda gestionar las solicitudes duplicadas. Analice la demanda general, la tasa de cambio y el tiempo de respuesta requerido para dimensionar correctamente la limitación o el búfer requeridos.

Pasos para la implementación

- Analice los requisitos del cliente: analice las solicitudes de los clientes para determinar si son capaces de realizar reintentos. Para los clientes que no puedan realizarlos, deberán implementarse búferes. Analice la demanda general, el ritmo de cambio y el tiempo de respuesta requerido para determinar el tamaño de la limitación o del búfer requeridos.
- Implemente un búfer o un limitador: implemente un búfer o un limitador en la carga de trabajo. Una cola como Amazon Simple Queue Service (Amazon SQS) puede proporcionar un búfer a sus componentes de la carga de trabajo. Amazon API Gateway puede proporcionar limitación a los componentes de la carga de trabajo.

Recursos

Prácticas recomendadas relacionadas:

- SUS02-BP06 Implementar el almacenamiento en búfer o la limitación para aplanar la curva de demanda
- REL05-BP02 Limitar las solicitudes

Documentos relacionados:

- AWS Auto Scaling
- AWS Instance Scheduler
- Amazon API Gateway
- Amazon Simple Queue Service
- «Getting started with Amazon SQS»
- Amazon Kinesis

Vídeos relacionados:

- Choosing the Right Messaging Service for Your Distributed App (Elección del servicio de mensajería correcto para su aplicación distribuida)

Ejemplos relacionados:

- Managing and monitoring API throttling in your workloads (Administrar y supervisar la limitación de las API en sus cargas de trabajo)
- «Throttling a tiered, multi-tenant REST API at scale using API Gateway»
- «Enabling Tiering and Throttling in a Multi-Tenant Amazon EKS SaaS Solution Using Amazon API Gateway»
- Application integration Using Queues and Messages (Integración de aplicaciones mediante colas y mensajes)

COST09-BP03 Aprovisionar recursos de forma dinámica

Los recursos se aprovisionan de manera planificada. Esto puede basarse en la demanda (por ejemplo, mediante el escalamiento automático) o en el tiempo, donde la demanda es predecible y

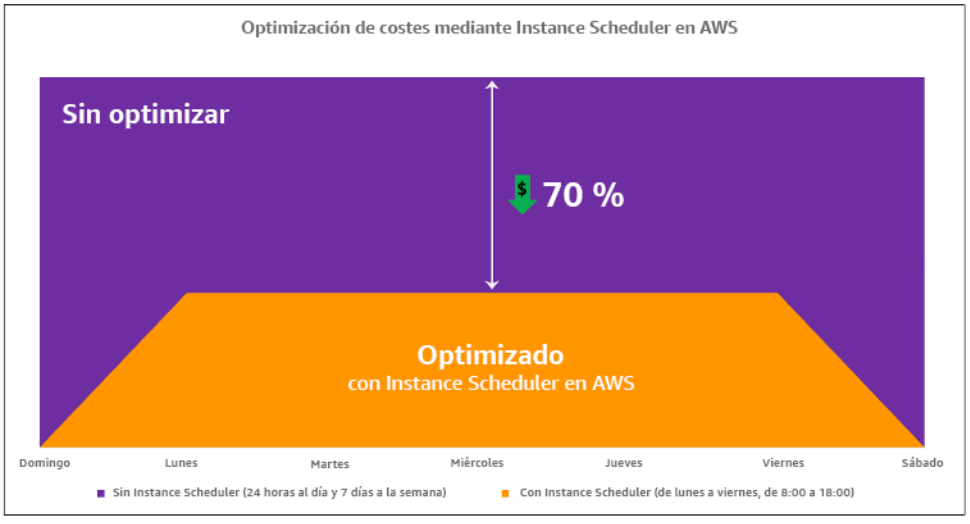
los recursos se proporcionan en función del tiempo. Estos métodos conllevan la menor cantidad de aprovisionamiento excesivo o insuficiente.

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

Guía para la implementación

Hay varias formas en que los clientes de AWS pueden aumentar los recursos disponibles para sus aplicaciones y suministrar recursos para satisfacer la demanda. Una de estas opciones es usar AWS Instance Scheduler, que automatiza el inicio y la parada de las instancias de Amazon Elastic Compute Cloud (Amazon EC2) y Amazon Relational Database Service (Amazon RDS). La otra opción es usar AWS Auto Scaling, lo que le permite escalar automáticamente los recursos de computación en función de la demanda de la aplicación o servicio. Suministrar recursos en función de la demanda le permitirá pagar únicamente por los recursos que utilice y reducir los costes, ya que solo lanza los recursos cuando se necesitan y los cancela cuando no.

AWS Instance Scheduler le permite configurar la detención y el inicio de sus instancias de Amazon EC2 y Amazon RDS en momentos definidos para que pueda satisfacer la demanda de los mismos recursos según un patrón temporal coherente; por ejemplo, que todos los días los usuarios accedan a las ocho de la mañana a instancias de Amazon EC2 que no se necesitan después de las seis de la tarde. Esta solución contribuye a reducir los costes operativos, ya que se detienen los recursos que no están en uso y se ponen en marcha cuando se necesitan.



Optimización de costes con AWS Instance Scheduler.

También puede configurar fácilmente los horarios de sus instancias de Amazon EC2 en todas sus cuentas y regiones con una interfaz de usuario (IU) sencilla mediante la configuración rápida de

AWS Systems Manager. Puede programar instancias de Amazon EC2 o Amazon RDS con AWS Instance Scheduler y detener e iniciar las instancias existentes. Sin embargo, no puede detener ni iniciar instancias que formen parte de su grupo de Auto Scaling (ASG) o que administren servicios como Amazon Redshift o Amazon OpenSearch Service. Los grupos de Auto Scaling tienen su propia programación para las instancias del grupo y la creación de estas instancias.

AWS Auto Scaling le ayuda a ajustar la capacidad para mantener un rendimiento predecible y estable al menor coste posible para satisfacer los cambios en la demanda. Es un servicio totalmente administrado y gratuito para escalar la capacidad de su aplicación, que se integra con las instancias y las flotas de spot de Amazon EC2, Amazon ECS, Amazon DynamoDB y Amazon Aurora. Auto Scaling detecta los recursos automáticamente, lo que le ayuda a buscar recursos en su carga de trabajo que se pueden configurar. Además, tiene estrategias de escalamiento integradas para optimizar el rendimiento y los costes, o un equilibrio entre ambos, y proporciona escalamiento predictivo para ayudar en los picos que se producen periódicamente.

Hay varias opciones de escalamiento disponibles para escalar su grupo de Auto Scaling:

- Mantener siempre los niveles de instancia actuales
- Escalar manualmente
- Escalar en función de una programación
- Escalar en función de la demanda
- Usar el escalamiento predictivo

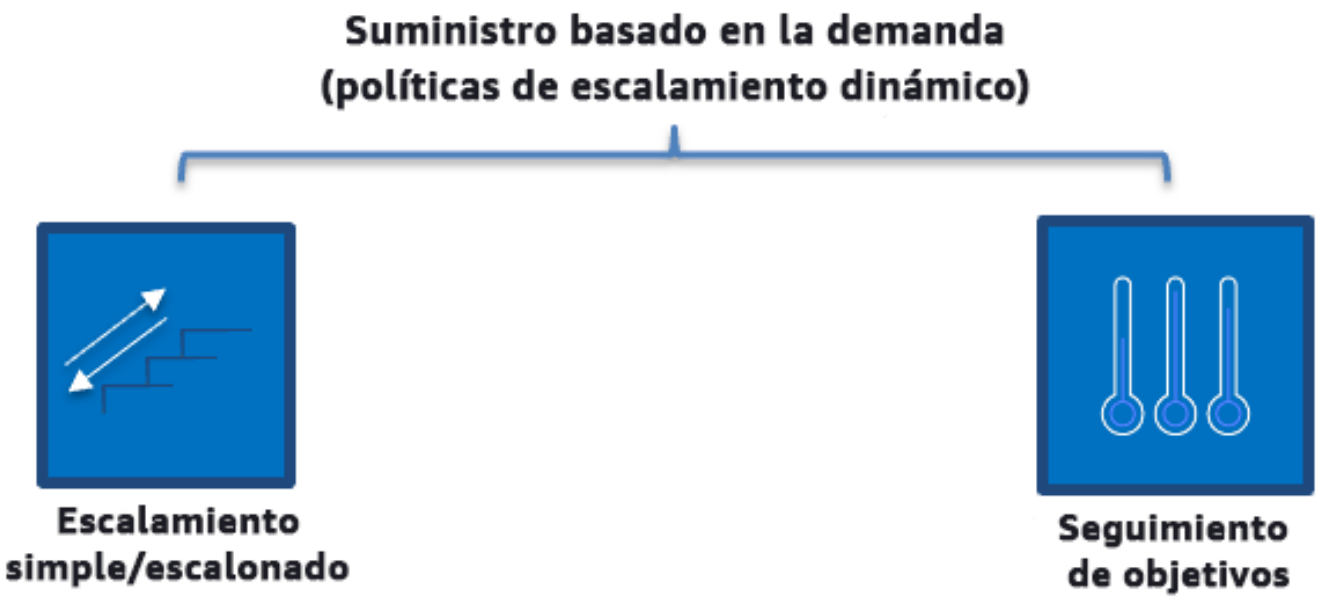
Existen diferentes políticas de Auto Scaling. Se pueden clasificar en políticas de escalamiento dinámicas y programadas. Las políticas dinámicas son escalamientos manuales o dinámicos, que pueden ser programados o predictivos. Puede utilizar políticas de escalamiento para un escalamiento dinámico, programado y predictivo. También puede usar las métricas y las alarmas de Amazon CloudWatch para desencadenar eventos de escalamiento para la carga de trabajo. Le recomendamos que utilice plantillas de lanzamiento, que permiten acceder a las características y mejoras más recientes. No todas las características de Auto Scaling están disponibles cuando se utilizan configuraciones de lanzamiento. Por ejemplo, no puede crear un grupo de Auto Scaling que lance instancias de spot y bajo demanda o que especifique varios tipos de instancias. Debe utilizar una plantilla de lanzamiento para configurar estas características. Cuando utilice plantillas de lanzamiento, le recomendamos que realice un control de versiones en cada una de ellas. Con el control de versiones de las plantillas de lanzamiento, puede crear un subconjunto del conjunto completo de parámetros. Luego, puede volver a utilizarlo para crear otras versiones de la misma plantilla de lanzamiento.

Puede usar AWS Auto Scaling o incorporar el escalamiento en su código con las API o SDK de AWS. Esto reduce los costes generales de la carga de trabajo al eliminar el coste operativo de realizar los cambios manualmente en su entorno. Además, los cambios se pueden realizar mucho más rápido. De este modo, también se adapta la dotación de recursos de la carga de trabajo a su demanda en cualquier momento. Para seguir esta práctica recomendada y suministrar recursos de forma dinámica a su organización, debe comprender el escalamiento horizontal y vertical en la Nube de AWS, así como la naturaleza de las aplicaciones que se ejecutan en las instancias de Amazon EC2. Es mejor que su equipo de administración financiera en la nube colabore con los equipos técnicos para seguir esta práctica recomendada.

Elastic Load Balancing (Elastic Load Balancing) le ayuda a escalar mediante la distribución de la demanda entre múltiples recursos. Con ASG y Elastic Load Balancing, puede administrar las solicitudes entrantes mediante el enrutamiento óptimo del tráfico para que ninguna instancia se sobrecargue en un grupo de Auto Scaling. Las solicitudes se distribuirían entre todos los destinatarios de un grupo objetivo por turnos, sin tener en cuenta la capacidad ni la utilización.

Las métricas habituales pueden ser métricas de Amazon EC2 estándar, como el uso de la CPU, el rendimiento de la red y la latencia de solicitud y respuesta observada de Elastic Load Balancing. Si es posible, debe usar una métrica indicativa de la experiencia del cliente. Suele ser una métrica personalizada que se puede originar en el código de la aplicación en la carga de trabajo. Para explicar cómo satisfacer la demanda de forma dinámica, vamos a agrupar Auto Scaling en dos categorías, modelos de suministro basados en la demanda y modelos de suministro basados en el tiempo, y analizaremos en profundidad cada una de ellas.

Suministro basado en la demanda: aproveche la elasticidad de la nube para suministrar recursos que satisfagan los cambios en la demanda utilizando el estado de la demanda casi en tiempo real. Para el suministro basado en la demanda, utilice las API o las características del servicio para cambiar mediante programación la cantidad de recursos en la nube de su arquitectura. De esta forma, puede escalar componentes en su arquitectura y aumentar la cantidad de recursos durante los picos de demanda para mantener el rendimiento, así como disminuir la capacidad cuando la demanda disminuya para reducir los costes.

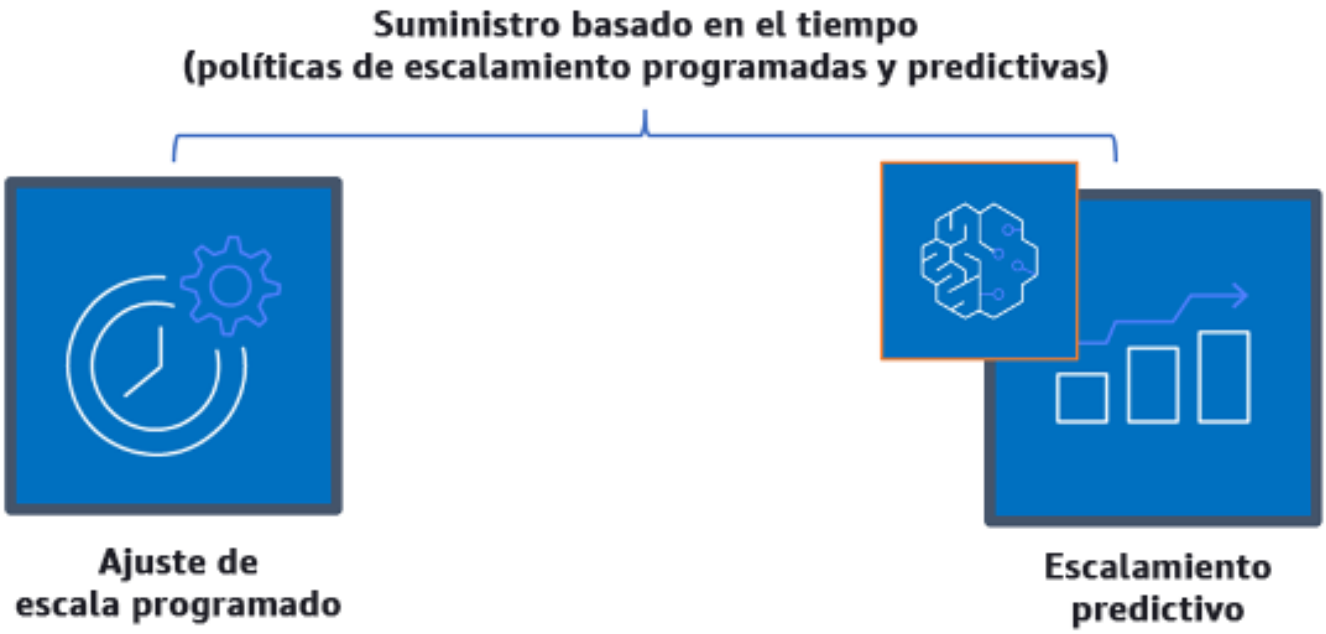


Políticas de escalamiento dinámico basadas en la demanda

- Escalamiento simple/escalonado: monitoriza las métricas y añade o elimina instancias de acuerdo con los pasos definidos manualmente por los clientes.
- Seguimiento de objetivos: mecanismo de control similar a un termostato que añade o elimina instancias automáticamente para mantener las métricas en un objetivo definido por el cliente.

Al diseñar con un enfoque basado en la demanda, tenga en cuenta dos consideraciones clave. La primera: debe conocer la rapidez con la que necesita aprovisionar recursos nuevos. La segunda: tenga en cuenta que el tamaño del margen entre la oferta y la demanda cambiará. Debe estar preparado para poder hacer frente a la velocidad del cambio en la demanda y también a los errores de recursos.

Suministro basado en el tiempo: el enfoque basado en el tiempo adapta la capacidad de los recursos a una demanda que es predecible o que está bien definida por el tiempo. Normalmente, este enfoque no depende de los niveles de utilización de los recursos. El enfoque basado en tiempo garantiza que los recursos estén disponibles en el momento específico en que se necesiten y que se puedan proporcionar sin retrasos debidos a los procedimientos de inicio y comprobaciones del sistema o de coherencia. Con el enfoque basado en tiempo, puede brindar recursos adicionales o aumentar la capacidad durante los periodos de mayor actividad.



Políticas de escalamiento basadas en el tiempo

Puede utilizar el escalamiento automático programado o predictivo para implementar un enfoque basado en el tiempo. Las cargas de trabajo se pueden programar para escalarse o desescalarse horizontalmente en momentos definidos (como el inicio del horario laboral). De este modo, los recursos están disponibles cuando lleguen los usuarios o aumente la demanda. El escalamiento predictivo utiliza patrones para escalar horizontalmente, mientras que el escalamiento programado utiliza tiempos predefinidos para escalar horizontalmente. También puede utilizar la estrategia de selección de tipos de instancias (ABS) basada en atributos en grupos de Auto Scaling, lo que permite expresar los requisitos de la instancia como un conjunto de atributos, como la vCPU, la memoria y el almacenamiento. De este modo, también puede utilizar automáticamente los tipos de instancia de nueva generación cuando se lancen y acceder a una gama más amplia de capacidad con las instancias de spot de Amazon EC2. La flota de Amazon EC2 y Amazon EC2 Auto Scaling seleccionan y lanzan instancias que se ajusten a los atributos especificados, por lo que no es necesario elegir manualmente los tipos de instancia.

También puede utilizar las API y los SDK de AWS y AWS CloudFormation para aprovisionar y retirar entornos completos de manera automática según sus necesidades. Este enfoque es ideal para los entornos de desarrollo o pruebas que se ejecutan únicamente en horarios laborales o periodos definidos. Puede usar API para escalar el tamaño de los recursos dentro de un entorno (escalado

vertical). Por ejemplo, puede escalar verticalmente una carga de trabajo de producción mediante el cambio del tamaño o la clase de instancia. Para ello, hay que detener o iniciar la instancia y seleccionar el tamaño o la clase de instancia diferente. Esta técnica también se puede aplicar a otros recursos, tales como los volúmenes elásticos de Amazon EBS, los cuales se pueden modificar para aumentar el tamaño, ajustar el rendimiento (IOPS) o cambiar el tipo de volumen mientras están en uso.

Al diseñar con un enfoque basado en tiempo, tenga en cuenta dos consideraciones clave. La primera: ¿qué grado de consistencia presenta el patrón? La segunda: ¿en qué afectaría el patrón si cambiara? Puede aumentar la precisión de las predicciones mediante la supervisión de sus cargas de trabajo y el uso de la inteligencia empresarial. Si observa cambios considerables en el patrón de uso, puede ajustar los tiempos para asegurarse de que se proporcione cobertura.

Pasos para la implementación

- Configure el escalamiento programado: en caso de cambios predecibles en la demanda, el escalamiento basado en el tiempo puede proporcionar el número correcto de recursos de manera oportuna. También es útil si la creación y configuración de recursos no es suficientemente rápida a la hora de responder a los cambios en la demanda. Use el análisis de las cargas de trabajo para configurar el escalamiento programado con AWS Auto Scaling. Para configurar la programación en función del tiempo, puede utilizar el escalamiento predictivo del escalamiento programado para aumentar por adelantado el número de instancias de Amazon EC2 de sus grupos de Auto Scaling en función de los cambios de carga previstos o predecibles.
- Configure el escalamiento predictivo: el escalamiento predictivo le permite aumentar el número de instancias de Amazon EC2 de su grupo de Auto Scaling según la previsión de los patrones diarios y semanales de los flujos de tráfico. Si tiene picos de tráfico regulares y aplicaciones que tardan mucho en iniciarse, debería plantearse el uso del escalamiento predictivo. El escalamiento predictivo puede ayudarle a escalar más rápidamente mediante la inicialización de la capacidad antes de la carga prevista si se compara con el escalamiento dinámico únicamente, que es de naturaleza reactiva. Por ejemplo, si los usuarios empiezan a utilizar su carga de trabajo con el inicio del horario laboral y no la utilizan fuera de dicho horario, el escalamiento predictivo puede añadir capacidad antes del horario laboral, lo que elimina el retraso del escalamiento dinámico para reaccionar ante los cambios en el tráfico.
- Configure el escalamiento automático dinámico: para configurar el escalamiento en función de las métricas de las cargas de trabajo activas, utilice Auto Scaling. Use los análisis y configure Auto Scaling para que se lance en los niveles de recursos correctos y verifique que la carga de trabajo se escala en el tiempo requerido. Puede lanzar y escalar automáticamente una flota de instancias

bajo demanda e instancias de spot en un mismo grupo de Auto Scaling. Además de beneficiarse de descuentos por utilizar instancias de spot, puede usar las instancias reservadas o un Savings Plan para obtener descuentos en los precios habituales de las instancias bajo demanda. Todos estos factores combinados le ayudarán a optimizar el ahorro de costes de las instancias de Amazon EC2 y a obtener la escala y el rendimiento que desea para su aplicación.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Instance Scheduler
- Escalar el tamaño de su grupo de Auto Scaling
- Getting Started with Amazon EC2 Auto Scaling (Introducción a Amazon EC2 Auto Scaling)
- Getting started with Amazon SQS (Introducción a Amazon SQS)
- Scheduled Scaling for Amazon EC2 Auto Scaling (Escalamiento programado para Amazon EC2 Auto Scaling)
- Predictive scaling for Amazon EC2 Auto Scaling (Escalamiento predictivo para Amazon EC2 Auto Scaling)

Vídeos relacionados:

- Target Tracking Scaling Policies for Auto Scaling (Políticas de escalamiento de seguimiento de destino para Auto Scaling)
- AWS Instance Scheduler

Ejemplos relacionados:

- Attribute based Instance Type Selection for Auto Scaling for Amazon EC2 Fleet (Selección de tipo de instancia basada en atributos para EC2 Auto Scaling y la flota de EC2)
- Optimizing Amazon Elastic Container Service for cost using scheduled scaling (Optimización de Amazon Elastic Container Service para coste mediante el escalamiento programado)
- Predictive scaling with Amazon EC2 Auto Scaling (Escalamiento predictivo con Amazon EC2 Auto Scaling)

- How do I use Instance Scheduler with CloudFormation to schedule Amazon EC2 instances? (¿Cómo uso Instance Scheduler con AWS CloudFormation para programar instancias de Amazon EC2)

Optimización a lo largo del tiempo

Preguntas

- COSTE 10. ¿Cómo evalúa los servicios nuevos?
- COSTE 11. ¿Cómo evalúa el coste del esfuerzo?

COSTE 10. ¿Cómo evalúa los servicios nuevos?

A medida que AWS presenta nuevos servicios y características, se recomienda que revise sus decisiones de diseño actuales para comprobar que sigan siendo las más rentables.

Prácticas recomendadas

- COST10-BP01 Desarrollo de un proceso de revisión de la carga de trabajo
- COST10-BP02 Revisión y análisis de esta carga de trabajo con regularidad

COST10-BP01 Desarrollo de un proceso de revisión de la carga de trabajo

Desarrolle un proceso que defina los criterios y el proceso para la revisión de las cargas de trabajo. El esfuerzo de revisión debe reflejar la ventaja potencial. Por ejemplo, las cargas de trabajo principales o las cargas de trabajo con un valor por encima del 10 % de la factura se revisan trimestral o semestralmente, mientras que las cargas de trabajo por debajo del 10 % se revisan anualmente.

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

Guía para la implementación

Para tener siempre la carga de trabajo más rentable, debe revisar periódicamente la carga de trabajo para saber si hay oportunidades de implementar nuevos servicios, características y componentes. Para conseguir reducir los costes totales, el proceso debe ser proporcional al volumen potencial de ahorro. Por ejemplo, las cargas de trabajo que suponen el 50 % de sus gastos totales se deben revisar con mayor regularidad y más a fondo que las cargas de trabajo que constituyen el 5 % de

sus gastos totales. Tenga en cuenta los factores externos o la volatilidad. Si la carga de trabajo da servicio a un segmento geográfico o de mercado específico y se prevén cambios en ese ámbito, unas revisiones más frecuentes podrían suponer un ahorro de costes. Otro factor de revisión es el esfuerzo para implementar los cambios. Si las pruebas y la validación de los cambios suponen un coste importante, las revisiones deberían ser menos frecuentes.

Hay que tener en cuenta el coste a largo plazo del mantenimiento de componentes y recursos obsoletos y heredados, y la imposibilidad de implementar nuevas características en ellos. El coste actual de las pruebas y la validación puede superar el beneficio propuesto. Sin embargo, con el tiempo, el coste de realizar el cambio puede aumentar significativamente a medida que se incrementa la brecha entre la carga de trabajo y las tecnologías actuales, lo que se traduce en costes aún mayores. Por ejemplo, el coste de pasar a un nuevo lenguaje de programación puede no ser rentable en la actualidad. No obstante, dentro de cinco años, el coste de las personas con competencias en ese lenguaje puede aumentar y, debido al crecimiento de la carga de trabajo, estaría trasladando un sistema aún mayor al nuevo lenguaje, lo que requeriría un esfuerzo aún mayor que el anterior.

Divida la carga de trabajo en componentes, asigne el coste del componente (basta con una estimación) y, a continuación, enumere los factores (por ejemplo, el esfuerzo y los mercados externos) junto a cada componente. Utilice estos indicadores para determinar la frecuencia de revisión de cada carga de trabajo. Por ejemplo, es posible que los servidores web tengan un coste elevado, un esfuerzo de cambio bajo y unos factores externos elevados, lo que da lugar a una frecuencia de revisión alta. Una base de datos central puede tener un coste medio, un esfuerzo de cambio alto y unos factores externos bajos, lo que da lugar a una frecuencia de revisión media.

Defina un proceso para evaluar nuevos servicios, patrones de diseño, tipos de recursos y configuraciones para optimizar el coste de la carga de trabajo a medida que estén disponibles. De forma similar a los procesos de revisión de los pilares de rendimiento y fiabilidad, identifique, valide y priorice las actividades de optimización y mejora, así como la corrección de problemas, e incorpórelas a sus tareas pendientes.

Pasos para la implementación

- Definir la frecuencia de revisión: defina con qué frecuencia se deben revisar la carga de trabajo y sus componentes. Asigne tiempo y recursos a la mejora continua y revise la frecuencia para mejorar la eficacia y la optimización de su carga de trabajo. Se trata de una combinación de factores y puede diferir de una carga de trabajo a otra en su organización y entre los componentes de la carga de trabajo. Entre los factores más comunes se encuentran la importancia para la organización medida en cuanto a los ingresos o la marca, el coste total de la ejecución de la carga

de trabajo (incluidos los costes de funcionamiento y de recursos), la complejidad de la carga de trabajo, la facilidad para implementar un cambio, cualquier acuerdo de licencia de software y si un cambio supusiera un aumento significativo de los costes de licencia debido a las licencias punitivas. Los componentes pueden definirse funcional o técnicamente, como servidores web y bases de datos, o recursos de computación y almacenamiento. Equilibre los factores de la forma correspondiente y desarrolle un periodo para la carga de trabajo y sus componentes. Puede decidir revisar toda la carga de trabajo cada 18 meses, revisar los servidores web cada 6 meses, la base de datos cada 12 meses, la computación y el almacenamiento a corto plazo cada 6 meses y el almacenamiento a largo plazo cada 12 meses.

- Definir la exhaustividad de la revisión: defina cuánto esfuerzo se dedica a la revisión de la carga de trabajo o de los componentes de la carga de trabajo. Al igual que sucede con la frecuencia de revisión, se trata de equilibrar múltiples factores. Evalúe y priorice las oportunidades de mejora para centrar los esfuerzos donde aporten los mayores beneficios, al tiempo que realiza una estimación del esfuerzo necesario para estas actividades. Si los resultados previstos no alcanzan los objetivos y el esfuerzo necesario cuesta más, repita el proceso con acciones alternativas. Sus procesos de revisión deben incluir tiempo y recursos de sus procesos para hacer posibles las mejoras incrementales continuas. Por ejemplo, puede decidir dedicar una semana de análisis al componente de base de datos, una semana de análisis a los recursos de computación y cuatro horas a las revisiones de almacenamiento.

Recursos

Documentos relacionados:

- Blog de noticias de AWS
- Tipos de informática en la nube
- Novedades de AWS

Ejemplos relacionados:

- AWS Support Proactive Services (Servicios proactivos de asistencia de AWS)
- Regular workload reviews for SAP workloads (Revisiones periódicas de las cargas de trabajo de SAP)

COST10-BP02 Revisión y análisis de esta carga de trabajo con regularidad

Las cargas de trabajo existentes se revisan periódicamente en función de cada proceso definido para averiguar si se pueden adoptar nuevos servicios, reemplazar los existentes o rediseñar las cargas de trabajo.

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

Guía para la implementación

AWS añade constantemente nuevas características para que pueda experimentar e innovar más rápidamente con la tecnología más reciente. En Novedades de AWS se detalla cómo AWS está haciendo esto y ofrece información general rápida de los servicios, las características y los anuncios de ampliaciones regionales de AWS a medida que se publican. Puede profundizar en los lanzamientos que se han anunciado y utilizarlos para revisar y analizar sus cargas de trabajo existentes. Para obtener las ventajas de los nuevos servicios y características de AWS, debe revisar sus cargas de trabajo e implementar los nuevos servicios y características según sea necesario. Esto significa que es posible que tenga que reemplazar los servicios existentes que utiliza para la carga de trabajo o modernizarla para adoptar estos nuevos servicios de AWS. Por ejemplo, podría revisar sus cargas de trabajo y reemplazar el componente de mensajería por Amazon Simple Email Service. Esto elimina el coste de utilizar y mantener una flota de instancias, a la vez que proporciona toda la funcionalidad a un coste reducido.

Para analizar su carga de trabajo y destacar las posibles oportunidades, debe tener en cuenta no solo nuevos servicios, sino también nuevas formas de crear soluciones. Vea los vídeos de This is My Architecture en AWS para conocer los diseños de arquitectura de otros clientes, sus desafíos y sus soluciones. Consulte All-In series para conocer las aplicaciones reales de los servicios de AWS y las historias de clientes. También puede ver la serie de vídeos Back to Basics, donde se explican, examinan y desglosan las prácticas recomendadas básicas de los patrones de arquitectura en la nube. Otra fuente son los vídeos How to Build This, diseñados para ayudar a las personas con grandes ideas a dar vida a su producto mínimo viable (MVP, por sus siglas en inglés) mediante servicios de AWS. Es una forma de que los creadores de todo el mundo que tengan una idea sólida reciban orientación sobre arquitectura de arquitectos de soluciones de AWS experimentados. Por último, puede consultar los materiales de recursos de Introducción, que contiene tutoriales paso a paso.

Antes de ejecutar el proceso de revisión, cumpla los requisitos de su empresa en cuanto a carga de trabajo, seguridad y privacidad de los datos para poder utilizar un servicio específico o los requisitos de la región y rendimiento mientras sigue el proceso de revisión acordado.

Pasos para la implementación

- Revisar periódicamente la carga de trabajo: mediante su proceso definido, realice las revisiones con la frecuencia especificada. Compruebe que dedica el esfuerzo adecuado a cada componente. Este proceso sería similar al del diseño inicial, en el que seleccionó los servicios para la optimización de costes. Analice los servicios y las ventajas que aportarían. Esta vez, tenga en cuenta el coste de realizar el cambio, no solo las ventajas a largo plazo.
- Implementar los nuevos servicios: si la conclusión del análisis es implementar cambios, realice primero una base de referencia de la carga de trabajo para conocer el coste actual de cada resultado. Implemente los cambios y, a continuación, realice un análisis para confirmar el nuevo coste de cada resultado.

Recursos

Documentos relacionados:

- Blog de noticias de AWS
- Novedades de AWS
- Documentación de AWS
- Introducción a AWS
- Recursos generales de AWS

Vídeos relacionados:

- AWS - This is My Architecture
- AWS - Back to Basics
- AWS - All-In series
- How to Build This

COSTE 11. ¿Cómo evalúa el coste del esfuerzo?

Prácticas recomendadas

- COST11-BP01 Realizar automatizaciones de las operaciones

COST11-BP01 Realizar automatizaciones de las operaciones

Evalúe el coste del esfuerzo de las operaciones en la nube. Cuantifique la reducción de tiempo y esfuerzo en las tareas de administración, despliegue y otras operaciones mediante la automatización. Evalúe el tiempo y el coste necesarios para el esfuerzo de las operaciones y automatice las tareas administrativas para reducir el esfuerzo manual en la medida de lo posible.

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

La automatización de las operaciones mejora la coherencia y la escalabilidad, proporciona más visibilidad, fiabilidad y flexibilidad, reduce los costes y acelera la innovación al liberar recursos humanos y mejorar las métricas. Reduce la frecuencia de las tareas manuales, mejora la eficacia y beneficia a las empresas ya que ofrece una experiencia coherente y fiable al desplegar, administrar u operar cargas de trabajo. Puede liberar recursos de infraestructura de las tareas operativas manuales y utilizarlos para tareas de mayor valor e innovaciones, con lo que se mejorarán los resultados empresariales. Las empresas necesitan una forma probada y contrastada de administrar sus cargas de trabajo en la nube. Esa solución debe ser segura, rápida y rentable, con mínimo riesgo y máxima fiabilidad.

Comience por priorizar sus operaciones en función del esfuerzo necesario, mediante el análisis del coste global de las operaciones en la nube. Por ejemplo, ¿cuánto tiempo se tarda en desplegar nuevos recursos en la nube, realizar cambios de optimización en los existentes o implementar las configuraciones necesarias? Analice el coste total de las acciones humanas teniendo en cuenta el coste de las operaciones y de la administración. Dé prioridad a las automatizaciones de las tareas administrativas para reducir el esfuerzo manual. El esfuerzo de revisión debe reflejar la ventaja potencial. Por ejemplo, el tiempo dedicado a realizar tareas manualmente frente al tiempo dedicado a realizarlas de forma automática. Dé prioridad a la automatización de las actividades repetitivas y de alto valor. Las actividades que entrañan un mayor riesgo de errores humanos suelen ser el mejor lugar para empezar a automatizar, ya que el riesgo suele suponer un coste operativo adicional no deseado (por ejemplo, que el equipo de operaciones trabaje horas extra).

Mediante servicios de AWS, herramientas o productos de terceros, puede elegir qué automatizaciones de AWS se implementarán y personalizarán según sus requisitos específicos. En la tabla siguiente se muestran algunas de las funciones y capacidades de funcionamiento básicas que puede conseguir con los servicios de AWS para automatizar la administración y el funcionamiento:

- AWS Audit Manager: audite su uso de AWS de forma continua para simplificar la evaluación del riesgo y el cumplimiento.

- AWS Backup: administre y automatice la protección de datos de forma centralizada.
- AWS Config: configure los recursos de computación, valore, audite y evalúe las configuraciones y el inventario de recursos.
- AWS CloudFormation: lance recursos de alta disponibilidad con la infraestructura como código.
- AWS CloudTrail: administración de cambios de TI, cumplimiento y control.
- Amazon EventBridge: programe eventos y desencadene la acción de AWS Lambda.
- AWS Lambda: automatice los procesos repetitivos desencadenándolos con eventos o ejecutándolos según una programación fija con Amazon EventBridge.
- AWS Systems Manager: inicie y detenga las cargas de trabajo, aplique revisiones a los sistemas operativos, automatice la configuración y la administración continua.
- AWS Step Functions: programe trabajos y automatice flujos de trabajo.
- AWS Service Catalog: consumo de plantillas e infraestructura como código con cumplimiento y control.

Considere el ahorro de tiempo que permitirá a su equipo centrarse en la retirada de la deuda técnica, la innovación y las características de valor añadido. Por ejemplo, es posible que deba migrar su entorno local mediante lift-and-shift a la nube lo más rápido posible y optimizarlo más adelante. Merece la pena explorar el ahorro que podría conseguir mediante el uso de servicios completamente administrados mediante AWS que eliminen o reduzcan los costes de las licencias como Amazon Relational Database Service, Amazon EMR, Amazon WorkSpaces y Amazon SageMaker AI. Los servicios administrados eliminan la carga operativa y administrativa del mantenimiento de un servicio, lo que le permite centrarse en la innovación. Y dado que los servicios administrados operan a la escala de la nube, pueden ofrecer un costo menor por transacción o servicio.

Si desea adoptar automatizaciones de forma inmediata con el uso de productos y servicios de AWS y si no dispone de competencias en su organización, póngase en contacto con AWS Managed Services (AMS), AWS Professional Services o los socios de AWS para aumentar la adopción de la automatización y mejorar su excelencia operativa en la nube.

AWS Managed Services (AMS) es un servicio que utiliza la infraestructura de AWS en nombre de los clientes y socios de la empresa. Proporciona un entorno seguro y conforme en el que puede desplegar sus cargas de trabajo. AMS utiliza modelos operativos de nube empresarial con automatización para permitirle cumplir con los requisitos de su organización, pasar a la nube más rápidamente y reducir los costes de administración continua.

AWS Professional Services también puede ayudarle a conseguir los resultados empresariales deseados y a automatizar las operaciones con AWS. AWS Professional Services proporciona prácticas especializadas globales para ayudarle en sus esfuerzos en áreas específicas de la computación en la nube empresarial. Las prácticas especializadas ofrecen orientación específica a través de prácticas recomendadas, marcos, herramientas y servicios en áreas temáticas de soluciones, tecnología y sectores. Ayudan a los clientes a desplegar operaciones de TI automatizadas, sólidas y ágiles, así como capacidades de gobernanza optimizadas para el centro en la nube.

Pasos para la implementación

- Crear una vez y desplegar muchas veces: utilice infraestructura como código como AWS CloudFormation, AWS SDK o AWS Command Line Interface (AWS CLI) para desplegar una vez y utilizar muchas veces para el mismo entorno o para escenarios de recuperación de desastres. Etiquete mientras despliega para realizar un seguimiento de su consumo, tal y como se define en otras prácticas recomendadas. Utilice AWS Launch Wizard para reducir el tiempo de despliegue de muchas cargas de trabajo empresariales populares. AWS Launch Wizard le guía a través del dimensionamiento, la configuración y el despliegue de cargas de trabajo empresariales según las prácticas recomendadas de AWS. También puede utilizar el AWS Service Catalog, que le ayuda a crear y administrar plantillas aprobadas de infraestructura como código para su uso en AWS para que cualquiera pueda descubrir recursos en la nube aprobados y de autoservicio.
- Automatizar operaciones: ejecute operaciones rutinarias automáticamente sin intervención manual. Con los servicios y las herramientas de AWS, puede elegir qué automatizaciones de AWS implementar y personalizar según sus requisitos específicos. Por ejemplo, utilice EC2 Image Builder para crear, probar y desplegar imágenes de máquinas virtuales y de contenedores para su uso en AWS o en un entorno local. Si la acción que desea llevar a cabo no se puede realizar con los servicios de AWS o necesita acciones más complejas con recursos de filtrado, automatice sus operaciones con las herramientas de AWS CLI o AWS SDK. AWS CLI proporciona la posibilidad de automatizar todo el proceso de control y administración de servicios de AWS mediante scripts sin necesidad de utilizar la consola de AWS. Seleccione sus SDK de AWS preferidos para interactuar con los servicios de AWS. Para ver otros ejemplos de código, consulte el repositorio de ejemplos de código de AWS SDK.

Recursos

Documentos relacionados:

- Modernización de las operaciones en la Nube de AWS

- Servicios de AWS para la automatización
- Automatización de AWS Systems Manager
- Automatizaciones de AWS para la administración y las operaciones de SAP
- AWS Managed Services
- AWS Professional Services
- Automatización de la infraestructura

Ejemplos relacionados:

- Reinventing automated operations (Part I) (Reinvención de las operaciones automatizadas [parte I])
- Reinventing automated operations (Part II) (Reinvención de las operaciones automatizadas [parte II])
- Automatizaciones de AWS para la administración y las operaciones de SAP
- IT Automations with AWS Lambda (Automatizaciones de TI con AWS Lambda)
- Repositorio de ejemplos de código de AWS
- Muestras de AWS

Sostenibilidad

El pilar de sostenibilidad incluye comprender las repercusiones de los servicios que se usan, cuantificar el impacto durante todo el ciclo de vida de la carga de trabajo y aplicar tanto principios de diseño como prácticas recomendadas para reducir estas repercusiones al diseñar cargas de trabajo en la nube. Encontrará recomendaciones de implementación en el documento técnico Pilar de sostenibilidad.

Áreas de prácticas recomendadas

- Selección de regiones
- Alineación con la demanda
- Software y arquitectura
- Almacenamiento
- Hardware y servicios
- Proceso y cultura

Selección de regiones

Pregunta

- SUS 1 ¿Cómo selecciona las regiones para la carga de trabajo?

SUS 1 ¿Cómo selecciona las regiones para la carga de trabajo?

La elección de la región para su carga de trabajo afecta significativamente a sus KPI, incluidos el rendimiento, el coste y la huella de carbono. Para mejorar eficazmente estos KPI, debe elegir las regiones para sus cargas de trabajo basándose tanto en los requisitos empresariales como en los objetivos de sostenibilidad.

Prácticas recomendadas

- SUS01-BP01 Elegir la región basándose tanto en los requisitos empresariales como en los objetivos de sostenibilidad

SUS01-BP01 Elegir la región basándose tanto en los requisitos empresariales como en los objetivos de sostenibilidad

Elija una región para su carga de trabajo basándose tanto en los requisitos empresariales como en los objetivos de sostenibilidad para optimizar sus KPI, incluidos el rendimiento, el coste y la huella de carbono.

Patrones comunes de uso no recomendados:

- Selecciona la región de la carga de trabajo en función de la propia ubicación.
- Consolida todos los recursos de la carga de trabajo en una ubicación geográfica.

Beneficios de establecer esta práctica recomendada: la colocación de una carga de trabajo cerca de proyectos de energías renovables de Amazon o de regiones con baja intensidad de carbono publicada puede ayudar a reducir la huella de carbono de una carga de trabajo en la nube.

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

Guía para la implementación

La Nube de AWS es una red en constante expansión de regiones y puntos de presencia (POP), con una infraestructura de red global que los une. La elección de la región para su carga de trabajo

afecta significativamente a sus KPI, incluidos el rendimiento, el coste y la huella de carbono. Para mejorar eficazmente estos KPI, debe elegir las regiones para su carga de trabajo basándose tanto en los requisitos empresariales como en los objetivos de sostenibilidad.

Pasos para la aplicación

- Siga estos pasos para evaluar y preseleccionar las posibles regiones para la carga de trabajo en función de los requisitos empresariales, incluido el cumplimiento, las características disponibles, el coste y la latencia:
  - Confirme que estas regiones cumplen con la normativa local vigente.
  - Utilice las listas de servicios regionales de AWS para comprobar si las regiones disponen de los servicios y las características que necesita para ejecutar su carga de trabajo.
  - Calcule el coste de la carga de trabajo en cada región mediante Calculadora de precios de AWS.
  - Pruebe la latencia de la red entre las ubicaciones de sus usuarios finales y cada Región de AWS.
- Elija regiones cerca de proyectos de energías renovables de Amazon y regiones en las que la intensidad de carbono recogida en la cuadrícula sea más baja que en otras ubicaciones (o regiones).
  - Determine las directrices de sostenibilidad pertinentes para realizar un seguimiento y comparar las emisiones de carbono de un año a otro basándose en el protocolo de gases de efecto invernadero (métodos basados en el mercado y en la ubicación).
  - Elija la región en función del método que utilice para hacer un seguimiento de las emisiones de carbono. Para obtener más detalles sobre la elección de una región en función de sus directrices de sostenibilidad, consulte How to select a Region for your workload based on sustainability goals (Cómo seleccionar una región para su carga de trabajo en función de los objetivos de sostenibilidad).

Recursos

Documentos relacionados:

- Descripción de las estimaciones de emisiones de carbono
- Amazon en todo el mundo
- Metodología de energía renovable
- Qué tener en cuenta al seleccionar una región para las cargas de trabajo

Vídeos relacionados:

- Architecting sustainably and reducing your AWS carbon footprint (Arquitectura sostenible y reducción de la huella de carbono de AWS)

Alineación con la demanda

Pregunta

- SUS 2 ¿Cómo alinea los recursos en la nube a su demanda?

SUS 2 ¿Cómo alinea los recursos en la nube a su demanda?

La forma en que los usuarios y las aplicaciones consumen sus cargas de trabajo y otros recursos puede ayudarle a identificar las mejoras necesarias para alcanzar sus objetivos de sostenibilidad. Escale la infraestructura para adaptarla continuamente a la demanda y compruebe que solo utiliza los recursos mínimos necesarios para prestar asistencia a sus usuarios. Alinee los niveles de servicio con las necesidades de los clientes. Posicione los recursos de forma que se limite el uso de red necesario para que los usuarios puedan consumirlos. Elimine los activos sin usar. Proporcione a los miembros de su equipo dispositivos que satisfagan sus necesidades con un impacto mínimo en la sostenibilidad.

Prácticas recomendadas

- SUS02-BP01 Escalar la infraestructura de la carga de trabajo dinámicamente
- SUS02-BP02: Alineación de los SLA con los objetivos de sostenibilidad
- SUS02-BP03: Detener la creación y el mantenimiento de los recursos no utilizados
- SUS02-BP04 Optimizar la ubicación geográfica de las cargas de trabajo en función de sus requisitos de red
- SUS02-BP05: Optimización de los recursos de los miembros del equipo para las actividades realizadas
- SUS02-BP06 Implementar el almacenamiento en búfer o la limitación para aplanar la curva de demanda

SUS02-BP01 Escalar la infraestructura de la carga de trabajo dinámicamente

Utilice la elasticidad de la nube y escale su infraestructura de forma dinámica para adaptar la oferta de recursos en la nube a la demanda y evitar un exceso de capacidad en su carga de trabajo.

Patrones comunes de uso no recomendados:

- No se escala la infraestructura con la carga de usuarios.
- La infraestructura se escala manualmente todo el tiempo.
- Deja la capacidad aumentada después de un evento de ajuste de escala en lugar de volver a desescalar verticalmente.

Ventajas de establecer esta práctica recomendada: la configuración y las pruebas de la elasticidad de la carga de trabajo contribuyen a ajustar de forma eficaz la oferta de recursos en la nube a la demanda y a evitar el exceso de capacidad aprovisionada. Puede aprovechar la elasticidad de la nube para escalar automáticamente la capacidad durante y después de los picos de demanda para asegurarse de que solo utiliza el número correcto de recursos necesarios para satisfacer los requisitos empresariales.

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

Guía para la implementación

La nube ofrece la flexibilidad de ampliar o reducir sus recursos de forma dinámica a través de diversos mecanismos para satisfacer los cambios en la demanda. La correspondencia óptima entre la oferta y la demanda ofrece el menor impacto medioambiental para una carga de trabajo.

La demanda puede ser fija o variable, lo que requiere métricas y automatización para garantizar que la administración no resulte difícil. Las aplicaciones pueden escalarse o desescalarse verticalmente mediante la modificación del tamaño de la instancia, escalarse o desescalarse horizontalmente mediante la modificación del número de instancias, o una combinación de ambas.

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

- Enfoque de seguimiento de objetivos: supervise su métrica de escalado y aumente o reduzca de forma automática la capacidad a medida que la necesite.
- Escalamiento predictivo: desescale horizontalmente para prever las tendencias diarias y semanales.
- Enfoque basado en la programación: configure su propia programación según los cambios de carga predecibles.
- Escalamiento de servicios: elija servicios (como los que son sin servidor) que se escalan de forma nativa por diseño o proporcione el escalamiento automático como una característica.

Identifique los períodos de uso reducido o inexistente y escale los recursos en consonancia para eliminar el exceso de capacidad y mejorar la eficiencia.

Pasos para la aplicación

- La elasticidad hace coincidir la oferta de los recursos que tiene con la demanda de esos recursos. Las instancias, los contenedores y las funciones proporcionan mecanismos de elasticidad, ya sea en combinación con el escalamiento automático o como características del servicio. AWS proporciona una serie de mecanismos de escalamiento automático para garantizar que las cargas de trabajo puedan desescalarse verticalmente de forma rápida y sencilla durante los periodos con poca carga de usuarios. A continuación, se presentan algunos ejemplos de mecanismos de escalamiento automático:

Auto scaling mechanism	Where to use
Amazon EC2 Auto Scaling	Se usa para verificar que tiene el número correcto de instancias de Amazon EC2 disponibles para gestionar la carga de usuarios de su aplicación.
Application Auto Scaling	Se usa para escalar automáticamente los recursos de servicios de AWS individuales más allá de Amazon EC2, como funciones de Lambda o servicios de Amazon Elastic Container Service (Amazon ECS).
Cluster Autoscaler de Kubernetes	Se usa para escalar automáticamente clústeres de Kubernetes en AWS.

- A menudo se habla de escalamiento en relación con servicios de computación como instancias de Amazon EC2 o funciones AWS Lambda. Considere la configuración de servicios no computacionales como unidades de capacidad de lectura y escritura de Amazon DynamoDB o particiones de Amazon Kinesis Data Streams para ajustarse a la demanda.
- Verifique que las métricas para escalar o desescalar verticalmente se validan con respecto al tipo de carga de trabajo que se está desplegando. Si está desplegando una aplicación de transcodificación de vídeo, se espera una utilización del 100 % de la CPU y no debería ser su métrica principal. Puede usar una métrica personalizada (como la utilización de la memoria) para

su política de escalamiento si es necesario. Para elegir las métricas adecuadas, tenga en cuenta las siguientes directrices para Amazon EC2:

  - La métrica debe ser una métrica de utilización válida y describir el grado de ocupación de una instancia.
  - El valor de la métrica debe aumentar o disminuir proporcionalmente al número de instancias del grupo de Auto Scaling.
- Utilice el escalamiento dinámico en lugar del manual para su grupo de Auto Scaling. También le recomendamos que use las políticas de escalamiento de seguimiento de destino en el escalamiento dinámico.
- Verifique que los despliegues de la carga de trabajo puedan manejar los eventos de escalamiento y desescalamiento horizontales. Cree escenarios de prueba para los eventos de escalamiento con el fin de verificar que la carga de trabajo se comporta del modo previsto y no afecta a la experiencia del usuario (como la pérdida de sesiones persistentes). Puede utilizar el historial de actividades para verificar una actividad de escalamiento correspondiente a un grupo de Auto Scaling.
- Evalúe los patrones predecibles de su carga de trabajo y escale de forma proactiva al anticiparse a los cambios previstos y planeados en la demanda. Con el escalamiento predictivo, puede eliminar la necesidad de aprovisionar capacidad en exceso. Para obtener más detalles, consulte Escalamiento predictivo con Amazon EC2 Auto Scaling.

Recursos

Documentos relacionados:

- Getting Started with Amazon EC2 Auto Scaling (Introducción a Amazon EC2 Auto Scaling)
- Predictive Scaling for EC2, Powered by Machine Learning (Escalado predictivo para EC2, impulsado por el aprendizaje automático)
- Analyze user behavior using Amazon OpenSearch Service, Amazon Data Firehose and Kibana (Análisis del comportamiento del usuario con Amazon OpenSearch Service, Amazon Data Firehose y Kibana)
- ¿Qué es Amazon CloudWatch?
- Supervisión de la carga de bases de datos con Información sobre rendimiento en Amazon RDS
- Introducing Native Support for Predictive Scaling with Amazon EC2 Auto Scaling (Introducción a la compatibilidad nativa para escalado predictivo con Amazon EC2 Auto Scaling)

- Introducing Karpenter - An Open-Source, High-Performance Kubernetes Cluster Autoscaler (Presentación de Karpenter: Cluster Autoscaler de Kubernetes de código abierto y alto rendimiento)
- Deep Dive on Amazon ECS Cluster Auto Scaling (Profundización en Auto Scaling de clúster de Amazon ECS)

Vídeos relacionados:

- Build a cost-, energy-, and resource-efficient compute environment (Crear un entorno de computación rentable, eficiente en términos de costes, energía y recursos)
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2 (Computación mejor, más rápida y más barata: Optimización de costes de Amazon EC2) (CMP202-R1)

Ejemplos relacionados:

- Lab: Amazon EC2 Auto Scaling Group Examples (Laboratorio: ejemplos de grupos de Amazon EC2 Auto Scaling)
- Lab: Implement Autoscaling with Karpenter (Laboratorio: Implementar escalado automático con Karpenter)

SUS02-BP02: Alineación de los SLA con los objetivos de sostenibilidad

Revise y optimice los acuerdos de nivel de servicio (SLA) de la carga de trabajo en función de sus objetivos de sostenibilidad a fin de minimizar los recursos necesarios para admitir la carga de trabajo sin dejar de satisfacer las necesidades empresariales.

Antipatrones usuales:

- Los SLA de carga de trabajo se desconocen o son ambiguos.
- Define su SLA solo para la disponibilidad y el rendimiento.
- Utiliza el mismo patrón de diseño (como la arquitectura multi-AZ) para todas sus cargas de trabajo.

Beneficios de establecer esta práctica recomendada: la alineación de los SLA con los objetivos de sostenibilidad conlleva un uso óptimo de los recursos, al tiempo que se satisfacen las necesidades empresariales.

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

Guía para la implementación

Los SLA definen el nivel de servicio que se espera de una carga de trabajo en la nube, como el tiempo de respuesta, la disponibilidad y la retención de datos. Influyen en la arquitectura, el uso de recursos y el impacto medioambiental de una carga de trabajo en la nube. Con una cadencia regular, revise los SLA y realice concesiones para reducir significativamente el uso de recursos a cambio de disminuciones aceptables en los niveles de servicio.

Pasos para la implementación

- Defina o rediseñe SLA que respalden sus objetivos de sostenibilidad y que, a la vez, cumplan sus requisitos empresariales sin superarlos.
- Haga concesiones para disminuir significativamente las repercusiones en la sostenibilidad a cambio de reducciones aceptables en los niveles de servicio.
  - Sostenibilidad y fiabilidad: las cargas de trabajo de alta disponibilidad tienden a consumir más recursos.
  - Sostenibilidad y rendimiento: el uso de más recursos para aumentar el rendimiento podría tener mayor impacto medioambiental.
  - Sostenibilidad y seguridad: las cargas de trabajo demasiado seguras podrían tener mayor impacto medioambiental.
- Use patrones de diseño, como microservicios en AWS, que den prioridad a las funciones fundamentales y permitan unos niveles de servicio más bajos (como objetivos de tiempo de respuesta o de tiempo de recuperación) para las funciones que no sean esenciales.

Recursos

Documentos relacionados:

- Acuerdos de nivel de servicios (SLA) de AWS
- Importance of Service Level Agreement for SaaS Providers (Importancia de los acuerdos de nivel de servicio para los proveedores de SaaS)

Vídeos relacionados:

- Delivering sustainable, high-performing architectures (Entrega de arquitecturas sostenibles y de alto rendimiento)

- Build a cost-, energy-, and resource-efficient compute environment (Crear un entorno de computación rentable, eficiente en términos de costes, energía y recursos)

SUS02-BP03: Detener la creación y el mantenimiento de los recursos no utilizados

Retire los activos no utilizados de su carga de trabajo para reducir el número de recursos en la nube necesarios para atender su demanda y minimizar los residuos.

Patrones comunes de uso no recomendados:

- No analiza su aplicación en busca de activos redundantes o que ya no son necesarios.
- No elimina los activos que son redundantes o que ya no son necesarios.

Beneficios de establecer esta práctica recomendada: la eliminación de los activos no utilizados libera recursos y mejora la eficacia general de la carga de trabajo.

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

Guía para la implementación

Los activos no utilizados consumen recursos de la nube, como espacio de almacenamiento y potencia de computación. Con la identificación y eliminación de estos activos, podrá liberar estos recursos, lo que dará lugar a una arquitectura en la nube más eficiente. Realice análisis periódicos en los activos de aplicaciones (como los informes precompilados, los conjuntos de datos y las imágenes estáticas) y los patrones de acceso a los activos para identificar cualquier tipo de redundancia, infrautilización y los posibles objetivos de retirada. Elimine esos activos redundantes para reducir el despilfarro de recursos en su carga de trabajo.

Pasos para la implementación

- Utilice herramientas de supervisión para identificar los activos estáticos que ya no sean necesarios.
- Antes de eliminar un activo, evalúe el impacto de su eliminación en la arquitectura.
- Desarrolle un plan y elimine los activos que ya no sean necesarios.
- Consolide los recursos generados superpuestos para eliminar el procesamiento redundante.
- Actualice las aplicaciones para que dejen de producir y almacenar activos que no sean necesarios.
- Indique a terceros que administren en su nombre recursos que ya no son necesarios que dejen de producirlos y almacenarlos.

- Indique a terceros que consoliden los recursos redundantes producidos en su nombre.
- Revise periódicamente la carga de trabajo para identificar y eliminar los activos no utilizados.

Recursos

Documentos relacionados:

- Optimización de la infraestructura de AWS para la sostenibilidad, parte II: almacenamiento
- ¿Cómo puedo terminar los recursos activos que ya no necesito en mi Cuenta de AWS?

Vídeos relacionados:

- How do I check for and then remove active resources that I no longer need on my Cuenta de AWS? (¿Cómo puedo comprobar y, a continuación, eliminar los recursos activos que ya no necesito en mi Cuenta de AWS?)

SUS02-BP04 Optimizar la ubicación geográfica de las cargas de trabajo en función de sus requisitos de red

Seleccione para su carga de trabajo una ubicación y unos servicios en la nube que acorten la distancia que debe recorrer el tráfico de red y reduzcan el total de recursos de red necesarios para admitir su carga de trabajo.

Patrones comunes de uso no recomendados:

- Se selecciona la región de la carga de trabajo en función de la propia ubicación.
- Consolida todos los recursos de la carga de trabajo en una ubicación geográfica.
- Todo el tráfico fluye a través de sus centros de datos existentes.

Beneficios de establecer esta práctica recomendada: Colocar una carga de trabajo cerca de sus usuarios permite obtener la menor latencia, al tiempo que disminuye el movimiento de datos a través de la red y reduce el impacto medioambiental.

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

Guía para la implementación

La infraestructura de Nube de AWS se crea en torno a opciones de ubicación como regiones, zonas de disponibilidad, grupos de ubicaciones y ubicaciones periféricas como AWS Outposts y Zonas locales de AWS. Estas opciones de ubicación son las responsables de mantener la conectividad entre los componentes de las aplicaciones, los servicios en la nube, las redes periféricas y los centros de datos locales.

Analice los patrones de acceso a la red en su carga de trabajo para identificar cómo utilizar estas opciones de ubicación en la nube y reducir la distancia que debe recorrer el tráfico de red.

Pasos para la implementación

- Analice los patrones de acceso de la red en su carga de trabajo para identificar cómo utilizan los usuarios su aplicación.
  - Use herramientas de monitorización como Amazon CloudWatch y AWS CloudTrailpara recopilar datos sobre las actividades de la red.
  - Analice los datos para identificar el patrón de acceso de la red.
- Seleccione las regiones para el despliegue de la carga de trabajo en función de los siguientes elementos clave:
  - Su objetivo de sostenibilidad: como se explica en la Selección de regiones.
  - Dónde se encuentran sus 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.
  - Dónde se encuentran sus 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 restricciones como la seguridad y el cumplimiento como se explica en Qué tener en cuenta al seleccionar una región para las cargas de trabajo.
- Utilice almacenamiento en caché local o Soluciones de almacenamiento en caché de AWS para los recursos de uso frecuente con el fin de mejorar el rendimiento, reducir el movimiento de datos y disminuir el impacto medioambiental.

Servicio	Cuándo usar
Amazon CloudFront	Se usa para almacenar en caché el contenido estático como imágenes, scripts y vídeos, así

Servicio	Cuándo usar
	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 añadir aceleración en memoria a sus tablas de DynamoDB.

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

Servicio	Cuándo 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.
Funciones de Amazon CloudFront	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.

- Use la agrupación de conexiones para permitir reutilizar las conexiones y reducir la cantidad de recursos necesarios.
- Use los almacenes de datos distribuidos que no se basen en conexiones persistentes y en actualizaciones sincrónicas por coherencia para atender a las poblaciones regionales.
- Reemplace la capacidad de red estática preaprovisionada por capacidad dinámica compartida y comparta el impacto en la sostenibilidad de la capacidad de red con otros suscriptores.

Recursos

Documentos relacionados:

- Optimización de la infraestructura de AWS para la sostenibilidad, parte III: redes
- Documentación de Amazon ElastiCache
- ¿Qué es Amazon CloudFront?
- Características clave de Amazon CloudFront

Vídeos relacionados:

- Demystifying data transfer on AWS (Desmitificación de la transferencia de datos en AWS)
- Scaling network performance on next-gen Amazon EC2 instances (Escalar el rendimiento de la red en instancias de EC2 de nueva generación)

Ejemplos relacionados:

- AWS Networking Workshops (Talleres de red de AWS)
- Architecting for sustainability - Minimize data movement across networks (Diseño de una arquitectura para la sostenibilidad: minimice el movimiento de datos entre las redes)

SUS02-BP05: Optimización de los recursos de los miembros del equipo para las actividades realizadas

Optimice los recursos proporcionados a los miembros del equipo para minimizar el impacto en la sostenibilidad medioambiental a la vez que se cubren sus necesidades.

Patrones comunes de uso no recomendados:

- Ignora el impacto de los dispositivos utilizados por los miembros de su equipo en la eficacia global de su aplicación en la nube.
- Administra y actualiza manualmente los recursos que utilizan los miembros del equipo.

Beneficios de establecer esta práctica recomendada: la optimización de los recursos de los miembros del equipo mejora la eficacia general de las aplicaciones basadas en la nube.

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

Guía para la implementación

Analice los dispositivos que usan los miembros de su equipo para consumir sus servicios, el ciclo de vida que se espera que tengan y el impacto económico y en la sostenibilidad. Implemente estrategias para optimizar estos recursos. Por ejemplo, realice las operaciones complejas (como la representación y la compilación) en escritorios en una infraestructura escalable con un uso intensivo, en lugar de hacerlo en sistemas de usuarios únicos de gran potencia infrautilizados.

Pasos para la implementación

- Aprovisione las estaciones de trabajo y otros dispositivos para alinearlos con la forma en que se usan.
- Use escritorios virtuales y streaming de aplicaciones para limitar los requisitos de dispositivos y actualizaciones.
- Traslade a la nube las tareas con un uso intensivo del procesador o la memoria para usar su elasticidad.
- Evalúe el impacto de los procesos y los sistemas en el ciclo de vida de los dispositivos y seleccione aquellas soluciones que minimizan los requisitos para el reemplazo de dispositivos a la vez que satisfacen los requisitos empresariales.
- Implemente la administración remota de los dispositivos para reducir la necesidad de realizar viajes de negocios.
  - Administrador de flotas de AWS Systems Manager es una experiencia de interfaz de usuario (IU) unificada que le ayuda a administrar de forma remota sus nodos que se ejecutan en AWS o en un entorno local.

Recursos

Documentos relacionados:

- ¿Qué es Amazon WorkSpaces?
- Optimizador de costes para Amazon WorkSpaces
- Documentación de Amazon AppStream 2.0
- NICE DCV

Vídeos relacionados:

- Managing cost for Amazon WorkSpaces on AWS (Administración de costes para Amazon WorkSpaces en AWS)

SUS02-BP06 Implementar el almacenamiento en búfer o la limitación para aplanar la curva de demanda

El almacenamiento en búfer y la limitación aplanan la curva de demanda y reducen la capacidad aprovisionada necesaria para su carga de trabajo.

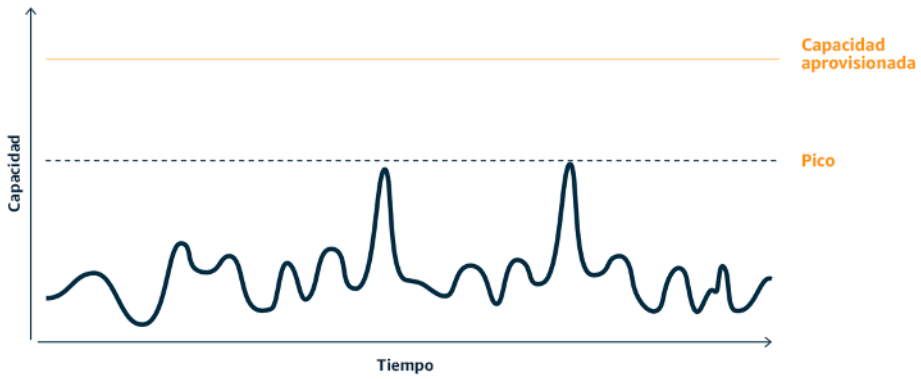
Patrones comunes de uso no recomendados:

- Procesa las solicitudes de los clientes inmediatamente mientras no es necesario.
- No analiza los requisitos de las solicitudes de los clientes.

Beneficios de establecer esta práctica recomendada: el aplanamiento de la curva de demanda reduce la capacidad aprovisionada necesaria para la carga de trabajo. La reducción de la capacidad aprovisionada implica un menor consumo de energía y un menor impacto medioambiental.

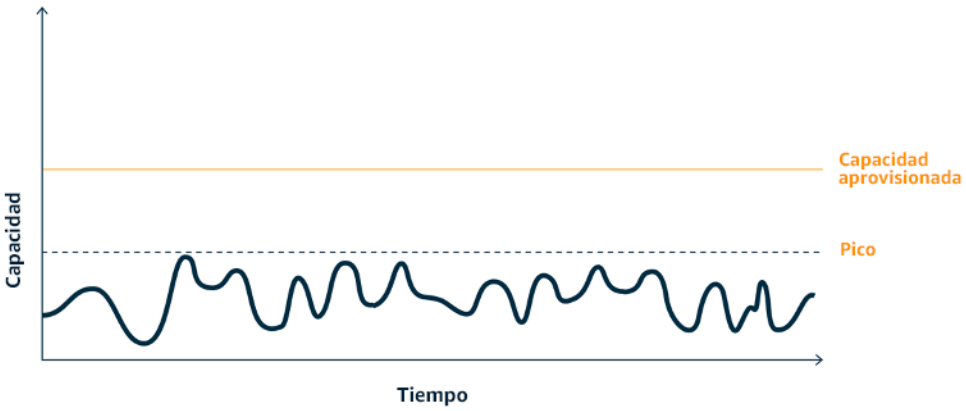
Nivel de riesgo expuesto si no se establece esta práctica recomendada: bajo

El aplanamiento de la curva de demanda de la carga de trabajo puede ayudarle a reducir la capacidad aprovisionada para una carga de trabajo y a reducir su impacto medioambiental. Supongamos una carga de trabajo con la curva de demanda que se muestra en la siguiente figura. Esta carga de trabajo tiene dos picos y, para gestionarlos, se aprovisiona la capacidad de recursos que muestra la línea naranja. Los recursos y la energía utilizados para esta carga de trabajo no están indicados por el área situada debajo de la curva de demanda, sino por el área situada debajo de la línea de capacidad aprovisionada, ya que esta capacidad se necesita para gestionar esos dos picos.



Curva de demanda con dos picos diferenciados que requieren una alta capacidad aprovisionada.

Puede utilizar el almacenamiento en búfer o la limitación para modificar la curva de demanda y suavizar los picos, lo que significa menos capacidad aprovisionada y menos energía consumida. Implemente limitaciones cuando sus clientes puedan realizar reintentos. Implemente el almacenamiento en búfer para almacenar la solicitud y aplazar el procesamiento para más adelante.



Efecto de la limitación sobre la curva de demanda y la capacidad aprovisionada.

Pasos para la implementación

- Analice las solicitudes de los clientes para determinar cómo responder a ellas. Entre las preguntas a tener en cuenta se incluyen las siguientes:
  - ¿Esta solicitud puede procesarse de forma asíncrona?
  - ¿El cliente tiene capacidad de reintentos?
- Si el cliente tiene capacidad de reintentos, puede implementar la limitación, que le indica al origen que si no puede atender la solicitud en el momento actual debe intentarlo más tarde.
  - Puede usar Amazon API Gateway para implementar la limitación.
- En el caso de los clientes que no pueden realizar reintentos, es necesario implementar un búfer para aplanar la curva de demanda. Un búfer aplaza el procesamiento de las solicitudes, por lo que permite a las aplicaciones que se ejecutan a diferentes ritmos comunicarse de forma efectiva. El enfoque basado en búfer utiliza una cola o una secuencia para aceptar mensajes de los productores. De este modo, los consumidores pueden leer y procesar los mensajes, lo que permite que dichos mensajes se ejecuten a la velocidad que cumpla con los requisitos empresariales de los consumidores.
  - Amazon Simple Queue Service (Amazon SQS) es un servicio administrado que proporciona colas que permiten que un solo consumidor lea mensajes individuales.

  - Amazon Kinesis ofrece una secuencia que permite que muchos consumidores lean los mismos mensajes.
- Analice la demanda general, la tasa de cambio y el tiempo de respuesta requerido para dimensionar correctamente la limitación o el búfer requeridos.

Recursos

Documentos relacionados:

- Getting started with Amazon SQS (Introducción a Amazon SQS)
- Application integration Using Queues and Messages (Integración de aplicaciones mediante colas y mensajes)

Vídeos relacionados:

- Choosing the Right Messaging Service for Your Distributed App (Elección del servicio de mensajería correcto para su aplicación distribuida)

Software y arquitectura

Pregunta

- SUS 3 ¿Cómo puede sacar partido de los patrones de software y de arquitectura para respaldar sus objetivos de sostenibilidad?

SUS 3 ¿Cómo puede sacar partido de los patrones de software y de arquitectura para respaldar sus objetivos de sostenibilidad?

Implemente patrones que permitan suavizar la carga y mantener un uso elevado consistente de los recursos implementados para minimizar los recursos consumidos. Puede haber componentes que queden inactivos debido a la falta de uso relacionada con los cambios en el comportamiento de los usuarios a lo largo del tiempo. Revise los patrones y la arquitectura para consolidar los componentes infrautilizados a fin de incrementar el uso general. Retire los componentes que ya no son necesarios. Analice el rendimiento de los componentes de su carga de trabajo y optimice aquellos que consumen la mayor cantidad de recursos. Tenga en cuenta los dispositivos que usan los clientes para acceder a sus servicios e implemente patrones para minimizar la necesidad de realizar actualizaciones de los dispositivos.

Prácticas recomendadas

- SUS03-BP01: Optimizar el software y la arquitectura para los trabajos asíncronos y programados
- SUS03-BP02 Eliminar o refactorizar los componentes de cargas de trabajo con uso reducido o nulo
- SUS03-BP03: Optimización de las áreas de código que consumen la mayor parte del tiempo o de los recursos
- SUS03-BP04 Optimizar el impacto en los dispositivos y equipos
- SUS03-BP05: Uso de los patrones de software y las arquitecturas que mejor respaldan los patrones de almacenamiento y el acceso a los datos

SUS03-BP01: Optimizar el software y la arquitectura para los trabajos asíncronos y programados

Utilice patrones de software y arquitectura eficientes, como los basados en colas, para mantener una utilización elevada y coherente de los recursos desplegados.

Patrones comunes de uso no recomendados:

- Realiza un aprovisionamiento excesivo de los recursos de su carga de trabajo en la nube para hacer frente a picos imprevistos de la demanda.
- Su arquitectura no desacopla los emisores y los receptores de mensajes asíncronos mediante un componente de mensajería.

Beneficios de establecer esta práctica recomendada:

- Los patrones de software y arquitectura eficientes minimizan los recursos no utilizados en la carga de trabajo y mejoran la eficiencia global.
- Puede escalar el procesamiento independientemente de la recepción de mensajes asíncronos.
- Mediante un componente de mensajería, tendrá unos requisitos de disponibilidad más relajados que podrá cumplir con menos recursos.

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

Guía para la implementación

Utilice patrones de arquitectura eficientes, como la arquitectura basada en eventos, que dan como resultado una utilización uniforme de los componentes y minimizan el aprovisionamiento excesivo en

la carga de trabajo. El uso de patrones de arquitectura eficientes minimiza los recursos inactivos por falta de uso debido a cambios en la demanda a lo largo del tiempo.

Comprenda los requisitos de los componentes de la carga de trabajo y adopte patrones de arquitectura que aumenten la utilización global de los recursos. Retire los componentes que ya no son necesarios.

Pasos para la aplicación

- Analice la demanda de su carga de trabajo para determinar cómo responder a ella.
- En el caso de solicitudes o trabajos que no requieran respuestas síncronas, utilice arquitecturas basadas en colas y empleados de escalamiento automático para maximizar la utilización. A continuación, encontrará algunos ejemplos de cuándo podría plantearse una arquitectura basada en colas:

Queuing mechanism	Description
Colas de trabajo de AWS Batch	Los trabajos de AWS Batch se envían a una cola de trabajos en la que permanecen hasta que pueden programarse para ejecutarse en un entorno de computación.
Instancias de spot de Amazon Simple Queue Service y Amazon EC2	Emparejamiento de Amazon SQS e instancias de spot para crear una arquitectura eficiente y tolerante a errores.

- En el caso de solicitudes o trabajos que puedan procesarse en cualquier momento, utilice mecanismos de programación para procesar los trabajos por lotes y obtener una mayor eficacia. A continuación, se presentan algunos ejemplos de mecanismos de programación en AWS:

Scheduling mechanism	Description
Programador de Amazon EventBridge	Una capacidad de Amazon EventBridge que le permite crear, ejecutar y administrar tareas programadas a escala.

Scheduling mechanism	Description
Programación basada en tiempo de AWS Glue	Defina una programación basada en el tiempo para sus rastreadores y trabajos en AWS Glue.
Tareas programadas de Amazon Elastic Container Service (Amazon ECS)	Amazon ECS admite la creación de tareas programadas. Las tareas programadas utilizan reglas de Amazon EventBridge para ejecutar tareas según una programación o en respuesta a un evento de EventBridge.
Instance Scheduler	Configure las programaciones de inicio y detención de sus instancias de Amazon EC2 y Amazon Relational Database Service.

- Si utiliza mecanismos de sondeo y webhooks en su arquitectura, reemplácelos por eventos. Utilice arquitecturas basadas en eventos para crear cargas de trabajo de elevada eficacia.
- Aproveche la tecnología sin servidor en AWS para eliminar la infraestructura aprovisionada en exceso.
- Dimensione correctamente los componentes individuales de su arquitectura para evitar recursos inactivos mientras se espera la entrada.

Recursos

Documentos relacionados:

- ¿Qué es Amazon Simple Queue Service?
- ¿Qué es Amazon MQ?
- Escalamiento basado en Amazon SQS
- ¿Qué es AWS Step Functions?
- ¿Qué es AWS Lambda?
- Uso de AWS Lambda con Amazon SQS
- ¿Qué es Amazon EventBridge?

Vídeos relacionados:

- Moving to event-driven architectures (Migración a arquitecturas basadas en eventos)

SUS03-BP02 Eliminar o refactorizar los componentes de cargas de trabajo con uso reducido o nulo

Elimine los componentes que ya no se usan ni se necesitan y refactorice aquellos con un uso reducido para minimizar el desperdicio en su carga de trabajo.

Patrones comunes de uso no recomendados:

- No comprueba periódicamente el nivel de utilización de los componentes individuales de la carga de trabajo.
- No comprueba ni analiza recomendaciones de herramientas de dimensionamiento de AWS como AWS Compute Optimizer.

Beneficios de establecer esta práctica recomendada: la eliminación de los activos no utilizados libera recursos y mejora la eficacia general de la carga de trabajo.

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

Guía para la implementación

Revise la carga de trabajo para identificar los componentes inactivos o no utilizados. Es un proceso de mejora iterativo que puede desencadenarse por cambios en la demanda o por el lanzamiento de un nuevo servicio en la nube. Por ejemplo, un descenso significativo del tiempo de ejecución de una función de AWS Lambda puede ser un indicador de la necesidad de reducir el tamaño de la memoria. Además, a medida que AWS lanza nuevos servicios y características, los servicios y la arquitectura óptimos para su carga de trabajo también pueden cambiar.

Supervise continuamente la actividad de la carga de trabajo y busque oportunidades para mejorar el nivel de uso de los componentes individuales. Con la eliminación de los componentes ociosos y la realización de actividades de redimensionamiento, cumplirá los requisitos de su empresa con el menor número de recursos en la nube.

Pasos para la implementación

- Supervise y capture las métricas de utilización de los componentes críticos de su carga de trabajo (como la utilización de la CPU, la utilización de la memoria o el rendimiento de la red en métricas de Amazon CloudWatch).

- Para cargas de trabajo estables, compruebe las herramientas de redimensionamiento de AWS como AWS Compute Optimizer a intervalos regulares para identificar los componentes ociosos, no utilizados o infrautilizados.
- En el caso de las cargas de trabajo efímeras, evalúe las métricas de utilización para identificar los componentes inactivos, no utilizados o infrautilizados.
- Retire los componentes y activos asociados (como las imágenes de Amazon ECR) que ya no sean necesarios.
- Refactorice o consolide los componentes infrautilizados con otros recursos para mejorar la eficiencia de uso. Por ejemplo, puede aprovisionar varias bases de datos pequeñas en una sola instancia de base de datos de Amazon RDS en vez de ejecutar bases de datos en instancias individuales infrautilizadas.
- Entienda los recursos aprovisionados por su carga de trabajo para completar una unidad de trabajo.

Recursos

Documentos relacionados:

- AWS Trusted Advisor
- ¿Qué es Amazon CloudWatch?
- Limpieza automatizada de imágenes no utilizadas en Amazon ECR

Ejemplos relacionados:

- Well-Architected Lab - Rightsizing with AWS Compute Optimizer(Laboratorio de Well-Architected: redimensionamiento con AWS Compute Optimizer)
- Well-Architected Lab - Optimize Hardware Patterns and Observe Sustainability KPIs (Laboratorio de Well-Architected: optimizar los patrones de hardware y observar los KPI de sostenibilidad)

SUS03-BP03: Optimización de las áreas de código que consumen la mayor parte del tiempo o de los recursos

Optimice el código que se ejecuta en los distintos componentes de su arquitectura para minimizar el uso de los recursos y, a la vez, maximizar el rendimiento.

Patrones comunes de uso no recomendados:

- Ignora la optimización del código para el uso de recursos.
- Normalmente responde a los problemas de rendimiento con un aumento de los recursos.
- Su proceso de revisión y desarrollo del código no realiza un seguimiento de los cambios de rendimiento.

Beneficios de establecer esta práctica recomendada: El uso de código eficiente minimiza el uso de recursos y mejora el rendimiento.

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

Guía para la implementación

Es fundamental examinar cada área funcional, incluido el código de una aplicación con arquitectura de nube, para optimizar su uso de recursos y su rendimiento. Supervise continuamente el rendimiento de la carga de trabajo en los entornos de creación y producción e identifique oportunidades para mejorar los fragmentos de código que tienen un uso de recursos especialmente elevado. Adopte un proceso de revisión periódico para identificar errores o antipatrones en su código que utilicen los recursos de forma ineficiente. Use algoritmos sencillos y eficaces que produzcan los mismos resultados para su caso de uso.

Pasos para la implementación

- Durante el desarrollo de sus cargas de trabajo, adopte un proceso automatizado de revisión del código para mejorar la calidad e identificar errores y antipatrones.
  - Automate code reviews with Amazon CodeGuru Reviewer (Revisiones automáticas de código con Amazon CodeGuru Reviewer)
  - Detecting concurrency bugs with Amazon CodeGuru (Detección de errores de simultaneidad con Amazon CodeGuru)
  - Raising code quality for Python applications using Amazon CodeGuru (Mejora de la calidad del código para aplicaciones Python con Amazon CodeGuru)
- A medida que ejecute las cargas de trabajo, supervise los recursos para identificar los componentes con elevados requisitos de recursos por unidad de trabajo como objetivos de las revisiones de código.
- Para las revisiones de código, use un generador de perfiles de código para identificar las áreas de código que emplean más tiempo o recursos como objetivo de la optimización.
  - Reducing your organization's carbon footprint with Amazon CodeGuru Profiler (Reducción de la huella de carbono de su organización con Amazon CodeGuru Profiler)

  - Understanding memory usage in your Java application with Amazon CodeGuru Profiler (Descripción del uso de memoria en su aplicación Java con Amazon CodeGuru Profiler)
  - Improving customer experience and reducing cost with Amazon CodeGuru Profiler (Mejora de la experiencia del cliente y reducción de costes con Amazon CodeGuru Profiler)
- Use el sistema operativo y el lenguaje de programación más eficaces para la carga de trabajo. Para obtener más información sobre los lenguajes de programación energéticamente eficientes (incluido Rust), consulte Sustainability with Rust (Sostenibilidad con Rust).
- Reemplace los algoritmos que hacen un uso intensivo de la computación por versiones más sencillas y eficientes que produzcan el mismo resultado.
- Elimine el código innecesario, como la ordenación y el formato.

Recursos

Documentos relacionados:

- ¿Qué es Amazon CodeGuru Profiler?
- Instancias de FPGA
- SDK de AWS en Herramientas para crear en AWS

Vídeos relacionados:

- Improve Code Efficiency Using Amazon CodeGuru Profiler (Mejora de la eficiencia del código con Amazon CodeGuru Profiler)
- Automate Code Reviews and Application Performance Recommendations with Amazon CodeGuru (Automatización de las revisiones de código y las recomendaciones de rendimiento de aplicaciones con Amazon CodeGuru)

SUS03-BP04 Optimizar el impacto en los dispositivos y equipos

Analice los dispositivos y los equipos empleados en la arquitectura y utilice estrategias para reducir su uso. Esto puede minimizar el impacto medioambiental global de su carga de trabajo en la nube.

Patrones comunes de uso no recomendados:

- Ignora el impacto medioambiental de los dispositivos que utilizan sus clientes.
- Administra y actualiza manualmente los recursos que utilizan los clientes.

Beneficios de establecer esta práctica recomendada: la implementación de patrones y características de software optimizados para el dispositivo del cliente puede reducir el impacto medioambiental general de la carga de trabajo en la nube.

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

Guía para la implementación

La implementación de patrones y características de software optimizados para el dispositivo del cliente puede reducir el impacto medioambiental general de la carga de trabajo en la nube.

- La implementación de nuevas características compatibles con versiones anteriores puede reducir el número de reemplazos de hardware.
- La optimización de una aplicación para que funcione de forma eficiente en los dispositivos puede contribuir a reducir su consumo de energía y a prolongar la duración de su batería (si funcionan con ella).
- La optimización de una aplicación para dispositivos también puede reducir la transferencia de datos a través de la red.

Comprenda los dispositivos y equipos utilizados en su arquitectura, su ciclo de vida previsto y el impacto de reemplazar esos componentes. Implemente patrones y características de software que puedan minimizar el consumo de energía del dispositivo, la necesidad de los clientes de reemplazarlo y también de actualizarlo manualmente.

Pasos para la implementación

- Realice un inventario de los dispositivos utilizados en su arquitectura. Los dispositivos pueden ser móviles, tabletas, dispositivos IoT, luces inteligentes o incluso dispositivos inteligentes en una fábrica.
- Optimice la aplicación que se ejecuta en los dispositivos:
  - Utilice estrategias como la ejecución de tareas en segundo plano para reducir su consumo de energía.
  - Tenga en cuenta la latencia y el ancho de banda de la red al crear cargas e implemente capacidades que ayuden al funcionamiento óptimo de las aplicaciones en enlaces de alta latencia y ancho de banda bajo.
  - Convierta las cargas útiles y los archivos a los formatos optimizados que requieren los dispositivos. Por ejemplo, puede utilizar Amazon Elastic Transcoder o AWS Elemental

MediaConvert para convertir archivos multimedia digitales de gran tamaño y alta calidad a formatos que los usuarios puedan reproducir en dispositivos móviles, tabletas, navegadores web y televisores conectados.

  - Realice las actividades con un uso intensivo de los recursos informáticos (como la representación de imágenes) en el lado del servidor o use el streaming de aplicaciones para mejorar la experiencia del usuario en los dispositivos más antiguos.
  - Segmente y pagine los resultados, sobre todo en las sesiones interactivas, para administrar las cargas y limitar los requisitos de almacenamiento local.
- Utilice el mecanismo automatizado vía inalámbrica (OTA) para desplegar actualizaciones en uno o varios dispositivos.
  - Puede utilizar una canalización de CI/CD para actualizar las aplicaciones móviles.
  - Puede utilizar AWS IoT Device Management para administrar a distancia los dispositivos conectados a escala.
- Para probar nuevas características y actualizaciones, utilice granjas de dispositivos administrados con conjuntos representativos de hardware e itere el desarrollo para maximizar los dispositivos admitidos. Para obtener más información, consulte SUS06-BP04 Usar granjas de dispositivos administrados para pruebas.

Recursos

Documentos relacionados:

- ¿Qué es AWS Device Farm?
- Documentación de Amazon AppStream 2.0
- NICE DCV
- Tutorial de OTA para actualizar el firmware en dispositivos que ejecutan FreeRTOS

Vídeos relacionados:

- Introduction to AWS Device Farm(Introducción a AWS Device Farm)

SUS03-BP05: Uso de los patrones de software y las arquitecturas que mejor respaldan los patrones de almacenamiento y el acceso a los datos

Analice cómo se usan los datos en la carga de trabajo, cómo los consumen los usuarios, cómo se transfieren y cómo se almacenan. Utilice patrones y arquitecturas de software que admitan mejor el acceso a los datos y el almacenamiento para minimizar los recursos de computación, redes y almacenamiento necesarios para admitir la carga de trabajo.

Patrones comunes de uso no recomendados:

- Supone que todas las cargas de trabajo tienen patrones similares de almacenamiento y acceso a los datos.
- Solo utiliza un nivel de almacenamiento, asumiendo que todas las cargas de trabajo encajan en ese nivel.
- Supone que los patrones de acceso a los datos se mantendrán coherentes a lo largo del tiempo.
- Su arquitectura admite una posible ampliación de acceso a los datos, lo que provoca que los recursos permanezcan inactivos la mayor parte del tiempo.

Beneficios de establecer esta práctica recomendada: la selección y la optimización de su arquitectura en función de los patrones de acceso y almacenamiento de datos le ayudará a disminuir la complejidad del desarrollo y a aumentar la utilización general. Saber cuándo utilizar las tablas globales, las particiones de datos y el almacenamiento en caché le ayudará a disminuir la sobrecarga operativa y a escalar en función de sus necesidades de carga de trabajo.

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

Guía para la implementación

Utilice los patrones de software y arquitectura que mejor se adapten a las características de sus datos y a sus patrones de acceso. Por ejemplo, utilice una arquitectura de datos moderna en AWS que le permita usar servicios personalizados optimizados para sus casos de uso de análisis exclusivos. Estos patrones de arquitectura permiten un procesamiento de datos eficaz y reducen el uso de recursos.

Pasos para la implementación

- Analice las características de los datos y los patrones de acceso para identificar la configuración correcta de sus recursos en la nube. Entre las características clave que se deben tener en cuenta se incluyen:

  - Tipo de datos: estructurados, semiestructurados, no estructurados
  - Crecimiento de datos: limitados, no limitados
  - Durabilidad de datos: persistentes, efímeros, transitorios
  - Patrones de acceso: lecturas o escrituras, frecuencia de actualización, con picos o constantes
- Utilice los patrones de arquitectura que mejor admitan los patrones de acceso y almacenamiento de datos.
  - Let's Architect! Modern data architectures (Let's Architect! Arquitecturas de datos modernas)
  - Databases on AWS: The Right Tool for the Right Job (Bases de datos de AWS: la herramienta adecuada para el trabajo adecuado.)
- Use tecnologías que funcionen de forma nativa con datos comprimidos.
- Utilice servicios de análisis personalizados para el procesamiento de datos en su arquitectura.
- Use el motor de base de datos que mejor admita su patrón de consulta dominante. Administre sus índices de base de datos para garantizar una ejecución eficaz de las consultas. Para más detalles, consulte Bases de datos de AWS.
- Seleccione protocolos de red que reduzcan la cantidad de capacidad de red consumida en su arquitectura.

Recursos

Documentos relacionados:

- Formatos de archivo de compatibilidad con la compresión de Athena
- Uso de COPY con formatos de datos de columnas con Amazon Redshift
- Conversión del formato de registro de entrada en Firehose
- Opciones de formato para las entradas y salidas de ETL en AWS Glue
- Mejora del rendimiento de las consultas en Amazon Athena con la conversión a formato de columnas
- Carga de archivos de datos comprimidos desde Amazon S3 con Amazon Redshift
- Supervisión de la carga de bases de datos con Información sobre rendimiento en Amazon Aurora
- Supervisión de la carga de bases de datos con Información sobre rendimiento en Amazon RDS
- Clase de almacenamiento de Amazon S3 Intelligent-Tiering

Vídeos relacionados:

- Building modern data architectures on AWS(Creación de arquitecturas de datos modernas en AWS)

Almacenamiento

Pregunta

- SUS 4 ¿Cómo puede aprovechar los patrones y las políticas de administración de datos para admitir sus objetivos de sostenibilidad?

SUS 4 ¿Cómo puede aprovechar los patrones y las políticas de administración de datos para admitir sus objetivos de sostenibilidad?

Implemente prácticas de administración de datos para reducir el almacenamiento aprovisionado que se necesita para admitir la carga de trabajo y los recursos necesarios para su uso. Comprenda sus datos y use las configuraciones y tecnologías de almacenamiento que respalden con más eficacia al valor empresarial de los datos y la forma en que se usan. Haga que el ciclo de vida de los datos incluya un almacenamiento más eficaz y de menor rendimiento cuando disminuyan los requisitos y elimine los datos que ya no se requieran.

Prácticas recomendadas

- SUS04-BP01: Implementación de una política de clasificación de datos
- SUS04-BP02 Usar tecnologías que admiten patrones de almacenamiento y acceso a los datos
- SUS04-BP03 Usar políticas para administrar el ciclo de vida de los conjuntos de datos
- SUS04-BP04 Utilice la elasticidad y la automatización para ampliar el almacenamiento de bloques o el sistema de archivos
- SUS04-BP05: Eliminación de datos innecesarios o redundantes
- SUS04-BP06 Usar sistemas de archivos o almacenamiento compartidos para acceder a datos comunes
- SUS04-BP07: Minimización del movimiento de datos entre redes
- SUS04-BP08: Realización de copias de seguridad de los datos solo cuando sea difícil volver a crearlos

SUS04-BP01: Implementación de una política de clasificación de datos

Clasifique los datos para comprender su criticidad para los resultados empresariales y elija el nivel de almacenamiento de bajo consumo adecuado para almacenar los datos.

Antipatrones usuales:

- No identifica activos de datos con características similares (como sensibilidad, criticidad empresarial o requisitos normativos) que se estén procesando o almacenando.
- No ha implementado un catálogo de datos para inventariar sus activos de datos.

Beneficios de establecer esta práctica recomendada: la implementación de una política de clasificación de datos le permite determinar el nivel de almacenamiento de bajo consumo para los datos.

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

Guía para la implementación

La clasificación de datos implica la identificación de los tipos de datos que se están procesando y almacenando en un sistema de información propiedad de una organización o controlado por ella. También implica tomar una determinación sobre la criticidad de los datos y la posible repercusión de su divulgación, pérdida o uso indebido.

Implemente la política de clasificación de datos mediante un trabajo en sentido inverso a partir del uso contextual de los datos y la creación de un esquema de categorización que tenga en cuenta el nivel de criticidad de un conjunto de datos determinado para las operaciones de una organización.

Pasos para la implementación

- Realice un inventario de los distintos tipos de datos que existen para su carga de trabajo.
  - Para obtener más detalles sobre las categorías de clasificación de datos, consulte el documento técnico Data Classification (Clasificación de datos).
- Determine la criticidad, la confidencialidad, la integridad y la disponibilidad de los datos en función del riesgo para la organización. Utilice estos requisitos para agrupar los datos en uno de los niveles de clasificación de datos que adopte.
  - Como ejemplo, consulte Four simple steps to classify your data and secure your startup (Cuatro sencillos pasos para clasificar sus datos y proteger su startup).

- Audite su entorno de forma periódica para detectar los datos no etiquetados y sin clasificar; a continuación, clasifique y etiquete los datos adecuadamente.
  - Como ejemplo, consulte Catálogo de datos y rastreadores en AWS Glue.
- Establezca un catálogo de datos que proporcione capacidades de auditoría y gobernanza.
- Determine y documente los procedimientos de tratamiento de cada clase de datos.
- Utilice la automatización para auditar continuamente su entorno con el fin de identificar los datos sin etiquetar y sin clasificar, y clasifíquelos y etiquételos adecuadamente.

Recursos

Documentos relacionados:

- Uso de Nube de AWS para respaldar la clasificación de datos
- Políticas de etiquetado deAWS Organizations

Vídeos relacionados:

- Enabling agility with data governance on AWS(Facilitar la agilidad con la gobernanza de los datos en AWS)

SUS04-BP02 Usar tecnologías que admiten patrones de almacenamiento y acceso a los datos

Use las tecnologías de almacenamiento que mejor respalden la forma en que accede y guarda sus datos a fin de minimizar los recursos aprovisionados para admitir la carga de trabajo.

Patrones comunes de uso no recomendados:

- Supone que todas las cargas de trabajo tienen patrones similares de almacenamiento y acceso a los datos.
- Solo utiliza un nivel de almacenamiento, asumiendo que todas las cargas de trabajo encajan en ese nivel.
- Supone que los patrones de acceso a los datos se mantendrán coherentes a lo largo del tiempo.

Beneficios de establecer esta práctica recomendada: seleccionar y optimizar sus tecnologías de almacenamiento en función de los patrones de acceso y almacenamiento de datos le ayudará

a reducir los recursos necesarios en la nube para satisfacer sus necesidades empresariales y a mejorar la eficacia general de la carga de trabajo en la nube.

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

Guía para la implementación

Seleccione la solución de almacenamiento que mejor se adapte a sus patrones de acceso, o bien considere cambiar sus patrones de acceso de modo que se adapten a la solución de almacenamiento, a fin de maximizar la eficiencia del rendimiento.

- Evalúe las características de sus datos y su patrón de acceso para recopilar las características clave de sus necesidades de almacenamiento. Entre las características clave que se deben tener en cuenta se incluyen:
  - Tipo de datos: estructurados, semiestructurados y no estructurados
  - Crecimiento de los datos: delimitado, no delimitado
  - Durabilidad de los datos: persistentes, efímeros, transitorios
  - Patrones de acceso: lecturas o escrituras, frecuencia, con picos o constantes
- Migre los datos a la tecnología de almacenamiento adecuada que sea compatible con las características de sus datos y su patrón de acceso. A continuación, le presentamos algunos ejemplos de tecnologías de almacenamiento de AWS y sus principales características:

Tipo	Tecnología	Características clave
Clases de almacenamiento	Amazon S3	Un servicio de almacenamiento de objetos con escalabilidad ilimitada, alta disponibilidad y varias opciones de accesibilidad. La transferencia y el acceso a los objetos dentro y fuera de Amazon S3 puede utilizar un servicio como Aceleración de transferencia o bien Puntos de accesopara respaldar su ubicación, necesidades

Tipo	Tecnología	Características clave
		de seguridad y patrones de acceso.
Almacenamiento de archivos	Amazon Glacier	Clase de almacenamiento de Amazon S3 desarrollada para el archivado de datos.
Sistema de archivos compartidos	Amazon Elastic File System (Amazon EFS)	Sistema de archivos montable al que pueden acceder varios tipos de soluciones de computaci ón. Amazon EFS aumenta y reduce automáticamente el almacenamiento y su rendimiento se ha optimizado para ofrecer latencias bajas y constantes.
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. En Amazon FSx, su latencia, rendimien to y E/S por segundo varían según el sistema de archivos y deben tenerse en cuenta a la hora de seleccionar el sistema de archivos adecuado para sus necesidad es de carga de trabajo.

Tipo	Tecnología	Características clave
Almacenamiento de bloques	Amazon Elastic Block Store (Amazon EBS)	Servicio de almacenamiento de bloques escalable y de alto rendimiento diseñado para Amazon Elastic Compute Cloud (Amazon EC2). Amazon EBS incluye almacenamiento respaldado por SSD para cargas de trabajo transaccionales y de IOPS intensivas, y almacenamiento respaldado por HDD para cargas de trabajo de rendimiento intensivo.
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 fuerte coherencia de datos. Muchas aplicaciones tradicionales, la planificación de recursos empresariales (ERP), la administración de las relaciones con los clientes (CRM) y los sistemas de comercio electrónico utilizan bases de datos relacionales para almacenar sus datos.

Tipo	Tecnología	Características clave
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.

- Para los sistemas de almacenamiento que tienen un tamaño fijo, como Amazon EBS o Amazon FSx, supervise el espacio de almacenamiento disponible y automatice la asignación de almacenamiento al alcanzar un umbral. Puede usar Amazon CloudWatch para recopilar y analizar diferentes métricas para Amazon EBS y Amazon FSx.
- Las clases de almacenamiento de Amazon S3 pueden configurarse en el nivel de objeto y un único bucket puede contener objetos almacenados en todas las clases de almacenamiento.
- También puede utilizar las políticas de ciclo de vida de Amazon S3 para realizar transiciones automáticas de objetos entre clases de almacenamiento o eliminar datos sin necesidad de realizar cambios en la aplicación. En general, tiene que equilibrar la eficiencia de los recursos, la latencia de acceso y la fiabilidad cuando considere estos mecanismos de almacenamiento.

Recursos

Documentos relacionados:

- Tipos de volumen de Amazon EBS
- Almacén de instancias de Amazon EC2
- Amazon S3 Intelligent-Tiering
- Características de E/S de Amazon EBS
- Uso de clases de almacenamiento de Amazon S3
- ¿Qué es Amazon Glacier?

Vídeos relacionados:

- Patrones de arquitectura para lagos de datos en AWS
- Deep dive on Amazon EBS (STG303-R1) (Conocer en profundidad Amazon EBS [STG303-R1])
- Optimize your storage performance with Amazon S3 (STG343) (Optimizar el rendimiento del almacenamiento con Amazon S3 [STG343])
- Building modern data architectures on AWS (Creación de arquitecturas de datos modernas en AWS)

Ejemplos relacionados:

- Amazon EFS CSI Driver (Controlador CSI de Amazon EFS)
- Amazon EBS CSI Driver (Controlador CSI de Amazon EBS)
- Amazon EFS Utilities (Utilidades de Amazon EFS)
- Amazon EBS Autoscale (Escala automática de Amazon EBS)
- Amazon S3 Examples (Ejemplos de Amazon S3)

SUS04-BP03 Usar políticas para administrar el ciclo de vida de los conjuntos de datos

Administre el ciclo de vida de todos sus datos y aplique automáticamente la eliminación para minimizar el almacenamiento total necesario para su carga de trabajo.

Patrones comunes de uso no recomendados:

- Elimina los datos manualmente.
- No elimina ningún dato de su carga de trabajo.
- No traslada los datos a niveles de almacenamiento de mayor eficiencia energética en función de sus requisitos de retención y acceso.

Beneficios de establecer esta práctica recomendada: el uso de políticas de ciclo de vida de los datos garantiza un acceso y una conservación eficaces de los datos en una carga de trabajo.

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

Guía para la implementación

Los conjuntos de datos suelen tener distintos requisitos de conservación y acceso durante su ciclo de vida. Por ejemplo, su aplicación puede necesitar acceso frecuente a algunos conjuntos de datos durante un periodo de tiempo limitado. Después, se accede a esos conjuntos de datos con poca frecuencia.

Para administrar eficazmente los conjuntos de datos a lo largo de su ciclo de vida, configure las políticas de ciclo de vida, que son reglas que definen cómo administrar los conjuntos de datos.

Con las reglas de configuración del ciclo de vida, puede indicar al servicio de almacenamiento específico que realice la transición de un conjunto de datos a niveles de almacenamiento de mayor eficiencia energética, que lo archive o que lo elimine.

Pasos para la aplicación

- Clasifique los conjuntos de datos de la carga de trabajo.
- Defina procedimientos de gestión para cada clase de datos.
- Establezca políticas de ciclo de vida automatizadas para la aplicación de reglas de ciclo de vida. A continuación, se ofrecen algunos ejemplos de cómo configurar políticas automatizadas de ciclo de vida para distintos servicios de almacenamiento de AWS:

Storage service	How to set automated lifecycle policies
Amazon S3	Puede usar Amazon S3 Lifecycle para administrar sus 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 coste. Puede aprovechar las métricas de Amazon S3 Storage Lens para identificar las oportunidades de optimización y las lagunas en la administración del ciclo de vida.

Storage service	How to set automated lifecycle policies
Amazon Elastic Block Store	Puede utilizar Amazon Data Lifecycle Manager para automatizar la creación, retención y eliminación de instantáneas de Amazon EBS y AMI respaldadas por Amazon EBS.
Amazon Elastic File System	La administración del ciclo de vida de Amazon EFS se ocupa automáticamente el almacenamiento de archivos para sus sistemas de archivos.
Amazon Elastic Container Registry	Las políticas de ciclo de vida de Amazon ECR automatizan la limpieza de las imágenes de contenedor al hacer caducar las imágenes por antigüedad o cantidad.
AWS Elemental MediaStore	Puede utilizar una política de ciclo de vida de objetos que controle durante cuánto tiempo deben almacenarse los objetos en el contenedor de MediaStore.

- Elimine los volúmenes, las instantáneas y los datos no utilizados que estén fuera de su periodo de retención. Aproveche las características nativas del servicio, como el tiempo de vida de Amazon DynamoDB o la retención de registros de Amazon CloudWatch para su eliminación.
- Agregue y comprima datos cuando proceda en función de las reglas de ciclo de vida.

Recursos

Documentos relacionados:

- Optimice sus reglas de Amazon S3 Lifecycle con el análisis de clases de almacenamiento de Amazon S3
- Evaluación de recursos con Reglas de AWS Config

Vídeos relacionados:

- Simplify Your Data Lifecycle and Optimize Storage Costs With Amazon S3 Lifecycle (Simplifique su ciclo de vida de datos y optimice los costes de almacenamiento con Amazon S3 Lifecycle)
- Reduce Your Storage Costs Using Amazon S3 Storage Lens (Reduzca sus costes de almacenamiento con Amazon S3 Storage Lens)

SUS04-BP04 Utilice la elasticidad y la automatización para ampliar el almacenamiento de bloques o el sistema de archivos

Utilice la elasticidad y la automatización para ampliar el almacenamiento de bloques o el sistema de archivos a medida que crecen los datos para minimizar el almacenamiento total aprovisionado.

Patrones comunes de uso no recomendados:

- Adquiere un almacenamiento de bloques grande o un sistema de archivos de gran tamaño para necesidades futuras.
- Aprovisiona en exceso las operaciones de entrada y salida por segundo (IOPS) de su sistema de archivos.
- No supervisa el uso de sus volúmenes de datos.

Beneficios de establecer esta práctica recomendada: minimizar el aprovisionamiento excesivo del sistema de almacenamiento reduce los recursos inactivos y mejora la eficacia general de su carga de trabajo.

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

Guía para la implementación

Cree almacenamiento de bloques y sistemas de archivos con una asignación de tamaño, rendimiento y latencia adecuados para su carga de trabajo. Utilice la elasticidad y la automatización para ampliar el almacenamiento de bloques o el sistema de archivos a medida que crecen los datos sin tener que aprovisionar en exceso estos servicios de almacenamiento.

Pasos para la implementación

- En el caso del almacenamiento de tamaño fijo, como Amazon EBS, asegúrese de supervisar la cantidad de almacenamiento utilizada en relación con el tamaño total del almacenamiento y cree una automatización, si es posible, para aumentar el tamaño del almacenamiento cuando se alcance un umbral.

- Use volúmenes elásticos y servicios administrados de datos en bloque para automatizar la asignación de almacenamiento adicional a medida que aumentan sus datos persistentes. Por ejemplo, puede utilizar Volúmenes elásticos de Amazon EBS para cambiar el tamaño de volumen, el tipo de volumen o ajustar el rendimiento de sus volúmenes de Amazon EBS.
- Elija la clase de almacenamiento, el modo de rendimiento y el modo de caudal adecuados para que su sistema de archivos responda a su necesidad empresarial, sin excederse.
  - Rendimiento de Amazon EFS
  - Rendimiento de los volúmenes de Amazon EBS en instancias de Linux
- Establezca niveles como objetivo de uso para los volúmenes de datos y ajuste el tamaño de los volúmenes que estén fuera de los intervalos esperados.
- Establezca el tamaño correcto de los volúmenes de solo lectura según los datos.
- Migre los datos a almacenes de objetos para evitar el aprovisionamiento del exceso de capacidad de los tamaños de volúmenes fijos en el almacenamiento en bloque.
- Revise periódicamente los volúmenes elásticos y los sistemas de archivos para terminar los volúmenes inactivos y reducir los recursos aprovisionados en exceso para ajustarlos al tamaño actual de los datos.

Recursos

Documentos relacionados:

- Documentación de Amazon FSx
- What is Amazon Elastic File System? (¿Qué es Amazon Elastic File System?)

Vídeos relacionados:

- Deep Dive on Amazon EBS Elastic Volumes (Profundización en los volúmenes elásticos de Amazon EBS)
- Amazon EBS and Snapshot Optimization Strategies for Better Performance and Cost Savings (Estrategias de optimización de Amazon EBS e instantáneas para mejorar el rendimiento y ahorrar costes)
- Optimizing Amazon EFS for cost and performance, using best practices (Optimización de Amazon EFS para costes y rendimiento, mediante prácticas recomendadas)

SUS04-BP05: Eliminación de datos innecesarios o redundantes

Elimine datos innecesarios o redundantes para minimizar los recursos de almacenamiento necesarios para guardar sus conjuntos de datos.

Patrones comunes de uso no recomendados:

- Duplica datos que se pueden obtener o recrear fácilmente.
- Realiza copia de seguridad de todos los datos sin tener en cuenta su criticidad.
- Solo elimina datos de forma irregular, en eventos operativos o no los elimina en absoluto.
- Almacena datos de forma redundante independientemente de la durabilidad del servicio de almacenamiento.
- Activa el control de versiones de Amazon S3 sin ninguna justificación empresarial.

Beneficios de establecer esta práctica recomendada: la eliminación de datos redundantes reduce el tamaño de almacenamiento necesario de la carga de trabajo y su impacto medioambiental.

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

Guía para la implementación

No almacene datos que no necesite. Automatice la eliminación de datos innecesarios. Use tecnologías que desdupliquen los datos en el nivel de archivo y de bloque. Aproveche las características de replicación y redundancia de datos nativos de los servicios.

Pasos para la aplicación

- Evalúe si puede evitar almacenar datos mediante los conjuntos de datos existentes de disponibilidad pública en AWS Data Exchange y Open Data on AWS (Datos abiertos en AWS).
- Use mecanismos que puedan desduplicar los datos en el nivel de bloque y de objeto. A continuación, se ofrecen algunos ejemplos de cómo desduplicar datos en AWS:

Storage service	Deduplication mechanism
Amazon S3	Use AWS Lake Formation FindMatches para encontrar registros coincidentes en un conjunto de datos (incluidos los que no tienen

Storage service	Deduplication mechanism
	identificadores) con la nueva transformación de ML FindMatches.
Amazon FSx	Active la desduplicación de datos en Amazon FSx para Windows.
Instantáneas de Amazon Elastic Block Store	Las instantáneas son copias de seguridad progresivas, lo que significa que solo se guardan los bloques del dispositivo que han cambiado después de la instantánea más reciente.

- Analice el acceso de datos para identificar los datos innecesarios. Automatice las políticas de ciclo de vida. Aproveche las características nativas del servicio, como el tiempo de vida de Amazon DynamoDB, Amazon S3 Lifecycle o la retención de registros de Amazon CloudWatch para su eliminación.
- Utilice las capacidades de virtualización de datos en AWS para mantener los datos en su origen y evitar la duplicación de datos.
  - Cloud Native Data Virtualization on AWS (Virtualización de datos nativos en la nube en AWS)
  - Lab: Optimize Data Pattern Using Amazon Redshift Data Sharing (Laboratorio: optimizar el patrón de datos mediante el uso compartido de datos de Amazon Redshift)
- Use una tecnología de copia de seguridad que pueda crear copias incrementales.
- Aproveche la durabilidad de Amazon S3 y la replicación de Amazon EBS para conseguir sus objetivos de durabilidad en lugar de tecnologías autoadministradas (como una matriz redundante de discos independientes [RAID]).
- Centralice los datos de registro y de seguimiento, desduplique las entradas de registro que sean idénticas y establezca mecanismos para ajustar los detalles cuando sea necesario.
- Rellene las memorias caché previamente solo cuando se justifique.
- Establezca la supervisión y automatización de la memoria caché para ajustar el tamaño de esta en consonancia.
- Quite los despliegues y los recursos desfasados de los almacenes de objetos y las memorias caché periféricas al introducir nuevas versiones de su carga de trabajo.

Recursos

Documentos relacionados:

- Cambio de la retención de datos de registro en CloudWatch Logs
- Desduplicación de datos en Amazon FSx para Windows File Server
- Características de Amazon FSx para ONTAP, incluida la desduplicación de datos
- Invalidación de archivos en Amazon CloudFront
- Uso de AWS Backup para hacer copias de seguridad y restaurar sistemas de archivos de Amazon EFS
- ¿Qué es Amazon CloudWatch Logs?
- Trabajar con copias de seguridad en Amazon RDS

Vídeos relacionados:

- Fuzzy Matching and Deduplicating Data with ML Transforms for AWS Lake Formation (Concordancia difusa y desduplicación de datos con transformaciones de ML para AWS Lake Formation)

Ejemplos relacionados:

- ¿Cómo analizo mis registros de acceso al servidor de Amazon S3 mediante Amazon Athena?

SUS04-BP06 Usar sistemas de archivos o almacenamiento compartidos para acceder a datos comunes

Adopte sistemas de archivos o almacenamiento compartidos para evitar la duplicación de datos y posibilitar una infraestructura más eficiente para su carga de trabajo.

Patrones comunes de uso no recomendados:

- Aprovisiona almacenamiento para cada cliente.
- No desconecta el volumen de datos de los clientes inactivos.
- No proporciona acceso al almacenamiento a través de plataformas y sistemas.

Beneficios de establecer esta práctica recomendada: el uso de sistemas de archivos o almacenamiento compartidos permite compartir los datos con uno o varios consumidores sin tener que copiarlos. De este modo, se reducen los recursos de almacenamiento necesarios para la carga de trabajo.

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

Guía para la implementación

Si tiene varios usuarios o aplicaciones que acceden a los mismos conjuntos de datos, el uso de la tecnología de almacenamiento compartido es esencial para habilitar una infraestructura eficiente para su carga de trabajo. La tecnología de almacenamiento compartido proporciona una ubicación central para almacenar y administrar conjuntos de datos y evitar la duplicación de datos. También refuerza la coherencia de los datos entre los distintos sistemas. Además, la tecnología de almacenamiento compartido permite un uso más eficaz de la potencia de computación, ya que varios recursos de computación pueden acceder a los datos y procesarlos simultáneamente en paralelo.

Obtenga datos de estos servicios de almacenamiento compartido solo cuando los necesite y desconecte los volúmenes que no utilice para liberar recursos.

Pasos para la implementación

- Migre los datos al almacenamiento compartido cuando tengan varios consumidores. A continuación le mostramos algunos ejemplos de tecnología de almacenamiento compartido en AWS:

Storage option	When to use
Amazon EBS Multi-Attach	Amazon EBS Multi-Attach le permite adjuntar un único volumen SSD IOPS aprovisio nadas (io1 o io2) a varias instancias que se encuentren en la misma zona de disponibi lidad.
Amazon EFS	Consulte When to Choose Amazon EFS (Cuándo elegir Amazon EFS).

Storage option	When to use
Amazon FSx	Consulte Choosing an Amazon FSx File System (Elección de un sistema de archivos de Amazon FSx).
Amazon S3	Las aplicaciones que no requieren una estructura de sistema de archivos y están diseñadas para colaborar con el almacenam iento de objetos pueden utilizar Amazon S3 como una solución de almacenamiento de objetos escalable de forma masiva, duradera y de bajo coste.

- Copie datos en sistemas de archivos compartidos, o recupérelos de ellos, solo cuando sea necesario. Por ejemplo, puede crear un sistema de archivos de Amazon FSx for Lustre respaldado por Amazon S3 y cargar solo el subconjunto de datos necesarios para procesar los trabajos en Amazon FSx.
- Elimine los datos según corresponda a sus patrones de uso, como se indica en SUS04-BP03 Usar políticas para administrar el ciclo de vida de los conjuntos de datos.
- Desconecte los volúmenes de los clientes que no los estén usando de forma activa.

Recursos

Documentos relacionados:

- Linking your file system to an Amazon S3 bucket (Vinculación del sistema de archivos a un bucket de Amazon S3)
- Using Amazon EFS for AWS Lambda in your serverless applications (Uso de Amazon EFS para AWS Lambda en las aplicaciones sin servidor)
- Amazon EFS Intelligent-Tiering Optimizes Costs for Workloads with Changing Access Patterns (Amazon EFS Intelligent-Tiering optimiza los costes para las cargas de trabajo con patrones de acceso cambiantes)
- Using Amazon FSx with your on-premises data repository (Uso de Amazon FSx con su repositorio de datos local)

Vídeos relacionados:

- Storage cost optimization with Amazon EFS (Optimización de costes de almacenamiento con Amazon EFS)

SUS04-BP07: Minimización del movimiento de datos entre redes

Utilice sistemas de archivos o almacenamiento de objetos compartidos para acceder a los datos comunes y minimizar el total de recursos de redes necesarios para admitir el movimiento de datos para su carga de trabajo.

Patrones comunes de uso no recomendados:

- Almacena todos los datos en la misma Región de AWS independientemente de dónde se encuentren los usuarios de los datos.
- No optimiza el tamaño ni el formato de los datos antes de moverlos por la red.

Beneficios de establecer esta práctica recomendada: la optimización del movimiento de datos por la red reduce los recursos de redes totales necesarios para la carga de trabajo y disminuye su impacto medioambiental.

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

Guía para la implementación

El movimiento de datos por la organización requiere recursos de computación, red y almacenamiento. Utilice técnicas para minimizar el movimiento de datos y mejorar la eficacia general de su carga de trabajo.

Pasos para la implementación

- Considere la proximidad a los datos o a los usuarios como un factor de decisión cuando seleccione una región para su carga de trabajo.
- Particione los servicios que se consumen regionalmente para que los datos específicos de una región se almacenen en la región en la que se consumen.
- Utilice formatos de archivo eficientes (como Parquet u ORC) y comprima los datos antes de moverlos por la red.
- No mueva los datos no utilizados. Algunos ejemplos que pueden ayudarle a evitar mover datos no utilizados:

  - Reduzca las respuestas de la API solo a los datos relevantes.
  - Agregue los datos cuando estén detallados (no se requiere información en el nivel de registro).
  - Consulte Well-Architected Lab: Optimize Data Pattern Using Amazon Redshift Data Sharing (Laboratorio de Well-Architected: optimizar el patrón de datos mediante el uso compartido de datos de Amazon Redshift)..
  - Considere el uso compartido de datos entre cuentas en AWS Lake Formation..
- Utilice servicios que puedan ayudarle a ejecutar el código más cerca de los usuarios de su carga de trabajo.

Servicio	Cuándo usar
Lambda@Edge	Se usa para las operaciones que utilizan muchos recursos de computación que se ejecutan cuando los objetos no están en la memoria caché.
Funciones de CloudFront	Se usan en 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	Ejecuta la computación local, la mensajería y el almacenamiento en caché de datos para los dispositivos conectados.

Recursos

Documentos relacionados:

- Optimización de la infraestructura de AWS para la sostenibilidad, parte III: redes
- Infraestructura global de AWS
- Características clave de Amazon CloudFront, incluida la red perimetral global de CloudFront
- Compresión de solicitudes HTTP en Amazon OpenSearch Service
- Compresión de datos intermedia con Amazon EMR
- Carga de archivos de datos comprimidos desde Amazon S3 en Amazon Redshift

- Entrega archivos comprimidos con Amazon CloudFront

Vídeos relacionados:

- Demystifying data transfer on AWS (Desmitificación de la transferencia de datos en AWS)

Ejemplos relacionados:

- Architecting for sustainability - Minimize data movement across networks (Diseño de una arquitectura para la sostenibilidad: minimice el movimiento de datos entre las redes)

SUS04-BP08: Realización de copias de seguridad de los datos solo cuando sea difícil volver a crearlos

Evite realizar copias de seguridad de datos que no tengan valor empresarial para minimizar los requisitos de recursos de almacenamiento para su carga de trabajo.

Patrones comunes de uso no recomendados:

- No dispone de una estrategia de copia de seguridad para los datos.
- Hace copias de seguridad de datos que pueden volver a crearse fácilmente.

Beneficios de establecer esta práctica recomendada: evitar las copias de seguridad de los datos que no son fundamentales reduce los recursos de almacenamiento necesarios para la carga de trabajo y disminuye su impacto medioambiental.

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

Guía para la implementación

Evitar la copia de seguridad de datos innecesarios puede contribuir a reducir los costes y los recursos de almacenamiento utilizados por la carga de trabajo. Realice copias de seguridad únicamente de aquellos datos que tengan valor empresarial o que sean necesarios para satisfacer los requisitos de cumplimiento. Examine las políticas de copia de seguridad y excluya el almacenamiento efímero que no proporcione valor alguno en un escenario de recuperación.

Pasos para la implementación

- Implemente la política de clasificación de datos tal y como se indica en SUS04-BP01: Implementación de una política de clasificación de datos.
- Utilice la criticidad de la clasificación de sus datos y diseñe la estrategia de copia de seguridad en función de su objetivo de tiempo de recuperación (RTO) y de su objetivo de punto de recuperación (RPO). Evite realizar copias de seguridad de datos no esenciales.
  - Excluya los datos que pueden volver a crearse fácilmente.
  - Excluya los datos efímeros de sus copias de seguridad.
  - Excluya las copias locales de los datos, a menos que el tiempo necesario para restaurar esos datos desde una ubicación común supere lo establecido en los acuerdos de nivel de servicio (SLA).
- Utilice una solución automatizada o un servicio administrado para realizar copias de seguridad de los datos fundamentales para la empresa.
  - AWS Backup es un servicio totalmente administrado que facilita la centralización y automatización de la protección de datos entre servicios de AWS, en la nube y en el entorno local. Para obtener orientación práctica sobre cómo crear copias de seguridad automatizadas con AWS Backup, consulte Well-Architected Labs - Testing Backup and Restore of Data (Laboratorios de Well-Architected: Pruebas de copia de seguridad y restauración de datos).
  - Automate backups and optimize backup costs for Amazon EFS using AWS Backup (Automatizar las copias de seguridad y optimizar los costes de copia de seguridad para Amazon EFS con AWS Backup).

Recursos

Prácticas recomendadas relacionadas:

- REL09-BP01 Identificar todos los datos de los que se debe hacer una copia de seguridad y crearla o reproducir los datos a partir de los orígenes
- REL09-BP03 Realizar copias de seguridad de los datos automáticamente
- REL13-BP02 Usar estrategias de recuperación definidas para cumplir los objetivos de recuperación

Documentos relacionados:

- Uso de AWS Backup para hacer copias de seguridad y restaurar sistemas de archivos de Amazon EFS

- Instantáneas de Amazon EBS
- Trabajar con copias de seguridad en Amazon Relational Database Service
- Socio de APN: socios que pueden ayudar con la copia de seguridad
- AWS Marketplace: productos que pueden usarse para la copia de seguridad
- Copia de seguridad de Amazon EFS
- Copia de seguridad de Amazon FSx para Windows File Server
- Copia de seguridad y restauración para Amazon ElastiCache (Redis OSS)

Vídeos relacionados:

- AWS re:Invent 2021 - Backup, disaster recovery, and ransomware protection with AWS (AWS re:Invent 2021: copia de seguridad, recuperación de desastres y protección contra ransomware con AWS)
- AWS Backup Demo: Cross-Account and Cross-Region Backup (Demostración de AWS Backup: copia de seguridad entre cuentas y entre regiones)
- AWS re:Invent 2019: Deep dive on AWS Backup, ft. Rackspace (STG341)

Ejemplos relacionados:

- Well-Architected Lab - Testing Backup and Restore of Data (Laboratorio de Well-Architected: probar la copia de seguridad y restauración de los datos)
- Well-Architected Lab - Backup and Restore with Failback for Analytics Workload (Laboratorio de Well-Architected: copia de seguridad y restauración con conmutación por recuperación para cargas de trabajo de análisis)
- Well-Architected Lab - Disaster Recovery - Backup and Restore (Laboratorio de Well-Architected: recuperación de desastres, copia de seguridad y restauración)

Hardware y servicios

Pregunta

- SUS 5 ¿Cómo selecciona y usa el hardware y los servicios en la nube de su arquitectura para lograr sus objetivos de sostenibilidad?

SUS 5 ¿Cómo selecciona y usa el hardware y los servicios en la nube de su arquitectura para lograr sus objetivos de sostenibilidad?

Realice cambios en sus prácticas de administración de hardware como forma de reducir el impacto en la sostenibilidad de las cargas de trabajo. Minimice la cantidad de hardware necesario para aprovisionar e implementar y seleccione el hardware y los servicios más eficaces para su carga de trabajo individual.

Prácticas recomendadas

- SUS05-BP01 Usar la mínima cantidad de hardware para cumplir sus necesidades
- SUS05-BP02: Uso de los tipos de instancia con el menor impacto
- SUS05-BP03 Usar servicios administrados
- SUS05-BP04 Optimizar el uso de aceleradores de computación basados en hardware

SUS05-BP01 Usar la mínima cantidad de hardware para cumplir sus necesidades

Utilice la cantidad mínima de hardware para su carga de trabajo a fin de satisfacer eficazmente sus necesidades empresariales.

Patrones comunes de uso no recomendados:

- No supervisa la utilización de los recursos.
- Tiene recursos con un bajo nivel de utilización en su arquitectura.
- No se revisa la utilización del hardware estático para determinar si debe redimensionarse.
- No establece objetivos de utilización de hardware para su infraestructura de computación en función de los KPI empresariales.

Beneficios de establecer esta práctica recomendada: el redimensionamiento de sus recursos en la nube contribuye a reducir el impacto medioambiental de una carga de trabajo, ahorrar dinero y mantener los niveles de rendimiento.

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

Guía para la implementación

Seleccione de forma óptima el número total de hardware necesario para su carga de trabajo con el fin de mejorar su eficacia global. La Nube de AWS ofrece la flexibilidad de ampliar o reducir el

número de recursos de forma dinámica a través de diversos mecanismos, como AWS Auto Scaling, y satisfacer los cambios en la demanda. También proporciona API y SDK que permiten modificar los recursos con un mínimo esfuerzo. Use estas capacidades para hacer cambios frecuentes en las implementaciones de su carga de trabajo. Además, utilice las directrices de dimensionamiento de las herramientas de AWS para usar eficazmente sus recursos en la nube y satisfacer sus necesidades empresariales.

Pasos para la implementación

- Elija el tipo de instancia que mejor se adapte a sus necesidades.
  - How do I choose the appropriate Amazon EC2 instance type for my workload? (¿Cómo elijo el tipo de instancia Amazon EC2 apropiado para mi carga de trabajo?)
  - Attribute based instance type selection for Amazon EC2 Fleet (Selección de tipo de instancia basada en atributos para Flota de Amazon EC2).
  - Crear un grupo de Auto Scaling mediante la selección del tipo de instancia basada en atributos.
- Escale mediante pequeños incrementos para las cargas de trabajo variables.
- Utilice varias opciones de compra de computación para equilibrar la flexibilidad de las instancias, la escalabilidad y el ahorro de costes.
  - Las instancias bajo demanda son las más adecuadas para las cargas de trabajo nuevas, con estado y con picos que no pueden ser flexibles en cuanto al tipo de instancia, la ubicación o el tiempo.
  - Las instancias de spot son una excelente forma de complementar las demás opciones para aplicaciones con tolerancia a errores y que son flexibles.
  - Use los Savings Plans para computación en cargas de trabajo en estado estable que permiten flexibilidad si cambian sus necesidades (como zona de disponibilidad, región, familias de instancias o tipos de instancia).
- Utilice la diversidad de instancias y zonas de disponibilidad para maximizar la disponibilidad de las aplicaciones y aprovechar el exceso de capacidad cuando sea posible.
- Use las recomendaciones de tamaño adecuado de las herramientas de AWS para adaptar su carga de trabajo.
  - AWS Compute Optimizer
  - AWS Trusted Advisor
- Negocie acuerdos de nivel de servicio (SLA) que permitan una reducción temporal de la capacidad mientras la automatización despliega recursos de reemplazo.

Recursos

Documentos relacionados:

- Optimizing your AWS Infrastructure for Sustainability, Part I: Compute (Optimización de la infraestructura de AWS para la sostenibilidad, parte I: computación)
- Attribute based Instance Type Selection for Auto Scaling for Amazon EC2 Fleet (Selección de tipo de instancia basada en atributos para Auto Scaling para Flota de Amazon EC2)
- Documentación de AWS Compute Optimizer
- Operating Lambda: Performance optimization (Operación de Lambda: optimización de rendimiento)
- Documentación de Auto Scaling

Vídeos relacionados:

- Build a cost-, energy-, and resource-efficient compute environment (Crear un entorno de computación rentable, eficiente en términos de costes, energía y recursos)

Ejemplos relacionados:

- Well-Architected Labs - Rightsizing with AWS Compute Optimizer and Memory Utilization Enabled (Level 200) (Laboratorios de Well-Architected: redimensionamiento con AWS Compute Optimizer y uso de memoria activado [nivel 200])

SUS05-BP02: Uso de los tipos de instancia con el menor impacto

Supervise y utilice continuamente nuevos tipos de instancias para aprovechar las mejoras de la eficiencia energética.

Patrones comunes de uso no recomendados:

- Solo utiliza una familia de instancias.
- Solo utiliza instancias x86.
- Especifica un tipo de instancia en su configuración de Amazon EC2 Auto Scaling.
- Utiliza instancias de AWS para fines para las que no fueron diseñadas (por ejemplo, utiliza instancias optimizadas para computación para una carga de trabajo de uso intensivo de memoria).
- No evalúa de forma regular nuevos tipos de instancia.

- No comprueba recomendaciones de herramientas de dimensionamiento de AWS como AWS Compute Optimizer.

Beneficios de establecer esta práctica recomendada: Al utilizar instancias energéticamente eficientes y del tamaño adecuado, podrá reducir en gran medida el impacto medioambiental y el coste de su carga de trabajo.

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

Guía para la implementación

El uso de instancias eficientes en la carga de trabajo en la nube es fundamental para el uso reducido de recursos y la rentabilidad. Supervise de forma continuada el lanzamiento de nuevos tipos de instancia y aproveche las mejoras de la eficiencia energética; se incluyen los tipos de instancia diseñados para admitir cargas de trabajo específicas, como el entrenamiento y la inferencia en machine learning y la transcodificación de vídeo.

Pasos para la implementación

- Conozca y explore los tipos de instancia que pueden reducir el impacto medioambiental de su carga de trabajo.
  - Suscríbase a Novedades de AWS para estar al día con las últimas tecnologías e instancias de AWS.
  - Conozca los diferentes tipos de instancias de AWS.
  - Conozca las instancias basadas en Graviton de AWS que ofrecen el mejor rendimiento por vatio de uso de energía en Amazon EC2 con re:Invent 2020 - Conocer en profundidad las instancias de Amazon EC2 con procesador AWS Graviton2 y Conocer en profundidad las instancias C7g de Amazon EC2 y AWS Graviton3.
- Planifique y realice la transición de su carga de trabajo a los tipos de instancia con el menor impacto.
  - Defina un proceso para evaluar nuevas funciones o instancias para su carga de trabajo. Aproveche la agilidad de la nube para probar rápidamente cómo los nuevos tipos de instancia pueden mejorar la sostenibilidad medioambiental de su carga de trabajo. Utilice las métricas proxy para medir cuántos recursos necesita para completar una unidad de trabajo.
  - Si es posible, modifique su carga de trabajo para que funcione con diversas cantidades de vCPU y de memoria para sacar el máximo partido de su elección de tipo de instancia.

- Considere la posibilidad de cambiar su carga de trabajo a instancias basadas en Graviton para mejorar la eficiencia del rendimiento de su carga de trabajo.
  - AWS Graviton Fast Start
  - Considerations when transitioning workloads to AWS Graviton-based Amazon Elastic Compute Cloud instances (Consideraciones al trasladar cargas de trabajo a instancias de Amazon EC2 basadas en AWS Graviton)
  - Graviton2 de AWS para ISV
- Considere la selección de la opción de Graviton de AWS en el uso de los servicios administrados de AWS.
- Migre su carga de trabajo a las regiones que ofrezcan las instancias con menor impacto en la sostenibilidad y que sigan cumpliendo sus requisitos empresariales.
- 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. Las instancias de AWS Inferentia, como las instancias Inf2, ofrecen hasta un 50 % más de rendimiento por vatio que las instancias de Amazon EC2 comparables.
- Utilice Amazon SageMaker AI Inference Recommender para el tamaño correcto del punto de conexión de inferencia de ML.
- Para cargas de trabajo con picos (cargas de trabajo con requisitos poco frecuentes de capacidad adicional), utilice instancias de rendimiento ampliable.
- Para cargas de trabajo sin estado y tolerantes a errores, utilice Instancias de spot de Amazon EC2 para incrementar el uso global de la nube y reducir el impacto en la sostenibilidad de los recursos no utilizados.

- Opere y optimice su instancia de carga de trabajo.
  - Para las cargas de trabajo efímeras, evalúe las métricas de Amazon CloudWatch de instancias como `CPUUtilization` para identificar si la instancia está inactiva o infrautilizada.
  - Para cargas de trabajo estables, compruebe las herramientas de redimensionamiento de AWS como AWS Compute Optimizer a intervalos regulares para identificar las oportunidades de optimizar y dimensionar las instancias.
    - Well-Architected Lab - Rightsizing Recommendations (Laboratorio de Well-Architected: recomendaciones de redimensionamiento)
    - Well-Architected Lab - Rightsizing with Compute Optimizer (Laboratorio de Well-Architected: redimensionamiento con Compute Optimizer)

- Well-Architected Lab - Optimize Hardware Patterns and Observice Sustainability KPIs (Laboratorio de Well-Architected: optimizar los patrones de hardware y observar los KPI de sostenibilidad)

Recursos

Documentos relacionados:

- Optimización de la infraestructura de AWS para la sostenibilidad, Parte I: computación
- AWS Graviton
- Amazon EC2 DL1
- Flotas de reservas de capacidad de Amazon EC2
- Flota de spot de Amazon EC2
- Funciones: configuración de funciones de Lambda
- Selección de tipo de instancia basada en atributos para la flota de Amazon EC2
- Building Sustainable, Efficient, and Cost-Optimized Applications on AWS (Creación de aplicaciones sostenibles, eficientes y con optimización de costes en AWS)
- How the Contino Sustainability Dashboard Helps Customers Optimize Their Carbon Footprint (Cómo el panel de sostenibilidad de Contino ayuda a los clientes a optimizar su huella de carbono)

Vídeos relacionados:

- Conocer en profundidad las instancias de Amazon EC2 con procesador AWS Graviton2
- Conocer en profundidad las instancias C7g de Amazon EC2 y AWS Graviton3
- Build a cost-, energy-, and resource-efficient compute environment (Crear un entorno de computación rentable, eficiente en términos de costes, energía y recursos)

Ejemplos relacionados:

- Solution: Guidance for Optimizing Deep Learning Workloads for Sustainability on AWS (Solución: guía para optimizar las cargas de trabajo de aprendizaje profundo para la sostenibilidad en AWS)
- Well-Architected Lab - Rightsizing Recommendations (Laboratorio de Well-Architected: recomendaciones de redimensionamiento)

- Well-Architected Lab - Rightsizing with Compute Optimizer (Laboratorio de Well-Architected: redimensionamiento con Compute Optimizer)
- Well-Architected Lab - Optimize Hardware Patterns and Observe Sustainability KPIs (Laboratorio de Well-Architected: optimizar los patrones de hardware y observar los KPI de sostenibilidad)
- Well-Architected Lab - Migrating Services to Graviton (Laboratorio de Well-Architected: migración de servicios a Graviton)

SUS05-BP03 Usar servicios administrados

Utilice los servicios administrados para operar con mayor eficacia en la nube.

Patrones comunes de uso no recomendados:

- Utiliza instancias de Amazon EC2 con baja utilización para ejecutar sus aplicaciones.
- Su equipo interno solo administra la carga de trabajo, sin tiempo para centrarse en la innovación o las simplificaciones.
- Despliega y mantiene tecnologías para tareas que pueden ejecutarse con mayor eficacia en servicios administrados.

Beneficios de establecer esta práctica recomendada:

- El uso de servicios administrados traslada la responsabilidad a AWS, que dispone de información sobre millones de clientes que puede ayudar a impulsar nuevas innovaciones y eficiencias.
- El servicio administrado distribuye el impacto medioambiental del servicio entre muchos usuarios gracias a los planos de control de varios principios.

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

Guía para la implementación

Los servicios administrados traspasan a AWS la responsabilidad del mantenimiento de un uso elevado y la optimización de la sostenibilidad del hardware desplegado. Los servicios administrados también eliminan la carga operativa y administrativa del mantenimiento de un servicio, lo que permite a su equipo tener más tiempo para centrarse en la innovación.

Revise su carga de trabajo para identificar los componentes que se pueden reemplazar por servicios administrados por AWS. Por ejemplo, Amazon RDS, Amazon Redshift y Amazon ElastiCache

proporcionan un servicio de base de datos administrada. Amazon Athena, Amazon EMR y Amazon OpenSearch Service proporcionan un servicio de análisis administrados.

Pasos para la implementación

1. Realice un inventario de su carga de trabajo para servicios y componentes.
2. Evalúe e identifique los componentes que se pueden reemplazar por servicios administrados. A continuación, encontrará algunos ejemplos de cuándo podría plantearse el uso de un servicio administrado:

Task	What to use on AWS
Alojamiento de una base de datos	Use instancias de Amazon Relational Database Service (Amazon RDS) administradas en vez de mantener sus propias instancias de Amazon RDS en Amazon Elastic Compute Cloud (Amazon EC2).
Alojamiento de una carga de trabajo de contenedores	Use AWS Fargate en vez de implementar su propia infraestructura de contenedores.
Alojamiento de aplicaciones web	Use AWS Amplify Hosting como servicio completamente administrado de CI/CD y de alojamiento para sitios web estáticos y aplicaciones web reproducidas en el servidor.

3. Identifique las dependencias y cree un plan de migraciones. Actualice los runbooks y las guías de estrategias según corresponda.
   - AWS Application Discovery Service recopila y presenta de modo automático la información detallada sobre el uso y las dependencias de aplicaciones para que pueda tomar decisiones más fundamentadas cuando planifique la migración.
4. Pruebe el servicio antes de migrar al servicio administrado.
5. Utilice el plan de migración para reemplazar los servicios autoadministrados por un servicio administrado.
6. Supervise continuamente el servicio una vez finalizada la migración para realizar los ajustes necesarios y optimizar el servicio.

Recursos

Documentos relacionados:

- Productos de Nube de AWS
- Calculadora de coste total de propiedad (TCO) de AWS
- Amazon DocumentDB
- Amazon Elastic Kubernetes Service (EKS)
- Amazon Managed Streaming for Apache Kafka (Amazon MSK)

Vídeos relacionados:

- Cloud operations at scale with AWS Managed Services(Operaciones en la nube a escala con AWS Managed Services)

SUS05-BP04 Optimizar el uso de aceleradores de computación basados en hardware

Optimice el uso de instancias de computación acelerada para reducir las demandas de infraestructura física de su carga de trabajo.

Patrones comunes de uso no recomendados:

- No supervisa el uso de GPU.
- Utiliza una instancia de uso general para la carga de trabajo, mientras que una instancia personalizada puede ofrecer mayor rendimiento, menor coste y mejor rendimiento por vatio.
- Utiliza aceleradores de computación basados en hardware para tareas en las que es más eficiente utilizar alternativas basadas en CPU.

Beneficios de establecer esta práctica recomendada: al optimizar el uso de los aceleradores basados en hardware, puede reducir las demandas de infraestructura física de su carga de trabajo.

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

Guía para la implementación

Si necesita una gran capacidad de procesamiento, puede beneficiarse del uso de instancias de computación acelerada, que proporcionan acceso a aceleradores de computación basados en hardware, como unidades de procesamiento gráfico (GPU) y matrices de puertas programables en

campo (FPGA). Estos aceleradores de hardware realizan ciertas funciones, como el procesamiento gráfico o la concordancia de patrones de datos, de forma más eficiente que las alternativas basadas en CPU. Muchas cargas de trabajo aceleradas, como el renderizado, la transcodificación y el machine learning, son muy variables en cuanto al uso de recursos. Ejecute este hardware solo durante el tiempo que sea necesario y retírelo mediante automatización cuando no se requiera para minimizar los recursos consumidos.

Pasos para la implementación

- Identifique qué instancias de computación acelerada pueden satisfacer sus necesidades.
- Para las cargas de trabajo de machine learning, utilice hardware personalizado específico para su carga de trabajo, como AWS Trainium, AWS Inferentiay Amazon EC2 DL1. Las instancias de AWS Inferentia, como las instancias Inf2, tienen hasta un 50 % más de rendimiento por vatio en comparación con instancias de Amazon EC2 comparables.
- Recopile la métrica de uso de sus instancias de computación acelerada. Por ejemplo, puede usar un agente de CloudWatch para recopilar métricas como `utilization_gpu` y `utilization_memory` para sus GPU, como se muestra en Collect NVIDIA GPU metrics with Amazon CloudWatch (Recopilación de métricas de CPU de 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.
  - Optimizar la configuración de GPU
  - GPU Monitoring and Optimization in the Deep Learning AMI (Supervisión y optimización de la GPU en la AMI de aprendizaje profundo)
  - Optimizing I/O for GPU performance tuning of deep learning training in Amazon SageMaker AI (Optimización de la E/S para el ajuste del rendimiento de la GPU en el entrenamiento del aprendizaje profundo en Amazon SageMaker)
- Utilice las bibliotecas de alto rendimiento y los controladores de GPU más recientes.
- Use la automatización para liberar instancias de GPU cuando no se estén usando.

Recursos

Documentos relacionados:

- Computación acelerada
- Let's Architect! Architecting with custom chips and accelerators (Arquitectura con chips y aceleradores personalizados)

- How do I choose the appropriate Amazon EC2 instance type for my workload? (¿Cómo elijo el tipo de instancia de EC2 apropiado para mi carga de trabajo?)
- Instancias VT1 de Amazon EC2
- Choose the best AI accelerator and model compilation for computer vision inference with Amazon SageMaker AI (Elija el mejor acelerador de IA y compilación de modelos para la inferencia de visión artificial con Amazon SageMaker)

Vídeos relacionados:

- How to select Amazon EC2 GPU instances for deep learning (Cómo seleccionar las instancias de GPU de Amazon EC2 para el aprendizaje profundo)
- Deploying Cost-Effective Deep Learning Inference (Despliegue rentable de la inferencia del aprendizaje profundo)

Proceso y cultura

Pregunta

- SUS 6 ¿Cómo respaldan sus procesos organizativos sus objetivos de sostenibilidad?

SUS 6 ¿Cómo respaldan sus procesos organizativos sus objetivos de sostenibilidad?

Realice cambios en sus prácticas de desarrollo, prueba e implementación como forma de reducir el impacto en la sostenibilidad.

Prácticas recomendadas

- SUS06-BP01 Adoptar métodos que permitan introducir mejoras en la sostenibilidad rápidamente
- SUS06-BP02: Mantenimiento de una carga de trabajo actualizada
- SUS06-BP03: Incremento de la utilización de los entornos de compilación
- SUS06-BP04 Usar granjas de dispositivos administrados para pruebas

SUS06-BP01 Adoptar métodos que permitan introducir mejoras en la sostenibilidad rápidamente

Adopte métodos y procesos para validar las mejoras potenciales, minimizar los costes de las pruebas y ofrecer pequeñas mejoras.

Patrones comunes de uso no recomendados:

- La revisión de su solicitud de sostenibilidad es una tarea que se realiza solo una vez al comienzo de un proyecto.
- Su carga de trabajo se ha quedado obsoleta, ya que el proceso de lanzamiento es demasiado complejo para incorporar pequeños cambios para la eficiencia de los recursos.
- No dispone de mecanismos para mejorar su carga de trabajo para la sostenibilidad.

Beneficios de establecer esta práctica recomendada: si establece un proceso para incorporar mejoras de sostenibilidad y realizar un seguimiento de ellas, podrá adoptar continuamente nuevas características y capacidades, eliminar problemas y mejorar la eficacia de la carga de trabajo.

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

Guía para la implementación

Pruebe y valide las mejoras de sostenibilidad potenciales antes de desplegarlas en producción. Tenga en cuenta el coste de las pruebas al calcular las posibles ventajas futuras de una mejora. Desarrolle métodos de prueba de bajo coste para ofrecer pequeñas mejoras.

Pasos para la implementación

- Añada requisitos para la mejora de la sostenibilidad a sus tareas pendientes de desarrollo.
- Utilice un proceso de mejora iterativo para identificar, evaluar, priorizar, probar y desplegar estas mejoras.
- Mejore y optimice continuamente sus procesos de desarrollo. Por ejemplo, automatice su proceso de entrega de software mediante canalizaciones de integración y entrega continuas (CI/CD) para probar y desplegar posibles mejoras con el fin de reducir el nivel de esfuerzo y limitar los errores provocados por los procesos manuales.
- Desarrolle y pruebe posibles mejoras con los componentes representativos mínimos viables para reducir el coste de las pruebas.
- Evalúe continuamente el impacto de las mejoras y realice los ajustes necesarios.

Recursos

Documentos relacionados:

- AWS habilita soluciones de sostenibilidad
- Scalable agile development practices based on AWS CodeCommit(Prácticas de desarrollo ágil escalables basadas en AWS CodeCommit)

Vídeos relacionados:

- Delivering sustainable, high-performing architectures (Entrega de arquitecturas sostenibles y de alto rendimiento)

Ejemplos relacionados:

- Well-Architected Lab - Turning cost & usage reports into efficiency reports (Laboratorio de Well-Architected: convertir informes sobre costes y uso en informes de eficiencia)

SUS06-BP02: Mantenimiento de una carga de trabajo actualizada

Mantenga actualizada su carga de trabajo para adoptar características eficaces, eliminar problemas y mejorar la eficacia general de su 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 su carga de trabajo.

Beneficios de establecer esta práctica recomendada: al establecer un proceso para mantener su carga de trabajo actualizada, puede adoptar nuevas funciones y capacidades, resolver problemas y mejorar la eficiencia de la carga de trabajo.

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

Guía para la implementación

La actualización de sistemas operativos, tiempos de ejecución, middlewares, bibliotecas y aplicaciones puede mejorar la eficacia de la carga de trabajo y facilitar la adopción de tecnologías más eficientes. Un software actualizado también puede incluir características que midan el impacto de su carga de trabajo en la sostenibilidad de forma más precisa, ya que los proveedores ofrecen

características para cumplir sus objetivos de sostenibilidad propios. Adopte una cadencia periódica para mantener su carga de trabajo al día de las últimas características y versiones.

Pasos para la implementación

- Defina un proceso y un calendario para evaluar nuevas funciones o instancias para su carga de trabajo. Aproveche la agilidad de la nube para probar rápidamente cómo las nuevas funciones pueden mejorar su carga de trabajo para:
  - Reducir el impacto en la sostenibilidad.
  - Lograr la eficacia operativa.
  - Eliminar las barreras para una mejora planificada.
  - Mejorar su capacidad a la hora de medir y administrar las repercusiones en la sostenibilidad.
- Inventariar el software y la arquitectura de su carga de trabajo e identificar los componentes que deben actualizarse.
  - 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 saber rápidamente qué instancias están ejecutando el software y las configuraciones requeridas por su política de software así como las instancias que deben actualizarse.
- Entienda cómo actualizar los componentes de su carga de trabajo.

Workload component	How to update
Imágenes de máquina	Use EC2 Image Builder para administrar las actualizaciones de las imágenes de máquina de Amazon (AMI) para imágenes de servidor Linux o Windows.
Imágenes de contenedor	Use Amazon Elastic Container Registry (Amazon ECR) con su canalización existente para administrar las imágenes de Amazon Elastic Container Service (Amazon ECS).
AWS Lambda	AWS Lambda incluye características de administración de versiones.

- Utilice la automatización del proceso de actualización para reducir el nivel de esfuerzo para desplegar nuevas funciones y limitar los errores causados por los procesos manuales.

- Puede utilizar CI/CD para actualizar automáticamente las AMI, las imágenes de contenedor y otros artefactos relacionados con su aplicación en la nube.
- Puede utilizar herramientas como AWS Systems Manager Patch Manager para automatizar el proceso de actualizaciones del sistema y programar la actividad con Ventanas de mantenimiento de AWS Systems Manager.

Recursos

Documentos relacionados:

- Centro de arquitectura de AWS
- Novedades de AWS
- Herramientas para desarrolladores de AWS

Ejemplos relacionados:

- Well-Architected Labs - Inventory and Patch Management (Laboratorios de Well-Architected: administración de inventario y parches)
- Laboratorio: AWS Systems Manager

SUS06-BP03: Incremento de la utilización de los entornos de compilación

Aumente la utilización de recursos para desarrollar, probar y crear sus cargas de trabajo.

Patrones comunes de uso no recomendados:

- Aprovisiona o finaliza manualmente sus entornos de compilación.
- Mantiene sus entornos de compilación en funcionamiento independientemente de las actividades de prueba, compilación o lanzamiento (por ejemplo, ejecución de un entorno fuera del horario laboral de los miembros de su equipo de desarrollo).
- Aprovisiona en exceso los recursos para sus entornos de compilación.

Beneficios de establecer esta práctica recomendada: al aumentar la utilización de los entornos de compilación, puede mejorar la eficacia general de su carga de trabajo en la nube y, al mismo tiempo, asignar los recursos a los creadores para que desarrollen, prueben y creen de forma eficaz.

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

Guía para la implementación

Use la automatización y la infraestructura como código para incorporar los entornos de compilación cuando sea necesario y retirarlos cuando no se utilicen. Un patrón común consiste en programar períodos de disponibilidad que coincidan con las horas de trabajo de los miembros del equipo de desarrollo. Sus entornos de prueba deben parecerse mucho a la configuración de producción. Sin embargo, busque oportunidades para utilizar tipos de instancia con capacidad de ampliación, instancias de spot de Amazon EC2, servicios de base de datos de escalamiento automático, contenedores y tecnologías sin servidor para coordinar el desarrollo y la capacidad de prueba con el uso. Limite el volumen de datos para cumplir únicamente los requisitos de prueba. Si utiliza datos de producción en las pruebas, estudie las posibilidades de compartir los datos de producción y no trasladarlos.

Pasos para la implementación

- Utilice la infraestructura como código para aprovisionar sus entornos de compilación.
- Use la automatización para administrar el ciclo de vida de sus entornos de desarrollo y pruebas, y maximizar la eficiencia de sus recursos de compilación.
- Utilice estrategias para maximizar la utilización de los entornos de desarrollo y prueba.
  - Use el mínimo viable de entornos representativos para desarrollar y probar mejoras potenciales.
  - Utilice tecnologías sin servidor si es posible.
  - Use instancias bajo demanda para complementar sus dispositivos de desarrollador.
  - Use tipos de instancia con capacidad de ampliación, instancias de spot y otras tecnologías para alinear la capacidad de creación con el uso.
  - Adopte servicios nativos en la nube para obtener un acceso seguro al shell de instancias en lugar de implementar flotas de hosts bastión.
  - Escale automáticamente sus recursos de compilación en función de sus tareas de compilación.

Recursos

Documentos relacionados:

- AWS Systems Manager Session Manager
- Instancias de rendimiento ampliable de Amazon EC2
- ¿Qué es AWS CloudFormation?
- ¿Qué es AWS CodeBuild?

- Instance Scheduler en AWS

Vídeos relacionados:

- Prácticas recomendadas de integración continua

SUS06-BP04 Usar granjas de dispositivos administrados para pruebas

Utilice granjas de dispositivos administrados para probar eficazmente una nueva característica en un conjunto representativo de hardware.

Patrones comunes de uso no recomendados:

- Prueba y despliega manualmente su aplicación en dispositivos físicos individuales.
- No utiliza el servicio de pruebas de aplicaciones para probar e interactuar con sus aplicaciones (por ejemplo, Android, iOS y aplicaciones web) en dispositivos físicos reales.

Beneficios de establecer esta práctica recomendada: el uso de granjas de dispositivos administrados para probar aplicaciones con tecnología basada en la nube proporciona una serie de ventajas:

- Incluyen características más eficaces para probar la aplicación en una amplia gama de dispositivos.
- Eliminan la necesidad de una infraestructura interna para las pruebas.
- Ofrecen diversos tipos de dispositivos, incluido el hardware más antiguo y menos popular, lo que elimina la necesidad de actualizaciones innecesarias de los dispositivos.

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

Guía para la implementación

El uso de granjas de dispositivos administrados puede ayudarle a agilizar el proceso de prueba de nuevas características en un conjunto representativo de hardware. Las granjas de dispositivos administrados ofrecen diversos tipos de dispositivos, incluido el hardware más antiguo y menos popular, y evitan el impacto en la sostenibilidad para el cliente que tienen las actualizaciones innecesarias de dispositivos.

Pasos para la implementación

- Defina los requisitos y el plan de pruebas (como el tipo de prueba, los sistemas operativos y el calendario de pruebas).
  - Puede utilizar Amazon CloudWatch RUM para recopilar y analizar datos en el cliente y dar forma a su plan de pruebas.
- Seleccione la granja de dispositivos administrada que pueda admitir sus requisitos de pruebas. Por ejemplo, puede utilizar AWS Device Farm para probar y comprender el impacto de sus cambios en un conjunto representativo de hardware.
- Utilice la integración continua/despliegue continuo (CI/CD) para programar y ejecutar sus pruebas.
  - Integrating AWS Device Farm with your CI/CD pipeline to run cross-browser Selenium tests (Integración de AWS Device Farm con su canalización de CI/CD para ejecutar pruebas de Selenium en varios navegadores)
  - Building and testing iOS and iPadOS apps with AWS DevOps and mobile services (Compilar y probar aplicaciones de iOS y iPadOS con AWS DevOps y servicios móviles)
- Revise continuamente los resultados de sus pruebas y efectúe las mejoras necesarias.

Recursos

Documentos relacionados:

- Lista de dispositivos de AWS Device Farm
- Visualización del panel de CloudWatch RUM

Ejemplos relacionados:

- Aplicación de muestra de AWS Device Farm para Android
- Aplicación de muestra de AWS Device Farm para iOS
- Pruebas web de Appium para AWS Device Farm

Vídeos relacionados:

- Optimize applications through end user insights with Amazon CloudWatch RUM (Optimizar las aplicaciones mediante la información del usuario final con Amazon CloudWatch RUM)

Avisos

Los clientes son responsables de realizar sus propias evaluaciones de la información contenida en este documento. Este documento: (a) solo tiene fines informativos, (b) representa las prácticas y las ofertas de productos vigentes de AWS, que están sujetas a cambios sin previo aviso, y (c) no crea ningún compromiso ni garantía de AWS y sus empresas afiliadas, proveedores o concesionarios de licencias. Los productos o servicios de AWS se proporcionan «tal cual», sin garantías, declaraciones ni condiciones de ningún tipo, ya sean explícitas o implícitas. Las responsabilidades y obligaciones de AWS en relación con sus clientes se rigen por los acuerdos de AWS, y este documento no modifica ni forma parte de ningún acuerdo entre AWS y sus clientes.

Copyright © 2021, Amazon Web Services, Inc. o sus empresas afiliadas.