Planificando para tener éxito MLOps

AWS Guía prescriptiva

AWS Guía prescriptiva: Planificando para tener éxito MLOps

Table of Contents

Planificar para tener éxito MLOps

Bruno Klein, Amazon Web Services (AWS)

Diciembre de 2021 (historial del documento)

La implementación de soluciones de aprendizaje automático (ML) en la producción presenta muchos desafíos que no surgen en los proyectos de desarrollo de software estándar. Las soluciones de aprendizaje automático son más complejas y difíciles de conseguir correctamente desde el principio. También existen en entornos generalmente volátiles, donde la distribución de los datos se desvía significativamente con el tiempo por diversas razones esperadas e inesperadas.

Estos problemas se agravan aún más por el hecho de que muchos profesionales del aprendizaje automático no tienen formación en ingeniería de software, por lo que es posible que no estén familiarizados con las mejores prácticas de este sector, como la escritura de código comprobable, la modularización de los componentes y el uso eficaz del control de versiones. Estos desafíos generan una deuda técnica y, con el tiempo, las soluciones se vuelven más complejas y difíciles de mantener, lo que se traduce en un efecto agravante para los equipos de aprendizaje automático.

Esta guía enumera las mejores prácticas de operaciones de aprendizaje automático (MLOps) que ayudan a mitigar estos desafíos en los proyectos y las cargas de trabajo de aprendizaje automático.

Dado que MLOps se trata de una preocupación transversal, estos problemas afectan no solo a los procesos de implementación y supervisión, sino también a todo el ciclo de vida del modelo. En esta guía, las MLOps mejores prácticas se organizan en cuatro áreas principales:

- Datos
- Entrenamiento
- Implementación
- Supervisión

Resultados empresariales específicos

La implementación de modelos de aprendizaje automático en la producción es una tarea que requiere un esfuerzo continuo y un equipo dedicado a mantener estos recursos durante toda su vida útil (en algunos casos, incluso años). Los modelos de aprendizaje automático pueden aportar un valor considerable a los datos empresariales, pero tienen costes elevados. Para minimizar los

costos, las empresas deben seguir las buenas prácticas en el desarrollo de software y la ciencia de datos. Deben ser conscientes de los matices de los sistemas de aprendizaje automático, como la desviación de los datos, que hace que los modelos funcionen inesperadamente después de un tiempo. Al ser conscientes de estas preocupaciones, las empresas pueden cumplir sus objetivos empresariales de forma segura y ágil a corto y largo plazo.

Existen varios tipos de modelos de aprendizaje automático y los sectores a los que se dirigen tienen diferentes tipos de tareas y problemas empresariales relacionados con el aprendizaje automático, por lo que es necesario tener en cuenta un conjunto diferente de preocupaciones para cada modelo y sector. Las prácticas expuestas en esta guía no son específicas de un modelo o negocio, sino que se aplican a un amplio conjunto de modelos e industrias para mejorar los tiempos de implementación, generar una mayor productividad y fomentar una gobernanza y una seguridad más sólidas.

Poner los modelos en producción es una tarea multidisciplinaria que requiere científicos de datos, ingenieros de aprendizaje automático, ingenieros de datos e ingenieros de software. Cuando cree su equipo de aprendizaje automático, le recomendamos que se centre en estas habilidades y antecedentes.

Datos

DevOps es una práctica de ingeniería de software que se ocupa de la operacionalización del software. Los elementos más comunes DevOps son el código con control de versiones, los procesos de integración y entrega continuas (CI/CD), las pruebas unitarias y la creación y el despliegue de código reproducible, todos ellos con código. Los modelos de aprendizaje automático son un producto de código y datos, por lo que los datos deben cumplir los mismos estándares que el código. MLOps deben abordar cuestiones relacionadas con los datos, como la forma de mantener la calidad de los datos, cómo identificar los casos extremos en los datos, cómo protegerlos y cómo hacer que los datos sean más fáciles de mantener.

Temas

Etiquetado

Proporcione instrucciones de etiquetado claras

Un conjunto de datos puede incluir muestras ambiguas que dan como resultado un etiquetado incoherente en todo el conjunto de datos. Por ejemplo, considere la tarea de etiquetar las imágenes que contienen un perro. Es posible que algunas muestras contengan solo una imagen del animal. ¿Deberían marcarse con una etiqueta positiva o negativa? Este tipo de problema podría resolverse proporcionando instrucciones claras y objetivas a los etiquetadores.

Utilice la votación por mayoría

Consideremos ahora la cuestión de etiquetar un speech-to-text conjunto de datos que contiene audio ruidoso con palabras fonéticamente similares o idénticas a otras, como saber y salir, zapato y dos, llorar y drogarse o derecha y escribir. En este caso, los etiquetadores podrían etiquetar estas muestras de forma incoherente.

Para mantener un alto grado de exactitud en el etiquetado, un enfoque común es utilizar la votación por mayoría, en la que se entrega la misma muestra de datos a varios trabajadores y sus resultados se agregan. Este método y sus variantes más sofisticadas se describen en la entrada del blog

Use the wisdom of crowd with Amazon SageMaker AI Ground Truth para anotar datos con mayor precisión en el blog AWS Machine Learning.

Divisiones y fugas de datos

La filtración de datos se produce cuando el modelo obtiene datos durante la inferencia (el momento en que el modelo está en producción y recibe solicitudes de predicción) a los que no debería tener acceso, como muestras de datos que se utilizaron para el entrenamiento o información que no estará disponible cuando el modelo se implemente en producción.

Si el modelo se prueba inadvertidamente con datos de entrenamiento, la filtración de datos podría provocar un sobreajuste. El sobreajuste significa que el modelo no se generaliza bien con datos invisibles. En esta sección, se proporcionan las mejores prácticas para evitar la fuga de datos y el sobreajuste.

Divida los datos en al menos tres conjuntos

Una fuente común de filtración de datos es dividir (dividir) los datos de forma incorrecta durante el entrenamiento. Por ejemplo, es posible que el científico de datos haya entrenado el modelo, consciente o inconscientemente, sobre la base de los datos que se utilizaron para las pruebas. En tales situaciones, es posible que observe métricas de éxito muy altas causadas por un sobreajuste. Para solucionar este problema, debes dividir los datos en al menos tres conjuntos: `trainingvalidation`, `ytesting`.

Al dividir los datos de esta manera, puede usar el `validation` conjunto para elegir y ajustar los parámetros que usa para controlar el proceso de aprendizaje (hiperparámetros). Cuando haya conseguido el resultado deseado o haya alcanzado un punto de mejora estable, evalúe el `testing` conjunto. Las métricas de rendimiento del `testing` conjunto deben ser similares a las métricas de los demás conjuntos. Esto indica que no hay ningún desajuste de distribución entre los conjuntos y se espera que su modelo se generalice bien en la producción.

Utilice un algoritmo de división estratificado

Al dividir los datos en y `testing` para conjuntos de datos pequeños `trainingvalidation`, o cuando trabaje con datos muy desequilibrados, asegúrese de utilizar un algoritmo de división estratificado. La estratificación garantiza que cada división contenga aproximadamente el mismo número o distribución de clases para cada división. La biblioteca Scikit-learn ML ya implementa la estratificación, al igual que Apache Spark.

En cuanto al tamaño de la muestra, asegúrate de que los conjuntos de validación y prueba tengan datos suficientes para la evaluación, de modo que puedas llegar a conclusiones estadísticamente significativas. Por ejemplo, un tamaño de división común para conjuntos de datos relativamente pequeños (menos de 1 millón de muestras) es del 70%, el 15% y el 15%, para `trainingvalidation`, `ytesting`. Para conjuntos de datos muy grandes (más de 1 millón de muestras), puede utilizar el 90%, el 5% y el 5% para maximizar los datos de entrenamiento disponibles.

En algunos casos de uso, resulta útil dividir los datos en conjuntos adicionales, ya que es posible que los datos de producción hayan experimentado cambios de distribución repentinos y radicales durante el período en el que se recopilaron. Por ejemplo, considere un proceso de recopilación de datos para crear un modelo de previsión de la demanda de artículos de las tiendas de abarrotes. Si el equipo de ciencia de datos recopiló los `training` datos durante 2019 y los `testing` datos desde enero de 2020 hasta marzo de 2020, un modelo probablemente obtendría una buena puntuación en el `testing` conjunto. Sin embargo, cuando el modelo entre en producción, el patrón de consumo de determinados artículos ya habría cambiado considerablemente debido a la pandemia de la COVID-19, y el modelo generaría resultados deficientes. En este escenario, tendría sentido añadir otro conjunto (por ejemplo,`recent_testing`) como salvaguardia adicional para la aprobación del modelo. Esta adición podría impedirle aprobar un modelo de producción que al instante tendría un mal rendimiento debido a un desajuste en la distribución.

En algunos casos, es posible que desee crear `testing` conjuntos adicionales `validation` o que incluyan tipos específicos de muestras, como datos asociados a poblaciones minoritarias. Es importante que estas muestras de datos sean correctas, pero es posible que no estén bien representadas en el conjunto de datos general. Estos subconjuntos de datos se denominan segmentos.

Consideremos el caso de un modelo de aprendizaje automático para el análisis crediticio que se basó en los datos de todo un país y se equilibró para tener en cuenta por igual todo el dominio de la variable objetivo. Además, tenga en cuenta que este modelo podría tener una `City` característica. Si el banco que usa este modelo expande sus negocios a una ciudad específica, podría interesarle saber cómo funciona el modelo en esa región. Por lo tanto, un proceso de aprobación no solo debe evaluar la calidad del modelo en función de los datos de las pruebas de todo el país, sino que también debe evaluar los datos de las pruebas de una zona urbana determinada.

Cuando los científicos de datos trabajan en un modelo nuevo, pueden evaluar fácilmente las capacidades del modelo y tener en cuenta los casos extremos mediante la integración de los segmentos infrarrepresentados en la fase de validación del modelo.

Considere la posibilidad de duplicar las muestras al hacer divisiones aleatorias

Otra fuente de fugas, menos común, se encuentra en los conjuntos de datos que pueden contener demasiadas muestras duplicadas. En este caso, incluso si divide los datos en subconjuntos, es posible que los distintos subconjuntos tengan muestras en común. Según el número de duplicados, el sobreajuste puede confundirse con una generalización.

Tenga en cuenta las funciones que podrían no estar disponibles al recibir inferencias en producción

La fuga de datos también se produce cuando los modelos se entrenan con funciones que no están disponibles en producción, en el momento en que se invocan las inferencias. Como los modelos suelen crearse a partir de datos históricos, estos datos pueden enriquecerse con columnas o valores adicionales que no estaban presentes en algún momento. Consideremos el caso de un modelo de aprobación de crédito que tiene una función que registra cuántos préstamos ha concedido un cliente al banco en los últimos seis meses. Existe el riesgo de que se produzcan filtraciones de datos si se implementa este modelo y se utiliza para la aprobación crediticia de un nuevo cliente que no tiene un historial de seis meses en el banco.

Amazon SageMaker AI Feature Store ayuda a resolver este problema. Puede probar sus modelos con mayor precisión mediante el uso de consultas de viaje en el tiempo, que puede utilizar para ver datos en momentos específicos.

Tienda de funciones

El uso de SageMaker AI Feature Store aumenta la productividad del equipo, ya que disocia los límites de los componentes (por ejemplo, el almacenamiento y el uso). También permite la reutilización de funciones en los diferentes equipos de ciencia de datos de la organización.

Utilice consultas sobre viajes en el tiempo

Las funciones de viaje en el tiempo de Feature Store ayudan a reproducir la creación de modelos y a respaldar prácticas de gobierno más sólidas. Esto puede resultar útil cuando una organización quiere evaluar el linaje de datos, de forma similar a como las herramientas de control de versiones, como Git, evalúan el código. Las consultas sobre viajes en el tiempo también ayudan a las organizaciones a proporcionar datos precisos para las comprobaciones de conformidad. Para obtener más

información, consulte Comprender las capacidades clave de Amazon SageMaker AI Feature Store en el blog AWS Machine Learning.

Uso de roles de IAM

Feature Store también ayuda a mejorar la seguridad sin afectar a la productividad ni a la innovación del equipo. Puedes usar funciones AWS Identity and Access Management (IAM) para conceder o restringir el acceso granular a funciones específicas para usuarios o grupos específicos.

Por ejemplo, la siguiente política restringe el acceso a una función confidencial de Feature Store.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "VisualEditor0",
            "Effect": "Deny",
            "Action": "*",
            "Resource": "arn:aws:s3:::amzn-s3-demo-bucket--usw2-az1--x-s3/12345678910/
sagemaker/us-east-2/offline-store/doctor-appointments"
        }
    ]
}
```

Para obtener más información sobre la seguridad y el cifrado de los datos mediante Feature Store, consulte Seguridad y control de acceso en la documentación sobre SageMaker IA.

Utilice las pruebas unitarias

Cuando los científicos de datos crean modelos basados en algunos datos, suelen hacer suposiciones sobre la distribución de los datos o realizan un análisis exhaustivo para comprender completamente las propiedades de los datos. Cuando se implementan estos modelos, con el tiempo se vuelven obsoletos. Cuando el conjunto de datos queda desactualizado, los científicos de datos, los ingenieros de aprendizaje automático y (en algunos casos) los sistemas automatizados rediseñan el modelo con nuevos datos que se obtienen de una tienda en línea o fuera de línea.

Sin embargo, es posible que la distribución de estos nuevos datos haya cambiado, lo que podría afectar al rendimiento del algoritmo actual. Una forma automática de comprobar este tipo de problemas consiste en utilizar el concepto de pruebas unitarias de la ingeniería de software. Entre los factores que más se suelen comprobar se incluyen el porcentaje de valores faltantes, la cardinalidad

de las variables categóricas y si las columnas con valores reales se ajustan a alguna distribución esperada mediante un marco como la estadística de las pruebas de hipótesis (prueba t). Es posible que también desee validar el esquema de datos para asegurarse de que no ha cambiado y que no generará entidades de entrada no válidas de forma silenciosa.

Las pruebas unitarias requieren comprender los datos y su dominio para poder planificar las afirmaciones exactas que se van a realizar como parte del proyecto de aprendizaje automático. Para obtener más información, consulte Probar la calidad de los datos a escala PyDeequ en el blog sobre AWS macrodatos.

Formación

MLOps se ocupa de la operacionalización del ciclo de vida del aprendizaje automático. Por lo tanto, debe facilitar la labor de los científicos e ingenieros de datos para crear modelos pragmáticos que satisfagan las necesidades empresariales y funcionen bien a largo plazo, sin incurrir en deudas técnicas.

Siga las mejores prácticas de esta sección para ayudar a abordar los desafíos de la formación de modelos.

Temas

Cree un modelo de referencia

Cuando los profesionales se enfrentan a un problema empresarial con una solución de aprendizaje automático, lo primero que suelen hacer es utilizar el state-of-the-art algoritmo. Esta práctica es arriesgada, porque es probable que el state-of-the-art algoritmo no haya sido probado en el tiempo. Además, el state-of-the-art algoritmo suele ser más complejo y no se entiende bien, por lo que podría suponer solo mejoras marginales en comparación con modelos alternativos más simples. Una mejor práctica es crear un modelo de referencia que sea relativamente rápido de validar e implementar y que pueda ganarse la confianza de las partes interesadas del proyecto.

Al crear una línea base, le recomendamos que evalúe su rendimiento métrico siempre que sea posible. Compare el rendimiento del modelo de referencia con otros sistemas automatizados o manuales para garantizar su éxito y asegurarse de que la implementación del modelo o el proyecto se puedan llevar a cabo a medio y largo plazo.

El modelo de referencia debe validarse más a fondo con los ingenieros de aprendizaje automático para confirmar que el modelo puede cumplir los requisitos no funcionales que se han establecido para el proyecto, como el tiempo de inferencia, la frecuencia con la que se espera que los datos

cambien de distribución, si el modelo se puede volver a entrenar fácilmente en estos casos y cómo se implementará, lo que afectará al costo de la solución. Obtenga puntos de vista multidisciplinarios sobre estas cuestiones para aumentar las posibilidades de desarrollar un modelo exitoso y duradero.

Los científicos de datos podrían inclinarse por añadir tantas funciones como sea posible a un modelo de referencia. Si bien esto aumenta la capacidad de un modelo para predecir el resultado deseado, es posible que algunas de estas características solo generen mejoras métricas incrementales. Muchas funciones, especialmente las que están altamente correlacionadas, pueden ser redundantes. Añadir demasiadas funciones aumenta los costes, ya que requiere más recursos informáticos y ajustes. El exceso de funciones también afecta a day-to-day las operaciones del modelo, ya que es más probable que los datos se desvíen o se produzcan con mayor rapidez.

Considere un modelo en el que dos entidades de entrada estén altamente correlacionadas, pero solo una entidad tenga causalidad. Por ejemplo, un modelo que predice si un préstamo va a dejar de pagar puede tener características como la edad del cliente y los ingresos, que pueden estar muy correlacionados, pero solo los ingresos deberían utilizarse para conceder o denegar un préstamo. Un modelo que se haya entrenado en estas dos características podría basarse en una característica que no tiene causalidad, como la edad, para generar el resultado de la predicción. Si, tras su puesta en producción, el modelo recibe solicitudes de inferencia de clientes mayores o menores que la edad media incluida en el conjunto de formación, podría empezar a tener un rendimiento deficiente.

Además, cada característica individual podría sufrir un cambio de distribución durante la producción y provocar que el modelo se comporte de forma inesperada. Por estas razones, cuantas más características tenga un modelo, más frágil será con respecto a la deriva y al estancamiento.

Los científicos de datos deben utilizar medidas de correlación y valores de Shapley para evaluar qué características añaden suficiente valor a la predicción y deben mantenerse. Tener modelos tan complejos aumenta la posibilidad de que se produzca un circuito de retroalimentación, en el que el modelo cambie el entorno para el que se modeló. Un ejemplo es un sistema de recomendaciones en el que el comportamiento del consumidor puede cambiar debido a las recomendaciones de un modelo. Los circuitos de retroalimentación que actúan en todos los modelos son menos comunes. Por ejemplo, considere un sistema de recomendaciones que recomiende películas y otro sistema que recomiende libros. Si ambos modelos se dirigen al mismo grupo de consumidores, se afectarían mutuamente.

Para cada modelo que desarrolle, considere qué factores podrían contribuir a estas dinámicas, de modo que sepa qué métricas debe monitorear en la producción.

Utilice un enfoque centrado en los datos y un análisis de errores

Si utilizas un modelo simple, tu equipo de aprendizaje automático puede centrarse en mejorar los datos en sí mismos y adoptar un enfoque centrado en los datos en lugar de uno centrado en los modelos. Si su proyecto utiliza datos no estructurados, como imágenes, texto, audio y otros formatos que puedan ser evaluados por personas (en comparación con los datos estructurados, que pueden ser más difíciles de asignar a una etiqueta de manera eficiente), una buena práctica para mejorar el rendimiento del modelo es realizar un análisis de errores.

El análisis de errores implica evaluar un modelo en un conjunto de validación y comprobar los errores más comunes. Esto ayuda a identificar posibles grupos de muestras de datos similares que el modelo podría tener dificultades para corregir. Para realizar un análisis de errores, puede enumerar las inferencias con errores de predicción más altos o clasificar los errores en los que se predijo que una muestra de una clase pertenecía a otra clase, por ejemplo.

Diseñe su modelo para una iteración rápida

Cuando los científicos de datos siguen las mejores prácticas, pueden experimentar con un nuevo algoritmo o combinar diferentes características de forma fácil y rápida durante la prueba de concepto o incluso el readiestramiento. Esta experimentación contribuye al éxito de la producción. Una buena práctica consiste en basarse en el modelo de referencia, emplear algoritmos un poco más complejos y añadir nuevas funciones de forma iterativa y, al mismo tiempo, supervisar el rendimiento del conjunto de entrenamiento y validación para comparar el comportamiento real con el comportamiento esperado. Este marco de formación puede proporcionar un equilibrio óptimo en cuanto al poder de predicción y ayudar a que los modelos sean lo más simples posible con una menor huella de deuda técnica.

Para una iteración rápida, los científicos de datos deben intercambiar diferentes implementaciones de modelos para determinar cuál es el mejor modelo para usar con datos específicos. Si tiene un equipo grande, un plazo corto y otros aspectos logísticos relacionados con la gestión de proyectos, la iteración rápida puede resultar difícil sin un método establecido.

En ingeniería de software, el principio de sustitución de Liskov es un mecanismo para diseñar las interacciones entre los componentes del software. Este principio establece que debe poder reemplazar una implementación de una interfaz por otra implementación sin interrumpir la aplicación cliente o la implementación. Cuando escribas código de entrenamiento para tu sistema de aprendizaje automático, puedes emplear este principio para establecer límites y encapsular el

código, de forma que puedas reemplazar el algoritmo con facilidad y probar nuevos algoritmos de forma más eficaz.

Por ejemplo, en el siguiente código, puedes añadir nuevos experimentos simplemente añadiendo una nueva implementación de clase.

```
from abc import ABC, abstractmethod

from pandas import DataFrame


class ExperimentRunner(object):

    def __init__(self, *experiments):
        self.experiments = experiments

    def run(self, df: DataFrame) -> None:
        for experiment in self.experiments:
            result = experiment.run(df)
            print(f'Experiment "{experiment.name}" gave result {result}')


class Experiment(ABC):

    @abstractmethod
    def run(self, df: DataFrame) -> float:
        pass

    @property
    @abstractmethod
    def name(self) -> str:
        pass


class Experiment1(Experiment):

    def run(self, df: DataFrame) -> float:
        print('performing experiment 1')
        return 0

    def name(self) -> str:
        return 'experiment 1'
```

```
class Experiment2(Experiment):

    def run(self, df: DataFrame) -> float:
        print('performing experiment 2')
        return 0

    def name(self) -> str:
        return 'experiment 2'


class Experiment3(Experiment):

    def run(self, df: DataFrame) -> float:
        print('performing experiment 3')
        return 0

    def name(self) -> str:
        return 'experiment 3'


if __name__ == '__main__':
    runner = ExperimentRunner(*[
        Experiment1(),
        Experiment2(),
        Experiment3()
    ])
    df = ...
    runner.run(df)
```

Realiza un seguimiento de tus experimentos de aprendizaje automático

Cuando trabajas con un gran número de experimentos, es importante evaluar si las mejoras observadas son producto de los cambios implementados o de la casualidad. Puede utilizar Amazon SageMaker AI Experiments para crear fácilmente experimentos y asociarles metadatos para su seguimiento, comparación y evaluación.

Reducir la aleatoriedad del proceso de creación del modelo resulta útil para depurar, solucionar problemas y mejorar la gobernanza, ya que permite predecir la inferencia del modelo de salida con mayor certeza si se utilizan el mismo código y los mismos datos.

A menudo, no es posible hacer que un código de entrenamiento sea totalmente reproducible, debido a la inicialización aleatoria del peso, la sincronicidad del cómputo paralelo, las complejidades internas de la GPU y factores no deterministas similares. Sin embargo, establecer correctamente los valores iniciales aleatorios para garantizar que cada sesión de entrenamiento comience desde el mismo punto y se comporte de manera similar, mejora considerablemente la previsibilidad de los resultados.

Solucione problemas relacionados con los trabajos de formación

En algunos casos, puede resultar difícil para los científicos de datos ajustar incluso un modelo de referencia muy simple. En este caso, podrían decidir que necesitan un algoritmo que pueda adaptarse mejor a funciones complejas. Una buena prueba consiste en utilizar la línea base de una parte muy pequeña del conjunto de datos (por ejemplo, unas 10 muestras) para asegurarse de que el algoritmo se ajusta demasiado a esta muestra. Esto ayuda a descartar problemas con los datos o el código.

Otra herramienta útil para depurar escenarios complejos es Amazon SageMaker AI Debugger, que puede detectar problemas relacionados con la corrección algorítmica y la infraestructura, como el uso óptimo de la computación.

Implementación

En la ingeniería de software, la puesta en producción del código requiere la debida diligencia, ya que el código puede comportarse de forma inesperada, un comportamiento imprevisto del usuario puede interrumpir el software y se pueden encontrar casos extremos inesperados. Los ingenieros y los DevOps ingenieros de software suelen emplear pruebas unitarias y estrategias de reversión para mitigar estos riesgos. Con el aprendizaje automático, la puesta en producción de los modelos requiere aún más planificación, ya que se espera que el entorno real cambie y, en muchas ocasiones, los modelos se validan con métricas que representan las métricas empresariales reales que se están intentando mejorar.

Siga las prácticas recomendadas de esta sección para ayudar a abordar estos desafíos.

Temas

Automatice el ciclo de implementación

El proceso de formación e implementación debe estar completamente automatizado para evitar errores humanos y garantizar que las comprobaciones de construcción se realicen de forma coherente. Los usuarios no deben tener permisos de acceso de escritura al entorno de producción.

Amazon SageMaker AI Pipelines y la AWS CodePipelineayuda a crear una CI/CD pipelines for ML projects. One of the advantages of using a CI/CD canalización permiten controlar las versiones de todo el código que se utiliza para ingerir datos, entrenar un modelo y realizar la supervisión mediante una herramienta como Git. A veces hay que volver a entrenar un modelo utilizando el mismo algoritmo e hiperparámetros, pero datos diferentes. La única forma de comprobar que estás utilizando la versión correcta del algoritmo es utilizar el control de código fuente y las etiquetas. Puedes usar las plantillas de proyecto predeterminadas que te proporciona la SageMaker IA como punto de partida para tu MLOps consulta.

Cuando crees canalizaciones de CI/CD para implementar tu modelo, asegúrate de etiquetar tus artefactos de construcción con un identificador de compilación, una versión o confirmación del código y una versión de datos. Esta práctica te ayuda a solucionar cualquier problema de implementación.

A veces, también es necesario etiquetar los modelos que realizan predicciones en campos muy regulados. La capacidad de trabajar de forma retrospectiva e identificar los datos, el código, la compilación, las comprobaciones y las aprobaciones exactos asociados a un modelo de aprendizaje automático puede ayudar a mejorar considerablemente la gobernanza.

Parte del trabajo de la cartera de CI/CD consiste en realizar pruebas con lo que está creando. Aunque se espera que las pruebas de unidades de datos se realicen antes de que un feature store ingiera los datos, la canalización sigue siendo responsable de realizar pruebas en la entrada y salida de un modelo determinado y de comprobar las métricas clave. Un ejemplo de este tipo de comprobación consiste en validar un modelo nuevo en un conjunto de validación fijo y confirmar que su rendimiento es similar al del modelo anterior mediante el uso de un umbral establecido. Si el rendimiento es significativamente inferior al esperado, la construcción debería fallar y el modelo no debería entrar en producción.

El uso generalizado de canalizaciones de CI/CD también admite solicitudes de cambios, lo que ayuda a evitar errores humanos. Cuando utilizas solicitudes de extracción, cada cambio de código debe ser revisado y aprobado por al menos otro miembro del equipo antes de que pueda pasar a producción. Las solicitudes de extracción también son útiles para identificar el código que no se ajusta a las normas empresariales y para difundir conocimientos entre el equipo.

Elige una estrategia de despliegue

MLOps las estrategias de despliegue incluyen blue/green, canary, shadow, and A/B las pruebas.

Azul/verde

Blue/green deployments are very common in software development. In this mode, two systems are kept running during development: blue is the old environment (in this case, the model that is being replaced) and green is the newly released model that is going to production. Changes can easily be rolled back with minimum downtime, because the old system is kept alive. For more in-depth information about blue/greendespliegues en el contexto de SageMaker, consulte la entrada del blog Implementación y supervisión seguras de los puntos de conexión de Amazon SageMaker AI con AWS CodePipeline y AWS CodeDeploy en el blog AWS Machine Learning.

Valor controlado

Las implementaciones de Canary son similares a blue/green deployments in that both keep two models running together. However, in canary deployments, the new model is rolled out to

users incrementally, until all traffic eventually shifts over to the new model. As in blue/green las implementaciones, pero el riesgo se mitiga porque el nuevo modelo (y potencialmente defectuoso) se supervisa de cerca durante la implementación inicial y se puede revertir en caso de problemas. En la SageMaker IA, puedes especificar la distribución inicial del tráfico mediante la API. InitialVariantWeight

Sombra

Puede utilizar despliegues paralelos para llevar un modelo a producción de forma segura. En este modo, el nuevo modelo funciona junto con un modelo o proceso empresarial anterior y realiza inferencias sin influir en ninguna decisión. Este modo puede resultar útil como una comprobación final o un experimento de mayor fidelidad antes de pasar el modelo a producción.

El modo Sombra es útil cuando no necesitas ningún comentario sobre las inferencias de los usuarios. Puede evaluar la calidad de las predicciones realizando un análisis de errores y comparando el modelo nuevo con el modelo anterior, y puede supervisar la distribución de la salida para comprobar que es la esperada. Para obtener información sobre cómo realizar un despliegue paralelo con SageMaker IA, consulte la entrada del blog Implemente modelos de aprendizaje automático en Amazon SageMaker AI en el blog AWS Machine Learning.

Prueba A/B

Cuando los profesionales del aprendizaje automático desarrollan modelos en sus entornos, las métricas para las que optimizan suelen reflejar las métricas empresariales que realmente importan. Esto hace que sea difícil determinar con certeza si un nuevo modelo realmente mejorará los resultados empresariales, como los ingresos y la tasa de clics, y reducirá el número de quejas de los usuarios.

Pensemos en un sitio web de comercio electrónico en el que el objetivo empresarial es vender tantos productos como sea posible. El equipo de revisión sabe que las ventas y la satisfacción del cliente se correlacionan directamente con reseñas informativas y precisas. Un miembro del equipo podría proponer un nuevo algoritmo de clasificación de reseñas para mejorar las ventas. Mediante las pruebas A/B, podrían extender los algoritmos antiguos y nuevos a grupos de usuarios diferentes pero similares, y supervisar los resultados para ver si los usuarios que recibieron predicciones del modelo más nuevo tienen más probabilidades de realizar compras.

Las pruebas A/B también ayudan a evaluar el impacto empresarial de la obsolescencia y la desviación de los modelos. Los equipos pueden poner en producción nuevos modelos con cierta

periodicidad, realizar pruebas A/B con cada modelo y crear un gráfico de edad en comparación con el rendimiento. Esto ayudaría al equipo a entender la deriva de los datos: la volatilidad de los datos de producción.

Para obtener más información sobre cómo realizar pruebas A/B con SageMaker IA, consulte la entrada del blog A/B Testing de modelos de aprendizaje automático en producción con Amazon SageMaker AI en el blog AWS Machine Learning.

Tenga en cuenta sus requisitos de inferencia

Con la SageMaker IA, puede elegir la infraestructura subyacente para implementar su modelo de diferentes maneras. Estas capacidades de invocación de inferencias admiten diferentes casos de uso y perfiles de costes. Sus opciones incluyen la inferencia en tiempo real, la inferencia asíncrona y la transformación por lotes, como se explica en las siguientes secciones.

Inferencia en tiempo real

La inferencia en tiempo real es ideal para las cargas de trabajo de inferencia en las que se requieren requisitos de baja latencia, interactivos y en tiempo real. Puede implementar su modelo en los servicios de alojamiento de SageMaker IA y obtener un punto final que pueda usarse para realizar inferencias. Estos puntos de conexión están totalmente gestionados, admiten el escalado automático (consulte Escalar automáticamente los modelos de Amazon SageMaker AI) y se pueden implementar en varias zonas de disponibilidad.

Si tiene un modelo de aprendizaje profundo creado con Apache MXNet PyTorch TensorFlow, o también puede utilizar Amazon SageMaker AI Elastic Inference (EI). Con la IE, puede adjuntar fracciones GPUs a cualquier instancia de SageMaker IA para acelerar la inferencia. Puede seleccionar la instancia de cliente para ejecutar su aplicación y adjuntar un acelerador de IE para utilizar la cantidad correcta de aceleración de GPU para sus necesidades de inferencia.

Otra opción es utilizar terminales multimodelo, que proporcionan una solución escalable y rentable para implementar un gran número de modelos. Estos puntos finales utilizan un contenedor de servicio compartido que permite alojar varios modelos. Los terminales multimodelo reducen los costes de alojamiento al mejorar la utilización de los terminales en comparación con el uso de terminales de un solo modelo. También reducen la sobrecarga de implementación, ya que la SageMaker IA gestiona la carga de los modelos en la memoria y su escalado en función de los patrones de tráfico.

Para obtener más información sobre las mejores prácticas para implementar modelos de aprendizaje automático en la SageMaker IA, consulte las mejores prácticas de implementación en la documentación sobre SageMaker IA.

Inferencia asíncrona

La inferencia asíncrona de Amazon SageMaker AI es una capacidad de la SageMaker IA que pone en cola las solicitudes entrantes y las procesa de forma asíncrona. Esta opción es ideal para solicitudes con grandes cargas útiles de hasta 1 GB, tiempos de procesamiento prolongados y requisitos de latencia prácticamente en tiempo real. La inferencia asíncrona le permite ahorrar costes al escalar automáticamente el recuento de instancias a cero cuando no hay solicitudes que procesar, de modo que solo paga cuando su terminal procesa las solicitudes.

Transformación por lotes

Usa la transformación por lotes cuando desees hacer lo siguiente:

- Procesar previamente los conjuntos de datos para acabar con el ruido o la compensación que interfiere con el entrenamiento o la inferencia del conjunto de datos.
- Obtener inferencias de los conjuntos de datos grandes.
- Ejecutar la inferencia cuando no sea necesario un punto de enlace persistente.
- Asociar registros de entrada con inferencias para ayudar a interpretar los resultados.

Monitorización

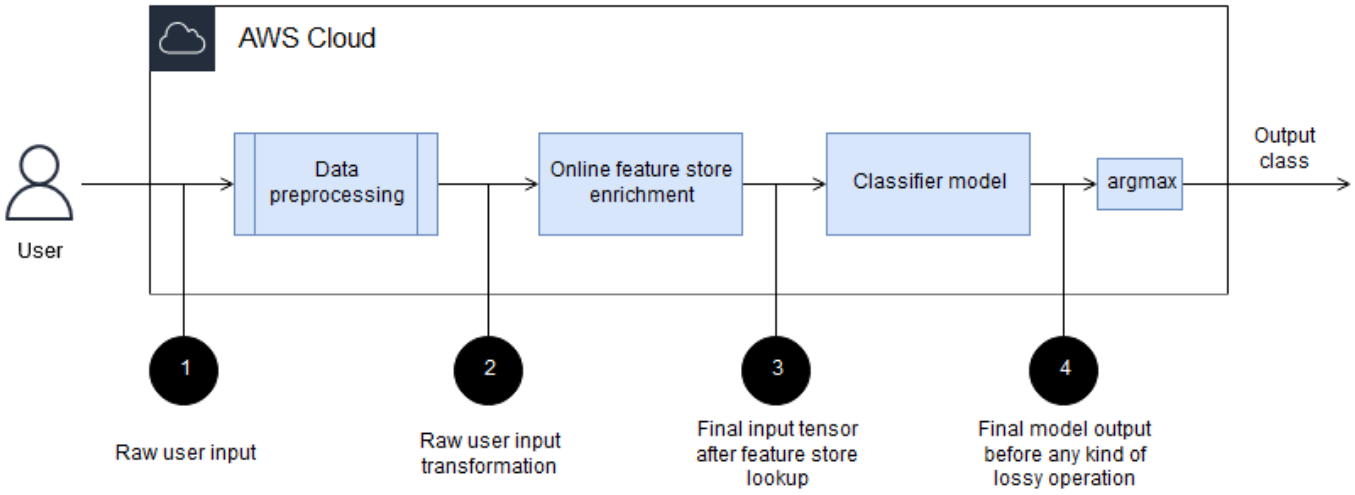
Cuando los modelos ya estén en producción y ofrezcan valor empresarial, realice comprobaciones continuas para identificar cuándo es necesario volver a capacitar los modelos o tomar medidas al respecto.

Su equipo de supervisión debe actuar de forma proactiva, no reactiva, para comprender mejor el comportamiento de los datos en el entorno e identificar la frecuencia, el ritmo y la brusquedad de las desviaciones de datos. El equipo debe identificar nuevos casos extremos en los datos que puedan estar infrarrepresentados en el conjunto de entrenamiento, el conjunto de validación y otros tipos de casos extremos. Deben almacenar las métricas de calidad de servicio (QoS), usar alarmas para tomar medidas de inmediato cuando surja un problema y definir una estrategia para incorporar y modificar los conjuntos de datos actuales. Estas prácticas comienzan por registrar las solicitudes y respuestas del modelo, a fin de proporcionar una referencia para la resolución de problemas o información adicional.

Lo ideal es que las transformaciones de datos se registren en algunas etapas clave durante el procesamiento:

- Antes de cualquier tipo de preprocesamiento
- Tras cualquier tipo de enriquecimiento de feature store
- Después de todas las etapas principales de un modelo
- Antes de cualquier tipo de función con pérdidas en la salida del modelo, como `argmax`

El siguiente diagrama ilustra estas etapas.



Puede utilizar SageMaker AI Model Monitor para capturar automáticamente los datos de entrada y salida y almacenarlos en Amazon Simple Storage Service (Amazon S3). Puede implementar otros tipos de registro intermedio añadiendo registros a un contenedor de servicio personalizado.

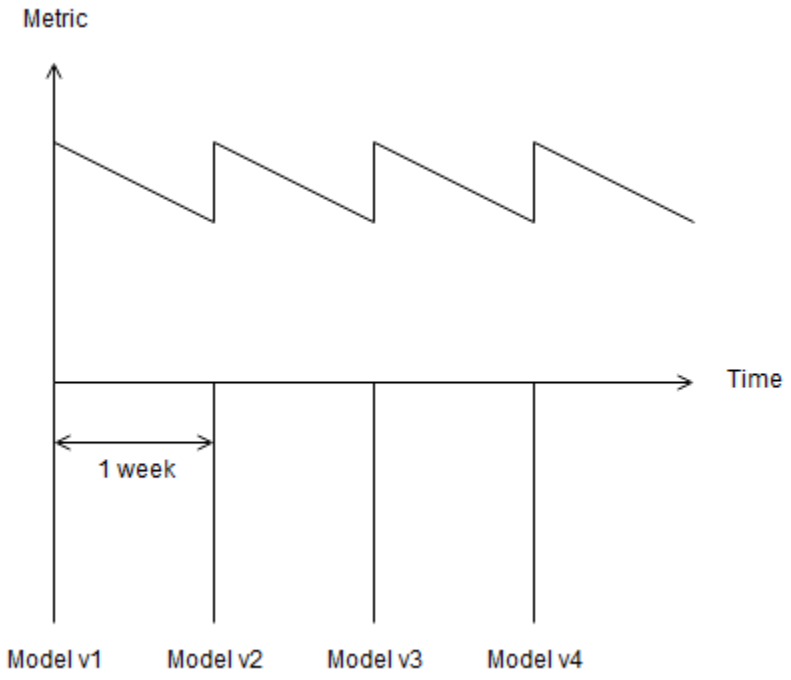
Después de registrar los datos de los modelos, puede supervisar la desviación de la distribución. En algunos casos, se puede obtener la verdad básica (datos correctamente etiquetados) poco después de la inferencia. Un ejemplo común de esto es un modelo que predice los anuncios más relevantes que se mostrarán a un usuario. En cuanto el usuario abandona la página, puedes determinar si ha hecho clic en el anuncio. Si el usuario ha hecho clic en el anuncio, puedes registrar esa información. En este sencillo ejemplo, puede cuantificar fácilmente el éxito de su modelo mediante una métrica, como la precisión o la F1, que se puede medir tanto en el entrenamiento como en la implementación. Para obtener más información sobre estos escenarios en los que ha etiquetado datos, consulte Supervisar la calidad del modelo en la documentación sobre SageMaker IA. Sin embargo, estos escenarios simples son poco frecuentes, ya que los modelos suelen diseñarse para optimizar métricas útiles desde el punto de vista matemático que solo representan los resultados empresariales reales. En estos casos, la mejor práctica consiste en supervisar los resultados empresariales cuando se implementa un modelo en producción.

Consideremos el caso de un modelo de clasificación de reseñas. Si el objetivo empresarial definido del modelo de aprendizaje automático es mostrar las opiniones más relevantes y útiles en la parte superior de la página web, puedes medir el éxito del modelo añadiendo un botón como «¿Te ha resultado útil?» para cada opinión. Medir la tasa de clics de este botón podría ser una medida de los resultados empresariales que le ayude a medir el rendimiento de su modelo en producción.

Para monitorizar la desviación de las etiquetas de entrada o salida en la SageMaker IA, puede utilizar las funciones de calidad de los datos de SageMaker AI Model Monitor, que monitoriza tanto la entrada como la salida. También puedes implementar tu propia lógica para SageMaker AI Model Monitor creando un contenedor personalizado.

Supervisar los datos que recibe un modelo tanto en tiempo de desarrollo como en tiempo de ejecución es fundamental. Los ingenieros deben monitorear los datos no solo para detectar cambios en el esquema, sino también para detectar desajustes en la distribución. Detectar los cambios en el esquema es más fácil y se puede implementar mediante un conjunto de reglas, pero la falta de coincidencia en la distribución suele ser más complicada, especialmente porque requiere definir un umbral para cuantificar cuándo se debe emitir una alarma. En los casos en los que se conoce la distribución monitorizada, lo más fácil suele ser monitorizar los parámetros de la distribución. En el caso de una distribución normal, esa sería la media y la desviación estándar. También son útiles otras métricas clave, como el porcentaje de valores faltantes, valores máximos y valores mínimos.

También puede crear trabajos de monitoreo continuo que muestreen los datos de entrenamiento y los datos de inferencia y comparen sus distribuciones. Puede crear estos trabajos tanto para la entrada como para la salida del modelo, y trazar los datos en función del tiempo para visualizar cualquier desviación repentina o gradual. Esto se ilustra en el siguiente gráfico.



Para comprender mejor el perfil de desviación de los datos, por ejemplo, la frecuencia con la que la distribución de los datos cambia significativamente, a qué ritmo o con qué rapidez, le recomendamos que implemente continuamente nuevas versiones del modelo y supervise su rendimiento. Por ejemplo, si su equipo implementa un modelo nuevo cada semana y observa que el rendimiento del modelo mejora significativamente cada vez, pueden decidir si deberían entregar nuevos modelos en menos de una semana como mínimo.

Próximos pasos y recursos

Esta guía explica algunas consideraciones a la hora de planificar el ciclo de vida de los modelos de aprendizaje automático que desea llevar a producción. Analiza los desafíos y las mejores prácticas en cuatro áreas (datos, formación, implementación y monitoreo) e incluye recursos adicionales relevantes.

AWS proporciona Well-Architected Framework, que ayuda a los arquitectos de la nube a crear infraestructuras seguras, de alto rendimiento, resilientes y eficientes para una variedad de aplicaciones, cargas de trabajo y dominios tecnológicos. Para obtener más información, consulte el Machine Learning Lens ofrecido por AWS Well-Architected.

Recursos

Documentación de Amazon SageMaker AI

- Tienda de funciones de Amazon SageMaker AI
- Seguridad y control de acceso de Feature Store
- Valores de Shapley
- Depurador Amazon SageMaker AI
- Amazon SageMaker AI Pipelines
- Plantillas de proyectos predeterminadas de Amazon SageMaker AI
- SageMaker Inferencia de IA en tiempo real
- Escale automáticamente los modelos de Amazon SageMaker AI
- Inferencia asíncrona de Amazon SageMaker AI
- SageMaker Monitor de modelos de IA

AWS herramientas para desarrolladores

- AWS CodePipeline

AWS publicaciones de blog

- Comprensión de las funciones clave de Amazon SageMaker AI Feature Store

- Pruebe la calidad de los datos a escala con PyDeequ
- Experimentos de Amazon SageMaker AI
- Implemente y supervise de forma segura los SageMaker puntos de conexión de Amazon con y CodePipeline AWS CodeDeploy
- Implemente modelos de aprendizaje automático alternativo en Amazon SageMaker AI
- Pruebas A/B de modelos de aprendizaje automático en producción con Amazon AI SageMaker

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	—	20 de diciembre de 2021

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: Migrar la base de datos de Oracle en las instalaciones a Amazon Aurora PostgreSQL-Compatible Edition.
- Redefinir la plataforma (transportar y redefinir): traslade una aplicación a la nube e introduzca algún nivel de optimización para aprovechar las capacidades de la nube. Ejemplo: Migrar la base de datos Oracle en las instalaciones a Amazon Relational Database Service (Amazon RDS) para Oracle en la nube de Nube de AWS.
- Recomprar (readquirir): cambie a un producto diferente, lo cual se suele llevar a cabo al pasar de una licencia tradicional a un modelo SaaS. Ejemplo: Migrar el sistema de administración de las 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: Migrar la base de datos de Oracle en las instalaciones a Oracle en una instancia de EC2 en la Nube de AWS.
- Reubicar: (migrar el hipervisor mediante lift and shift): traslade la infraestructura a la nube sin comprar equipo nuevo, reescribir aplicaciones o modificar las operaciones actuales. Los servidores se migran de una plataforma en las instalaciones a un servicio en la nube para la misma plataforma. Ejemplo: migrar una Microsoft Hyper-V aplicación a AWS.
- Retener (revisitar): conserve las aplicaciones en el entorno de origen. Estas pueden incluir las aplicaciones que requieren una refactorización importante, que desee posponer para más adelante, y las aplicaciones heredadas que desee retener, ya que no hay ninguna justificación empresarial para migrarlas.

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

A

ABAC

Consulte control de acceso basado en atributos.

servicios abstractos

Consulte servicios administrados.

ACID

Consulte atomicidad, consistencia, aislamiento, durabilidad.

migración activa-activa

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

migración activa-pasiva

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

función de agregación

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

IA

Consulte inteligencia artificial.

AIOps

Consulte operaciones de inteligencia artificial

anonimización

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

antipatrones

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

control de aplicaciones

Enfoque de seguridad que permite usar de manera exclusiva aplicaciones aprobadas para ayudar a proteger un sistema contra el malware.

cartera de aplicaciones

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

inteligencia artificial (IA)

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

operaciones de inteligencia artificial (AIOps)

El proceso de utilizar técnicas de machine learning para resolver problemas operativos, reducir los incidentes operativos y la intervención humana, y mejorar la calidad del servicio. Para obtener más información sobre cómo 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

bot malicioso

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

BCP

Consulte planificación de la continuidad del negocio.

gráfico de comportamiento

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

sistema big-endian

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

clasificación binaria

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

filtro de floración

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

implementación azul/verde

Estrategia de implementación en la que se crean dos entornos separados, pero idénticos. La versión actual de la aplicación se ejecuta en un entorno (azul) y la nueva versión de la aplicación se ejecuta en el otro entorno (verde). Esta estrategia lo ayuda a hacer reversiones rápidas con un impacto mínimo.

bot

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

botnet

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

branch

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

acceso de emergencia

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

estrategia de implementación sobre infraestructura existente

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

caché de búfer

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

capacidad empresarial

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

planificación de la continuidad del negocio (BCP)

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

C

CAF

Consulte AWS Cloud Adoption Framework.

implementación canario

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

CCoE

Consulte Centro de excelencia en la nube.

CDC

Consulte captura de datos de cambios.

captura de datos de cambio (CDC)

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

ingeniería del caos

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

CI/CD

Consulte integración continua y entrega continua.

clasificación

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

cifrado del cliente

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

Centro de excelencia en la nube (CCoE)

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

computación en la nube

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

modelo operativo en la nube

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

etapas de adopción de la nube

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

- Proyecto: ejecución de algunos proyectos relacionados con la nube con fines de prueba de concepto y aprendizaje
- Fundamento: realizar inversiones fundamentales para escalar su adopción de la nube (p. ej., crear una landing zone, definir una CCo E, establecer un modelo de operaciones)

- Migración: migración de aplicaciones individuales
- Reinvención: optimización de productos y servicios e innovación en la nube

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

CMDB

Consulte base de datos de administración de configuración.

repositorio de código

Una ubicación donde el código fuente y otros activos, como documentación, muestras y scripts, se almacenan y actualizan mediante procesos de control de versiones. Algunos repositorios en la nube comunes son GitHub o Bitbucket Cloud. Cada versión del código se denomina rama. En una estructura de microservicios, cada repositorio se encuentra dedicado a una única funcionalidad. Una sola 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 machine learning para analizar y extraer información de formatos visuales, como imágenes y videos digitales. Por ejemplo, Amazon SageMaker AI proporciona algoritmos de procesamiento de imágenes para CV.

deriva de configuración

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

base de datos de administración de configuración (CMDB)

Repositorio que almacena y administra información sobre una base de datos y su entorno de TI, incluidos los componentes de hardware y software y sus configuraciones. Por lo general, los datos de una CMDB se utilizan en la etapa de detección y análisis de la cartera de productos durante la migración.

paquete de conformidad

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

integración y entrega continuas (CI/CD)

El proceso de automatización de las etapas de origen, compilación, prueba, puesta en escena y producción del proceso de publicación del software. CI/CD se describe comúnmente como una canalización. CI/CD puede ayudarlo a automatizar los procesos, mejorar la productividad, mejorar la calidad del código y entregar más rápido. Para obtener más información, consulte Beneficios de la entrega continua. CD también puede significar implementación continua. Para obtener más información, consulte Entrega continua frente a implementación continua.

CV

Consulte visión artificial.

D

datos en reposo

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

clasificación de datos

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

deriva de datos

Una variación significativa entre los datos de producción y los datos que se utilizaron para entrenar un modelo de machine learning, o un cambio significativo en los datos de entrada a lo largo del tiempo. La deriva de datos puede reducir la calidad, la precisión y la imparcialidad generales de las predicciones de los modelos de machine learning.

datos en tránsito

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

malla de datos

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

minimización de datos

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

perímetro de datos

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

preprocesamiento de datos

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

procedencia de los datos

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

titular de los datos

Persona cuyos datos se recopilan y procesan.

almacenamiento de datos

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

lenguaje de definición de datos (DDL)

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

lenguaje de manipulación de datos (DML)

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

DDL

Consulte lenguaje de definición de bases de datos.

conjunto profundo

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

aprendizaje profundo

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

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 en una tabla de hechos. Los atributos de la tabla de dimensiones suelen ser campos de texto o números discretos que se comportan como texto. Estos atributos se suelen utilizar para restringir consultas, filtrarlas y etiquetar los conjuntos de resultados.

desastre

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

recuperación de desastres (DR)

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

DML

Consulte lenguaje de manipulación de bases de datos.

diseño basado en el dominio

Un enfoque para desarrollar un sistema de software complejo mediante la conexión de sus componentes a dominios en evolución, o a los objetivos empresariales principales, a los que sirve cada componente. 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 desviaciones

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

DVSM

Consulte asignación de flujos de valor para el desarrollo.

E

EDA

Consulte análisis de datos de tipo exploratorio.

EDI

Consulte intercambio electrónico de datos.

computación en la periferia

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

intercambio electrónico de datos (EDI)

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

cifrado

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

clave de cifrado

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

endianidad

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

punto de conexión

Consulte punto de conexión de servicio.

servicio de punto de conexión

Servicio que puede alojar en una nube privada virtual (VPC) para compartir con otros usuarios. Puede crear un servicio de punto final AWS PrivateLink y conceder permisos a otros Cuentas de AWS o a responsables AWS Identity and Access Management (de 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)

Sistema que automatiza y administra los procesos empresariales clave (como la contabilidad, MES y la administración de proyectos) de una empresa.

cifrado de sobre

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

entorno

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

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

epopeya

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

ERP

Consulte planificación de recursos empresariales.

análisis de datos de tipo exploratorio (EDA)

El proceso de analizar un conjunto de datos para comprender sus características principales. Se recopilan o agregan datos y, a continuación, se realizan las investigaciones iniciales para encontrar patrones, detectar anomalías y comprobar las suposiciones. El EDA se realiza mediante el cálculo de estadísticas resumidas y la creación de visualizaciones de datos.

F

tabla de hechos

Tabla central de un esquema en estrella. Almacena datos cuantitativos sobre operaciones empresariales. Por lo general, una tabla de hechos contiene dos tipos de columnas: las que contienen medidas y las que contienen una clave externa para una tabla de dimensiones.

Fail Fast

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

límite de aislamiento de errores

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

rama de característica

Consulte rama.

características

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

importancia de las características

La importancia que tiene una característica para las predicciones de un modelo. Por lo general, esto se expresa como una puntuación numérica que se puede calcular mediante diversas

técnicas, como las explicaciones aditivas de Shapley (SHAP) y los gradientes integrados. Para obtener más información, consulte Interpretabilidad del modelo de aprendizaje automático con. AWS

transformación de funciones

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

peticiones con pocos pasos

Proporcionar a un LLM una pequeña cantidad de ejemplos que demuestren la tarea y el resultado deseado antes de pedirle que lleve a cabo una tarea similar. Esta técnica es una aplicación del aprendizaje contextual, mediante el que los modelos aprenden a partir de ejemplos (pasos) incrustados en las peticiones. La técnica de peticiones con pocos pasos puede ser eficaz para las tareas que requieren un formato, un razonamiento o un conocimiento del dominio específicos. Consulte también peticiones desde cero.

FGAC

Consulte control de acceso detallado.

control de acceso preciso (FGAC)

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

migración relámpago

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

FM

Consulte modelo fundacional.

Modelo fundacional (FM)

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 más información, consulte ¿Qué son los modelos fundacionales?

G

IA generativa

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

bloqueo geográfico

Consulte restricciones geográficas.

restricciones geográficas (bloqueo geográfico)

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

Flujo de trabajo de Gitflow

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

imagen dorada

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

estrategia de implementación desde cero

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

barrera de protección

Una regla de alto nivel que ayuda a regular los recursos, las políticas y 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 CSPM GuardDuty AWS Trusted Advisor, Amazon Inspector y AWS Lambda cheques personalizados.

H

HA

Consulte alta disponibilidad.

migración heterogénea de bases de datos

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

alta disponibilidad (HA)

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

modernización histórica

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

datos de reserva

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

migración homogénea de bases de datos

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

datos recientes

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

hotfix

Una solución urgente para un problema crítico en un entorno de producción. Debido a su urgencia, una revisión suele realizarse fuera del flujo de trabajo de DevOps publicación típico.

periodo de hiperatención

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

I

IaC

Consulte infraestructura como código.

políticas basadas en identidades

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

aplicación inactiva

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

IIoT

Consulte Internet de las cosas industrial.

infraestructura inmutable

Modelo que implementa una nueva infraestructura para las cargas de trabajo de producción en lugar de actualizar o modificar la infraestructura existente o aplicarle revisiones. Las infraestructuras inmutables son de manera intrínseca más coherentes, fiables y predecibles que las infraestructuras mutables. Para más información, consulte la práctica recomendada Implementación mediante una infraestructura inmutable en el Marco de AWS Well-Architected.

VPC entrante (de entrada)

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

migración gradual

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

Industria 4.0

Término que introdujo Klaus Schwab en 2016 para referirse a la modernización de los procesos de fabricación mediante los avances en la conectividad, los datos en tiempo real, la automatización, el análisis, la IA y el ML.

infraestructura

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

infraestructura como código (IaC)

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

Internet de las cosas industrial (T) IIo

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

VPC de inspección

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

Internet de las cosas (IoT)

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

interpretabilidad

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

IoT

Consulte Internet de las cosas.

biblioteca de información de TI (ITIL)

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

administración de servicios de TI (ITSM)

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

ITIL

Consulte biblioteca de información de TI.

ITSM

Consulte administración de servicios de TI.

L

control de acceso basado en etiquetas (LBAC)

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

zona de aterrizaje

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

modelo de lenguaje de gran tamaño (LLM)

Modelo de IA de aprendizaje profundo que se entrenó previamente con una gran cantidad de datos. Un LLM puede llevar a cabo varias tareas, como responder preguntas, resumir documentos, traducir textos a otros idiomas y completar oraciones. Para 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

Consulte Las 7 R.

sistema little-endian

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

LLM

Consulte modelo de lenguaje de gran tamaño.

entornos inferiores

Consulte entorno.

M

machine learning (ML)

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

rama principal

Consulte rama.

malware

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

Servicios administrados

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

sistema de ejecución de fabricación (MES)

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

MAP

Consulte Programa de aceleración de la migración.

mecanismo

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

Message Queuing Telemetry Transport (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: rehospede la migración a Amazon EC2 AWS con Application Migration Service.

Migration Portfolio Assessment (MPA)

Herramienta en línea que proporciona información a fin de validar los argumentos comerciales necesarios para migrar a la Nube de AWS. La MPA ofrece una evaluación detallada de la cartera

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

Evaluación de la preparación para la migración (MRA)

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

estrategia de migración

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

ML

Consulte machine learning.

modernización

Transformar una aplicación obsoleta (antigua o monolítica) y su infraestructura en un sistema ágil, elástico y de alta disponibilidad en la nube para reducir los gastos, aumentar la eficiencia y aprovechar las innovaciones. Para más información, consulte Strategy for modernizing applications in the Nube de AWS.

evaluación de la preparación para la modernización

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

aplicaciones monolíticas (monolitos)

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

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

MPA

Consulte Migration Portfolio Assessment.

MQTT

Consulte Message Queuing Telemetry Transport.

clasificación multiclase

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

infraestructura mutable

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

O

OAC

Consulte control de acceso de origen.

OAI

Consulte identidad de acceso de origen.

OCM

Consulte administración del cambio organizacional.

migración fuera de línea

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

OI

Consulte integración de operaciones.

OLA

Consulte acuerdo de nivel operativo.

migración en línea

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

OPC-UA

Consulte Open Process Communications: arquitectura unificada.

Open Process Communications: arquitectura unificada (OPC-UA)

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

acuerdo de nivel operativo (OLA)

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

revisión de la preparación operativa (ORR)

Lista de comprobación de preguntas y prácticas recomendadas asociadas que son útiles para comprender, evaluar, prevenir o reducir el alcance de los incidentes y posibles errores. Para más información, consulte Operational Readiness Reviews (ORR) en el Marco de AWS Well-Architected.

tecnología operativa (TO)

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

integración de operaciones (OI)

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

registro de seguimiento organizativo

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

administración del cambio organizacional (OCM)

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

control de acceso de origen (OAC)

En CloudFront, una opción mejorada para restringir el acceso y proteger el contenido del Amazon Simple Storage Service (Amazon S3). El OAC admite todos los buckets de S3 Regiones de AWS, el cifrado del lado del servidor 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 revisión de la preparación operativa.

OT

Consulte tecnología operativa.

VPC saliente (de salida)

En una arquitectura de AWS cuentas múltiples, una VPC que gestiona las conexiones de red que se inician desde una aplicación. La arquitectura 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 información de identificación personal.

manual de estrategias

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

PLC

Consulte controlador lógico programable.

PLM

Consulte administración del ciclo de vida del producto.

policy

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

persistencia políglota

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

evaluación de cartera

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

predicate

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

inserción de predicados

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

control preventivo

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

entidad principal

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

Privacidad desde el diseño

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

zonas alojadas privadas

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

Control de seguridad que se diseñó para evitar la implementación de recursos que no cumplan con la normativa. Estos controles analizan los recursos antes de aprovisionarlos. Si el recurso no cumple con los requisitos del control, no se aprovisiona. Para obtener más información, consulte la guía de referencia de controles en la AWS Control Tower documentación y consulte Controles proactivos en la sección Implementación de controles de seguridad en AWS.

administración del ciclo de vida del producto (PLM)

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

entorno de producción

Consulte entorno.

controlador lógico programable (PLC)

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

encadenamiento de peticiones

Uso de la salida de una petición de LLM como entrada para la siguiente petición a fin de generar mejores respuestas. Esta técnica se utiliza para dividir una tarea compleja en tareas secundarias o para refinar o ampliar de forma iterativa una respuesta preliminar. Ayuda a mejorar la precisión y la relevancia de las respuestas de un modelo y permite obtener resultados más detallados y personalizados.

seudonimización

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

publish/subscribe (pub/sub)

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

Q

plan de consulta

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

regresión del plan de consulta

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

R

Matriz RACI

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

RAG

Consulte generación aumentada por recuperación.

ransomware

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

Matriz RASCI

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

RCAC

Consulte control de acceso por filas y columnas.

réplica de lectura

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

rediseñar

Consulte Las 7 R.

objetivo de punto de recuperación (RPO)

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

objetivo de tiempo de recuperación (RTO)

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

refactorizar

Consulte Las 7 R.

Region

Conjunto de AWS recursos en un área geográfica. Cada uno Región de AWS está aislado e independiente de los demás para proporcionar tolerancia a las fallas, estabilidad y resiliencia. Para más información, consulte Specify which Regiones de AWS your account can use.

regresión

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

volver a alojar

Consulte Las 7 R.

versión

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

reubicar

Consulte Las 7 R.

redefinir la plataforma

Consulte Las 7 R.

recomprar

Consulte Las 7 R.

resiliencia

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

política basada en recursos

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

matriz responsable, confiable, consultada e informada (RACI)

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

control receptivo

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

retain

Consulte Las 7 R.

retirar

Consulte Las 7 R.

Generación aumentada de recuperación (RAG)

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

rotación

Proceso mediante el que periódicamente se actualiza un secreto para que resulte más difícil que un atacante pueda acceder a las credenciales.

control de acceso por filas y columnas (RCAC)

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

RPO

Consulte objetivo de punto de recuperación.

RTO

Consulte objetivo de tiempo de recuperación.

manual de procedimientos

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

S

SAML 2.0

Un estándar abierto que utilizan muchos proveedores de identidad (IdPs). Esta función permite el inicio de sesión único (SSO) federado, de modo que los usuarios pueden iniciar sesión Consola de administración de AWS o llamar a las operaciones de la AWS API sin tener que crear un

usuario en IAM para todos los miembros de la organización. Para obtener más información sobre la federación basada en SAML 2.0, consulte Acerca de la federación basada en SAML 2.0 en la documentación de IAM.

SCADA

Consulte control de supervisión y adquisición de datos.

SCP

Consulte política de control de servicio.

secreta

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

seguridad desde el diseño

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

control de seguridad

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

refuerzo de la seguridad

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

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

Herramientas y servicios que combinan sistemas de administración de información sobre seguridad (SIM) y de administración de eventos de seguridad (SEM). Un sistema de SIEM recopila, monitorea y analiza los datos de servidores, redes, dispositivos y otras fuentes para detectar amenazas y brechas de seguridad y generar alertas.

automatización de la respuesta de seguridad

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

cifrado del servidor

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

política de control de servicio (SCP)

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)

Métrica objetivo que representa el estado de un servicio medido mediante un indicador de nivel de servicio.

modelo de responsabilidad compartida

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

SIEM

Consulte sistema de administración de eventos e información de seguridad.

único punto de error (SPOF)

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

SLA

Consulte acuerdo de nivel de servicio.

SLI

Consulte indicador de nivel de servicio.

SLO

Consulte objetivo de nivel de servicio.

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

SPOF

Consulte único punto de error.

esquema en estrella

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

patrón de higo estrangulador

Un enfoque para modernizar los sistemas monolíticos mediante la reescritura y el reemplazo gradual de las funciones del sistema hasta que se pueda desmantelar el sistema heredado. Este patrón utiliza la analogía de una higuera que crece hasta convertirse en un árbol estable y, finalmente, se apodera y reemplaza a su host. El patrón fue presentado por Martin Fowler como una forma de gestionar el riesgo al reescribir sistemas monolíticos. Para ver un ejemplo 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.

control de supervisión y adquisición de datos (SCADA)

En el sector de la fabricación, sistema que utiliza hardware y software para supervisar los activos físicos y las operaciones de producción.

cifrado simétrico

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

pruebas sintéticas

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

petición del sistema

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

T

etiquetas

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

variable de destino

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

lista de tareas

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

entorno de prueba

Consulte entorno.

entrenamiento

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

puerta de enlace de tránsito

Un centro de tránsito de red que puede usar para interconectar sus redes con VPCs 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

Consulte entorno.

V

succión

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

control de versión

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

Emparejamiento de VPC

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

carga de trabajo

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

flujo de trabajo

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

WORM

Consulte escritura única y lectura múltiple.

WQF

Consulte AWS Workload Qualification Framework.

escritura única y lectura múltiple (WORM)

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

Z

ataque de día cero

Ataque, normalmente de malware, que se aprovecha de una vulnerabilidad de día cero.

vulnerabilidad de día cero

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

peticiones desde cero

Proporcionar a un LLM instrucciones para llevar a cabo una tarea, pero sin ejemplos (pasos) que puedan ayudar a guiarlo. El LLM debe usar los conocimientos del entrenamiento previo para

llevar a cabo la tarea. La eficacia de la petición desde cero depende de la complejidad de la tarea y de la calidad de la petición. Consulte también peticiones con pocos pasos.

aplicación zombi

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

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