Escalar la infraestructura de Amazon EKS para optimizar la computación, las cargas de trabajo y el rendimiento de la red

AWS Guía prescriptiva

AWS Guía prescriptiva: Escalar la infraestructura de Amazon EKS para optimizar la computación, las cargas de trabajo y el rendimiento de la red

Table of Contents

Escalar la infraestructura de Amazon EKS para optimizar la computación, las cargas de trabajo y el rendimiento de la red

Aniket Dekate, Aniket Kurzadkar e Ishwar Chaauthaiwale, Amazon Web Services ()AWS

Noviembre de 2024 (historial del documento)

Amazon Elastic Kubernetes Service (Amazon EKS) es un servicio de Kubernetes administrado. Con Amazon EKS, puede ejecutar pods de Kubernetes en un entorno de nube contenerizado sin necesidad de instalar y operar su propio plano de control. Con la AWS administración del plano de control, Amazon EKS reduce la administración operativa de la organización. Otros beneficios del uso de Amazon EKS incluyen el escalado, la confiabilidad y la seguridad en el entorno de nube.

Esta guía está diseñada para ayudar a las organizaciones a optimizar su infraestructura de Amazon EKS en las siguientes áreas:

- El escalado informático es un componente fundamental para el rendimiento de las aplicaciones en un entorno dinámico de Kubernetes:
  - Asignación eficiente de recursos: conozca las técnicas para asignar los recursos informáticos de forma dinámica a fin de satisfacer la demanda variable.
  - Herramientas de automatización: obtenga una visión general de las herramientas y los servicios que automatizan el escalado de la computación y reducen la necesidad de intervención manual.
- El escalado de la carga de trabajo ayuda a garantizar que las aplicaciones puedan gestionar diferentes cargas de trabajo sin degradar el rendimiento:
  - Escalador automático de módulos horizontales: analice en profundidad cómo un HPA ayuda a escalar las cargas de trabajo en función de métricas en tiempo real.
  - Escalador automático proporcional de clústeres: descubra cómo el CPA escala automáticamente y mantiene una relación proporcional entre los nodos y las réplicas, aumentando o reduciendo las cargas de trabajo a medida que cambia el tamaño del clúster.
  - Escalado basado en eventos: revise las estrategias para escalar las aplicaciones en respuesta a eventos o factores desencadenantes específicos.
- El escalado de la red ayuda a mantener una comunicación fluida entre los servicios y un flujo de datos eficiente en entornos dinámicos:
  - Complemento CNI de Amazon VPC: descubra cómo el complemento CNI de VPC permite la creación de redes escalables dentro de los clústeres de Amazon EKS.

  - Redes personalizadas: revise la administración de direcciones IP y la segregación del tráfico de red en los clústeres de Amazon EKS.
  - Delegación de prefijos: obtenga información general sobre cómo optimizar la administración de IP en clústeres Amazon EKS grandes y escalables.
  - Amazon VPC Lattice: obtenga información general sobre cómo VPC Lattice puede gestionar redes y VPC cruzadas para lograr un escalado perfecto. service-to-service
- La optimización de costos ayuda a las empresas a ver dónde se gastan sus recursos y a asignar los gastos de manera adecuada a los departamentos o proyectos:
  - Asignar el tamaño adecuado a los recursos: considere técnicas para dimensionar los recursos de la nube de forma adecuada a la carga de trabajo.
  - Supervisión y control de costos: revise las herramientas y las mejores prácticas para rastrear y optimizar los gastos en la nube.

Cada sección se centra en los objetivos particulares que son necesarios para crear un entorno de nube fiable, eficaz y asequible.

Objetivos

Esta guía puede ayudarle a usted y a su organización a alcanzar los siguientes objetivos empresariales:

- Mejora de la eficiencia de los recursos: logre una utilización óptima de los recursos mediante el escalado dinámico de los recursos informáticos, de las cargas de trabajo y de la red en función de las demandas en tiempo real.

  Este objetivo hace hincapié en la importancia de aumentar y reducir los recursos en respuesta a los patrones de uso reales. Herramientas como los escaladores automáticos de cápsulas horizontales y el complemento Amazon VPC CNI ayudan a las organizaciones a utilizar solo los recursos que necesitan, lo que minimiza el desperdicio y maximiza el rendimiento.

- Rendimiento mejorado de las aplicaciones: mantenga un alto rendimiento y capacidad de respuesta de las aplicaciones, incluso con cargas de trabajo y patrones de tráfico fluctuantes.

  Este objetivo se centra en las estrategias que ayudan a garantizar que las aplicaciones puedan gestionar los picos de tráfico y las cargas de trabajo pesadas sin comprometer el rendimiento.

Técnicas como el escalado de la carga de trabajo basado en eventos, la asignación eficiente del cómputo y las arquitecturas de red escalables son fundamentales para lograr este objetivo.

- Escalabilidad perfecta: permita un escalado fluido de los componentes de la infraestructura, lo que permitirá un crecimiento y una adaptación sin esfuerzo a las cambiantes necesidades empresariales.

  La escalabilidad perfecta es crucial para las organizaciones que anticipan el crecimiento o experimentan niveles de tráfico variables. Este objetivo aborda la importancia de implementar soluciones escalables en todos los recursos de cómputo, carga de trabajo y red, de modo que el escalado sea automático, eficiente y transparente.

- Optimización de costes: minimice los costes de la nube y, al mismo tiempo, mantenga o mejore el rendimiento y la escalabilidad.

  La optimización de los costes puede incluir la reducción de los gastos, como el dimensionamiento adecuado de los recursos, el uso de soluciones de escalado rentables y la supervisión de los gastos. El objetivo es equilibrar el ahorro de costes con la necesidad de un alto rendimiento y escalabilidad.

Escalado de computación

El escalado de la computación es un componente fundamental para el rendimiento de las aplicaciones en un entorno dinámico de Kubernetes. Kubernetes reduce el desperdicio mediante el ajuste dinámico de los recursos informáticos (como la CPU y la memoria) en respuesta a la demanda en tiempo real. Esta capacidad ayuda a evitar el aprovisionamiento excesivo o insuficiente, lo que también puede ahorrar gastos operativos. Kubernetes elimina de manera efectiva la necesidad de intervención manual al permitir que la infraestructura se amplíe automáticamente durante las horas pico y disminuya durante los períodos de menor actividad.

El escalado informático general de Kubernetes automatiza el proceso de escalado, lo que aumenta la flexibilidad y la escalabilidad de la aplicación y mejora su comportamiento tolerante a errores. En última instancia, las capacidades de Kubernetes mejoran la excelencia operativa y la productividad.

En esta sección se analizan los siguientes tipos de escalado de cómputo:

- Escalador automático de clústeres
- Escalador automático de clústeres con sobreaprovisionamiento
- Karpenter

Clúster AutoScaler

Según las necesidades de los módulos, la herramienta Cluster Autoscaler modifica automáticamente el tamaño añadiendo nodos cuando es necesario o quitándolos cuando no son necesarios y están infrautilizados.

Considere la herramienta Cluster Autoscaler como una solución de escalado para cargas de trabajo en las que la demanda aumenta gradualmente y la latencia en el escalado no es un problema importante.

La herramienta Cluster Autoscaler ofrece las siguientes funciones clave:

- Escalado: escala los nodos hacia arriba y hacia abajo de forma dinámica en respuesta a las demandas reales de recursos.
- Programación de módulos: ayuda a garantizar que todos los módulos estén operativos y cuenten con los recursos que necesitan para funcionar, lo que evita la escasez de recursos.

- Rentabilidad: elimina los gastos innecesarios de operar nodos infrautilizados al eliminarlos.

Escalador automático de clústeres con sobreaprovisionamiento

El escalador automático de clústeres con sobreaprovisionamiento funciona de manera similar al escalador automático de clústeres, ya que despliega los nodos de manera eficiente y ahorra tiempo al ejecutar módulos de baja prioridad en los nodos. Con esta técnica, el tráfico se redirige a estos módulos en respuesta a picos repentinos de demanda, lo que permite que la aplicación siga funcionando sin interrupciones.

El escalador automático de clústeres con sobreaprovisionamiento ofrece las funciones de los módulos ficticios que se pueden utilizar para implementar y ejecutar nodos fácilmente cuando la carga de trabajo es muy grande, no se necesita latencia y el escalado debe ser rápido.

El escalador automático de clústeres con sobreaprovisionamiento ofrece las siguientes funciones clave:

- Mejor capacidad de respuesta: al hacer que el exceso de capacidad esté constantemente accesible, se tarda menos tiempo en ampliar el clúster en respuesta a los picos de demanda.
- Reserva de recursos: la gestión eficaz de los picos de tráfico inesperados contribuye a una gestión correcta con poco tiempo de inactividad.
- Escalado fluido: minimizar los retrasos en la asignación de recursos facilita un proceso de escalado más fluido.

Karpenter

Karpenter for Kubernetes supera a la herramienta tradicional de escalado automático de clústeres en términos de código abierto, rendimiento y personalización. Con Karpenter, puede lanzar automáticamente solo los recursos informáticos necesarios para gestionar las demandas de su clúster en tiempo real. Karpenter está diseñado para ofrecer un escalado más eficiente y con mayor capacidad de respuesta.

Las aplicaciones con cargas de trabajo extremadamente variables o complejas, en las que las decisiones de escalado rápidas son esenciales, se benefician enormemente del uso de Karpenter. Se integra y ofrece una mejor AWS optimización de la implementación y la selección de nodos.

Karpenter incluye las siguientes características clave:

- Aprovisionamiento dinámico: Karpenter proporciona las instancias y los tamaños correctos para cada propósito y aprovisiona nuevos nodos de forma dinámica en función de los requisitos particulares de los pods.
- Programación avanzada: mediante una ubicación inteligente de los módulos, Karpenter organiza los nodos de manera que los recursos como la GPU, la CPU, la memoria y el almacenamiento se utilicen de la forma más eficaz posible.
- Escalado rápido: Karpenter puede escalar rápidamente y, con frecuencia, reacciona en cuestión de segundos. Esta capacidad de respuesta es útil para los patrones de tráfico repentino o cuando la carga de trabajo exige un escalado inmediato
- Rentabilidad: si elige cuidadosamente la instancia más eficaz, puede reducir los costos operativos y aprovechar las alternativas adicionales de ahorro de costos que ofrecen AWS, como las instancias bajo demanda, las instancias puntuales y las instancias reservadas.

Escalamiento de la carga

El escalado de la carga de trabajo en Kubernetes es esencial para mantener el rendimiento de las aplicaciones y la eficiencia de los recursos en entornos dinámicos. El escalado ayuda a garantizar que las aplicaciones puedan gestionar cargas de trabajo variables sin degradar el rendimiento. Kubernetes ofrece la posibilidad de aumentar o reducir automáticamente los recursos en función de métricas en tiempo real, lo que permite a las organizaciones responder rápidamente a los cambios en el tráfico. Esta elasticidad no solo mejora la experiencia del usuario, sino que también optimiza la utilización de los recursos, lo que ayuda a minimizar los costes asociados a los recursos infrautilizados o sobreaprovisionados.

Además, el escalado efectivo de la carga de trabajo permite una alta disponibilidad, lo que garantiza que las aplicaciones sigan respondiendo incluso durante los períodos de máxima demanda. El escalado de la carga de trabajo en Kubernetes permite a las organizaciones hacer un mejor uso de los recursos de la nube al ajustar la capacidad de forma dinámica para satisfacer las necesidades actuales.

En esta sección se analizan los siguientes tipos de escalado de la carga de trabajo:

- Escalador automático de cápsulas horizontal
- Escalador automático proporcional de clústeres
- Escalador automático basado en eventos basado en Kubernetes

Escalador automático de pods horizontales

El escalador automático de pods horizontal (HPA) es una función de Kubernetes que ajusta automáticamente la cantidad de réplicas de pods en una implementación, controlador de replicación o conjunto con estado, en función del uso observado de la CPU u otras métricas seleccionadas. El HPA garantiza que las aplicaciones puedan gestionar los niveles fluctuantes de tráfico y carga de trabajo sin necesidad de intervención manual. La HPA ofrece un medio para preservar un rendimiento óptimo y, al mismo tiempo, hacer un uso eficaz de los recursos disponibles.

En contextos en los que la demanda de los usuarios puede fluctuar considerablemente con el tiempo, las aplicaciones web, los microservicios y APIs la HPA son especialmente útiles.

El escalador automático de cápsulas horizontales ofrece las siguientes funciones clave:

- Escalado automático: HPA aumenta o reduce automáticamente la cantidad de réplicas de pods en respuesta a las métricas en tiempo real, lo que garantiza que las aplicaciones puedan escalarse para satisfacer la demanda de los usuarios.
- Decisiones basadas en métricas: de forma predeterminada, HPA escala en función del uso de la CPU. Sin embargo, también puede usar métricas personalizadas, como el uso de memoria o métricas específicas de la aplicación, lo que permite estrategias de escalado más personalizadas.
- Parámetros configurables: puede elegir los recuentos mínimo y máximo de réplicas y los porcentajes de utilización deseados, lo que le da autoridad sobre la severidad del escalado.
- Integración con Kubernetes: para monitorear y modificar los recursos, HPA funciona en conjunto con otros elementos del ecosistema de Kubernetes, como el servidor de métricas, la API de Kubernetes y los adaptadores de métricas personalizados.
- Mejor utilización de los recursos: el HPA ayuda a garantizar que los recursos se utilicen de forma eficaz, lo que reduce los costes y mejora el rendimiento, al modificar de forma dinámica la cantidad de módulos.

Escalador automático proporcional de clústeres

El escalador automático proporcional de clústeres (CPA) es un componente de Kubernetes diseñado para ajustar automáticamente la cantidad de réplicas de pods en un clúster en función de la cantidad de nodos disponibles. A diferencia de los escaladores automáticos tradicionales, que se escalan en función de las métricas de uso de los recursos (como la CPU y la memoria), el CPA escala las cargas de trabajo en proporción al tamaño del propio clúster.

Este enfoque es particularmente útil para las aplicaciones que necesitan mantener un cierto nivel de redundancia o disponibilidad en relación con el tamaño del clúster, como CoredNS y otros servicios de infraestructura. Algunos de los principales casos de uso de la CPA son los siguientes:

- Sobreaprovisionamiento
- Amplíe los servicios de la plataforma principal
- Amplíe las cargas de trabajo porque la CPA no requiere un servidor de métricas ni un adaptador Prometheus

Al automatizar el proceso de escalado, la CPA ayuda a las empresas a mantener una distribución equilibrada de la carga de trabajo, a aumentar la eficiencia de los recursos y a garantizar que las aplicaciones se aprovisionen adecuadamente para satisfacer la demanda de los usuarios.

El escalador automático proporcional de clústeres ofrece las siguientes funciones clave:

- Escalado basado en nodos: el CPA escala las réplicas según la cantidad de nodos del clúster que se pueden programar, lo que permite que las aplicaciones se expandan o contraigan en proporción al tamaño del clúster.
- Ajuste proporcional: para garantizar que la aplicación pueda escalar de acuerdo con los cambios en el tamaño del clúster, el escalador automático establece una relación proporcional entre el número de nodos y el número de réplicas. Esta relación se utiliza para calcular el número deseado de réplicas para una carga de trabajo.
- Integración con los componentes de Kubernetes: el CPA funciona con componentes estándar de Kubernetes, como el escalador automático de pods horizontales (HPA), pero se centra específicamente en el recuento de nodos más que en las métricas de utilización de los recursos. Esta integración permite una estrategia de escalado más completa.
- Clientes de API de Golang: para supervisar la cantidad de nodos y sus núcleos disponibles, CPA utiliza clientes de API de Golang que se ejecutan dentro de los pods y se comunican con el servidor de API de Kubernetes.
- Parámetros configurables: con un`ConfigMap`, los usuarios pueden establecer umbrales y parámetros de escalado que CPA utiliza para modificar su comportamiento y asegurarse de que sigue el plan de escalado previsto.

Escalador automático basado en eventos basado en Kubernetes

El escalador automático impulsado por eventos (KEDA) basado en Kubernetes es un proyecto de código abierto que permite que las cargas de trabajo de Kubernetes se escalen en función de la cantidad de eventos que deben procesarse. KEDA mejora la escalabilidad de las aplicaciones al permitirles responder de forma dinámica a las diferentes cargas de trabajo, especialmente a las que se basan en eventos.

Al automatizar el proceso de escalado en función de los eventos, KEDA ayuda a las organizaciones a optimizar la utilización de los recursos, mejorar el rendimiento de las aplicaciones y reducir los costes asociados al sobreaprovisionamiento. Este enfoque es especialmente valioso para las aplicaciones que experimentan patrones de tráfico variables, como los microservicios, las funciones sin servidor y los sistemas de procesamiento de datos en tiempo real.

KEDA ofrece las siguientes funciones clave:

- Escalado basado en eventos: KEDA le permite definir reglas de escalado en función de fuentes de eventos externas, como colas de mensajes, solicitudes HTTP o métricas personalizadas. Esta capacidad ayuda a garantizar que las aplicaciones se escalen en respuesta a la demanda en tiempo real.
- Componente liviano: KEDA es un componente liviano y de un solo propósito que no requiere mucha configuración ni sobrecarga para integrarse fácilmente en los clústeres de Kubernetes existentes.
- Integración con Kubernetes: KEDA amplía las capacidades de los componentes nativos de Kubernetes, como el escalador automático de módulos horizontales (HPA). KEDA agrega capacidades de escalado basadas en eventos a estos componentes, mejorándolos en lugar de reemplazarlos.
- Soporte para múltiples fuentes de eventos: KEDA es compatible con una amplia gama de fuentes de eventos, incluidas las plataformas de mensajería populares como RabbitMQ, Apache Kafka y otras. Gracias a esta adaptabilidad, puede personalizar el escalado para adaptarlo a su arquitectura única basada en eventos.
- Escaladores personalizados: con los escaladores personalizados, puede designar métricas específicas que KEDA puede utilizar para iniciar acciones de escalado en respuesta a requisitos o lógicas empresariales específicos.
- Configuración declarativa: de acuerdo con los principios de Kubernetes, puedes usar KEDA para describir el comportamiento de escalado de forma declarativa utilizando los recursos personalizados de Kubernetes para definir cómo debe realizarse el escalado.

Escalado de red

El escalado de la red en Kubernetes es fundamental para mantener una comunicación fluida entre los servicios y permitir un flujo de datos eficiente en entornos dinámicos. Escalar la infraestructura de red ayuda a garantizar que el clúster pueda gestionar diferentes niveles de tráfico sin sufrir cuellos de botella o problemas de latencia. Kubernetes proporciona herramientas y mecanismos para escalar los recursos de la red, lo que permite a las organizaciones mantener un rendimiento óptimo a medida que cambian los patrones de tráfico.

Esta elasticidad en el escalado de la red mejora la experiencia general del usuario al garantizar conexiones rápidas y confiables. El escalado de la red también optimiza el uso de los recursos de la red, lo que ayuda a reducir los costos asociados a los componentes de la red infrautilizados o sobrecargados.

Además, un escalado efectivo de la red es vital para respaldar la alta disponibilidad y la resiliencia. Al ajustar dinámicamente la capacidad y el enrutamiento de la red, las organizaciones pueden garantizar que los servicios sigan siendo accesibles y responsivos incluso durante los períodos de máxima demanda o picos de tráfico inesperados. Este enfoque permite una mejor utilización de los recursos de red en la nube, lo que garantiza que la infraestructura esté siempre alineada con los requisitos actuales.

En esta sección se analizan los siguientes tipos de escalado de redes:

- Complemento CNI de Amazon VPC para Kubernetes
- Redes personalizadas
- Delegación de prefijos
- Amazon VPC Lattice

Complemento CNI de Amazon VPC para Kubernetes

El complemento Amazon VPC Container Network Interface (CNI) para Kubernetes es un componente fundamental de Amazon EKS. El complemento VPC CNI proporciona capacidades de red avanzadas al integrar los pods de Kubernetes con Amazon VPC. Con este complemento, a cada pod se le asigna una dirección IP única desde la nube privada virtual (VPC), lo que mejora el aislamiento y el rendimiento de la red. A medida que los clústeres crecen y las demandas de la red fluctúan, el

complemento CNI de Amazon VPC desempeña un papel clave a la hora de garantizar operaciones de red eficientes y escalables.

El complemento administra automáticamente la asignación y el enrutamiento de las direcciones IP dentro de la VPC, lo que simplifica la administración de la red y reduce el riesgo de conflictos de IP. Admite funciones como la delegación de prefijos, lo que permite una administración de IP más flexible.

El complemento CNI de VPC ayuda a las organizaciones a optimizar el rendimiento de la red, mejorar la seguridad y reducir el riesgo de agotamiento de la IP. Estas capacidades son especialmente valiosas para entornos dinámicos y de gran escala en los que la demanda de la red fluctúa, como las arquitecturas de microservicios, las cargas de trabajo de alta densidad y las aplicaciones multiusuario.

El complemento CNI de Amazon VPC ofrece las siguientes características clave:

- Redes mejoradas: el complemento CNI de VPC permite que cada pod reciba su propia dirección IP directamente de la VPC, lo que proporciona un aislamiento y un rendimiento de red sólidos. Este enfoque es fundamental para las cargas de trabajo que requieren un alto rendimiento de red y una latencia baja.
- Delegación de prefijos: para superar los problemas de agotamiento de las direcciones IP en clústeres grandes, la delegación de prefijos asigna de forma dinámica bloques más grandes a los nodos, que luego se IPs subdividen para utilizarlos en módulos. Este enfoque garantiza una utilización eficiente de la IP y simplifica el escalado de la red.
- Redes personalizadas: los usuarios pueden configurar interfaces de red personalizadas (ENIs) para los pods, lo que ayuda a distribuir el tráfico de los pods entre varias interfaces, lo que reduce la congestión de la red y mejora la escalabilidad.
- Support for IPv6: al habilitar IPv6 los clústeres de Amazon EKS, los usuarios pueden ampliar significativamente el espacio de direcciones IP disponible, lo que facilita el escalado de aplicaciones distribuidas de gran tamaño sin las restricciones ni las IPv4 limitaciones.
- Integración con Kubernetes: el complemento CNI de VPC funciona a la perfección con los componentes de red de Kubernetes, lo que garantiza que IPs se administren de manera eficiente en todos los pods, servicios y puntos finales externos, y es compatible con funciones avanzadas, como grupos de seguridad para pods.

Redes personalizadas

Las redes personalizadas en Amazon EKS permiten asignar interfaces de red específicas a los pods, lo que proporciona un mayor control sobre la administración de direcciones IP y el tráfico de red. Este enfoque resulta especialmente útil en situaciones en las que el agotamiento de las direcciones IP es motivo de preocupación o cuando es necesario segregar el tráfico de la red por motivos de seguridad, conformidad o rendimiento. Las redes personalizadas ayudan a las organizaciones a administrar de manera eficiente el espacio de direcciones IP, segregar el tráfico y garantizar un rendimiento de red escalable.

Con las redes personalizadas, los administradores pueden administrar los recursos de la red de manera más eficiente. Los administradores pueden usar redes personalizadas para asegurarse de que los pods tengan el aislamiento de red necesario y de que el clúster pueda ampliarse sin limitaciones de direcciones IP.

Las redes personalizadas ofrecen las siguientes funciones clave:

- Administración de IP mejorada: las redes personalizadas permiten asignar interfaces de red específicas (ENIs) a los pods, lo que ayuda a gestionar el agotamiento de las direcciones IP al distribuir el tráfico de los pods entre varios ENIs. Esta capacidad es especialmente importante en clústeres con cargas de trabajo de alta densidad.
- Segregación del tráfico: con las interfaces de red personalizadas, puede separar el tráfico de los pods en función de criterios específicos, como el tipo de aplicación o los requisitos de seguridad. Este enfoque proporciona un mayor control sobre cómo fluye el tráfico dentro y fuera del clúster.
- Soporte para IPv6: las redes personalizadas en Amazon EKS también son compatibles IPv6, lo que ofrece una solución a las limitaciones de IPv4 las direcciones. La red se puede escalar de manera eficiente sin conflictos de direcciones IP, incluso en implementaciones a gran escala.
- Escalabilidad y flexibilidad: a medida que el clúster se amplía, las redes personalizadas permiten la administración dinámica de las interfaces de red. A los nuevos pods se les asignan los recursos de red adecuados sin intervención manual. Este enfoque ayuda a mantener un entorno de red flexible y escalable que puede adaptarse a las cargas de trabajo cambiantes.

Delegación de prefijos

La delegación de prefijos en Kubernetes, especialmente en Amazon EKS, está diseñada para agilizar y optimizar la administración de direcciones IP a medida que los clústeres se escalan. Al asignar

dinámicamente bloques más grandes de direcciones IP (prefijos) a los nodos, la delegación de prefijos reduce el riesgo de agotamiento de la IP y simplifica la administración del espacio IP.

Este enfoque mejora la eficiencia de la red, minimiza la fragmentación y ayuda a que los clústeres se escalen sin problemas sin necesidad de ajustar manualmente el rango de IP. La delegación de prefijos resulta especialmente útil para las implementaciones a gran escala, las cargas de trabajo de alta densidad y los entornos en los que la administración de IP dinámica y flexible es fundamental para mantener el rendimiento y la escalabilidad de la red.

La delegación de prefijos ofrece las siguientes funciones clave:

- Administración eficiente de direcciones IP: la delegación de prefijos permite la asignación dinámica de los rangos de IP, lo que reduce el riesgo de agotamiento de la IP y garantiza un uso eficiente del espacio IP disponible.
- Administración de red simplificada: al permitir que los nodos gestionen sus propias asignaciones de IP, la delegación de prefijos minimiza la fragmentación de la red y simplifica el proceso de enrutamiento, lo que facilita el escalado de los clústeres según sea necesario.
- Support para despliegues a gran escala: en clústeres grandes con cargas de trabajo de alta densidad, la delegación de prefijos permite una escalabilidad perfecta al permitir que nuevos nodos se unan al clúster sin necesidad de ajustar manualmente el rango de IP.

Amazon VPC Lattice

Amazon VPC Lattice permite una service-to-service comunicación eficiente y segura dentro y fuera de ella VPCs, especialmente en arquitecturas de microservicios. VPC Lattice utiliza medidas de seguridad, como grupos de seguridad y listas de control de acceso a la red (red ACLs), además de la integración AWS Identity and Access Management (IAM) para una autenticación de aplicaciones detallada. Un servicio de proxy de capa 7 en el centro de VPC Lattice ofrece conexión, equilibrio de carga, autenticación, autorización, observabilidad, gestión del tráfico y descubrimiento de servicios.

Al simplificar las configuraciones de redes y seguridad, VPC Lattice ayuda a las organizaciones a optimizar la administración del tráfico, mejorar el rendimiento de las aplicaciones y escalar sin problemas entre múltiples y. VPCs Regiones de AWS Esto resulta especialmente valioso para las aplicaciones distribuidas que requieren una red coherente y fiable, como los microservicios, las implementaciones entre regiones y los complejos entornos nativos de la nube.

Amazon VPC Lattice ofrece las siguientes características clave:

- Service-to-service redes: VPC Lattice simplifica la configuración de redes y seguridad entre los servicios dentro de una arquitectura de microservicios. Proporciona una plataforma unificada para administrar la comunicación, de modo que los servicios pueden escalarse de forma independiente y, al mismo tiempo, mantener un alto rendimiento y seguridad.
- Redes entre VPC: VPC Lattice es crucial para administrar el tráfico en varias regiones. VPCs Proporciona un marco de red coherente que permite que los servicios se comuniquen sin problemas, independientemente de su ubicación física. Esta capacidad es particularmente importante para aplicaciones a gran escala que abarcan regiones múltiples VPCs o geográficas.
- Gestión de seguridad mejorada: al integrar las políticas de seguridad directamente en la capa de red, VPC Lattice permite una service-to-service comunicación segura y eficiente. Esta función reduce la complejidad de la administración de la seguridad en un entorno distribuido, lo que permite un escalado más sencillo y una reducción de los gastos operativos.
- Administración del tráfico simplificada: VPC Lattice ofrece funciones avanzadas de administración del tráfico, que incluyen mecanismos de enrutamiento, equilibrio de carga y conmutación por error. Con estas funciones, el tráfico se distribuye de manera eficiente entre los servicios, lo que optimiza el rendimiento de la red y mejora la escalabilidad de la aplicación.

Optimización de costos

Para respaldar un control efectivo de los recursos, la minimización de los costos de Kubernetes es crucial para las empresas que utilizan esta tecnología de organización de contenedores. Resulta difícil realizar un seguimiento adecuado del gasto en los entornos de Kubernetes debido a su complejidad, que incluye varios componentes, como módulos y nodos. Mediante la aplicación de técnicas de optimización de costes, las empresas pueden ver en qué se gastan sus recursos y asignar los gastos de forma adecuada a los departamentos o proyectos.

Si bien el escalamiento dinámico tiene ventajas, si no se gestiona adecuadamente, puede generar gastos imprevistos. La gestión eficiente de los costes ayuda a asignar los recursos solo cuando realmente se necesitan, lo que evita un aumento imprevisto de los gastos.

En esta sección se analizan los siguientes enfoques para la optimización de costos:

- Kubecost
- Ricitos de oro
- AWS Fargate
- Spot Instances
- Instancias reservadas
- AWS Instancias de gravitón

Kubecost

Kubecost es una solución de administración de costos que ayuda a las empresas a rastrear, controlar y maximizar sus gastos en infraestructura en la nube. Está diseñada específicamente para los clústeres de Kubernetes. Kubecost le brinda información sobre la utilización de los recursos y un conocimiento de los costos en tiempo real, lo que le permite comprender mejor dónde y qué cantidad de sus recursos en la nube se utilizan. Con esta información, puede optimizar su gasto en infraestructura, mejorar la eficiencia de los recursos y tomar decisiones más informadas sobre sus inversiones en la nube.

Kubecost ofrece las siguientes funciones clave:

- Asignación de costos: Kubecost ofrece una asignación exhaustiva de los costos de los recursos de Kubernetes, incluidas las cargas de trabajo, los servicios, los espacios de nombres y las etiquetas. Esta función ayuda a los equipos a monitorear los costos por entorno, proyecto o equipo.
- Supervisión de costes en tiempo real: ofrece una supervisión en tiempo real de los costes de la nube, lo que proporciona a las organizaciones información inmediata sobre los patrones de gasto y ayuda a evitar sobrecostes inesperados.
- Recomendaciones de optimización: Kubecost ofrece sugerencias prácticas para minimizar la utilización de los recursos, como reducir los recursos inactivos, ajustar el tamaño de las cargas de trabajo y maximizar los gastos de almacenamiento.
- Elaboración de presupuestos y alertas: los usuarios de Kubecost pueden crear presupuestos y recibir recordatorios cuando un gasto se acerque o supere los criterios predeterminados. Esta función ayuda a los equipos a cumplir con las restricciones financieras.

Ricitos de oro

Goldilocks es una utilidad de Kubernetes diseñada para ayudar a los usuarios a optimizar sus solicitudes de recursos y sus límites para las cargas de trabajo de Kubernetes. Proporciona recomendaciones sobre cómo configurar los recursos de CPU y memoria para los contenedores que se ejecutan en un clúster de Kubernetes. Estas recomendaciones le ayudan a asegurarse de que las aplicaciones tienen la cantidad adecuada de recursos para funcionar de manera eficiente y sin desperdicios. Esta optimización puede suponer un ahorro de costes, una mejora del rendimiento y un uso más eficiente de los clústeres de Kubernetes.

Goldilocks ofrece las siguientes funciones clave:

- Recomendaciones de recursos: Goldilocks determina la configuración ideal para las solicitudes y restricciones de recursos mediante el análisis de las estadísticas anteriores de consumo de CPU y memoria de las cargas de trabajo de Kubernetes. De este modo, es más fácil evitar el aprovisionamiento insuficiente o excesivo, lo que puede provocar problemas de rendimiento y desperdicio de recursos.
- Integración con VPA: Goldilocks aprovecha el escalador automático de pods verticales (VPA) de Kubernetes para recopilar datos y ofrecer recomendaciones. Se ejecuta en un «modo de recomendación», lo que significa que en realidad no cambia la configuración de los recursos, sino que ofrece orientación sobre cuáles deberían ser esas configuraciones.

- Análisis basado en el espacio de nombres: Goldilocks le permite regular con precisión qué cargas de trabajo se optimizan y supervisan, ya que le permite centrarse en espacios de nombres específicos para su análisis.
- Panel visual: el panel basado en la web muestra visualmente las solicitudes de recursos y las restricciones sugeridas, lo que facilita la comprensión de los datos y la adopción de medidas al respecto.
- Funcionamiento no intrusivo: Goldilocks no altera la configuración del clúster porque funciona en modo de recomendación. Si lo desea, puede aplicar manualmente la configuración de recursos recomendada después de revisar las recomendaciones.

AWS Fargate

En el contexto de Amazon EKS, https://docs.aws.amazon.com/eks/latest/userguide/fargate.htmlAWS Fargate le permite ejecutar pods de Kubernetes sin administrar las instancias de Amazon subyacentes. EC2 Se trata de un motor de cómputo sin servidor que le permite centrarse en implementar y escalar aplicaciones en contenedores sin preocuparse por la infraestructura.

AWS Fargate ofrece las siguientes funciones clave:

- Sin administración de infraestructura: Fargate elimina la necesidad de aprovisionar, administrar o escalar EC2 instancias de Amazon o nodos de Kubernetes. AWS gestiona toda la gestión de la infraestructura, incluidos los parches y el escalado.
- Aislamiento a nivel de pod: a diferencia de los nodos de trabajo que se basan en Amazon, EC2 Fargate proporciona aislamiento a nivel de tareas o de pod. Cada pod se ejecuta en su propio entorno informático aislado, lo que mejora la seguridad y el rendimiento.
- Escalado automático: Fargate escala automáticamente los pods de Kubernetes en función de la demanda. No es necesario gestionar las políticas de escalado ni los grupos de nodos.
- Facturación por segundo: solo paga por los recursos de vCPU y memoria que consuma cada pod durante el tiempo exacto de ejecución, lo que constituye una opción rentable para determinadas cargas de trabajo.
- Reducción de los gastos generales: al eliminar la necesidad de administrar EC2 instancias, Fargate le permite centrarse en crear y administrar sus aplicaciones en lugar de en las operaciones de infraestructura.

Spot Instances

Las instancias puntuales ofrecen ahorros significativos en comparación con los precios de las instancias bajo demanda y son una opción asequible para ejecutar EC2 nodos de trabajo de Amazon en un clúster de Amazon EKS. Sin embargo, AWS pueden interrumpir las instancias puntuales en caso de que se necesite capacidad de instancia bajo demanda. AWS pueden recuperar instancias puntuales con un aviso de 2 minutos cuando se necesita la capacidad necesaria, lo que las hace menos fiables para cargas de trabajo críticas e informales.

Para las cargas de trabajo sensibles a los costes y que pueden soportar interrupciones, las instancias puntuales de Amazon EKS son una buena opción. El uso de una combinación de instancias puntuales e instancias bajo demanda en un clúster de Kubernetes le ayuda a ahorrar dinero sin sacrificar la disponibilidad de las cargas de trabajo esenciales.

Las instancias puntuales ofrecen las siguientes funciones clave:

- Ahorro de costes: las instancias puntuales pueden ser más económicas que las instancias bajo demanda, lo que las hace ideales para cargas de trabajo sensibles a los costes.
- Ideal para cargas de trabajo tolerantes a errores: ideal para cargas de trabajo sin estado y tolerantes a errores, como el procesamiento por lotes, los trabajos de CI/CD, el aprendizaje automático o el procesamiento de datos a gran escala, donde las instancias se pueden reemplazar sin grandes interrupciones.
- Integración de grupos con escalado automático: Amazon EKS integra las instancias puntuales con el escalador automático de clústeres de Kubernetes, que puede sustituir automáticamente los nodos de instancias puntuales interrumpidos por otras instancias puntuales o instancias bajo demanda disponibles.

instancias reservadas

En Amazon EKS, Reserved Instances es un modelo de precios para los EC2 nodos de trabajo de Amazon que ejecutan las cargas de trabajo de Kubernetes. Al usar instancias reservadas, se compromete a usar tipos de instancias específicos durante un período de 1 o 3 años, a cambio de ahorrar costos en comparación con los precios de las instancias bajo demanda. La reserva de instancias en Amazon EKS es una forma asequible de realizar cargas de trabajo consistentes y a largo plazo en los EC2 nodos de trabajo de Amazon.

Las instancias reservadas se utilizan habitualmente en Amazon EC2. Sin embargo, los nodos de trabajo de su clúster de Amazon EKS (que son EC2 instancias) también pueden beneficiarse de este modelo de ahorro de costes, siempre que la carga de trabajo requiera un uso predecible y a largo plazo.

Los servicios de producción, las bases de datos y otras aplicaciones activas que necesitan una alta disponibilidad y un rendimiento uniforme son ejemplos de cargas de trabajo estables que son ideales para las instancias reservadas.

Las instancias reservadas ofrecen las siguientes funciones clave:

- Ahorro de costos: las instancias reservadas ofrecen ahorros en comparación con las instancias bajo demanda, según la duración (1 o 3 años) y el plan de pago (todo por adelantado, con pago parcial o sin pago inicial).
- Compromiso a largo plazo: se compromete a un plazo de 1 o 3 años para un tipo, tamaño y tamaño de instancia específicos. Región de AWS Esto es ideal para cargas de trabajo estables y que se ejecutan de forma continua a lo largo del tiempo.
- Precios predecibles: dado que se compromete a cumplir un plazo específico, las instancias reservadas ofrecen costos mensuales o iniciales predecibles, lo que facilita la presupuestación para las cargas de trabajo a largo plazo.
- Flexibilidad de instancias: con las instancias reservadas convertibles, puede cambiar el tipo, la familia o el tamaño de las instancias durante el período de reserva. Las instancias reservadas convertibles ofrecen más flexibilidad que las instancias reservadas estándar, que no permiten cambios.
- Capacidad garantizada: las instancias reservadas garantizan que la capacidad esté disponible en la zona de disponibilidad en la que se realiza la reserva, lo cual es crucial para las cargas de trabajo críticas que necesitan una potencia informática constante.
- Sin riesgo de interrupción: a diferencia de las instancias puntuales, las instancias reservadas no están sujetas a ninguna interrupción. AWS Esto las hace ideales para ejecutar cargas de trabajo de misión crítica que requieren un tiempo de actividad garantizado.

AWS Instancias de Graviton

AWS Graviton es una familia de procesadores basados en ARM diseñada para mejorar el rendimiento y AWS la rentabilidad de las cargas de trabajo en la nube. En el contexto de Amazon

EKS, puede usar las instancias de Graviton como nodos de trabajo para ejecutar sus cargas de trabajo de Kubernetes, lo que ofrece importantes mejoras de rendimiento y ahorros de costos.

Las instancias Graviton son una excelente opción para las aplicaciones nativas de la nube y las que requieren un uso intensivo de cómputo, ya que ofrecen una relación precio-rendimiento superior a la de las instancias x86. Sin embargo, cuando se plantee adoptar instancias Graviton, tenga en cuenta la compatibilidad con ARM.

AWS Las instancias Graviton ofrecen las siguientes funciones clave:

- Arquitectura basada en ARM: los procesadores AWS Graviton se basan en la arquitectura ARM, que es diferente de las arquitecturas x86 tradicionales, pero muy eficiente para muchas cargas de trabajo.
- Rentables: las EC2 instancias de Amazon basadas en Graviton suelen ofrecer una mejor relación precio-rendimiento en comparación con las instancias basadas en x86. EC2 Esto los convierte en una opción atractiva para los clústeres de Kubernetes que ejecutan Amazon EKS.
- Rendimiento: los procesadores Graviton2, la segunda generación de AWS Graviton, ofrecen mejoras significativas en términos de rendimiento informático, rendimiento de la memoria y eficiencia energética. Son ideales para cargas de trabajo con un uso intensivo de la CPU y de la memoria.
- Diversos tipos de instancias: las instancias de Graviton vienen en varias familias, como las t4g, m7g, c7g y r7g, y abarcan una variedad de casos de uso, desde cargas de trabajo de uso general hasta cargas de trabajo optimizadas para la computación, optimizadas para la memoria y estables.
- Grupos de nodos de Amazon EKS: puede configurar grupos de nodos gestionados por Amazon EKS o grupos de nodos autogestionados para incluir instancias basadas en Graviton. Con este enfoque, puede ejecutar cargas de trabajo optimizadas para la arquitectura ARM en el mismo clúster de Kubernetes junto con instancias basadas en x86.

Pasos a seguir a continuación

Esta guía proporciona información que le ayudará a optimizar Amazon EKS con respecto al escalado de cómputo, escalado de carga de trabajo, escalado de red y optimización de costos. Al comprender y aplicar estos conceptos, las organizaciones pueden lograr un entorno de nube altamente eficiente, escalable y rentable que satisfaga sus necesidades dinámicas.

La implementación efectiva del escalado de la carga de trabajo y la computación ayuda a garantizar que los recursos se utilicen de manera eficiente y que las aplicaciones mantengan un alto rendimiento incluso durante las horas punta. La adopción de técnicas de escalado de redes, como las redes personalizadas y la delegación de prefijos, facilita la administración de los recursos de la red y permite una escalabilidad perfecta. Hacer hincapié en la optimización de los costes ayuda a las organizaciones a equilibrar el rendimiento con la eficiencia financiera.

La integración de esta guía en su estrategia de nube puede ayudarlo a mejorar el rendimiento y la escalabilidad de su infraestructura y a ahorrar costos. Este enfoque integral puede permitirle crear un entorno de nube sólido que respalde el crecimiento de su organización y se adapte a las cambiantes demandas empresariales.

Recursos

AWS blogs

- Aprovechando la optimización de costes y la resiliencia de EKS con instancias puntuales
- Combinación de AWS Graviton con x86 CPUs para optimizar los costes y la resiliencia mediante Amazon EKS

AWS documentación

- CNI de Amazon VPC
- Amazon Elastic Kubernetes Service AWS (documento técnico: Descripción general de las opciones de implementación activadas) AWS
- Guía de prácticas recomendadas de Amazon EKS
- Karpenter
- Más información sobre Kubecost
- Simplifique la administración de la computación con AWS Fargate

Otros recursos

- Escalado automático de clústeres (documentación de Kubernetes)
- Goldilocks: una herramienta de código abierto para recomendar solicitudes de recursos (blog de Fairwinds)
- Escalado automático de pods horizontales (documentación de Kubernetes)
- Kubecost (documentación de Kubecost)
- Escalado automático basado en eventos de Kubernetes (documentación de KEDA)

Historial de documentos

En la siguiente tabla se describen los cambios importantes en esta guía, Escalar la infraestructura de Amazon EKS para optimizar el procesamiento, las cargas de trabajo y el rendimiento de la red. Si quiere recibir notificaciones de futuras actualizaciones, puede suscribirse a las notificaciones RSS.

Cambio	Descripción	Fecha
Publicación inicial	—	11 de noviembre de 2024

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:

- Refactor/re-architect — Mueva una aplicación y modifique su arquitectura aprovechando al máximo las funciones nativas de 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 PostgreSQL-Compatible edición 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: Migrar la base de datos Oracle en las instalaciones a Amazon Relational Database Service (Amazon RDS) para Oracle en la nube de 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 hacer cambios para aprovechar las funcionalidades de la nube. Ejemplo: Migrar la base de datos de Oracle en las instalaciones a Oracle en una instancia de EC2 en la 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 en las instalaciones 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

A2A () Agent-to-Agent

Un protocolo completo para la colaboración entre agentes que facilita la delegación de tareas y la transferencia de estados.

ABAC

Consulte control de acceso basado en atributos.

servicios abstractos

Consulte servicios administrados.

ACID

Consulte atomicidad, consistencia, aislamiento, 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 una 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 de origen gestiona las transacciones de las aplicaciones conectadas, mientras los datos se replican en la de destino. La base de datos de destino no acepta ninguna transacción durante la migración.

Agente

Un sistema de IA que puede razonar, planificar y tomar medidas de forma autónoma utilizando herramientas para alcanzar los objetivos.

Agent Ops

Prácticas operativas para crear, probar, implementar y ejecutar agentes de IA en producción a escala.

función de agregación

Función SQL que actúa en un grupo de filas y calcula un único valor de devolución para el grupo. Entre los ejemplos de funciones de agregación se incluyen `SUM` y `MAX`.

IA

Consulte inteligencia artificial.

AIOps

Consulte 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

Enfoque de seguridad que permite usar de manera exclusiva 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 se utiliza AIOps en la estrategia de migración de AWS , 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

bot malicioso

Bot destinado a causar interrupciones o daños a personas u organizaciones.

BCP

Consulte 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. Consulte también endianidad.

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.

blue/green despliegue

Estrategia de implementación 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 se ejecuta en el otro entorno (verde). Esta estrategia lo ayuda a hacer reversiones rápidas 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 la información de Internet. Otros bots, conocidos como bots maliciosos, tienen como objetivo causar interrupciones o daños a personas u organizaciones.

botnet

Redes de bots infectadas por malware y que están bajo el control de una sola parte, conocida como pastor de bots 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 de emergencia

En circunstancias excepcionales y mediante un proceso aprobado, es una forma rápida de que un usuario pueda acceder a un Cuenta de AWS sitio al que normalmente no tiene permisos de acceso. Para obtener más información, consulte el indicador de implementación de procedimientos rompe-cristales en la AWS Well-Architected guía.

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 AWS Cloud Adoption Framework.

implementación canario

Lanzamiento lento e incremental de una versión para los usuarios finales. Cuando tenga mayor confianza en la nueva versión, la implementa y reemplaza la versión actual en su totalidad.

CCoE

Consulte Centro de excelencia en la nube.

CDC

Consulte 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

Introducción intencionada de 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 integración continua y 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.

Desarrollador ciudadano

Un usuario empresarial que crea aplicaciones de IA utilizando plataformas sin code/low código sin conocimientos técnicos especializados.

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 CCoE 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 relacionada con la tecnología de computación de periferia.

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 siguientes son las cuatro fases por las que suelen pasar las empresas cuando migran a la Nube de AWS:

- Proyecto: ejecución de algunos proyectos relacionados con la nube con fines de prueba de concepto y aprendizaje
- Fundamento: realización de inversiones fundamentales para escalar la adopción de la nube (p. ej., crear una zona de aterrizaje, definir un CCoE, establecer un modelo de operaciones)
- Migración: migración de aplicaciones individuales
- Re-invention — Optimizar los productos y servicios e innovar en la nube

Stephen Orban definió estas etapas en la entrada del blog The Journey Toward Cloud-First & the Stages of Adoption del 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 base de datos de administración de 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. Algunos repositorios en la nube comunes son GitHub o Bitbucket Cloud. Cada versión del código se denomina rama. En una estructura de microservicios, cada repositorio se encuentra dedicado a una única funcionalidad. Una sola CI/CD canalización 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 machine learning para analizar y extraer información de formatos visuales, como imágenes y videos digitales. Por ejemplo, Amazon SageMaker AI proporciona algoritmos de procesamiento de imágenes para CV.

deriva de configuración

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

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

Un conjunto de AWS Config reglas y medidas correctivas que puede reunir para personalizar sus controles 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, creación, prueba, puesta en escena y producción del proceso de publicación del software. CI/CD se describe comúnmente como una canalización. CI/CD puede ayudarlo a automatizar los procesos, mejorar la productividad, mejorar

la calidad del código y entregar más rápido. 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

Consulte 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 los datos es un componente del pilar de seguridad del AWS Well-Architected Framework. Para obtener más información, consulte Clasificación de datos.

deriva 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 deriva de 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

Marco de arquitectura que proporciona una propiedad de datos distribuida y descentralizada con una administración y una gobernanza centralizadas.

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

Sistema de administración de datos que respalda la inteligencia empresarial, como los análisis. Los almacenes de datos suelen contener grandes cantidades de datos históricos y, por lo general, se utilizan para las consultas y los 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 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.

defensa en profundidad

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 enfoque de defensa en profundidad podría combinar la autenticación multifactor, 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 en 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 suelen utilizar para restringir consultas, filtrarlas y etiquetar los 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)

Estrategia y proceso que utiliza para minimizar el tiempo de inactividad y la pérdida de datos a causa de un desastre. Para obtener más información, consulte Recuperación de cargas de trabajo ante desastres en AWS: Recuperación en la nube en el AWS Well-Architected marco.

DML

Consulte 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. Eric Evans introdujo este concepto en su libro Domain-Driven Design: Tackling Complexity in the Heart of 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 ASP.NET Microsoft (ASMX) mediante contenedores y Amazon API Gateway.

DR

Consulte recuperación ante desastres.

Detección de desviaciones

Seguimiento de las desviaciones con respecto a una configuración con línea de base. 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 asignación de flujos de valor para el desarrollo.

E

EDA

Consulte análisis de datos de tipo exploratorio.

EDI

Consulte 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 la nube, la computación de periferia puede reducir la latencia de la comunicación y mejorar el tiempo de respuesta.

intercambio electrónico de datos (EDI)

Intercambio automatizado de documentos comerciales entre organizaciones. Para más información, consulte ¿Qué es el intercambio electrónico de datos?

cifrado

Proceso de computación que transforma datos de texto plano, que son 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. Big-endian los sistemas almacenan primero el byte más significativo. Little-endian los sistemas almacenan primero el byte menos significativo.

punto de conexión

Consulte punto de conexión de 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 con AWS PrivateLink entidades principales Cuentas de AWS o AWS Identity and Access Management (de IAM) y conceder permisos a ellas. 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)

Sistema que automatiza y administra los procesos empresariales clave (como la contabilidad, MES y la administración 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 un CI/CD proceso, 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 hechos

Tabla central de un esquema en estrella. Almacena datos cuantitativos sobre operaciones empresariales. Por lo general, 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.

Fail Fast

Filosofía que utiliza pruebas frecuentes e incrementales para reducir el ciclo de vida del desarrollo. Es una parte fundamental de los enfoques ágiles.

límite de aislamiento de errores

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 más información, consulte AWS Fault Isolation Boundaries.

rama de característica

Consulte rama.

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.

peticiones con pocos pasos

Proporcionar a un LLM una pequeña cantidad de ejemplos que demuestren la tarea y el resultado deseado antes de pedirle que lleve a cabo una tarea similar. Esta técnica es una aplicación del aprendizaje contextual, en el que los modelos aprenden a partir de ejemplos (tomas) integrados en las instrucciones. Few-shot Las indicaciones pueden ser eficaces para tareas que requieren

un formato, un razonamiento o un conocimiento del dominio específicos. Consulte también peticiones desde cero.

FGAC

Consulte 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 de cambio 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 modelo fundacional.

Modelo fundacional (FM)

Gran red neuronal de aprendizaje profundo que se entrenó con conjuntos de datos masivos de datos generalizados y no etiquetados. Los FM pueden hacer una amplia variedad de tareas generales, como comprender el lenguaje, generar texto e imágenes y conversar en lenguaje natural. Para más información, consulte ¿Qué son los modelos fundacionales?

Puerta de enlace FM

Un intermediario centralizado que controla y normaliza el acceso a los modelos básicos. También se conoce como puerta de enlace LLM.

G

IA generativa

Subconjunto de modelos de IA que se entrenaron con grandes cantidades de datos y que pueden utilizar una simple petición de texto para crear contenido y artefactos nuevos, como imágenes, videos, texto y audio. Para más información, consulte ¿Qué es la IA generativa?

bloqueo geográfico

Consulte 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, mientras que el flujo de trabajo basado en enlaces troncales es el enfoque moderno preferido.

imagen dorada

Instantánea de un sistema o software que se usa 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 la conformidad en todas las unidades organizativas (OU). 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 CSPM GuardDuty AWS Trusted Advisor, Amazon Inspector y AWS Lambda cheques personalizados.

barandas (AI)

Mecanismos de seguridad que filtran, validan y restringen las entradas y salidas de los agentes para ayudar a garantizar un comportamiento responsable y seguro de la IA.

H

HA

Consulte 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 de reserva

Parte de los datos históricos etiquetados que se ocultan de un conjunto de datos que se utiliza para entrenar un modelo de machine learning. Puede utilizar los datos de reserva para evaluar el rendimiento del modelo mediante la comparación de las predicciones del modelo con los datos de reserva.

human-in-the-loop (HiTL)

Un patrón de flujo de trabajo en el que la ejecución de los agentes se detiene para su revisión y aprobación por parte de una persona en los puntos de decisión críticos.

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, una revisión suele realizarse fuera del flujo de trabajo habitual de las DevOps versiones.

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

Consulte infraestructura como código.

políticas basadas en identidades

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

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

Modelo que implementa una nueva infraestructura para las cargas de trabajo de producción en lugar de actualizar o modificar la infraestructura existente o aplicarle revisiones. Las infraestructuras inmutables son de manera intrínseca más coherentes, fiables y predecibles que las infraestructuras mutables. Para obtener más información, consulte las mejores prácticas del Framework para implementar con una infraestructura inmutable. AWS Well-Architected

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 de referencia de seguridad de AWS recomienda configurar su cuenta de red con VPC entrantes, salientes y de inspección para proteger la interfaz bidireccional entre su aplicación e Internet en general.

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. AI/ML

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 (IIoT)

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 del Internet de las cosas industrial (IIoT).

VPC de inspección

En una arquitectura de AWS cuentas múltiples, una VPC centralizada que gestiona las inspecciones del tráfico de red entre las VPC (iguales o Regiones de AWS diferentes), Internet y las redes locales. La Arquitectura de referencia de seguridad de AWS recomienda configurar su cuenta de red con VPC entrantes, salientes y de inspección para proteger la interfaz bidireccional entre su 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 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 de gran tamaño (LLM)

: Modelo de IA de aprendizaje profundo que se entrenó previamente con una gran cantidad de datos. Un LLM puede llevar a cabo varias tareas, como responder preguntas, resumir documentos, traducir textos a otros idiomas y completar oraciones. Para más información, consulte ¿Qué es un LLM (modelo de lenguaje de gran tamaño)?

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

: Consulte Las 7 R.

sistema little-endian

: Un sistema que almacena primero el byte menos significativo. Consulte también endianidad.

LLM

Consulte modelo de lenguaje de gran tamaño.

entornos inferiores

Consulte 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

Consulte rama.

malware

Software diseñado para comprometer la seguridad o la privacidad de la computadora. El malware podría 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 administrados

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

sistema de ejecución de fabricación (MES)

Sistema de software para seguir, supervisar, documentar y controlar los procesos de producción que convierten las materias primas en productos acabados en la zona de producción.

MAP

Consulte Programa de aceleración de la migración.

MCP

Consulte Model Context Protocol.

Protocolo de contexto para modelos (MCP)

Un protocolo sin estado para la comunicación entre el agente y la herramienta.

Servidor MCP

Un servicio que expone una o más herramientas a través del protocolo Model Context.

mecanismo

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

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 sistema de ejecución de fabricación.

Message Queuing Telemetry Transport (MQTT)

Un protocolo de comunicación ligero de máquina a máquina (M2M), basado en el publish/subscribepatrón, para dispositivos de IoT con recursos limitados.

microservicio

Un servicio pequeño e independiente que se comunica a través de API bien definidas 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 servicios sin servidor. AWS

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 API ligeras. 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

Cross-functional equipos que agilizan la migración de las cargas de trabajo mediante enfoques ágiles y automatizados. Los equipos de las fábricas de migración suelen estar compuestos por analistas y propietarios de operaciones, 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: rehospede la migración a Amazon EC2 AWS con Application Migration Service.

Migration Portfolio Assessment (MPA)

Herramienta en línea que proporciona información a fin de validar los argumentos comerciales necesarios para migrar a la 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 de los socios 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

Enfoque utilizado para migrar una carga de trabajo a la Nube de AWS. Para más información, consulte la entrada Las 7 R de este glosario y también Mobilize your organization to accelerate large-scale migrations.

ML

Consulte machine learning.

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 más información, consulte Strategy for modernizing applications in the 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 más información, consulte Evaluating modernization readiness for applications in the 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.

MPA

Consulte Migration Portfolio Assessment.

MQTT

Consulte Message Queuing Telemetry Transport.

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

Modelo que actualiza y modifica la infraestructura actual para las cargas de trabajo de producción. Para mejorar la coherencia, la confiabilidad y la previsibilidad, el AWS Well-Architected Marco recomienda el uso de una infraestructura inmutable como práctica recomendada.

O

OAC

Consulte control de acceso de origen.

OAI

Consulte identidad de acceso de origen.

OCM

Consulte administración 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

Consulte acuerdo de 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 de máquina a máquina (M2M) para la automatización industrial. 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)

Lista de comprobación de preguntas y prácticas recomendadas asociadas que son útiles para comprender, evaluar, prevenir o reducir el alcance de los incidentes y posibles errores. Para

obtener más información, consulte las revisiones de preparación operativa (ORR) en el AWS Well-Architected marco.

tecnología operativa (TO)

Sistemas de hardware y software que funcionan con el entorno físico para controlar las operaciones, los equipos y la infraestructura industriales. En el sector de la fabricación, 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 y AWS CloudTrail que registra todos los eventos Cuentas de AWS 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 con AWS KMS (SSE-KMS) y `DELETE` las solicitudes `PUT` y dinámicas 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 revisión de la preparación operativa.

OT

Consulte 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 de referencia de seguridad de AWS recomienda configurar su cuenta de red con VPC entrantes, salientes y de inspección para proteger la interfaz bidireccional entre su 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 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 administración del ciclo de vida del producto.

policy

Objeto que puede definir permisos (consulte política basada en identidad), especificar las condiciones de acceso (consulte política basada en recursos) o definir los permisos máximos para todas las cuentas de una organización de AWS Organizations (consulte política de control de servicio).

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.

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

Condición de consulta que devuelve `true` o `false`. En general, se encuentra en una cláusula `WHERE`.

inserción de predicados

Técnica de optimización de consultas en bases de datos que filtra los datos de la consulta antes de transferirlos. Esta técnica reduce la cantidad de datos de la base de datos relacional que se tienen que recuperar y procesar. Además, 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

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

zonas alojadas privadas

Contenedor que aloja información acerca de cómo desea que responda Amazon Route 53 a las consultas de DNS de un dominio y sus subdominios en una o varias VPC. Para obtener más información, consulte Uso de zonas alojadas privadas en la documentación de Route 53.

control proactivo

Control de seguridad que se diseñó para evitar la implementación de recursos que no cumplan con la normativa. Estos controles analizan los recursos antes de aprovisionarlos. Si el recurso no cumple con los requisitos del control, no se aprovisiona. 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.

administración del ciclo de vida del producto (PLM)

Administración 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 la reducción de su uso y su retirada.

entorno de producción

Consulte entorno.

controlador lógico programable (PLC)

En el sector de la fabricación, computadora adaptable y altamente fiable que supervisa las máquinas y automatiza los procesos de fabricación.

encadenamiento de peticiones

Uso de la salida de una petición de LLM como entrada para la siguiente petición a fin de generar mejores respuestas. Esta técnica se utiliza para dividir una tarea compleja en tareas secundarias 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)

Patrón que permite establecer 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 pueden suscribir otros microservicios. El sistema puede agregar 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

Consulte responsable, fiable, consultada e informada (RACI).

RAG

Consulte generación aumentada por recuperación.

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

Consulte responsable, fiable, consultada e informada (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

Consulte Las 7 R.

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

Consulte Las 7 R.

Region

Conjunto 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 más información, consulte Specify which Regiones de AWS your account can use.

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 Las 7 R.

versión

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

reubicar

Consulte Las 7 R.

redefinir la plataforma

Consulte Las 7 R.

recomprar

Consulte Las 7 R.

resiliencia

Capacidad de una aplicación para resistir interrupciones o recuperarse de ellas. Al planificar la resiliencia en la Nube de AWS, la alta disponibilidad y la recuperación ante desastres son consideraciones comunes. Para más información, consulte Resiliencia en la Nube de AWS.

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 Las 7 R.

retirar

Consulte Las 7 R.

Generación aumentada de recuperación (RAG)

Tecnología de IA generativa mediante la que un LLM hace referencia a un origen de datos autorizado que se encuentra fuera de sus orígenes de datos de entrenamiento antes de generar una respuesta. Por ejemplo, un modelo de RAG podría hacer una búsqueda semántica en la base de conocimientos o en los datos personalizados de una organización. Para más información, consulte ¿Qué es RAG (generación aumentada por recuperación)?

rotación

Proceso mediante el que periódicamente se actualiza un secreto para que resulte más difícil que un atacante pueda acceder 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 objetivo de punto de recuperación.

RTO

Consulte 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 Consola de administración de AWS 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 control de supervisión y adquisición de datos.

SCP

Consulte política de control de servicio.

secreta

En AWS Secrets Manager, información confidencial o restringida, como una contraseña o credenciales de usuario, que se almacena de forma cifrada. Se compone del valor del secreto y de sus metadatos. El valor del secreto puede ser binario, una sola cadena o varias cadenas. Para más información, consulte What's in a Secrets Manager secret? en la documentación de Secrets Manager.

seguridad desde el diseño

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 de controles de seguridad principales: preventivos, de detección, 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

Acción predefinida y programada que está diseñada para responder automáticamente a un evento de seguridad o corregirlo. Estas automatizaciones sirven como controles de seguridad preventivos o adaptables que le ayudan a implementar las mejores prácticas AWS de seguridad. La modificación de un grupo de seguridad de VPC, la aplicación de revisiones a una instancia de Amazon EC2 o la rotación de credenciales son algunos ejemplos de acciones de respuesta automatizadas.

cifrado del servidor

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

política de control de servicio (SCP)

Una política que proporciona un control centralizado de los permisos de todas las cuentas de una organización en AWS Organizations. Las SCP definen barreras de protección o establecen límites a las acciones que un administrador puede delegar en los usuarios o roles. Puede utilizar las SCP como listas de permitidos o rechazados, para especificar qué servicios o acciones se encuentra permitidos o prohibidos. Para obtener más información, consulte las políticas de control del servicio 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)

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 con AWS la que compartes la seguridad y 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.

Shadow AI

Aplicaciones de IA no autorizadas creadas o utilizadas fuera de los canales regulados dentro de una organización.

SIEM

Consulte sistema de administración de eventos e información de seguridad.

único punto de error (SPOF)

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

SLA

Consulte acuerdo de nivel de servicio.

SLI

Consulte indicador de nivel de servicio.

SLO

Consulte objetivo de nivel de servicio.

modelo de dividir y sembrar

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 más información, consulte Phased approach to modernizing applications in the Nube de AWS.

SPOF

Consulte único punto de error.

esquema en estrella

Estructura organizativa de una base de datos que utiliza una tabla de hechos de gran tamaño para almacenar datos transaccionales o medidos y una o varias tablas dimensionales más pequeñas para almacenar los atributos de los datos. Esta estructura está diseñada para utilizarse 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 de cómo aplicar 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.

control de supervisión y adquisición de datos (SCADA)

En el sector de la fabricación, sistema que utiliza hardware y software para supervisar 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

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

petición del sistema

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

T

etiquetas

Key-value pares que actúan como metadatos para organizar sus AWS recursos. Las etiquetas pueden ayudar a administrar, identificar, organizar, buscar y filtrar recursos de . 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.

herramienta

Una función o API que un agente puede invocar para realizar operaciones en sistemas externos.

puerta de enlace de tránsito

Centro de tránsito de red que puede utilizar para interconectar las VPC y las redes en las instalaciones. 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.

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

Consulte 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

Conexión entre dos VPC que 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 hace un cálculo en un grupo de filas que se relacionan de alguna manera con el registro actual. Las funciones de ventana son útiles para las tareas de procesamiento, 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.

WORM

Consulte escritura única y lectura múltiple.

WQF

Consulte AWS Workload Qualification Framework.

escritura única y lectura múltiple (WORM)

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

Z

ataque de día cero

Ataque, normalmente de malware, que se aprovecha de 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.

peticiones desde cero

Proporcionar a un LLM instrucciones para llevar a cabo una tarea, pero sin ejemplos (pasos) que puedan ayudar a guiarlo. El LLM debe usar los conocimientos del entrenamiento previo para llevar a cabo la tarea. La eficacia de la petición desde cero depende de la complejidad de la tarea y de la calidad de la petición. Consulte también peticiones con 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.