Uso de Amazon Comprehend Medical LLMs y para la salud y las ciencias de la vida

AWS Guía prescriptiva

AWS Guía prescriptiva: Uso de Amazon Comprehend Medical LLMs y para la salud y las ciencias de la vida

Table of Contents

Uso de Amazon Comprehend Medical LLMs y para la salud y las ciencias de la vida

Amazon Web Services (???colaboradores)

Diciembre de 2025 (historial del documento)

Descripción general de

El volumen cada vez mayor de datos médicos y la necesidad de un procesamiento eficiente y preciso han impulsado la adopción del procesamiento del lenguaje natural (PNL) con tecnologías de inteligencia artificial y aprendizaje automático (AI/ML). Los modelos clasificadores previamente entrenados y los modelos de lenguajes extensos (LLMs) se han convertido en herramientas poderosas para diversas tareas de la PNL médica, como la respuesta a preguntas clínicas, el resumen de informes y la generación de información. Sin embargo, el ámbito de la salud y las ciencias de la vida presenta desafíos únicos debido a la complejidad de la terminología médica, los conocimientos específicos del campo y los requisitos reglamentarios. El uso efectivo de clasificadores previamente entrenados o LLMs en este dominio requiere un enfoque bien diseñado que combine los puntos fuertes de estos modelos con recursos y técnicas específicos del campo.

Las prácticas de la industria en los ámbitos de la salud y las ciencias de la vida se han basado tradicionalmente en sistemas basados en reglas, en la codificación manual y en los procesos de revisión por parte de expertos. Estos sistemas y procesos requieren mucho tiempo y son propensos a errores. La integración de las tecnologías de IA y PNL, como Amazon Comprehend Medical y los modelos básicos de Amazon Bedrock, ofrece soluciones eficientes y escalables para procesar datos médicos y, al mismo tiempo, mejorar la precisión y la coherencia.

Esta guía explora el uso de Amazon Comprehend Medical LLMs y la automatización inteligente en el sector de la salud. Describe las mejores prácticas, los desafíos y los enfoques prácticos para agilizar los procesos de codificación médica, extracción de información de los pacientes y resumen de registros. Al utilizar las capacidades de Amazon Comprehend Medical LLMs y, las organizaciones de atención médica pueden alcanzar nuevos niveles de eficiencia operativa, reducir los costos y, potencialmente, mejorar la atención a los pacientes.

La guía detalla las consideraciones únicas del ámbito de la salud, como comprender la terminología médica, utilizar dominios específicos LLMs y abordar las limitaciones de los sistemas. AI/ML

Proporciona una ruta integral de toma de decisiones para que los administradores de TI, los arquitectos y los líderes técnicos del sector sanitario evalúen la preparación de la organización, evalúen las opciones de implementación y utilicen las herramientas adecuadas Servicios de AWS para una automatización exitosa.

Al seguir las directrices y las mejores prácticas descritas en esta guía, las organizaciones sanitarias pueden aprovechar el poder de las AI/ML tecnologías y, al mismo tiempo, sortear las complejidades del ámbito médico. Este enfoque apoya el cumplimiento de las directrices éticas y reglamentarias y promueve el uso responsable de los sistemas de IA en la atención médica. Está diseñado para generar información precisa y privada.

Destinatarios previstos

Esta guía está dirigida a las partes interesadas en la tecnología, los arquitectos, los líderes técnicos y los responsables de la toma de decisiones que desean implementar soluciones de procesamiento del lenguaje natural impulsadas por la IA para el análisis y la automatización de los datos médicos.

Objetivos

Las organizaciones sanitarias y de ciencias de la vida pueden cumplir varios objetivos empresariales con Amazon Comprehend Medical LLMs y. Estos resultados suelen incluir el aumento de la eficiencia operativa, la reducción de los costes y la mejora de la atención a los pacientes. En esta sección se describen los objetivos empresariales clave y los beneficios asociados a la implementación de las estrategias y las mejores prácticas descritas en esta guía.

Los siguientes son algunos de los objetivos que las organizaciones pueden alcanzar al implementar las directrices y las mejores prácticas de esta guía:

- Reducir el tiempo de desarrollo: el objetivo final de esta guía es reducir el tiempo de desarrollo con los costos, disminuir la deuda técnica y mitigar los posibles fracasos de los proyectos derivados de la POC. Al comprender AI/ML los servicios clave, como Amazon Comprehend Medical, y las ventajas y limitaciones del uso de la LLM para las tareas de atención médica, las empresas pueden lograr una comercialización más rápida y acelerar el cumplimiento de los objetivos empresariales.
- Extraiga información para automatizar las tareas de codificación médica: tras las visitas de los pacientes, los especialistas en codificación y los proveedores pueden extraer información de los textos médicos, como notas subjetivas, objetivas, de evaluación y de planes (SOAP). Esto

puede reducir los esfuerzos de documentación manual y ayudar al proveedor a centrarse en las necesidades del paciente. Al combinar las capacidades de reconocimiento de entidades de Amazon Comprehend Medical LLMs, las organizaciones pueden extraer información médica relevante de los registros de los pacientes, las notas clínicas y otras fuentes de datos de atención médica. Esto puede minimizar los errores humanos y promover prácticas coherentes.

- Resuma los registros de los pacientes y la documentación clínica: el resumen automático del historial del paciente, los planes de tratamiento y los resultados médicos puede ahorrar un tiempo valioso a los proveedores de atención médica. LLMs puede ayudar a generar una documentación clínica completa y estructurada. Puede obtener más contexto con Amazon Comprehend Medical, utilizar un LLM de dominio médico o ajustar un LLM con datos médicos. Estos enfoques pueden ayudar a proporcionar resúmenes precisos y a garantizar que la documentación cumpla con los requisitos y estándares de conformidad.
- Apoye las decisiones clínicas y la atención de los pacientes: mediante el uso de enlaces ontológicos en Amazon Comprehend Medical y mediante el uso LLMs, los proveedores pueden responder a preguntas médicas o solicitar recomendaciones sobre la atención de los pacientes. Esto permite a los profesionales de la salud tomar decisiones informadas que mejoran los resultados de los pacientes y reducen el riesgo de errores médicos.

Enfoques generativos de IA y PNL para la salud y las ciencias de la vida

El procesamiento del lenguaje natural (PNL) es una tecnología de aprendizaje automático que permite a las computadoras interpretar, manipular y comprender el lenguaje humano. Las organizaciones sanitarias y de ciencias de la vida disponen de grandes volúmenes de datos procedentes de las historias clínicas de los pacientes. Pueden usar el software de PNL para procesar automáticamente estos datos. Por ejemplo, pueden combinar la PNL con la IA generativa para agilizar la codificación médica, extraer información de los pacientes y resumir los registros.

Según la tarea de PNL que desee realizar, es posible que las diferentes arquitecturas sean las más adecuadas para su caso de uso. Esta guía aborda las siguientes opciones generativas de IA y PNL para aplicaciones sanitarias y de ciencias de la vida en: AWS

- Uso de Amazon Comprehend Medical— Aprenda a utilizar Amazon Comprehend Medical de forma independiente, sin necesidad de integrarlo con un modelo de lenguaje grande (LLM).
- Combinación de Amazon Comprehend Medical con modelos lingüísticos de gran tamaño— Obtenga información sobre cómo combinar Amazon Comprehend Medical con un LLM en una arquitectura de generación aumentada de recuperación (RAG).
- Uso de modelos de lenguaje extensos para casos de uso de la salud y las ciencias de la vida— Obtenga información sobre cómo usar un LLM para aplicaciones de salud y ciencias de la vida, ya sea mediante una arquitectura LLM ajustada o una RAG.

Uso de Amazon Comprehend Medical

Amazon Comprehend Medical detecta Servicio de AWS y devuelve información útil en textos clínicos no estructurados, como notas del médico, resúmenes de alta, resultados de pruebas y notas de casos. Utiliza modelos de procesamiento del lenguaje natural (PNL) para detectar entidades. Las entidades son referencias textuales a información médica, como afecciones médicas, medicamentos o información de salud protegida (PHI).

> **Important**
>
> Amazon Comprehend Medical no sustituye el asesoramiento, el diagnóstico ni el tratamiento médico profesional. Amazon Comprehend Medical proporciona puntuaciones de confianza

que indican el nivel de confianza en la precisión de las entidades detectadas. Identifique el umbral de confianza adecuado para su caso de uso y utilice umbrales de confianza altos en situaciones que requieran una alta precisión. En ciertos casos de uso, los resultados deberán ser revisados y verificados por revisores humanos debidamente entrenados. Por ejemplo, Amazon Comprehend Medical solo debe utilizarse en escenarios de atención al paciente después de que un profesional médico debidamente formado haya revisado su exactitud y buen juicio médico.

Puede acceder a Amazon Comprehend Medical a través Consola de administración de AWS del, AWS Command Line Interface el AWS CLI() o mediante AWS SDKs el. AWS SDKs Están disponibles para varios lenguajes de programación y plataformas, como Java, Python, Ruby, .NET, iOS y Android. Puede utilizarla para acceder mediante programación SDKs a Amazon Comprehend Medical desde su aplicación cliente.

En esta sección se analizan las principales capacidades de Amazon Comprehend Medical. También se analizan las ventajas de utilizar este servicio en comparación con un modelo de lenguaje amplio (LLM).

Capacidades de Amazon Comprehend Medical

Amazon Comprehend Medical APIs ofrece inferencias por lotes y casi en tiempo real. APIs Pueden asimilar textos médicos y proporcionar resultados para las tareas de PNL médicas mediante el reconocimiento de entidades médicas y la identificación de las relaciones entre entidades. Puede realizar análisis tanto en archivos individuales como en lotes en varios archivos almacenados en un bucket de Amazon Simple Storage Service (Amazon S3). Amazon Comprehend Medical ofrece las siguientes operaciones de API de análisis de texto para la detección de entidades sincrónicas:

- Detecta entidades: detecta categorías médicas generales, como la anatomía, la afección médica, la categoría de PHI, los procedimientos y las expresiones horarias.
- Detectar la PHI: detecta entidades específicas, como la edad, la fecha, el nombre e información personal similar.

Amazon Comprehend Medical también incluye varias operaciones de API que puede utilizar para realizar análisis de texto por lotes en documentos clínicos. Para obtener más información sobre cómo utilizar estas operaciones de API, consulte Análisis de texto por lotes APIs.

Utilice Amazon Comprehend Medical para detectar entidades en textos clínicos y vincular esas entidades con conceptos de ontologías médicas estandarizadas, incluidas RxNorm las bases de conocimiento ICD-10-CM y SNOMED CT. Puede realizar análisis tanto en archivos individuales como en lotes en documentos grandes o en varios archivos almacenados en un bucket de Amazon S3. Amazon Comprehend Medical ofrece la siguiente ontología que vincula las operaciones de la API:

- Infer ICD10 CM: la operación Infer ICD10 CM detecta posibles afecciones médicas y las vincula a los códigos de la versión de 2019 de la décima revisión, modificación clínica (ICD-10-CM) de la Clasificación Internacional de Enfermedades. Para cada posible afección médica detectada, Amazon Comprehend Medical muestra los códigos y las descripciones correspondientes de la ICD-10-CM. Las afecciones médicas que aparecen en los resultados incluyen una puntuación de confianza, que indica la confianza que Amazon Comprehend Medical tiene en la precisión de las entidades asociadas a los conceptos correspondientes de los resultados.
- InferRxNorm— La InferRxNormoperación identifica como entidades los medicamentos que figuran en la historia clínica de un paciente. Vincula las entidades con los identificadores conceptuales (RxCUI) de la RxNorm base de datos de la Biblioteca Nacional de Medicina. Cada RxCUI es único para diferentes concentraciones y formas de dosificación. Los medicamentos incluidos en los resultados incluyen una puntuación de confianza, que indica la confianza que Amazon Comprehend Medical tiene en la precisión de las entidades que coinciden con los conceptos de RxNorm la base de conocimientos. Amazon Comprehend Medical enumera los mejores medicamentos CUIs recetados que podrían coincidir con cada medicamento que detecte en orden descendente según la puntuación de confianza.
- InfersnomedCT: la operación InfersnomedCT identifica los posibles conceptos médicos como entidades y los vincula a los códigos de la versión 2021-03 de la Nomenclatura Sistematizada de Términos Clínicos de Medicina (SNOMED CT). SNOMED CT proporciona un vocabulario completo de conceptos médicos, que incluye afecciones médicas y anatomía, así como pruebas, tratamientos y procedimientos médicos. Para cada identificador de concepto coincidente, Amazon Comprehend Medical muestra los cinco conceptos médicos principales, cada uno con una puntuación de confianza e información contextual, como características y atributos. El concepto SNOMED CT se IDs puede utilizar entonces para estructurar los datos clínicos de los pacientes con fines de codificación médica, elaboración de informes o análisis clínicos si se utiliza junto con la polijerarquía de SNOMED CT.

Para obtener más información, consulte Análisis de texto APIs y enlace de ontologías APIs en la documentación de Amazon Comprehend Medical.

Casos de uso de Amazon Comprehend Medical

Como servicio independiente, Amazon Comprehend Medical podría abordar el caso de uso de su organización. Amazon Comprehend Medical puede realizar tareas como las siguientes:

- Ayuda con la codificación médica en los registros de los pacientes
- Detecte datos de información de salud protegida (PHI)
- Validar la medicación, incluidos atributos como la dosis, la frecuencia y la forma

Los resultados de Amazon Comprehend Medical son digeribles para la mayoría de los consultorios médicos. Sin embargo, es posible que deba considerar alternativas si tiene limitaciones como las siguientes:

- Distintas definiciones de entidad: por ejemplo, su definición `FREQUENCY` de entidad farmacológica puede diferir. En cuanto a la frecuencia, Amazon Comprehend Medical realiza las predicciones necesarias, pero su organización podría utilizar el término pro re nata (PRN).
- Cantidad abrumadora de resultados: por ejemplo, las notas de los pacientes suelen contener varios síntomas y palabras clave que se corresponden con varios códigos ICD-10-CM. Sin embargo, varias de las palabras clave no son aplicables al diagnóstico. En este caso, el proveedor debe evaluar numerosas entidades de la ICD-10-CM y sus puntuaciones de confianza, lo que requiere un tiempo de procesamiento manual.
- Entidades personalizadas o tareas de PNL: por ejemplo, es posible que los proveedores deseen extraer pruebas de la PRN, por ejemplo, tomarlas según sea necesario en caso de dolor. Como no está disponible a través de Amazon Comprehend Medical, se necesita un modelo AI/ML diferente. Se requiere una AI/ML solución diferente si la tarea de PNL está fuera del reconocimiento de la entidad, como el resumen, la respuesta a preguntas y el análisis de opiniones.

Combinación de Amazon Comprehend Medical con modelos lingüísticos de gran tamaño

Un estudio realizado en 2024 por NEJM AI demostró que, por lo general, el uso de un LLM, sin necesidad de preguntar nada, para tareas de codificación médica se traduce en un rendimiento deficiente. El uso de Amazon Comprehend Medical con un LLM puede ayudar a mitigar estos problemas de rendimiento. Los resultados de Amazon Comprehend Medical son un contexto útil para

un LLM que realiza tareas de PNL. Por ejemplo, proporcionar el contexto de Amazon Comprehend Medical al modelo de lenguaje amplio puede ayudarle a:

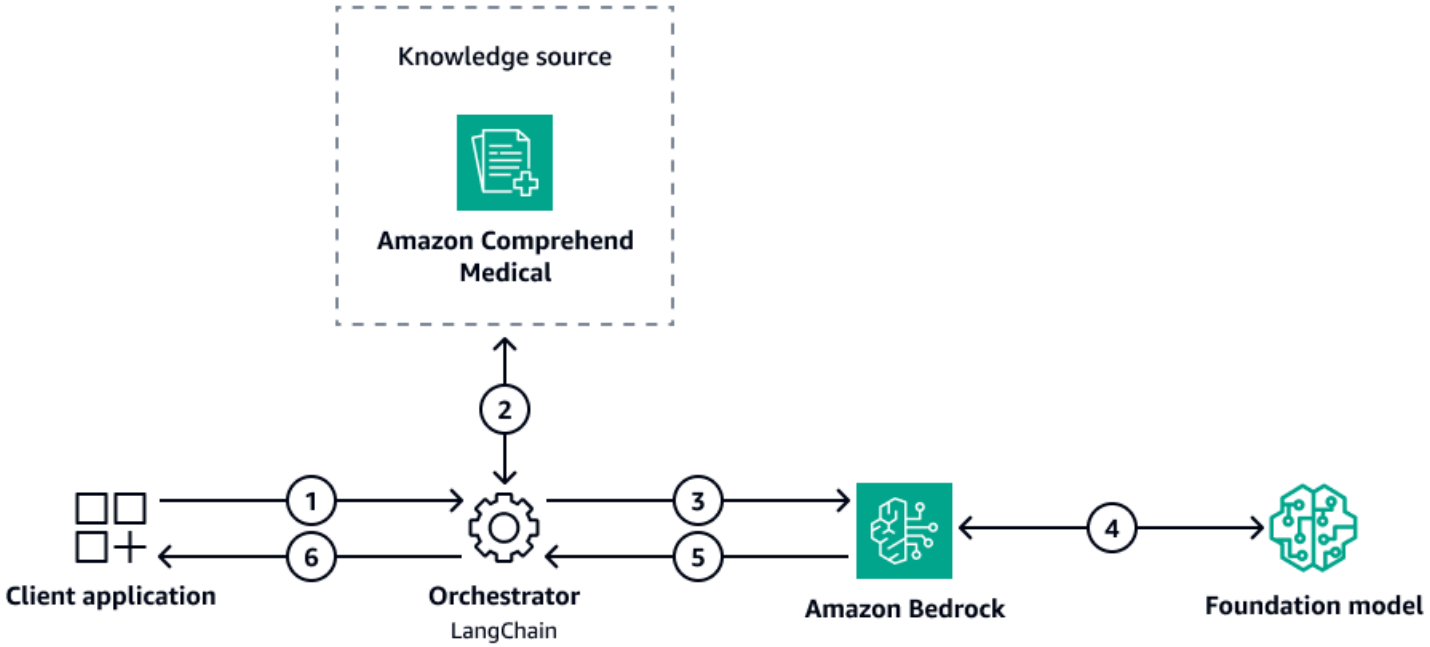
- Mejore la precisión de la selección de entidades utilizando los resultados iniciales de Amazon Comprehend Medical como contexto para el LLM
- Implemente el reconocimiento de entidades, el resumen, la respuesta a preguntas y otros casos de uso personalizados

En esta sección se describe cómo puede combinar Amazon Comprehend Medical con un LLM mediante un enfoque de generación aumentada de recuperación (RAG). La generación aumentada de recuperación (RAG) es una tecnología de IA generativa en la que un LLM hace referencia a una fuente de datos autorizada que se encuentra fuera de sus fuentes de datos de entrenamiento antes de generar una respuesta. Para más información, consulte ¿Qué es RAG (generación aumentada por recuperación)?

Para ilustrar este enfoque, en esta sección se utiliza el ejemplo de codificación médica (de diagnóstico) relacionada con la ICD-10-CM. Incluye un ejemplo de arquitectura y plantillas de ingeniería rápidas para ayudar a acelerar la innovación. También incluye prácticas recomendadas para usar Amazon Comprehend Medical en un flujo de trabajo de RAG.

Arquitectura basada en RAG con Amazon Comprehend Medical

El siguiente diagrama ilustra un enfoque RAG para identificar los códigos de diagnóstico ICD-10-CM a partir de las notas de los pacientes. Utiliza Amazon Comprehend Medical como fuente de conocimiento. En un enfoque RAG, el método de recuperación suele recuperar información de una base de datos vectorial que contiene los conocimientos aplicables. En lugar de una base de datos vectorial, esta arquitectura utiliza Amazon Comprehend Medical para la tarea de recuperación. El orquestador envía la información de las notas del paciente a Amazon Comprehend Medical y recupera la información del código ICD-10-CM. El orquestador envía este contexto al modelo de base descendente (LLM), a través de Amazon Bedrock. El LLM genera una respuesta mediante la información del código ICD-10-CM y esa respuesta se devuelve a la aplicación cliente.



El diagrama muestra el siguiente flujo de trabajo de RAG:

1. La aplicación cliente envía las notas del paciente como una consulta al orquestador. Un ejemplo de estas notas de un paciente podría ser: «La paciente es una paciente de 71 años del Dr. X. La paciente acudió a la sala de emergencias anoche con un historial de dolor abdominal persistente de aproximadamente 7 a 8 días. No ha tenido fiebres ni escalofríos definidos ni antecedentes de ictericia. The patient denies any significant recent weight loss".
2. El orquestador utiliza Amazon Comprehend Medical para recuperar los códigos ICD-10-CM relevantes para la información médica de la consulta. Utiliza la API Infer ICD10CM para extraer e inferir los códigos ICD-10-CM de las notas de los pacientes.
3. El orquestador crea un mensaje que incluye la plantilla del mensaje, la consulta original y los códigos ICD-10-CM recuperados de Amazon Comprehend Medical. Envía este contexto mejorado a Amazon Bedrock.
4. Amazon Bedrock procesa la entrada y utiliza un modelo básico para generar una respuesta que incluye los códigos ICD-10-CM y las correspondientes pruebas de la consulta. La respuesta generada incluye los códigos ICD-10-CM identificados y las pruebas de las notas del paciente que respaldan cada código. A continuación, se muestra una respuesta de ejemplo:

```
<response>
<icd10>
<code>R10.9</code>
<evidence>history of abdominal pain</evidence>
```

```
</icd10>
<icd10>
<code>R10.30</code>
<evidence>history of abdominal pain</evidence>
</icd10>
</response>
```

5. Amazon Bedrock envía la respuesta generada al orquestador.
6. El orquestador devuelve la respuesta a la aplicación cliente, donde el usuario puede revisarla.

Casos de uso del uso de Amazon Comprehend Medical en un flujo de trabajo de RAG

Amazon Comprehend Medical puede realizar tareas específicas de PNL. Para obtener más información, consulte Casos de uso de Amazon Comprehend Medical.

Es posible que desee integrar Amazon Comprehend Medical en un flujo de trabajo de RAG para casos de uso avanzados, como los siguientes:

- Genere resúmenes clínicos detallados combinando datos médicos extraídos con información contextual de los historiales de los pacientes
- Automatice la codificación médica para casos complejos mediante el uso de entidades extraídas con información vinculada a ontologías para la asignación de códigos
- Automatice la creación de notas clínicas estructuradas a partir de texto no estructurado mediante el uso de entidades médicas extraídas
- Analice los efectos secundarios de los medicamentos en función de los nombres y atributos de los medicamentos extraídos
- Desarrolle sistemas inteligentes de apoyo clínico que combinen la información médica extraída con up-to-date investigaciones y directrices

Mejores prácticas para usar Amazon Comprehend Medical en un flujo de trabajo de RAG

Al integrar los resultados de Amazon Comprehend Medical en una solicitud para obtener un LLM, es esencial seguir las mejores prácticas. Esto puede mejorar el rendimiento y la precisión. Las siguientes son recomendaciones clave:

- Comprenda las puntuaciones de confianza de Amazon Comprehend Medical: Amazon Comprehend Medical proporciona las puntuaciones de confianza para cada entidad detectada y cada enlace ontológico. Es fundamental entender el significado de estas puntuaciones y establecer los umbrales adecuados para su caso de uso específico. Las puntuaciones de confianza ayudan a filtrar las entidades de baja confianza, lo que reduce el ruido y mejora la calidad de las aportaciones del LLM.
- Utilice las puntuaciones de confianza en la ingeniería rápida: cuando elabore las indicaciones para el LLM, considere la posibilidad de incorporar las puntuaciones de confianza de Amazon Comprehend Medical como contexto adicional. Esto ayuda al LLM a priorizar o sopesar las entidades en función de sus niveles de confianza, lo que podría mejorar la calidad del resultado.
- Evalúe los resultados de Amazon Comprehend Medical con datos reales: los datos reales son información que se sabe que es verdadera. Se pueden usar para validar que una AI/ML aplicación produce resultados precisos. Antes de integrar los resultados de Amazon Comprehend Medical en su flujo de trabajo de LLM, evalúe el rendimiento del servicio en una muestra representativa de sus datos. Compare los resultados con anotaciones basadas en datos básicos para identificar posibles discrepancias o áreas de mejora. Esta evaluación le ayuda a comprender los puntos fuertes y las limitaciones de Amazon Comprehend Medical para su caso de uso.
- Seleccione estratégicamente la información relevante: Amazon Comprehend Medical puede proporcionarle una gran cantidad de información, pero es posible que no toda sea relevante para su tarea. Seleccione cuidadosamente las entidades, los atributos y los metadatos que sean más relevantes para su caso de uso. Proporcionar demasiada información irrelevante al LLM puede generar ruido y, potencialmente, disminuir el rendimiento.
- Alinee las definiciones de entidades: asegúrese de que las definiciones de entidades y atributos utilizadas por Amazon Comprehend Medical se ajusten a su interpretación. Si hay discrepancias, considere la posibilidad de proporcionar un contexto o una aclaración adicionales al LLM para cerrar la brecha entre la producción de Amazon Comprehend Medical y sus requisitos. Si la entidad Amazon Comprehend Medical no cumple sus expectativas, puede implementar una detección de entidad personalizada incluyendo instrucciones adicionales (y posibles ejemplos) en el mensaje.
- Proporcione conocimientos específicos del dominio: si bien Amazon Comprehend Medical proporciona información médica valiosa, es posible que no capture todos los matices de su dominio específico. Considere la posibilidad de complementar los resultados de Amazon Comprehend Medical con fuentes de conocimiento adicionales específicas del dominio, como ontologías, terminologías o conjuntos de datos seleccionados por expertos. Esto proporciona un contexto más completo al LLM.

- Cumpla con las directrices éticas y reglamentarias: cuando se trate de datos médicos, es importante cumplir con los principios éticos y las directrices reglamentarias, como las relacionadas con la privacidad de los datos, la seguridad y el uso responsable de los sistemas de IA en la atención médica. Asegúrese de que su implementación cumpla con las leyes pertinentes y las mejores prácticas del sector.

Al seguir estas mejores prácticas, AI/ML los profesionales pueden utilizar eficazmente los puntos fuertes de Amazon Comprehend Medical LLMs y. En el caso de las tareas de PNL médica, estas mejores prácticas ayudan a mitigar los posibles riesgos y pueden mejorar el rendimiento.

Ingeniería rápida para el contexto de Amazon Comprehend Medical

La ingeniería rápida es el proceso de diseñar y refinar las indicaciones para guiar una solución de IA generativa a fin de generar los resultados deseados. Tú eliges los formatos, frases, palabras y símbolos más adecuados para guiar a la IA a interactuar con tus usuarios de forma más significativa.

En función de la operación de API que realice, Amazon Comprehend Medical devuelve las entidades detectadas, los códigos y descripciones de ontología y las puntuaciones de confianza. Estos resultados se contextualizan en el mensaje cuando la solución invoca el LLM de destino. Debe diseñar la solicitud para que presente el contexto dentro de la plantilla de solicitud.

> **Note**
>
> Los ejemplos de instrucciones de esta sección siguen la guía antrópica. Si utilizas un proveedor de LLM diferente, sigue las recomendaciones de ese proveedor.

En general, en el mensaje se insertan tanto el texto médico original como los resultados de Amazon Comprehend Medical. La siguiente es una estructura de pronósticos común:

```
<medical_text>
medical text
</medical_text>

<comprehend_medical_text_results>
comprehend medical text results
</comprehend_medical_text_results>
```

```
<prompt_instructions>
prompt instructions
</prompt_instructions>
```

En esta sección se proporcionan estrategias para incluir los resultados de Amazon Comprehend Medical como contexto rápido para las siguientes tareas habituales de la PNL médica:

- Filtrar los resultados de Amazon Comprehend Medical
- Amplíe las tareas de PNL médica con Amazon Comprehend Medical
- Aplique barandas con Amazon Comprehend Medical

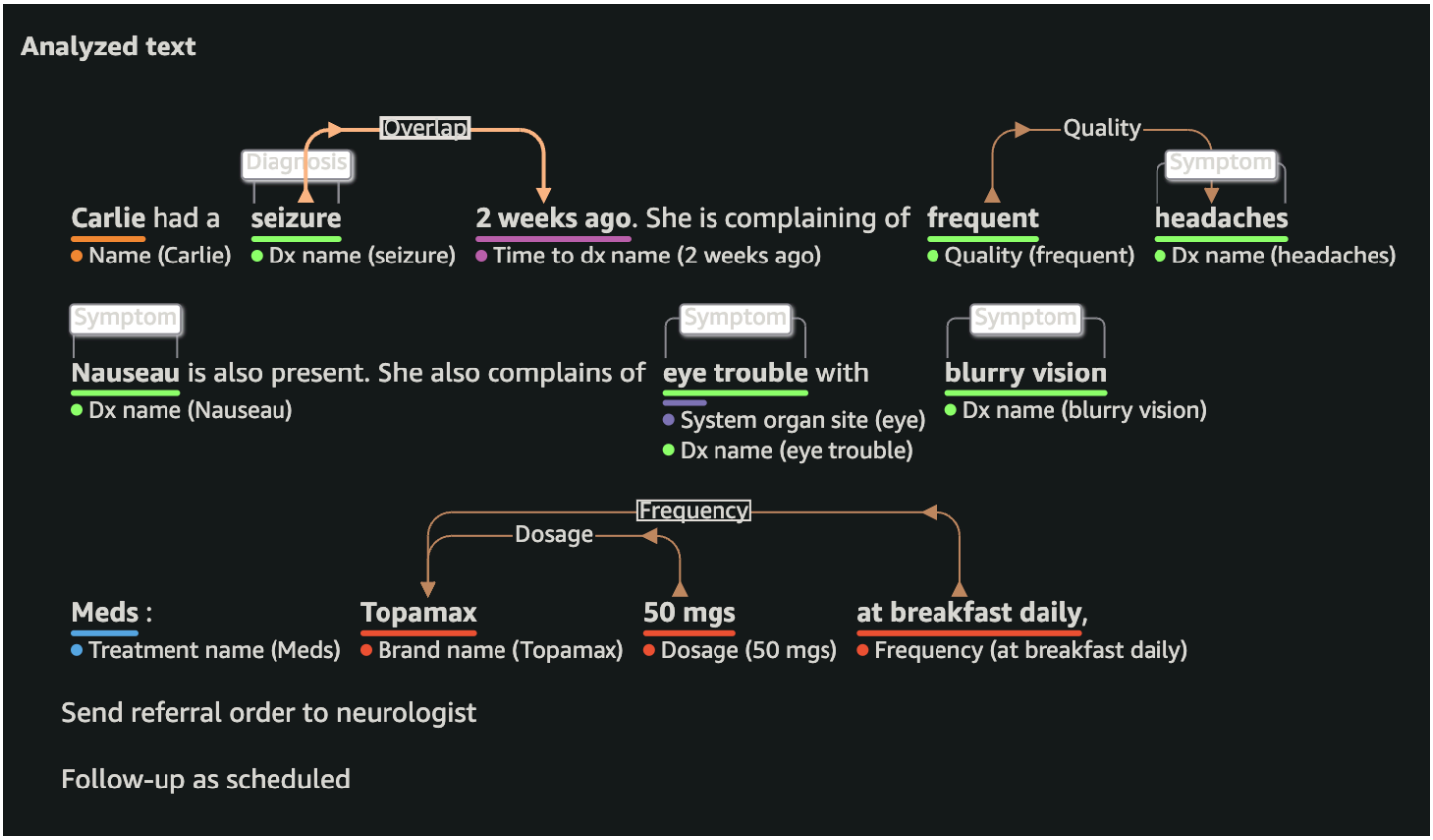
Filtrar los resultados de Amazon Comprehend Medical

Amazon Comprehend Medical suele proporcionar una gran cantidad de información. Es posible que desee reducir la cantidad de resultados que el profesional médico debe revisar. En este caso, puede usar un LLM para filtrar estos resultados. Las entidades Amazon Comprehend Medical incluyen una puntuación de confianza que puede utilizar como mecanismo de filtrado al diseñar el mensaje.

El siguiente es un ejemplo de nota para un paciente:

```
Carlie had a seizure 2 weeks ago. She is complaining of frequent headaches
Nausea is also present. She also complains of eye trouble with blurry vision
Meds : Topamax 50 mgs at breakfast daily,
Send referral order to neurologist
Follow-up as scheduled
```

En esta nota para un paciente, Amazon Comprehend Medical detecta las siguientes entidades.



Las entidades enlazan con los siguientes códigos ICD-10-CM para detectar convulsiones y cefaleas.

Categoría	Código ICD-10-CM	Descripción del ICD-10-CM	Puntuación de confianza
Convulsión	R56.9	Convulsiones no especificadas	0.8348
Convulsión	G40.909	Epilepsia, no especificada, no intratable, sin estado epiléptico	0.5424
Convulsión	56,00 R	Convulsiones febriles simples	0.4937
Convulsión	G40.09	Otras convulsiones	0.4397
Convulsión	G40.409	Otras epilepsias generalizadas y	0.4138

		síndromes epiléptic os, no intratables, sin estado epiléptico	
Dolor de cabeza	R51	Dolor de cabeza	0.4067
Dolor de cabeza	R51.9	Dolor de cabeza, sin especificar	0.3844
Dolor de cabeza	G 44.52	Nueva cefalea persistente diaria (NDPH)	0,3005
Dolor de cabeza	G44	Otro síndrome de cefalea	0.2670
Dolor de cabeza	G44.8	Otros síndromes de cefalea específicos	0,2542

Puede pasar los códigos ICD-10-CM al indicador para aumentar la precisión del LLM. Para reducir el ruido, puede filtrar los códigos ICD-10-CM utilizando la puntuación de confianza incluida en los resultados de Amazon Comprehend Medical. El siguiente es un ejemplo de mensaje que incluye únicamente los códigos ICD-10-CM que tienen una puntuación de confianza superior a 0,4:

```
<patient_note>
Carlie had a seizure 2 weeks ago. She is complaining of frequent headaches
Nausea is also present. She also complains of eye trouble with blurry vision
Meds : Topamax 50 mgs at breakfast daily,
Send referral order to neurologist
Follow-up as scheduled
</patient_note>

<comprehend_medical_results>
<icd-10>
  <entity>
    <text>seizure</text>
    <code>
      <description>Unspecified convulsions</description>
      <code_value>R56.9</code_value>
      <score>0.8347607851028442</score>
```

```
    </code>
    <code>
      <description>Epilepsy, unspecified, not intractable, without status epilepticus</
description>
      <code_value>G40.909</code_value>
      <score>0.542376697063446</score>
    </code>
    <code>
      <description>Other seizures</description>
      <code_value>G40.89</code_value>
      <score>0.4396627545356750 5</score>
    </code>
    <code>
      <description>Other generalized epilepsy and epileptic syndromes, not intractable,
 without status epilepticus</description>
      <code_value>G40.409</code_value>
      <score>0.41382506489753723</score>
    </code>
  </entity>
  <entity>
    <text>headaches</text>
    <code>
      <description>Headache</description>
      <code_value>R51</code_value>
      <score>0.4066613018512726</score>
    </code>
  </entity>
  <entity>
    <text>Nausea</text>
    <code>
      <description>Nausea</description>
      <code_value>R11.0</code_value>
      <score>0.6460834741592407</score>
    </code>
  </entity>
  <entity>
    <text>eye trouble</text>
    <code>
      <description>Unspecified disorder of eye and adnexa</description>
      <code_value>H57.9</code_value>
      <score>0.6780954599380493</score>
    </code>
    <code>
```

```
      <description>Unspecified visual disturbance</description>
      <code_value>H53.9</code_value>
      <score>0.5871203541755676</score>
    </code>
    <code>
      <description>Unspecified disorder of binocular vision</description>
      <code_value>H53.30</code_value>
      <score>0.5539672374725342</score>
    </code>
  </entity>
  <entity>
    <text>blurry vision</text>
    <code>
      <description>Other visual disturbances</description>
      <code_value>H53.8</code_value>
      <score>0.9001834392547607</score>
    </code>
  </entity>
</icd-10>
</comprehend_medical_results>

<prompt>
Given the patient note and Amazon Comprehend Medical ICD-10-CM code results above,
 please select the most relevant ICD-10-CM diagnosis codes for the patient.
For each selected code, provide a brief explanation of why it is relevant based on the
 information in the patient note.
</prompt>
```

Amplíe las tareas de PNL médica con Amazon Comprehend Medical

Al procesar textos médicos, el contexto de Amazon Comprehend Medical puede ayudar al LLM a seleccionar mejores fichas. En este ejemplo, desea hacer coincidir los síntomas del diagnóstico con los medicamentos. También querrá buscar texto relacionado con las pruebas médicas, como los términos relacionados con una prueba de análisis de sangre. Puede utilizar Amazon Comprehend Medical para detectar las entidades y los nombres de los medicamentos. En este caso, utilizaría la DetectEntitiesV2 y InferRxNorm APIs Amazon Comprehend Medical.

El siguiente es un ejemplo de nota para un paciente:

```
Carlie had a seizure 2 weeks ago. She is complaining of increased frequent headaches
Given lyme disease symptoms such as muscle ache and stiff neck will order prescription.
Meds : Topamax 50 mgs at breakfast daily. Amoxicillan 25 mg by mouth twice a day
```

```
Place MRI radiology order at RadNet
```

Para centrarse en el código de diagnóstico, en la solicitud solo se `DX_NAME` utilizan las entidades relacionadas `MEDICAL_CONDITION` con el tipo. Se excluyen otros metadatos debido a su irrelevancia. En el caso de las entidades medicamentosas, se incluye el nombre del medicamento junto con los atributos extraídos. Se excluyen otros metadatos de entidades farmacéuticas de Amazon Comprehend Medical por irrelevancia. El siguiente es un ejemplo de mensaje que utiliza los resultados filtrados de Amazon Comprehend Medical. El mensaje se centra en `MEDICAL_CONDITION` las entidades que tienen ese `DX_NAME` tipo. Este mensaje está diseñado para vincular con mayor precisión los códigos de diagnóstico con los medicamentos y extraer con mayor precisión las pruebas solicitadas por los médicos:

```
<patient_note>
Carlie had a seizure 2 weeks ago. She is complaining of increased freqeunt headaches
Given lyme disease symptoms such as muscle ache and stiff neck will order
 prescription.
Meds : Topamax 50 mgs at breakfast daily. Amoxicillan 25 mg by mouth twice a day
Place MRI radiology order at RadNet
</patient_note>

<detect_entity_results>
<entity>
    <text>seizure</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>headaches</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>lyme disease</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>muscle ache</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
```

```
<entity>
    <text>stiff neck</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
</detect_entity_results>

<rx_results>
<entity>
    <text>Topamax</text>
    <category>MEDICATION</category>
    <type>BRAND_NAME</type>
    <attributes>
        <attribute>
            <type>FREQUENCY</type>
            <text>at breakfast daily</text>
        </attribute>
        <attribute>
            <type>DOSAGE</type>
            <text>50 mgs</text>
        </attribute>
        <attribute>
            <type>ROUTE_OR_MODE</type>
            <text>by mouth</text>
        </attribute>
    </attributes>
</entity>
<entity>
    <text>Amoxicillan</text>
    <category>MEDICATION</category>
    <type>GENERIC_NAME</type>
    <attributes>
        <attribute>
            <type>ROUTE_OR_MODE</type>
            <text>by mouth</text>
        </attribute>
        <attribute>
            <type>DOSAGE</type>
            <text>25 mg</text>
        </attribute>
        <attribute>
            <type>FREQUENCY</type>
            <text>twice a day</text>
```

```
        </attribute>
    </attributes>
</entity>
</rx_results>

<prompt>
Based on the patient note and the detected entities, can you please:
1. Link the diagnosis symptoms with the medications prescribed.
Provide your reasoning for the linkages.
2. Extract any entities related to medical order tests mentioned in the note.
</prompt>
```

Aplique barandas con Amazon Comprehend Medical

Puede utilizar un LLM y Amazon Comprehend Medical para crear barandas antes de utilizar la respuesta generada. Puede ejecutar este flujo de trabajo en textos médicos no modificados o posprocesados. Los casos de uso incluyen abordar la información de salud protegida (PHI), detectar alucinaciones o implementar políticas personalizadas para publicar los resultados. Por ejemplo, puede utilizar el contexto de Amazon Comprehend Medical para identificar los datos de la PHI y, a continuación, utilizar el LLM para eliminar esos datos de la PHI.

El siguiente es un ejemplo de información de la historia clínica de un paciente que incluye la PHI:

```
Patient name: John Doe
Patient SSN: 123-34-5678
Patient DOB: 01/01/2024
Patient address: 123 Main St, Anytown USA
Exam details: good health. Pulse is 60 bpm. needs to work on diet with BMI of 190
```

El siguiente es un ejemplo de mensaje que incluye los resultados de Amazon Comprehend Medical como contexto:

```
<original_text>
Patient name: John Doe
Patient SSN: 123-34-5678 Patient DOB: 01/01/2024
Patient address: 123 Main St, Anytown USA
Exam details: good health. Pulse is 60 bpm. needs to work on diet with BMI of 190
</original_text>

<comprehend_medical_phi_entities>
<entity>
```

```
  <text>John Doe</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9967944025993347</score>
  <type>NAME</type>
</entity>
<entity>
  <text>123-34-5678</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9998034834861755</score>
  <type>ID</type>
</entity>
<entity>
  <text>01/01/2000</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9964448809623718</score>
  <type>DATE</type>
</entity>
</comprehend_medical_phi_entities>

<instructions>
Using the provided original text and the Amazon Comprehend Medical PHI entities
 detected, please analyze the text to determine if it contains any additional protected
 health information (PHI) beyond the entities already identified. If additional PHI is
 found, please list and categorize it. If no additional PHI is found, please state that
 explicitly.
In addition if PHI is found, generate updated text with the PHI removed.
</instructions>
```

Uso de modelos de lenguaje extensos para casos de uso de la salud y las ciencias de la vida

Aquí se describe cómo puede utilizar modelos de lenguaje de gran tamaño (LLMs) para aplicaciones sanitarias y de ciencias de la vida. Algunos casos de uso requieren el uso de un modelo de lenguaje amplio para las capacidades generativas de IA. Existen ventajas y limitaciones incluso para la mayoría state-of-the-art LLMs, y las recomendaciones de esta sección están diseñadas para ayudarlo a lograr los resultados esperados.

Puede utilizar el proceso de toma de decisiones para determinar la solución LLM adecuada para su caso de uso, teniendo en cuenta factores como el conocimiento del dominio y los datos de formación disponibles. Además, en esta sección se analizan las mejores prácticas médicas populares LLMs y

previamente capacitadas para su selección y uso. También se analizan las ventajas y desventajas entre soluciones complejas y de alto rendimiento y enfoques más simples y de menor costo.

Casos de uso de un LLM

Amazon Comprehend Medical puede realizar tareas específicas de PNL. Para obtener más información, consulte Casos de uso de Amazon Comprehend Medical.

Las capacidades de IA lógica y generativa de un LLM pueden ser necesarias para los casos de uso avanzados de la sanidad y las ciencias de la vida, como los siguientes:

- Clasificar entidades médicas personalizadas o categorías de texto
- Responder a preguntas clínicas
- Resumir los informes médicos
- Generar y detectar información a partir de información médica

Enfoques de personalización

Es fundamental entender cómo LLMs se implementan. LLMs por lo general, se entrenan con miles de millones de parámetros, incluidos datos de entrenamiento de muchos dominios. Esta formación permite al LLM abordar las tareas más generalizadas. Sin embargo, a menudo surgen desafíos cuando se requieren conocimientos específicos de un dominio. Los códigos clínicos, la terminología médica y la información de salud que se requieren para generar respuestas precisas son algunos ejemplos de conocimientos especializados en el ámbito de la salud y las ciencias de la vida. Por lo tanto, utilizar el LLM tal cual (sin necesidad de complementar el conocimiento del dominio) para estos casos de uso probablemente arroje resultados inexactos. Existen varios enfoques populares que puede utilizar para superar este desafío: la ingeniería rápida, la generación aumentada de recuperación (RAG) y el ajuste preciso.

Ingeniería de peticiones

La ingeniería rápida es el proceso en el que se guían las soluciones de IA generativa para crear los resultados deseados ajustando las entradas al LLM. Al elaborar indicaciones precisas con un contexto relevante, es posible guiar el modelo hacia la realización de tareas sanitarias especializadas que requieren razonamiento. Una ingeniería rápida y eficaz puede mejorar considerablemente el rendimiento del modelo para los casos de uso del sector sanitario sin necesidad de modificar el

modelo. Para obtener más información sobre la ingeniería rápida, consulte Implementación de la ingeniería rápida avanzada con Amazon Bedrock (entrada AWS del blog). En la ingeniería rápida se pueden utilizar técnicas de chain-of-thought creación rápida de imágenes y la generación de solicitudes.

Peticiones con pocos pasos

Las indicaciones con pocas tomas son una técnica en la que se proporcionan al LLM algunos ejemplos de la entrada/salida deseada antes de pedirle que realice una tarea similar. En contextos de atención médica, este enfoque es particularmente valioso para tareas especializadas, como el reconocimiento de entidades médicas o el resumen de notas clínicas. Al incluir de 3 a 5 ejemplos de alta calidad en su solicitud, puede mejorar significativamente la comprensión del modelo de la terminología médica y los patrones de dominios específicos. Para ver un ejemplo de indicaciones de pocas tomas, consulte Ingeniería y ajuste preciso de señales de pocas tomas en LLMs Amazon Bedrock (entrada del blog).AWS

Por ejemplo, al extraer las dosis de los medicamentos de las notas clínicas, puede proporcionar ejemplos de diferentes estilos de notación que ayuden al modelo a reconocer las variaciones en la forma en que los profesionales de la salud documentan las recetas. Este enfoque es especialmente eficaz cuando se trabaja con formatos de documentación estandarizados o cuando existen patrones consistentes en los datos.

Chain-of-thought pidiéndole

Chain-of-thought Las indicaciones (CoT) guían al LLM a través de un step-by-step proceso de razonamiento. Esto lo hace valioso para tareas complejas de apoyo a las decisiones médicas y de razonamiento diagnóstico. Al indicar explícitamente al modelo que «piense paso a paso» al analizar los escenarios clínicos, puede mejorar su capacidad para seguir los protocolos de razonamiento médico y reducir los errores de diagnóstico.

Esta técnica es excelente cuando el razonamiento clínico requiere varios pasos lógicos, como el diagnóstico diferencial o la planificación del tratamiento. Sin embargo, este enfoque tiene limitaciones cuando se trata de conocimientos médicos altamente especializados ajenos a los datos de formación del modelo o cuando se requiere una precisión absoluta para tomar decisiones de cuidados intensivos.

En estos casos, la combinación de la CoT con otro enfoque puede producir mejores resultados. Una opción es combinar el CoT con un sistema de orientación autocoherente. Para obtener más información, consulte Mejorar el rendimiento de los modelos de lenguaje generativo con mensajes

de autocoherencia en Amazon Bedrock (AWS entrada del blog). Otra opción es combinar los marcos de razonamiento, como las ReAct indicaciones, con el RAG. Para obtener más información, consulte Desarrollar asistentes de IA generativos avanzados basados en el chat mediante el uso de RAG y la generación de ReAct mensajes (Guía prescriptiva).AWS

Generación aumentada de recuperación

La generación aumentada de recuperación (RAG) es una tecnología de IA generativa en la que un LLM hace referencia a una fuente de datos autorizada que se encuentra fuera de sus fuentes de datos de entrenamiento antes de generar una respuesta. Un sistema RAG puede recuperar información ontológica médica (como las clasificaciones internacionales de enfermedades, los archivos nacionales de medicamentos y los epígrafes de temas médicos) a partir de una fuente de conocimiento. Esto proporciona un contexto adicional al LLM para respaldar la tarea de la PNL médica.

Como se explica en la Combinación de Amazon Comprehend Medical con modelos lingüísticos de gran tamaño sección, puede utilizar un enfoque RAG para recuperar el contexto de Amazon Comprehend Medical. Otras fuentes de conocimiento comunes incluyen los datos de dominio médico que se almacenan en un servicio de base de datos, como Amazon OpenSearch Service, Amazon Kendra o Amazon Aurora. La extracción de información de estas fuentes de conocimiento puede afectar al rendimiento de la recuperación, especialmente en el caso de consultas semánticas que utilizan una base de datos vectorial.

Otra opción para almacenar y recuperar información específica del dominio es utilizar Amazon Q Business en el flujo de trabajo de RAG. Amazon Q Business puede indexar repositorios de documentos internos o sitios web públicos (como CMS.gov para datos de la ICD-10). Amazon Q Business puede entonces extraer la información relevante de estas fuentes antes de pasar la consulta al LLM.

Existen varias formas de crear un flujo de trabajo de RAG personalizado. Por ejemplo, hay muchas formas de recuperar datos de una fuente de conocimiento. Para simplificar, recomendamos el enfoque de recuperación habitual que consiste en utilizar una base de datos vectorial, como Amazon OpenSearch Service, para almacenar el conocimiento como incrustaciones. Esto requiere que utilice un modelo de incrustación, como un transformador de oraciones, para generar incrustaciones para la consulta y para el conocimiento almacenado en la base de datos vectorial.

Para obtener más información sobre los enfoques de RAG totalmente gestionados y personalizados, consulte las opciones y arquitecturas de Retrieval Augmented Generation en. AWS

Ajuste

El ajuste de un modelo existente implica tomar un LLM, como un modelo Amazon Titan, Mistral o Llama, y luego adaptar el modelo a sus datos personalizados. Existen varias técnicas de ajuste, la mayoría de las cuales implican modificar solo unos pocos parámetros en lugar de modificar todos los parámetros del modelo. Esto se denomina ajuste fino con eficiencia de parámetros (PEFT). Para obtener más información, consulte Hugging Face GitHub PEFT en.

Los siguientes son dos casos de uso comunes en los que puedes optar por ajustar un LLM para una tarea de PNL médica:

- Tarea generativa: los modelos basados en decodificadores realizan tareas generativas de IA. AI/ML los profesionales utilizan datos basados en datos básicos para afinar un LLM existente. Por ejemplo, puede entrenar el LLM utilizando MedQuAD, un conjunto de datos públicos de preguntas y respuestas médicas. Cuando se invoca una consulta al LLM ajustado, no se necesita un enfoque RAG para proporcionar el contexto adicional al LLM.
- Incrustaciones: los modelos basados en codificadores generan incrustaciones al transformar el texto en vectores numéricos. Estos modelos basados en codificadores suelen denominarse modelos de incrustación. Un modelo de transformador de oraciones es un tipo específico de modelo de incrustación que está optimizado para oraciones. El objetivo es generar incrustaciones a partir del texto introducido. Las incrustaciones se utilizan luego para el análisis semántico o en tareas de recuperación. Para afinar el modelo de integración, es necesario disponer de un corpus de conocimientos médicos, como documentos, que pueda utilizar como datos de formación. Esto se logra con pares de texto basados en la similitud o el sentimiento para afinar un modelo de transformador de oraciones. Para obtener más información, consulte Entrenamiento y ajuste de modelos de incrustación con Sentence Transformers v3 en Hugging Face.

Puedes usar Amazon SageMaker Ground Truth para crear un conjunto de datos de entrenamiento etiquetado de alta calidad. Puede utilizar la salida del conjunto de datos etiquetados de Ground Truth para entrenar sus propios modelos. También puedes usar el resultado como un conjunto de datos de entrenamiento para un modelo de Amazon SageMaker AI. Para obtener más información sobre el reconocimiento de entidades nombradas, la clasificación del texto de una sola etiqueta y la clasificación del texto de varias etiquetas, consulte Etiquetado de texto con Ground Truth en la documentación de Amazon SageMaker AI.

Para obtener más información sobre el ajuste preciso, consulte esta guíaPerfeccionamiento de modelos lingüísticos de gran tamaño en el sector sanitario.

Cómo elegir un LLM

Amazon Bedrock es el punto de partida recomendado para evaluar el alto rendimiento LLMs. Para obtener más información, consulte Modelos de base compatibles en Amazon Bedrock. Puede utilizar los trabajos de evaluación de modelos en Amazon Bedrock para comparar los resultados de varios resultados y, a continuación, elegir el modelo que mejor se adapte a su caso de uso. Para obtener más información, consulte Elegir el modelo con mejor rendimiento mediante las evaluaciones de Amazon Bedrock en la documentación de Amazon Bedrock.

Algunos LLMs tienen una formación limitada sobre datos de dominio médico. Si su caso de uso requiere ajustar un LLM o un LLM que Amazon Bedrock no admite, considere la posibilidad de utilizar Amazon AI. SageMaker En el SageMaker caso de la IA, puede utilizar un LLM ajustado con precisión o elegir un LLM personalizado que se haya formado con datos del ámbito médico.

En la siguiente tabla se enumeran las personas más populares LLMs que se han formado con datos del dominio médico.

LLM	Tareas	Conocimiento	Arquitectura
BioBert	Recuperación de información, clasificación de textos y reconocimiento de entidades nombradas	Resúmenes PubMed, artículos de texto completo y conocimientos generales del PubMedCentral dominio	Codificador
Clínica Albert	Recuperación de información, clasificación de textos y reconocimiento de entidades nombradas	Amplio conjunto de datos multicéntrico, junto con más de 3 000 000 de historias clínicas de pacientes procedentes de sistemas de historiales médicos electrónicos (EHR)	Codificador

GPT clínico	Resumen, respuesta a preguntas y generación de texto	Conjuntos de datos médicos extensos y diversos, que incluyen registros médicos, conocimientos específicos del campo y consultas de diálogo de múltiples rondas	Decodificador
GatorTron-OG	Resumen, respuesta a preguntas, generación de texto y recuperación de información	Notas clínicas y literatura biomédica	Codificador
Med-Bert	Recuperación de información, clasificación de textos y reconocimiento de entidades nombradas	Amplio conjunto de datos de textos médicos, notas clínicas, trabajos de investigación y documentos relacionados con la asistencia sanitaria	Codificador
Med-Palm	Preguntas y respuestas con fines médicos	Conjuntos de datos de textos médicos y biomédicos	Decodificador
Medalla Paca	Tareas de respuesta a preguntas y diálogo médico	Una variedad de textos médicos, que incluyen recursos como tarjetas didácticas médicas, wikis y conjuntos de datos de diálogos	Decodificador

BioMedbert	Recuperación de información, clasificación de textos y reconocimiento de entidades nombradas	Exclusivamente resúmenes PubMed y artículos a texto completo de PubMedCentral	Codificador
BioMedLM	Resumen, respuesta a preguntas y generación de texto	Literatura biomédica a partir de fuentes de conocimiento PubMed	Decodificador

Las siguientes son las mejores prácticas para utilizar medicamentos previamente entrenados: LLMs

- Comprenda los datos de entrenamiento y su relevancia para su tarea médica de PNL.
- Identifique la arquitectura LLM y su propósito. Los codificadores son adecuados para las incrustaciones y las tareas de PNL. Los decodificadores son para tareas de generación.
- Evalúe los requisitos de infraestructura, rendimiento y costo para alojar el LLM médico previamente formado.
- Si es necesario realizar un ajuste preciso, asegúrese de que los datos de entrenamiento estén bien fundamentados o estén bien informados. Asegúrese de ocultar o borrar cualquier información de identificación personal (PII) o información de salud protegida (PHI).

Las tareas de la PNL médica en el mundo real pueden diferir de las previamente entrenadas LLMs en términos de conocimiento o casos de uso previstos. Si un LLM específico para un dominio específico no cumple con los parámetros de evaluación, puedes ajustar un LLM con tu propio conjunto de datos o puedes desarrollar un nuevo modelo básico. Formar un nuevo modelo básico es una tarea ambiciosa y, a menudo, costosa. Para la mayoría de los casos de uso, recomendamos ajustar un modelo existente.

Al utilizar o ajustar un LLM médico previamente entrenado, es importante tener en cuenta la infraestructura, la seguridad y las barreras.

Infraestructura

En comparación con el uso de Amazon Bedrock para la inferencia por lotes o bajo demanda, alojar LLM médicos previamente entrenados (generalmente de Hugging Face) requiere recursos significativos. Para alojar LLM médicos previamente entrenados, es habitual utilizar una imagen de

Amazon SageMaker AI que se ejecute en una instancia de Amazon Elastic Compute Cloud (Amazon EC2) con una o GPUs más instancias, como las instancias ml.g5 para la computación acelerada o las instancias ml.inf2 para. AWS Inferentia Esto se debe a que LLMs consumen una gran cantidad de memoria y espacio en disco.

Seguridad y barandas

Según los requisitos de conformidad de su empresa, considere la posibilidad de utilizar Amazon Comprehend y Amazon Comprehend Medical para ocultar o redactar la información de identificación personal (PII) y la información de salud protegida (PHI) de los datos de formación. Esto ayuda a evitar que el LLM utilice datos confidenciales al generar respuestas.

Te recomendamos que consideres y evalúes los prejuicios, la imparcialidad y las alucinaciones en tus aplicaciones de IA generativa. Ya sea que utilices un LLM preexistente o uno que estés perfeccionando, implementa barreras para evitar respuestas dañinas. Las barandillas son dispositivos de protección que puede personalizar para adaptarlos a los requisitos generativos de las aplicaciones de IA y a las políticas de IA responsables. Por ejemplo, puede usar Amazon Bedrock Guardrails.

Perfeccionamiento de modelos lingüísticos de gran tamaño en el sector sanitario

El enfoque de ajuste detallado que se describe en esta sección respalda el cumplimiento de las directrices éticas y reglamentarias y promueve el uso responsable de los sistemas de IA en el sector sanitario. Está diseñado para generar información precisa y privada. La IA generativa está revolucionando la prestación de servicios de salud, pero off-the-shelf los modelos suelen ser insuficientes en entornos clínicos en los que la precisión es fundamental y el cumplimiento no es negociable. El ajuste preciso de los modelos básicos con datos de dominios específicos colma esta brecha. Le ayuda a crear sistemas de IA que hablen el idioma de la medicina y, al mismo tiempo, cumplan con estrictos estándares reglamentarios. Sin embargo, el camino hacia el éxito de los ajustes requiere abordar con cuidado los desafíos únicos de la atención médica: proteger los datos confidenciales, justificar las inversiones en inteligencia artificial con resultados mensurables y mantener la relevancia clínica en un panorama médico en rápida evolución.

Cuando los enfoques más livianos alcanzan sus límites, los ajustes se convierten en una inversión estratégica. La expectativa es que las ganancias en precisión, latencia o eficiencia operativa compensen los importantes costos de computación e ingeniería necesarios. Es importante recordar

que el ritmo de progreso en los modelos básicos es rápido, por lo que la ventaja de un modelo ajustado podría durar solo hasta la próxima versión importante del modelo.

En esta sección, el análisis se centra en los siguientes dos casos de uso de gran impacto de clientes del sector sanitario: AWS

- Sistemas de apoyo a la toma de decisiones clínicas: mejore la precisión del diagnóstico mediante modelos que entiendan las historias clínicas complejas de los pacientes y las pautas en evolución. Los ajustes precisos pueden ayudar a los modelos a comprender en profundidad las historias clínicas complejas de los pacientes e integrar pautas especializadas, lo que podría reducir los errores de predicción del modelo. Sin embargo, es necesario sopesar estos beneficios con el coste de la formación sobre conjuntos de datos confidenciales de gran tamaño y con la infraestructura necesaria para las aplicaciones clínicas de alto nivel. ¿La mejora de la precisión y el conocimiento del contexto justificarán la inversión, especialmente cuando se lanzan nuevos modelos con frecuencia?
- Análisis de documentos médicos: automatice el procesamiento de las notas clínicas, los informes de imágenes y los documentos de seguro mientras mantiene el cumplimiento de la Ley de Portabilidad y Responsabilidad de los Seguros de Salud (HIPAA). En este caso, el ajuste preciso puede permitir que el modelo gestione formatos únicos, abreviaturas especializadas y requisitos reglamentarios de forma más eficaz. Los beneficios suelen traducirse en la reducción del tiempo de revisión manual y en la mejora del cumplimiento. Sin embargo, es esencial evaluar si estas mejoras son lo suficientemente sustanciales como para justificar los recursos de ajuste. Determine si la ingeniería rápida y la organización del flujo de trabajo pueden satisfacer sus necesidades.

Estos escenarios reales ilustran el proceso de ajuste, desde la experimentación inicial hasta la implementación del modelo, al tiempo que abordan los requisitos únicos de la atención médica en cada etapa.

Estimación de los costes y el retorno de la inversión

Los siguientes son los factores de costo que debe tener en cuenta al ajustar un LLM:

- Tamaño del modelo: el ajuste de los modelos más grandes cuesta más
- Tamaño del conjunto de datos: los costos y el tiempo de cómputo aumentan con el tamaño del conjunto de datos para su ajuste
- Estrategia de ajuste: los métodos eficientes en cuanto a los parámetros pueden reducir los costes en comparación con las actualizaciones completas de los parámetros

Al calcular el retorno de la inversión (ROI), tenga en cuenta la mejora de las métricas elegidas (como la precisión) multiplicada por el volumen de solicitudes (la frecuencia con la que se utilizará el modelo) y el tiempo esperado antes de que las versiones más recientes superen el modelo.

Además, tenga en cuenta la vida útil de su LLM base. Cada 6 a 12 meses aparecen nuevos modelos base. Si su detector de enfermedades raras tarda 8 meses en afinarse y validarse, es posible que solo obtenga 4 meses de rendimiento superior antes de que los modelos más nuevos cierren la brecha.

Al calcular los costes, el ROI y la vida útil potencial de su caso de uso, podrá tomar una decisión basada en datos. Por ejemplo, si ajustar el modelo de apoyo a las decisiones clínicas conduce a una reducción mensurable de los errores de diagnóstico en miles de casos al año, la inversión podría amortizarse rápidamente. Por el contrario, si la ingeniería rápida por sí sola hace que el flujo de trabajo de análisis de documentos se acerque a la precisión deseada, sería aconsejable posponer los ajustes hasta que llegue la próxima generación de modelos.

El ajuste fino no lo es. one-size-fits-all Si decide realizar ajustes, el enfoque correcto depende del caso de uso, los datos y los recursos.

Elegir una estrategia de ajuste

Una vez que haya determinado que el ajuste es el enfoque correcto para su caso de uso de la atención médica, el siguiente paso es seleccionar la estrategia de ajuste más adecuada. Hay varios enfoques disponibles. Cada uno tiene ventajas y desventajas distintas para las aplicaciones de atención médica. La elección entre estos métodos depende de sus objetivos específicos, de los datos disponibles y de las limitaciones de recursos.

Objetivos de formación

La formación previa adaptada a un dominio (DAPT) es un método no supervisado que implica la formación previa del modelo sobre una gran cantidad de texto no etiquetado y específico de un dominio específico (como millones de documentos médicos). Este enfoque es ideal para mejorar la capacidad de los modelos de entender las abreviaturas de las especialidades médicas y la terminología utilizada por radiólogos, neurólogos y otros proveedores especializados. Sin embargo, el DAPT requiere grandes cantidades de datos y no aborda resultados de tareas específicas.

El ajuste preciso supervisado (SFT) enseña al modelo a seguir instrucciones explícitas mediante ejemplos estructurados de entrada y salida. Este enfoque es excelente para los flujos de trabajo

de análisis de documentos médicos, como el resumen de documentos o la codificación clínica. El ajuste de instrucciones es una forma común de SFT en la que el modelo se entrena con ejemplos que incluyen instrucciones explícitas combinadas con los resultados deseados. Esto mejora la capacidad del modelo para comprender y seguir diversas indicaciones del usuario. Esta técnica es particularmente valiosa en los entornos de atención médica porque entrena al modelo con ejemplos clínicos específicos. El principal inconveniente es que requiere ejemplos cuidadosamente etiquetados. Además, el modelo ajustado puede tener problemas con casos extremos en los que no hay ejemplos. Para obtener instrucciones sobre cómo realizar ajustes con Amazon SageMaker Jumpstart, consulte Instrucciones de ajuste del FLAN T5 XL con Amazon SageMaker Jumpstart (entrada del blog).AWS

El aprendizaje por refuerzo a partir de la retroalimentación humana (RLHF) optimiza el comportamiento del modelo en función de los comentarios y preferencias de los expertos. Utilice un modelo de recompensas basado en las preferencias y métodos humanos, como la optimización proximal de políticas (PPO) o la optimización de preferencias directas (DPO), para optimizar el modelo y evitar actualizaciones destructivas. El RLHF es ideal para alinear los resultados con las directrices clínicas y garantizar que las recomendaciones se ajusten a los protocolos aprobados. Este enfoque requiere una cantidad considerable de tiempo para que los médicos envíen sus comentarios e implica una compleja cartera de formación. Sin embargo, la RLHF es particularmente valiosa en el sector de la salud porque ayuda a los expertos médicos a determinar la forma en que los sistemas de IA se comunican y hacen recomendaciones. Por ejemplo, los médicos pueden dar su opinión para asegurarse de que el modelo se adapta adecuadamente a los pacientes, sabe cuándo expresar su incertidumbre y se ajusta a las directrices clínicas. Técnicas como la PPO optimizan de forma iterativa el comportamiento del modelo en función de los comentarios de los expertos y, al mismo tiempo, limitan las actualizaciones de los parámetros para preservar los conocimientos médicos básicos. Esto permite que los modelos transmitan diagnósticos complejos en un lenguaje fácil de entender para el paciente y, al mismo tiempo, detectar enfermedades graves para su atención médica inmediata. Esto es crucial para la asistencia sanitaria, donde tanto la precisión como el estilo de comunicación son importantes. Para obtener más información sobre la RLHF, consulte Ajustar modelos lingüísticos de gran tamaño con el aprendizaje reforzado a partir de comentarios humanos o de la IA (AWS entrada del blog).

Métodos de implementación

Una actualización completa de los parámetros implica actualizar todos los parámetros del modelo durante el entrenamiento. Este enfoque funciona mejor para los sistemas de apoyo a la toma de decisiones clínicas que requieren una integración profunda de las historias clínicas de los pacientes,

los resultados de laboratorio y las pautas en evolución. Los inconvenientes incluyen el alto costo de cómputo y el riesgo de sobreajuste si el conjunto de datos no es grande y diverso.

Los métodos de ajuste preciso con eficiencia de parámetros (PEFT) actualizan solo un subconjunto de parámetros para evitar un sobreajuste o una pérdida catastrófica de las capacidades lingüísticas. Los tipos incluyen la adaptación de rango bajo (LoRa), los adaptadores y el ajuste de prefijos. Los métodos PEFT ofrecen un menor costo computacional, una capacitación más rápida y son excelentes para experimentos como la adaptación de un modelo de apoyo a la toma de decisiones clínicas a los protocolos o la terminología de un nuevo hospital. La principal limitación es la posible reducción del rendimiento en comparación con las actualizaciones completas de los parámetros.

Para obtener más información sobre los métodos de ajuste, consulte Métodos de ajuste avanzados en Amazon SageMaker AI (AWS entrada del blog).

Creación de un conjunto de datos de ajuste

La calidad y la diversidad del conjunto de datos de ajuste fino son fundamentales para el rendimiento del modelo, la seguridad y la prevención de sesgos. Las siguientes son tres áreas críticas que se deben tener en cuenta al crear este conjunto de datos:

- El volumen se basa en un enfoque de ajuste
- Anotación de datos de un experto en el campo
- Diversidad del conjunto de datos

Como se muestra en la siguiente tabla, los requisitos de tamaño del conjunto de datos para el ajuste fino varían según el tipo de ajuste fino que se realice.

Estrategia de ajuste	Tamaño del conjunto de datos
Capacitación previa adaptada al dominio	Más de 100 000 textos de dominio
Ajustes supervisados	Más de 10.000 pares etiquetados
Refuerzo del aprendizaje a partir de la retroalimentación humana	Más de 1000 pares de preferencias de expertos

Puede usar AWS GlueAmazon EMR y Amazon SageMaker Data Wrangler para automatizar el proceso de extracción y transformación de datos a fin de conservar un conjunto de datos de su propiedad. Si no puede conservar un conjunto de datos lo suficientemente grande, puede descubrirlos y descargarlos directamente en su propia cuenta. Cuenta de AWS AWS Data Exchange Consulte a su asesor legal antes de utilizar conjuntos de datos de terceros.

Los anotadores expertos con conocimientos en el campo, como médicos, biólogos y químicos, deberían formar parte del proceso de conservación de datos para incorporar los matices de los datos médicos y biológicos en el resultado del modelo. Amazon SageMaker Ground Truth proporciona una interfaz de usuario de bajo código para que los expertos anoten el conjunto de datos.

Un conjunto de datos que represente a la población humana es esencial para que la sanidad y las ciencias de la vida ajusten los casos de uso a fin de evitar sesgos y reflejar los resultados del mundo real. AWS Glue las sesiones interactivas o las instancias de Amazon SageMaker Notebook ofrecen una forma eficaz de explorar conjuntos de datos de forma iterativa y ajustar las transformaciones mediante el uso de cuadernos compatibles con Jupyter. Las sesiones interactivas le permiten trabajar con una variedad de entornos de desarrollo integrados populares () en su entorno local. IDEs Como alternativa, puedes trabajar con AWS Glue o con las libretas de Amazon SageMaker Studio a través del Consola de administración de AWS.

Ajustando el modelo

AWS ofrece servicios como Amazon SageMaker AI y Amazon Bedrock, que son cruciales para el éxito de los ajustes.

SageMaker La IA es un servicio de aprendizaje automático totalmente gestionado que ayuda a los desarrolladores y científicos de datos a crear, entrenar e implementar modelos de aprendizaje automático con rapidez. Entre las tres funciones útiles de la SageMaker IA para realizar ajustes, se incluyen las siguientes:

- SageMakerCapacitación: una función de aprendizaje automático totalmente gestionada que le ayuda a entrenar de manera eficiente una amplia gama de modelos a escala
- SageMaker JumpStart— Una capacidad que se basa en los trabajos de SageMaker formación para proporcionar modelos previamente entrenados, algoritmos integrados y plantillas de soluciones para las tareas de aprendizaje automático
- SageMaker HyperPod— Una solución de infraestructura diseñada específicamente para la formación distribuida de modelos básicos y LLMs

Amazon Bedrock es un servicio totalmente gestionado que proporciona acceso a modelos básicos de alto rendimiento a través de una API, con funciones integradas de seguridad, privacidad y escalabilidad. El servicio ofrece la capacidad de ajustar varios modelos fundamentales disponibles. Para obtener más información, consulte los modelos y regiones compatibles para obtener más información sobre los ajustes y la formación previa continua en la documentación de Amazon Bedrock.

Al abordar el proceso de ajuste con cualquiera de los dos servicios, tenga en cuenta el modelo base, la estrategia de ajuste y la infraestructura.

Elección del modelo base

Los modelos de código cerrado, como Anthropic Claude, Meta Llama y Amazon Nova, ofrecen un out-of-the-box rendimiento sólido con un cumplimiento gestionado, pero limitan la flexibilidad de ajuste a las opciones compatibles con los proveedores, como las gestionadas, como Amazon Bedrock. APIs Esto limita la capacidad de personalización, especialmente en los casos de uso de la sanidad regulados. Por el contrario, los modelos de código abierto, como Meta Llama, proporcionan control y flexibilidad totales en todos los servicios de SageMaker IA de Amazon, lo que los hace ideales cuando necesitas personalizar, auditar o adaptar en profundidad un modelo a tus requisitos específicos de datos o flujo de trabajo.

Estrategia de ajuste

El ajuste de instrucciones simples se puede realizar mediante la personalización del modelo Amazon Bedrock o Amazon SageMaker JumpStart. Los enfoques PEFT complejos, como LoRa o los adaptadores, requieren trabajos de SageMaker formación o una función de ajuste personalizada en Amazon Bedrock. Se admite la formación distribuida para modelos muy grandes. SageMaker HyperPod

Escalabilidad y control de la infraestructura

Los servicios totalmente gestionados, como Amazon Bedrock, minimizan la administración de la infraestructura y son ideales para las organizaciones que priorizan la facilidad de uso y el cumplimiento. Las opciones semigestionadas, por ejemplo SageMaker JumpStart, ofrecen cierta flexibilidad con menos complejidad. Estas opciones son adecuadas para la creación rápida de prototipos o cuando se utilizan flujos de trabajo prediseñados. Los trabajos de SageMaker formación vienen acompañados de un control y una personalización HyperPod totales, pero estos requieren más experiencia y son ideales cuando es necesario ampliarlos para grandes conjuntos de datos o si se requieren procesos personalizados.

Supervisión de modelos ajustados

En el sector de la salud y las ciencias de la vida, la supervisión de los ajustes del LLM requiere el seguimiento de varios indicadores clave de rendimiento. La precisión proporciona una medida de referencia, pero debe equilibrarse con la precisión y la memoria, especialmente en aplicaciones en las que las clasificaciones erróneas tienen consecuencias importantes. La puntuación F1 ayuda a abordar los problemas de desequilibrio de clases que pueden ser comunes en los conjuntos de datos médicos. Para obtener más información, consulte la sección Evaluación LLMs de aplicaciones sanitarias y de ciencias de la vida de esta guía.

Las métricas de calibración ayudan a garantizar que los niveles de confianza del modelo coincidan con las probabilidades del mundo real. Las métricas de imparcialidad pueden ayudarle a detectar posibles sesgos en diferentes grupos demográficos de pacientes.

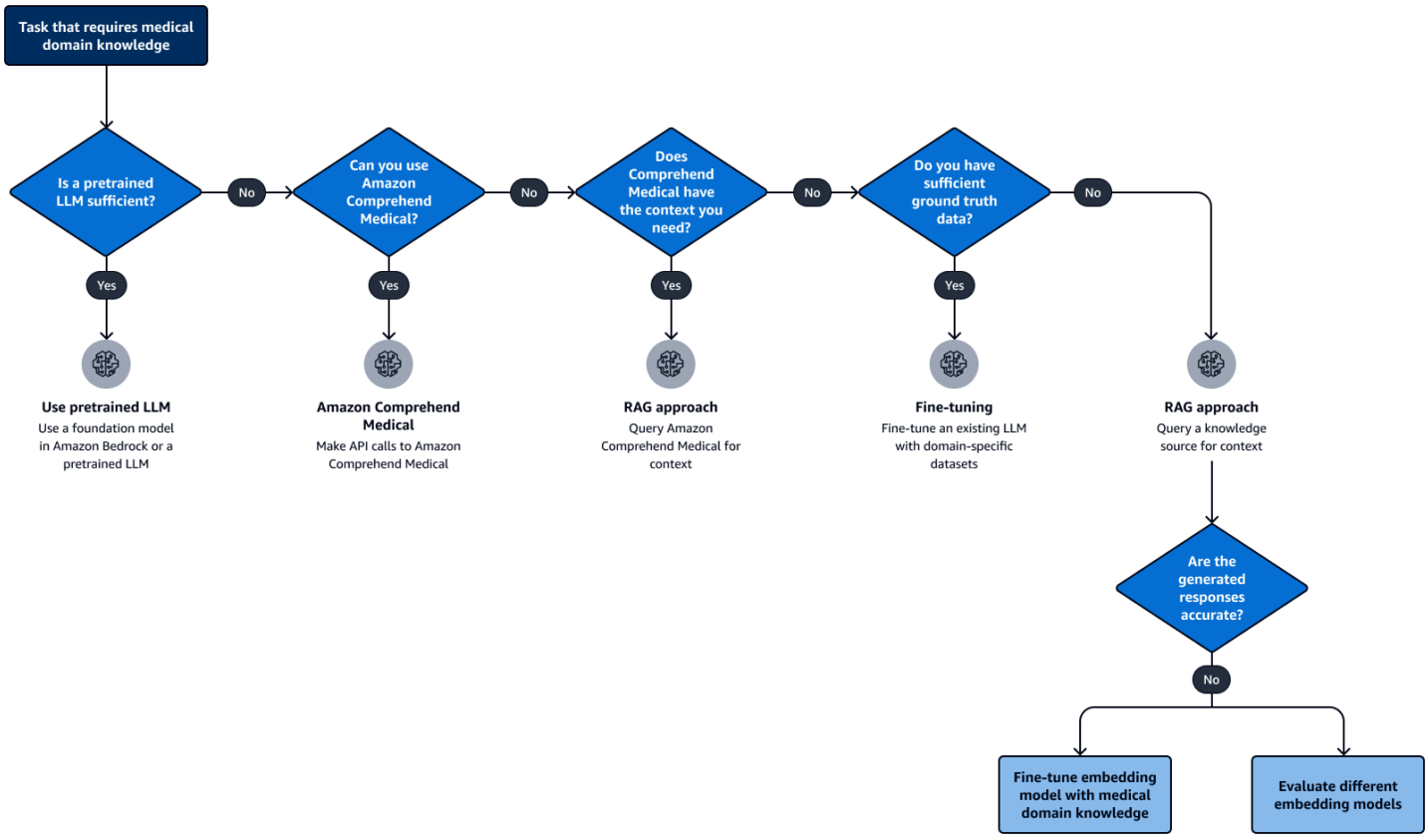
MLflowes una solución de código abierto que puede ayudarle a realizar un seguimiento de los experimentos de ajuste. MLflow es compatible de forma nativa con Amazon SageMaker AI, lo que le ayuda a comparar visualmente las métricas de las sesiones de entrenamiento. Para ajustar los trabajos en Amazon Bedrock, las métricas se transmiten a CloudWatch Amazon para que pueda visualizarlas en la consola. CloudWatch

Elegir un enfoque de PNL para la salud y las ciencias de la vida

Enfoques generativos de IA y PNL para la salud y las ciencias de la vidaEn la sección se describen los siguientes enfoques para abordar las tareas del procesamiento del lenguaje natural (PNL) en aplicaciones sanitarias y de ciencias de la vida:

- Uso de Amazon Comprehend Medical
- Combinación de Amazon Comprehend Medical con un máster en un flujo de trabajo de recuperación aumentada (RAG)
- Uso de un LLM ajustado
- Uso de un flujo de trabajo RAG

Al evaluar las limitaciones conocidas de LLMs las tareas del ámbito médico y su caso de uso, puede elegir el enfoque que mejor se adapte a su tarea. El siguiente árbol de decisiones puede ayudarte a elegir un enfoque de LLM para tu tarea de PNL médica:



En el diagrama, se muestra el siguiente flujo de trabajo:

1. Para los casos de uso de la salud y las ciencias de la vida, identifique si la tarea de la PNL requiere un conocimiento específico del dominio. Según sea necesario, coordine con expertos en la materia (SMEs).
2. Si puede utilizar un LLM general o un modelo que se haya formado con conjuntos de datos médicos, utilice un modelo básico disponible en Amazon Bedrock o el LLM previamente entrenado. Para obtener más información, consulte la sección Cómo elegir un LLM de esta guía.
3. Si las capacidades de detección de entidades y enlace ontológico de Amazon Comprehend Medical se adaptan a su caso de uso, utilice Amazon Comprehend Medical. APIs Para obtener más información, consulte la sección Uso de Amazon Comprehend Medical de esta guía.
4. A veces, Amazon Comprehend Medical tiene el contexto necesario, pero no es compatible con su caso de uso. Por ejemplo, es posible que necesite definiciones de entidades diferentes, reciba una cantidad abrumadora de resultados, necesite entidades personalizadas o necesite una tarea de PNL personalizada. Si este es el caso, utilice un enfoque RAG para consultar el contexto en Amazon Comprehend Medical. Para obtener más información, consulte la sección Combinación de Amazon Comprehend Medical con modelos lingüísticos de gran tamaño de esta guía.
5. Si dispone de una cantidad suficiente de datos fiables, ajuste un LLM existente. Para obtener más información, consulte la sección Enfoques de personalización de esta guía.
6. Si los otros enfoques no cumplen con los objetivos médicos de su tarea de PNL, implemente una solución RAG. Para obtener más información, consulte la sección Enfoques de personalización de esta guía.
7. Después de implementar la solución RAG, evalúe si las respuestas generadas son precisas. Para obtener más información, consulte la sección Evaluación LLMs de aplicaciones sanitarias y de ciencias de la vida de esta guía. Es habitual empezar con un modelo de incrustaciones de texto de Amazon Titan o un modelo de transformador de oraciones general, como el modelo All-MiniLM-L6-V2. Sin embargo, debido a la falta de contexto del dominio, es posible que estos modelos no capten la terminología médica del texto. Si es necesario, considere los siguientes ajustes:
   a. Evalúe otros modelos de incrustación
   b. Ajuste el modelo de incrustación con conjuntos de datos específicos del dominio

Consideraciones sobre la madurez empresarial

La madurez empresarial es fundamental a la hora de adaptar las soluciones LLM para aplicaciones de sanidad y ciencias de la vida. Estas organizaciones se enfrentan a distintos niveles de complejidad a la hora de LLMs implementarlas, según sus criterios de aceptación. Con frecuencia, las organizaciones que carecen de AI/ML recursos invierten en el apoyo de los contratistas para crear soluciones LLM. En estas situaciones, es importante entender las siguientes desventajas:

- Alto rendimiento a un coste y un mantenimiento elevados: es posible que necesite una solución compleja que requiera ajustes precisos o personalizados LLMs para cumplir con los estrictos estándares de rendimiento. Sin embargo, esto conlleva mayores costes y requisitos de mantenimiento. Es posible que necesite contratar recursos especializados o asociarse con contratistas para mantener estas sofisticadas soluciones. Esto puede retrasar el desarrollo.
- Buen rendimiento a un bajo coste y mantenimiento: también puede que descubras que servicios como Amazon Bedrock o Amazon Comprehend Medical ofrecen un rendimiento aceptable. Si bien estos LLMs enfoques pueden proporcionar resultados perfectos, estas soluciones suelen ofrecer resultados consistentes y de alta calidad. Estas soluciones son de menor costo y reducen la carga de mantenimiento. Esto puede acelerar el desarrollo.

Si un enfoque más simple y de menor costo ofrece de manera consistente resultados de alta calidad que cumplen con sus criterios de aceptación, considere si el aumento del rendimiento compensa el costo, el mantenimiento y el tiempo. Sin embargo, si la solución más sencilla está muy por debajo del rendimiento objetivo y si su organización carece de la capacidad de inversión necesaria para soluciones complejas y sus requisitos de mantenimiento, considere la posibilidad de posponer el AI/ML desarrollo hasta que haya más recursos o soluciones alternativas disponibles.

Además, para cualquier solución de PNL médica que se base en un LLM, le recomendamos que realice una supervisión y una evaluación continuas. Evalúe los comentarios de los usuarios a lo largo del tiempo e implemente evaluaciones periódicas para asegurarse de que la solución sigue cumpliendo sus objetivos empresariales.

Evaluación LLMs de aplicaciones sanitarias y de ciencias de la vida

En esta sección se ofrece una visión general completa de los requisitos y las consideraciones para evaluar modelos lingüísticos extensos (LLMs) en los casos de uso de la sanidad y las ciencias de la vida.

Es importante utilizar datos básicos y comentarios de las pymes para mitigar los sesgos y validar la precisión de la respuesta generada por la LLM. En esta sección se describen las mejores prácticas para recopilar y conservar los datos de formación y pruebas. También le ayuda a implementar barreras y a medir el sesgo y la imparcialidad de los datos. También analiza las tareas médicas más comunes del procesamiento del lenguaje natural (PNL), como la clasificación de textos, el reconocimiento de entidades nombradas y la generación de textos, y las métricas de evaluación asociadas a ellas.

También presenta los flujos de trabajo para realizar la evaluación del LLM durante la fase de experimentación de la formación y la fase de posproducción. El monitoreo de modelos y las operaciones de LLM son elementos importantes de este proceso de evaluación.

Datos de entrenamiento y pruebas para tareas de PNL médica

Las tareas de la PNL médica suelen utilizar corpus médicos (por ejemplo PubMed) o información del paciente (como las notas sobre las visitas de los pacientes a la clínica) para clasificar, resumir y generar información. El personal médico, como los médicos, los administradores de atención médica o los técnicos, varía en cuanto a experiencia y puntos de vista. Debido a la subjetividad entre este personal médico, los conjuntos de datos de formación y pruebas más pequeños representan un riesgo de sesgo. Para mitigar este riesgo, recomendamos las siguientes prácticas recomendadas:

- Cuando utilice una solución LLM previamente entrenada, asegúrese de disponer de una cantidad adecuada de datos de prueba. Los datos de las pruebas deben parecerse mucho a los datos médicos reales. Según la tarea, puede oscilar entre 20 y más de 100 registros.
- Al afinar un LLM, recopile un número suficiente de registros etiquetados (veraces) de diversos ámbitos SMEs de la medicina objetivo. Un punto de partida general son al menos 100 registros de alta calidad. Sin embargo, dada la complejidad de la tarea y sus criterios de aceptación de la precisión, es posible que se necesiten más registros.

- Si es necesario para su caso de uso médico, implemente barreras y mida el sesgo y la imparcialidad de los datos. Por ejemplo, asegúrese de que el LLM evite los diagnósticos erróneos debidos a los perfiles raciales de los pacientes. Para más información, consulte la sección Seguridad y barandas de esta guía.

Muchas empresas de investigación y desarrollo de la IA, como Anthropic, ya han incorporado barreras en sus modelos básicos para evitar la toxicidad. Puede utilizar la detección de toxicidad para comprobar las indicaciones de entrada y las respuestas de salida. LLMs Para obtener más información, consulte Detección de toxicidad en la documentación de Amazon Comprehend y consulte Guardrails en la documentación de Amazon Bedrock.

En cualquier tarea de IA generativa, existe el riesgo de alucinaciones. Puedes mitigar este riesgo realizando tareas de PNL, como la clasificación. También puede utilizar técnicas más avanzadas, como las métricas de similitud de texto. BertScorees una métrica de similitud de texto que se utiliza habitualmente. Para obtener más información sobre las técnicas que puede utilizar para mitigar las alucinaciones, consulte Una encuesta exhaustiva sobre las técnicas de mitigación de las alucinaciones en modelos lingüísticos extensos.

Métricas para las tareas médicas de PNL

Puede crear métricas cuantificables después de establecer los datos básicos y las etiquetas proporcionadas por las pymes para la formación y las pruebas. Comprobar la calidad mediante procesos cualitativos, como las pruebas de stress y la revisión de los resultados del LLM, es útil para un desarrollo rápido. Sin embargo, las métricas actúan como puntos de referencia cuantitativos que respaldan las futuras operaciones de LLM y actúan como puntos de referencia de rendimiento para cada versión de producción.

Entender la tarea médica es fundamental. Por lo general, las métricas se asignan a una de las siguientes tareas generales de la PNL:

- Clasificación del texto: el LLM clasifica el texto en una o más categorías predefinidas, según la solicitud de entrada y el contexto proporcionado. Un ejemplo es clasificar una categoría de dolor mediante una escala de dolor. Algunos ejemplos de métricas de clasificación de textos son:
  - Precisión
  - Precisión, también conocida como macroprecisión
  - Recuperación, también conocida como recuperación macroscópica

  - Puntuación F1, también conocida como puntuación macro F1
  - ¡Pérdida de Hamming
- Reconocimiento de entidades con nombre (NER): también conocido como extracción de texto, el reconocimiento de entidades con nombre es el proceso de localizar y clasificar las entidades nombradas que se mencionan en un texto no estructurado en categorías predefinidas. Un ejemplo es extraer los nombres de los medicamentos de los registros de los pacientes. Algunos ejemplos de métricas de NER incluyen:
  - Precisión
  - Precisión
  - Exhaustividad
  - Puntuación de F1
  - Pérdida de Hamming
- Generación: el LLM genera texto nuevo procesando el mensaje y el contexto proporcionado. La generación incluye tareas de resumen o tareas de respuesta a preguntas. Algunos ejemplos de métricas de generación son:
  - Suplente orientado al retiro del mercado para la evaluación de Gisting (ROUGE)
  - Métrica para la evaluación de la traducción con Explicit (METEOR) ORdering
  - Estudiante de evaluación bilingüe (BLEU) (para traducciones)
  - Distancia entre cuerdas, también conocida como similitud de coseno

Preguntas frecuentes sobre casos de uso de la sanidad y las ciencias de la vida

Las siguientes son preguntas frecuentes relacionadas con el uso de Amazon Comprehend Medical LLMs o para tareas de PNL médicas.

¿Cómo elijo entre Amazon Comprehend Medical y un LLM?

Si su tarea consiste en detectar entidades médicas en su texto médico, consulte la documentación de Amazon Comprehend Medical para saber qué entidades médicas se pueden extraer y si alguna de las ontologías aborda su caso de uso. Si no es así, considere la posibilidad de utilizar un LLM. Para obtener más información, consulte Casos de uso de Amazon Comprehend Medical y Casos de uso de un LLM en esta guía.

¿Cómo puedo proporcionar los resultados de Amazon Comprehend Medical a un LLM?

Puede incorporar los resultados de Amazon Comprehend Medical como contexto en sus solicitudes de LLM. Esto proporciona conocimientos y terminología médicos adicionales al LLM. El contexto proporcionado puede mejorar el desempeño del LLM en tareas como el reconocimiento de entidades, el resumen o la respuesta a preguntas. La guía proporciona varios ejemplos de cómo estructurar las indicaciones con los resultados de Amazon Comprehend Medical. Para obtener más información, consulte la sección Combinación de Amazon Comprehend Medical con modelos lingüísticos de gran tamaño de esta guía.

¿Cuáles son algunas de las mejores prácticas a la hora de utilizar Amazon Comprehend Medical LLMs con?

Te recomendamos que utilices las puntuaciones de confianza de Amazon Comprehend Medical para filtrar o priorizar las entidades según tus indicaciones. También es importante evaluar su rendimiento en función de sus datos específicos y validar que las definiciones de las entidades se ajusten a sus requisitos. La combinación de Amazon Comprehend Medical con fuentes de conocimiento específicas del dominio puede mejorar aún más el rendimiento del LLM. Para obtener

más información, consulte la sección Mejores prácticas para usar Amazon Comprehend Medical en un flujo de trabajo de RAG de esta guía.

¿Debo utilizar un LLM médico previamente formado o ajustar un LLM general para mi caso de uso en el sector de la salud?

La decisión depende de sus requisitos específicos y de la disponibilidad de datos de formación de alta calidad. Un médico previamente formado LLMs puede ser un buen punto de partida. Sin embargo, es posible que aún tengas que ajustarlos con los datos específicos de tu dominio. Si tienes suficientes datos etiquetados, ajustar un LLM general puede ser una opción viable. Para obtener más información, consulte Cómo elegir un LLM y Elegir un enfoque de PNL para la salud y las ciencias de la vida en esta guía.

¿Cómo puedo evaluar el desempeño de las tareas médicas LLMs de PNL?

Recomendamos utilizar métricas cuantitativas, como la exactitud, la precisión, la recuperación y la puntuación F1 para las tareas de clasificación de textos y reconocimiento de entidades nombradas. Puede utilizar ROUGE y METEOR para tareas de generación de texto. Es importante contar con datos fiables sobre el terreno etiquetados por expertos en la materia e implementar procesos para monitorear el rendimiento de los modelos a lo largo del tiempo. Para obtener más información, consulte la sección Evaluación LLMs de aplicaciones sanitarias y de ciencias de la vida de esta guía.

¿Cuáles son las ventajas y desventajas entre las soluciones LLM de alta y baja complejidad?

Ajustar un LLM o crear un LLM personalizado son soluciones muy complejas. Estos enfoques pueden mejorar el rendimiento, pero conllevan costos y requisitos de mantenimiento más altos. Las soluciones más sencillas, como el uso de Amazon Comprehend Medical LLMs o de Amazon Comprehend Medical, pueden ofrecer un rendimiento aceptable con costes más bajos y ciclos de desarrollo más rápidos. Sin embargo, es posible que estos enfoques no cumplan con los estrictos requisitos de precisión en algunos casos de uso. Para obtener más información, consulte la sección Consideraciones sobre la madurez empresarial de esta guía.

Próximos pasos y recursos

Esta guía le ayuda Servicios de AWS a automatizar la PNL médica y las tareas de IA generativa para aplicaciones del mundo real en entornos de producción. Describe cómo puede utilizar Amazon Comprehend Medical, LLMs con el respaldo de Amazon Bedrock, con formación LLMs médica previa o con LLMs ajustes precisos para lograr sus objetivos empresariales de salud y ciencias de la vida. En esta guía se describen las ventajas y limitaciones de los siguientes enfoques:

- Uso de Amazon Comprehend Medical de forma independiente
- Proporcionar los resultados de Amazon Comprehend Medical a un LLM
- Utilizar un LLM general previamente entrenado o un LLM médico en un enfoque de generación aumentada de recuperación (RAG)
- Perfeccionar un LLM general o un LLM médico

Utilice el árbol de decisiones y las consideraciones sobre la madurez empresarial de esta guía para elegir entre estos enfoques en función del nivel de madurez de su organización. AI/ML Aunque Amazon Comprehend Medical y Amazon LLMs Bedrock ofrecen potentes capacidades, solo tienen éxito si se implementan y evalúan adecuadamente. Utilice la información y las métricas de evaluación que se describen en esta guía para validar el rendimiento de la solución.

Para los próximos pasos, recomendamos que los administradores de TI, los arquitectos y los líderes técnicos del sector sanitario colaboren con AI/ML los profesionales para identificar su tarea médica relacionada con la PNL. Utilice esta guía para elegir una ruta de desarrollo y, a continuación, utilice las funciones Servicios de AWS y funciones adecuadas para implementar con éxito una solución automatizada. AWS

AWS recursos

- Documentación médica de Amazon Comprehend Medical:
  - Guía para desarrolladores
  - Referencia de la API
- Documentación de Amazon Bedrock
  - Evaluación del modelo Amazon Bedrock
  - Perfeccionamiento en Amazon Bedrock

- Ajuste un modelo en Amazon AI SageMaker
- Amazon SageMaker Ground Truth
- Amazon Comprehend: detección de toxicidad
- AWS Socios con competencias sanitarias

Otros recursos de

- Tabla de clasificación de Open Medical-LLM
- Una encuesta sobre modelos lingüísticos extensos para el cuidado de la salud: desde los datos, la tecnología y las aplicaciones hasta la responsabilidad y la ética
- Los modelos lingüísticos extensos son malos codificadores médicos: comparación de la consulta de códigos médicos
- De principiante a experto: modelar el conocimiento médico en general LLMs

Colaboradores

Creación

- Joe King, científico de datos AWS sénior
- Ankith Ede, arquitecto de soluciones AWS
- Clement Perrot, estratega sénior de IA generativa AWS
- Jillian Forde, arquitecta sénior de soluciones AWS
- Rajesh Sitaraman, consultor sénior de entregas AWS
- Ross Claytor, científico aplicado principal AWS
- Shivesh Ummat, arquitecto de soluciones AWS

Revisión

- Dilshad Raihan Akkam Veettil, científico de datos sénior AWS
- Joseph Cottingham, AWS arquitecto de aprendizaje profundo

Redacción técnica

- Lilly AbouHarb, escritora técnica AWS sénior

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
Nuevas secciones	Hemos añadido la sección sobre cómo ajustar los modelos lingüísticos de gran tamaño en el sector sanitario y la sección de ingeniería rápida.	5 de diciembre de 2025
Publicación inicial	—	16 de diciembre de 2024

AWS Glosario de orientación prescriptiva

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

Números

Las 7 R

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

- 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.

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.