



Evaluar su proyecto de aprendizaje automático con la lista de verificación MLOps

AWS Guía prescriptiva



AWS Guía prescriptiva: Evaluar su proyecto de aprendizaje automático con la lista de verificación MLOps

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Las marcas comerciales y la imagen comercial de Amazon no se pueden utilizar en relación con ningún producto o servicio que no sea de Amazon, de ninguna manera que pueda causar confusión entre los clientes y que menosprecie o desacredite a Amazon. Todas las demás marcas registradas que no son propiedad de Amazon son propiedad de sus respectivos propietarios, que pueden o no estar afiliados, conectados o patrocinados por Amazon.

Table of Contents

Introducción	1
MLOps componentes de la lista de verificación	2
1. Administración centrada en los datos	2
2. Experimentación	4
3. Observabilidad y gestión de modelos	6
4. Procesos robustos y promoción	9
5. Integración continua	12
6. Monitorización continua	15
7. Implementación continua	19
8. Formación continua	21
9. Gobernanza	23
Ejemplo de uso	25
Recursos	26
Historial de documentos	27
Glosario	28
#	28
A	29
B	32
C	34
D	37
E	42
F	44
G	46
H	47
I	49
L	51
M	52
O	57
P	60
Q	63
R	63
S	66
T	70
U	72

V	72
W	73
Z	74
.....	lxxv

Evaluación de su proyecto de aprendizaje automático con la MLOps lista de verificación

Charles Frenzel, Sharath Nagaraja y Spencer Romo Amazon Web Services (AWS)

julio 2023 ([historial de documentos](#))

La [MLOps lista de verificación es una lista](#) de verificación viable que puede utilizar en cualquier fase de su proyecto de aprendizaje automático (ML). La lista de verificación es una herramienta para evaluar la preparación general, examinar la cobertura del sistema e identificar nuevas áreas de oportunidad en los sistemas de aprendizaje automático distribuidos. MLOps es la combinación de personas, tecnología y procesos para ofrecer soluciones de aprendizaje automático. Una buena arquitectura MLOps ayuda a las empresas a implementar los modelos de aprendizaje automático en la producción de forma eficaz y coherente, y puede ofrecer valor empresarial.

El uso de la MLOps lista de verificación le ayuda a hacer lo siguiente:

- Evalúe su MLOps sistema.
- Encuentre áreas de oportunidad.
- Encuentre áreas de mejora.
- Evalúe y actualice su hoja de ruta estratégica en AWS.
- Genere elementos pendientes.

Te recomendamos que utilices la MLOps lista de verificación al principio del MLOps proyecto, pero es posible utilizar partes de ella durante cualquier fase.

MLOps componentes de la lista de verificación

La [MLOps lista de verificación](#) consta de nueve componentes diferentes. Cada componente representa un aspecto de un MLOps sistema maduro. Esta sección cubre los componentes de la lista de verificación en detalle. Tenga en cuenta que hay cierta superposición en esta sección porque algunos componentes dependen de otros componentes.

La lista de verificación cubre las siguientes áreas:

- [1. Administración centrada en los datos](#)
- [2. Experimentación](#)
- [3. Observabilidad y gestión de modelos](#)
- [4. Procesos robustos y promoción](#)
- [5. Integración continua](#)
- [6. Monitorización continua](#)
- [7. Implementación continua](#)
- [8. Formación continua](#)
- [9. Gobernanza](#)

Nota: La seguridad ya está bien cubierta por AWS. Recomendamos consultar las guías y patrones de AWS Prescriptive [Guidance](#) relacionados con la seguridad [y](#) revisar las prácticas recomendadas de seguridad de [AWS](#).

1. Administración centrada en los datos

La gestión de datos consiste en garantizar que los datos utilizados en la formación, las pruebas y la inferencia se gestionen, protejan y validen adecuadamente. Cuando se crean modelos a escala, los datos son el principal recurso que permite un alto rendimiento del modelo.

1.1 Repositorio de datos

Un repositorio de datos requiere la capacidad de rastrear los datos y ver su punto de origen. Cuando se agregan o eliminan datos nuevos, el repositorio de datos registra esos cambios durante la point-in-time recuperación. El

repositorio de datos debe tener en cuenta cómo se rastrean y procesan los datos de las etiquetas y cómo se rastrean los artefactos de datos intermedios.

1.2 Integración de diversas fuentes de datos

Según la aplicación, el entrenamiento del modelo puede requerir datos de muchas fuentes. Diseñar y mantener un manifiesto que informe a los profesionales del aprendizaje automático sobre las fuentes de datos disponibles y cómo se relacionan entre sí es fundamental para crear modelos.

1.3 Validación del esquema de datos

Para alimentar los datos de los modelos, es importante que los datos de entrenamiento sean homogéneos. Es posible que se requieran transformaciones u otros análisis exploratorios para los datos que se almacenan en soluciones de lagos de datos, como Amazon Simple Storage Service (Amazon S3), o en almacenes de datos de documentos.

1.4 Control de versiones y linaje de los datos

Al entrenar modelos que puedan usarse en producción, debe poder reproducir los resultados y disponer de una forma fiable de realizar los [estudios de ablación](#) para comprender mejor el rendimiento general del modelo. El seguimiento del estado de los datos de entrenamiento es fundamental para lograr esta reproducibilidad. Herramientas como el [control de versiones de datos \(DVC\)](#) pueden ayudar en este sentido.

1.5 Flujo de trabajo de etiquetado

En los casos en que los datos etiquetados no estén disponibles al inicio del proyecto, la creación de los datos etiquetados suele ser un paso necesario. Herramientas como Amazon SageMaker Ground Truth requieren que los datos de entrada estén estructurados adecuadamente y requieren un trabajo de etiquetado definido y probado. Se debe utilizar una plantilla de etiquetadores internos o externos. Luego, los datos deben validarse utilizando enfoques de etiquetado redundante o aprendizaje automático para identificar valores atípicos o errores en el conjunto de datos de entrenamiento.

1.6 Almacenamiento de funciones en línea y fuera de línea

El sistema ML tiene un almacén de [funciones o un almacén](#) centralizado para las funciones y los metadatos asociados, de modo que es posible reutilizar las funciones o las entradas del modelo. Puede crear un almacenamiento en línea o sin conexión. Utilice una tienda en línea para casos de uso de inferencias en tiempo real y de baja latencia. Use una tienda fuera de línea para el entrenamiento y la inferencia por lotes.

2. Experimentación

La experimentación abarca el registro, el seguimiento y las métricas de los experimentos. Esto se traduce en la integración de los metadatos de los experimentos en la plataforma, en el control de código fuente y en los entornos de desarrollo. La experimentación también incluye la posibilidad de optimizar el rendimiento y la precisión del modelo mediante la depuración.

2.1 Entornos de desarrollo integrados

Un entorno de desarrollo integrado (IDE) se integra directamente con la nube. El IDE puede

interactuar con el sistema más grande y enviar comandos a él. Lo ideal es que sea compatible con lo siguiente:

- Desarrollo local
- Integración de control de versiones
- La depuración está implementada y todos los registros y artefactos generados pasan al control de versiones

2.2 Control de versiones de código

Para garantizar la reproducibilidad y la reutilización, todo el código se archiva en el repositorio fuente con el control de versiones adecuado. Esto incluye el código de infraestructura, el código de aplicación, el código de modelo e incluso los cuadernos (si opta por usarlos).

2.3 Seguimiento

Un proyecto de aprendizaje automático requiere una herramienta que pueda rastrear y analizar los experimentos de aprendizaje automático. Esta herramienta debe registrar todas las métricas, parámetros y artefactos durante la ejecución de un experimento de aprendizaje automático, registrando todos los metadatos en una ubicación central. La ubicación central permitirá analizar, visualizar y auditar todos los experimentos que ejecute.

2.4 Integración multiplataforma

Los resultados históricos de los experimentos y todos sus metadatos están accesibles en otras partes del sistema. Por ejemplo, las canalizaciones de orquestación existentes pueden acceder a estos datos, al igual que las herramientas de monitoreo.

2.5 Depuración: precisión y rendimiento del sistema

Existe un marco completo de depuración de modelos para examinar las ejecuciones en busca de lo siguiente:

- Encuentre cuellos de botella
- Alerta sobre anomalías
- Maximice la utilización de recursos
- Ayuda en el análisis de los experimentos

Cuando la formación es intensiva, la capacidad de maximizar el rendimiento es crucial y la convierte en una herramienta necesaria para la optimización de costes.

3. Observabilidad y gestión de modelos

La sección de observabilidad y gestión de modelos de la lista de verificación abarca el control de las versiones de los modelos y el seguimiento del linaje en todo el sistema de aprendizaje automático. El control de versiones de los modelos ayuda a rastrear y controlar todos los cambios aplicados a un modelo para que pueda recuperar una versión anterior cuando sea necesario. El seguimiento del linaje proporciona una vista de las entradas y salidas del modelo. Otro beneficio clave del seguimiento del linaje es la point-in-time recuperación (PITR), que automatiza la implementación y la recuperación del sistema.

3.1 Registro de modelos versionados

En general, un registro de modelos permite el control de versiones y el seguimiento del linaje de los componentes del modelo. Un buen registro puede asociar metadatos al modelo versionado, incluidos los siguientes:

- Los datos utilizados
- Información sobre el modelo
- Resultados de las métricas de evaluación
- Código de modelo asociado

3.2 Sesgo, imparcialidad y explicabilidad

Como mínimo, un sistema de aprendizaje automático debe tener un proceso mediante el cual las predicciones de un modelo sean explicables a otras partes. Los usuarios deberían poder comprobar si los resultados están sesgados por cada característica. Lo ideal es medir el sesgo de los datos antes de introducirlos en el modelo de aprendizaje automático y registrar estas métricas para las tarjetas modelo y las auditorías.

3.3 Seguimiento del linaje: entradas y salidas de datos

El seguimiento se realiza para seguir el flujo de datos que entran y salen del sistema (por ejemplo, desde el lago de datos hasta el proceso de formación). Este seguimiento actúa como un registro a partir del cual se pueden recrear todos los procesos del sistema y proporciona una pista de auditoría para su análisis.

3.4 Seguimiento del linaje: información ambiental

Este seguimiento captura información sobre la configuración del entorno de ejecución, como las imágenes de los contenedores para todo el código del modelo y las dependencias asociadas a los contenedores.

3.5 Seguimiento del linaje: modelo

Este seguimiento captura información sobre el modelo. Incluye todo, desde información sobre el algoritmo del modelo hasta los parámetros e hiperparámetros incluidos en el modelo.

3.6 Integración con el despliegue y la supervisión

El sistema debe estar vinculado directamente con los subsistemas de monitoreo y despliegue del PITR. En el caso de la supervisión, esto significa comparar el rendimiento del modelo con sus ciclos de entrenamiento para detectar el deterioro de la calidad del modelo. Para la implementación, esto es compatible con la PITR y la posibilidad de volver a una versión anterior del modelo según sea necesario.

3.7 Configuración de parámetros de canalización

Técnicamente, la configuración de los parámetros de la canalización se incluye tanto en el seguimiento del linaje como en el seguimiento de los experimentos, ya que la configuración de la canalización debe estar versionada y asociada directamente a un modelo. La configuración de los parámetros de canalización se detalla en esta sección porque es imprescindible realizar un seguimiento de todas las configuraciones de orquestación del sistema y versionarlas.

3.8 Los problemas se pueden rastrear, depurar y reproducir.

Un ingeniero puede rastrear, depurar y reproducir todos los problemas del sistema sin mucho esfuerzo. Esto implica que existe un nivel suficiente de observabilidad. Esta verificación se basa principalmente en el cumplimiento de los demás elementos de la sección de observabilidad y gestión de modelos.

3.9 Visualización del rendimiento

El sistema puede capturar y recopilar registros en un formato tipo base de datos de series temporales e ingerirlos directamente en el panel de control. El panel de control proporciona una visión holística de las métricas del modelo y de la computadora, con la posibilidad de desglosar y realizar consultas.

4. Procesos robustos y promoción

Las canalizaciones ofrecen muchas opciones para el ajuste de hiperparámetros, el AutoML y las rutinas de procesamiento. Las canalizaciones se registran de un extremo a otro. Los pipelines robustos pueden ejecutar la capacitación en paralelo en múltiples instancias y marcos, escalando los tamaños de carga según sea necesario. Las canalizaciones sólidas pueden promover los modelos para que pasen a la fase de producción, se desplieguen en tiempo real, se transmitan o se distribuyan por lotes. Estas implementaciones pueden admitir la inferencia de un solo modelo o de varios modelos.

4.1 Capacitación distribuida y a gran escala

Un sistema de aprendizaje automático maduro admite la capacidad de ejecutar el entrenamiento en grandes instancias optimizadas para la computación en paralelo. Cuenta con las herramientas necesarias para garantizar que estos recursos se utilicen en su totalidad y que la capacitación se amplíe de manera uniforme en todo el clúster de procesamiento.

4.2 Support para múltiples marcos

Los desarrolladores pueden portar diferentes marcos de plataformas, como PyTorch Flax, para realizar trabajos de formación e inferencia. Del mismo modo, se admiten y se pueden utilizar diferentes idiomas y versiones. Cambiar a otro marco no estropeará el sistema.

4.3 Ajuste de hiperparámetros

Un paso de ajuste de hiperparámetros forma parte del proceso de formación. Los hiperpará

metros de los modelos implementados se ajustan. Hay varias opciones disponibles para ajustar los hiperparámetros. Para mejorar la precisión, al menos una de las opciones de ajuste debe tener un enfoque o inferencia bayesiano.

4.4 Opción AutoML

Para reducir la experimentación y la comparación manuales, un sistema de aprendizaje automático avanzado admite la ejecución de AutoML, que selecciona automáticamente la mejor canalización de funciones, hiperparámetros y modelo. Tenga en cuenta que AutoML es una función que se puede usar de forma pragmática, pero no es una panacea.

4.5 Soporte de inferencias: en tiempo real

Esto se denomina comúnmente modelo como servicio (MaaS). El sistema admite la inferencia en tiempo real con las operaciones de la API REST, para realizar solicitudes de inferencia a pedido. Es capaz de ofrecer una infraestructura MaaS en la que el modelo puede escalarse tanto horizontal como verticalmente como una API independiente o como un punto final asociado a otras aplicaciones. Como alternativa, es posible implementarla mediante tecnología sin servidor.

4.6 Soporte de inferencias: streaming

Los modelos se pueden promover a un formato de inferencia en tiempo real, como Amazon Kinesis o Amazon Managed Streaming for Apache Kafka, mediante el cual la inferencia se ejecuta en streaming en el modelo. Esto requiere completar al menos el 90 por ciento de la lista de verificación, ya que las barreras, la observabilidad y el monitoreo son esenciales para la inferencia en tiempo real.

4.7 Soporte de inferencias: por lotes

El sistema admite el despliegue por lotes de modelos como trabajos programados o iniciados. El sistema puede ejecutar modelos como parte de un proceso de extracción, transformación y carga (ETL) o de forma aislada. Los trabajos por lotes registran el estado de cada paso y se ejecutan siguiendo un patrón ordenado, como un gráfico acíclico dirigido. Como alternativa, los trabajos pueden escribir en una base de datos, que actúa como servidor de inferencia de modelos.

4.8 Rutinas de preprocesamiento y posprocesamiento

Cuando es necesario, los datos se incluyen como parte del proceso de admisión del modelo o de los trabajos por lotes. Si hay varios modelos o varios pasos en juego, las rutinas de posprocesamiento se encargan de caracterizar los datos.

4.9 Capacidad para invocar modelos jerárquicos o simultáneos

El sistema ML puede implementar muchos modelos juntos o ejecutarlos de forma secuencial. Lo primero significa alojar una flota de recursos en un único modelo de punto final. Esto último implica que varios modelos deben ejecutarse de forma encadenada uno tras otro. El sistema puede gestionar estos dos tipos de complejidad de forma resiliente.

4.10 Estrategias de escalado horizontal y vertical

Una canalización debe tener la capacidad de soportar ambos tipos de estrategias de escalado, tanto para la formación como para la inferencia. El sistema ML puede aumentar su tamaño y distribuir el tráfico entre varias máquinas cuando aumentan la latencia o el rendimiento. Se establecen políticas para este tipo de comportamiento y tienen en cuenta la asignación óptima de los recursos.

4.11 registro End-to-end

El equipo de desarrollo debe tener el registro integrado en todo el código de canalización para que el registro pueda capturar las entradas, las salidas y los pasos intermedios del sistema. El registro debería permitir el seguimiento de las ejecuciones en proceso y la depuración de errores.

5. Integración continua

El sistema ML ejecuta pruebas para validar que el sistema funciona de principio a fin, comprobando los posibles puntos de fallo. Las pruebas se ejecutan automáticamente al confirmarse, y las pruebas más largas se ejecutan según un cronograma fijo. Las pruebas comprueban las áreas tradicionales de la ingeniería de software, como a nivel de unidad y sistema. Además, las pruebas capturan los detalles del aprendizaje automático al comprobar los datos, las características y el modelo.

5.1 Verificaciones de códigos locales

Antes de guardar el código en un repositorio de código centralizado, los desarrolladores realizan comprobaciones a nivel local, como pruebas unitarias básicas y análisis estáticos. Al ejecutar estas comprobaciones antes de confirmarlas, se aumenta la calidad general del código y se detectan los problemas antes de que se introduzcan en el control de versiones.

5.2 Análisis de código estático

El repositorio de código central tiene herramientas de análisis de código estático que se ejecutan rápidamente al confirmarse. Estas herramientas deberían mejorar el estilo y el formato del código. También debería comprobar si hay vulnerabilidades de seguridad comunes en el código fuente y de infraestructura, errores comunes y otros puntos débiles del código.

5.3 Pruebas de calidad de datos

Las pruebas de calidad de los datos deberían comprobar, como mínimo, que los datos no hayan infringido un esquema fijo. Un enfoque más completo consiste en calcular las estadísticas de los datos en el momento de la ingesta, establecer restricciones a los datos y realizar pruebas con ellas.

Es posible configurar las pruebas de calidad de los datos de forma independiente o como parte del proceso. Las estadísticas y las restricciones se reutilizan para la supervisión.

5.4 Pruebas de características

Como parte de un proceso completo, se genera la importancia de las características. Las pruebas de características afirman que la importancia de las características, o la forma en que el modelo atribuye los valores de las características, no cambia. Las pruebas de características pueden contribuir a la supervisión porque pueden alertar y rastrear las infracciones en las entradas de un modelo.

5.5 Pruebas unitarias

Las pruebas unitarias de todo el código (modelo, aplicación e infraestructura) se ejecutan antes de la confirmación y durante la confirmación. Cada prueba unitaria permite comprobar una parte importante del código para confirmar que funciona según lo esperado. En el caso del código ML, se pueden ejecutar pruebas para comprobar la corrección algorítmica.

5.6 Pruebas de integración

Una prueba de integración verifica que la tubería funcione correctamente de principio a fin, incluida la instalación de la infraestructura asociada a la tubería. Esta prueba valida que el sistema funciona y se registra según lo esperado. Si la implementación es independiente, también debería realizarse una end-to-end prueba para garantizar que la implementación funcione.

5.7 Pruebas de humo

El sistema cuenta con pruebas de humo que se ejecutan en una regresión mínima y rápida de cada una de las funcionalidades. Las pruebas de humo forman parte de la integración continua y pueden ejecutarse en un entorno contenerizado para imitar la funcionalidad de la nube.

5.8 Pruebas de carga

Se están realizando pruebas de carga bajo demanda. Además de capturar el comportamiento del sistema de aprendizaje automático con cargas altas y bajas, las pruebas de carga proporcionan estadísticas sobre el rendimiento o la latencia de todo el sistema. Los datos recopilados mediante las pruebas de carga proporcionan información sobre el tamaño de los recursos y las políticas de escalado.

5.9 Modele pruebas funcionales

Las entradas y salidas del modelo se someten a pruebas funcionales automatizadas. Para comprobar un comportamiento dentro de una capacidad, tanto las salidas como las entradas del modelo se prueban con datos reales o falsos con ejemplos básicos.

5.10 Modele pruebas de inferencia con casos extremos

Como parte de las pruebas de funcionalidad mínima, las pruebas del modelo deberían comprobar si se ha producido un comportamiento extremo con determinadas variables antes de promocionarlo. Esto coloca una barandilla adicional para ayudar a prevenir comportamientos inesperados.

6. Monitorización continua

En el monitoreo continuo, los procesos automatizados observan y detectan los problemas de rendimiento y los problemas del modelo. De este modo, los propietarios pueden identificar posibles problemas y amenazas en tiempo real para abordarlos rápidamente.

El monitoreo continuo revela posibles problemas en el modelo, como la calidad de los datos, el cambio de distribución, el cambio de concepto del modelo y la degradación de la calidad del modelo. La supervisión continua también incluye un registro exhaustivo de las medidas tradicionales del sistema, como la saturación, la latencia, el tráfico y los errores. Se ha establecido una práctica estrategia de notificación y alerta para notificar a los propietarios cuando surjan problemas.

6.1 Monitorización de modelos: detección de la calidad de los datos

La monitorización basada en reglas permite saber cuándo los datos entrantes se desvían de los datos de entrenamiento del modelo. Este tipo de supervisión crea un esquema a partir de los datos de entrenamiento, establece restricciones en función de ese esquema y, a continuación, ejecuta excepciones cuando se produce una infracción.

6.2 Monitorización de modelos: cambio de distribución

La supervisión está configurada para observar la distribución de los datos entrantes y comprobar que no se ha desviado de la distribución de los datos de entrenamiento del modelo. Por ejemplo, los datos entrantes se muestrean [como una ventana móvil](#) sobre los datos de inferencia. Luego, se ejecuta un trabajo para probar la distribución muestreada y la distribución de entrenamiento para ver si son iguales.

6.3 Supervisión del modelo: desviación conceptual del modelo

Una verificación de la desviación conceptual busca que la relación entre las entradas de un modelo y la variable objetivo permanezca inalterada con respecto a los datos de entrenamiento. Una comprobación adicional consiste en confirmar que las características relativas y su importancia no cambian.

6.4 Supervisión del modelo: comprobación de la evaluación del modelo

Se trata de un control de seguimiento que evalúa si la calidad del modelo se ha degradado. La verificación de evaluación del modelo compara las métricas de evaluación de referencia del tiempo de entrenamiento con los resultados entrantes para evaluar si el nivel de precisión del modelo ha disminuido con los nuevos datos. Como calcula las métricas de

	<p>precisión, esta comprobación requiere disponer de la veracidad fundamental de los nuevos datos tras la inferencia.</p>
6.5 Capturas del sistema: esquemas de entrada	<p>El sistema ML captura el esquema de los datos de entrenamiento, pruebas y validación. Además de proporcionar información sobre las entradas, los esquemas proporcionan estadísticas sobre su sesgo e integridad. Los esquemas se utilizan para realizar pruebas inmediatas y comprobar la calidad de los datos en la producción.</p>
6.6 Capturas del sistema: resultados de la evaluación y estadísticas	<p>El sistema ML genera información precisa sobre los datos de validación y entrenamiento. Puede generar las predicciones y las etiquetas reales de las ejecuciones de validación y entrenamiento. Se utilizan como restricciones de monitoreo para el modelo de producción en vivo.</p>
6.7 Capturas del sistema: anomalías	<p>Existe un mecanismo de seguimiento para detectar las anomalías en los flujos de datos entrantes. Si se producen valores atípicos en los datos entrantes o si durante un período de tiempo específico cambia la distribución de las características clave, el sistema reconoce que se trata de una anomalía y la marca.</p>

6.8 Registro: saturación y recursos

Se está registrando el grado de llenado del sistema. Las métricas de recursos y saturación deben centrarse en el uso de la CPU, el uso de la unidad de procesamiento gráfico (GPU), el uso de la memoria y el uso del disco. Estas métricas deben estar disponibles en formato de series temporales con la capacidad de medirse en percentiles. En el caso de los trabajos por lotes, proporciona información sobre el rendimiento, que muestra cuántas unidades de información puede procesar el sistema en cada período de tiempo.

6.9 Registro: latencia

Debe existir un registro para medir el retraso en la comunicación de la red o el tiempo que se tarda en atender una solicitud. Un ingeniero debería poder determinar cuánto tardan los modelos de inferencia en realizar las predicciones y cuánto tardan en cargarse.

6.10 Registro: tráfico

La configuración de registro del tráfico mide el volumen de tráfico en cada instancia. El tráfico se mide por la cantidad de solicitudes HTTP y de bytes o paquetes enviados o recibidos durante un período de tiempo determinado. El registro del tráfico proporciona información sobre la carga de trabajo total que se deposita en un sistema.

6.11 Registro: errores

La configuración de registro de errores captura el número de solicitudes que fallan. Los errores son de los siguientes tipos:

- Explícitos (por ejemplo, errores HTTP 500)
- Implícito (por ejemplo, una respuesta correcta de HTTP 200 acompañada de un contenido incorrecto)
- Política (por ejemplo, si te comprometes a un tiempo de respuesta de un segundo, cualquier solicitud superior a un segundo será un error)

Si los códigos de respuesta del protocolo no son suficientes para expresar todas las condiciones de error, es posible que se necesiten protocolos secundarios (internos) para rastrear los modos de falla parcial.

6.12 Notificaciones y alertas

Las notificaciones y alertas se configuran a partir de la supervisión. Las notificaciones incluyen la posibilidad de recibir Slack, notificaciones por correo electrónico, páginas y mensajes del Servicio de Mensajes Cortos (SMS). Alertar no significa enviar notificaciones sobre todas las posibles infracciones. Por el contrario, significa establecer alertas sobre excepciones específicas que sean significativas e importantes para el equipo de desarrollo. De esta forma, se evita la fatiga de las alertas.

7. Implementación continua

Para que un sistema de aprendizaje automático se implemente de forma continua, debe poder desviar el tráfico desde o entre los modelos activos. Un sistema que se implementa de forma

continua tiene al menos una forma de promover los modelos a producción: canary, shadow, blue/green, or A/B. Confirme que en el sistema ML hay al menos una forma de revertir los modelos.

7.1 Cambio de modelo

El sistema puede cambiar entre modelos versionados durante la puesta en escena y la producción. Puede desviar el tráfico de una sola vez o de forma incremental hacia nuevas variantes de producción.

7.2 Procesos de promoción de modelos

Para la promoción de modelos existe un proceso de validación por etapas. El proceso utiliza pruebas fuera de línea que no afectan al sistema de producción, como la ejecución de pruebas con datos de validación en un entorno provisional. Se establecen un manual y métricas para la promoción de modelos. La promoción sigue una de las estrategias de lanzamiento.

7.3 Estrategias de reversión

Existe una estrategia de reversión para que, cuando se produzca un error o el modelo se desvíe del comportamiento esperado, se produzca una reversión, una alternativa o una reversión. En una reversión, el modelo vuelve a una versión de implementación anterior. Como alternativa, el modelo se sustituye por una heurística sólida. Roll Through promoverá la producción del siguiente modelo, al igual que el modelo anterior. Existen guías para todas estas estrategias.

7.4 Despliegue de Canary

El sistema se puede implementar mediante un canario. Inicialmente, una pequeña parte del tráfico se envía al nuevo modelo. Con el tiempo, todo el tráfico pasa al nuevo modelo. Este cambio se supervisa de cerca porque las

pruebas se realizan en el entorno de producción.

7.5 Modele un despliegue paralelo

El sistema puede ejecutar un despliegue paralelo en el que el nuevo modelo funcione junto con el modelo existente. Ambos modelos reciben tráfico, pero solo el modelo anterior genera inferencias. Las evaluaciones se ejecutan en el nuevo modelo en comparación con el modelo existente y, a continuación, el nuevo modelo se promueve manualmente.

7.6 Despliegue azul/verde

El sistema se puede implementar con un modelo nuevo (verde, que está en fase de pruebas) y con el modelo anterior (azul, que es de producción), y ambos funcionan al mismo tiempo. Una vez finalizadas las pruebas, el tráfico se desvía del entorno azul al verde. Esta estrategia evita el tiempo de inactividad debido a que se crean entornos idénticos.

7.7 Support para pruebas A/B o más

El sistema admite el uso de versiones de modelos en el entorno implementado para ejecutar pruebas A/B en el tráfico entrante. Esto puede incluir la posibilidad de promocionarse automáticamente en función del modelo más nuevo que gane en las pruebas. En las configuraciones más avanzadas se utilizará un proceso de [bandido con varios brazos](#).

8. Formación continua

La formación continua significa que el sistema de aprendizaje automático reentrena de forma automática y continua los modelos de aprendizaje automático para que se adapten a los cambios en los datos antes de volver a distribuirlos. Entre los posibles factores desencadenantes de la

reconstrucción se incluyen los cambios en los datos, los cambios en el modelo o los cambios en el código.

8.1 Comprobaciones: validación de las entradas del modelo

Existen comprobaciones para verificar que la entrada de un modelo no se desvíe de un estándar determinado. La validación de entradas implica realizar pruebas funcionales durante la promoción del modelo. También implica la verificación inmediata de las solicitudes de entrada, por ejemplo, mediante el uso de aserciones y tipos enumerados.

8.2 Activación del reentrenamiento: tareas programadas

Esta es la forma más básica de automatización de la formación. El reentrenamiento del modelo se establece según un cronograma (por ejemplo, cada semana). En este escenario, es probable que la automatización sea baja y que los resultados se revisen manualmente y se comprueben al azar antes de la promoción del modelo.

8.3 Activación del reentrenamiento: nuevos datos de entrenamiento

El reentrenamiento se inicia con un umbral de datos entrantes. El modelo puede volver a entrenarse desde cero o ejecutar las actualizaciones de forma incremental. Si se dispone de una cantidad específica de datos, se inicia un trabajo de formación.

8.4 Activación del reentrenamiento: degradación del rendimiento del modelo

Esta técnica utiliza la supervisión y la observabilidad para ejecutar el reentrenamiento del modelo y requiere un nivel avanzado de automatización. Por ejemplo, la precisión disminuye a partir de un rango determinado, lo que desencadena el reentrenamiento de un modelo en todos los datos o en parte de ellos.

8.5 Activación del reentrenamiento: cambio en la distribución de datos

La supervisión del cambio en la distribución de los datos proporciona una forma de reentrenar el modelo cuando cambian sus datos subyacentes. Una infracción basada en un cambio de [concepto o un cambio de distribución de datos inicia un trabajo](#) de readiestramiento del modelo.

9. Gobernanza

La gobernanza del aprendizaje automático abarca un conjunto de procesos y marcos que ayudan a implementar los modelos de aprendizaje automático. Incluye la explicabilidad del modelo, la auditabilidad, la trazabilidad y otros requisitos más abstractos pero esenciales para un ciclo de vida exitoso del aprendizaje automático. end-to-end

9.1 Calidad y conformidad de los datos

El sistema ML tiene en cuenta los aspectos de la información de identificación personal (PII), incluida la anonimización. Ha documentado y revisado el linaje a nivel de columna para comprender el origen, la calidad y la idoneidad de los datos. También cuenta con controles automatizados de calidad de los datos para detectar anomalías.

9.2 Auditoría y documentación

El sistema ML tiene un registro completo de todos los cambios durante el desarrollo, incluidos los experimentos realizados y los motivos por los que se han tomado decisiones para cumplir con la normativa.

9.3 Reproducibilidad y trazabilidad

El sistema ML incluye una instantánea completa de los datos para una reinstanciación precisa y rápida del modelo, o tiene la capacidad de recrear el entorno y volver a entrenarlo con una muestra de datos.

9.4 de aprobación Human-in-the-loop

El sistema ML cuenta con verificación y autorización manuales para garantizar el cumplimiento de las normas. El sistema requiere la aprobación de cada cambio de entorno (por ejemplo, desarrollo, control de calidad, preproducción y producción).

9.5 Pruebas de sesgo y ataques adversarios

El sistema de aprendizaje automático cuenta con pruebas contradictorias de Red Team que utilizan múltiples herramientas y vectores de ataque, y una verificación automática del sesgo en subpoblaciones específicas. Este componente está relacionado con la sección de observabilidad y gestión de modelos.

Ejemplo de uso de la MLOps lista de verificación

[No hay una única forma de utilizar la lista de verificación. MLOps](#) Sin embargo, un patrón común consiste en repetirla hasta que la calidad y el rendimiento del sistema estén en los niveles deseados.

Un ciclo de iteración puede tener el siguiente aspecto:

1. Revisa la lista de verificación elemento por elemento.
2. Revise los resultados de la lista de verificación existente.
3. Identifique cualquier laguna.
4. Priorice las brechas para subsanarlas.
5. Desarrolle las implementaciones de las nuevas funciones que desee.
6. Repita el ciclo de iteración.

Recursos

Referencias

- David Nigenda, Zohar Karnin, Muhammad Bilal Zafar, Raghu Ramesha, Alan Tan, Michele Donini y Krishnaram Kenthapadi. 2022. «Amazon SageMaker Model Monitor: un sistema para obtener información en tiempo real sobre los modelos de Machine Learning implementados»: [arXiv: 2111.13657](#).
- Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau y Tobias Meisen. 2019. «Estudios de ablación en redes neuronales artificiales»: [arXiv: 1901.08644](#).
- Brian Babcock, Mayur Datar y Rajeev Motwani. [«Toma de muestras desde una ventana móvil mediante transmisión de datos»](#): Universidad de Princeton.
- Chip Huyen. [Diseño de sistemas Machine Learning: un proceso iterativo para aplicaciones listas para la producción](#). Sebastopol, California: O'Reilly Media, 2022.
- Cathy Chen, Niall Murphy, Kranti Parisa, D. Sculley y Todd Underwood. [Reliable Machine Learning: aplicación de los principios de la SRE al aprendizaje automático en la producción](#). Sebastopol, California: O'Reilly Media, 2022.

Lecturas adicionales

- [AWS Well-Architected Framework: Machine Learning Lens](#)
- [Detecte la desviación de datos de la PNL con Amazon SageMaker Model Monitor personalizado](#) (entrada del blog)
- [Pruebas A/B dinámicas para modelos de aprendizaje automático con SageMaker MLOps proyectos de Amazon](#) (entrada del blog)

Herramientas

- [Amazon SageMaker](#)
- [Control de versiones de datos \(DVC\)](#)

Consultoría

- [Servicios profesionales de AWS](#)

Historial de documentos

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

Cambio	Descripción	Fecha
Publicación inicial	—	26 de julio 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.

antipatronos

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 CCoE](#) en el blog de estrategia Nube de AWS empresarial.

computación en la nube

La tecnología en la nube que se utiliza normalmente para la administración de dispositivos de IoT y el almacenamiento de datos de forma remota. La computación en la nube suele estar 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 CCoE, establecer un modelo de operaciones)

- Migración: migración de aplicaciones individuales
- Reinención: 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 o Bitbucket Cloud. Cada versión del código se denomina rama. En una estructura de microservicios, cada repositorio se encuentra dedicado a una única funcionalidad. Una sola 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/CD puede 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](#) 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.

I

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 (IIoT)

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

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 IIoT?](#).

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 dismantelar 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 versión original de inglés, prevalecerá la versión en inglés.