aws

Aplicación del marco AWS Well-Architected para Amazon Neptune

AWS Guía prescriptiva

AWS Guía prescriptiva: Aplicación del marco AWS Well-Architected para Amazon Neptune

Table of Contents

Introducción 1
Destinatarios previstos 1
Objetivos 2
Excelencia operativa 3
Automatice la implementación mediante un enfoque de IaC 3
Realice cambios frecuentes, pequeños y reversibles 4
Anticipe el fracaso 4
Aprenda de todos los fallos operativos 5
Utilice las funciones de registro para supervisar la actividad no autorizada o anómala 6
Seguridad 7
Implementar la seguridad de los datos 8
Proteja sus redes 9
Implemente la autenticación y la autorización 9
Fiabilidad 11
Comprenda las cuotas de servicio de Neptune 11
Comprenda los patrones de despliegue de Neptune 12
Gestione y escale los clústeres de Neptune 13
Gestione las copias de seguridad y los eventos de conmutación por error 14
Eficiencia del rendimiento 16
Comprenda el modelado de gráficos 16
Optimización de consultas 17
Clústeres del tamaño correcto 19
Optimice las escrituras 20
Optimización de costos 22
Comprenda los patrones de uso y los servicios necesarios 22
Seleccione los recursos prestando atención al costo 23
Elija la mejor configuración de instancias de Neptune para su carga de trabajo 24
Almacenamiento y transferencia de datos del tamaño adecuado 26
Sostenibilidad 27
Selección de regiones de AWS 27
El consumo se basa en los patrones de comportamiento de los usuarios 28
Optimice los patrones de arquitectura y desarrollo de software 28
Recursos 30
Colaboradores 31

Historial de documentos 32
Glosario 33
..... 33
A 34
B 37
C 39
D 42
E 46
F 49
G 51
H 52
I 53
L 56
M 57
O 61
P 64
Q 67
R 67
S 70
T 74
U 76
V 77
W 77
Z 78
..... lxxx

Aplicación del marco AWS Well-Architected para Amazon Neptune

Amazon Web Services (colaboradores)

Septiembre de 2023 (historial del documento)

Puede crear soluciones basadas en gráficos en Amazon Web Services (AWS) mediante Amazon Neptune. Esta guía proporciona orientación prescriptiva para aplicar los principios del AWS Well-Architected Framework al planificar la implementación de Neptune. La aplicación del AWS marco Well-Architected para Amazon Neptune Analytics trata el mismo tema para el motor de análisis gráfico Neptune.

El AWS Well-Architected Framework le ayuda a crear infraestructuras seguras, de alto rendimiento, resilientes y eficientes para una variedad de aplicaciones y cargas de trabajo. También proporciona un enfoque coherente para evaluar las arquitecturas e implementar diseños escalables.

El AWS Well-Architected Framework se basa en los seis pilares siguientes:

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

Esta guía proporciona información sobre los pilares de diseño y las prácticas recomendadas de Well-Architected Framework, así como consideraciones que se deben tener en cuenta al implementar Neptune on. AWS

Destinatarios previstos

Esta guía está destinada a ingenieros de datos, arquitectos de soluciones y analistas de datos que diseñan e implementan soluciones que utilizan gráficos. AWS

Objetivos

Esta guía puede ayudarle a usted y a su organización a hacer lo siguiente:

- Elija entre las opciones de implementación y los lenguajes de consulta compatibles, según su caso de uso y sus patrones de consulta.
- Siga los patrones de diseño de Well-Architected de AWS que le ayudarán a mejorar la resiliencia y la seguridad.
- Diseñe sus consultas para obtener un rendimiento y un ahorro de costes óptimos.
- Aprenda a ser eficiente desde el punto de vista operativo al gestionar su clúster de Neptune en producción.

Pilar de excelencia operativa

El pilar de excelencia operativa del AWS Well-Architected Framework se centra en ejecutar y monitorear los sistemas, y en mejorar continuamente los procesos y procedimientos. Incluye la capacidad de respaldar el desarrollo y ejecutar las cargas de trabajo de manera eficaz, obtener información sobre su funcionamiento y mejorar continuamente los procesos y procedimientos de apoyo para ofrecer valor empresarial. Puede reducir la complejidad operativa mediante la autorreparación de las cargas de trabajo, que detectan y solucionan la mayoría de los problemas sin intervención humana. Puede trabajar para lograr este objetivo siguiendo las prácticas recomendadas que se describen en esta sección. Utilice las métricas y los mecanismos de Amazon Neptune para responder adecuadamente cuando su carga de trabajo se desvíe del comportamiento esperado. APIs

Este análisis del pilar de la excelencia operativa se centra en las siguientes áreas clave:

- Infraestructura como código (IaC)
- Administración de cambios
- Estrategias de resiliencia
- Administración de incidentes
- Informes de auditoría para garantizar el cumplimiento
- Registro y supervisión

Automatice la implementación mediante un enfoque de IaC

Entre las prácticas recomendadas para automatizar el despliegue en Neptune mediante IaC se incluyen las siguientes:

- Aplique la infraestructura como código (IaC) para implementar los clústeres de Neptune siempre que sea posible. Para una configuración coherente del entorno, utilice una AWS CloudFormationplantilla o HashiCorp Terraform para crear todos los recursos necesarios para su clúster. AWS Cloud Development Kit (AWS CDK)
- Automatice los procedimientos operativos de Neptune, como el cambio de tamaño de las instancias, la adición o eliminación de réplicas de lectura o la realización de conmutaciones por error manuales en tablas globales, siempre que sea posible.

- Almacene las cadenas de conexión de forma externa a su cliente. Utilice los procesos de extracción, transformación y carga (ETL) para facilitar las estrategias de blue/green implementación, la recuperación ante desastres (DR) y las migraciones a nuevos clústeres con un tiempo de inactividad prácticamente nulo. Las cadenas de conexión se pueden almacenar en AWS Secrets ManagerAmazon DynamoDB o en cualquier ubicación en la que se puedan cambiar de forma dinámica.
- Usa etiquetas para añadir metadatos a tus recursos de Neptune y realiza un seguimiento del uso en función de las etiquetas. Para obtener más información, consulte Etiquetado de los recursos de Amazon Neptune.

Realice cambios frecuentes, pequeños y reversibles

Las siguientes recomendaciones se centran en cambios pequeños y reversibles para minimizar la complejidad y reducir la probabilidad de que se interrumpa la carga de trabajo:

- Guarde las plantillas y scripts de IaC en un servicio de control de código fuente, como GitHub o GitLab.

> **Important**
>
> No almacene AWS las credenciales en el control de código fuente.

- Exija que las implementaciones de iAC utilicen un servicio de integración y entrega continuas (CI/CD), como o. AWS CodeDeployAWS CodeBuild Estos servicios compilan, prueban e implementan código en un entorno no de producción que contiene un clúster efímero de Neptuno antes de afectar al clúster de Amazon Neptune de producción.
- Pruebe las consultas de infraestructura y aplicaciones en un entorno inferior antes de implementarlas en producción. Esto minimizará la probabilidad de una interrupción y ayudará a garantizar que funcionen bien con su carga de trabajo y su escalabilidad.

Anticipe el fracaso

Una infraestructura que se recupere automáticamente ejemplifica la excelencia operativa al anticipar las fallas e intentar resolver cualquier problema sin intervención. Las siguientes recomendaciones le ayudarán a alcanzar esa madurez con Neptune:

- Cree un plan de supervisión que utilice CloudWatch las métricas de Amazon para supervisar el uso de la CPU y la memoria de la instancia de base de datos y comprender los patrones de uso. Cree CloudWatch paneles y alarmas para las métricas clave y las respuestas del cliente de Neptune que se encuentran en los registros de sus aplicaciones. Para obtener más información sobre los indicadores de un uso elevado o bajo de la CPU, consulte Uso CloudWatch para supervisar el rendimiento de una instancia de base de datos en Neptuno en la documentación de Neptuno.

  Si con frecuencia recibe out-of-memory excepciones en sus consultas cuando el nivel `FreeableMemory` es bajo, considere la posibilidad de utilizar una instancia de la familia X2.
- Configure las notificaciones para supervisar el estado del cúmulo de Neptune. Por ejemplo, `BufferCacheHitRatio` debe estar constantemente alta (superior al 99,9 por ciento), mientras que `MainRequestQueuePendingRequests` debe estar constantemente baja (idealmente 0, pero en función de sus requisitos y de la tolerancia a la latencia).
- Considere la posibilidad de utilizar réplicas de lectura para lograr una alta disponibilidad en Neptune. Debe tener al menos dos réplicas de lectura en zonas de disponibilidad diferentes a las de la instancia de grabación para garantizar que siempre haya una instancia disponible para atender las consultas de lectura durante un evento de conmutación por error.
- Escale automáticamente las réplicas de lectura en función de las métricas de uso. Para obtener más información, consulte Escalar automáticamente el número de réplicas en un clúster de base de datos de Amazon Neptune.
- Pruebe la conmutación por error de la instancia de base de datos para comprender cuánto tiempo tarda el proceso en su caso de uso.
- Si su aplicación necesita sobrevivir a una Región de AWS interrupción total, considere la posibilidad de utilizar bases de datos globales como parte de sus planes de recuperación ante desastres.

Aprenda de todos los fallos operativos

Una infraestructura que se recupere automáticamente es un esfuerzo a largo plazo que se desarrolla de forma iterativa a medida que se producen problemas poco frecuentes o las respuestas no son tan eficaces como se desearía. La adopción de las siguientes prácticas impulsa la concentración hacia ese objetivo:

- Impulse la mejora aprendiendo de todos los fracasos.

- Comparta lo aprendido entre los equipos y la organización. Si varios equipos de una organización utilizan Neptune, cree una sala de chat o un grupo de usuarios común para compartir los aprendizajes y las mejores prácticas.

Utilice las funciones de registro para supervisar la actividad no autorizada o anómala

Para observar patrones anómalos de rendimiento y actividad, almacene los registros en Amazon CloudWatch Logs. Tenga en cuenta las siguientes prácticas recomendadas:

- Habilite el registro de consultas lentas. Revise periódicamente el registro y diagnostique por qué determinadas consultas son lentas. Utilice los puntos finales de Neptune para explicar y perfilar Gremlin, SPARQL u OpenCypher para comprender por qué estas consultas son lentas.
- Habilite los registros de auditoría de Neptune y revíselos periódicamente para detectar anomalías o accesos no autorizados.
- Si utiliza el registro de consultas lentas o el registro de auditoría, habilite la publicación en los registros. CloudWatch Esto le ayudará a evitar quedarse sin espacio en disco en las instancias. Las instancias de Neptune tienen una capacidad de almacenamiento de registros limitada y sobrescribirán los archivos de registro más antiguos cuando se exceda el espacio de registro. CloudWatch Los registros permiten la retención de registros a largo plazo. Las capacidades de supervisión mejoradas de CloudWatch los registros mejorarán su capacidad para consultar los registros y diagnosticar problemas.
- Para facilitar mejores herramientas de análisis para sus registros de auditoría, puede configurar un clúster de base de datos de Neptune para publicar los datos del registro de auditoría en CloudWatch un grupo de registros de Logs. Con CloudWatch los registros, puede realizar un análisis en tiempo real de los datos de registro, CloudWatch utilizarlos para crear alarmas y ver métricas, y utilizar CloudWatch los registros para almacenar los registros en un lugar de almacenamiento muy duradero. Para obtener más información, consulte Publicar registros de Neptune en Amazon CloudWatch Logs.
- Neptune admite el registro de las acciones del plano de control mediante. AWS CloudTrail Para obtener más información, consulte Registrar llamadas a la API de Amazon Neptune con. AWS CloudTrail

Pilar de seguridad

La seguridad en la nube AWS es la máxima prioridad. Como AWS cliente, usted se beneficia de una arquitectura de centro de datos y red diseñada para cumplir con los requisitos de las organizaciones más sensibles a la seguridad.

La seguridad es una responsabilidad compartida entre usted AWS y usted. El modelo de responsabilidad compartida la describe como seguridad de la nube y seguridad en la nube:

- Seguridad de la nube: AWS es responsable de proteger la infraestructura que se ejecuta Servicios de AWS en la Nube de AWS. AWS también le proporciona servicios que puede utilizar de forma segura. Los auditores externos prueban y verifican periódicamente la eficacia de la AWS seguridad como parte de los programas de AWS cumplimiento. Para obtener más información acerca de los programas de conformidad que se aplican a Amazon Neptune, consulte Servicios de AWS en el ámbito del programa de conformidad.
- Seguridad en la nube: su responsabilidad viene determinada por lo Servicio de AWS que utilice. También es responsable de otros factores, incluida la confidencialidad de los datos, los requisitos de la empresa y la legislación y los reglamentos aplicables. Para obtener más información sobre la privacidad de datos, consulte Preguntas frecuentes sobre la privacidad de datos. Para obtener información sobre la protección de datos en Europa, consulte el modelo de responsabilidad AWS compartida y la entrada del blog sobre el RGPD.

El pilar de seguridad del AWS Well-Architected Framework le ayuda a entender cómo aplicar el modelo de responsabilidad compartida al utilizar Neptune. En los siguientes temas, se le mostrará cómo configurar Neptune para satisfacer sus objetivos de seguridad y conformidad. También aprenderá a usar otros Servicios de AWS que le ayuden a monitorear y proteger sus recursos de Neptune.

El pilar de seguridad incluye las siguientes áreas de enfoque clave:

- Seguridad de los datos
- Seguridad de la red
- Autenticación y autorización

Implementar la seguridad de los datos

Las filtraciones y filtraciones de datos ponen en riesgo a sus clientes y pueden tener un impacto negativo sustancial en su empresa. Las siguientes prácticas recomendadas ayudan a proteger los datos de sus clientes de una exposición inadvertida o malintencionada:

- Los nombres de los clústeres, las etiquetas, los grupos de parámetros, las funciones AWS Identity and Access Management (IAM) y otros metadatos no deben contener información confidencial o delicada, ya que esos datos pueden aparecer en los registros de facturación o diagnóstico.
- URIs o los enlaces a servidores externos almacenados como datos en Neptune no deben contener información sobre credenciales para validar las solicitudes.
- Las instancias cifradas de Neptune ofrecen una capa adicional de protección de datos al ayudarle a proteger los datos del acceso no autorizado al almacenamiento subyacente. Puede utilizar el cifrado de Neptune para aumentar la protección de datos de las aplicaciones implementadas en la nube. También puede utilizar el cifrado de Neptune para cumplir con los requisitos de conformidad de los datos en reposo.

  Para habilitar el cifrado para una nueva instancia de base de datos de Neptune, elija Sí en la sección Habilitar el cifrado de la consola Neptune (seleccionada de forma predeterminada) o configurando la propiedad en. AWS::Neptune::DBCluster::StorageEncrypted AWS CloudFormationSi el cifrado está activado, Neptune utilizará la clave gestionada por AWS del Amazon Relational Database Service (Amazon RDS) de forma predeterminada, o puede crear una clave gestionada por el cliente. Para obtener información sobre la creación de una instancia de base de datos Neptune, consulte Creación de un nuevo clúster de base de datos Neptune. Para obtener más información, consulte Cifrar los recursos de Neptune en reposo. Las instantáneas automatizadas y manuales utilizan el mismo cifrado que seleccionó para el clúster de Neptune.
- Cuando utilice los lenguajes SPARQL y OpenCypher, practique las técnicas adecuadas de validación y parametrización de las entradas para evitar la inyección de SQL y otras formas de ataques. Evite crear consultas que utilicen la concatenación de cadenas con entradas proporcionadas por el usuario. Utilice consultas parametrizadas o sentencias preparadas para pasar de forma segura los parámetros de entrada a la base de datos de gráficos. Para obtener más información, consulte Ejemplos de consultas parametrizadas de OpenCypher y SPARQL Injection Defence.
- Para el lenguaje Gremlin, utilice variantes del lenguaje Gremlin en lugar de pasar directamente scripts Gremlin basados en cadenas para evitar posibles problemas de inyección.

Proteja sus redes

Solo se puede crear un clúster de base de datos de Amazon Neptune en una nube privada virtual (VPC). AWS Solo se puede acceder a los puntos de enlace del clúster de base de datos de Neptune dentro de esa VPC, normalmente desde una instancia de Amazon Elastic Compute Cloud (Amazon EC2) que se ejecute en esa VPC. Puede proteger sus datos de Neptune limitando el acceso a la VPC en la que se encuentra su clúster de base de datos de Neptune. Para obtener más información, consulte Conectarse a su gráfico de Amazon Neptune.

Para proteger sus datos en tránsito, Neptune aplica conexiones SSL a través de HTTPS a cualquier instancia o punto final del clúster mediante protocolos y cifrados seguros. Neptune proporciona certificados SSL para sus instancias de base de datos de Neptune. Los certificados SSL de Neptune solo admiten nombres de host de punto final de clúster, punto final de lector y punto final de instancia. Si utilizas un balanceador de carga o un servidor proxy (por ejemplo HAProxy), debes usar la terminación SSL y tener tu propio certificado SSL en el servidor proxy. El acceso directo SSL no funciona porque los certificados SSL proporcionados no coinciden con el nombre de host del servidor proxy. Para obtener más información sobre la conexión a los puntos de enlace de Neptuno con SSL, consulte Uso del punto de enlace HTTP REST para conectarse a una instancia de base de datos de Neptune.

Implemente la autenticación y la autorización

Para controlar quién puede realizar las acciones de administración de Neptune en los clústeres y las instancias de base de datos de Neptune, utilice las credenciales de IAM. Cuando se conecta para AWS utilizar credenciales de IAM, su función de IAM debe tener políticas de IAM que concedan los permisos necesarios para realizar las operaciones de administración de Neptune. Asegúrese de seguir el principio de privilegios mínimos y conceder solo los permisos necesarios para completar una tarea. Para obtener más información, consulte Uso de diferentes tipos de políticas de IAM para controlar el acceso a Neptune y Autenticación de IAM mediante credenciales temporales.

Para controlar quién puede conectarse a un clúster de Neptune y consultar los datos, puede usar IAM para autenticarse en su instancia de base de datos o clúster de base de datos de Neptune. Si habilita la autenticación de IAM en un clúster de base de datos de Neptune, cualquier persona que acceda al clúster de base de datos debe autenticarse primero. Para obtener más información, consulte Habilitar la autenticación de bases de datos de IAM en Neptune para ver los pasos para habilitar la autenticación de IAM.

Cuando la autenticación de bases de datos de IAM está habilitada, cada una de las solicitudes debe firmarse con AWS Signature Version 4. Para saber cómo enviar solicitudes firmadas a todos los puntos finales de Neptune con la autenticación de IAM habilitada, consulte Conexión y firma con AWS firma, versión 4. Muchas bibliotecas y herramientas, como awscurl, ya admiten la versión 4 de Signature. AWS

Para interactuar con otros Servicios de AWS, Amazon Neptune utiliza funciones vinculadas a servicios de IAM. Un rol vinculado a un servicio es un tipo único de rol de IAM que está vinculado directamente a Neptune. Neptune predefine los roles vinculados al servicio e incluyen todos los permisos que el servicio requiere para llamar a otros Servicios de AWS en su nombre. Para obtener más información, consulte Uso de funciones vinculadas a servicios para Neptune.

Pilar de fiabilidad

El pilar de confiabilidad del Marco de AWS Trabajo de Buena Arquitectura abarca la capacidad de una carga de trabajo para realizar su función prevista de manera correcta y consistente 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.

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 afectarán al comportamiento de su carga de trabajo en todos los pilares de Well-Architected. Para la fiabilidad, debe seguir patrones específicos.

El pilar de confiabilidad se centra en las siguientes áreas clave:

- Arquitectura de carga de trabajo, incluidas las cuotas de servicio y los patrones de implementación
- Administración de cambios
- Administración de errores

Comprenda las cuotas de servicio de Neptune

El volumen de un clúster de Neptune puede crecer hasta un tamaño máximo de 128 tebibytes (TiB) en todos los Regiones de AWS casos admitidos, excepto en China GovCloud y, donde la cuota es de 64 TiB.

La cuota de 128 TiB es suficiente para almacenar aproximadamente entre 200 y 400 000 millones de objetos en el gráfico. En un gráfico de propiedades etiquetadas (LPG), un objeto es un nodo, una arista o una propiedad de un nodo o arista. En un gráfico del marco de descripción de recursos (RDF), un objeto es un cuadrilátero.

Para cualquier clúster Neptune Serverless, debe establecer el número mínimo y máximo de unidades de capacidad de Neptune (). NCUs Cada NCU consta de 2 gibibytes (GiB) de memoria y la vCPU y la red asociadas. Los valores mínimo y máximo de la NCU se aplican a todas las instancias sin servidor del clúster. El valor máximo de NCU más alto que puede establecer es 128,0 NCUs y el mínimo más bajo es 1,0. NCUs Optimice el rango de NCU que mejor se adapte `NCUUtilization` a su aplicación observando las CloudWatch métricas de Amazon `ServerlessDatabaseCapacity` y capturando el rango en el que se encuentra habitualmente y correlacionando el comportamiento

no deseado o los costos dentro de ese rango. Si descubre que su carga de trabajo no se amplía lo suficientemente rápido, aumente el mínimo NCUs para proporcionar suficiente procesamiento para el aumento inicial mientras se amplía.

Cada una de ellas Cuenta de AWS tiene cuotas para cada región en cuanto a la cantidad de recursos de base de datos que puede crear. Estos recursos incluyen las instancias y clústeres de base de datos. Después de que alcance el límite de un recurso, las llamadas adicionales para crear ese recurso dejan de funcionar con una excepción. Algunas cuotas son cuotas flexibles que se pueden aumentar si se solicita. Para obtener una lista de las cuotas compartidas entre Amazon Neptune y Amazon RDS, Amazon Aurora y Amazon DocumentDB (con compatibilidad con MongoDB), junto con enlaces para solicitar aumentos de cuota cuando estén disponibles, consulte Cuotas en Amazon RDS.

Comprenda los patrones de despliegue de Neptune

En los clústeres de base de datos de Neptune, hay una instancia de base de datos principal y hasta 15 réplicas de Neptune. La instancia de base de datos principal admite operaciones de lectura y escritura y realiza todas las modificaciones de datos en el volumen del clúster. Las réplicas de Neptune se conectan al mismo volumen de almacenamiento que la instancia de base de datos principal y solo admiten operaciones de lectura. Las réplicas de Neptune pueden descargar las cargas de trabajo de lectura de la instancia de base de datos principal.

Para lograr una alta disponibilidad, utilice réplicas de lectura. Tener una o más instancias de réplica de lectura disponibles en diferentes zonas de disponibilidad puede aumentar la disponibilidad, ya que las réplicas de lectura sirven como destinos de conmutación por error para la instancia principal. Si la instancia de escritura falla, Neptune convierte una instancia de réplica de lectura en la instancia principal. Cuando esto ocurre, se produce una breve interrupción (generalmente de menos de 30 segundos) mientras se reinicia la instancia promocionada, durante la cual las solicitudes de lectura y escritura realizadas a la instancia principal fallan con una excepción. Para obtener la máxima fiabilidad, considere la posibilidad de utilizar dos réplicas de lectura en distintas zonas de disponibilidad. Si la instancia principal de la zona de disponibilidad 1 se desconecta, la instancia de la zona de disponibilidad 2 pasa a ser principal, pero no podrá gestionar las consultas mientras eso suceda. Por lo tanto, se necesita una instancia en la zona de disponibilidad 3 para gestionar las consultas de lectura durante la transición.

Si utiliza Neptune Serverless, las instancias de lectura y escritura de todas las zonas de disponibilidad se ampliarán y reducirán, independientemente unas de otras, en función de la carga de la base de datos. Puede establecer el nivel de promoción de una instancia de lectura en 0 o 1

para que se amplíe o disminuya según la capacidad de la instancia de escritura. Esto la prepara para asumir la carga de trabajo actual en cualquier momento.

Gestione y escale los clústeres de Neptune

Puede usar el autoscalamiento de Neptune para ajustar automáticamente el número de réplicas de Neptuno en un clúster de base de datos para cumplir sus requisitos de conectividad y carga de trabajo en función de los umbrales de uso de la CPU. Con el autoscalamiento, su clúster de base de datos Neptune puede gestionar los aumentos repentinos de la carga de trabajo. Cuando la carga de trabajo disminuye, el autoscaling elimina las réplicas innecesarias para no tener que pagar por la capacidad no utilizada. Tenga en cuenta que el inicio de una nueva instancia puede tardar hasta 15 minutos, por lo que el autoscalamiento por sí solo no es una solución suficiente para los cambios rápidos de la demanda.

Puede usar el autoscalamiento solo con un clúster de base de datos de Neptune que ya tenga una instancia de escritura principal y al menos una instancia de réplica de lectura (consulte Instancias y clústeres de bases de datos de Amazon Neptune). Además, todas las instancias de réplica de lectura del clúster deben estar en un estado disponible. Si alguna réplica de lectura está en un estado diferente al disponible, el autoscalamiento de Neptune no hace nada hasta que todas las réplicas de lectura del clúster estén disponibles.

Si experimenta cambios rápidos en la demanda, considere la posibilidad de utilizar instancias sin servidor. Las instancias sin servidor se pueden escalar verticalmente durante períodos cortos, mientras que el autoescalado se escala horizontalmente durante períodos más largos. Esta configuración proporciona una escalabilidad óptima porque las instancias sin servidor se escalan verticalmente, mientras que el autoscalamiento crea instancias de nuevas réplicas de lectura para gestionar la carga de trabajo más allá de la capacidad máxima de una sola instancia sin servidor. Para obtener más información sobre el escalado de capacidad de Amazon Neptune Serverless, consulte Escalado de capacidad en un clúster de base de datos de Neptune Serverless.

Si sus necesidades de escalado cambian en momentos predecibles, puede programar cambios en las instancias mínimas, máximas y umbrales para gestionar mejor esas necesidades cambiantes. Recuerde programar los eventos de escalamiento horizontal con al menos 15 minutos de antelación para permitir que esas instancias se conecten cuando sea necesario.

Use los parámetros de un grupo de parámetros para administrar la configuración de la base de datos en Amazon Neptune. Los grupos de parámetros sirven de contenedor para los valores de configuración del motor que se aplican a una o varias instancias de bases de datos. Al modificar

los parámetros del clúster en grupos de parámetros, comprenda la diferencia entre los parámetros estáticos y dinámicos, y cómo y cuándo se aplican. Utilice el punto final de estado para ver la configuración aplicada actualmente.

Gestione las copias de seguridad y los eventos de conmutación por error

Neptune realiza automáticamente una copia de seguridad del volumen del clúster y conserva los datos de la copia de seguridad durante el período de retención de la copia de seguridad. Las copias de seguridad de Neptune son continuas y progresivas para que se puedan restaurar con rapidez a cualquier punto durante el periodo de retención de copia de seguridad. Puede especificar un período de retención de la copia de seguridad de 1 a 35 días al crear o modificar un clúster de base de datos.

Para conservar una copia de seguridad más allá del período de retención de la copia de seguridad, también puede tomar una instantánea de los datos del volumen de su clúster. Al almacenar instantáneas, se generan los cargos de almacenamiento estándar para Neptune.

Cuando crea una instantánea de Amazon Neptune de un clúster de base de datos, Neptune crea una instantánea del volumen de almacenamiento del clúster y hace copias de seguridad de todos sus datos, no solo de instancias individuales. Para crear posteriormente un clúster de base de datos nuevo, restaure esa instantánea de clúster de base de datos. Al restaurar el clúster de base de datos, proporciona el nombre de la instantánea del clúster de base de datos desde la que desea realizar la restauración y, a continuación, proporciona un nombre para el nuevo clúster de base de datos que se crea mediante la restauración.

Compruebe cómo responde el sistema a los eventos de conmutación por error. Usa la API de Neptune para forzar un evento de conmutación por error. Reiniciar con conmutación por error resulta útil cuando se quiere simular el fallo de una instancia de base de datos para realizar pruebas o restaurar operaciones en la zona de disponibilidad original tras producirse una conmutación por error. Para obtener más información, consulte Configuración y administración de una implementación Multi-AZ. Al reiniciar una instancia de escritura de bases de datos, se conmuta por error a la réplica en espera. El reinicio de una réplica de Neptune no inicia una conmutación por error.

Diseñe sus clientes para que sean fiables. Pruebe su comportamiento durante los eventos de conmutación por error. Implemente la lógica de reintento en su cliente con una lógica de retroceso exponencial. Los ejemplos de código que implementan esta lógica se encuentran en los ejemplos de AWS Lambda funciones de Amazon Neptune.

Considere la posibilidad de utilizarla AWS Backupsi tiene un conjunto común de requisitos de respaldo que se aplican a varios motores de bases de datos.

Pilar de eficiencia de rendimiento

El pilar de eficiencia del rendimiento del AWS Well-Architected Framework se centra en cómo optimizar el rendimiento al ingerir o consultar datos. La optimización del rendimiento es un proceso gradual y continuo que consiste en lo siguiente:

- Confirmar los requisitos empresariales
- Medir el rendimiento de la carga de trabajo
- Identificar los componentes de bajo rendimiento
- Ajustar los componentes para que se adapten a las necesidades de su empresa

El pilar de la eficiencia del rendimiento proporciona pautas específicas para cada caso de uso que pueden ayudar a identificar el modelo de datos gráfico y los lenguajes de consulta correctos que se deben utilizar. También incluye las prácticas recomendadas que se deben seguir al incorporar y consumir datos de Amazon Neptune.

El pilar de la eficiencia del desempeño se centra en las siguientes áreas clave:

- Modelado gráfico
- Optimización de las consultas
- Dimensionamiento correcto de clústeres
- Optimización de escritura

Comprenda el modelado de gráficos

Comprenda la diferencia entre los modelos Labeled Property Graph (LPG) y Resource Description Framework (RDF). En la mayoría de los casos, es una cuestión de preferencia. Sin embargo, hay varios casos de uso en los que un modelo es más adecuado que el otro. Si necesita conocer la ruta que conecta dos nodos de su gráfico, elija LPG. Si desea federar datos entre clústeres de Neptune u otros almacenes triples de gráficos, elija RDF.

Si está creando una aplicación de software como servicio (SaaS) o una aplicación que requiere varios inquilinos, considere la posibilidad de incorporar la separación lógica de los inquilinos en su modelo de datos en lugar de tener un inquilino para cada clúster. Para lograr ese tipo de diseño, puede utilizar gráficos con nombres y estrategias de etiquetado de SPARQL, como anteponer los identificadores de los clientes a las etiquetas o añadir pares clave-valor de la propiedad que

representen los identificadores de los inquilinos. Asegúrese de que su capa de clientes incorpore estos valores para mantener esa separación lógica.

El rendimiento de las consultas depende de la cantidad de objetos gráficos (nodos, bordes, propiedades) que deban evaluarse al procesar la consulta. Por lo tanto, el modelo gráfico puede tener un impacto significativo en el rendimiento de la aplicación. Utilice etiquetas granulares siempre que sea posible y almacene solo las propiedades que necesite para determinar la ruta o filtrar. Para lograr un mayor rendimiento, considere la posibilidad de calcular previamente partes del gráfico, como crear nodos de resumen o bordes más directos que conecten rutas comunes.

Intente evitar navegar por nodos que tengan un número anormalmente alto de aristas con la misma etiqueta. Estos nodos suelen tener miles de aristas (mientras que la mayoría de los nodos tienen un número de aristas de decenas). El resultado es una complejidad informática y de datos mucho mayor. Es posible que estos nodos no sean problemáticos en algunos patrones de consulta, pero recomendamos modelar los datos de forma diferente para evitarlos, especialmente si va a navegar por el nodo como paso intermedio. Puedes usar registros de consultas lentas para ayudar a identificar las consultas que navegan por estos nodos. Es probable que observes métricas de latencia y acceso a los datos mucho más altas que los patrones de consulta habituales, especialmente si utilizas el modo de depuración.

Utilice un nodo determinista IDs para los nodos y las aristas si su caso de uso lo admite en lugar de utilizar Neptune para asignar valores GUID aleatorios. IDs Acceder a los nodos por ID es el método más eficaz.

Optimización de consultas

Los lenguajes OpenCypher y Gremlin se pueden usar indistintamente en los modelos GLP. Si el rendimiento es una de las principales preocupaciones, considere la posibilidad de utilizar los dos lenguajes indistintamente, ya que uno podría funcionar mejor que el otro para patrones de consulta específicos.

Neptune está en proceso de conversión a su motor de consultas alternativo (DFE). OpenCypher solo se ejecuta en el DFE, pero las consultas de Gremlin y SPARQL se pueden configurar opcionalmente para que se ejecuten en el DFE mediante anotaciones de consulta. Considere la posibilidad de probar las consultas con el DFE activado y comparar el rendimiento del patrón de consulta cuando no utilice el DFE.

Neptune está optimizado para consultas de tipo transaccional que comienzan en un solo nodo o conjunto de nodos y se despliegan desde allí, en lugar de consultas analíticas que evalúan todo el

gráfico. Para sus cargas de trabajo de consultas analíticas, considere la posibilidad de utilizar el SDK de AWS para Pandas o utilizar neptune-export en combinación con Amazon EMR. AWS Glue

Para identificar las ineficiencias y los cuellos de botella en sus modelos y consultas, utilice el y `explain` APIs para cada lenguaje de consulta a fin de obtener `profile` explicaciones detalladas del plan de consulta y las métricas de consulta. Para obtener más información, consulte el perfil de Gremlin, la explicación de OpenCypher y la explicación de SPARQL.

Comprenda sus patrones de consulta. Si el número de bordes distintos de un gráfico aumenta, la estrategia de acceso a Neptune predeterminada puede resultar ineficiente. Las siguientes consultas pueden resultar bastante ineficientes:

- Consultas que se desplazan hacia atrás a través de los bordes cuando no se proporciona ninguna etiqueta de borde.
- Cláusulas que utilizan este mismo patrón internamente, como `.both()` en Gremlin, o cláusulas que eliminan nodos en cualquier idioma (lo que requiere eliminar los bordes entrantes sin conocer las etiquetas).
- Consultas que acceden a los valores de las propiedades sin especificar las etiquetas de las propiedades. Estas consultas pueden resultar bastante ineficientes. Si esto coincide con su patrón de uso, considere habilitar el índice OSGP (objeto, sujeto, gráfico, predicado).

Utilice el registro de consultas lentas para identificar las consultas lentas. La lentitud de las consultas puede deberse a planes de consultas no optimizados o a un número innecesariamente elevado de búsquedas en los índices, lo que puede aumentar los costes de E/S. Los puntos finales explicativos y perfilados de Neptune para Gremlin, SPARQL u OpenCypher pueden ayudarle a entender por qué estas consultas son lentas. Entre las causas se pueden incluir las siguientes:

- Los nodos con un número de aristas anormalmente alto en comparación con el nodo promedio del gráfico (por ejemplo, miles en lugar de decenas) pueden añadir complejidad computacional y, por lo tanto, una latencia más prolongada y un mayor consumo de recursos. Determine si estos nodos están modelados correctamente o si se pueden mejorar los patrones de acceso para reducir la cantidad de bordes que deben atravesarse.
- Las consultas no optimizadas contendrán una advertencia de que algunos pasos específicos no están optimizados. Reescribir estas consultas para usar pasos optimizados podría mejorar el rendimiento.
- Los filtros redundantes pueden provocar búsquedas de índices innecesarias. Del mismo modo, los patrones redundantes pueden provocar búsquedas de índices duplicadas que pueden optimizarse

mejorando la consulta (consulte `Index Operations - Duplication ratio` el resultado del perfil).

- Algunos lenguajes, como Gremlin, no tienen valores numéricos bien escritos y, en su lugar, utilizan la promoción tipográfica. Por ejemplo, si el valor es 55, Neptune busca valores enteros, largos, flotantes y otros tipos numéricos equivalentes a 55. Esto da como resultado operaciones adicionales. Si sabe de antemano que sus tipos coinciden, puede evitarlo utilizando una sugerencia de consulta.
- Su modelo gráfico puede tener un gran impacto en el rendimiento. Considere la posibilidad de reducir la cantidad de objetos que deben evaluarse utilizando etiquetas más granulares o calculando previamente los atajos para las rutas lineales de saltos múltiples.

Si la optimización de consultas por sí sola no le permite alcanzar sus requisitos de rendimiento, considere la posibilidad de utilizar diversas técnicas de almacenamiento en caché con Neptune para cumplir esos requisitos.

Clústeres del tamaño correcto

Ajuste el tamaño del clúster a sus requisitos de simultaneidad y rendimiento. El número de consultas simultáneas que puede gestionar cada instancia del clúster es igual a dos veces el número de consultas virtuales CPUs (vCPUs) de esa instancia. Las consultas adicionales que llegan mientras todos los subprocesos de trabajo están ocupados se colocan en una cola del lado del servidor. Estas consultas se gestionan mediante FIFO cuando los subprocesos de trabajo están disponibles. first-in-first-out La CloudWatch métrica de `MainRequestQueuePendingRequests` Amazon muestra la profundidad de cola actual de cada instancia. Si este valor suele estar por encima de cero, considera la posibilidad de elegir una instancia con más vCPUs. Si la profundidad de la cola supera los 8.192, Neptune devolverá un error. `ThrottlingException`

Aproximadamente el 65 por ciento de la RAM de cada instancia se reserva para la memoria caché del búfer. La memoria caché del búfer contiene el conjunto de datos de trabajo (no todo el gráfico, solo los datos que se están consultando). Para determinar qué porcentaje de datos se obtiene de la caché del búfer en lugar del almacenamiento, supervise la CloudWatch métrica. `BufferCacheHitRatio` Si esta métrica suele caer por debajo del 99,9 por ciento, considere la posibilidad de probar una instancia con más memoria para determinar si reduce la latencia y los costes de E/S.

Las réplicas de lectura no tienen que tener el mismo tamaño que la instancia de grabación. Sin embargo, las cargas de trabajo de escritura pesadas pueden provocar que las réplicas más

pequeñas se retrasen y se reinicien porque no pueden seguir el ritmo de la replicación. Por lo tanto, recomendamos hacer réplicas iguales o mayores que la instancia de grabación.

Cuando utilices el autoescalado para tus réplicas de lectura, recuerda que poner una nueva réplica de lectura en línea puede tardar hasta 15 minutos. Cuando el tráfico de clientes aumente de forma rápida pero predecible, considere la posibilidad de utilizar el escalado programado para establecer un número mínimo de réplicas de lectura más alto para tener en cuenta ese tiempo de inicialización.

Las instancias sin servidor admiten varios casos de uso y cargas de trabajo diferentes. Considere la posibilidad de utilizar instancias sin servidor en lugar de aprovisionadas en los siguientes escenarios:

- Su carga de trabajo fluctúa con frecuencia a lo largo del día.
- Ha creado una nueva aplicación y no está seguro del tamaño de la carga de trabajo.
- Está realizando el desarrollo y las pruebas.

Es importante tener en cuenta que las instancias sin servidor son más caras que las instancias aprovisionadas equivalentes en términos de dólar por GB de RAM. Cada instancia sin servidor consta de 2 GB de RAM junto con la vCPU y la red asociadas. Realice un análisis de costos entre sus opciones para evitar facturas inesperadas. En general, solo podrá ahorrar costes con la tecnología sin servidor si su carga de trabajo es muy intensa durante unas pocas horas al día y prácticamente nula durante el resto del día o si su carga de trabajo fluctúa considerablemente a lo largo del día.

Optimice las escrituras

Para optimizar las escrituras, ten en cuenta lo siguiente:

- El cargador masivo Neptune es la forma óptima de cargar inicialmente la base de datos o añadirla a los datos existentes. El cargador Neptune no es transaccional y no puede eliminar datos, así que no lo utilice si estos son sus requisitos.
- Las actualizaciones transaccionales se pueden realizar mediante los lenguajes de consulta compatibles. Para optimizar las operaciones de E/S de escritura, escriba los datos en lotes de 50 a 100 objetos por confirmación. Un objeto es un nodo, una arista o una propiedad en un nodo o una arista en LPG, o un almacén triple o un cuadrilátero en RDF.
- Todas las operaciones de escritura de Neptune son de un solo hilo para cada conexión. Al enviar una gran cantidad de datos a Neptune, considere la posibilidad de tener varias conexiones paralelas, cada una de las cuales escriba datos. Cuando eliges una instancia aprovisionada

por Neptune, el tamaño de la instancia se asocia a un número de v. CPUs Neptune crea dos subprocesos de base de datos para cada vCPU de la instancia, así que comience con el doble de v CPUs cuando pruebe la paralelización óptima. Las instancias sin servidor escalan el número de v CPUs a una tasa de aproximadamente uno por cada 4. NCUs

- Planifique y ConcurrentModificationExceptionsgestione de manera eficiente todos los procesos de escritura, incluso si solo una conexión escribe datos en cualquier momento. Diseñe sus clientes para que sean fiables cuando `ConcurrentModificationExceptions` se produzcan.
- Si desea eliminar todos sus datos, considere la posibilidad de utilizar la API de restablecimiento rápido en lugar de emitir consultas de eliminación simultáneas. Esta última llevará mucho más tiempo e incurrirá en un costo de E/S sustancial en comparación con la primera.
- Si desea eliminar la mayoría de los datos, considere la posibilidad de exportar los datos que desee conservar mediante neptune-export para cargar los datos en un nuevo clúster. A continuación, elimine el clúster original.

Pilar de optimización de costos

El pilar de optimización de costes del AWS Well-Architected Framework se centra en evitar costes innecesarios. Las siguientes recomendaciones pueden ayudarle a cumplir los principios de diseño de optimización de costes y las prácticas recomendadas de arquitectura de Amazon Neptune.

El pilar de la optimización de costos se centra en las siguientes áreas clave:

- Comprender el gasto a lo largo del tiempo y controlar la asignación de fondos
- Seleccionar los recursos del tipo y la cantidad correctos
- Escalar para satisfacer las necesidades empresariales sin gastar de más

Comprenda los patrones de uso y los servicios necesarios

Neptune es una buena opción para su carga de trabajo si su modelo de datos tiene una estructura gráfica discernible y sus consultas necesitan explorar relaciones y recorrer varios saltos. Una base de datos de gráficos no es adecuada para los siguientes patrones:

- Principalmente consultas de un solo salto (considera si tus datos podrían representarse mejor como atributos de un objeto)
- Datos JSON o BLOB almacenados como propiedades
- Consultas que se agregan en un conjunto de datos, como calcular la suma de una propiedad numérica en un gran número de nodos

Considere si el uso conjunto de varias bases de datos diseñadas específicamente para patrones de acceso específicos podría satisfacer todas sus necesidades. Por ejemplo:

- Una API que requiera navegaciones gráficas complejas menos frecuentes junto con una recuperación altamente simultánea de las propiedades de un único nodo podría presentarse mejor utilizando uno o más de Neptune, DynamoDB o Amazon DocumentDB.
- Las bases de datos relacionales pueden coexistir con Neptuno para mantener su funcionalidad actual, pero utilice Neptune solo para recorridos de saltos múltiples que no funcionen ni escalen bien en bases de datos relacionales.

Comprenda los costos asociados a los servicios que interactúan con Neptune y lo complementan, incluidos los siguientes:

- Costos de almacenamiento del Amazon Simple Storage Service (Amazon S3) para los archivos de datos que se cargan de forma masiva en Neptune
- Funciones Lambda utilizadas para consultas de inserción o inserción, consultas de lectura y procesamiento de flujos de Neptune
- La capa de API creada en Neptune para interactuar con la aplicación cliente (en lugar de tener conexiones directas a la base de datos) en o AWS AppSync
- AWS Glue trabajos utilizados para transferir datos hacia y desde Neptuno
- Instancias de Amazon Kinesis o Amazon Managed Streaming for Apache Kafka (Amazon MSK) que reciben datos de streaming para incorporarlos casi en tiempo real a Neptune.
- AWS Database Migration Service para la migración de datos relacionales a Neptune
- Costos SageMaker de Amazon Runtime para los modelos de aprendizaje automático de Jupyter notebooks y Deep Graph Library

Seleccione los recursos prestando atención al costo

Los precios de Neptune se basan en el costo por hora de la instancia (o las unidades de cómputo de Neptune consumidas si no hay servidor), la E/S de datos y el uso del almacenamiento. Las instancias representan, en promedio, el 85 por ciento del costo total, por lo que el tamaño adecuado puede tener importantes implicaciones en los costos. La mejor manera de ajustar el tamaño de las instancias es probar el rendimiento de las aplicaciones en una variedad de instancias y comparar los siguientes factores:

- ¿La `MainRequestQueuePendingRequests` CloudWatch métrica se mantiene en un número consistentemente bajo, cercano a cero?
- ¿La `BufferCacheHitRatio` CloudWatch métrica se mantiene igual o superior al 99,9 por ciento la mayoría de las veces?
- ¿Cuáles son las curvas de costo y rendimiento, por ejemplo, los costos y los costos de E/S de datos asociados? Los costes de lectura de datos pueden aumentar considerablemente si se trata de una instancia de tamaño insuficiente que requiera el intercambio frecuente de la memoria caché del búfer por la del almacenamiento. `BufferCacheHitRatio`disminuirán con frecuencia en estos escenarios.

Los costos de las instancias se escalan linealmente con el tamaño dentro de la misma familia de instancias. El costo por hora de la `db.r6i.2xlarge` instancia es el doble que el de la `db.r6i.xlarge` instancia y también tiene el doble de la asignación de recursos. La `db.r6i.24xlarge` instancia es 24 veces el costo por hora de la `db.r6i.xlarge` instancia.

Calcule el número de consultas simultáneas que debe admitir. Puede tener entre cero y quince réplicas de lectura para procesar las consultas de solo lectura. Si sus requisitos varían según la hora del día, la semana o el mes, puede usar varias instancias más pequeñas para escalar según un cronograma. Cada vCPU de una instancia proporciona dos subprocesos para gestionar consultas simultáneas. Tres réplicas de `db.r6i.xlarge` lectura, con 4 vCPU cada una, pueden gestionar 24 consultas simultáneas.

Si, por el contrario, el volumen de tráfico se mide en consultas por segundo (QPS), debe experimentar para determinar la latencia media de las consultas. El número de consultas por segundo que admite un clúster de Neptune es igual a. `vCPU × 2 × (1 second/average query latency)` Por ejemplo, si tiene 4 vCPU y una latencia de consulta de 100 milisegundos (0,1 segundos),. `QPS = 4 × 2 × (1s/0.1s) = 80 queries per second`

Las instancias aprovisionadas son más baratas que las sin servidor para cargas de trabajo continuas, estables y predecibles. La tecnología sin servidor ofrece oportunidades para optimizar los costes cuando se tiene una carga de trabajo que requiere un uso muy elevado durante unas pocas horas al día (por ejemplo`db.r6i.4xlarge`) y, después, prácticamente no hay tráfico durante el resto del día (por ejemplo, 1 unidad de cómputo de Neptune). Una instancia sin servidor que se amplíe durante unas horas y luego se vuelva a reducir será más económica que usar una instancia aprovisionada `db.r6i.4xlarge` todo el día.

Elija la mejor configuración de instancias de Neptune para su carga de trabajo

Para experimentar con Neptune a nivel básico, puede utilizar la capa gratuita de AWS. Las 750 horas gratuitas de `db.t3.medium` uso de una `db.t4g.medium` instancia son suficientes para que comprenda bien Neptuno a baja escala. El clúster se conservará una vez finalizado el período de prueba gratuito, aunque a partir de ese momento se te cobrará por su uso.

Las `db.t4g.medium` instancias `db.t3.medium` y son adecuadas para entornos de desarrollo de bajo coste, pero tenga en cuenta que tienen una relación RAM a vCPU menor (2:1) que las instancias de la familia R (8:1) o las instancias de la familia X (16:1). Los perfiles de rendimiento pueden diferir de los de esas clases, especialmente cuando las consultas se desplazan por una parte

importante del gráfico `OutOfMemoryExceptions` y cuando lo hacen. Para determinar si esta última condición podría verse afectada, compruebe la `BufferCacheHitRatio` CloudWatch métrica.

Recomendamos encarecidamente no realizar ninguna prueba de rendimiento o carga con instancias de la familia T, ya que es posible que se obtengan resultados incoherentes que no sean indicativos de un entorno de producción.

Las instancias aprovisionadas ofrecen la mejor combinación de coste y rendimiento cuando la carga de trabajo es bastante estable y predecible. Elija el tamaño de la instancia en función de la simultaneidad de solicitudes requerida y de la complejidad de la consulta. Una mayor simultaneidad requiere más v. CPUs Una mayor complejidad de consulta requiere más RAM. Utilice la `MainRequestQueuePendingRequests` CloudWatch métrica para determinar el impacto de la primera (un valor superior a cero representa más solicitudes simultáneas de las que se pueden gestionar). Usa la `BufferCacheHitRatio` CloudWatch métrica para determinar el impacto de la segunda. Una proporción que suele caer por debajo del 99,9 por ciento indica que no hay suficiente RAM para contener la parte funcional del gráfico que se está evaluando, lo que provoca un intercambio de caché más frecuente. Si la familia de instancias R proporciona suficiente simultaneidad pero no suficiente RAM, considere la posibilidad de probar la familia de instancias X.

Los casos de uso ideales para las instancias sin servidor se describen en la documentación de Neptune. Si no está seguro de si la solución aprovisionada o sin servidor es lo mejor para usted y su principal preocupación es el costo, pruebe su carga de trabajo en la modalidad sin servidor para determinar la cantidad de unidades NCUs utilizadas y compare el costo de las soluciones aprovisionadas () con las aplicaciones sin servidor ()`N hours × hourly provisioned cost`. `sum of NCUs × hourly cost per NCU` Si no está seguro de cuál es la instancia de aprovisionamiento de tamaño equivalente, una NCU equivale aproximadamente a 2 GB de RAM y la vCPU y la red asociadas. Si la instancia aprovisionada pertenece a la familia R6i, la proporción es de 1 vCPU por cada 8 GB de RAM, o 4 NCUs, junto con la red asociada.

Cuando utilices una instancia sin servidor para las instancias principales y de réplica, recuerda que las réplicas de lectura de los niveles de promoción 0 y 1 se escalarán NCUs en función de la instancia de grabación para que se escalen correctamente en caso de que se produzca una conmutación por error. Establezca los límites de la NCU para estas instancias en función de las instancias (grabadoras o lectoras) que reciban más tráfico.

En entornos en los que no se necesite el clúster las 24 horas del día, los 7 días de la semana, considere la posibilidad de escribir scripts que apaguen las instancias de Neptune cuando no estén en uso y las vuelvan a iniciar antes de utilizarlas. Las instancias de Neptune se reiniciarán automáticamente cada 7 días para garantizar que se apliquen las actualizaciones de mantenimiento

necesarias. Si piensa dejar las instancias apagadas durante períodos prolongados, utilice un script semanal para volver a cerrarlas.

Almacenamiento y transferencia de datos del tamaño adecuado

Las consultas más eficientes (por ejemplo, las consultas que necesitan tocar menos nodos, bordes y propiedades del gráfico) requieren menos transferencia de E/S y, potencialmente, pueden utilizar instancias más pequeñas porque se requiere menos memoria caché de búfer. Utilice el perfil o explique los puntos finales del lenguaje de consulta para optimizar la consulta y considere la posibilidad de optimizar el modelo gráfico para el rendimiento de la consulta.

Neptune usa la codificación de diccionarios en cadenas grandes, y ese diccionario está optimizado para el rendimiento, no para la eficiencia. Si sus propiedades tienen cadenas JSON grandes BLOBs o que cambian con frecuencia, considere la posibilidad de guardarlas fuera de Neptuno en Amazon S3, Amazon DynamoDB o Amazon DocumentDB, y almacenar solo una referencia dentro del nodo Neptune.

En algunos casos, elegir un tamaño de instancia más grande puede resultar más económico. Si los costes de E/S son muy altos debido a que son bajos`BufferCacheHitRatio`, es posible que una caché de búfer más grande reduzca considerablemente ese coste. Esto se debe a que todos los datos cabrían en la memoria caché en lugar de ser intercambiados frecuentemente del almacenamiento e incurrir en la velocidad de transferencia de E/S.

Neptune utiliza copy-on-write la clonación. Al clonar para dividir un gráfico en varios fragmentos, puede ser más eficaz no eliminar los datos no deseados del clúster clonado, ya que ello implicará la creación de nuevas páginas de datos, lo que aumentará los costes de almacenamiento. Los datos que no hayan cambiado desde antes del evento de clonación existirán en una sola página de datos compartida entre los dos clústeres y solo se cobrará por esa única copia.

No habilite el índice OSGP ni utilice instancias R5d a menos que haya realizado pruebas para confirmar que suponen una diferencia sustancial en su carga de trabajo. Ambas están diseñadas para situaciones poco frecuentes y pueden aumentar los costes y obtener beneficios mínimos o nulos.

Pilar de sostenibilidad

El pilar de sostenibilidad del AWS Well-Architected Framework se centra en minimizar los impactos ambientales de la ejecución de cargas de trabajo en la nube. Los temas clave incluyen un modelo de responsabilidad compartida para la sostenibilidad, comprender el impacto y maximizar el uso para minimizar los recursos necesarios y reducir los impactos posteriores.

El pilar de la sostenibilidad contiene las siguientes áreas de enfoque clave:

- ¿Su impacto
- Objetivos de sostenibilidad
- Uso maximizado
- Anticipar y adoptar ofertas de hardware y software nuevas y más eficientes
- Uso de servicios gestionados
- Reducción del impacto descendente

Esta guía se centra en su impacto. Para obtener más información sobre los demás principios de diseño de sostenibilidad, consulte el AWS Well-Architected Framework.

Sus elecciones y requisitos tienen un impacto en el medio ambiente. Si puede elegir Regiones de AWS que tengan una menor intensidad de carbono y si sus requisitos reflejan las necesidades reales de carga de trabajo en lugar de limitarse a maximizar el tiempo de actividad y la durabilidad, la sostenibilidad de la carga de trabajo aumenta. En las siguientes secciones, se analizan las mejores prácticas y las consideraciones bien pensadas que, si se adoptan en el diseño de la carga de trabajo y en las operaciones en curso, tendrán un impacto medioambiental positivo.

Selección de regiones de AWS

Algunos Regiones de AWS están cerca de los proyectos de energía renovable de Amazon o ubicados donde la red tiene una intensidad de carbono publicada inferior a la de otros. Considera el impacto en la sostenibilidad de las regiones que podrían ser viables para tu carga de trabajo y compara tu lista con las regiones en las que Neptune está disponible.

El consumo se basa en los patrones de comportamiento de los usuarios

Ajustar el consumo al tráfico y al comportamiento de los usuarios ayuda a AWS minimizar el impacto de los servicios en el medio ambiente. Tenga en cuenta las siguientes prácticas recomendadas al diseñar la solución:

- Supervise CloudWatch las métricas de Amazon `CPUUtilization` `MainRequestQueuePendingRequests`, por ejemplo, y `TotalRequestsPerSec` determine cuándo su demanda es mayor o menor, y asegúrese de que los recursos de su clúster tengan el tamaño adecuado en esos momentos.
- Automatice la interrupción de los entornos que no son de producción durante las horas en que no se utilizan. Para obtener más información, consulte la entrada del blog Automatice la detención y el inicio de los recursos del entorno de Amazon Neptune mediante etiquetas de recursos.
- Si sus patrones de tráfico varían con frecuencia y de forma impredecible, considere la posibilidad de utilizar instancias Neptune Serverless que se amplíen o disminuyan según la demanda, en lugar de utilizar una instancia aprovisionada para los picos de tráfico.
- Considere la posibilidad de alinear sus acuerdos de nivel de servicio con los objetivos de sostenibilidad, además de con los objetivos de continuidad empresarial. Reducir los requisitos, como la recuperación ante desastres en varias regiones, la alta disponibilidad o la retención de copias de seguridad a largo plazo, especialmente para los entornos que no son de producción o las cargas de trabajo que no son esenciales para la misión, puede reducir la cantidad de recursos necesarios para cumplir esos objetivos.

Optimice los patrones de arquitectura y desarrollo de software

Para evitar el desperdicio, optimice sus modelos y consultas, y comparta los recursos de cómputo para utilizar todos los recursos disponibles en las instancias y los clústeres de Neptune. Entre las prácticas recomendadas específicas se incluyen las siguientes:

- Haga que los desarrolladores compartan las instancias de Neptune y las instancias de aplicación de Jupyter Notebook en lugar de que cada uno cree las suyas propias. Proporcione a cada desarrollador su propia partición lógica en un único clúster de Neptune mediante el uso de estrategias de particionamiento multiusuario y cree carpetas de bloc de notas independientes para cada desarrollador en una sola instancia de Jupyter.

- Implemente patrones que maximicen el uso de los recursos y minimicen el tiempo de inactividad, como subprocesos paralelos para cargar datos y agrupar registros en lotes en una transacción más grande.
- Optimice sus consultas y su modelo gráfico para minimizar los recursos necesarios para calcular los resultados.
- Para los resultados de las consultas de Gremlin, utilice la función de caché de resultados para minimizar los recursos que se gastan en volver a calcular las consultas paginadas o que se repiten con frecuencia.
- Mantenga sus entornos de Neptune actualizados. Las versiones más recientes de Neptune admiten las EC2 instancias más recientes, como Graviton, que son más eficientes. También incluyen mejoras en la optimización de las consultas y correcciones de errores que reducen la cantidad de recursos necesarios para calcular las consultas.

Recursos

Referencias

- AWS Well-Architected
- AWS Documentación de Well-Architected Framework
- Aplicación del marco AWS Well-Architected para Amazon Neptune Analytics
- Últimas actualizaciones de Amazon Neptune
- Mejores prácticas: sacar el máximo provecho de Neptuno

Publicaciones de blog

- Pruebas automatizadas del acceso a los datos de Amazon Neptune con Apache Gremlin TinkerPop
- Automatice la detención y el inicio de los recursos del entorno Amazon Neptune mediante etiquetas de recursos
- Control de acceso detallado para las acciones del plano de datos de Amazon Neptune
- Utilice el razonamiento semántico para deducir nuevos hechos a partir de su gráfico RDF mediante la integración con Amazon Neptune RDFox
- Cree una solución de detección de fraudes en tiempo real con Amazon Neptune ML

Cursos gratuitos de AWS Skill Builder

- Introducción a Amazon Neptune
- Creación de aplicaciones en Amazon Neptune
- Modelado de datos para Amazon Neptune

Colaboradores

Entre los colaboradores de esta guía se encuentran:

- Brian O'Keefe, arquitecto principal de soluciones de Neptune, AWS
- Abhishek Mishra, arquitecto sénior de soluciones para Neptune, AWS
- Ganesh Sawhney, jefe de equipo y arquitecto de Strategic Partner Success Solutions, AWS
- Michael Havey, arquitecto sénior de soluciones de Neptune, AWS
- Kevin Phillips, arquitecto de soluciones de Neptune, AWS
- Melissa Kwok, arquitecta de soluciones para Neptune, AWS
- Sakti Mishra, arquitecta principal de soluciones AWS
- Javed Ali, arquitecto sénior de soluciones, AWS

Historial de documentos

En la siguiente tabla, se describen cambios significativos de esta guía. Si quiere recibir notificaciones de futuras actualizaciones, puede suscribirse a las notificaciones RSS.

Cambio	Descripción	Fecha
Publicación inicial	—	27 de septiembre de 2023

AWS Glosario de orientación prescriptiva

Los siguientes son términos de uso común en las estrategias, guías y patrones proporcionados por la Guía AWS prescriptiva. Para sugerir entradas, utilice el enlace Enviar comentarios al final del glosario.

Números

Las 7 R

Siete estrategias de migración comunes para trasladar aplicaciones a la nube. Estas estrategias se basan en las 5 R que Gartner identificó en 2011 y consisten en lo siguiente:

- Refactorizar/rediseñar: traslade una aplicación y modifique su arquitectura mediante el máximo aprovechamiento de las características nativas en la nube para mejorar la agilidad, el rendimiento y la escalabilidad. Por lo general, esto implica trasladar el sistema operativo y la base de datos. Ejemplo: migre su base de datos Oracle local a la edición compatible con PostgreSQL de Amazon Aurora.
- Redefinir la plataforma (transportar y redefinir): traslade una aplicación a la nube e introduzca algún nivel de optimización para aprovechar las capacidades de la nube. Ejemplo: migre su base de datos Oracle local a Amazon Relational Database Service (Amazon RDS) para Oracle en el. Nube de AWS
- Recomprar (readquirir): cambie a un producto diferente, lo cual se suele llevar a cabo al pasar de una licencia tradicional a un modelo SaaS. Ejemplo: migre su sistema de gestión de relaciones con los clientes (CRM) a Salesforce.com.
- Volver a alojar (migrar mediante lift-and-shift): traslade una aplicación a la nube sin realizar cambios para aprovechar las capacidades de la nube. Ejemplo: migre su base de datos Oracle local a Oracle en una EC2 instancia del. Nube de AWS
- Reubicar: (migrar el hipervisor mediante lift and shift): traslade la infraestructura a la nube sin comprar equipo nuevo, reescribir aplicaciones o modificar las operaciones actuales. Los servidores se migran de una plataforma local a un servicio en la nube para la misma plataforma. Ejemplo: migrar una Microsoft Hyper-V aplicación a AWS.
- Retener (revisitar): conserve las aplicaciones en el entorno de origen. Estas pueden incluir las aplicaciones que requieren una refactorización importante, que desee posponer para más adelante, y las aplicaciones heredadas que desee retener, ya que no hay ninguna justificación empresarial para migrarlas.

- Retirar: retire o elimine las aplicaciones que ya no sean necesarias en un entorno de origen.

A

ABAC

Consulte control de acceso basado en atributos.

servicios abstractos

Consulte servicios gestionados.

ACID

Consulte atomicidad, consistencia, aislamiento y durabilidad.

migración activa-activa

Método de migración de bases de datos en el que las bases de datos de origen y destino se mantienen sincronizadas (mediante una herramienta de replicación bidireccional o mediante operaciones de escritura doble) y ambas bases de datos gestionan las transacciones de las aplicaciones conectadas durante la migración. Este método permite la migración en lotes pequeños y controlados, en lugar de requerir una transición única. Es más flexible, pero requiere más trabajo que la migración activa-pasiva.

migración activa-pasiva

Método de migración de bases de datos en el que las bases de datos de origen y destino se mantienen sincronizadas, pero solo la base de datos de origen gestiona las transacciones de las aplicaciones conectadas, mientras los datos se replican en la base de datos de destino. La base de datos de destino no acepta ninguna transacción durante la migración.

función agregada

Función SQL que opera en un grupo de filas y calcula un único valor de retorno para el grupo. Algunos ejemplos de funciones agregadas incluyen `SUM` y`MAX`.

IA

Véase inteligencia artificial.

AIOps

Consulte las operaciones de inteligencia artificial.

anonimización

El proceso de eliminar permanentemente la información personal de un conjunto de datos. La anonimización puede ayudar a proteger la privacidad personal. Los datos anonimizados ya no se consideran datos personales.

antipatrones

Una solución que se utiliza con frecuencia para un problema recurrente en el que la solución es contraproducente, ineficaz o menos eficaz que una alternativa.

control de aplicaciones

Un enfoque de seguridad que permite el uso únicamente de aplicaciones aprobadas para ayudar a proteger un sistema contra el malware.

cartera de aplicaciones

Recopilación de información detallada sobre cada aplicación que utiliza una organización, incluido el costo de creación y mantenimiento de la aplicación y su valor empresarial. Esta información es clave para el proceso de detección y análisis de la cartera y ayuda a identificar y priorizar las aplicaciones que se van a migrar, modernizar y optimizar.

inteligencia artificial (IA)

El campo de la informática que se dedica al uso de tecnologías informáticas para realizar funciones cognitivas que suelen estar asociadas a los seres humanos, como el aprendizaje, la resolución de problemas y el reconocimiento de patrones. Para más información, consulte ¿Qué es la inteligencia artificial?

operaciones de inteligencia artificial (AIOps)

El proceso de utilizar técnicas de machine learning para resolver problemas operativos, reducir los incidentes operativos y la intervención humana, y mejorar la calidad del servicio. Para obtener más información sobre cómo AIOps se utiliza en la estrategia de AWS migración, consulte la guía de integración de operaciones.

cifrado asimétrico

Algoritmo de cifrado que utiliza un par de claves, una clave pública para el cifrado y una clave privada para el descifrado. Puede compartir la clave pública porque no se utiliza para el descifrado, pero el acceso a la clave privada debe estar sumamente restringido.

atomicidad, consistencia, aislamiento, durabilidad (ACID)

Conjunto de propiedades de software que garantizan la validez de los datos y la fiabilidad operativa de una base de datos, incluso en caso de errores, cortes de energía u otros problemas.

control de acceso basado en atributos (ABAC)

La práctica de crear permisos detallados basados en los atributos del usuario, como el departamento, el puesto de trabajo y el nombre del equipo. Para obtener más información, consulte ABAC AWS en la documentación AWS Identity and Access Management (IAM).

origen de datos fidedigno

Ubicación en la que se almacena la versión principal de los datos, que se considera la fuente de información más fiable. Puede copiar los datos del origen de datos autorizado a otras ubicaciones con el fin de procesarlos o modificarlos, por ejemplo, anonimizarlos, redactarlos o seudonimizarlos.

Zona de disponibilidad

Una ubicación distinta dentro de una Región de AWS que está aislada de los fallos en otras zonas de disponibilidad y que proporciona una conectividad de red económica y de baja latencia a otras zonas de disponibilidad de la misma región.

AWS Marco de adopción de la nube (AWS CAF)

Un marco de directrices y mejores prácticas AWS para ayudar a las organizaciones a desarrollar un plan eficiente y eficaz para migrar con éxito a la nube. AWS CAF organiza la orientación en seis áreas de enfoque denominadas perspectivas: negocios, personas, gobierno, plataforma, seguridad y operaciones. Las perspectivas empresariales, humanas y de gobernanza se centran en las habilidades y los procesos empresariales; las perspectivas de plataforma, seguridad y operaciones se centran en las habilidades y los procesos técnicos. Por ejemplo, la perspectiva humana se dirige a las partes interesadas que se ocupan de los Recursos Humanos (RR. HH.), las funciones del personal y la administración de las personas. Desde esta perspectiva, AWS CAF proporciona orientación para el desarrollo, la formación y la comunicación de las personas a fin de preparar a la organización para una adopción exitosa de la nube. Para obtener más información, consulte la Página web de AWS CAF y el Documento técnico de AWS CAF.

AWS Marco de calificación de la carga de trabajo (AWS WQF)

Herramienta que evalúa las cargas de trabajo de migración de bases de datos, recomienda estrategias de migración y proporciona estimaciones de trabajo. AWS WQF se incluye con AWS

Schema Conversion Tool ().AWS SCT Analiza los esquemas de bases de datos y los objetos de código, el código de las aplicaciones, las dependencias y las características de rendimiento y proporciona informes de evaluación.

B

Un bot malo

Un bot destinado a interrumpir o causar daño a personas u organizaciones.

BCP

Consulte la planificación de la continuidad del negocio.

gráfico de comportamiento

Una vista unificada e interactiva del comportamiento de los recursos y de las interacciones a lo largo del tiempo. Puede utilizar un gráfico de comportamiento con Amazon Detective para examinar los intentos de inicio de sesión fallidos, las llamadas sospechosas a la API y acciones similares. Para obtener más información, consulte Datos en un gráfico de comportamiento en la documentación de Detective.

sistema big-endian

Un sistema que almacena primero el byte más significativo. Véase también endianness.

clasificación binaria

Un proceso que predice un resultado binario (una de las dos clases posibles). Por ejemplo, es posible que su modelo de ML necesite predecir problemas como “¿Este correo electrónico es spam o no es spam?” o “¿Este producto es un libro o un automóvil?”.

filtro de floración

Estructura de datos probabilística y eficiente en términos de memoria que se utiliza para comprobar si un elemento es miembro de un conjunto.

implementación azul/verde

Una estrategia de despliegue en la que se crean dos entornos separados pero idénticos. La versión actual de la aplicación se ejecuta en un entorno (azul) y la nueva versión de la aplicación en el otro entorno (verde). Esta estrategia le ayuda a revertirla rápidamente con un impacto mínimo.

bot

Aplicación de software que ejecuta tareas automatizadas a través de Internet y simula la actividad o interacción humana. Algunos bots son útiles o beneficiosos, como los rastreadores web que indexan información en Internet. Algunos otros bots, conocidos como bots malos, tienen como objetivo interrumpir o causar daños a personas u organizaciones.

botnet

Redes de bots que están infectadas por malware y que están bajo el control de una sola parte, conocida como pastor u operador de bots. Las botnets son el mecanismo más conocido para escalar los bots y su impacto.

branch

Área contenida de un repositorio de código. La primera rama que se crea en un repositorio es la rama principal. Puede crear una rama nueva a partir de una rama existente y, a continuación, desarrollar características o corregir errores en la rama nueva. Una rama que se genera para crear una característica se denomina comúnmente rama de característica. Cuando la característica se encuentra lista para su lanzamiento, se vuelve a combinar la rama de característica con la rama principal. Para obtener más información, consulte Acerca de las sucursales (GitHub documentación).

acceso con cristales rotos

En circunstancias excepcionales y mediante un proceso aprobado, un usuario puede acceder rápidamente a un sitio para el Cuenta de AWS que normalmente no tiene permisos de acceso. Para obtener más información, consulte el indicador Implemente procedimientos de rotura de cristales en la guía Well-Architected AWS .

estrategia de implementación sobre infraestructura existente

La infraestructura existente en su entorno. Al adoptar una estrategia de implementación sobre infraestructura existente para una arquitectura de sistemas, se diseña la arquitectura en función de las limitaciones de los sistemas y la infraestructura actuales. Si está ampliando la infraestructura existente, puede combinar las estrategias de implementación sobre infraestructuras existentes y de implementación desde cero.

caché de búfer

El área de memoria donde se almacenan los datos a los que se accede con más frecuencia.

capacidad empresarial

Lo que hace una empresa para generar valor (por ejemplo, ventas, servicio al cliente o marketing). Las arquitecturas de microservicios y las decisiones de desarrollo pueden estar impulsadas por las capacidades empresariales. Para obtener más información, consulte la sección Organizado en torno a las capacidades empresariales del documento técnico Ejecutar microservicios en contenedores en AWS.

planificación de la continuidad del negocio (BCP)

Plan que aborda el posible impacto de un evento disruptivo, como una migración a gran escala en las operaciones y permite a la empresa reanudar las operaciones rápidamente.

C

CAF

Consulte el marco AWS de adopción de la nube.

despliegue canario

El lanzamiento lento e incremental de una versión para los usuarios finales. Cuando está seguro, despliega la nueva versión y reemplaza la versión actual en su totalidad.

CCoE

Consulte Cloud Center of Excellence.

CDC

Consulte la captura de datos de cambios.

captura de datos de cambio (CDC)

Proceso de seguimiento de los cambios en un origen de datos, como una tabla de base de datos, y registro de los metadatos relacionados con el cambio. Puede utilizar los CDC para diversos fines, como auditar o replicar los cambios en un sistema de destino para mantener la sincronización.

ingeniería del caos

Introducir intencionalmente fallos o eventos disruptivos para poner a prueba la resiliencia de un sistema. Puedes usar AWS Fault Injection Service (AWS FIS) para realizar experimentos que estresen tus AWS cargas de trabajo y evalúen su respuesta.

CI/CD

Consulte la integración continua y la entrega continua.

clasificación

Un proceso de categorización que permite generar predicciones. Los modelos de ML para problemas de clasificación predicen un valor discreto. Los valores discretos siempre son distintos entre sí. Por ejemplo, es posible que un modelo necesite evaluar si hay o no un automóvil en una imagen.

cifrado del cliente

Cifrado de datos localmente, antes de que el objetivo los Servicio de AWS reciba.

Centro de excelencia en la nube (CCoE)

Equipo multidisciplinario que impulsa los esfuerzos de adopción de la nube en toda la organización, incluido el desarrollo de las prácticas recomendadas en la nube, la movilización de recursos, el establecimiento de plazos de migración y la dirección de la organización durante las transformaciones a gran escala. Para obtener más información, consulte las publicaciones de CCo E en el blog de estrategia Nube de AWS empresarial.

computación en la nube

La tecnología en la nube que se utiliza normalmente para la administración de dispositivos de IoT y el almacenamiento de datos de forma remota. La computación en la nube suele estar conectada a la tecnología de computación perimetral.

modelo operativo en la nube

En una organización de TI, el modelo operativo que se utiliza para crear, madurar y optimizar uno o más entornos de nube. Para obtener más información, consulte Creación de su modelo operativo de nube.

etapas de adopción de la nube

Las cuatro fases por las que suelen pasar las organizaciones cuando migran a Nube de AWS:

- Proyecto: ejecución de algunos proyectos relacionados con la nube con fines de prueba de concepto y aprendizaje
- Fundamento: realizar inversiones fundamentales para escalar su adopción de la nube (p. ej., crear una landing zone, definir una CCo E, establecer un modelo de operaciones)
- Migración: migración de aplicaciones individuales
- Reinvención: optimización de productos y servicios e innovación en la nube

Stephen Orban definió estas etapas en la entrada del blog The Journey Toward Cloud-First & the Stages of Adoption en el blog Nube de AWS Enterprise Strategy. Para obtener información sobre su relación con la estrategia de AWS migración, consulte la guía de preparación para la migración.

CMDB

Consulte la base de datos de administración de la configuración.

repositorio de código

Una ubicación donde el código fuente y otros activos, como documentación, muestras y scripts, se almacenan y actualizan mediante procesos de control de versiones. Los repositorios en la nube más comunes incluyen GitHub oBitbucket Cloud. Cada versión del código se denomina rama. En una estructura de microservicios, cada repositorio se encuentra dedicado a una única funcionalidad. Una sola canalización de CI/CD puede utilizar varios repositorios.

caché en frío

Una caché de búfer que está vacía no está bien poblada o contiene datos obsoletos o irrelevantes. Esto afecta al rendimiento, ya que la instancia de la base de datos debe leer desde la memoria principal o el disco, lo que es más lento que leer desde la memoria caché del búfer.

datos fríos

Datos a los que se accede con poca frecuencia y que suelen ser históricos. Al consultar este tipo de datos, normalmente se aceptan consultas lentas. Trasladar estos datos a niveles o clases de almacenamiento de menor rendimiento y menos costosos puede reducir los costos.

visión artificial (CV)

Campo de la IA que utiliza el aprendizaje automático para analizar y extraer información de formatos visuales, como imágenes y vídeos digitales. Por ejemplo, Amazon SageMaker AI proporciona algoritmos de procesamiento de imágenes para CV.

desviación de configuración

En el caso de una carga de trabajo, un cambio de configuración con respecto al estado esperado. Puede provocar que la carga de trabajo deje de cumplir las normas y, por lo general, es gradual e involuntario.

base de datos de administración de configuración (CMDB)

Repositorio que almacena y administra información sobre una base de datos y su entorno de TI, incluidos los componentes de hardware y software y sus configuraciones. Por lo general, los

datos de una CMDB se utilizan en la etapa de detección y análisis de la cartera de productos durante la migración.

paquete de conformidad

Conjunto de AWS Config reglas y medidas correctivas que puede reunir para personalizar sus comprobaciones de conformidad y seguridad. Puede implementar un paquete de conformidad como una entidad única en una región Cuenta de AWS y, o en una organización, mediante una plantilla YAML. Para obtener más información, consulta los paquetes de conformidad en la documentación. AWS Config

integración y entrega continuas (CI/CD)

El proceso de automatización de las etapas de origen, compilación, prueba, puesta en escena y producción del proceso de publicación del software. CI/CD is commonly described as a pipeline. CI/CDpuede ayudarlo a automatizar los procesos, mejorar la productividad, mejorar la calidad del código y entregar con mayor rapidez. Para obtener más información, consulte Beneficios de la entrega continua. CD también puede significar implementación continua. Para obtener más información, consulte Entrega continua frente a implementación continua.

CV

Vea la visión artificial.

D

datos en reposo

Datos que están estacionarios en la red, como los datos que se encuentran almacenados.

clasificación de datos

Un proceso para identificar y clasificar los datos de su red en función de su importancia y sensibilidad. Es un componente fundamental de cualquier estrategia de administración de riesgos de ciberseguridad porque lo ayuda a determinar los controles de protección y retención adecuados para los datos. La clasificación de datos es un componente del pilar de seguridad del AWS Well-Architected Framework. Para obtener más información, consulte Clasificación de datos.

desviación de datos

Una variación significativa entre los datos de producción y los datos que se utilizaron para entrenar un modelo de machine learning, o un cambio significativo en los datos de entrada

a lo largo del tiempo. La desviación de los datos puede reducir la calidad, la precisión y la imparcialidad generales de las predicciones de los modelos de machine learning.

datos en tránsito

Datos que se mueven de forma activa por la red, por ejemplo, entre los recursos de la red.

malla de datos

Un marco arquitectónico que proporciona una propiedad de datos distribuida y descentralizada con una administración y un gobierno centralizados.

minimización de datos

El principio de recopilar y procesar solo los datos estrictamente necesarios. Practicar la minimización de los datos Nube de AWS puede reducir los riesgos de privacidad, los costos y la huella de carbono de la analítica.

perímetro de datos

Un conjunto de barreras preventivas en su AWS entorno que ayudan a garantizar que solo las identidades confiables accedan a los recursos confiables desde las redes esperadas. Para obtener más información, consulte Crear un perímetro de datos sobre. AWS

preprocesamiento de datos

Transformar los datos sin procesar en un formato que su modelo de ML pueda analizar fácilmente. El preprocesamiento de datos puede implicar eliminar determinadas columnas o filas y corregir los valores faltantes, incoherentes o duplicados.

procedencia de los datos

El proceso de rastrear el origen y el historial de los datos a lo largo de su ciclo de vida, por ejemplo, la forma en que se generaron, transmitieron y almacenaron los datos.

titular de los datos

Persona cuyos datos se recopilan y procesan.

almacenamiento de datos

Un sistema de administración de datos que respalde la inteligencia empresarial, como la analítica. Los almacenes de datos suelen contener grandes cantidades de datos históricos y, por lo general, se utilizan para consultas y análisis.

lenguaje de definición de datos (DDL)

Instrucciones o comandos para crear o modificar la estructura de tablas y objetos de una base de datos.

lenguaje de manipulación de datos (DML)

Instrucciones o comandos para modificar (insertar, actualizar y eliminar) la información de una base de datos.

DDL

Consulte el lenguaje de definición de bases de datos.

conjunto profundo

Combinar varios modelos de aprendizaje profundo para la predicción. Puede utilizar conjuntos profundos para obtener una predicción más precisa o para estimar la incertidumbre de las predicciones.

aprendizaje profundo

Un subcampo del ML que utiliza múltiples capas de redes neuronales artificiales para identificar el mapeo entre los datos de entrada y las variables objetivo de interés.

defense-in-depth

Un enfoque de seguridad de la información en el que se distribuyen cuidadosamente una serie de mecanismos y controles de seguridad en una red informática para proteger la confidencialidad, la integridad y la disponibilidad de la red y de los datos que contiene. Al adoptar esta estrategia AWS, se añaden varios controles en diferentes capas de la AWS Organizations estructura para ayudar a proteger los recursos. Por ejemplo, un defense-in-depth enfoque podría combinar la autenticación multifactorial, la segmentación de la red y el cifrado.

administrador delegado

En AWS Organizations, un servicio compatible puede registrar una cuenta de AWS miembro para administrar las cuentas de la organización y gestionar los permisos de ese servicio. Esta cuenta se denomina administrador delegado para ese servicio. Para obtener más información y una lista de servicios compatibles, consulte Servicios que funcionan con AWS Organizations en la documentación de AWS Organizations .

Implementación

El proceso de hacer que una aplicación, características nuevas o correcciones de código se encuentren disponibles en el entorno de destino. La implementación abarca implementar

cambios en una base de código y, a continuación, crear y ejecutar esa base en los entornos de la aplicación.

entorno de desarrollo

Consulte entorno.

control de detección

Un control de seguridad que se ha diseñado para detectar, registrar y alertar después de que se produzca un evento. Estos controles son una segunda línea de defensa, ya que lo advierten sobre los eventos de seguridad que han eludido los controles preventivos establecidos. Para obtener más información, consulte Controles de detección en Implementación de controles de seguridad en AWS.

asignación de flujos de valor para el desarrollo (DVSM)

Proceso que se utiliza para identificar y priorizar las restricciones que afectan negativamente a la velocidad y la calidad en el ciclo de vida del desarrollo de software. DVSM amplía el proceso de asignación del flujo de valor diseñado originalmente para las prácticas de fabricación ajustada. Se centra en los pasos y los equipos necesarios para crear y transferir valor a través del proceso de desarrollo de software.

gemelo digital

Representación virtual de un sistema del mundo real, como un edificio, una fábrica, un equipo industrial o una línea de producción. Los gemelos digitales son compatibles con el mantenimiento predictivo, la supervisión remota y la optimización de la producción.

tabla de dimensiones

En un esquema en estrella, tabla más pequeña que contiene los atributos de datos sobre los datos cuantitativos de una tabla de hechos. Los atributos de la tabla de dimensiones suelen ser campos de texto o números discretos que se comportan como texto. Estos atributos se utilizan habitualmente para restringir consultas, filtrar y etiquetar conjuntos de resultados.

desastre

Un evento que impide que una carga de trabajo o un sistema cumplan sus objetivos empresariales en su ubicación principal de implementación. Estos eventos pueden ser desastres naturales, fallos técnicos o el resultado de acciones humanas, como una configuración incorrecta involuntaria o un ataque de malware.

recuperación de desastres (DR)

La estrategia y el proceso que se utilizan para minimizar el tiempo de inactividad y la pérdida de datos ocasionados por un desastre. Para obtener más información, consulte Recuperación ante desastres de cargas de trabajo en AWS: Recovery in the Cloud in the AWS Well-Architected Framework.

DML

Consulte el lenguaje de manipulación de bases de datos.

diseño basado en el dominio

Un enfoque para desarrollar un sistema de software complejo mediante la conexión de sus componentes a dominios en evolución, o a los objetivos empresariales principales, a los que sirve cada componente. Este concepto lo introdujo Eric Evans en su libro, Diseño impulsado por el dominio: abordando la complejidad en el corazón del software (Boston: Addison-Wesley Professional, 2003). Para obtener información sobre cómo utilizar el diseño basado en dominios con el patrón de higos estranguladores, consulte Modernización gradual de los servicios web antiguos de Microsoft ASP.NET (ASMX) mediante contenedores y Amazon API Gateway.

DR

Consulte recuperación ante desastres.

detección de deriva

Seguimiento de las desviaciones con respecto a una configuración de referencia. Por ejemplo, puedes usarlo AWS CloudFormation para detectar desviaciones en los recursos del sistema o puedes usarlo AWS Control Tower para detectar cambios en tu landing zone que puedan afectar al cumplimiento de los requisitos de gobierno.

DVSM

Consulte el mapeo del flujo de valor del desarrollo.

E

EDA

Consulte el análisis exploratorio de datos.

EDI

Véase intercambio electrónico de datos.

computación en la periferia

La tecnología que aumenta la potencia de cálculo de los dispositivos inteligentes en la periferia de una red de IoT. En comparación con la computación en nube, la computación perimetral puede reducir la latencia de la comunicación y mejorar el tiempo de respuesta.

intercambio electrónico de datos (EDI)

El intercambio automatizado de documentos comerciales entre organizaciones. Para obtener más información, consulte Qué es el intercambio electrónico de datos.

cifrado

Proceso informático que transforma datos de texto plano, legibles por humanos, en texto cifrado.

clave de cifrado

Cadena criptográfica de bits aleatorios que se genera mediante un algoritmo de cifrado. Las claves pueden variar en longitud y cada una se ha diseñado para ser impredecible y única.

endianidad

El orden en el que se almacenan los bytes en la memoria del ordenador. Los sistemas big-endianos almacenan primero el byte más significativo. Los sistemas Little-Endian almacenan primero el byte menos significativo.

punto de conexión

Consulte el punto final del servicio.

servicio de punto de conexión

Servicio que puede alojar en una nube privada virtual (VPC) para compartir con otros usuarios. Puede crear un servicio de punto final AWS PrivateLink y conceder permisos a otros directores Cuentas de AWS o a AWS Identity and Access Management (IAM). Estas cuentas o entidades principales pueden conectarse a su servicio de punto de conexión de forma privada mediante la creación de puntos de conexión de VPC de interfaz. Para obtener más información, consulte Creación de un servicio de punto de conexión en la documentación de Amazon Virtual Private Cloud (Amazon VPC).

planificación de recursos empresariales (ERP)

Un sistema que automatiza y gestiona los procesos empresariales clave (como la contabilidad, el MES y la gestión de proyectos) de una empresa.

cifrado de sobre

El proceso de cifrar una clave de cifrado con otra clave de cifrado. Para obtener más información, consulte el cifrado de sobres en la documentación de AWS Key Management Service (AWS KMS).

entorno

Una instancia de una aplicación en ejecución. Los siguientes son los tipos de entornos más comunes en la computación en la nube:

- entorno de desarrollo: instancia de una aplicación en ejecución que solo se encuentra disponible para el equipo principal responsable del mantenimiento de la aplicación. Los entornos de desarrollo se utilizan para probar los cambios antes de promocionarlos a los entornos superiores. Este tipo de entorno a veces se denomina entorno de prueba.
- entornos inferiores: todos los entornos de desarrollo de una aplicación, como los que se utilizan para las compilaciones y pruebas iniciales.
- entorno de producción: instancia de una aplicación en ejecución a la que pueden acceder los usuarios finales. En una canalización de CI/CD, el entorno de producción es el último entorno de implementación.
- entornos superiores: todos los entornos a los que pueden acceder usuarios que no sean del equipo de desarrollo principal. Esto puede incluir un entorno de producción, entornos de preproducción y entornos para las pruebas de aceptación por parte de los usuarios.

epopeya

En las metodologías ágiles, son categorías funcionales que ayudan a organizar y priorizar el trabajo. Las epopeyas brindan una descripción detallada de los requisitos y las tareas de implementación. Por ejemplo, las epopeyas AWS de seguridad de CAF incluyen la gestión de identidades y accesos, los controles de detección, la seguridad de la infraestructura, la protección de datos y la respuesta a incidentes. Para obtener más información sobre las epopeyas en la estrategia de migración de AWS , consulte la Guía de implementación del programa.

ERP

Consulte planificación de recursos empresariales.

análisis de datos de tipo exploratorio (EDA)

El proceso de analizar un conjunto de datos para comprender sus características principales. Se recopilan o agregan datos y, a continuación, se realizan las investigaciones iniciales para

encontrar patrones, detectar anomalías y comprobar las suposiciones. El EDA se realiza mediante el cálculo de estadísticas resumidas y la creación de visualizaciones de datos.

F

tabla de datos

La tabla central de un esquema en forma de estrella. Almacena datos cuantitativos sobre las operaciones comerciales. Normalmente, una tabla de hechos contiene dos tipos de columnas: las que contienen medidas y las que contienen una clave externa para una tabla de dimensiones.

fallan rápidamente

Una filosofía que utiliza pruebas frecuentes e incrementales para reducir el ciclo de vida del desarrollo. Es una parte fundamental de un enfoque ágil.

límite de aislamiento de fallas

En el Nube de AWS, un límite, como una zona de disponibilidad Región de AWS, un plano de control o un plano de datos, que limita el efecto de una falla y ayuda a mejorar la resiliencia de las cargas de trabajo. Para obtener más información, consulte Límites de AWS aislamiento de errores.

rama de característica

Consulte la sucursal.

características

Los datos de entrada que se utilizan para hacer una predicción. Por ejemplo, en un contexto de fabricación, las características pueden ser imágenes que se capturan periódicamente desde la línea de fabricación.

importancia de las características

La importancia que tiene una característica para las predicciones de un modelo. Por lo general, esto se expresa como una puntuación numérica que se puede calcular mediante diversas técnicas, como las explicaciones aditivas de Shapley (SHAP) y los gradientes integrados. Para obtener más información, consulte Interpretabilidad del modelo de aprendizaje automático con AWS.

transformación de funciones

Optimizar los datos para el proceso de ML, lo que incluye enriquecer los datos con fuentes adicionales, escalar los valores o extraer varios conjuntos de información de un solo campo de datos. Esto permite que el modelo de ML se beneficie de los datos. Por ejemplo, si divide la fecha del “27 de mayo de 2021 00:15:37” en “jueves”, “mayo”, “2021” y “15”, puede ayudar al algoritmo de aprendizaje a aprender patrones matizados asociados a los diferentes componentes de los datos.

indicaciones de unos pocos pasos

Proporcionar a un LLM un pequeño número de ejemplos que demuestren la tarea y el resultado deseado antes de pedirle que realice una tarea similar. Esta técnica es una aplicación del aprendizaje contextual, en el que los modelos aprenden a partir de ejemplos (planos) integrados en las instrucciones. Las indicaciones con pocas tomas pueden ser eficaces para tareas que requieren un formato, un razonamiento o un conocimiento del dominio específicos. Consulte también el apartado de mensajes sin intervención.

FGAC

Consulte el control de acceso detallado.

control de acceso preciso (FGAC)

El uso de varias condiciones que tienen por objetivo permitir o denegar una solicitud de acceso.

migración relámpago

Método de migración de bases de datos que utiliza la replicación continua de datos mediante la captura de datos modificados para migrar los datos en el menor tiempo posible, en lugar de utilizar un enfoque gradual. El objetivo es reducir al mínimo el tiempo de inactividad.

FM

Consulte el modelo básico.

modelo de base (FM)

Una gran red neuronal de aprendizaje profundo que se ha estado entrenando con conjuntos de datos masivos de datos generalizados y sin etiquetar. FMs son capaces de realizar una amplia variedad de tareas generales, como comprender el lenguaje, generar texto e imágenes y conversar en lenguaje natural. Para obtener más información, consulte Qué son los modelos básicos.

G

IA generativa

Un subconjunto de modelos de IA que se han entrenado con grandes cantidades de datos y que pueden utilizar un simple mensaje de texto para crear contenido y artefactos nuevos, como imágenes, vídeos, texto y audio. Para obtener más información, consulte Qué es la IA generativa.

bloqueo geográfico

Consulta las restricciones geográficas.

restricciones geográficas (bloqueo geográfico)

En Amazon CloudFront, una opción para impedir que los usuarios de países específicos accedan a las distribuciones de contenido. Puede utilizar una lista de permitidos o bloqueados para especificar los países aprobados y prohibidos. Para obtener más información, consulta Restringir la distribución geográfica del contenido en la CloudFront documentación.

Flujo de trabajo de Gitflow

Un enfoque en el que los entornos inferiores y superiores utilizan diferentes ramas en un repositorio de código fuente. El flujo de trabajo de Gitflow se considera heredado, y el flujo de trabajo basado en enlaces troncales es el enfoque moderno preferido.

imagen dorada

Instantánea de un sistema o software que se utiliza como plantilla para implementar nuevas instancias de ese sistema o software. Por ejemplo, en la fabricación, una imagen dorada se puede utilizar para aprovisionar software en varios dispositivos y ayuda a mejorar la velocidad, la escalabilidad y la productividad de las operaciones de fabricación de dispositivos.

estrategia de implementación desde cero

La ausencia de infraestructura existente en un entorno nuevo. Al adoptar una estrategia de implementación desde cero para una arquitectura de sistemas, puede seleccionar todas las tecnologías nuevas sin que estas deban ser compatibles con una infraestructura existente, lo que también se conoce como implementación sobre infraestructura existente. Si está ampliando la infraestructura existente, puede combinar las estrategias de implementación sobre infraestructuras existentes y de implementación desde cero.

barrera de protección

Una regla de alto nivel que ayuda a regular los recursos, las políticas y el cumplimiento en todas las unidades organizativas (OUs). Las barreras de protección preventivas aplican políticas para garantizar la alineación con los estándares de conformidad. Se implementan mediante políticas de control de servicios y límites de permisos de IAM. Las barreras de protección de detección detectan las vulneraciones de las políticas y los problemas de conformidad, y generan alertas para su corrección. Se implementan mediante Amazon AWS Config AWS Security Hub GuardDuty AWS Trusted Advisor, Amazon Inspector y AWS Lambda cheques personalizados.

H

HA

Consulte la alta disponibilidad.

migración heterogénea de bases de datos

Migración de la base de datos de origen a una base de datos de destino que utilice un motor de base de datos diferente (por ejemplo, de Oracle a Amazon Aurora). La migración heterogénea suele ser parte de un esfuerzo de rediseño de la arquitectura y convertir el esquema puede ser una tarea compleja. AWS ofrece AWS SCT, lo cual ayuda con las conversiones de esquemas.

alta disponibilidad (HA)

La capacidad de una carga de trabajo para funcionar de forma continua, sin intervención, en caso de desafíos o desastres. Los sistemas de alta disponibilidad están diseñados para realizar una conmutación por error automática, ofrecer un rendimiento de alta calidad de forma constante y gestionar diferentes cargas y fallos con un impacto mínimo en el rendimiento.

modernización histórica

Un enfoque utilizado para modernizar y actualizar los sistemas de tecnología operativa (TO) a fin de satisfacer mejor las necesidades de la industria manufacturera. Un histórico es un tipo de base de datos que se utiliza para recopilar y almacenar datos de diversas fuentes en una fábrica.

datos retenidos

Parte de los datos históricos etiquetados que se ocultan de un conjunto de datos que se utiliza para entrenar un modelo de aprendizaje automático. Puede utilizar los datos de reserva para evaluar el rendimiento del modelo comparando las predicciones del modelo con los datos de reserva.

migración homogénea de bases de datos

Migración de la base de datos de origen a una base de datos de destino que comparte el mismo motor de base de datos (por ejemplo, Microsoft SQL Server a Amazon RDS para SQL Server). La migración homogénea suele formar parte de un esfuerzo para volver a alojar o redefinir la plataforma. Puede utilizar las utilidades de bases de datos nativas para migrar el esquema.

datos recientes

Datos a los que se accede con frecuencia, como datos en tiempo real o datos traslacionales recientes. Por lo general, estos datos requieren un nivel o una clase de almacenamiento de alto rendimiento para proporcionar respuestas rápidas a las consultas.

hotfix

Una solución urgente para un problema crítico en un entorno de producción. Debido a su urgencia, las revisiones suelen realizarse fuera del flujo de trabajo habitual de las versiones. DevOps

periodo de hiperatención

Periodo, inmediatamente después de la transición, durante el cual un equipo de migración administra y monitorea las aplicaciones migradas en la nube para solucionar cualquier problema. Por lo general, este periodo dura de 1 a 4 días. Al final del periodo de hiperatención, el equipo de migración suele transferir la responsabilidad de las aplicaciones al equipo de operaciones en la nube.

I

IaC

Vea la infraestructura como código.

políticas basadas en identidad

Política asociada a uno o más directores de IAM que define sus permisos en el Nube de AWS entorno.

aplicación inactiva

Aplicación que utiliza un promedio de CPU y memoria de entre 5 y 20 por ciento durante un periodo de 90 días. En un proyecto de migración, es habitual retirar estas aplicaciones o mantenerlas en las instalaciones.

IIoT

Consulte Internet de las cosas industrial.

infraestructura inmutable

Un modelo que implementa una nueva infraestructura para las cargas de trabajo de producción en lugar de actualizar, aplicar parches o modificar la infraestructura existente. Las infraestructuras inmutables son intrínsecamente más consistentes, fiables y predecibles que las infraestructuras mutables. Para obtener más información, consulte las prácticas recomendadas para implementar con una infraestructura inmutable en Well-Architected Framework AWS .

VPC entrante (de entrada)

En una arquitectura de AWS cuentas múltiples, una VPC que acepta, inspecciona y enruta las conexiones de red desde fuera de una aplicación. La arquitectura AWS de referencia de seguridad recomienda configurar la cuenta de red con entradas, salidas e inspección VPCs para proteger la interfaz bidireccional entre la aplicación y el resto de Internet.

migración gradual

Estrategia de transición en la que se migra la aplicación en partes pequeñas en lugar de realizar una transición única y completa. Por ejemplo, puede trasladar inicialmente solo unos pocos microservicios o usuarios al nuevo sistema. Tras comprobar que todo funciona correctamente, puede trasladar microservicios o usuarios adicionales de forma gradual hasta que pueda retirar su sistema heredado. Esta estrategia reduce los riesgos asociados a las grandes migraciones.

Industria 4.0

Un término que Klaus Schwab introdujo en 2016 para referirse a la modernización de los procesos de fabricación mediante avances en la conectividad, los datos en tiempo real, la automatización, el análisis y la inteligencia artificial/aprendizaje automático.

infraestructura

Todos los recursos y activos que se encuentran en el entorno de una aplicación.

infraestructura como código (IaC)

Proceso de aprovisionamiento y administración de la infraestructura de una aplicación mediante un conjunto de archivos de configuración. La IaC se ha diseñado para ayudarlo a centralizar la administración de la infraestructura, estandarizar los recursos y escalar con rapidez a fin de que los entornos nuevos sean repetibles, fiables y consistentes.

Internet de las cosas industrial (T) IIo

El uso de sensores y dispositivos conectados a Internet en los sectores industriales, como el productivo, el eléctrico, el automotriz, el sanitario, el de las ciencias de la vida y el de la agricultura. Para obtener más información, consulte Creación de una estrategia de transformación digital de la Internet de las cosas (IIoT) industrial.

VPC de inspección

En una arquitectura de AWS cuentas múltiples, una VPC centralizada que gestiona las inspecciones del tráfico de red VPCs entre Internet y las redes locales (en una misma o Regiones de AWS diferente). La arquitectura AWS de referencia de seguridad recomienda configurar su cuenta de red con entrada, salida e inspección VPCs para proteger la interfaz bidireccional entre la aplicación e Internet en general.

Internet de las cosas (IoT)

Red de objetos físicos conectados con sensores o procesadores integrados que se comunican con otros dispositivos y sistemas a través de Internet o de una red de comunicación local. Para obtener más información, consulte ¿Qué es IoT?.

interpretabilidad

Característica de un modelo de machine learning que describe el grado en que un ser humano puede entender cómo las predicciones del modelo dependen de sus entradas. Para obtener más información, consulte Interpretabilidad del modelo de aprendizaje automático con. AWS

IoT

Consulte Internet de las cosas.

biblioteca de información de TI (ITIL)

Conjunto de prácticas recomendadas para ofrecer servicios de TI y alinearlos con los requisitos empresariales. La ITIL proporciona la base para la ITSM.

administración de servicios de TI (ITSM)

Actividades asociadas con el diseño, la implementación, la administración y el soporte de los servicios de TI para una organización. Para obtener información sobre la integración de las operaciones en la nube con las herramientas de ITSM, consulte la Guía de integración de operaciones.

ITIL

Consulte la biblioteca de información de TI.

ITSM

Consulte Administración de servicios de TI.

L

control de acceso basado en etiquetas (LBAC)

Una implementación del control de acceso obligatorio (MAC) en la que a los usuarios y a los propios datos se les asigna explícitamente un valor de etiqueta de seguridad. La intersección entre la etiqueta de seguridad del usuario y la etiqueta de seguridad de los datos determina qué filas y columnas puede ver el usuario.

zona de aterrizaje

Una landing zone es un AWS entorno multicuenta bien diseñado, escalable y seguro. Este es un punto de partida desde el cual las empresas pueden lanzar e implementar rápidamente cargas de trabajo y aplicaciones con confianza en su entorno de seguridad e infraestructura. Para obtener más información sobre las zonas de aterrizaje, consulte Configuración de un entorno de AWS seguro y escalable con varias cuentas.

modelo de lenguaje grande (LLM)

Un modelo de IA de aprendizaje profundo que se entrena previamente con una gran cantidad de datos. Un LLM puede realizar múltiples tareas, como responder preguntas, resumir documentos, traducir textos a otros idiomas y completar oraciones. Para obtener más información, consulte Qué son. LLMs

migración grande

Migración de 300 servidores o más.

LBAC

Consulte control de acceso basado en etiquetas.

privilegio mínimo

La práctica recomendada de seguridad que consiste en conceder los permisos mínimos necesarios para realizar una tarea. Para obtener más información, consulte Aplicar permisos de privilegio mínimo en la documentación de IAM.

migrar mediante lift-and-shift

Ver 7 Rs.

sistema little-endian

Un sistema que almacena primero el byte menos significativo. Véase también endianness.

LLM

Véase un modelo de lenguaje amplio.

entornos inferiores

Véase entorno.

M

machine learning (ML)

Un tipo de inteligencia artificial que utiliza algoritmos y técnicas para el reconocimiento y el aprendizaje de patrones. El ML analiza y aprende de los datos registrados, como los datos del Internet de las cosas (IoT), para generar un modelo estadístico basado en patrones. Para más información, consulte Machine learning.

rama principal

Ver sucursal.

malware

Software diseñado para comprometer la seguridad o la privacidad de la computadora. El malware puede interrumpir los sistemas informáticos, filtrar información confidencial u obtener acceso no autorizado. Algunos ejemplos de malware son los virus, los gusanos, el ransomware, los troyanos, el spyware y los registradores de pulsaciones de teclas.

servicios gestionados

Servicios de AWS para los que AWS opera la capa de infraestructura, el sistema operativo y las plataformas, y usted accede a los puntos finales para almacenar y recuperar datos. Amazon Simple Storage Service (Amazon S3) y Amazon DynamoDB son ejemplos de servicios gestionados. También se conocen como servicios abstractos.

sistema de ejecución de fabricación (MES)

Un sistema de software para rastrear, monitorear, documentar y controlar los procesos de producción que convierten las materias primas en productos terminados en el taller.

MAP

Consulte Migration Acceleration Program.

mecanismo

Un proceso completo en el que se crea una herramienta, se impulsa su adopción y, a continuación, se inspeccionan los resultados para realizar ajustes. Un mecanismo es un ciclo que se refuerza y mejora a sí mismo a medida que funciona. Para obtener más información, consulte Creación de mecanismos en el AWS Well-Architected Framework.

cuenta de miembro

Todas las Cuentas de AWS demás cuentas, excepto la de administración, que forman parte de una organización. AWS Organizations Una cuenta no puede pertenecer a más de una organización a la vez.

MES

Consulte el sistema de ejecución de la fabricación.

Transporte telemétrico de Message Queue Queue (MQTT)

Un protocolo de comunicación ligero machine-to-machine (M2M), basado en el patrón de publicación/suscripción, para dispositivos de IoT con recursos limitados.

microservicio

Un servicio pequeño e independiente que se comunica a través de una red bien definida APIs y que, por lo general, es propiedad de equipos pequeños e independientes. Por ejemplo, un sistema de seguros puede incluir microservicios que se adapten a las capacidades empresariales, como las de ventas o marketing, o a subdominios, como las de compras, reclamaciones o análisis. Los beneficios de los microservicios incluyen la agilidad, la escalabilidad flexible, la facilidad de implementación, el código reutilizable y la resiliencia. Para obtener más información, consulte Integrar microservicios mediante AWS servicios sin servidor.

arquitectura de microservicios

Un enfoque para crear una aplicación con componentes independientes que ejecutan cada proceso de la aplicación como un microservicio. Estos microservicios se comunican a través de una interfaz bien definida mediante un uso ligero. APIs Cada microservicio de esta arquitectura se puede actualizar, implementar y escalar para satisfacer la demanda de funciones específicas de una aplicación. Para obtener más información, consulte Implementación de microservicios en. AWS

Programa de aceleración de la migración (MAP)

Un AWS programa que proporciona soporte de consultoría, formación y servicios para ayudar a las organizaciones a crear una base operativa sólida para migrar a la nube y para ayudar a compensar el costo inicial de las migraciones. El MAP incluye una metodología de migración para ejecutar las migraciones antiguas de forma metódica y un conjunto de herramientas para automatizar y acelerar los escenarios de migración más comunes.

migración a escala

Proceso de transferencia de la mayoría de la cartera de aplicaciones a la nube en oleadas, con más aplicaciones desplazadas a un ritmo más rápido en cada oleada. En esta fase, se utilizan las prácticas recomendadas y las lecciones aprendidas en las fases anteriores para implementar una fábrica de migración de equipos, herramientas y procesos con el fin de agilizar la migración de las cargas de trabajo mediante la automatización y la entrega ágil. Esta es la tercera fase de la estrategia de migración de AWS.

fábrica de migración

Equipos multifuncionales que agilizan la migración de las cargas de trabajo mediante enfoques automatizados y ágiles. Los equipos de las fábricas de migración suelen incluir a analistas y propietarios de operaciones, empresas, ingenieros de migración, desarrolladores y DevOps profesionales que trabajan a pasos agigantados. Entre el 20 y el 50 por ciento de la cartera de aplicaciones empresariales se compone de patrones repetidos que pueden optimizarse mediante un enfoque de fábrica. Para obtener más información, consulte la discusión sobre las fábricas de migración y la Guía de fábricas de migración a la nube en este contenido.

metadatos de migración

Información sobre la aplicación y el servidor que se necesita para completar la migración. Cada patrón de migración requiere un conjunto diferente de metadatos de migración. Algunos ejemplos de metadatos de migración son la subred de destino, el grupo de seguridad y AWS la cuenta.

patrón de migración

Tarea de migración repetible que detalla la estrategia de migración, el destino de la migración y la aplicación o el servicio de migración utilizados. Ejemplo: realoje la migración a Amazon EC2 con AWS Application Migration Service.

Migration Portfolio Assessment (MPA)

Una herramienta en línea que proporciona información para validar el modelo de negocio para migrar a. Nube de AWS La MPA ofrece una evaluación detallada de la cartera (adecuación del

tamaño de los servidores, precios, comparaciones del costo total de propiedad, análisis de los costos de migración), así como una planificación de la migración (análisis y recopilación de datos de aplicaciones, agrupación de aplicaciones, priorización de la migración y planificación de oleadas). La herramienta MPA (requiere iniciar sesión) está disponible de forma gratuita para todos los AWS consultores y consultores asociados de APN.

Evaluación de la preparación para la migración (MRA)

Proceso que consiste en obtener información sobre el estado de preparación de una organización para la nube, identificar sus puntos fuertes y débiles y elaborar un plan de acción para cerrar las brechas identificadas mediante el AWS CAF. Para obtener más información, consulte la Guía de preparación para la migración. La MRA es la primera fase de la estrategia de migración de AWS.

estrategia de migración

El enfoque utilizado para migrar una carga de trabajo a. Nube de AWS Para obtener más información, consulte la entrada de las 7 R de este glosario y consulte Movilice a su organización para acelerar las migraciones a gran escala.

ML

Consulte el aprendizaje automático.

modernización

Transformar una aplicación obsoleta (antigua o monolítica) y su infraestructura en un sistema ágil, elástico y de alta disponibilidad en la nube para reducir los gastos, aumentar la eficiencia y aprovechar las innovaciones. Para obtener más información, consulte Estrategia para modernizar las aplicaciones en el Nube de AWS.

evaluación de la preparación para la modernización

Evaluación que ayuda a determinar la preparación para la modernización de las aplicaciones de una organización; identifica los beneficios, los riesgos y las dependencias; y determina qué tan bien la organización puede soportar el estado futuro de esas aplicaciones. El resultado de la evaluación es un esquema de la arquitectura objetivo, una hoja de ruta que detalla las fases de desarrollo y los hitos del proceso de modernización y un plan de acción para abordar las brechas identificadas. Para obtener más información, consulte Evaluación de la preparación para la modernización de las aplicaciones en el Nube de AWS.

aplicaciones monolíticas (monolitos)

Aplicaciones que se ejecutan como un único servicio con procesos estrechamente acoplados. Las aplicaciones monolíticas presentan varios inconvenientes. Si una característica de la

aplicación experimenta un aumento en la demanda, se debe escalar toda la arquitectura. Agregar o mejorar las características de una aplicación monolítica también se vuelve más complejo a medida que crece la base de código. Para solucionar problemas con la aplicación, puede utilizar una arquitectura de microservicios. Para obtener más información, consulte Descomposición de monolitos en microservicios.

MAPA

Consulte la evaluación de la cartera de migración.

MQTT

Consulte Message Queue Queue Telemetría y Transporte.

clasificación multiclase

Un proceso que ayuda a generar predicciones para varias clases (predice uno de más de dos resultados). Por ejemplo, un modelo de ML podría preguntar "¿Este producto es un libro, un automóvil o un teléfono?" o "¿Qué categoría de productos es más interesante para este cliente?".

infraestructura mutable

Un modelo que actualiza y modifica la infraestructura existente para las cargas de trabajo de producción. Para mejorar la coherencia, la fiabilidad y la previsibilidad, el AWS Well-Architected Framework recomienda el uso de una infraestructura inmutable como práctica recomendada.

O

OAC

Consulte el control de acceso de origen.

OAI

Consulte la identidad de acceso de origen.

OCM

Consulte gestión del cambio organizacional.

migración fuera de línea

Método de migración en el que la carga de trabajo de origen se elimina durante el proceso de migración. Este método implica un tiempo de inactividad prolongado y, por lo general, se utiliza para cargas de trabajo pequeñas y no críticas.

OI

Consulte integración de operaciones.

OLA

Véase el acuerdo a nivel operativo.

migración en línea

Método de migración en el que la carga de trabajo de origen se copia al sistema de destino sin que se desconecte. Las aplicaciones que están conectadas a la carga de trabajo pueden seguir funcionando durante la migración. Este método implica un tiempo de inactividad nulo o mínimo y, por lo general, se utiliza para cargas de trabajo de producción críticas.

OPC-UA

Consulte Open Process Communications: arquitectura unificada.

Comunicaciones de proceso abierto: arquitectura unificada (OPC-UA)

Un protocolo de comunicación machine-to-machine (M2M) para la automatización industrial. El OPC-UA proporciona un estándar de interoperabilidad con esquemas de cifrado, autenticación y autorización de datos.

acuerdo de nivel operativo (OLA)

Acuerdo que aclara lo que los grupos de TI operativos se comprometen a ofrecerse entre sí, para respaldar un acuerdo de nivel de servicio (SLA).

revisión de la preparación operativa (ORR)

Una lista de preguntas y las mejores prácticas asociadas que le ayudan a comprender, evaluar, prevenir o reducir el alcance de los incidentes y posibles fallos. Para obtener más información, consulte Operational Readiness Reviews (ORR) en AWS Well-Architected Framework.

tecnología operativa (OT)

Sistemas de hardware y software que funcionan con el entorno físico para controlar las operaciones, los equipos y la infraestructura industriales. En la industria manufacturera, la integración de los sistemas de TO y tecnología de la información (TI) es un enfoque clave para las transformaciones de la industria 4.0.

integración de operaciones (OI)

Proceso de modernización de las operaciones en la nube, que implica la planificación de la preparación, la automatización y la integración. Para obtener más información, consulte la Guía de integración de las operaciones.

registro de seguimiento organizativo

Un registro creado por el AWS CloudTrail que se registran todos los eventos para todos Cuentas de AWS los miembros de una organización AWS Organizations. Este registro de seguimiento se crea en cada Cuenta de AWS que forma parte de la organización y realiza un seguimiento de la actividad en cada cuenta. Para obtener más información, consulte Crear un registro para una organización en la CloudTrail documentación.

administración del cambio organizacional (OCM)

Marco para administrar las transformaciones empresariales importantes y disruptivas desde la perspectiva de las personas, la cultura y el liderazgo. La OCM ayuda a las empresas a prepararse para nuevos sistemas y estrategias y a realizar la transición a ellos, al acelerar la adopción de cambios, abordar los problemas de transición e impulsar cambios culturales y organizacionales. En la estrategia de AWS migración, este marco se denomina aceleración de personal, debido a la velocidad de cambio que requieren los proyectos de adopción de la nube. Para obtener más información, consulte la Guía de OCM.

control de acceso de origen (OAC)

En CloudFront, una opción mejorada para restringir el acceso y proteger el contenido del Amazon Simple Storage Service (Amazon S3). El OAC admite todos los buckets de S3 Regiones de AWS, el cifrado del lado del servidor AWS KMS (SSE-KMS) y las solicitudes dinámicas `PUT` y `DELETE` dirigidas al bucket de S3.

identidad de acceso de origen (OAI)

En CloudFront, una opción para restringir el acceso y proteger el contenido de Amazon S3. Cuando utiliza OAI, CloudFront crea un principal con el que Amazon S3 puede autenticarse. Los directores autenticados solo pueden acceder al contenido de un bucket de S3 a través de una distribución específica. CloudFront Consulte también el OAC, que proporciona un control de acceso más detallado y mejorado.

ORR

Consulte la revisión de la preparación operativa.

OT

Consulte la tecnología operativa.

VPC saliente (de salida)

En una arquitectura de AWS cuentas múltiples, una VPC que gestiona las conexiones de red que se inician desde una aplicación. La arquitectura AWS de referencia de seguridad recomienda configurar la cuenta de red con entradas, salidas e inspección VPCs para proteger la interfaz bidireccional entre la aplicación e Internet en general.

P

límite de permisos

Una política de administración de IAM que se adjunta a las entidades principales de IAM para establecer los permisos máximos que puede tener el usuario o el rol. Para obtener más información, consulte Límites de permisos en la documentación de IAM.

información de identificación personal (PII)

Información que, vista directamente o combinada con otros datos relacionados, puede utilizarse para deducir de manera razonable la identidad de una persona. Algunos ejemplos de información de identificación personal son los nombres, las direcciones y la información de contacto.

PII

Consulte la información de identificación personal.

manual de estrategias

Conjunto de pasos predefinidos que capturan el trabajo asociado a las migraciones, como la entrega de las funciones de operaciones principales en la nube. Un manual puede adoptar la forma de scripts, manuales de procedimientos automatizados o resúmenes de los procesos o pasos necesarios para operar un entorno modernizado.

PLC

Consulte controlador lógico programable.

PLM

Consulte la gestión del ciclo de vida del producto.

policy

Un objeto que puede definir los permisos (consulte la política basada en la identidad), especifique las condiciones de acceso (consulte la política basada en los recursos) o defina los permisos máximos para todas las cuentas de una organización AWS Organizations (consulte la política de control de servicios).

persistencia políglota

Elegir de forma independiente la tecnología de almacenamiento de datos de un microservicio en función de los patrones de acceso a los datos y otros requisitos. Si sus microservicios tienen la misma tecnología de almacenamiento de datos, pueden enfrentarse a desafíos de implementación o experimentar un rendimiento deficiente. Los microservicios se implementan más fácilmente y logran un mejor rendimiento y escalabilidad si utilizan el almacén de datos que mejor se adapte a sus necesidades. Para obtener más información, consulte Habilitación de la persistencia de datos en los microservicios.

evaluación de cartera

Proceso de detección, análisis y priorización de la cartera de aplicaciones para planificar la migración. Para obtener más información, consulte la Evaluación de la preparación para la migración.

predicate

Una condición de consulta que devuelve `true` o`false`, por lo general, se encuentra en una cláusula. `WHERE`

pulsar un predicado

Técnica de optimización de consultas de bases de datos que filtra los datos de la consulta antes de transferirlos. Esto reduce la cantidad de datos que se deben recuperar y procesar de la base de datos relacional y mejora el rendimiento de las consultas.

control preventivo

Un control de seguridad diseñado para evitar que ocurra un evento. Estos controles son la primera línea de defensa para evitar el acceso no autorizado o los cambios no deseados en la red. Para obtener más información, consulte Controles preventivos en Implementación de controles de seguridad en AWS.

entidad principal

Una entidad AWS que puede realizar acciones y acceder a los recursos. Esta entidad suele ser un usuario raíz para un Cuenta de AWS rol de IAM o un usuario. Para obtener más información, consulte Entidad principal en Términos y conceptos de roles en la documentación de IAM.

privacidad desde el diseño

Un enfoque de ingeniería de sistemas que tiene en cuenta la privacidad durante todo el proceso de desarrollo.

zonas alojadas privadas

Un contenedor que contiene información sobre cómo desea que Amazon Route 53 responda a las consultas de DNS de un dominio y sus subdominios dentro de uno o más VPCs. Para obtener más información, consulte Uso de zonas alojadas privadas en la documentación de Route 53.

control proactivo

Un control de seguridad diseñado para evitar el despliegue de recursos no conformes. Estos controles escanean los recursos antes de aprovisionarlos. Si el recurso no cumple con el control, significa que no está aprovisionado. Para obtener más información, consulte la guía de referencia de controles en la AWS Control Tower documentación y consulte Controles proactivos en Implementación de controles de seguridad en AWS.

gestión del ciclo de vida del producto (PLM)

La gestión de los datos y los procesos de un producto a lo largo de todo su ciclo de vida, desde el diseño, el desarrollo y el lanzamiento, pasando por el crecimiento y la madurez, hasta el rechazo y la retirada.

entorno de producción

Consulte el entorno.

controlador lógico programable (PLC)

En la fabricación, una computadora adaptable y altamente confiable que monitorea las máquinas y automatiza los procesos de fabricación.

encadenamiento rápido

Utilizar la salida de una solicitud de LLM como entrada para la siguiente solicitud para generar mejores respuestas. Esta técnica se utiliza para dividir una tarea compleja en subtareas o para

refinar o ampliar de forma iterativa una respuesta preliminar. Ayuda a mejorar la precisión y la relevancia de las respuestas de un modelo y permite obtener resultados más detallados y personalizados.

seudonimización

El proceso de reemplazar los identificadores personales de un conjunto de datos por valores de marcadores de posición. La seudonimización puede ayudar a proteger la privacidad personal. Los datos seudonimizados siguen considerándose datos personales.

publish/subscribe (pub/sub)

Un patrón que permite las comunicaciones asíncronas entre microservicios para mejorar la escalabilidad y la capacidad de respuesta. Por ejemplo, en un MES basado en microservicios, un microservicio puede publicar mensajes de eventos en un canal al que se puedan suscribir otros microservicios. El sistema puede añadir nuevos microservicios sin cambiar el servicio de publicación.

Q

plan de consulta

Serie de pasos, como instrucciones, que se utilizan para acceder a los datos de un sistema de base de datos relacional SQL.

regresión del plan de consulta

El optimizador de servicios de la base de datos elige un plan menos óptimo que antes de un cambio determinado en el entorno de la base de datos. Los cambios en estadísticas, restricciones, configuración del entorno, enlaces de parámetros de consultas y actualizaciones del motor de base de datos PostgreSQL pueden provocar una regresión del plan.

R

Matriz RACI

Véase responsable, responsable, consultado, informado (RACI).

RAG

Consulte Retrieval Augmented Generation.

ransomware

Software malicioso que se ha diseñado para bloquear el acceso a un sistema informático o a los datos hasta que se efectúe un pago.

Matriz RASCI

Véase responsable, responsable, consultado, informado (RACI).

RCAC

Consulte control de acceso por filas y columnas.

réplica de lectura

Una copia de una base de datos que se utiliza con fines de solo lectura. Puede enrutar las consultas a la réplica de lectura para reducir la carga en la base de datos principal.

rediseñar

Ver 7 Rs.

objetivo de punto de recuperación (RPO)

La cantidad de tiempo máximo aceptable desde el último punto de recuperación de datos. Esto determina qué se considera una pérdida de datos aceptable entre el último punto de recuperación y la interrupción del servicio.

objetivo de tiempo de recuperación (RTO)

La demora máxima aceptable entre la interrupción del servicio y el restablecimiento del servicio.

refactorizar

Ver 7 Rs.

Región

Una colección de AWS recursos en un área geográfica. Cada uno Región de AWS está aislado e independiente de los demás para proporcionar tolerancia a las fallas, estabilidad y resiliencia. Para obtener más información, consulte Regiones de AWS Especificar qué cuenta puede usar.

regresión

Una técnica de ML que predice un valor numérico. Por ejemplo, para resolver el problema de “¿A qué precio se venderá esta casa?”, un modelo de ML podría utilizar un modelo de regresión lineal para predecir el precio de venta de una vivienda en función de datos conocidos sobre ella (por ejemplo, los metros cuadrados).

volver a alojar

Consulte 7 Rs.

versión

En un proceso de implementación, el acto de promover cambios en un entorno de producción.

trasladarse

Ver 7 Rs.

redefinir la plataforma

Ver 7 Rs.

recompra

Ver 7 Rs.

resiliencia

La capacidad de una aplicación para resistir las interrupciones o recuperarse de ellas. La alta disponibilidad y la recuperación ante desastres son consideraciones comunes a la hora de planificar la resiliencia en el. Nube de AWS Para obtener más información, consulte Nube de AWS Resiliencia.

política basada en recursos

Una política asociada a un recurso, como un bucket de Amazon S3, un punto de conexión o una clave de cifrado. Este tipo de política especifica a qué entidades principales se les permite el acceso, las acciones compatibles y cualquier otra condición que deba cumplirse.

matriz responsable, confiable, consultada e informada (RACI)

Una matriz que define las funciones y responsabilidades de todas las partes involucradas en las actividades de migración y las operaciones de la nube. El nombre de la matriz se deriva de los tipos de responsabilidad definidos en la matriz: responsable (R), contable (A), consultado (C) e informado (I). El tipo de soporte (S) es opcional. Si incluye el soporte, la matriz se denomina matriz RASCI y, si la excluye, se denomina matriz RACI.

control receptivo

Un control de seguridad que se ha diseñado para corregir los eventos adversos o las desviaciones con respecto a su base de seguridad. Para obtener más información, consulte Controles receptivos en Implementación de controles de seguridad en AWS.

retain

Consulte 7 Rs.

jubilarse

Ver 7 Rs.

Generación aumentada de recuperación (RAG)

Tecnología de inteligencia artificial generativa en la que un máster hace referencia a una fuente de datos autorizada que se encuentra fuera de sus fuentes de datos de formación antes de generar una respuesta. Por ejemplo, un modelo RAG podría realizar una búsqueda semántica en la base de conocimientos o en los datos personalizados de una organización. Para obtener más información, consulte Qué es el RAG.

rotación

Proceso de actualizar periódicamente un secreto para dificultar el acceso de un atacante a las credenciales.

control de acceso por filas y columnas (RCAC)

El uso de expresiones SQL básicas y flexibles que tienen reglas de acceso definidas. El RCAC consta de permisos de fila y máscaras de columnas.

RPO

Consulte el objetivo del punto de recuperación.

RTO

Consulte el objetivo de tiempo de recuperación.

manual de procedimientos

Conjunto de procedimientos manuales o automatizados necesarios para realizar una tarea específica. Por lo general, se diseñan para agilizar las operaciones o los procedimientos repetitivos con altas tasas de error.

S

SAML 2.0

Un estándar abierto que utilizan muchos proveedores de identidad (IdPs). Esta función permite el inicio de sesión único (SSO) federado, de modo que los usuarios pueden iniciar sesión AWS

Management Console o llamar a las operaciones de la AWS API sin tener que crear un usuario en IAM para todos los miembros de la organización. Para obtener más información sobre la federación basada en SAML 2.0, consulte Acerca de la federación basada en SAML 2.0 en la documentación de IAM.

SCADA

Consulte el control de supervisión y la adquisición de datos.

SCP

Consulte la política de control de servicios.

secreta

Información confidencial o restringida, como una contraseña o credenciales de usuario, que almacene de forma cifrada. AWS Secrets Manager Se compone del valor secreto y sus metadatos. El valor secreto puede ser binario, una sola cadena o varias cadenas. Para obtener más información, consulta ¿Qué hay en un secreto de Secrets Manager? en la documentación de Secrets Manager.

seguridad desde el diseño

Un enfoque de ingeniería de sistemas que tiene en cuenta la seguridad durante todo el proceso de desarrollo.

control de seguridad

Barrera de protección técnica o administrativa que impide, detecta o reduce la capacidad de un agente de amenazas para aprovechar una vulnerabilidad de seguridad. Existen cuatro tipos principales de controles de seguridad: preventivos, de detección, con capacidad de respuesta y proactivos.

refuerzo de la seguridad

Proceso de reducir la superficie expuesta a ataques para hacerla más resistente a los ataques. Esto puede incluir acciones, como la eliminación de los recursos que ya no se necesitan, la implementación de prácticas recomendadas de seguridad consistente en conceder privilegios mínimos o la desactivación de características innecesarias en los archivos de configuración.

sistema de información sobre seguridad y administración de eventos (SIEM)

Herramientas y servicios que combinan sistemas de administración de información sobre seguridad (SIM) y de administración de eventos de seguridad (SEM). Un sistema de SIEM

recopila, monitorea y analiza los datos de servidores, redes, dispositivos y otras fuentes para detectar amenazas y brechas de seguridad y generar alertas.

automatización de la respuesta de seguridad

Una acción predefinida y programada que está diseñada para responder automáticamente a un evento de seguridad o remediarlo. Estas automatizaciones sirven como controles de seguridad detectables o adaptables que le ayudan a implementar las mejores prácticas AWS de seguridad. Algunos ejemplos de acciones de respuesta automatizadas incluyen la modificación de un grupo de seguridad de VPC, la aplicación de parches a una EC2 instancia de Amazon o la rotación de credenciales.

cifrado del servidor

Cifrado de los datos en su destino, por parte de quien Servicio de AWS los recibe.

política de control de servicio (SCP)

Política que proporciona un control centralizado de los permisos de todas las cuentas de una organización en AWS Organizations. SCPs defina barreras o establezca límites a las acciones que un administrador puede delegar en usuarios o roles. Puede utilizarlas SCPs como listas de permitidos o rechazados para especificar qué servicios o acciones están permitidos o prohibidos. Para obtener más información, consulte las políticas de control de servicios en la AWS Organizations documentación.

punto de enlace de servicio

La URL del punto de entrada de un Servicio de AWS. Para conectarse mediante programación a un servicio de destino, puede utilizar un punto de conexión. Para obtener más información, consulte Puntos de conexión de Servicio de AWS en Referencia general de AWS.

acuerdo de nivel de servicio (SLA)

Acuerdo que aclara lo que un equipo de TI se compromete a ofrecer a los clientes, como el tiempo de actividad y el rendimiento del servicio.

indicador de nivel de servicio (SLI)

Medición de un aspecto del rendimiento de un servicio, como la tasa de errores, la disponibilidad o el rendimiento.

objetivo de nivel de servicio (SLO)

Una métrica objetivo que representa el estado de un servicio, medido mediante un indicador de nivel de servicio.

modelo de responsabilidad compartida

Un modelo que describe la responsabilidad que compartes con respecto a la seguridad y AWS el cumplimiento de la nube. AWS es responsable de la seguridad de la nube, mientras que usted es responsable de la seguridad en la nube. Para obtener más información, consulte el Modelo de responsabilidad compartida.

SIEM

Consulte la información de seguridad y el sistema de gestión de eventos.

punto único de fallo (SPOF)

Una falla en un único componente crítico de una aplicación que puede interrumpir el sistema.

SLA

Consulte el acuerdo de nivel de servicio.

SLI

Consulte el indicador de nivel de servicio.

SLO

Consulte el objetivo de nivel de servicio.

split-and-seed modelo

Un patrón para escalar y acelerar los proyectos de modernización. A medida que se definen las nuevas funciones y los lanzamientos de los productos, el equipo principal se divide para crear nuevos equipos de productos. Esto ayuda a ampliar las capacidades y los servicios de su organización, mejora la productividad de los desarrolladores y apoya la innovación rápida. Para obtener más información, consulte Enfoque gradual para modernizar las aplicaciones en el. Nube de AWS

SPOF

Consulte el punto único de falla.

esquema en forma de estrella

Estructura organizativa de una base de datos que utiliza una tabla de hechos grande para almacenar datos medidos o transaccionales y una o más tablas dimensionales más pequeñas para almacenar los atributos de los datos. Esta estructura está diseñada para usarse en un almacén de datos o con fines de inteligencia empresarial.

patrón de higo estrangulador

Un enfoque para modernizar los sistemas monolíticos mediante la reescritura y el reemplazo gradual de las funciones del sistema hasta que se pueda desmantelar el sistema heredado. Este patrón utiliza la analogía de una higuera que crece hasta convertirse en un árbol estable y, finalmente, se apodera y reemplaza a su host. El patrón fue presentado por Martin Fowler como una forma de gestionar el riesgo al reescribir sistemas monolíticos. Para ver un ejemplo con la aplicación de este patrón, consulte Modernización gradual de los servicios web antiguos de Microsoft ASP.NET (ASMX) mediante contenedores y Amazon API Gateway.

subred

Un intervalo de direcciones IP en la VPC. Una subred debe residir en una sola zona de disponibilidad.

supervisión, control y adquisición de datos (SCADA)

En la industria manufacturera, un sistema que utiliza hardware y software para monitorear los activos físicos y las operaciones de producción.

cifrado simétrico

Un algoritmo de cifrado que utiliza la misma clave para cifrar y descifrar los datos.

pruebas sintéticas

Probar un sistema de manera que simule las interacciones de los usuarios para detectar posibles problemas o monitorear el rendimiento. Puede usar Amazon CloudWatch Synthetics para crear estas pruebas.

indicador del sistema

Una técnica para proporcionar contexto, instrucciones o pautas a un LLM para dirigir su comportamiento. Las indicaciones del sistema ayudan a establecer el contexto y las reglas para las interacciones con los usuarios.

T

etiquetas

Pares clave-valor que actúan como metadatos para organizar los recursos. AWS Las etiquetas pueden ayudarle a administrar, identificar, organizar, buscar y filtrar recursos. Para obtener más información, consulte Etiquetado de los recursos de AWS.

variable de destino

El valor que intenta predecir en el ML supervisado. Esto también se conoce como variable de resultado. Por ejemplo, en un entorno de fabricación, la variable objetivo podría ser un defecto del producto.

lista de tareas

Herramienta que se utiliza para hacer un seguimiento del progreso mediante un manual de procedimientos. La lista de tareas contiene una descripción general del manual de procedimientos y una lista de las tareas generales que deben completarse. Para cada tarea general, se incluye la cantidad estimada de tiempo necesario, el propietario y el progreso.

entorno de prueba

Consulte entorno.

entrenamiento

Proporcionar datos de los que pueda aprender su modelo de ML. Los datos de entrenamiento deben contener la respuesta correcta. El algoritmo de aprendizaje encuentra patrones en los datos de entrenamiento que asignan los atributos de los datos de entrada al destino (la respuesta que desea predecir). Genera un modelo de ML que captura estos patrones. Luego, el modelo de ML se puede utilizar para obtener predicciones sobre datos nuevos para los que no se conoce el destino.

puerta de enlace de tránsito

Un centro de tránsito de red que puede usar para interconectar sus VPCs redes con las locales. Para obtener más información, consulte Qué es una pasarela de tránsito en la AWS Transit Gateway documentación.

flujo de trabajo basado en enlaces troncales

Un enfoque en el que los desarrolladores crean y prueban características de forma local en una rama de característica y, a continuación, combinan esos cambios en la rama principal. Luego, la rama principal se adapta a los entornos de desarrollo, preproducción y producción, de forma secuencial.

acceso de confianza

Otorgar permisos a un servicio que especifique para realizar tareas en su organización AWS Organizations y en sus cuentas en su nombre. El servicio de confianza crea un rol vinculado al servicio en cada cuenta, cuando ese rol es necesario, para realizar las tareas de administración

por usted. Para obtener más información, consulte AWS Organizations Utilización con otros AWS servicios en la AWS Organizations documentación.

ajuste

Cambiar aspectos de su proceso de formación a fin de mejorar la precisión del modelo de ML. Por ejemplo, puede entrenar el modelo de ML al generar un conjunto de etiquetas, incorporar etiquetas y, luego, repetir estos pasos varias veces con diferentes ajustes para optimizar el modelo.

equipo de dos pizzas

Un DevOps equipo pequeño al que puedes alimentar con dos pizzas. Un equipo formado por dos integrantes garantiza la mejor oportunidad posible de colaboración en el desarrollo de software.

U

incertidumbre

Un concepto que hace referencia a información imprecisa, incompleta o desconocida que puede socavar la fiabilidad de los modelos predictivos de ML. Hay dos tipos de incertidumbre: la incertidumbre epistémica se debe a datos limitados e incompletos, mientras que la incertidumbre aleatoria se debe al ruido y la aleatoriedad inherentes a los datos. Para más información, consulte la guía Cuantificación de la incertidumbre en los sistemas de aprendizaje profundo.

tareas indiferenciadas

También conocido como tareas arduas, es el trabajo que es necesario para crear y operar una aplicación, pero que no proporciona un valor directo al usuario final ni proporciona una ventaja competitiva. Algunos ejemplos de tareas indiferenciadas son la adquisición, el mantenimiento y la planificación de la capacidad.

entornos superiores

Ver entorno.

V

succión

Una operación de mantenimiento de bases de datos que implica limpiar después de las actualizaciones incrementales para recuperar espacio de almacenamiento y mejorar el rendimiento.

control de versión

Procesos y herramientas que realizan un seguimiento de los cambios, como los cambios en el código fuente de un repositorio.

Emparejamiento de VPC

Una conexión entre dos VPCs que le permite enrutar el tráfico mediante direcciones IP privadas. Para obtener más información, consulte ¿Qué es una interconexión de VPC? en la documentación de Amazon VPC.

vulnerabilidad

Defecto de software o hardware que pone en peligro la seguridad del sistema.

W

caché caliente

Un búfer caché que contiene datos actuales y relevantes a los que se accede con frecuencia. La instancia de base de datos puede leer desde la caché del búfer, lo que es más rápido que leer desde la memoria principal o el disco.

datos templados

Datos a los que el acceso es infrecuente. Al consultar este tipo de datos, normalmente se aceptan consultas moderadamente lentas.

función de ventana

Función SQL que realiza un cálculo en un grupo de filas que se relacionan de alguna manera con el registro actual. Las funciones de ventana son útiles para procesar tareas, como calcular una media móvil o acceder al valor de las filas en función de la posición relativa de la fila actual.

carga de trabajo

Conjunto de recursos y código que ofrece valor comercial, como una aplicación orientada al cliente o un proceso de backend.

flujo de trabajo

Grupos funcionales de un proyecto de migración que son responsables de un conjunto específico de tareas. Cada flujo de trabajo es independiente, pero respalda a los demás flujos de trabajo del proyecto. Por ejemplo, el flujo de trabajo de la cartera es responsable de priorizar las aplicaciones, planificar las oleadas y recopilar los metadatos de migración. El flujo de trabajo de la cartera entrega estos recursos al flujo de trabajo de migración, que luego migra los servidores y las aplicaciones.

GUSANO

Mira, escribe una vez, lee muchas.

WQF

Consulte el marco AWS de calificación de la carga de trabajo.

escribe una vez, lee muchas (WORM)

Un modelo de almacenamiento que escribe los datos una sola vez y evita que los datos se eliminen o modifiquen. Los usuarios autorizados pueden leer los datos tantas veces como sea necesario, pero no pueden cambiarlos. Esta infraestructura de almacenamiento de datos se considera inmutable.

Z

ataque de día cero

Un ataque, normalmente de malware, que aprovecha una vulnerabilidad de día cero.

vulnerabilidad de día cero

Un defecto o una vulnerabilidad sin mitigación en un sistema de producción. Los agentes de amenazas pueden usar este tipo de vulnerabilidad para atacar el sistema. Los desarrolladores suelen darse cuenta de la vulnerabilidad a raíz del ataque.

aviso de tiro cero

Proporcionar a un LLM instrucciones para realizar una tarea, pero sin ejemplos (imágenes) que puedan ayudar a guiarla. El LLM debe utilizar sus conocimientos previamente entrenados para

realizar la tarea. La eficacia de las indicaciones cero depende de la complejidad de la tarea y de la calidad de las indicaciones. Consulte también las indicaciones de pocos pasos.

aplicación zombi

Aplicación que utiliza un promedio de CPU y memoria menor al 5 por ciento. En un proyecto de migración, es habitual retirar estas aplicaciones.

Las traducciones son generadas a través de traducción automática. En caso de conflicto entre la traducción y la version original de inglés, prevalecerá la version en inglés.