Uso de tablas globales de Amazon DynamoDB

AWS Guía prescriptiva

AWS Guía prescriptiva: Uso de tablas globales de Amazon DynamoDB

Table of Contents

Uso de tablas globales de Amazon DynamoDB

Jason Hunter, Amazon Web Services (AWS)

Marzo de 2024 (historia del documento)

Las tablas globales se crean en la huella global de Amazon DynamoDB para proporcionarle una base de datos totalmente administrada, multirregión y multiactiva que ofrece un rendimiento rápido y local, tanto de lectura como de escritura, para aplicaciones globales de escalado masivo. Las tablas globales replican automáticamente las tablas de DynamoDB en todas las que elija. Regiones de AWS No es necesario realizar cambios en la aplicación porque las tablas globales utilizan DynamoDB APIs existente. No hay costos iniciales ni compromisos por utilizar las tablas globales y solo pagará por los recursos que utilice.

En esta guía se explica cómo utilizar las tablas globales de DynamoDB de forma eficaz. Se proporcionan datos clave sobre las tablas globales, se explican los principales casos de uso de la característica, se presenta una taxonomía de tres modelos de escritura diferentes que debe tener en cuenta, se analizan las cuatro opciones principales de enrutamiento de solicitudes que podría implementar, se analizan las formas de evacuar una región activa o una región sin conexión, se explica cómo pensar en la planificación de la capacidad de rendimiento y se proporciona una lista de aspectos que debe tener en cuenta al implementar tablas globales.

Esta guía se inscribe en un contexto más amplio de las implementaciones en AWS varias regiones, tal como se describe en el documento técnico sobre los fundamentos de AWS varias regiones y en el vídeo sobre los patrones de diseño de la resiliencia de los datos. AWS

Contenido

Descripción general de las tablas globales

Datos clave

- Hay dos versiones de las tablas globales: la versión 2017.11.29 (antigua) (a veces denominada v1) y la versión 2019.11.21 (actual) (a veces denominada v2). Esta guía se centra exclusivamente en la versión actual.
- DynamoDB (sin tablas globales) es un servicio regional, lo que significa que tiene una alta disponibilidad y es intrínsecamente resistente a los fallos de la infraestructura, incluidos los fallos de una zona de disponibilidad completa. Una tabla de DynamoDB de una sola región está diseñada para ofrecer una disponibilidad del 99,99%. Para obtener más información, consulte el acuerdo de nivel de servicio (SLA) de DynamoDB.
- Una tabla global de DynamoDB replica sus datos entre dos o más regiones. Una tabla de DynamoDB multirregional está diseñada para ofrecer una disponibilidad del 99,999%. Con una planificación adecuada, las tablas globales pueden ayudar a crear una arquitectura resistente a los errores regionales.
- Las tablas globales emplean un modelo de replicación activa-activa. Desde la perspectiva de DynamoDB, la tabla en cada región tiene la misma capacidad para aceptar solicitudes de lectura y escritura. Tras recibir una solicitud de escritura, la tabla de réplica local replica la operación de escritura en otras regiones remotas participantes en segundo plano.
- Los elementos se replican de forma individual. Es posible que los elementos que se actualizan en una sola transacción no se repliquen juntos.
- Cada partición de tabla de la región de origen replica sus operaciones de escritura en paralelo con todas las demás particiones. Es posible que la secuencia de operaciones de escritura en una región remota no coincida con la secuencia de operaciones de escritura que se realizaron en la región de origen. Para obtener más información sobre las particiones de tablas, consulte la entrada de blog Escalado de DynamoDB: cómo afectan al rendimiento las particiones, las claves activas y la división por actividad.
- Un elemento que se acaba de escribir se propagará normalmente a todas las réplicas de tabla en cuestión de segundos. Las regiones cercanas tienden a propagarse más rápido.
- Amazon CloudWatch proporciona una `ReplicationLatency` métrica para cada par de regiones. Se calcula observando los artículos que llegan, comparando su hora de llegada con su tiempo de escritura inicial y calculando una media. Los tiempos se almacenan CloudWatch en la región

de origen. La visualización de los tiempos medio y máximo puede resultar útil para determinar el retraso medio y en el peor de los casos de la réplica. No hay SLA para esta latencia.

- Si un elemento individual se actualiza aproximadamente al mismo tiempo (dentro de esta `ReplicationLatency` ventana) en dos regiones diferentes y la segunda operación de escritura se realiza antes de que se replique la primera operación de escritura, existe la posibilidad de que se produzcan conflictos de escritura. Las tablas globales resuelven estos conflictos mediante el mecanismo del último escritor, que se basa en la marca temporal de las operaciones de escritura. La primera operación «pierde» frente a la segunda. Estos conflictos no se registran en CloudWatch o AWS CloudTrail.
- Cada elemento tiene una marca de tiempo de última escritura que se mantiene como una propiedad de sistema privada. El enfoque de último escritor gana se implementa mediante una operación de escritura condicional que requiere que la marca de tiempo del elemento entrante sea mayor que la marca de tiempo del elemento existente.
- Una tabla global reproduce todos los elementos en todas las regiones participantes. Si desea tener distintos ámbitos de replicación, puede crear varias tablas globales y asignar a cada tabla diferentes regiones participantes.
- La región local acepta operaciones de escritura incluso si la región de réplica está fuera de línea o `ReplicationLatency` crece. La tabla local sigue intentando replicar elementos en la tabla remota hasta que cada elemento lo consigue.
- En el improbable caso de que una región quede completamente desconectada, cuando vuelva a estar en línea más adelante, se volverán a intentar todas las replicaciones entrantes y salientes pendientes. No es necesaria ninguna acción especial para volver a sincronizar las tablas. El mecanismo que gana el último autor garantiza que los datos acaben siendo coherentes.
- En cualquier momento, puede agregar una nueva región a una tabla de DynamoDB. DynamoDB gestiona la sincronización inicial y la replicación continua. También puede eliminar una región (incluso la región original) y, de este modo, se eliminará la tabla local de esa región.
- DynamoDB no dispone de un punto de conexión global. Todas las solicitudes se realizan a un punto final regional que accede a la instancia de la tabla global local de esa región.
- Las llamadas a DynamoDB no deben ir de una región a otra. La práctica recomendada es que una aplicación alojada en una región acceda directamente solo al punto final de DynamoDB local de su región. Si se detectan problemas en una región (en la capa de DynamoDB o en la pila circundante), el tráfico de los usuarios finales debe enrutarse a un punto de enlace de aplicación diferente que esté alojado en una región diferente. Las tablas globales garantizan que la aplicación alojada en cada región tenga acceso a los mismos datos.

Casos de uso

Las tablas globales ofrecen las siguientes ventajas comunes:

- Operaciones de lectura de baja latencia. Puede colocar una copia de los datos más cerca del usuario final para reducir la latencia de la red durante las operaciones de lectura. Los datos se mantienen tan actualizados como el `ReplicationLatency` valor.
- Operaciones de escritura de baja latencia. Un usuario final puede escribir en una región cercana para reducir la latencia de la red y el tiempo necesario para completar la operación de escritura. El tráfico de escritura debe enrutarse cuidadosamente para garantizar que no haya conflictos. Las técnicas de enrutamiento se analizan en una sección posterior.
- Mayor resiliencia y recuperación de desastres. Si una región presenta un rendimiento inferior o una interrupción total, puede evacuarla (retirar algunas o todas las solicitudes dirigidas a esa región) y cumplir un objetivo de punto de recuperación (RPO) y un objetivo de tiempo de recuperación (RTO) medidos en segundos. El uso de tablas globales también aumenta el SLA de DynamoDB para el porcentaje de tiempo de actividad mensual del 99,99% al 99,999%.
- Migración de región sin problemas. Puede añadir una nueva región y, a continuación, eliminar la antigua para migrar una implementación de una región a otra, sin ningún tiempo de inactividad en la capa de datos.

Por ejemplo, Fidelity Investments presentó en re:Invent 2022 cómo utilizan las tablas globales de DynamoDB para su sistema de gestión de pedidos. Su objetivo era lograr un procesamiento fiable y de baja latencia a una escala que no podrían alcanzar con el procesamiento local y, al mismo tiempo, mantener la resiliencia ante los fallos regionales y zonales de disponibilidad.

Modos de escritura para tablas globales

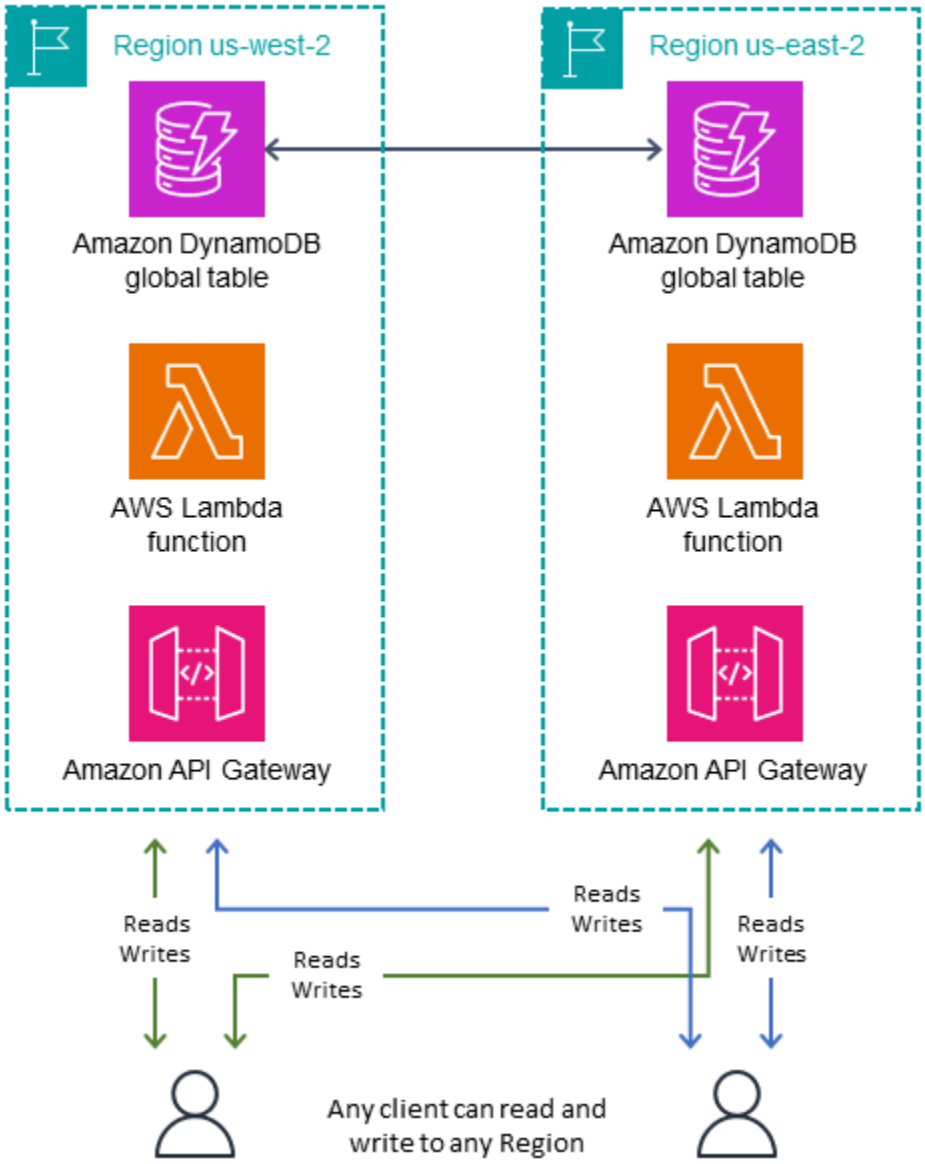
Las tablas globales son siempre activas-activas en el nivel de tabla. No obstante, es recomendable tratarlas como activas-pasivas mediante el control del enrutamiento de las solicitudes de escritura. Por ejemplo, puede decidir enrutar las solicitudes de escritura a una sola región para evitar posibles conflictos de escritura.

Hay tres patrones principales de escritura gestionada, como se explica en las tres secciones siguientes. Debe considerar qué patrón de escritura se ajusta a su caso de uso. Esta elección afecta a la forma de enrutar las solicitudes, evacuar una región y gestionar la recuperación de desastres. Las instrucciones de las secciones posteriores dependen del modo de escritura de la aplicación.

Temas

Modo de escritura en cualquier región (no principal)

El modo de escritura en cualquier región es totalmente activo-activo y no impone restricciones en cuanto al lugar en el que se puede realizar una operación de escritura. Cualquier región puede aceptar una solicitud de escritura en cualquier momento. Este es el modo más simple; sin embargo, solo se puede usar con algunos tipos de aplicaciones. Es adecuado cuando todas las operaciones de escritura son idempotentes. Idempotentes significa que se pueden repetir de forma segura, de modo que las operaciones de escritura simultáneas o repetidas en todas las regiones no entren en conflicto, por ejemplo, cuando un usuario actualiza sus datos de contacto. También funciona bien para un conjunto de datos solo anexado en el que todas las operaciones de escritura son inserciones únicas bajo una clave principal determinista, lo que constituye un caso especial de idempotencia. Por último, este modo es adecuado cuando el riesgo de operaciones de escritura conflictivas es aceptable.



El modo de escritura en cualquier región es la arquitectura más sencilla de implementar. El enrutamiento es más sencillo porque cualquier región puede ser el destino de escritura en cualquier momento. La conmutación por error es más fácil, ya que cualquier operación de escritura reciente se puede reproducir cualquier número de veces en cualquier región secundaria. Siempre que sea posible, debe efectuar el diseño para este modo de escritura.

Por ejemplo, varios servicios de streaming de vídeo utilizan tablas globales para realizar un seguimiento de los marcadores, las reseñas, los indicadores de estado de las reproducciones, etc. Estas implementaciones pueden utilizar el modo de escritura en cualquier región siempre que garanticen que todas las operaciones de escritura sean idempotentes. Este será el caso si cada actualización (por ejemplo, si se establece un nuevo código de tiempo más reciente, se asigna una

nueva opinión o se establece un nuevo estado de visualización) se asigna directamente el nuevo estado del usuario y el siguiente valor correcto de un elemento no depende de su valor actual. Si, por casualidad, las solicitudes de escritura del usuario se redirigen a distintas regiones, la última operación de escritura persistirá y el estado global se liquidará en función de la última asignación. Las operaciones de lectura en este modo acabarán siendo coherentes y se retrasarán según el último `ReplicationLatency` valor.

En otro ejemplo, una empresa de servicios financieros utiliza tablas globales como parte de un sistema para mantener un recuento continuo de las compras con tarjeta de débito de cada cliente, con el fin de calcular las recompensas en efectivo de ese cliente. Las nuevas transacciones llegan de todo el mundo y se envían a varias regiones. Esta empresa pudo utilizar el modo de escritura para cualquier región con un rediseño cuidadoso. El boceto de diseño inicial mantenía un solo `RunningBalance` artículo por cliente. Las acciones del cliente actualizaban el saldo con una `ADD` expresión, que no es idempotente (ya que el nuevo valor correcto depende del valor actual), y el saldo se desincroniza si hay dos operaciones de escritura en el mismo saldo aproximadamente al mismo tiempo y en distintas regiones. El rediseño utiliza la transmisión de eventos, que funciona como un libro de contabilidad con un flujo de trabajo solo de anexos. Cada acción de cliente añade un nuevo elemento a la colección de elementos que se mantiene para ese cliente. (Una colección de elementos es el conjunto de elementos que comparten una clave principal pero tienen claves de clasificación diferentes). Cada operación de escritura es una inserción idempotente que utiliza el identificador de cliente como clave de partición y el identificador de transacción como clave de clasificación. Este diseño dificulta el cálculo del saldo, ya que requiere extraer los elementos y luego realizar algunos cálculos `Query` desde el lado del cliente, pero hace que todas las operaciones de escritura sean idempotentes y logra simplificaciones significativas en el enrutamiento y la conmutación por error. (Esto se analiza con más detalle más adelante en esta guía).

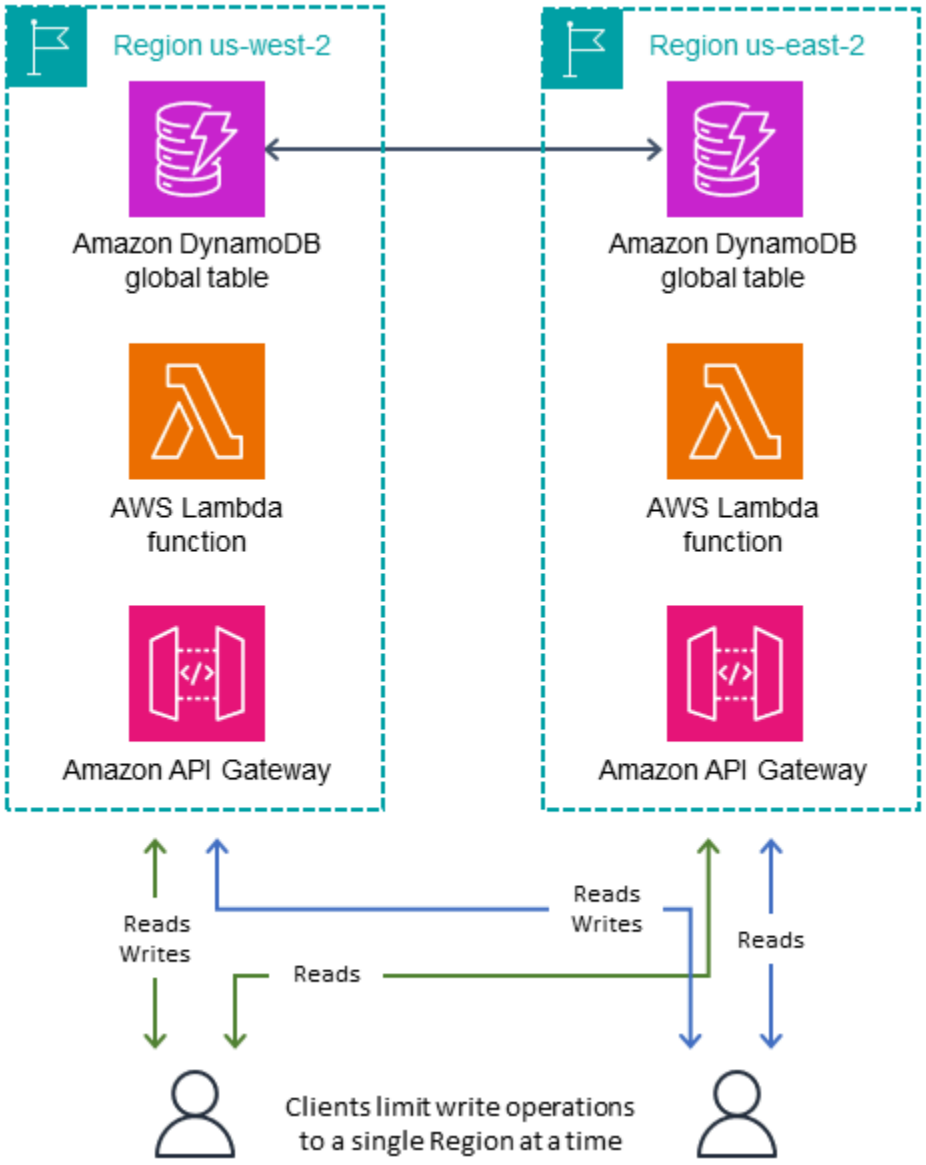
Un tercer ejemplo es el de una empresa que ofrece servicios de colocación de anuncios en línea. Esta empresa decidió que un bajo riesgo de pérdida de datos sería aceptable para lograr las simplificaciones de diseño del modo de escritura a cualquier región. Cuando publican anuncios, solo disponen de unos pocos milisegundos para recuperar los metadatos suficientes para determinar qué anuncio mostrar y, a continuación, registrar la impresión del anuncio para que no repitan el mismo anuncio pronto. Utilizan tablas globales para realizar operaciones de lectura de baja latencia para los usuarios finales de todo el mundo y operaciones de escritura de baja latencia. Registran todas las impresiones de anuncios de un usuario en un único elemento, que se representa como una lista creciente. Utilizan un elemento en lugar de añadirlo a una colección de artículos, por lo que pueden eliminar las impresiones de anuncios antiguas como parte de cada operación de redacción sin tener que pagar por una operación de eliminación. Esta operación de escritura no es idempotente; si el

mismo usuario final ve anuncios publicados en varias regiones aproximadamente al mismo tiempo, existe la posibilidad de que una operación de escritura para una impresión de anuncio sobrescriba a otra. El riesgo es que un usuario vea un anuncio repetido de vez en cuando. Decidieron que esto es aceptable.

Modo de escritura en una región (principal único)

El modo de escritura en una región es activo-pasivo y dirige todas las operaciones de escritura de la tabla a una sola región activa. (DynamoDB no tiene la noción de una sola región activa; esto lo gestiona la capa externa a DynamoDB). El modo de escritura en una región evita los conflictos de escritura al garantizar que las operaciones de escritura fluyan solo a una región a la vez. Este modo de escritura es útil cuando se desean utilizar expresiones o transacciones condicionales. Estas expresiones no son posibles a menos que sepas que estás actuando en función de los datos más recientes, por lo que es necesario enviar todas las solicitudes de escritura a una sola región que tenga los datos más recientes.

En última instancia, las operaciones de lectura coherentes pueden ir a cualquiera de las regiones de la réplica para lograr latencias más bajas. Las operaciones de lectura altamente consistentes deben ir a la única región principal.



A veces es necesario cambiar la región activa en respuesta a un error regional, como se explica más adelante. Algunos usuarios cambian la región actualmente activa de forma periódica, por ejemplo, al implementar una follow-the-sunimplementación. De este modo, la región activa se sitúa cerca de la zona geográfica con más actividad (normalmente, cuando es de día, de ahí viene el nombre), lo que se traduce en las operaciones de lectura y escritura con la latencia más baja. También tiene la ventaja adicional de comprobar el código que cambia de región todos los días y comprobar que está bien probado antes de cualquier recuperación ante un desastre.

Las regiones pasivas pueden mantener una infraestructura reducida en torno a DynamoDB, que solo se crea si se convierte en la región activa. Esta guía no cubre los diseños de luces piloto y sistemas

de espera cálidos. Para obtener más información, puede leer la entrada del blog sobre la arquitectura de recuperación ante desastres (DR) AWS, parte III: Pilot Light and Warm Standby.

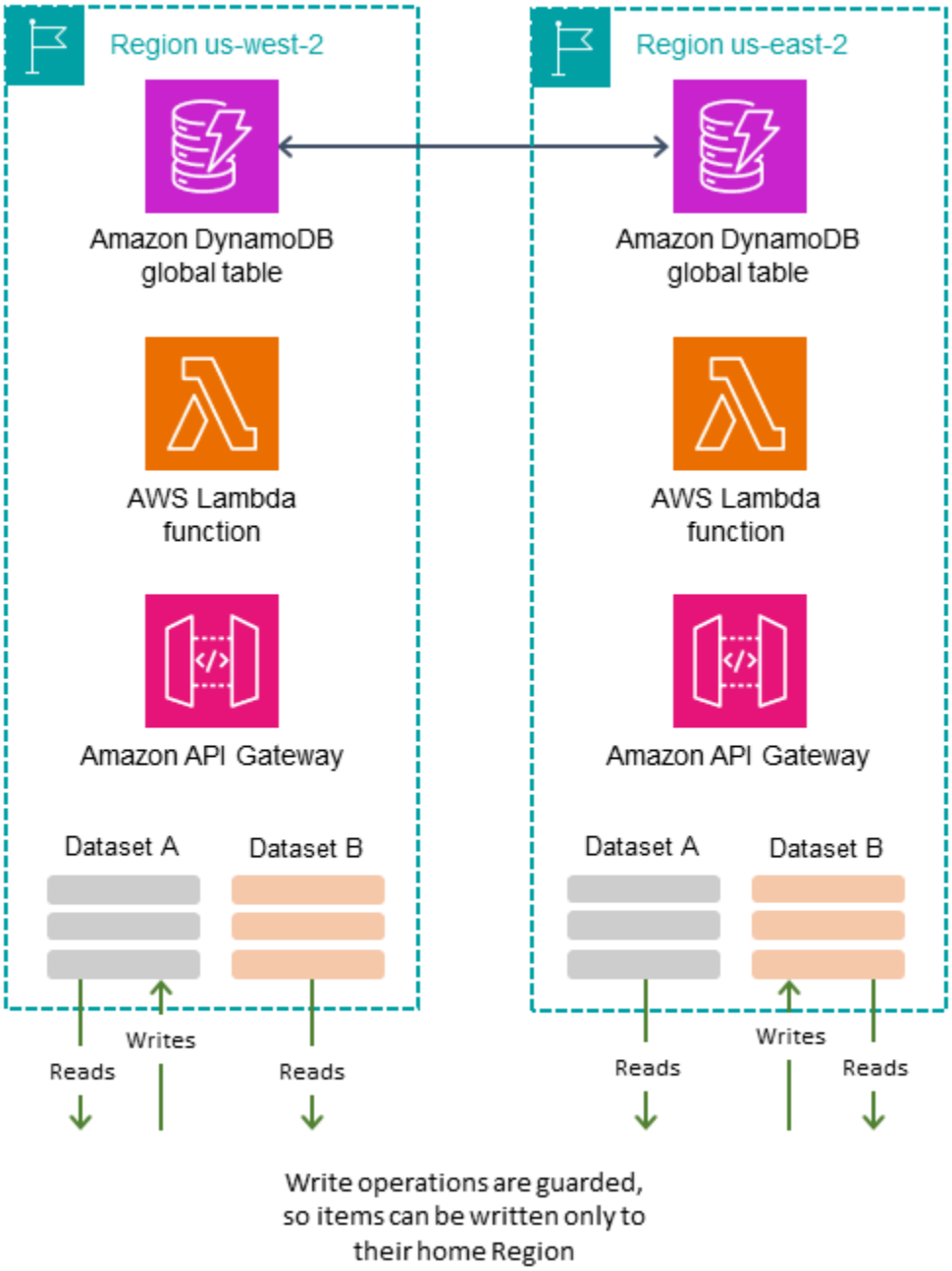
El modo de escritura en una región funciona bien cuando se utilizan tablas globales para operaciones de lectura distribuidas globalmente y de baja latencia. Un ejemplo es una gran empresa de redes sociales que necesita disponer de los mismos datos de referencia en todas las regiones del mundo. No actualizan los datos con frecuencia, pero cuando lo hacen, escriben solo en una región para evitar posibles conflictos de escritura. Las operaciones de lectura siempre están permitidas desde cualquier región.

Como otro ejemplo, pensemos en la empresa de servicios financieros mencionada anteriormente que implementó el cálculo de la devolución de efectivo diaria. Utilizaron el modo de escribir en cualquier región para calcular el saldo, pero utilizaron el modo de escritura en una región para realizar un seguimiento de los pagos de devolución de efectivo. Si quieren recompensar un centavo por cada 10$ gastados, tienen que hacer `Query` todas las transacciones del día anterior, calcular el total gastado, anotar la decisión de devolución de efectivo en una nueva tabla, eliminar el conjunto de artículos consultados para marcarlos como consumidos y sustituirlos por un artículo único que almacene cualquier resto que deba incluirse en los cálculos del día siguiente. Este trabajo requiere transacciones, por lo que funciona mejor con el modo escribir en una región. Una aplicación puede mezclar modos de escritura, incluso en la misma mesa, siempre que las cargas de trabajo no tengan ninguna posibilidad de superponerse.

Modo de escritura en su región (principal mixto)

El modo de escritura en su región asigna distintos subconjuntos de datos a distintas regiones de origen y permite realizar operaciones de escritura en un elemento únicamente a través de su región de origen. Este modo es activo-pasivo, pero asigna la región activa en función del elemento. Cada región es principal para su propio conjunto de datos que no se superponga, y las operaciones de escritura deben estar protegidas para garantizar la ubicación adecuada.

Este modo es similar al de escribir en una región, excepto que permite operaciones de escritura de menor latencia, ya que los datos asociados a cada usuario se pueden colocar más cerca de la red de ese usuario. También distribuye la infraestructura circundante de manera más uniforme entre las regiones y requiere menos trabajo para construir la infraestructura durante un escenario de conmutación por error, ya que todas las regiones tienen una parte de su infraestructura ya activa.



Puede determinar la región de origen de los elementos de varias maneras:

- Intrínseco: algún aspecto de los datos, como un atributo especial o un valor incrustado en su clave de partición, aclara la región de origen. Esta técnica se describe en la entrada del blog Use Region Pinning para establecer una región de inicio para los artículos de una tabla global de Amazon DynamoDB.

- Negociado: la región de origen de cada conjunto de datos se negocia de alguna manera externa, por ejemplo, con un servicio global independiente que mantiene las asignaciones. La asignación puede tener una duración limitada, después de la cual está sujeta a renegociación.
- Orientado a tablas: en lugar de crear una única tabla global replicante, se crea el mismo número de tablas globales que las regiones replicantes. El nombre de cada tabla indica su región de origen. En las operaciones estándar, todos los datos se escriben en la región de origen, mientras que las demás regiones conservan una copia de solo lectura. Durante una conmutación por error, otra región adopta temporalmente tareas de escritura para esa tabla.

Por ejemplo, imagina que trabajas para una empresa de juegos. Necesitas operaciones de lectura y escritura de baja latencia para todos los jugadores de todo el mundo. Asignas a cada jugador a la región que esté más cerca de él. Esa región realiza todas sus operaciones de lectura y escritura, lo que garantiza una gran read-after-write coherencia. Sin embargo, cuando un jugador viaja o si su región de origen sufre una interrupción, hay una copia completa de sus datos disponible en otras regiones y se le puede asignar al jugador a una región de origen diferente.

Como otro ejemplo, imagina que trabajas en una empresa de videoconferencias. Los metadatos de cada conferencia telefónica se asignan a una región concreta. Las personas que llaman pueden usar la región más cercana a ellas para obtener la latencia más baja. Si se produce una interrupción en una región, el uso de tablas globales permite una recuperación rápida, ya que el sistema puede trasladar el procesamiento de la llamada a otra región en la que ya existe una copia replicada de los datos.

Estrategias de enrutamiento para tablas globales

Quizá la parte más compleja de la implementación de una tabla global sea administrar el enrutamiento de las solicitudes. Las solicitudes deben pasar primero de un usuario final a una región elegida y enrutada de alguna manera. La solicitud encuentra una pila de servicios en esa región, incluida una capa de procesamiento que quizás consista en un balanceador de carga respaldado por una AWS Lambda función, un contenedor o un nodo de Amazon Elastic Compute Cloud (Amazon EC2) y, posiblemente, otros servicios, incluida quizás otra base de datos. Esa capa de procesamiento se comunica con DynamoDB. Debería hacerlo mediante el punto final local de esa región. Los datos de la tabla global se replican en las demás regiones participantes y cada región dispone de una pila similar de servicios en torno a su tabla de DynamoDB.

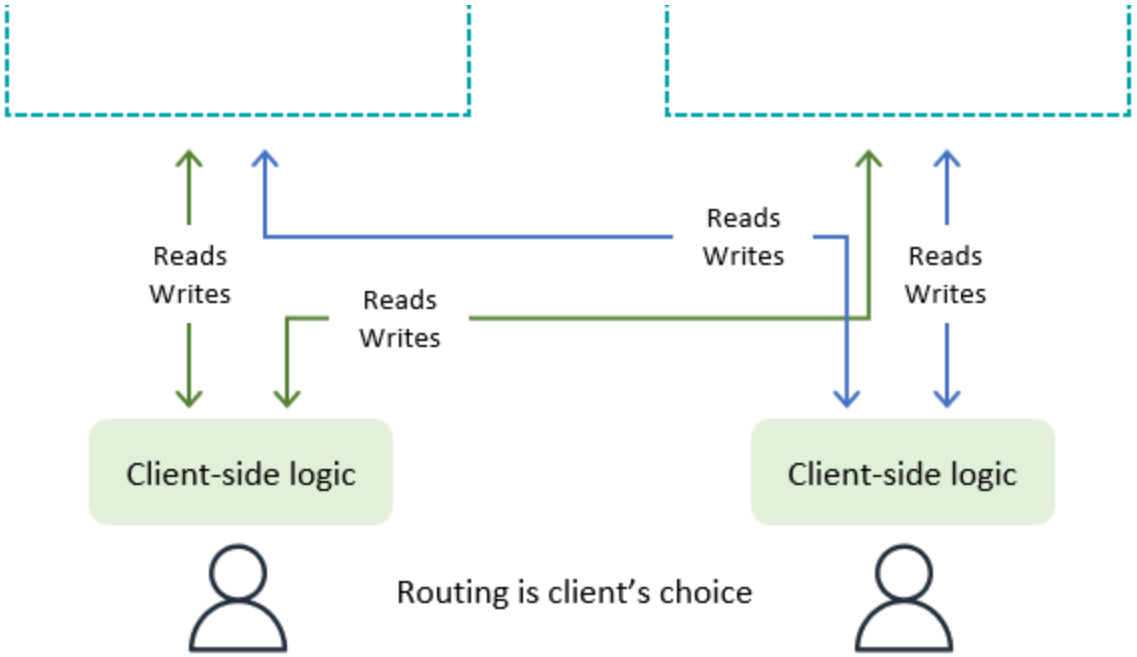
La tabla global proporciona a cada pila de las distintas regiones una copia local de los mismos datos. Podría considerar la posibilidad de diseñar para una única pila en una única Región y prever la realización de llamadas remotas al punto de conexión de DynamoDB de una Región secundaria si se produce algún problema con la tabla de DynamoDB local. Esta no es la mejor práctica. Las latencias asociadas al acceso entre regiones pueden ser 100 veces superiores a las del acceso local. Una back-and-forth serie de 5 solicitudes puede tardar milisegundos cuando se realiza de forma local, pero segundos cuando se cruza el mundo. Es mejor enrutar al usuario final a otra región para que se procese. Para garantizar la resiliencia, necesita la replicación en varias regiones: la replicación de la capa de procesamiento y la capa de datos.

Existen numerosas técnicas para enrutar una solicitud de un usuario final a una región para su procesamiento. La elección correcta depende del modo de escritura y de las consideraciones de conmutación por error. En esta sección se analizan cuatro opciones: impulsada por el cliente, capa de cómputo, Amazon Route 53 y. AWS Global Accelerator

Temas

Enrutamiento de solicitudes basado en el cliente

Con el enrutamiento de solicitudes impulsado por el cliente, el cliente del usuario final (una aplicación, una página web u otro cliente) realiza un seguimiento de los puntos de enlace de la aplicación válidos (por ejemplo, un punto de enlace de Amazon API Gateway en lugar de un punto de enlace literal de DynamoDB) y utiliza su propia lógica integrada para elegir la región con la que comunicarse. JavaScript Puede elegir en función de una selección aleatoria, las latencias más bajas observadas, las mediciones de ancho de banda más altas observadas o las comprobaciones de estado realizadas localmente.



Como ventaja, el enrutamiento de solicitudes impulsado por el cliente puede adaptarse a factores como las condiciones reales del tráfico público de Internet para cambiar de región si detecta algún deterioro en el rendimiento. El cliente debe conocer todos los posibles puntos de conexión, pero lanzar un nuevo punto de conexión regional no es algo frecuente.

Con el modo de escritura en cualquier región, un cliente puede seleccionar unilateralmente su punto final preferido. Si su acceso a una región se ve afectado, el cliente puede enrutarse a otro punto de conexión.

Con el modo de escritura en una región, el cliente necesita un mecanismo para dirigir sus solicitudes de escritura a la región actualmente activa. Este podría ser un mecanismo básico, como probar empíricamente qué región acepta actualmente las solicitudes de escritura (anotando cualquier rechazo de escritura y recurriendo a una alternativa). O puede tratarse de un mecanismo complejo,

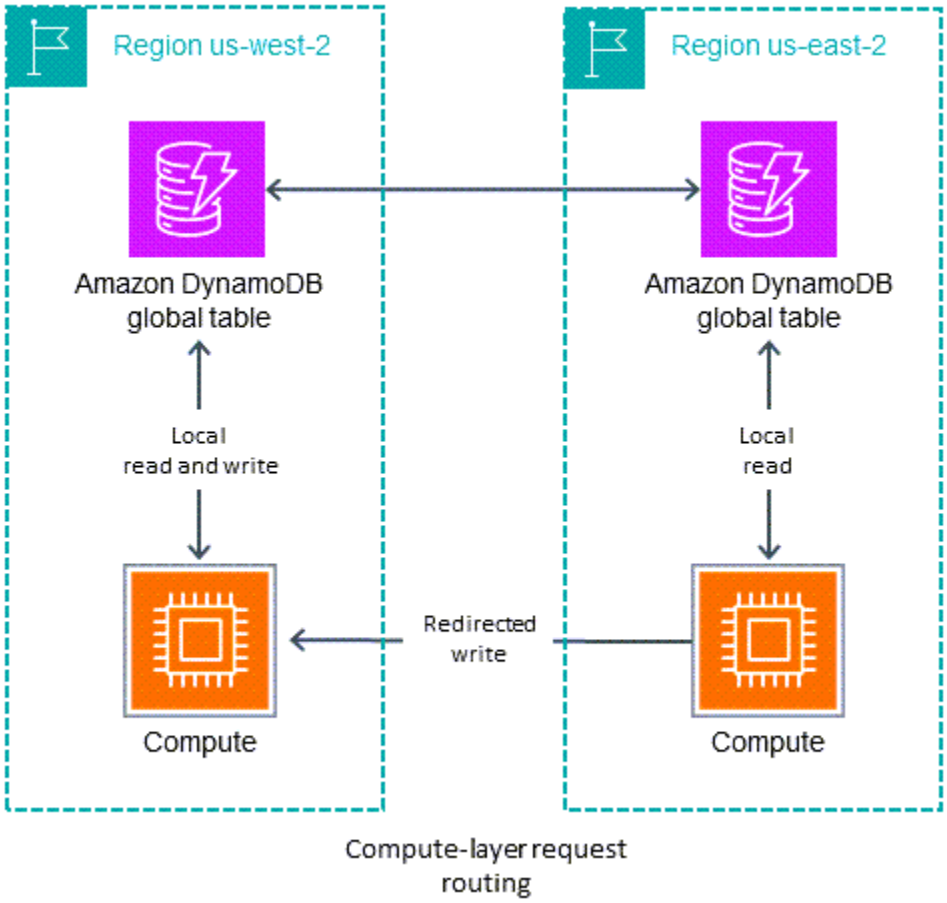
como el uso de un coordinador global para consultar el estado actual de la aplicación (quizás basado en el control de enrutamiento de Amazon Application Recovery Controller (ARC) (ARC), que proporciona un sistema de cinco regiones controlado por quórum para mantener el estado global para necesidades como esta). El cliente puede decidir si las solicitudes de lectura pueden enviarse a cualquier región para garantizar una coherencia definitiva o si deben enviarse a la región activa para garantizar una coherencia sólida.

Con el modo de escritura en su región, el cliente debe determinar la región de origen del conjunto de datos con el que está trabajando. Por ejemplo, si el cliente corresponde a una cuenta de usuario y cada cuenta de usuario está alojada en una región, el cliente puede solicitar la asignación de punto final adecuada para utilizarla con sus credenciales desde un sistema de inicio de sesión global.

Por ejemplo, una empresa de servicios financieros que ayuda a los usuarios a gestionar las finanzas de su empresa a través de la web utiliza tablas globales con el modo «Escribe en tu región». Cada usuario debe iniciar sesión en un servicio central. Ese servicio devuelve las credenciales, así como el punto final de la región en la que funcionarán esas credenciales. La región que se devuelve se basa en el lugar donde se encuentra actualmente el conjunto de datos del usuario. Las credenciales son válidas durante un período breve. Después de eso, la página web negocia automáticamente un nuevo inicio de sesión, lo que brinda la oportunidad de redirigir la actividad del usuario a una nueva región.

Enrutamiento de solicitudes de la capa de computación

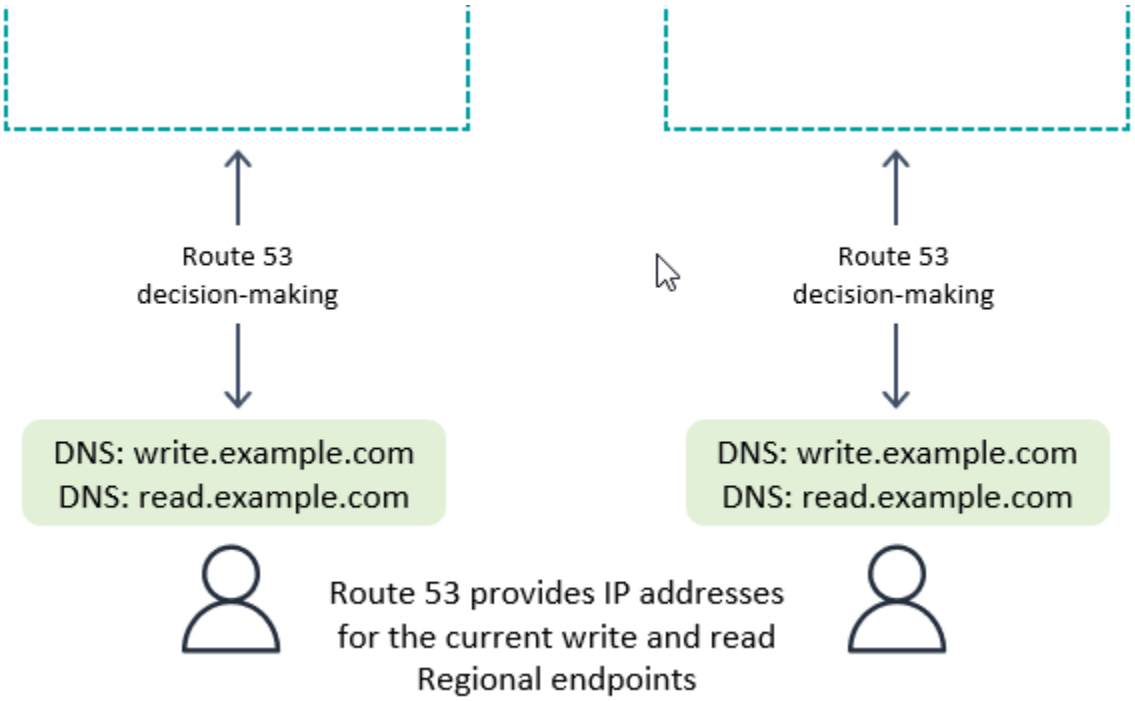
Con el enrutamiento de solicitudes en la capa de cómputo, el código que se ejecuta en la capa de cómputo determina si se debe procesar la solicitud localmente o pasarla a una copia suya que se esté ejecutando en otra región. Al utilizar el modo de escritura en una región, la capa de procesamiento puede detectar que no es la región activa y permitir las operaciones de lectura locales y, al mismo tiempo, reenviar todas las operaciones de escritura a otra región. Este código de capa de cómputo debe conocer la topología de los datos y las reglas de enrutamiento y aplicarlas de manera confiable, en función de la configuración más reciente que especifica qué regiones están activas para qué datos. La pila de software externa en la región no tiene por qué conocer cómo el microservicio enruta las solicitudes de lectura y de escritura. En un diseño sólido, la región receptora valida si es la principal actual para la operación de escritura. Si no lo es, genera un error que indica que es necesario corregir el estado global. La Región receptora también podría almacenar en búfer la operación de escritura durante un tiempo si la región principal está en proceso de cambiar. En todos los casos, la pila de computación de una región escribe solo en su punto de conexión de DynamoDB local, pero las pilas de computación podrían comunicarse entre sí.



El Grupo Vanguard utiliza un sistema denominado Global Orchestration and Status Tool (GOaST) y una biblioteca denominada Global Multi-Region library (GMRlib) para este proceso de enrutamiento, tal como se presentó en re:Invent 2022. Utilizan un único modelo primario. follow-the-sun GOaEl ST mantiene el estado global, de forma similar al control de enrutamiento ARC descrito en la sección anterior. Utiliza una tabla global para rastrear qué región es la región principal y cuándo está programado el siguiente conmutador principal. Se realizan todas las operaciones de lectura y escritura GMRlib, que se coordinan con GOa ST. GMRlib permite que las operaciones de lectura se realicen localmente, con baja latencia. Para las operaciones de escritura, GMRlib comprueba si la región local es la región principal actual. Si es así, la operación de escritura se completa directamente. Si no, GMRlib reenvía la tarea de escritura a GMRlib la región principal. La biblioteca receptora confirma que también se considera la región principal y, si no lo es, genera un error, lo que indica un retraso de propagación con el estado global. Este enfoque proporciona un beneficio de validación al no escribir directamente en un punto de conexión de DynamoDB remoto.

Enrutamiento de solicitudes de Route 53

Amazon Route 53 es una tecnología de servicio de nombres de dominio (DNS). Con Route 53, el cliente solicita su punto final buscando un nombre de dominio DNS conocido y Route 53 devuelve la dirección IP que corresponde a los puntos finales regionales que considera más adecuados. Route 53 tiene una larga lista de políticas de enrutamiento que utiliza para determinar la región adecuada. También puede realizar un enrutamiento de conmutación por error para desviar el tráfico de las regiones que no pasen las comprobaciones de estado.



Con el modo de escritura a cualquier región, o si se combina con el enrutamiento de solicitudes de nivel de cómputo en el backend, Route 53 tiene total libertad para devolver la región en función de cualquier regla interna compleja, como elegir la región más cercana a la red o la proximidad geográfica, o cualquier otra opción.

Con el modo de escritura en una región, puede configurar Route 53 para que devuelva la región actualmente activa (mediante ARC). Si el cliente quiere conectarse a una región pasiva (por ejemplo, para operaciones de lectura), puede buscar un nombre de DNS diferente.

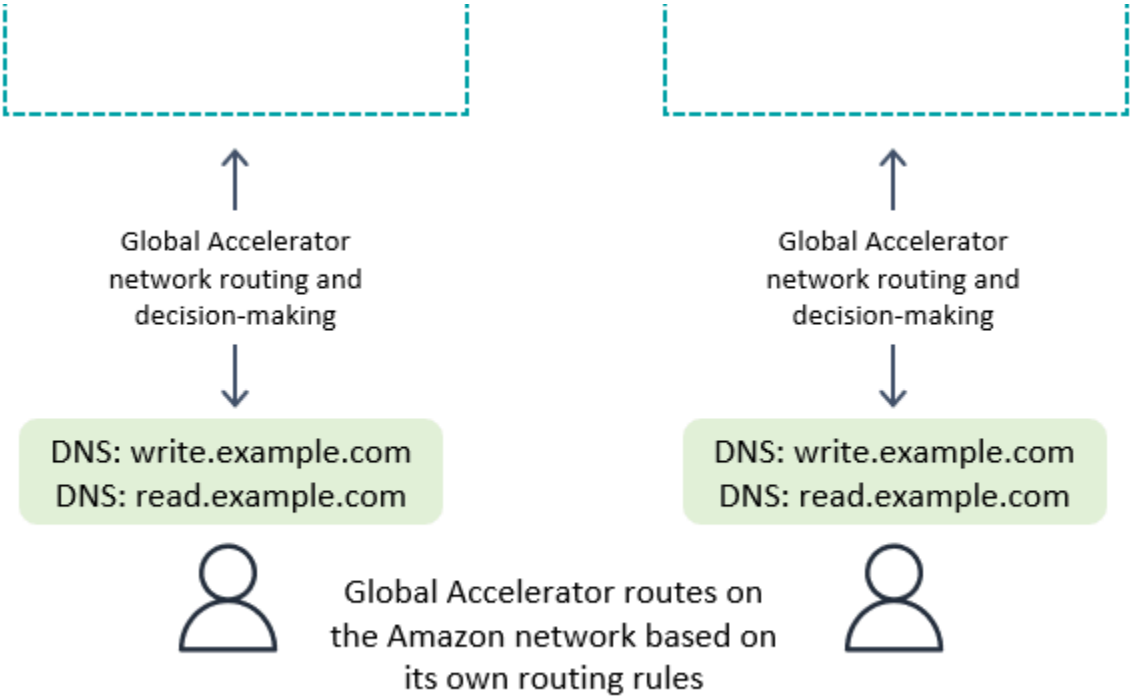
> **Note**
>
> Los clientes almacenan en caché las direcciones IP en la respuesta de Route 53 durante un tiempo indicado por la configuración de tiempo de vida (TTL) del nombre de dominio. Un TTL

más largo amplía el objetivo de tiempo de recuperación (RTO) para que todos los clientes reconozcan el nuevo punto de conexión. Un valor de 60 segundos es típico para utilizar en la conmutación por error. No todo el software cumple perfectamente con la caducidad del TTL del DNS, y es posible que haya varios niveles de almacenamiento en caché del DNS, por ejemplo, en el sistema operativo, la máquina virtual y la aplicación.

Con el modo de escritura en tu región, es mejor evitar Route 53, a menos que también utilices el enrutamiento de solicitudes a nivel de cómputo.

Enrutamiento de solicitudes de Global Accelerator

Con AWS Global Acceleratorun cliente, busca el nombre de dominio conocido en Route 53. Sin embargo, en lugar de recuperar una dirección IP que corresponde a un punto final regional, el cliente recupera una dirección IP estática de transmisión automática que se dirige a la ubicación AWS perimetral más cercana. A partir de esa ubicación de borde, todo el tráfico de la AWS red privada se dirige a algún punto final (balanceadores de carga de red, balanceadores de carga de aplicaciones, EC2 instancias o direcciones IP elásticas) en una región elegida mediante las reglas de enrutamiento que se mantienen en Global Accelerator. En comparación con el enrutamiento basado en las reglas de Route 53, el enrutamiento de solicitudes de Global Accelerator tiene latencias más bajas porque reduce la cantidad de tráfico en el Internet público. Además, dado que Global Accelerator no depende de la caducidad del TTL del DNS para cambiar las reglas de enrutamiento, puede ajustar el enrutamiento más rápidamente.



Con el modo de escritura en cualquier región, o si se combina con el enrutamiento de solicitudes de la capa de cómputo en el backend, Global Accelerator funciona a la perfección. El cliente se conecta a la ubicación perimetral más cercana y no tiene que preocuparse por la región que reciba la solicitud.

Con el modo de escritura en una región, las reglas de enrutamiento de Global Accelerator deben enviar las solicitudes a la región actualmente activa. Puede utilizar comprobaciones de estado que informen artificialmente de un error en cualquier región que su sistema global no considere la activa. Al igual que con el DNS, es posible usar un nombre de dominio DNS alternativo para enrutar las solicitudes de lectura, si las solicitudes pueden provenir de cualquier región.

Con el modo de escritura en tu región, es mejor evitar Global Accelerator, a menos que también utilices el enrutamiento de solicitudes a nivel de cómputo.

Procesos de evacuación para mesas globales

La evacuación de una región es el proceso de migrar una actividad (normalmente actividad de escritura, posiblemente actividad de lectura) fuera de esa región.

Evacuación de una región activa

Puede decidir evacuar una región activa por varios motivos: como parte de su actividad empresarial habitual (por ejemplo, si utiliza el modo de escribir a una región) follow-the-sun, debido a una decisión empresarial de cambiar la región actualmente activa, en respuesta a fallos en el paquete de software ajeno a DynamoDB o porque se encuentra con problemas generales, como latencias más altas de lo habitual dentro de la región.

Con el modo de escritura en cualquier región, evacuar una región activa es sencillo. Puede enrutar el tráfico a regiones alternativas mediante cualquier sistema de enrutamiento y dejar que las operaciones de escritura que ya se han realizado en la región evacuada se repliquen como de costumbre.

Con los modos escribir en una región y escribir en su región, debe asegurarse de que todas las operaciones de escritura en la región activa se hayan grabado por completo, procesado en flujo y propagado globalmente antes de iniciar las operaciones de escritura en la nueva región activa, para garantizar que las futuras operaciones de escritura se procesen con la última versión de los datos.

Supongamos que la región A es activa y la región B es pasiva (para la tabla completa o para los elementos asignados a la región A). El mecanismo típico para llevar a cabo una evacuación consiste en pausar las operaciones de escritura en A, esperar el tiempo suficiente para que esas operaciones se hayan propagado completamente a B, actualizar la pila de la arquitectura para que reconozca B como activa y, a continuación, reanudar las operaciones de escritura en B. No existe ninguna métrica que indique con absoluta certeza que la región A ha replicado completamente sus datos a la región B. Si la región A está en buen estado, pausar las operaciones de escritura en la región A y esperar diez veces el valor máximo reciente de la métrica `ReplicationLatency` normalmente sería suficiente para determinar que la replicación se ha completado. Si el estado de la región A no es correcto y muestra otras zonas de latencias aumentadas, elegiría un múltiplo mayor para el tiempo de espera.

Evacuación de una región sin conexión

Hay un caso especial que hay que tener en cuenta: ¿qué pasa si la región A se desconecta por completo sin previo aviso? Esto es muy poco probable, pero debe tenerse en cuenta de todos modos. Si esto ocurre, cualquier operación de escritura en la región A que aún no se haya propagado se retiene y se propaga después de que la región A vuelva a estar en línea. Las operaciones de escritura no se pierden, pero su propagación se retrasa indefinidamente.

La aplicación decide cómo continuar en este caso. Por motivos de continuidad empresarial, es posible que las operaciones de escritura deban continuar hacia la nueva región primaria B. No obstante, si un elemento de la región B recibe una actualización mientras hay una propagación pendiente de una operación de escritura para ese elemento desde la región A, la propagación se suprime según el modelo de último escritor gana. Cualquier actualización en la región B podría suprimir una solicitud de escritura entrante.

Con el modo de escritura en cualquier región, las operaciones de lectura y escritura pueden continuar en la región B, confiando en que los objetos de la región A se propagarán eventualmente a la región B y reconociendo la posibilidad de que falten elementos hasta que la región A vuelva a funcionar. Siempre que sea posible, por ejemplo, con operaciones de escritura idempotentes, debería plantearse la posibilidad de reproducir el tráfico de escritura reciente (por ejemplo, mediante una fuente de eventos ascendente) para cubrir el vacío de cualquier operación de escritura que pueda faltar y dejar que el último escritor gane. La resolución de conflictos suprima la eventual propagación de la operación de escritura entrante.

En el caso de los demás modos de escritura, hay que tener en cuenta hasta qué punto se puede continuar trabajando con una ligera out-of-date visión del mundo. Faltará una pequeña duración de las operaciones de escritura, según el seguimiento de `ReplicationLatency`, hasta que la región A vuelva a estar en línea. ¿Puede avanzar la actividad? En algunos casos de uso puede que sí, pero en otros puede que no si no hay mecanismos de mitigación adicionales.

Por ejemplo, imagine que tiene que mantener un saldo de crédito disponible sin interrupción incluso después de una interrupción total en una región. Podrías dividir el saldo en dos partidas diferentes, una reservada en la región A y otra en la región B, y empezar cada una con la mitad del saldo disponible. De este modo, se utilizaría el modo de escritura en su región. Las actualizaciones transaccionales procesadas en cada región se escribirían según la copia local del saldo. Si la región A se queda totalmente fuera de línea, se podría seguir trabajando con el procesamiento de transacciones en la región B y las operaciones de escritura se limitarían a la parte del saldo que se mantiene en la región B. Dividir el saldo de esta manera conlleva complejidades cuando el saldo baja

o hay que reequilibrar el crédito, pero proporciona un ejemplo de recuperación segura de la actividad incluso con operaciones de escritura pendientes inciertas.

Como otro ejemplo, imagina que estás capturando datos de formularios web. Puede usar el control de simultaneidad optimista (OCC) para asignar versiones a los elementos de datos e incrustar la última versión en el formulario web como un campo oculto. En cada envío, la operación de escritura solo se realiza correctamente si la versión de la base de datos sigue coincidiendo con la versión con la que se creó el formulario. Si las versiones no coinciden, el formulario web puede actualizarse (o combinarse cuidadosamente) en función de la versión actual de la base de datos y el usuario puede continuar de nuevo. El modelo OCC suele proteger contra el hecho de que otro cliente sobrescriba los datos y produzca una nueva versión de ellos, pero también puede ser de ayuda durante la conmutación por error, cuando un cliente puede encontrarse con versiones más antiguas de los datos. Imaginemos que utiliza la marca de tiempo como versión. El formulario se creó por primera vez para la región A a las 12:00 pero (tras una conmutación por error) intenta escribir en la región B y observa que la última versión de la base de datos es a las 11:59. En este escenario, el cliente puede esperar a que la versión de las 12:00 se propague a la región B y escribir en esa versión, o basarse en la de las 11:59 y crear una nueva versión de las 12:01 (que, después de la escritura, suprimiría la versión entrante una vez que se recupere la región A).

Como tercer ejemplo, una empresa de servicios financieros guarda datos sobre las cuentas de los clientes y sus transacciones financieras en una base de datos de DynamoDB. En el caso de que se produzca una interrupción total en la Región A, querrá asegurarse de que cualquier actividad de escritura relacionada con sus cuentas esté totalmente disponible en la Región B, o bien poner sus cuentas en cuarentena, de forma parcial, hasta que la Región A vuelva a funcionar. En lugar de pausar toda la actividad, decide pausarla solo para la minúscula fracción de cuentas que ha determinado que tienen transacciones no propagadas. Para lograrlo, utilizan una tercera región, a la que llamaremos región C. Antes de procesar cualquier operación de escritura en la región A, incluyen un resumen sucinto de esas operaciones pendientes (por ejemplo, un nuevo recuento de transacciones para una cuenta) en la región C. Este resumen es suficiente para que la región B determine si su vista está totalmente actualizada. Esta acción bloquea de un modo efectivo la cuenta desde el momento de la escritura en la región C hasta que la región A acepta las operaciones de escritura y la región B las recibe. Los datos de la región C no se utilizan excepto como parte de un proceso de conmutación por error, tras el cual la región B puede cruzar sus datos con los de la región C para comprobar si alguna de sus cuentas está desactualizada. Esas cuentas se marcarían como puestas en cuarentena hasta que la empresa de recuperación de la región A propagase los datos parciales a la región B. Si la región C fallara, se podría crear una nueva región D para utilizarla en su lugar. Los datos de la región C eran muy transitorios y, al cabo de unos minutos, la región

D tendría un up-to-date registro suficiente de las operaciones de escritura durante el vuelo como para ser totalmente útil. Si se produce un error en la región B, la región A puede seguir aceptando solicitudes de escritura en cooperación con la región C. Esta empresa estaba dispuesta a aceptar escrituras de latencia más alta (a dos regiones: C y luego A) y tuvo la suerte de contar con un modelo de datos en el que se podía resumir sucintamente el estado de una cuenta.

Planificación de la capacidad de rendimiento para tablas globales

La migración del tráfico de una región a otra requiere un examen cuidadoso de la configuración de las tablas de DynamoDB en lo que respecta a la capacidad.

Estas son algunas consideraciones para administrar la capacidad de escritura:

- Una tabla global debe estar en modo bajo demanda o aprovisionada con el escalado automático activado.
- Si se aprovisiona con escalado automático, la configuración de escritura (utilización mínima, máxima y objetivo) se replica en todas las regiones. Aunque la configuración del escalado automático esté sincronizada, la capacidad de escritura real aprovisionada podría flotar independientemente entre las regiones.
- Una de las razones por las que es posible que veas una capacidad de escritura aprovisionada diferente se debe a la función de tiempo de vida (TTL). Al habilitar el TTL en DynamoDB, puede especificar un nombre de atributo cuyo valor indique la hora de caducidad del elemento, en formato de época de Unix en segundos. Transcurrido ese tiempo, DynamoDB puede eliminar el elemento sin incurrir en costos de escritura. Con las tablas globales, puede configurar TTL en cualquier región y la configuración se replica automáticamente a otras regiones que están asociadas a la tabla global. Cuando un elemento reúne las condiciones para eliminarse mediante una regla TTL, esa tarea puede realizarse en cualquier región. La operación de eliminación se realiza sin consumir unidades de escritura en la tabla de origen, pero las tablas de réplica obtendrán una escritura replicada de esa operación de eliminación e incurrirán en costes por unidad de escritura replicada.
- Si utiliza el escalado automático, asegúrese de que la configuración de la capacidad máxima de escritura aprovisionada es lo suficientemente alta como para gestionar todas las operaciones de escritura, así como todas las posibles operaciones de eliminación de TTL. El escalado automático ajusta cada región en función de su consumo de escritura. Las tablas bajo demanda no tienen una configuración de capacidad de escritura máxima aprovisionada, pero el límite de rendimiento de escritura máxima en el nivel de tabla especifica la capacidad de escritura máxima sostenida que permitirá la tabla bajo demanda. El límite predeterminado es de 40 000, pero se puede ajustar. Le recomendamos que lo establezca lo suficientemente alto como para gestionar todas las operaciones de escritura (incluidas las operaciones de escritura TTL) que pueda necesitar la tabla

bajo demanda. Este valor debe ser el mismo en todas las regiones participantes cuando configure tablas globales.

Estas son algunas consideraciones para administrar la capacidad de lectura:

- Se permite que la configuración de la administración de la capacidad de lectura difiera de una región a otra porque se supone que las distintas regiones pueden tener patrones de lectura independientes. Al agregar por primera vez una réplica global a una tabla, se propaga la capacidad de la región de origen. Tras la creación, puede ajustar la configuración de la capacidad de lectura, que no se transfiere al otro lado.
- Cuando utilice el escalado automático de DynamoDB, asegúrese de que la configuración de la capacidad máxima de lectura aprovisionada es lo suficientemente alta como para gestionar todas las operaciones de lectura en todas las regiones. Durante las operaciones estándar, la capacidad de lectura quizá se reparta entre las regiones, pero durante la conmutación por error la tabla debería poder adaptarse automáticamente al aumento de la carga de trabajo de lectura. Las tablas bajo demanda no tienen una configuración de capacidad de lectura máxima aprovisionada, pero el límite de rendimiento de lectura máxima en el nivel de tabla especifica la capacidad de lectura máxima sostenida que permitirá la tabla bajo demanda. El límite predeterminado es de 40 000, pero se puede ajustar. Le recomendamos que lo establezca lo suficientemente alto como para gestionar todas las operaciones de lectura que podría necesitar la tabla si todas las operaciones de lectura tuvieran que enrutarse a esta única región.
- Si una tabla de una región no suele recibir tráfico de lectura pero podría tener que absorber una gran cantidad de tráfico de lectura tras una conmutación por error, puede aumentar la capacidad de lectura aprovisionada de la tabla, esperar a que la tabla termine de actualizarse y volver a reducir la capacidad. Puede dejar la tabla en modo aprovisionado o cambiarla a modo bajo demanda. Esto prepara la tabla para aceptar un mayor nivel de tráfico de lectura.

ARC cuenta con comprobaciones de disponibilidad que pueden resultar útiles para confirmar que las regiones de DynamoDB tienen configuraciones de tablas y cuotas de cuenta similares, independientemente de que utilice Route 53 para enrutar las solicitudes o no. Estas comprobaciones de disponibilidad también le ayudan a ajustar las cuotas a nivel de cuenta para que coincidan.

Lista de verificación de preparación para tablas globales

Utilice la siguiente lista de comprobación para tomar decisiones y realizar tareas cuando despliegue tablas globales.

- Determine cuántas y qué regiones deben participar en la tabla global.
- Determine el modo de escritura de la aplicación.
- Planifique su estrategia de enrutamiento en función de su modo de escritura.
- Defina su plan de evacuación en función de su modo de escritura y su estrategia de enrutamiento.
- Capture métricas sobre el estado, la latencia y los errores de cada región. Para obtener una lista de las métricas de DynamoDB, consulte la entrada AWS del blog Monitoring Amazon DynamoDB for Operational Awareness. También debería utilizar canarios sintéticos (solicitudes artificiales diseñadas para detectar errores) y observar en tiempo real el tráfico de clientes. No todos los problemas aparecen en las métricas de DynamoDB.
- Establezca alarmas para cualquier aumento sostenido de `ReplicationLatency`. Un aumento podría indicar un error de configuración accidental en el que la tabla global tiene diferentes opciones de escritura en distintas regiones, lo que da lugar a solicitudes replicadas con errores y a un aumento de las latencias. También podría indicar que existe una interrupción regional. Un buen ejemplo sería generar una alerta si el promedio reciente supera los 180 000 milisegundos. También puede vigilar si `ReplicationLatency` cae a 0, lo que indica que la replicación se ha estancado.
- Asigne una configuración máxima de lectura y escritura suficiente para cada tabla global.
- Identifique las condiciones en las que evacuaría una región. Si la decisión implica una evaluación manual, documente todas las consideraciones. Este trabajo debe realizarse cuidadosamente con antelación, no bajo estrés.
- Mantenga un manual de procedimientos para cada acción que deba llevarse a cabo cuando evacúe una región. Normalmente se requiere muy poco trabajo para las tablas globales, pero trasladar el resto de la pila puede resultar complejo.

> **Note**
>
> Con los procedimientos de conmutación por error, se recomienda confiar únicamente en las operaciones del plano de datos y no en las del plano de control, ya que algunas operaciones del plano de control pueden verse degradadas durante los fallos de la región.

Para obtener más información, consulte la entrada del AWS blog Cree aplicaciones resilientes con las tablas globales de Amazon DynamoDB: Parte 4.

- Pruebe periódicamente todos los aspectos del manual de procedimientos, incluidas las evacuaciones de región. Un manual de procedimientos no probado es un manual poco fiable.
- Considere utilizarlas AWS Resilience Hubpara evaluar la resiliencia de toda la aplicación (incluidas las tablas globales). Este servicio proporciona una visión completa del estado de resiliencia de su cartera de aplicaciones a través de su panel de control.
- Considere la posibilidad de utilizar las comprobaciones de aptitud para el ARC para evaluar la configuración actual de su aplicación y realizar un seguimiento de cualquier desviación respecto a las mejores prácticas.
- Cuando realice comprobaciones de estado para utilizarlas con Route 53 o Global Accelerator, realice una serie de llamadas que abarquen todo el flujo de la base de datos. Si limita la comprobación para confirmar únicamente que el punto final de DynamoDB está activo, no podrá cubrir muchos modos de error, AWS Identity and Access Management como errores de configuración (IAM), problemas de despliegue de código, errores en la pila fuera de DynamoDB, latencias de lectura o escritura superiores a la media, etc.

Preguntas frecuentes sobre tablas globales

En esta sección, se brindan respuestas a las preguntas frecuentes sobre las tablas globales de DynamoDB.

¿Cuál es el precio de las tablas globales?

- El precio de una operación de escritura en una tabla de DynamoDB tradicional se expresa en unidades de capacidad de escritura WCUs () para las tablas aprovisionadas o en unidades de solicitud de escritura WRUs () para las tablas bajo demanda. Si escribe un elemento de 5 KB, se genera un cargo de 5 unidades. El precio de la escritura en una tabla global se expresa en unidades de capacidad de escritura replicadas (rWCUs) para las tablas aprovisionadas o en unidades de solicitud de escritura replicadas (r) para las tablas bajo demanda. WRUs
- Los gastos de RWCu y RWru se cobran en todas las regiones en las que el elemento se escriba directamente o mediante replicación.
- La escritura en un índice secundario global (GSI) se considera una operación de escritura local y utiliza unidades de escritura normales.
- No hay capacidad reservada disponible para r WCUs en este momento. La compra de capacidad reservada WCUs puede seguir siendo beneficiosa para las tablas en las que se GSIs consumen unidades de escritura.
- Al agregar una nueva región a una tabla global, DynamoDB inicia la nueva región automáticamente y le cobra como si se tratara de una restauración de tabla, en función del tamaño de GB de la tabla. También cobra tarifas de transferencia de datos entre regiones.

¿Qué regiones admiten las tablas globales?

Las tablas globales son compatibles con todas las Regiones de AWS.

¿Cómo se GSIs gestionan las tablas globales?

En las tablas globales (actualmente, versión 2019) cuando se crea un GSI en una región, se crea automáticamente en otras regiones participantes y se repone de forma automática.

¿Cómo detengo la replicación de una tabla global?

Puede eliminar una tabla de réplica del mismo modo que eliminaría cualquier otra tabla. Al eliminar la tabla global se detiene la replicación en esa región y se elimina la copia de la tabla guardada en dicha región. No obstante, no se puede detener la replicación mientras se mantienen copias de la tabla como entidades independientes, ni tampoco se puede pausar.

¿Cómo interactúa Amazon DynamoDB Streams con las tablas globales?

Cada tabla global produce un flujo independiente basado en todas sus operaciones de escritura, independientemente de dónde hayan comenzado. Puede elegir consumir el flujo de DynamoDB en una región o en todas las regiones (de forma independiente). Si desea procesar operaciones de escritura locales pero no replicadas, puede agregar su propio atributo de región a cada elemento para identificar la región de escritura. A continuación, puede utilizar un filtro de eventos AWS Lambda para llamar a la función de Lambda solo para las operaciones de escritura en la región local. Esta acción ayuda en las operaciones de inserción y actualización, pero no en las de eliminación.

¿Cómo gestionan las transacciones las tablas globales?

Las operaciones transaccionales proporcionan garantías de atomicidad, uniformidad, aislamiento y durabilidad (ACID, por sus siglas en inglés) solo en la región en la que se creó la operación de escritura originalmente. No se admiten las transacciones entre regiones en las tablas globales. Por ejemplo, si tiene una tabla global con réplicas en las regiones Este de EE. UU. (Ohio) y Oeste de EE. UU. (Oregón) y realiza una operación `TransactWriteItems` en la región Este de EE. UU. (Ohio), puede observar transacciones completadas parcialmente en la región Oeste de EE. UU. (Oregón) a medida que los cambios se replican. Los cambios se replican en otras regiones solo cuando se han confirmado en la región de origen.

¿Cómo interactúan las tablas globales con la memoria caché de DynamoDB Accelerator (DAX)?

Las tablas globales eluden DAX mediante la actualización directa de DynamoDB, por lo que DAX no tiene constancia de que está almacenando datos obsoletos. La memoria caché de DAX solo se actualiza cuando caduca el TTL de la memoria caché.

¿Se propagan las etiquetas de las tablas?

No, las etiquetas no se propagan automáticamente.

¿Debo hacer copias de seguridad de las tablas de todas las regiones o solo de una?

La respuesta depende de la finalidad de la copia de seguridad.

- Si desea garantizar la durabilidad de los datos, DynamoDB ya proporciona esa protección. El servicio garantiza la durabilidad.
- Si desea conservar una instantánea para los registros históricos (por ejemplo, para cumplir los requisitos normativos), la copia de seguridad en una región debería ser suficiente. Puede copiar la copia de seguridad a más regiones mediante AWS Backup.
- Si desea recuperar datos borrados o modificados por error, utilice la point-in-timerecuperación de DynamoDB (PITR) en una región.

¿Cómo puedo implementar tablas globales mediante? AWS CloudFormation

- CloudFormation representa una tabla de DynamoDB y una tabla global como dos recursos independientes: y. `AWS::DynamoDB::Table` `AWS::DynamoDB::GlobalTable` Un enfoque consiste en crear todas las tablas que puedan ser potencialmente globales mediante el constructo `GlobalTable`, mantenerlas inicialmente como tablas independientes y agregar regiones después, si es necesario.
- En CloudFormation, cada tabla global está controlada por una sola pila, en una sola región, independientemente del número de réplicas. Al implementar la plantilla, CloudFormation crea y actualiza todas las réplicas como parte de una operación de pila única. No debe desplegar el mismo recurso AWS::DynamoDB::GlobalTable en varias regiones. Se producirán errores y no se admite. Si despliega la plantilla de aplicación en varias regiones, puede usar condiciones para crear el recurso `AWS::DynamoDB::GlobalTable` en una sola región. O bien, puede optar por definir recursos `AWS::DynamoDB::GlobalTable` en una pila que sea independiente de la pila de aplicaciones y asegurarse de que despliega en una sola región.
- Si tiene una tabla normal y quiere convertirla en una tabla global y, al mismo tiempo, administrarla de la siguiente manera CloudFormation: defina la política de eliminación como`Retain`, elimine

la tabla de la pila, conviértala en una tabla global en la consola y, a continuación, importe la tabla global como un nuevo recurso a la pila. Para obtener más información, consulta el AWS GitHub repositorio amazon-dynamodb-table-to- global-table-cdk.

- En este momento no se admite la replicación entre cuentas.

Conclusión y recursos

Las tablas globales de DynamoDB tienen muy pocos controles, pero aun así requieren una cuidadosa consideración. Debe determinar su modo de escritura, el modelo de enrutamiento y los procesos de evacuación. Debe instrumentar la aplicación en todas las regiones y estar preparado para ajustar el enrutamiento o realizar una evacuación para mantener el estado global. La recompensa es disponer de un conjunto de datos distribuido por todo el mundo con operaciones de lectura y escritura de baja latencia y diseñado para ofrecer una disponibilidad del 99,999%.

Para obtener más información sobre las tablas globales de DynamoDB, consulte los siguientes recursos:

- Documentación de Amazon DynamoDB
- Amazon Route 53 Application Recovery Controller
- Comprobaciones de preparación para ARC (documentación)AWS
- Políticas de enrutamiento de Route 53 (AWS documentación)
- AWS Global Accelerator
- Acuerdo de nivel de servicio de DynamoDB
- AWS Conceptos básicos sobre varias regiones (documento técnico)AWS
- Diseñe patrones de resiliencia de datos con AWS(presentación de re:Invent 2022)AWS
- Cómo se modernizaron Fidelity Investments y Reltio con Amazon AWS DynamoDB (presentación de re:Invent 2022)
- Patrones de diseño y prácticas recomendadas para varias regiones (presentación de re:Invent 2022)AWS
- Arranca la arquitectura de recuperación ante desastres (DR) AWS, tercera parte: Pilot Light and Warm Standby (AWS entrada del blog)
- Utilice la fijación por regiones para establecer una región de origen para los artículos de una tabla global AWS de Amazon DynamoDB (entrada del blog)
- Supervisión de Amazon DynamoDB para garantizar el conocimiento AWS operativo (entrada del blog)
- Escalar DynamoDB: cómo afectan al rendimiento AWS las particiones, las teclas de acceso rápido y la división en función del calor (entrada del blog)

Historial de documentos

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

Cambio	Descripción	Fecha
AWS Global Accelerator Información actualizada	Se corrigieron los puntos finales del enrutamiento de solicitudes de Global Accelerator.	14 de marzo de 2024
Información de soporte actualizada Región de AWS	Se han actualizado las preguntas frecuentes para indicar que las tablas globales ahora son compatibles con todas las Regiones de AWS.	15 de noviembre de 2023
Publicación inicial	—	19 de mayo de 2023

AWS Glosario de orientación prescriptiva

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

Números

Las 7 R

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

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

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

A

ABAC

Consulte control de acceso basado en atributos.

servicios abstractos

Consulte servicios gestionados.

ACID

Consulte atomicidad, consistencia, aislamiento y durabilidad.

migración activa-activa

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

migración activa-pasiva

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

función agregada

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

IA

Véase inteligencia artificial.

AIOps

Consulte las operaciones de inteligencia artificial.

anonimización

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

antipatrones

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

control de aplicaciones

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

cartera de aplicaciones

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

inteligencia artificial (IA)

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

operaciones de inteligencia artificial (AIOps)

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

cifrado asimétrico

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

atomicidad, consistencia, aislamiento, durabilidad (ACID)

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

control de acceso basado en atributos (ABAC)

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

origen de datos fidedigno

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

Zona de disponibilidad

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

AWS Marco de adopción de la nube (AWS CAF)

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

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

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

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

B

Un bot malo

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

BCP

Consulte la planificación de la continuidad del negocio.

gráfico de comportamiento

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

sistema big-endian

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

clasificación binaria

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

filtro de floración

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

implementación azul/verde

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

bot

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

botnet

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

branch

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

acceso con cristales rotos

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

estrategia de implementación sobre infraestructura existente

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

caché de búfer

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

capacidad empresarial

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

planificación de la continuidad del negocio (BCP)

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

C

CAF

Consulte el marco AWS de adopción de la nube.

despliegue canario

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

CCoE

Consulte Cloud Center of Excellence.

CDC

Consulte la captura de datos de cambios.

captura de datos de cambio (CDC)

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

ingeniería del caos

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

CI/CD

Consulte la integración continua y la entrega continua.

clasificación

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

cifrado del cliente

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

Centro de excelencia en la nube (CCoE)

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

computación en la nube

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

modelo operativo en la nube

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

etapas de adopción de la nube

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

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

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

CMDB

Consulte la base de datos de administración de la configuración.

repositorio de código

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

caché en frío

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

datos fríos

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

visión artificial (CV)

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

desviación de configuración

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

base de datos de administración de configuración (CMDB)

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

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

paquete de conformidad

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

integración y entrega continuas (CI/CD)

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

CV

Vea la visión artificial.

D

datos en reposo

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

clasificación de datos

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

desviación de datos

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

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

datos en tránsito

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

malla de datos

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

minimización de datos

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

perímetro de datos

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

preprocesamiento de datos

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

procedencia de los datos

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

titular de los datos

Persona cuyos datos se recopilan y procesan.

almacenamiento de datos

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

lenguaje de definición de datos (DDL)

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

lenguaje de manipulación de datos (DML)

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

DDL

Consulte el lenguaje de definición de bases de datos.

conjunto profundo

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

aprendizaje profundo

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

defense-in-depth

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

administrador delegado

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

Implementación

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

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

entorno de desarrollo

Consulte entorno.

control de detección

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

asignación de flujos de valor para el desarrollo (DVSM)

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

gemelo digital

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

tabla de dimensiones

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

desastre

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

recuperación de desastres (DR)

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

DML

Consulte el lenguaje de manipulación de bases de datos.

diseño basado en el dominio

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

DR

Consulte recuperación ante desastres.

detección de deriva

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

DVSM

Consulte el mapeo del flujo de valor del desarrollo.

E

EDA

Consulte el análisis exploratorio de datos.

EDI

Véase intercambio electrónico de datos.

computación en la periferia

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

intercambio electrónico de datos (EDI)

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

cifrado

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

clave de cifrado

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

endianidad

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

punto de conexión

Consulte el punto final del servicio.

servicio de punto de conexión

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

planificación de recursos empresariales (ERP)

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

cifrado de sobre

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

entorno

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

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

epopeya

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

ERP

Consulte planificación de recursos empresariales.

análisis de datos de tipo exploratorio (EDA)

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

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

F

tabla de datos

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

fallan rápidamente

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

límite de aislamiento de fallas

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

rama de característica

Consulte la sucursal.

características

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

importancia de las características

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

transformación de funciones

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

indicaciones de unos pocos pasos

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

FGAC

Consulte el control de acceso detallado.

control de acceso preciso (FGAC)

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

migración relámpago

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

FM

Consulte el modelo básico.

modelo de base (FM)

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

G

IA generativa

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

bloqueo geográfico

Consulta las restricciones geográficas.

restricciones geográficas (bloqueo geográfico)

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

Flujo de trabajo de Gitflow

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

imagen dorada

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

estrategia de implementación desde cero

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

barrera de protección

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

H

HA

Consulte la alta disponibilidad.

migración heterogénea de bases de datos

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

alta disponibilidad (HA)

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

modernización histórica

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

datos retenidos

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

migración homogénea de bases de datos

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

datos recientes

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

hotfix

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

periodo de hiperatención

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

I

IaC

Vea la infraestructura como código.

políticas basadas en identidad

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

aplicación inactiva

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

IIoT

Consulte Internet de las cosas industrial.

infraestructura inmutable

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

VPC entrante (de entrada)

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

migración gradual

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

Industria 4.0

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

infraestructura

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

infraestructura como código (IaC)

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

Internet de las cosas industrial (T) IIo

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

VPC de inspección

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

Internet de las cosas (IoT)

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

interpretabilidad

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

IoT

Consulte Internet de las cosas.

biblioteca de información de TI (ITIL)

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

administración de servicios de TI (ITSM)

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

ITIL

Consulte la biblioteca de información de TI.

ITSM

Consulte Administración de servicios de TI.

L

control de acceso basado en etiquetas (LBAC)

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

zona de aterrizaje

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

modelo de lenguaje grande (LLM)

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

migración grande

Migración de 300 servidores o más.

LBAC

Consulte control de acceso basado en etiquetas.

privilegio mínimo

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

migrar mediante lift-and-shift

Ver 7 Rs.

sistema little-endian

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

LLM

Véase un modelo de lenguaje amplio.

entornos inferiores

Véase entorno.

M

machine learning (ML)

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

rama principal

Ver sucursal.

malware

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

servicios gestionados

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

sistema de ejecución de fabricación (MES)

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

MAP

Consulte Migration Acceleration Program.

mecanismo

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

cuenta de miembro

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

MES

Consulte el sistema de ejecución de la fabricación.

Transporte telemétrico de Message Queue Queue (MQTT)

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

microservicio

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

arquitectura de microservicios

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

Programa de aceleración de la migración (MAP)

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

migración a escala

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

fábrica de migración

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

metadatos de migración

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

patrón de migración

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

Migration Portfolio Assessment (MPA)

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

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

Evaluación de la preparación para la migración (MRA)

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

estrategia de migración

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

ML

Consulte el aprendizaje automático.

modernización

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

evaluación de la preparación para la modernización

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

aplicaciones monolíticas (monolitos)

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

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

MAPA

Consulte la evaluación de la cartera de migración.

MQTT

Consulte Message Queue Queue Telemetría y Transporte.

clasificación multiclase

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

infraestructura mutable

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

O

OAC

Consulte el control de acceso de origen.

OAI

Consulte la identidad de acceso de origen.

OCM

Consulte gestión del cambio organizacional.

migración fuera de línea

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

OI

Consulte integración de operaciones.

OLA

Véase el acuerdo a nivel operativo.

migración en línea

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

OPC-UA

Consulte Open Process Communications: arquitectura unificada.

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

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

acuerdo de nivel operativo (OLA)

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

revisión de la preparación operativa (ORR)

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

tecnología operativa (OT)

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

integración de operaciones (OI)

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

registro de seguimiento organizativo

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

administración del cambio organizacional (OCM)

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

control de acceso de origen (OAC)

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

identidad de acceso de origen (OAI)

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

ORR

Consulte la revisión de la preparación operativa.

OT

Consulte la tecnología operativa.

VPC saliente (de salida)

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

P

límite de permisos

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

información de identificación personal (PII)

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

PII

Consulte la información de identificación personal.

manual de estrategias

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

PLC

Consulte controlador lógico programable.

PLM

Consulte la gestión del ciclo de vida del producto.

policy

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

persistencia políglota

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

evaluación de cartera

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

predicate

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

pulsar un predicado

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

control preventivo

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

entidad principal

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

privacidad desde el diseño

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

zonas alojadas privadas

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

control proactivo

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

gestión del ciclo de vida del producto (PLM)

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

entorno de producción

Consulte el entorno.

controlador lógico programable (PLC)

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

encadenamiento rápido

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

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

seudonimización

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

publish/subscribe (pub/sub)

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

Q

plan de consulta

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

regresión del plan de consulta

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

R

Matriz RACI

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

RAG

Consulte Retrieval Augmented Generation.

ransomware

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

Matriz RASCI

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

RCAC

Consulte control de acceso por filas y columnas.

réplica de lectura

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

rediseñar

Ver 7 Rs.

objetivo de punto de recuperación (RPO)

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

objetivo de tiempo de recuperación (RTO)

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

refactorizar

Ver 7 Rs.

Región

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

regresión

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

volver a alojar

Consulte 7 Rs.

versión

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

trasladarse

Ver 7 Rs.

redefinir la plataforma

Ver 7 Rs.

recompra

Ver 7 Rs.

resiliencia

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

política basada en recursos

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

matriz responsable, confiable, consultada e informada (RACI)

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

control receptivo

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

retain

Consulte 7 Rs.

jubilarse

Ver 7 Rs.

Generación aumentada de recuperación (RAG)

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

rotación

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

control de acceso por filas y columnas (RCAC)

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

RPO

Consulte el objetivo del punto de recuperación.

RTO

Consulte el objetivo de tiempo de recuperación.

manual de procedimientos

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

S

SAML 2.0

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

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

SCADA

Consulte el control de supervisión y la adquisición de datos.

SCP

Consulte la política de control de servicios.

secreta

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

seguridad desde el diseño

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

control de seguridad

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

refuerzo de la seguridad

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

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

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

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

automatización de la respuesta de seguridad

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

cifrado del servidor

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

política de control de servicio (SCP)

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

punto de enlace de servicio

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

acuerdo de nivel de servicio (SLA)

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

indicador de nivel de servicio (SLI)

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

objetivo de nivel de servicio (SLO)

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

modelo de responsabilidad compartida

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

SIEM

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

punto único de fallo (SPOF)

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

SLA

Consulte el acuerdo de nivel de servicio.

SLI

Consulte el indicador de nivel de servicio.

SLO

Consulte el objetivo de nivel de servicio.

split-and-seed modelo

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

SPOF

Consulte el punto único de falla.

esquema en forma de estrella

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

patrón de higo estrangulador

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

subred

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

supervisión, control y adquisición de datos (SCADA)

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

cifrado simétrico

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

pruebas sintéticas

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

indicador del sistema

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

T

etiquetas

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

variable de destino

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

lista de tareas

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

entorno de prueba

Consulte entorno.

entrenamiento

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

puerta de enlace de tránsito

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

flujo de trabajo basado en enlaces troncales

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

acceso de confianza

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

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

ajuste

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

equipo de dos pizzas

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

U

incertidumbre

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

tareas indiferenciadas

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

entornos superiores

Ver entorno.

V

succión

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

control de versión

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

Emparejamiento de VPC

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

vulnerabilidad

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

W

caché caliente

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

datos templados

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

función de ventana

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

carga de trabajo

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

flujo de trabajo

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

GUSANO

Mira, escribe una vez, lee muchas.

WQF

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

escribe una vez, lee muchas (WORM)

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

Z

ataque de día cero

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

vulnerabilidad de día cero

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

aviso de tiro cero

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

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

aplicación zombi

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

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