Recupere las opciones y arquitecturas de generación aumentada en AWS

AWS Guía prescriptiva

AWS Guía prescriptiva: Recupere las opciones y arquitecturas de generación aumentada en AWS

Table of Contents

Recuperación de opciones y arquitecturas de generación aumentada en AWS

Mithil Shah, Rajeev Muralidhar y Natacha Fort, Amazon Web Services

Octubre de 2024 (historia del documento)

La IA generativa se refiere a un subconjunto de modelos de IA que pueden crear nuevos contenidos y artefactos, como imágenes, vídeos, texto y audio, a partir de un simple mensaje de texto. Los modelos de IA generativa se entrenan con grandes cantidades de datos que abarcan una amplia gama de temas y tareas. Esto les permite demostrar una notable versatilidad a la hora de realizar diversas tareas, incluso aquellas para las que no han recibido formación explícita. Debido a la capacidad de un solo modelo para realizar múltiples tareas, estos modelos suelen denominarse modelos básicos (FMs).

Una de las aplicaciones más notables de los modelos de IA generativa es su habilidad para responder preguntas. Sin embargo, existen desafíos específicos que surgen cuando estos modelos se utilizan para responder preguntas basadas en documentos personalizados. Los documentos personalizados pueden incluir información confidencial, sitios web internos, documentación interna, Confluence SharePoint páginas, páginas y otros. Una opción es utilizar la generación aumentada de recuperación (RAG). Con el RAG, el modelo básico hace referencia a una fuente de datos autorizada que se encuentra fuera de sus fuentes de datos de entrenamiento (como los documentos personalizados) antes de generar una respuesta.

Esta guía describe las distintas opciones de IA generativa disponibles para responder a las preguntas de la documentación personalizada, incluidos los sistemas de recuperación y generación aumentada (RAG). También proporciona información general sobre la creación de sistemas RAG en Amazon Web Services (AWS). Al revisar las opciones y arquitecturas de RAG, puede elegir entre servicios totalmente gestionados o arquitecturas RAG personalizadas. AWS

Destinatarios previstos

Los destinatarios de esta guía son arquitectos y administradores de IA generativa que desean crear una solución de RAG, revisar las arquitecturas disponibles y comprender las ventajas y desventajas de cada opción.

Objetivos

Esta guía lo ayuda a hacer lo siguiente:

- Conozca las opciones de IA generativa disponibles para responder a las preguntas de los documentos personalizados
- Revise las opciones de arquitectura de los sistemas RAG en AWS
- Comprenda las ventajas y desventajas de cada opción de RAG
- Elija una arquitectura RAG para su entorno AWS

Opciones de IA generativa para consultar documentos personalizados

Las organizaciones suelen tener varias fuentes de datos estructurados y no estructurados. Esta guía se centra en cómo utilizar la IA generativa para responder preguntas a partir de datos no estructurados.

Los datos no estructurados de su organización pueden provenir de diversas fuentes. Pueden ser archivos de texto PDFs, wikis internas, documentos técnicos, sitios web públicos, bases de conocimiento u otros. Si desea un modelo básico que pueda responder a las preguntas sobre los datos no estructurados, tiene a su disposición las siguientes opciones:

- Prepare un nuevo modelo básico utilizando sus documentos personalizados y otros datos de formación
- Perfeccione un modelo básico existente utilizando los datos de sus documentos personalizados
- Utilice el aprendizaje contextual para pasar un documento al modelo básico cuando haga una pregunta
- Utilice un enfoque de recuperación y generación aumentada (RAG)

Formar un nuevo modelo básico desde cero que incluya sus datos personalizados es una tarea ambiciosa. Algunas empresas lo han conseguido con éxito, por ejemplo, Bloomberg con su BloombergGPTmodelo. Otro ejemplo es el EXAONEmodelo multimodalLG AI Research, que se entrenó con 600 mil millones de obras de arte y 250 millones de imágenes de alta resolución, acompañadas de texto. Según The Cost of AI: Should You Build or Buy Your Foundation Model (LinkedIn), la formación de un modelo similar Meta Llama 2 cuesta alrededor de 4,8 millones de dólares. Hay dos requisitos principales para formar un modelo desde cero: el acceso a los recursos (financieros, técnicos y de tiempo) y un claro retorno de la inversión. Si esto no parece ser lo adecuado, la siguiente opción es afinar un modelo básico existente.

El ajuste preciso de un modelo existente implica tomar un modelo, como un modelo Amazon Titan, Mistral o Llama, y luego adaptarlo a los 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 preciso con eficiencia de parámetros. Existen dos métodos principales para el ajuste fino:

- El ajuste supervisado utiliza datos etiquetados y le ayuda a entrenar el modelo para un nuevo tipo de tarea. Por ejemplo, si desea generar un informe basado en un formulario PDF, puede que tenga que enseñarle al modelo cómo hacerlo proporcionando suficientes ejemplos.
- Los ajustes sin supervisión son independientes de las tareas y adaptan el modelo básico a sus propios datos. Entrena al modelo para que comprenda el contexto de los documentos. Luego, el modelo ajustado crea contenido, como un informe, mediante un estilo más personalizado para su organización.

Sin embargo, es posible que el ajuste fino no sea ideal para los casos de uso de preguntas y respuestas. Para obtener más información, consulte Comparación del RAG y el ajuste fino en esta guía.

Al hacer una pregunta, puede pasar a un documento el modelo básico y utilizar el aprendizaje contextual del modelo para obtener respuestas del documento. Esta opción es adecuada para la consulta ad hoc de un solo documento. Sin embargo, esta solución no funciona bien para consultar varios documentos o para consultar sistemas y aplicaciones, como Microsoft SharePoint o Atlassian Confluence.

La última opción es usar RAG. Con RAG, el modelo básico hace referencia a sus documentos personalizados antes de generar una respuesta. RAG amplía las capacidades del modelo a la base de conocimientos interna de su organización, sin necesidad de volver a entrenar el modelo. Es un enfoque rentable para mejorar el resultado del modelo para que siga siendo relevante, preciso y útil en varios contextos.

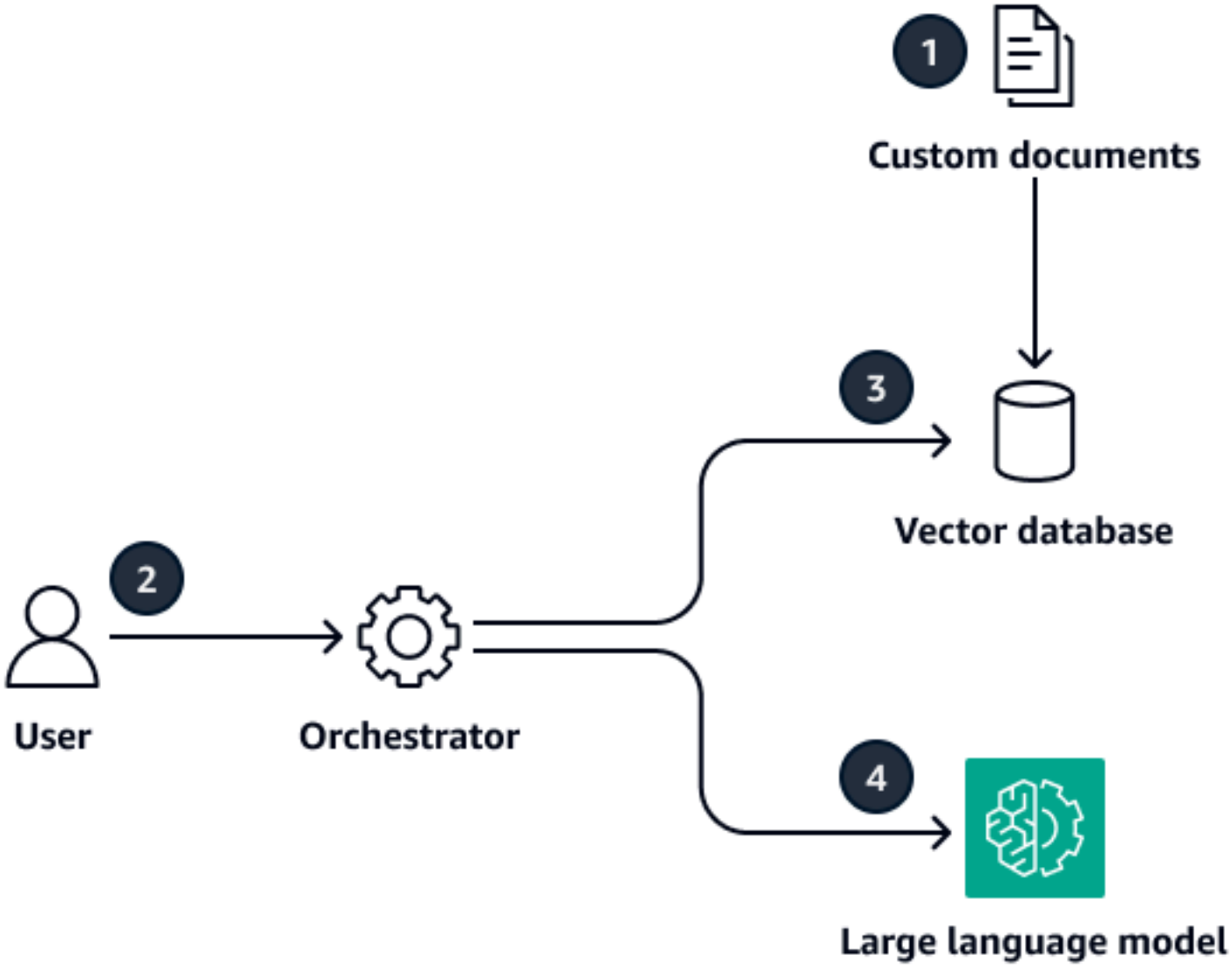
Temas de esta sección:

- Comprensión de la recuperación y la generación aumentada
- Comparación entre la recuperación, la generación aumentada y el ajuste preciso
- Casos de uso de Retrieval Augmented Generation

Comprensión de la recuperación y la generación aumentada

La generación aumentada de recuperación (RAG) es una técnica que se utiliza para complementar un gran modelo de lenguaje (LLM) con datos externos, como los documentos internos de una empresa. Esto proporciona al modelo el contexto que necesita para producir resultados precisos y útiles para su caso de uso específico. El RAG es un enfoque pragmático y eficaz para su uso

LLMs en una empresa. El siguiente diagrama muestra una descripción general de alto nivel de cómo funciona un enfoque RAG.



En términos generales, el proceso RAG consta de cuatro pasos. El primer paso se realiza una vez y los otros tres pasos se realizan tantas veces como sea necesario:

1. Se crean incrustaciones para incorporar los documentos internos a una base de datos vectorial. Las incrustaciones son representaciones numéricas del texto de los documentos que capturan el significado semántico o contextual de los datos. Una base de datos vectorial es básicamente una base de datos de estas incrustaciones y, a veces, se denomina almacén vectorial o índice vectorial. Este paso requiere limpiar, formatear y fragmentar los datos, pero se trata de una actividad inicial que se realiza una sola vez.

2. Un humano envía una consulta en lenguaje natural.
3. Un orquestador realiza una búsqueda de similitudes en la base de datos vectorial y recupera los datos relevantes. El orquestador agrega los datos recuperados (también conocidos como contexto) a la línea de comandos que contiene la consulta.
4. El orquestador envía la consulta y el contexto al LLM. El LLM genera una respuesta a la consulta utilizando el contexto adicional.

Desde la perspectiva del usuario, RAG parece interactuar con cualquier LLM. Sin embargo, el sistema sabe mucho más sobre el contenido en cuestión y proporciona respuestas ajustadas a la base de conocimientos de la organización.

Para obtener más información sobre cómo funciona un enfoque RAG, consulte Qué es el RAG en el sitio web. AWS

Componentes de los sistemas RAG de nivel de producción

La creación de un sistema RAG a nivel de producción requiere pensar en varios aspectos diferentes del flujo de trabajo de RAG. Conceptualmente, un flujo de trabajo de RAG a nivel de producción requiere las siguientes capacidades y componentes, independientemente de la implementación específica:

- Conectores: conectan diferentes fuentes de datos empresariales con la base de datos vectorial. Entre los ejemplos de fuentes de datos estructurados se incluyen las bases de datos transaccionales y analíticas. Algunos ejemplos de fuentes de datos no estructurados son los almacenes de objetos, las bases de código y las plataformas de software como servicio (SaaS). Cada fuente de datos puede requerir patrones de conectividad, licencias y configuraciones diferentes.
- Procesamiento de datos: los datos se presentan en muchas formas y formatos PDFs, como imágenes escaneadas, documentos, presentaciones y Microsoft SharePoint archivos. Debe utilizar técnicas de procesamiento de datos para extraer, procesar y preparar los datos para la indexación.
- Incrustaciones: para realizar una búsqueda de relevancia, debe convertir los documentos y las consultas de los usuarios a un formato compatible. Al utilizar modelos de lenguaje incrustados, se convierten los documentos en una representación numérica. Básicamente, se trata de entradas para el modelo básico subyacente.
- Base de datos vectorial: la base de datos vectorial es un índice de las incrustaciones, el texto asociado y los metadatos. El índice está optimizado para la búsqueda y la recuperación.

- Recuperador: para la consulta del usuario, el recuperador busca el contexto relevante de la base de datos vectorial y clasifica las respuestas en función de los requisitos empresariales.
- Modelo básico: el modelo básico de un sistema RAG suele ser un LLM. Al procesar el contexto y la solicitud, el modelo básico genera y formatea una respuesta para el usuario.
- Barandillas: las barandillas están diseñadas para garantizar que la consulta, el mensaje, el contexto recuperado y la respuesta de LLM sean precisas, responsables, éticas y estén libres de alucinaciones y sesgos.
- Orquestador: el orquestador es responsable de programar y administrar el flujo de trabajo. end-to-end
- Experiencia de usuario: normalmente, el usuario interactúa con una interfaz de chat conversacional que cuenta con numerosas funciones, como mostrar el historial de chats y recopilar los comentarios de los usuarios sobre las respuestas.
- Administración de identidades y usuarios: es fundamental controlar el acceso de los usuarios a la aplicación con precisión. En el sistema Nube de AWS, las políticas, las funciones y los permisos se gestionan normalmente mediante AWS Identity and Access Management (IAM).

Evidentemente, hay una cantidad significativa de trabajo para planificar, desarrollar, lanzar y administrar un sistema RAG. Los servicios totalmente gestionados, como Amazon Bedrock o Amazon Q Business, pueden ayudarle a gestionar parte del trabajo pesado indiferenciado. Sin embargo, las arquitecturas RAG personalizadas pueden proporcionar un mayor control sobre los componentes, como el recuperador o la base de datos vectorial.

Comparación entre la recuperación, la generación aumentada y el ajuste preciso

En la siguiente tabla se describen las ventajas y desventajas de los enfoques de ajuste preciso y basados en RAG.

Enfoque	Ventajas	Desventajas
Ajuste	• Si un modelo ajustado se entrena con un enfoque no supervisado, podrá crear contenido que se	• El ajuste fino puede tardar unas horas o días, según el tamaño del modelo. Por lo tanto, no es una buena

Enfoque	Ventajas	Desventajas
	ajuste más al estilo de su organización. • Un modelo ajustado que se base en datos patentados o reglamentarios puede ayudar a su organización a cumplir con los estándares de cumplimiento y datos internos o específicos de la industria.	solución si sus documentos personalizados cambian con frecuencia. • El ajuste preciso requiere una comprensión de las técnicas, como la adaptación de bajo rango (LoRa) y el ajuste fino con eficiencia de parámetros (PEFT). El ajuste fino puede requerir un científico de datos. • Es posible que el ajuste fino no esté disponible para todos los modelos. • Los modelos ajustados con precisión no incluyen una referencia a la fuente en sus respuestas. • Se puede aumentar el riesgo de alucinaciones cuando se utiliza un modelo ajustado para responder a las preguntas.

Enfoque	Ventajas	Desventajas
RAG	• RAG le permite crear un sistema de preguntas y respuestas para sus documentos personalizados sin necesidad de realizar ajustes. • RAG puede incorporar los documentos más recientes en unos minutos. • AWS ofrece soluciones RAG totalmente gestionad as. Por lo tanto, no se requiere ningún científico de datos ni conocimientos especializados en aprendiza je automático. • En su respuesta, un modelo RAG proporciona una referencia a la fuente de información. • Como RAG utiliza el contexto de la búsqueda vectorial como base de la respuesta generada, se reduce el riesgo de alucinaci ones.	• El RAG no funciona bien al resumir información de documentos completos.

Si necesita crear una solución de preguntas y respuestas que haga referencia a sus documentos personalizados, le recomendamos que comience con un enfoque basado en el RAG. Utilice los ajustes precisos si necesita que el modelo realice tareas adicionales, como el resumen.

Puede combinar los enfoques de ajuste fino y RAG en un único modelo. En este caso, la arquitectura RAG no cambia, pero el LLM que genera la respuesta también se ajusta con los documentos

personalizados. Esto combina lo mejor de ambos mundos y podría ser una solución óptima para su caso de uso. Para obtener más información sobre cómo combinar el ajuste preciso supervisado con el RAG, consulte el estudio RAFT: Adaptación del modelo lingüístico a un dominio específico del RAG del. University of California, Berkeley

Casos de uso de Retrieval Augmented Generation

Los siguientes son casos de uso comunes para utilizar un enfoque de RAG:

- Motores de búsqueda: los motores de búsqueda compatibles con RAG pueden proporcionar fragmentos más precisos y up-to-date destacados en sus resultados de búsqueda.
- Sistemas de preguntas y respuestas: el RAG puede mejorar la calidad de las respuestas en los sistemas de preguntas y respuestas. El modelo basado en la recuperación utiliza la búsqueda por similitud para encontrar pasajes o documentos relevantes que contengan la respuesta. Luego, genera una respuesta concisa y relevante basada en esa información.
- Venta minorista o comercio electrónico: RAG puede mejorar la experiencia del usuario en el comercio electrónico al proporcionar recomendaciones de productos más relevantes y personalizadas. Al recuperar e incorporar información sobre las preferencias del usuario y los detalles del producto, RAG puede generar recomendaciones más precisas y útiles para los clientes.
- Industrial o de fabricación: en la industria manufacturera, RAG le ayuda a acceder rápidamente a información crítica, como las operaciones de la planta de fabricación. También puede ayudar con los procesos de toma de decisiones, la solución de problemas y la innovación organizacional. Para los fabricantes que operan dentro de marcos regulatorios estrictos, RAG puede recuperar rápidamente las regulaciones y estándares de cumplimiento actualizados de fuentes internas y externas, como las normas del sector o las agencias reguladoras.
- Atención médica: RAG tiene potencial en el sector de la salud, donde el acceso a información precisa y oportuna es crucial. Al recuperar e incorporar el conocimiento médico relevante de fuentes externas, el RAG puede proporcionar respuestas más precisas y contextuales en las aplicaciones de atención médica. Estas aplicaciones aumentan la información a la que puede acceder un médico humano, quien, en última instancia, es quien toma la decisión y no el modelo.
- Legal: el RAG se puede aplicar con eficacia en situaciones legales, como fusiones y adquisiciones, en las que los documentos legales complejos proporcionan un contexto para las consultas. Esto puede ayudar a los profesionales del derecho a abordar rápidamente cuestiones reglamentarias complejas.

Opciones de generación aumentada de recuperación totalmente gestionadas en AWS

Para gestionar los flujos de trabajo de Retrieval Augmented Generation (RAG) AWS, puede utilizar canalizaciones de RAG personalizadas o utilizar algunas de las funciones de servicios totalmente gestionados que ofrece. AWS Como incluyen muchos de los componentes principales de un sistema basado en RAG, los servicios totalmente gestionados pueden ayudarle a gestionar parte del trabajo pesado indiferenciado. Sin embargo, estos servicios ofrecen menos oportunidades de personalización.

Los totalmente gestionados Servicios de AWS utilizan conectores para ingerir datos de fuentes de datos externas, como sitios web, Atlassian Confluence o Microsoft. SharePoint Las fuentes de datos compatibles varían según. Servicio de AWS

En esta sección, se analizan las siguientes opciones totalmente gestionadas para crear flujos de trabajo de RAG: AWS

- Bases de conocimiento de Amazon Bedrock
- Amazon Q Business
- Amazon SageMaker AI Canvas

Para obtener más información sobre cómo elegir entre estas opciones, consulte Elegir una opción de generación aumentada de recuperación en AWS esta guía.

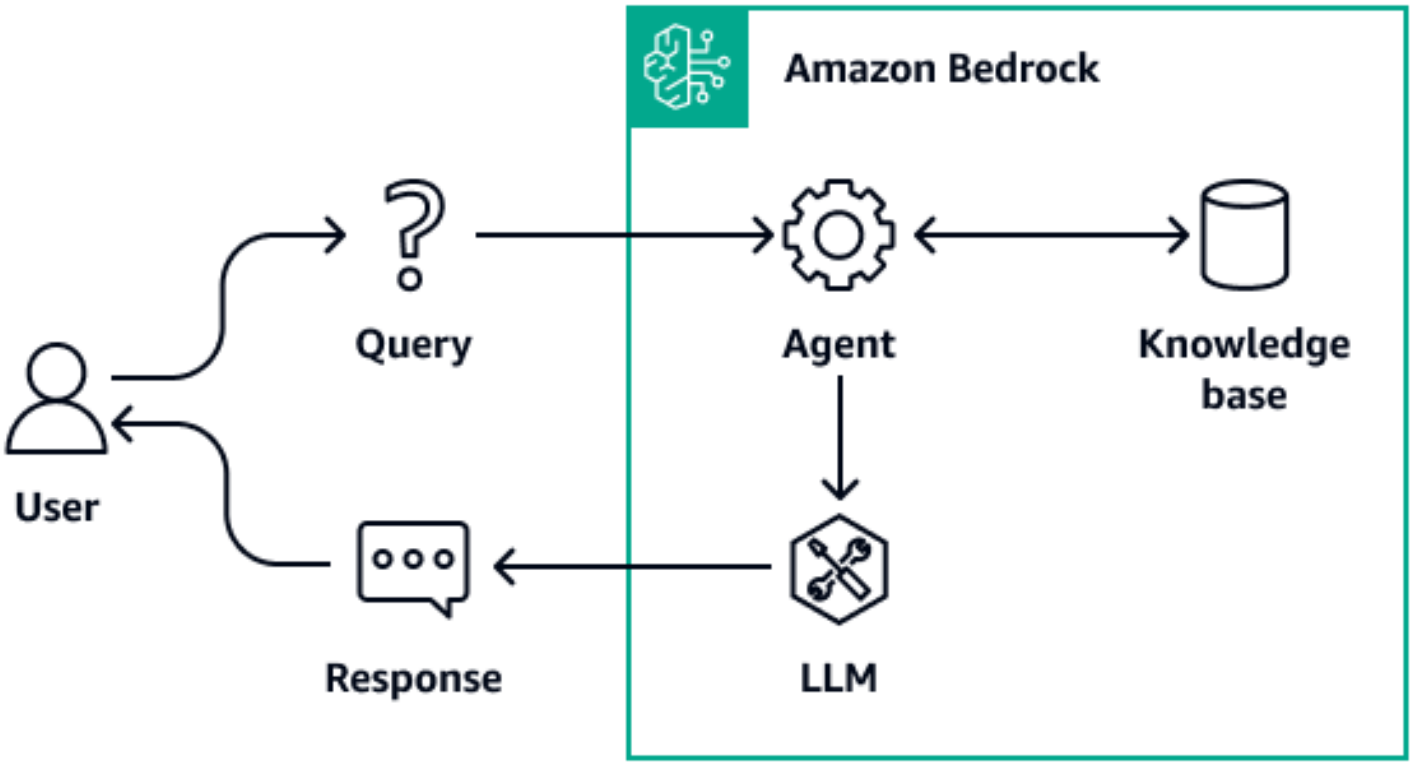
Bases de conocimiento de Amazon Bedrock

Amazon Bedrock es un servicio totalmente gestionado que pone a su disposición modelos básicos de alto rendimiento (FMs) de las principales empresas emergentes de IA y Amazon a través de una API unificada. Las bases de conocimiento son una funcionalidad de Amazon Bedrock que le ayuda a implementar todo el flujo de trabajo de RAG, desde la ingesta hasta la recuperación y el rápido aumento. No es necesario crear integraciones personalizadas con las fuentes de datos ni gestionar los flujos de datos. La gestión del contexto de las sesiones está integrada para que su aplicación de IA generativa pueda admitir fácilmente conversaciones en varios turnos.

Tras especificar la ubicación de los datos, las bases de conocimiento de Amazon Bedrock extraen internamente los documentos, los divide en bloques de texto, convierte el texto en incrustaciones

y, a continuación, las almacena en la base de datos vectorial que elija. Amazon Bedrock administra y actualiza las incrustaciones, manteniendo la base de datos vectorial sincronizada con los datos. Para obtener más información sobre cómo funcionan las bases de conocimiento, consulte Cómo funcionan las bases de conocimiento de Amazon Bedrock.

Si añade bases de conocimiento a un agente de Amazon Bedrock, el agente identifica la base de conocimientos adecuada en función de las entradas del usuario. El agente recupera la información relevante y la añade a la solicitud de entrada. La solicitud actualizada proporciona al modelo más información contextual para generar una respuesta. Para mejorar la transparencia y minimizar las alucinaciones, la información recuperada de la base de conocimientos se puede rastrear hasta su origen.



Amazon Bedrock admite los dos siguientes APIs para RAG:

- RetrieveAndGenerate— Puede utilizar esta API para consultar su base de conocimientos y generar respuestas a partir de la información que recupera. Internamente, Amazon Bedrock convierte las consultas en incrustaciones, consulta la base de conocimientos, amplía la solicitud con los resultados de la búsqueda como información de contexto y devuelve la respuesta generada por LLM. Amazon Bedrock también gestiona la memoria a corto plazo de la conversación para ofrecer resultados más contextuales.

- Recuperar: puede usar esta API para consultar su base de conocimientos con información obtenida directamente de la base de conocimientos. Puedes usar la información devuelta por esta API para procesar el texto recuperado, evaluar su relevancia o desarrollar un flujo de trabajo independiente para generar respuestas. Internamente, Amazon Bedrock convierte las consultas en incrustaciones, busca en la base de conocimientos y devuelve los resultados pertinentes. Puede crear flujos de trabajo adicionales sobre los resultados de búsqueda. Por ejemplo, puedes usar el LangChain`AmazonKnowledgeBasesRetriever`complemento para integrar los flujos de trabajo de RAG en aplicaciones de IA generativa.

Para ver ejemplos de patrones arquitectónicos e step-by-step instrucciones de uso APIs, consulte Knowledge Bases que ahora ofrece una experiencia de RAG totalmente gestionada en Amazon Bedrock (entrada del AWS blog). Para obtener más información sobre cómo usar la `RetrieveAndGenerate` API para crear un flujo de trabajo RAG para una aplicación inteligente basada en chat, consulte Creación de una aplicación de chatbot contextual con Amazon Bedrock Knowledge Bases (AWS entrada del blog).

Orígenes de datos para bases de conocimientos

Puede conectar los datos que son de su propiedad a una base de conocimientos. Después de configurar un conector de fuente de datos, puede sincronizar o mantener los datos actualizados con su base de conocimientos y hacer que estén disponibles para su consulta. Las bases de conocimiento de Amazon Bedrock admiten conexiones a las siguientes fuentes de datos:

- Amazon Simple Storage Service (Amazon S3): puede conectar un bucket de Amazon S3 a una base de conocimientos de Amazon Bedrock mediante la consola o la API. La base de conocimientos ingiere e indexa los archivos del bucket. Este tipo de fuente de datos admite las siguientes funciones:
  - Campos de metadatos del documento: puede incluir un archivo independiente para especificar los metadatos de los archivos del bucket de Amazon S3. A continuación, puede utilizar estos campos de metadatos para filtrar y mejorar la relevancia de las respuestas.
  - Filtros de inclusión o exclusión: puedes incluir o excluir cierto contenido al rastrear.
  - Sincronización incremental: se realiza un seguimiento de los cambios en el contenido y solo se rastrea el contenido que ha cambiado desde la última sincronización.
- Confluence— Puede conectar una Atlassian Confluence instancia a una base de conocimientos de Amazon Bedrock mediante la consola o la API. Este tipo de fuente de datos admite las siguientes funciones:

- Detección automática de los campos principales del documento: los campos de metadatos se detectan y rastrean automáticamente. Puede utilizar estos campos para filtrar.
- Filtros de inclusión o exclusión de contenido: puede incluir o excluir cierto contenido mediante un prefijo o un patrón de expresión regular en el espacio, el título de la página, el título del blog, el comentario, el nombre del archivo adjunto o la extensión.
- Sincronización incremental: se realiza un seguimiento de los cambios en el contenido y solo se rastrea el contenido que ha cambiado desde la última sincronización.
- OAuth Autenticación 2.0, autenticación con token de Confluence API: las credenciales de autenticación se almacenan en. AWS Secrets Manager

- Microsoft SharePoint— Puedes conectar una SharePoint instancia a una base de conocimientos mediante la consola o la API. Este tipo de fuente de datos admite las siguientes funciones:
  - Detección automática de los campos principales del documento: los campos de metadatos se detectan y rastrean automáticamente. Puede utilizar estos campos para filtrar.
  - Filtros de inclusión o exclusión de contenido: puede incluir o excluir determinado contenido mediante un prefijo o un patrón de expresión regular en el título de la página principal, el nombre del evento y el nombre del archivo (incluida su extensión).
  - Sincronización incremental: se realiza un seguimiento de los cambios en el contenido y solo se rastrea el contenido que ha cambiado desde la última sincronización.
  - OAuth Autenticación 2.0: las credenciales de autenticación se almacenan en. AWS Secrets Manager
- Salesforce— Puede conectar una Salesforce instancia a una base de conocimientos mediante la consola o la API. Este tipo de fuente de datos admite las siguientes funciones:
  - Detección automática de los campos principales del documento: los campos de metadatos se detectan y rastrean automáticamente. Puede utilizar estos campos para filtrar.
  - Filtros de inclusión o exclusión de contenido: puede incluir o excluir cierto contenido mediante un prefijo o un patrón de expresión regular. Para obtener una lista de los tipos de contenido a los que puede aplicar filtros, consulte los filtros de inclusión/exclusión en la documentación de Amazon Bedrock.
  - Sincronización incremental: se realiza un seguimiento de los cambios en el contenido y solo se rastrea el contenido que ha cambiado desde la última sincronización.
  - OAuth Autenticación 2.0: las credenciales de autenticación se almacenan en. AWS Secrets Manager

- Rastreador web: un rastreador web de Amazon Bedrock se conecta y rastrea lo que usted proporciona. URLs Se admiten las siguientes características:
  - Seleccione varios para rastrearlos URLs
  - Respeta las directivas estándar de robots.txt, como y `Allow` `Disallow`
  - Excluya URLs los que coincidan con un patrón
  - Limite la velocidad de rastreo
  - En Amazon CloudWatch, consulta el estado de cada URL rastreada

Para obtener más información sobre las fuentes de datos que puede conectar a su base de conocimiento de Amazon Bedrock, consulte Crear un conector de fuente de datos para su base de conocimientos.

Bases de datos vectoriales para bases de conocimiento

Al configurar una conexión entre la base de conocimientos y la fuente de datos, debe configurar una base de datos vectorial, también conocida como almacén vectorial. Una base de datos vectorial es el lugar donde Amazon Bedrock almacena, actualiza y administra las incrustaciones que representan sus datos. Cada fuente de datos admite distintos tipos de bases de datos vectoriales. Para determinar qué bases de datos vectoriales están disponibles para su fuente de datos, consulte los tipos de fuentes de datos.

Si prefiere que Amazon Bedrock cree automáticamente una base de datos vectorial en Amazon OpenSearch Serverless, puede elegir esta opción al crear la base de conocimientos. Sin embargo, también puede optar por configurar su propia base de datos vectorial. Si configura su propia base de datos vectoriales, consulte Requisitos previos para su propio almacén de vectores para obtener una base de conocimientos. Cada tipo de base de datos vectorial tiene sus propios requisitos previos.

Según el tipo de fuente de datos, las bases de conocimiento de Amazon Bedrock admiten las siguientes bases de datos vectoriales:

- Amazon OpenSearch Serverless
- Amazon Aurora PostgreSQL-Compatible Edition
- Pinecone (documentación de Pinecone)
- Redis Enterprise Cloud (documentación de Redis)
- MongoDB Atlas (documentación de MongoDB)

Amazon Q Business

Amazon Q Business es un asistente totalmente gestionado y basado en IA generativa que puede configurar para responder preguntas, proporcionar resúmenes, generar contenido y completar tareas en función de los datos de su empresa. Permite a los usuarios finales recibir respuestas inmediatas y basadas en los permisos de fuentes de datos empresariales con citas.

Características principales de

Las siguientes funciones de Amazon Q Business pueden ayudarle a crear una aplicación de IA generativa basada en RAG apta para producción:

- Conectores integrados: Amazon Q Business admite más de 40 tipos de conectores, como conectores para Adobe Experience Manager (AEM) SalesforceJira, yMicrosoft SharePoint. Para obtener una lista completa, consulte Conectores compatibles. Si necesitas un conector que no sea compatible, puedes usar Amazon AppFlow para extraer datos de tu fuente de datos y llevarlos a Amazon Simple Storage Service (Amazon S3) y, a continuación, conectar Amazon Q Business al bucket de Amazon S3. Para ver una lista completa de las fuentes de datos AppFlow compatibles con Amazon, consulta Aplicaciones compatibles.
- Canalizaciones de indexación integradas: Amazon Q Business proporciona una canalización integrada para indexar datos en una base de datos vectorial. Puede utilizar una AWS Lambda función para añadir una lógica de preprocesamiento a su proceso de indexación.
- Opciones de índice: puede crear y aprovisionar un índice nativo en Amazon Q Business y utilizar un recuperador de Amazon Q Business para extraer datos de ese índice. Como alternativa, puede utilizar un índice de Amazon Kendra preconfigurado como recuperador. Para obtener más información, consulte Creación de un recuperador para una aplicación de Amazon Q Business.
- Modelos básicos: Amazon Q Business utiliza los modelos básicos compatibles con Amazon Bedrock. Para obtener una lista completa, consulte los modelos de cimentación compatibles en Amazon Bedrock.
- Plugins: Amazon Q Business ofrece la posibilidad de utilizar complementos para integrarse con los sistemas de destino, como una forma automática de resumir la información de los tickets y la creación de tickets en Jira ellos. Una vez configurados, los complementos pueden facilitar acciones de lectura y escritura que pueden ayudarlo a aumentar la productividad de los usuarios finales. Amazon Q Business admite dos tipos de complementos: complementos integrados y complementos personalizados.

- Barandillas: Amazon Q Business admite controles globales y controles a nivel temático. Por ejemplo, estos controles pueden detectar información de identificación personal (PII), usos indebidos o información confidencial en las solicitudes. Para obtener más información, consulte Controles de administración y barandas de protección en Amazon Q Business.
- Gestión de identidades: con Amazon Q Business, puede gestionar los usuarios y su acceso a la aplicación de IA generativa basada en RAG. Para obtener más información, consulte Administración de identidad y acceso para Amazon Q Business. Además, los conectores de Amazon Q Business indexan la información de la lista de control de acceso (ACL) que se adjunta a un documento junto con el propio documento. A continuación, Amazon Q Business almacena la información de ACL que indexa en la tienda de usuarios de Amazon Q Business para crear asignaciones de usuarios y grupos y filtrar las respuestas de chat en función del acceso del usuario final a los documentos. Para obtener más información, consulte Conceptos de conectores de fuentes de datos.
- Enriquecimiento de documentos: la función de enriquecimiento de documentos le ayuda a controlar qué documentos y atributos de los documentos se incorporan al índice y también cómo se ingieren. Esto se puede lograr mediante dos enfoques:
  - Configure las operaciones básicas: utilice las operaciones básicas para añadir, actualizar o eliminar los atributos del documento de los datos. Por ejemplo, puede eliminar los datos de PII si elige eliminar cualquier atributo del documento relacionado con la PII.
  - Configuración de funciones Lambda: utilice una función Lambda preconfigurada para aplicar a sus datos una lógica de manipulación de atributos de documentos más avanzada y personalizada. Por ejemplo, los datos de su empresa pueden almacenarse como imágenes escaneadas. En ese caso, puede utilizar una función Lambda para ejecutar el reconocimiento óptico de caracteres (OCR) en los documentos escaneados para extraer texto de ellos. A continuación, cada documento escaneado se trata como un documento de texto durante la ingestión. Por último, durante el chat, Amazon Q tendrá en cuenta los datos textuales extraídos de los documentos escaneados cuando genere las respuestas.

  Al implementar la solución, puede optar por combinar ambos enfoques de enriquecimiento de documentos. Puede utilizar operaciones básicas para analizar primero los datos y, a continuación, utilizar una función Lambda para operaciones más complejas. Para obtener más información, consulte Enriquecimiento de documentos en Amazon Q Business.
- Integración: después de crear la aplicación Amazon Q Business, puede integrarla en otras aplicaciones, como Slack oMicrosoft Teams. Por ejemplo, consulte Implementar una Slack puerta

de enlace para Amazon Q Business e Implementar una Microsoft Teams puerta de enlace para Amazon Q Business (publicaciones de AWS blog).

Personalización para el usuario final

Amazon Q Business admite la carga de documentos que podrían no estar almacenados en las fuentes de datos y el índice de su organización. Los documentos cargados no se almacenan. Solo están disponibles para su uso en la conversación en la que se cargan los documentos. Amazon Q Business admite la carga de tipos de documentos específicos. Para obtener más información, consulta Cómo subir archivos y chatear en Amazon Q Business.

Amazon Q Business incluye una función de filtrado por atributo de documento. Tanto los administradores como los usuarios finales pueden utilizar esta función. Los administradores pueden personalizar y controlar las respuestas de chat para los usuarios finales mediante el uso de atributos. Por ejemplo, si el tipo de origen de datos es un atributo asociado a sus documentos, puede especificar que las respuestas de chat se generen únicamente a partir de un origen de datos específico. O bien, puede permitir que los usuarios finales restrinjan el alcance de las respuestas del chat mediante los filtros de atributos que haya seleccionado.

Los usuarios finales pueden crear Amazon Q Apps ligeras y diseñadas específicamente dentro de su entorno más amplio de aplicaciones de Amazon Q Business. Las aplicaciones Amazon Q permiten la automatización de tareas para un dominio específico, como una aplicación diseñada específicamente para el equipo de marketing.

Amazon SageMaker AI Canvas

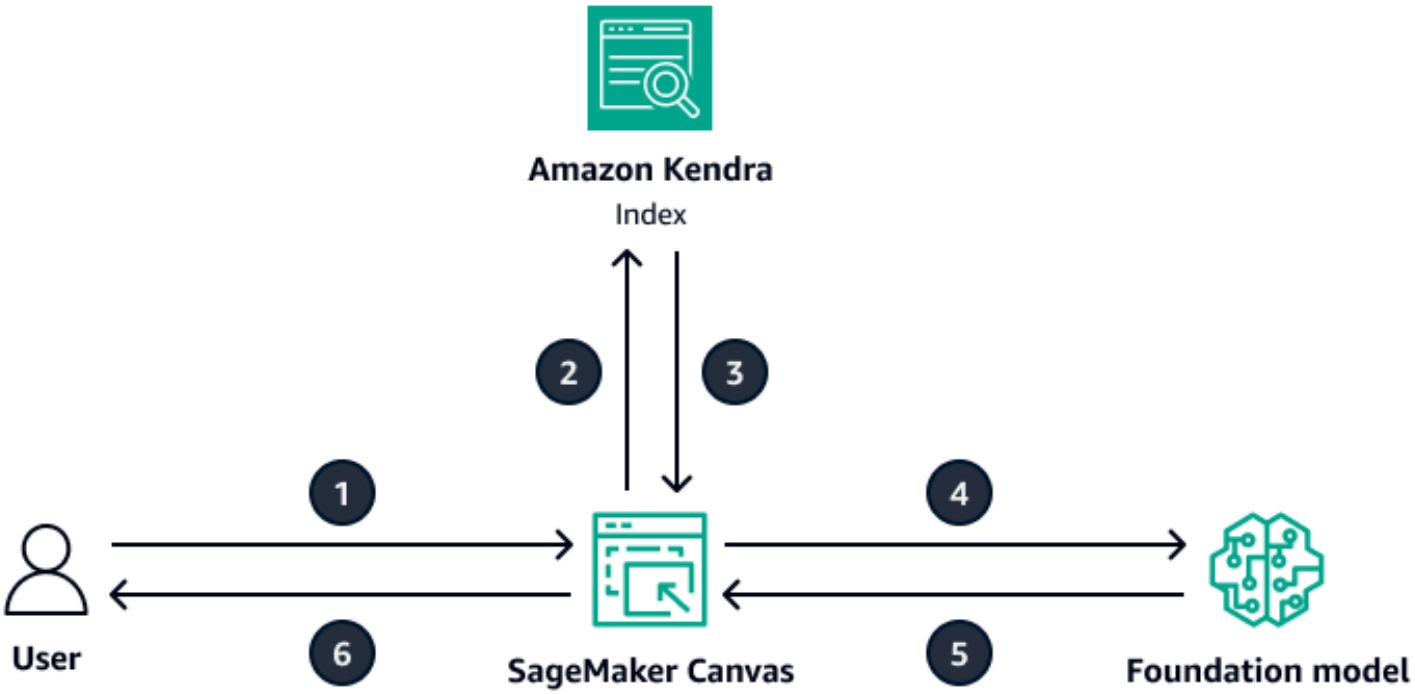
Amazon SageMaker AI Canvas le ayuda a utilizar el aprendizaje automático para generar predicciones sin necesidad de escribir ningún código. Proporciona una interfaz visual sin código que le permite preparar datos, crear e implementar modelos de aprendizaje automático, lo que agiliza el ciclo de vida del end-to-end aprendizaje automático en un entorno unificado. Las complejidades de la preparación de los datos, el desarrollo de modelos, la detección de sesgos, la explicabilidad y la supervisión se resumen en una interfaz intuitiva. Los usuarios no necesitan ser expertos en SageMaker inteligencia artificial o en operaciones de aprendizaje automático (MLOps) para desarrollar, operacionalizar y monitorear modelos con AI Canvas. SageMaker

Con SageMaker AI Canvas, la funcionalidad RAG se proporciona a través de una función de consulta de documentos sin código. Puedes enriquecer la experiencia de chat en SageMaker AI Canvas

utilizando un índice de Amazon Kendra como búsqueda empresarial subyacente. Para obtener más información, consulte Extraer información de documentos mediante consultas de documentos.

La conexión de SageMaker AI Canvas al índice de Amazon Kendra requiere una configuración única. Como parte de la configuración del dominio, un administrador de la nube puede elegir uno o más índices de Kendra que el usuario puede consultar al interactuar con Canvas. SageMaker Para obtener instrucciones sobre cómo habilitar la función de consulta de documentos, consulte Cómo empezar a usar Amazon SageMaker AI Canvas.

SageMaker AI Canvas gestiona la comunicación subyacente entre Amazon Kendra y el modelo base seleccionado. Para obtener más información sobre los modelos básicos compatibles con SageMaker AI Canvas, consulte los modelos básicos de IA generativa en SageMaker AI Canvas. El siguiente diagrama muestra cómo funciona la función de consulta de documentos después de que el administrador de la nube haya conectado SageMaker AI Canvas a un índice de Amazon Kendra.



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

1. El usuario inicia un nuevo chat en SageMaker AI Canvas, activa la consulta de documentos, selecciona el índice objetivo y, a continuación, envía una pregunta.
2. SageMaker AI Canvas utiliza la consulta para buscar datos relevantes en el índice de Amazon Kendra.

3. SageMaker AI Canvas recupera los datos y sus fuentes del índice Amazon Kendra.
4. SageMaker AI Canvas actualiza la solicitud para incluir el contexto recuperado del índice Amazon Kendra y envía la solicitud al modelo base.
5. El modelo básico utiliza la pregunta original y el contexto recuperado para generar una respuesta.
6. SageMaker AI Canvas proporciona la respuesta generada al usuario. Incluye referencias a las fuentes de datos, como los documentos, que se utilizaron para generar la respuesta.

Arquitecturas de generación aumentada de recuperación personalizada en AWS

En la sección anterior, se describe cómo utilizar una generación aumentada de recuperación (Servicio de AWS RAG) totalmente gestionada. Sin embargo, algunos casos de uso requieren un mayor control sobre los componentes del sistema, como el recuperador o el LLM (también denominado generador). Por ejemplo, es posible que necesite la flexibilidad necesaria para elegir su propia base de datos vectorial o acceder a una fuente de datos no compatible. Para estos casos de uso, puede crear una arquitectura RAG personalizada.

Esta sección contiene los siguientes temas:

- Recuperadores para flujos de trabajo RAG
- Generadores para flujos de trabajo RAG

Para obtener más información sobre cómo elegir entre las opciones de recuperación y generador de esta sección, consulte esta Elegir una opción de generación aumentada de recuperación en AWS guía.

Recuperadores para flujos de trabajo RAG

En esta sección se explica cómo crear un recuperador. Puede utilizar una solución de búsqueda semántica totalmente gestionada, como Amazon Kendra, o puede crear una búsqueda semántica personalizada mediante AWS una base de datos vectorial.

Antes de revisar las opciones del recuperador, asegúrese de entender los tres pasos del proceso de búsqueda vectorial:

1. Separa los documentos que deben indexarse en partes más pequeñas. Esto se denomina fragmentación.
2. Se utiliza un proceso llamado incrustación para convertir cada fragmento en un vector matemático. A continuación, indexa cada vector en una base de datos vectorial. El enfoque que utilice para indexar los documentos influye en la velocidad y precisión de la búsqueda. El enfoque de indexación depende de la base de datos vectorial y de las opciones de configuración que proporciona.

3. La consulta del usuario se convierte en un vector mediante el mismo proceso. El recuperador busca en la base de datos vectoriales vectores que sean similares al vector de consulta del usuario. La similitud se calcula mediante métricas como la distancia euclidiana, la distancia por coseno o el producto de puntos.

En esta guía se describe cómo utilizar los siguientes servicios Servicios de AWS o los de terceros para crear una capa de recuperación personalizada en: AWS

- Amazon Kendra
- OpenSearch Servicio Amazon
- Amazon Aurora, PostgreSQL y pgvector
- Análisis por Amazon Neptune
- Amazon MemoryDB
- Amazon DocumentDB
- Pinecone
- MongoDB Atlas
- Weaviate

Amazon Kendra

Amazon Kendra es un servicio de búsqueda inteligente y totalmente gestionado que utiliza el procesamiento del lenguaje natural y algoritmos avanzados de aprendizaje automático para devolver respuestas específicas a las preguntas de búsqueda a partir de sus datos. Amazon Kendra le ayuda a ingerir directamente documentos de varias fuentes y a consultarlos una vez que se hayan sincronizado correctamente. El proceso de sincronización crea la infraestructura necesaria para crear una búsqueda vectorial en el documento ingerido. Por lo tanto, Amazon Kendra no requiere los tres pasos tradicionales del proceso de búsqueda vectorial. Tras la sincronización inicial, puede utilizar un programa definido para gestionar la ingesta continua.

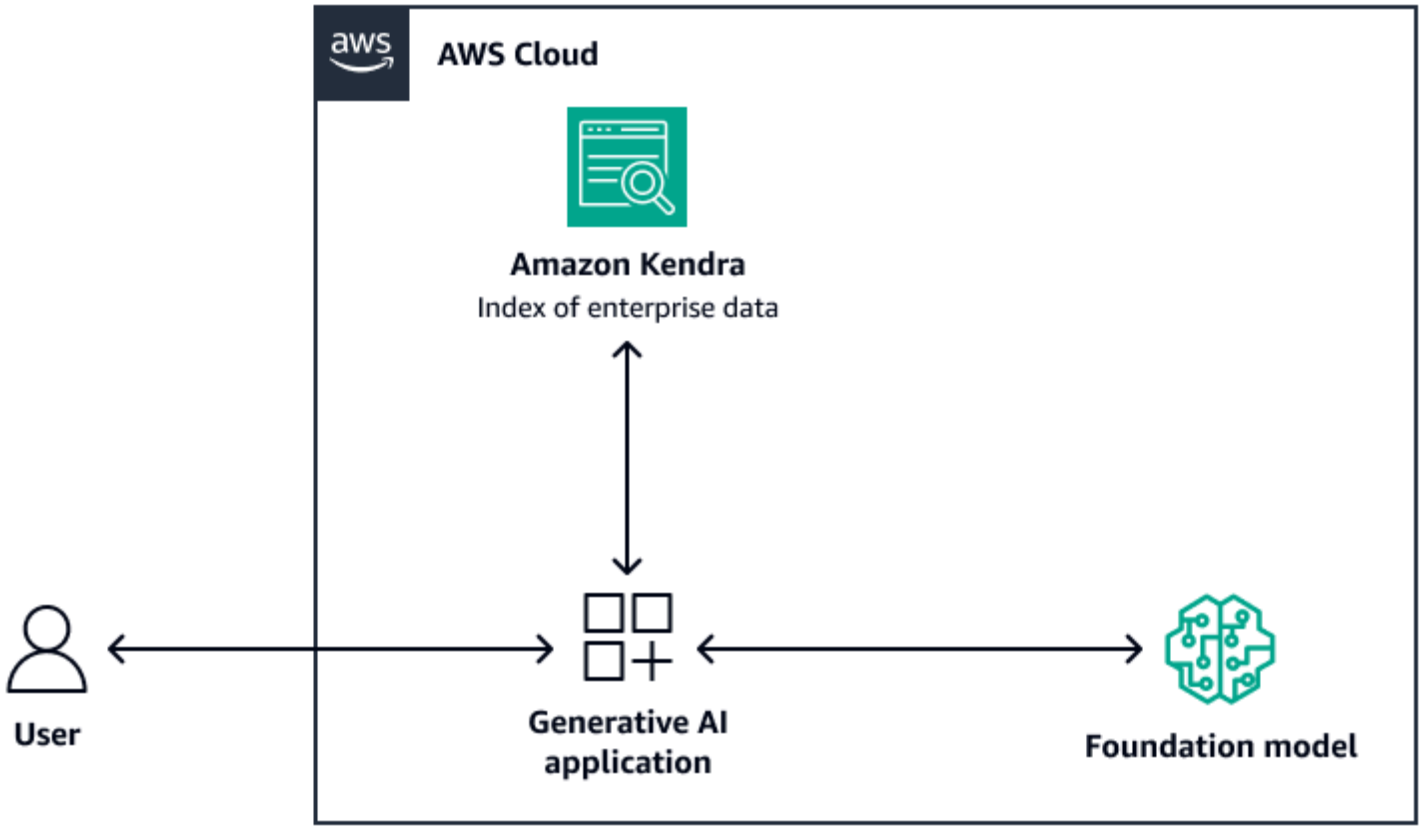
Las siguientes son las ventajas de usar Amazon Kendra para RAG:

- No es necesario mantener una base de datos vectorial porque Amazon Kendra se encarga de todo el proceso de búsqueda vectorial.
- Amazon Kendra contiene conectores prediseñados para fuentes de datos populares, como bases de datos, rastreadores de sitios web, buckets e instancias de Amazon S3. Microsoft SharePoint

Atlassian Confluence Están disponibles conectores desarrollados por AWS los socios, como conectores para y. Box GitLab

- Amazon Kendra proporciona un filtrado de listas de control de acceso (ACL) que devuelve solo los documentos a los que tiene acceso el usuario final.
- Amazon Kendra puede impulsar las respuestas en función de los metadatos, como la fecha o el repositorio de origen.

La siguiente imagen muestra un ejemplo de arquitectura que utiliza Amazon Kendra como capa de recuperación del sistema RAG. Para obtener más información, consulte Cree rápidamente aplicaciones de IA generativa de alta precisión a partir de datos empresariales mediante Amazon Kendra LangChain y modelos de lenguaje de gran tamañoAWS (entrada del blog).



Para el modelo básico, puede usar Amazon Bedrock o un LLM implementado a través de Amazon SageMaker AI. JumpStart Puede usarlo AWS Lambda con LangChainpara organizar el flujo entre el usuario, Amazon Kendra y el LLM. Para crear un sistema RAG que utilice Amazon, Kendra y LLMs variosLangChain, consulte el repositorio de Amazon LangChain Kendra Extensions. GitHub

OpenSearch Servicio Amazon

Amazon OpenSearch Service proporciona algoritmos de aprendizaje automático integrados para la búsqueda de k-vecinos más cercanos (k-NN) con el fin de realizar una búsqueda vectorial. OpenSearch El servicio también proporciona un motor vectorial para Amazon EMR Serverless. Puede usar este motor vectorial para crear un sistema RAG que tenga capacidades de búsqueda y almacenamiento vectorial escalables y de alto rendimiento. Para obtener más información sobre cómo crear un sistema RAG mediante OpenSearch Serverless, consulte Crear flujos de trabajo RAG escalables y sin servidor con un motor vectorial para los modelos Amazon Serverless OpenSearch y Amazon Bedrock Claude (entrada del blog).AWS

Las ventajas de utilizar Service para la búsqueda vectorial son las siguientes: OpenSearch

- Proporciona un control total sobre la base de datos vectoriales, incluida la creación de una búsqueda vectorial escalable mediante OpenSearch Serverless.
- Proporciona control sobre la estrategia de fragmentación.
- Utiliza algoritmos de vecino más cercano (ANN) aproximados de las bibliotecas Non-Metric Space Library (NMSLIB), Faiss y Apache Lucene para impulsar una búsqueda k-NN. Puede cambiar el algoritmo en función del caso de uso. Para obtener más información sobre las opciones para personalizar la búsqueda vectorial mediante el OpenSearch Servicio, consulta la explicación de las capacidades de las bases de datos vectoriales de Amazon OpenSearch Service (entrada AWS del blog).
- OpenSearch Serverless se integra con las bases de conocimiento de Amazon Bedrock como un índice vectorial.

Amazon Aurora, PostgreSQL y pgvector

La edición compatible con PostgreSQL de Amazon Aurora es un motor de base de datos relacional totalmente administrado que le ayuda a configurar, operar y escalar las implementaciones de PostgreSQL. pgvector es una extensión de código abierto para PostgreSQL que proporciona capacidades de búsqueda de similitudes vectoriales. Esta extensión está disponible tanto para la versión compatible con Aurora PostgreSQL como para Amazon Relational Database Service (Amazon RDS) para PostgreSQL. Para obtener más información sobre cómo crear un sistema basado en RAG que utilice pgvector y Aurora, compatible con PostgreSQL, consulte las siguientes publicaciones del blog: AWS

- Creación de búsquedas basadas en IA en PostgreSQL con Amazon AI y pgvector SageMaker

- Aproveche pgvector y Amazon Aurora PostgreSQL para el procesamiento de lenguaje natural, los chatbots y el análisis de opiniones

Las siguientes son las ventajas de usar pgvector y Aurora compatible con PostgreSQL:

- Soporta la búsqueda exacta y aproximada del vecino más cercano. También admite las siguientes métricas de similitud: distancia L2, producto interior y distancia del coseno.
- Es compatible con archivos invertidos con compresión plana (IVFFlat) e indexación jerárquica de mundos pequeños navegables (HNSW).
- Puede combinar la búsqueda vectorial con consultas sobre datos específicos del dominio que estén disponibles en la misma instancia de PostgreSQL.
- Aurora, compatible con PostgreSQL, está optimizado para el almacenamiento en caché por niveles I/O y lo proporciona. Para cargas de trabajo que superen la memoria de instancia disponible, pgvector puede aumentar las consultas por segundo para la búsqueda vectorial hasta 8 veces.

Análisis por Amazon Neptune

Amazon Neptune Analytics es un motor de base de datos de gráficos con memoria optimizada para análisis. Es compatible con una biblioteca de algoritmos analíticos de gráficos optimizados, consultas gráficas de baja latencia y capacidades de búsqueda vectorial en los recorridos de gráficos. También tiene una búsqueda de similitud vectorial integrada. Proporciona un punto final para crear un gráfico, cargar datos, invocar consultas y realizar búsquedas de similitud vectorial. Para obtener más información sobre cómo crear un sistema basado en RAG que utilice Neptune Analytics, consulte Uso de gráficos de conocimiento para crear aplicaciones de GraphRag con Amazon Bedrock y Amazon Neptune AWS (entrada del blog).

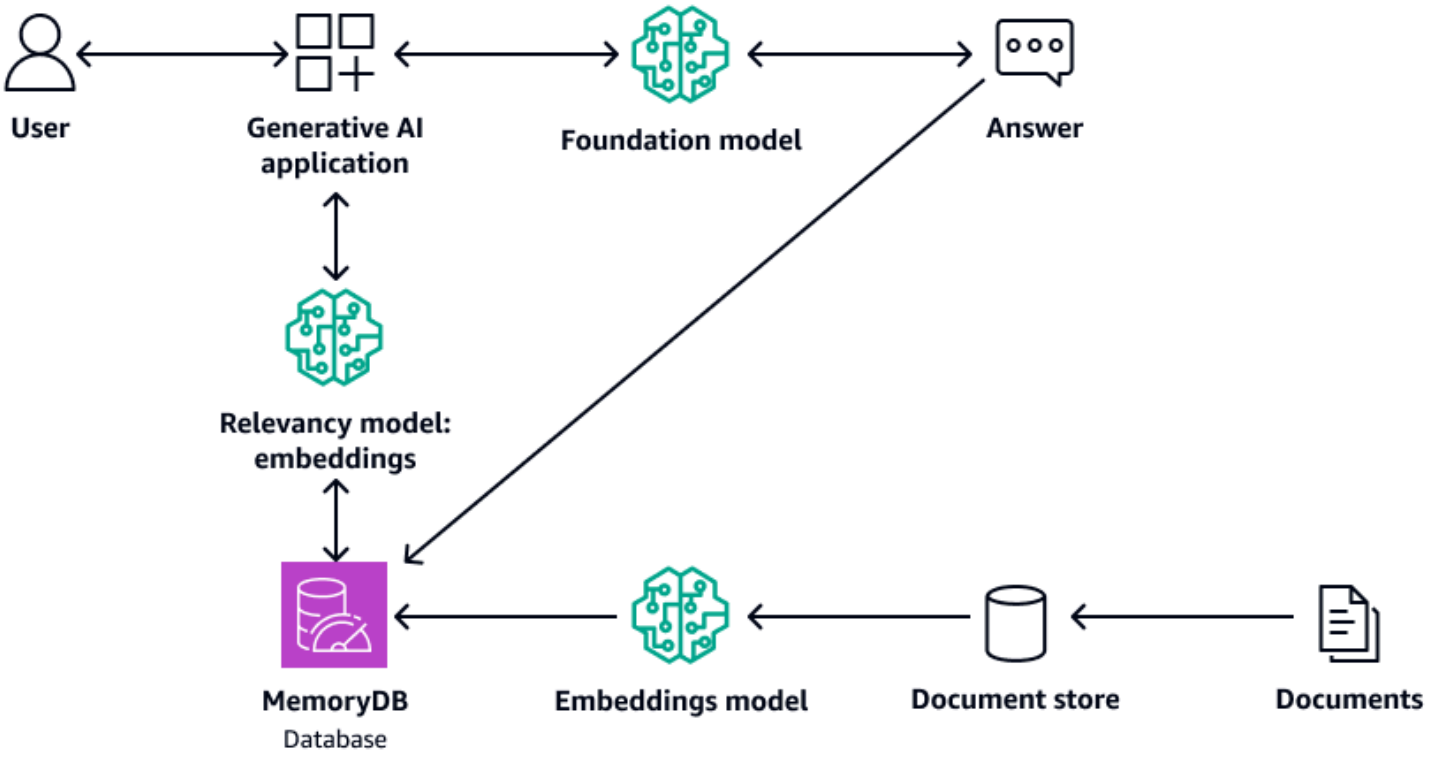
Las ventajas de utilizar Neptune Analytics son las siguientes:

- Puede almacenar y buscar incrustaciones en las consultas de gráficos.
- Si integra Neptune Analytics conLangChain, esta arquitectura admite consultas gráficas en lenguaje natural.
- Esta arquitectura almacena grandes conjuntos de datos de gráficos en la memoria.

Amazon MemoryDB

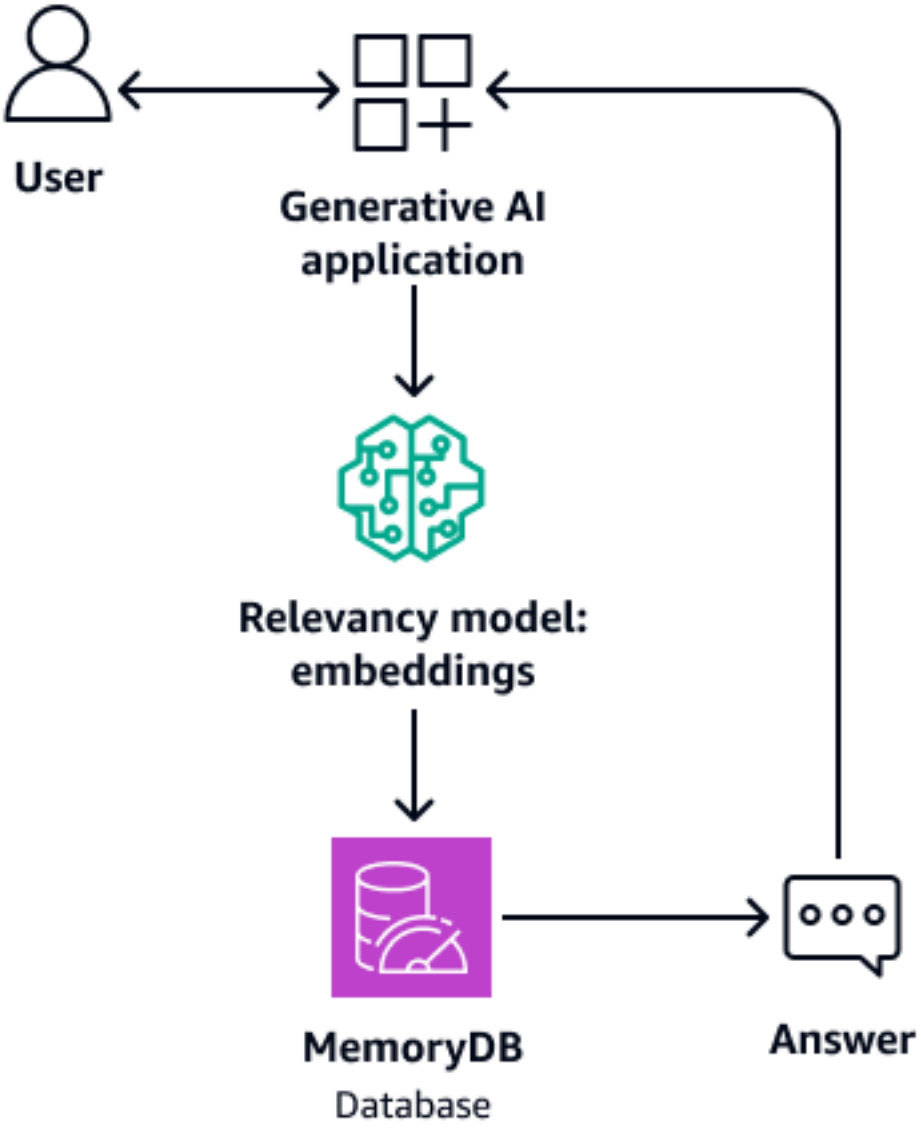
Amazon MemoryDB es un servicio de base de datos en memoria duradero que ofrece un rendimiento ultrarrápido. Todos los datos se almacenan en la memoria, lo que admite una lectura de microsegundos, una latencia de escritura de milisegundos de un solo dígito y un alto rendimiento. La búsqueda vectorial de MemoryDB amplía la funcionalidad de MemoryDB y se puede utilizar junto con la funcionalidad de MemoryDB existente. Para obtener más información, consulte el repositorio Preguntas y respuestas con LLM y RAG en. GitHub

El siguiente diagrama muestra un ejemplo de arquitectura que utiliza MemoryDB como base de datos vectorial.



Las ventajas de utilizar MemoryDB son las siguientes:

- Es compatible con los algoritmos de indexación Flat y HNSW. Para obtener más información, consulte La búsqueda vectorial de Amazon MemoryDB ya está disponible de forma general en el blog de noticias AWS
- También puede actuar como memoria intermedia para el modelo básico. Esto significa que las preguntas respondidas anteriormente se recuperan del búfer en lugar de volver a pasar por el proceso de recuperación y generación. El siguiente diagrama muestra este proceso.



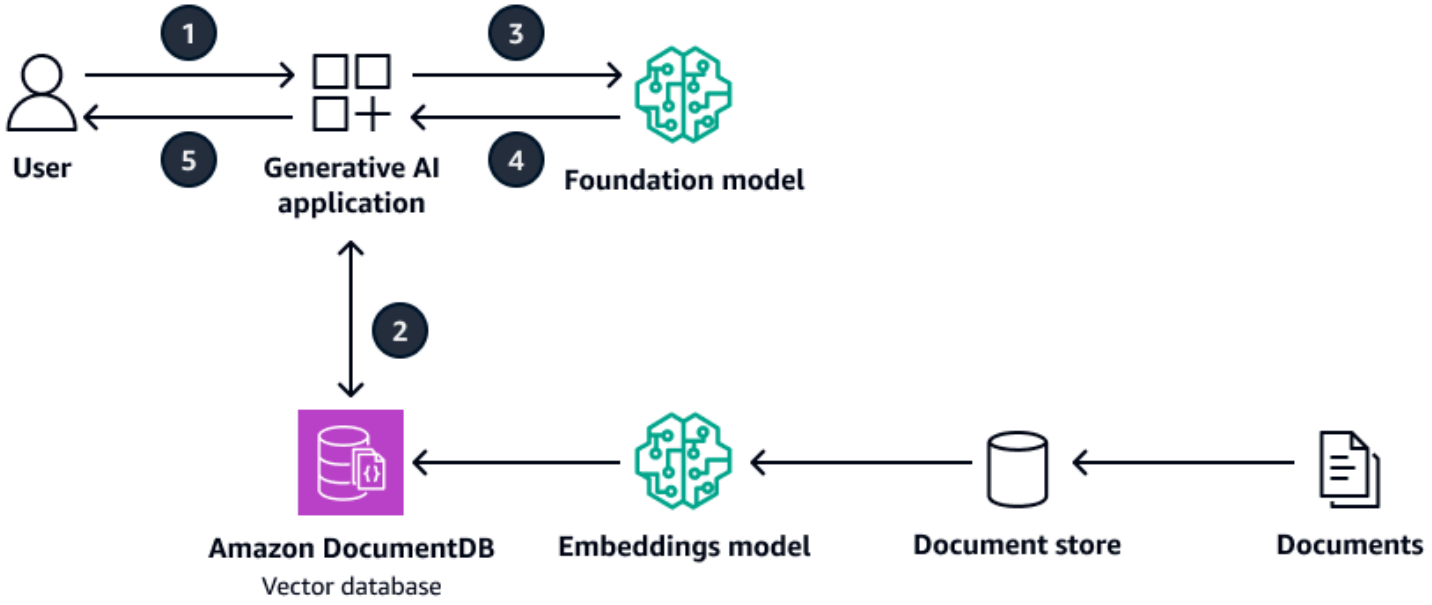
- Como utiliza una base de datos en memoria, esta arquitectura proporciona un tiempo de consulta de milisegundos de un solo dígito para la búsqueda semántica.
- Proporciona hasta 33 000 consultas por segundo con una recuperación del 95 al 99% y 26 500 consultas por segundo con una recuperación superior al 99%. Para obtener más información, consulte el vídeo AWS re:Invent 2023: búsqueda vectorial de latencia ultrabaja para Amazon MemoryDB en. YouTube

Amazon DocumentDB

Amazon DocumentDB (con compatibilidad con MongoDB) es un servicio rápido, de confianza y completamente administrado. Facilita la configuración, el funcionamiento y el MongoDB escalado de

bases de datos compatibles en la nube. La búsqueda vectorial para Amazon DocumentDB combina la flexibilidad y la amplia capacidad de consulta de una base de datos de documentos basada en JSON con la potencia de la búsqueda vectorial. Para obtener más información, consulte el repositorio Preguntas y respuestas con LLM y RAG en. GitHub

En el siguiente diagrama se muestra un ejemplo de arquitectura que utiliza Amazon DocumentDB como base de datos vectorial.



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

1. El usuario envía una consulta a la aplicación de IA generativa.
2. La aplicación de IA generativa realiza una búsqueda de similitudes en la base de datos vectorial Amazon DocumentDB y recupera los extractos de documentos pertinentes.
3. La aplicación de IA generativa actualiza la consulta del usuario con el contexto recuperado y envía la solicitud al modelo base objetivo.
4. El modelo básico utiliza el contexto para generar una respuesta a la pregunta del usuario y devuelve la respuesta.
5. La aplicación de IA generativa devuelve la respuesta al usuario.

Las ventajas de utilizar Amazon DocumentDB son las siguientes:

- Es compatible con los métodos HNSW y de IVFFlat indexación.

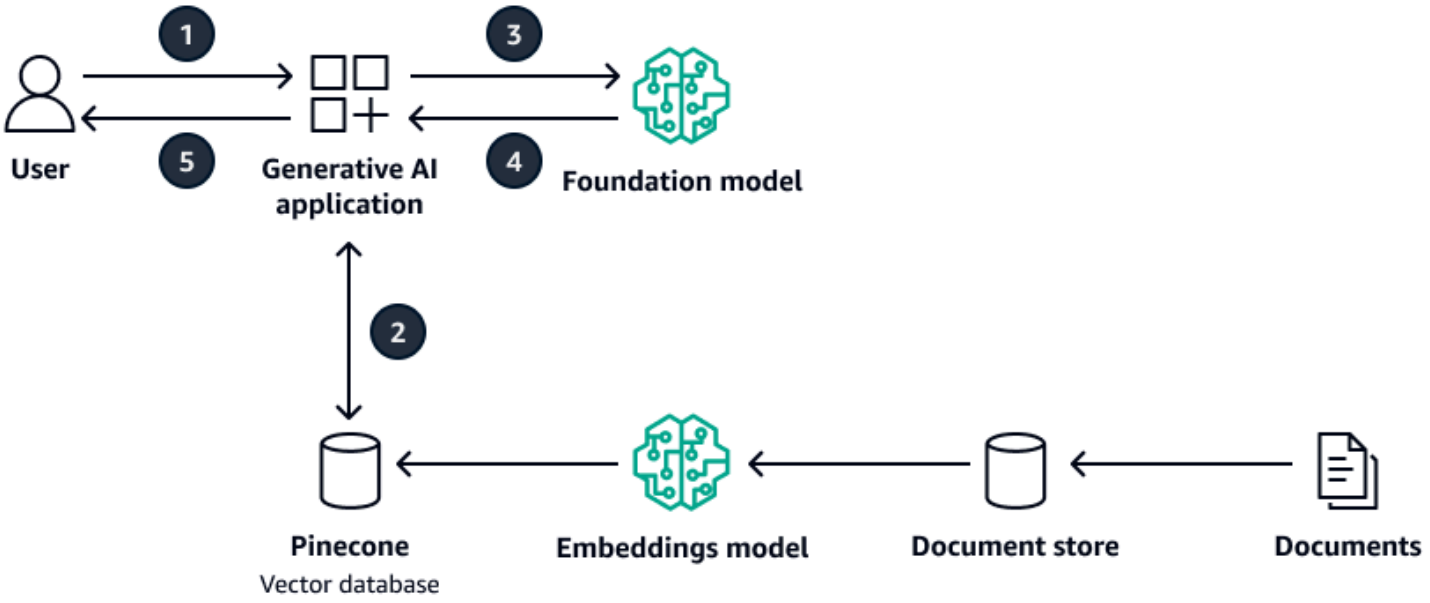
- Admite hasta 2000 dimensiones en los datos vectoriales y admite las métricas de distancia euclidiana, coseno y producto puntual.
- Proporciona tiempos de respuesta en milisegundos.

Pinecone

Pineconees una base de datos vectorial totalmente gestionada que le ayuda a añadir la búsqueda vectorial a las aplicaciones de producción. Está disponible a través de AWS Marketplace. La facturación se basa en el uso y los cargos se calculan multiplicando el precio del pod por el número de pods. Para obtener más información sobre cómo crear un sistema basado en RAG que utilicePinecone, consulta las siguientes entradas del blog: AWS

- Mitigue las alucinaciones mediante RAG utilizando la base de datos Pinecone vectorial y Llama-2 de Amazon AI SageMaker JumpStart
- Utilice Amazon SageMaker AI Studio para crear una solución de respuesta a preguntas RAG con Llama 2 y Pinecone para experimentar rápidamente LangChain

El siguiente diagrama muestra un ejemplo de arquitectura que se utiliza Pinecone como base de datos vectorial.



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

1. El usuario envía una consulta a la aplicación de IA generativa.
2. La aplicación de IA generativa realiza una búsqueda de similitudes en la base de datos Pinecone vectoriales y recupera los extractos de documentos relevantes.
3. La aplicación de IA generativa actualiza la consulta del usuario con el contexto recuperado y envía la solicitud al modelo básico objetivo.
4. El modelo básico utiliza el contexto para generar una respuesta a la pregunta del usuario y devuelve la respuesta.
5. La aplicación de IA generativa devuelve la respuesta al usuario.

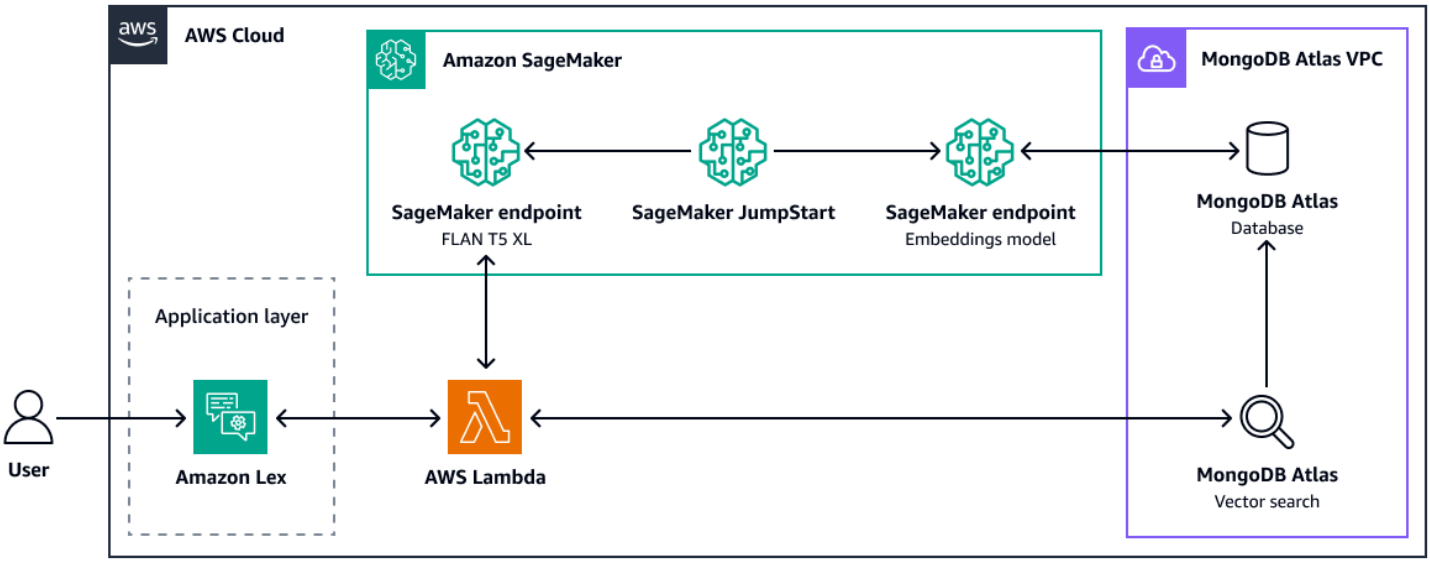
Las siguientes son las ventajas de usarPinecone:

- Es una base de datos vectorial completamente administrada y elimina la sobrecarga de administrar su propia infraestructura.
- Ofrece funciones adicionales como el filtrado, las actualizaciones de índices en tiempo real y la mejora de las palabras clave (búsqueda híbrida).

MongoDB Atlas

MongoDB Atlases una base de datos en la nube totalmente gestionada que gestiona toda la complejidad de la implementación y la gestión de sus despliegues. AWS Puede utilizar la búsqueda vectorial MongoDB Atlas para almacenar incrustaciones vectoriales en su base de datos. MongoDB Las bases de conocimiento de Amazon Bedrock admiten MongoDB Atlas el almacenamiento vectorial. Para obtener más información, consulte Introducción a la integración de la base de conocimientos de Amazon Bedrock en la MongoDB documentación.

Para obtener más información sobre cómo utilizar la búsqueda MongoDB Atlas vectorial para RAG, consulte Retrieval-Augmented Generation with, LangChain Amazon SageMaker AI JumpStart y MongoDB Atlas Semantic Search (entrada del blog).AWS El siguiente diagrama muestra la arquitectura de la solución que se detalla en esta entrada de blog.



Las ventajas de utilizar la búsqueda MongoDB Atlas vectorial son las siguientes:

- Puede utilizar su implementación actual de MongoDB Atlas para almacenar y buscar incrustaciones vectoriales.
- Puede utilizar la API de consultas para MongoDB consultar las incrustaciones vectoriales.
- Puede escalar de forma independiente la búsqueda vectorial y la base de datos.
- Las incrustaciones vectoriales se almacenan cerca de los datos de origen (documentos), lo que mejora el rendimiento de la indexación.

Weaviate

Weaviatees una popular base de datos vectorial de código abierto y baja latencia que admite tipos de medios multimodales, como texto e imágenes. La base de datos almacena objetos y vectores, lo que combina la búsqueda vectorial con el filtrado estructurado. Para obtener más información sobre el uso Weaviate de Amazon Bedrock para crear un flujo de trabajo de RAG, consulte Cree soluciones de IA generativa listas para empresas con modelos básicos de Cohere en Amazon Bedrock Weaviate y una base de datos vectorial en (entrada del blog). AWS MarketplaceAWS

Las ventajas de su uso son las siguientes: Weaviate

- Es de código abierto y está respaldado por una comunidad sólida.
- Está diseñado para búsquedas híbridas (tanto de vectores como de palabras clave).

- Puede implementarlo AWS como una oferta de software como servicio (SaaS) gestionado o como un clúster de Kubernetes.

Generadores para flujos de trabajo RAG

Los modelos de lenguaje grandes (LLMs) son modelos de aprendizaje profundo muy grandes que se entrenan previamente con grandes cantidades de datos. Son increíblemente flexibles. LLMs pueden realizar diversas tareas, como responder preguntas, resumir documentos, traducir idiomas y completar oraciones. Tienen el potencial de interrumpir la creación de contenido y la forma en que las personas utilizan los motores de búsqueda y los asistentes virtuales. Si bien no son perfectos, LLMs demuestran una notable habilidad para hacer predicciones basadas en un indicador o en un número de entradas relativamente pequeño.

LLMs son un componente fundamental de una solución RAG. En el caso de las arquitecturas RAG personalizadas, hay dos Servicios de AWS que sirven como opciones principales:

- Amazon Bedrock es un servicio totalmente gestionado que permite que las principales empresas LLMs de IA y Amazon estén disponibles para su uso a través de una API unificada.
- Amazon SageMaker AI JumpStart es un centro de aprendizaje automático que ofrece modelos básicos, algoritmos integrados y soluciones de aprendizaje automático prediseñadas. Con la SageMaker IA JumpStart, puede acceder a modelos previamente entrenados, incluidos los modelos básicos. También puede usar sus propios datos para ajustar los modelos previamente entrenados.

Amazon Bedrock

Amazon Bedrock ofrece modelos líderes del sector deAnthropic,Stability AI,Meta, Cohere AI21 LabsMistral AI, y Amazon. Para obtener una lista completa, consulte los modelos de cimentación compatibles en Amazon Bedrock. Amazon Bedrock también le permite personalizar los modelos con sus propios datos.

Puede evaluar el rendimiento del modelo para determinar cuáles son los más adecuados para su caso de uso de RAG. Puede probar los modelos más recientes y también comprobar qué capacidades y características ofrecen los mejores resultados y al mejor precio. El modelo Anthropic Claude Sonnet es una opción común para las aplicaciones RAG porque sobresale en una amplia gama de tareas y proporciona un alto grado de confiabilidad y previsibilidad.

SageMaker IA JumpStart

SageMaker JumpStart La IA proporciona modelos de código abierto previamente entrenados para una amplia gama de tipos de problemas. Puede entrenar y ajustar estos modelos de forma gradual antes de implementarlos. Puede acceder a los modelos, plantillas de soluciones y ejemplos previamente entrenados a través de la página de JumpStart inicio de SageMaker IA en Amazon SageMaker AI Studio o usar el SDK de Python para SageMaker IA.

SageMaker La IA JumpStart ofrece modelos state-of-the-art básicos para casos de uso como la redacción de contenido, la generación de código, la respuesta a preguntas, la redacción de textos publicitarios, el resumen, la clasificación, la recuperación de información y más. Utilice los JumpStart modelos básicos para crear sus propias soluciones generativas de IA e integre soluciones personalizadas con funciones de IA adicionales. SageMaker Para obtener más información, consulte Introducción a Amazon SageMaker AI JumpStart.

SageMaker La IA JumpStart incorpora y mantiene modelos básicos disponibles públicamente para que pueda acceder a ellos, personalizarlos e integrarlos en sus ciclos de vida del aprendizaje automático. Para obtener más información, consulte los modelos básicos disponibles públicamente. SageMaker La IA JumpStart también incluye modelos básicos patentados de proveedores externos. Para obtener más información, consulte Modelos básicos patentados.

Elegir una opción de generación aumentada de recuperación en AWS

Las secciones de opciones de RAG totalmente gestionadas y arquitecturas de RAG personalizadas de esta guía describen varios enfoques para crear una solución de búsqueda basada en RAG. AWS En esta sección se describe cómo seleccionar entre estas opciones en función de su caso de uso. En algunas situaciones, puede que funcione más de una opción. En ese escenario, la elección depende de la facilidad de implementación, de las habilidades disponibles en la organización y de las políticas y estándares de la empresa.

Le recomendamos que considere las opciones de RAG personalizadas y totalmente gestionadas en la siguiente secuencia y que elija la primera opción que se adapte a su caso de uso:

1. Usa Amazon Q Business a menos que:
   - Este servicio no está disponible en su país Región de AWS y sus datos no se pueden mover a una región en la que estén disponibles
   - Tiene un motivo específico para personalizar el flujo de trabajo de RAG
   - Desea utilizar una base de datos vectorial existente o un LLM específico
2. Utilice las bases de conocimiento de Amazon Bedrock a menos que:
   - Tiene una base de datos vectorial que no es compatible
   - Tiene un motivo específico para personalizar el flujo de trabajo de RAG
3. Combine Amazon Kendra con el generador que elija, a menos que:
   - ¿Desea elegir su propia base de datos vectorial
   - ¿Quieres personalizar la estrategia de fragmentación
4. Si quieres tener más control sobre el recuperador y quieres seleccionar tu propia base de datos vectoriales:
   - Si no tienes una base de datos vectorial existente y no necesitas consultas de gráficos o de baja latencia, considera usar Amazon OpenSearch Service.
   - Si ya tiene una base de datos PostgreSQL vectorial, considere la posibilidad de utilizar Amazon Aurora PostgreSQL and option pgvector.
   - Si necesita una latencia baja, considere una opción en memoria, como Amazon MemoryDB o Amazon DocumentDB.

- Si desea combinar la búsqueda vectorial con una consulta de gráficos, considere Amazon Neptune Analytics.
- Si ya utiliza una base de datos vectorial de terceros o encuentra alguna ventaja específica en una, considere la opción PineconeMongoDB Atlas, y Weaviate.

5. Si quieres elegir un LLM:
   - Si utilizas Amazon Q Business, no puedes elegir el LLM.
   - Si utiliza Amazon Bedrock, puede elegir uno de los modelos de base compatibles.
   - Si usa Amazon Kendra o una base de datos vectorial personalizada, puede usar uno de los generadores descritos en esta guía o usar un LLM personalizado.

   > **Note**
   >
   > También puede usar sus documentos personalizados para ajustar un LLM existente a fin de aumentar la precisión de sus respuestas. Para obtener más información, consulte la sección Comparación entre el RAG y el ajuste fino de esta guía.

6. Si ya tiene una implementación de Amazon SageMaker AI Canvas que desea utilizar o si quiere comparar las respuestas de RAG de diferentes tipos LLMs, considere Amazon SageMaker AI Canvas.

Conclusión

Esta guía describe las distintas opciones para crear un sistema de generación aumentada de recuperación (RAG). AWS Puede empezar con servicios totalmente gestionados, como las bases de conocimiento de Amazon Q Business y Amazon Bedrock. Si quieres tener más control sobre el flujo de trabajo de RAG, puedes elegir un recuperador personalizado. En el caso de un generador, puede utilizar una API para llamar a un LLM compatible en Amazon Bedrock o puede implementar su propio LLM mediante Amazon AI. SageMaker JumpStart Consulte las recomendaciones de la sección Cómo elegir una opción de RAG para determinar qué opción es la más adecuada para su caso de uso. Tras seleccionar la mejor opción para su caso de uso, utilice las referencias que se proporcionan en esta guía para empezar a crear su aplicación basada en RAG.

Historial de documentos

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

Cambio	Descripción	Fecha
Publicación inicial	—	28 de octubre de 2024

AWS Glosario de orientación prescriptiva

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

Números

Las 7 R

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

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

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

A

A2A () Agent-to-Agent

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

ABAC

Consulte control de acceso basado en atributos.

servicios abstractos

Consulte servicios administrados.

ACID

Consulte atomicidad, consistencia, aislamiento, durabilidad.

migración activa-activa

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

migración activa-pasiva

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

Agente

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

Agent Ops

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

función de agregación

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

IA

Consulte inteligencia artificial.

AIOps

Consulte operaciones de inteligencia artificial

anonimización

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

antipatrones

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

control de aplicaciones

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

cartera de aplicaciones

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

inteligencia artificial (IA)

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

operaciones de inteligencia artificial (AIOps)

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

cifrado asimétrico

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

atomicidad, consistencia, aislamiento, durabilidad (ACID)

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

control de acceso basado en atributos (ABAC)

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

origen de datos fidedigno

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

Zona de disponibilidad

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

AWS Marco de adopción de la nube (AWS CAF)

Un marco de directrices y mejores prácticas AWS para ayudar a las organizaciones a desarrollar un plan eficiente y eficaz para migrar con éxito a la nube. AWS CAF organiza la orientación en seis áreas de enfoque denominadas perspectivas: negocios, personas, gobierno, plataforma, seguridad y operaciones. Las perspectivas empresariales, humanas y de gobernanza se centran en las habilidades y los procesos empresariales; las perspectivas de plataforma, seguridad y

operaciones se centran en las habilidades y los procesos técnicos. Por ejemplo, la perspectiva humana se dirige a las partes interesadas que se ocupan de los Recursos Humanos (RR. HH.), las funciones del personal y la administración de las personas. Desde esta perspectiva, AWS CAF proporciona orientación para el desarrollo, la formación y la comunicación de las personas a fin de preparar a la organización para una adopción exitosa de la nube. Para obtener más información, consulte la Página web de AWS CAF y el Documento técnico de AWS CAF.

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

Herramienta que evalúa las cargas de trabajo de migración de bases de datos, recomienda estrategias de migración y proporciona estimaciones de trabajo. AWS WQF se incluye con AWS Schema Conversion Tool ().AWS SCT Analiza los esquemas de bases de datos y los objetos de código, el código de las aplicaciones, las dependencias y las características de rendimiento y proporciona informes de evaluación.

B

bot malicioso

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

BCP

Consulte planificación de la continuidad del negocio.

gráfico de comportamiento

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

sistema big-endian

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

clasificación binaria

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

filtro de floración

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

blue/green despliegue

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

bot

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

botnet

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

branch

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

acceso de emergencia

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

estrategia de implementación sobre infraestructura existente

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

caché de búfer

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

capacidad empresarial

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

planificación de la continuidad del negocio (BCP)

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

C

CAF

Consulte AWS Cloud Adoption Framework.

implementación canario

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

CCoE

Consulte Centro de excelencia en la nube.

CDC

Consulte captura de datos de cambios.

captura de datos de cambio (CDC)

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

ingeniería del caos

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

CI/CD

Consulte integración continua y entrega continua.

clasificación

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

Desarrollador ciudadano

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

cifrado del cliente

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

Centro de excelencia en la nube (CCoE)

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

computación en la nube

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

modelo operativo en la nube

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

etapas de adopción de la nube

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

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

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

CMDB

Consulte base de datos de administración de configuración.

repositorio de código

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

caché en frío

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

datos fríos

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

visión artificial (CV)

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

deriva de configuración

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

base de datos de administración de configuración (CMDB)

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

paquete de conformidad

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

integración y entrega continuas () CI/CD

El proceso de automatización de las etapas de origen, creación, prueba, puesta en escena y producción del proceso de publicación del software. CI/CD se describe comúnmente como una canalización. CI/CD puede ayudarlo a automatizar los procesos, mejorar la productividad, mejorar

la calidad del código y entregar más rápido. Para obtener más información, consulte Beneficios de la entrega continua. CD también puede significar implementación continua. Para obtener más información, consulte Entrega continua frente a implementación continua.

CV

Consulte visión artificial.

D

datos en reposo

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

clasificación de datos

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

deriva de datos

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

datos en tránsito

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

malla de datos

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

minimización de datos

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

perímetro de datos

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

preprocesamiento de datos

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

procedencia de los datos

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

titular de los datos

Persona cuyos datos se recopilan y procesan.

almacenamiento de datos

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

lenguaje de definición de datos (DDL)

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

lenguaje de manipulación de datos (DML)

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

DDL

Consulte lenguaje de definición de bases de datos.

conjunto profundo

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

aprendizaje profundo

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

defensa en profundidad

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

administrador delegado

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

Implementación

El proceso de hacer que una aplicación, características nuevas o correcciones de código se encuentren disponibles en el entorno de destino. La implementación abarca implementar cambios en una base de código y, a continuación, crear y ejecutar esa base en los entornos de la aplicación.

entorno de desarrollo

Consulte entorno.

control de detección

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

asignación de flujos de valor para el desarrollo (DVSM)

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

gemelo digital

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

tabla de dimensiones

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

desastre

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

recuperación de desastres (DR)

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

DML

Consulte lenguaje de manipulación de bases de datos.

diseño basado en el dominio

Un enfoque para desarrollar un sistema de software complejo mediante la conexión de sus componentes a dominios en evolución, o a los objetivos empresariales principales, a los que sirve cada componente. Eric Evans introdujo este concepto en su libro Domain-Driven Design: Tackling Complexity in the Heart of Software (Boston: Addison-Wesley Professional, 2003). Para

obtener información sobre cómo utilizar el diseño basado en dominios con el patrón de higos estranguladores, consulte Modernización gradual de los servicios web antiguos de ASP.NET Microsoft (ASMX) mediante contenedores y Amazon API Gateway.

DR

Consulte recuperación ante desastres.

Detección de desviaciones

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

DVSM

Consulte asignación de flujos de valor para el desarrollo.

E

EDA

Consulte análisis de datos de tipo exploratorio.

EDI

Consulte intercambio electrónico de datos.

computación en la periferia

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

intercambio electrónico de datos (EDI)

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

cifrado

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

clave de cifrado

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

endianidad

El orden en el que se almacenan los bytes en la memoria del ordenador. Big-endian los sistemas almacenan primero el byte más significativo. Little-endian los sistemas almacenan primero el byte menos significativo.

punto de conexión

Consulte punto de conexión de servicio.

servicio de punto de conexión

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

planificación de recursos empresariales (ERP)

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

cifrado de sobre

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

entorno

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

- entorno de desarrollo: instancia de una aplicación en ejecución que solo se encuentra disponible para el equipo principal responsable del mantenimiento de la aplicación. Los

entornos de desarrollo se utilizan para probar los cambios antes de promocionarlos a los entornos superiores. Este tipo de entorno a veces se denomina entorno de prueba.

- entornos inferiores: todos los entornos de desarrollo de una aplicación, como los que se utilizan para las compilaciones y pruebas iniciales.
- entorno de producción: instancia de una aplicación en ejecución a la que pueden acceder los usuarios finales. En un CI/CD proceso, el entorno de producción es el último entorno de implementación.
- entornos superiores: todos los entornos a los que pueden acceder usuarios que no sean del equipo de desarrollo principal. Esto puede incluir un entorno de producción, entornos de preproducción y entornos para las pruebas de aceptación por parte de los usuarios.

epopeya

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

ERP

Consulte planificación de recursos empresariales.

análisis de datos de tipo exploratorio (EDA)

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

F

tabla de hechos

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

Fail Fast

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

límite de aislamiento de errores

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

rama de característica

Consulte rama.

características

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

importancia de las características

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

transformación de funciones

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

peticiones con pocos pasos

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

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

FGAC

Consulte control de acceso detallado.

control de acceso preciso (FGAC)

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

migración relámpago

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

FM

Consulte modelo fundacional.

Modelo fundacional (FM)

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

Puerta de enlace FM

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

G

IA generativa

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

bloqueo geográfico

Consulte restricciones geográficas.

restricciones geográficas (bloqueo geográfico)

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

Flujo de trabajo de Gitflow

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

imagen dorada

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

estrategia de implementación desde cero

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

barrera de protección

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

barandas (AI)

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

H

HA

Consulte alta disponibilidad.

migración heterogénea de bases de datos

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

alta disponibilidad (HA)

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

modernización histórica

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

datos de reserva

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

human-in-the-loop (HiTL)

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

migración homogénea de bases de datos

Migración de la base de datos de origen a una base de datos de destino que comparte el mismo motor de base de datos (por ejemplo, Microsoft SQL Server a Amazon RDS para SQL Server).

La migración homogénea suele formar parte de un esfuerzo para volver a alojar o redefinir la plataforma. Puede utilizar las utilidades de bases de datos nativas para migrar el esquema.

datos recientes

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

hotfix

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

periodo de hiperatención

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

I

IaC

Consulte infraestructura como código.

políticas basadas en identidades

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

aplicación inactiva

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

IIoT

Consulte Internet de las cosas industrial.

infraestructura inmutable

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

VPC entrante (de entrada)

En una arquitectura de AWS cuentas múltiples, una VPC que acepta, inspecciona y enruta las conexiones de red desde fuera de una aplicación. La Arquitectura de referencia de seguridad de AWS recomienda configurar su cuenta de red con VPC entrantes, salientes y de inspección para proteger la interfaz bidireccional entre su aplicación e Internet en general.

migración gradual

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

Industria 4.0

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

infraestructura

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

infraestructura como código (IaC)

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

Internet de las cosas industrial (IIoT)

El uso de sensores y dispositivos conectados a Internet en los sectores industriales, como el productivo, el eléctrico, el automotriz, el sanitario, el de las ciencias de la vida y el de la

agricultura. Para obtener más información, consulte Creación de una estrategia de transformación digital del Internet de las cosas industrial (IIoT).

VPC de inspección

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

Internet de las cosas (IoT)

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

interpretabilidad

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

IoT

Consulte Internet de las cosas.

biblioteca de información de TI (ITIL)

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

administración de servicios de TI (ITSM)

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

ITIL

Consulte biblioteca de información de TI.

ITSM

Consulte administración de servicios de TI.

L

control de acceso basado en etiquetas (LBAC)

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

zona de aterrizaje

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

modelo de lenguaje de gran tamaño (LLM)

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

migración grande

Migración de 300 servidores o más.

LBAC

Consulte control de acceso basado en etiquetas.

privilegio mínimo

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

migrar mediante lift-and-shift

Consulte Las 7 R.

sistema little-endian

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

LLM

Consulte modelo de lenguaje de gran tamaño.

entornos inferiores

Consulte entorno.

M

machine learning (ML)

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

rama principal

Consulte rama.

malware

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

Servicios administrados

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

sistema de ejecución de fabricación (MES)

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

MAP

Consulte Programa de aceleración de la migración.

MCP

Consulte Model Context Protocol.

Protocolo de contexto para modelos (MCP)

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

Servidor MCP

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

mecanismo

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

cuenta de miembro

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

MES

Consulte sistema de ejecución de fabricación.

Message Queuing Telemetry Transport (MQTT)

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

microservicio

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

arquitectura de microservicios

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

Programa de aceleración de la migración (MAP)

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

migración a escala

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

fábrica de migración

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

metadatos de migración

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

patrón de migración

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

Migration Portfolio Assessment (MPA)

Herramienta en línea que proporciona información a fin de validar los argumentos comerciales necesarios para migrar a la Nube de AWS. La MPA ofrece una evaluación detallada de la cartera (adecuación del tamaño de los servidores, precios, comparaciones del costo total de propiedad, análisis de los costos de migración), así como una planificación de la migración (análisis y recopilación de datos de aplicaciones, agrupación de aplicaciones, priorización de la migración y planificación de oleadas). La herramienta MPA (requiere iniciar sesión) está disponible de forma gratuita para todos los AWS consultores y consultores de los socios de APN.

Evaluación de la preparación para la migración (MRA)

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

estrategia de migración

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

ML

Consulte machine learning.

modernización

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

evaluación de la preparación para la modernización

Evaluación que ayuda a determinar la preparación para la modernización de las aplicaciones de una organización; identifica los beneficios, los riesgos y las dependencias; y determina qué

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

aplicaciones monolíticas (monolitos)

Aplicaciones que se ejecutan como un único servicio con procesos estrechamente acoplados. Las aplicaciones monolíticas presentan varios inconvenientes. Si una característica de la aplicación experimenta un aumento en la demanda, se debe escalar toda la arquitectura. Agregar o mejorar las características de una aplicación monolítica también se vuelve más complejo a medida que crece la base de código. Para solucionar problemas con la aplicación, puede utilizar una arquitectura de microservicios. Para obtener más información, consulte Descomposición de monolitos en microservicios.

MPA

Consulte Migration Portfolio Assessment.

MQTT

Consulte Message Queuing Telemetry Transport.

clasificación multiclase

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

infraestructura mutable

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

O

OAC

Consulte control de acceso de origen.

OAI

Consulte identidad de acceso de origen.

OCM

Consulte administración del cambio organizacional.

migración fuera de línea

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

OI

Consulte integración de operaciones.

OLA

Consulte acuerdo de nivel operativo.

migración en línea

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

OPC-UA

Consulte Open Process Communications: arquitectura unificada.

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

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

acuerdo de nivel operativo (OLA)

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

revisión de la preparación operativa (ORR)

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

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

tecnología operativa (TO)

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

integración de operaciones (OI)

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

registro de seguimiento organizativo

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

administración del cambio organizacional (OCM)

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

control de acceso de origen (OAC)

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

identidad de acceso de origen (OAI)

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

ORR

Consulte revisión de la preparación operativa.

OT

Consulte tecnología operativa.

VPC saliente (de salida)

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

P

límite de permisos

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

información de identificación personal (PII)

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

PII

Consulte información de identificación personal.

manual de estrategias

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

PLC

Consulte controlador lógico programable.

PLM

Consulte administración del ciclo de vida del producto.

policy

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

persistencia políglota

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

evaluación de cartera

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

predicate

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

inserción de predicados

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

control preventivo

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

entidad principal

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

Privacidad desde el diseño

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

zonas alojadas privadas

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

control proactivo

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

administración del ciclo de vida del producto (PLM)

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

entorno de producción

Consulte entorno.

controlador lógico programable (PLC)

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

encadenamiento de peticiones

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

seudonimización

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

publish/subscribe (pub/sub)

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

Q

plan de consulta

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

regresión del plan de consulta

El optimizador de servicios de la base de datos elige un plan menos óptimo que antes de un cambio determinado en el entorno de la base de datos. Los cambios en estadísticas,

restricciones, configuración del entorno, enlaces de parámetros de consultas y actualizaciones del motor de base de datos PostgreSQL pueden provocar una regresión del plan.

R

Matriz RACI

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

RAG

Consulte generación aumentada por recuperación.

ransomware

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

Matriz RASCI

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

RCAC

Consulte control de acceso por filas y columnas.

réplica de lectura

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

rediseñar

Consulte Las 7 R.

objetivo de punto de recuperación (RPO)

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

objetivo de tiempo de recuperación (RTO)

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

refactorizar

Consulte Las 7 R.

Region

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

regresión

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

volver a alojar

Consulte Las 7 R.

versión

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

reubicar

Consulte Las 7 R.

redefinir la plataforma

Consulte Las 7 R.

recomprar

Consulte Las 7 R.

resiliencia

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

política basada en recursos

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

matriz responsable, confiable, consultada e informada (RACI)

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

control receptivo

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

retain

Consulte Las 7 R.

retirar

Consulte Las 7 R.

Generación aumentada de recuperación (RAG)

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

rotación

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

control de acceso por filas y columnas (RCAC)

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

RPO

Consulte objetivo de punto de recuperación.

RTO

Consulte objetivo de tiempo de recuperación.

manual de procedimientos

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

S

SAML 2.0

Un estándar abierto que utilizan muchos proveedores de identidad (IdPs). Esta función permite el inicio de sesión único (SSO) federado, de modo que los usuarios pueden iniciar sesión Consola de administración de AWS o llamar a las operaciones de la AWS API sin tener que crear un usuario en IAM para todos los miembros de la organización. Para obtener más información sobre la federación basada en SAML 2.0, consulte Acerca de la federación basada en SAML 2.0 en la documentación de IAM.

SCADA

Consulte control de supervisión y adquisición de datos.

SCP

Consulte política de control de servicio.

secreta

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

seguridad desde el diseño

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

control de seguridad

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

refuerzo de la seguridad

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

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

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

automatización de la respuesta de seguridad

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

cifrado del servidor

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

política de control de servicio (SCP)

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

punto de enlace de servicio

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

acuerdo de nivel de servicio (SLA)

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

indicador de nivel de servicio (SLI)

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

objetivo de nivel de servicio (SLO)

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

modelo de responsabilidad compartida

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

Shadow AI

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

SIEM

Consulte sistema de administración de eventos e información de seguridad.

único punto de error (SPOF)

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

SLA

Consulte acuerdo de nivel de servicio.

SLI

Consulte indicador de nivel de servicio.

SLO

Consulte objetivo de nivel de servicio.

modelo de dividir y sembrar

Un patrón para escalar y acelerar los proyectos de modernización. A medida que se definen las nuevas funciones y los lanzamientos de los productos, el equipo principal se divide para crear nuevos equipos de productos. Esto ayuda a ampliar las capacidades y los servicios de su organización, mejora la productividad de los desarrolladores y apoya la innovación rápida. Para más información, consulte Phased approach to modernizing applications in the Nube de AWS.

SPOF

Consulte único punto de error.

esquema en estrella

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

patrón de higo estrangulador

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

subred

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

control de supervisión y adquisición de datos (SCADA)

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

cifrado simétrico

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

pruebas sintéticas

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

petición del sistema

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

T

etiquetas

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

variable de destino

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

lista de tareas

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

entorno de prueba

Consulte entorno.

entrenamiento

Proporcionar datos de los que pueda aprender su modelo de ML. Los datos de entrenamiento deben contener la respuesta correcta. El algoritmo de aprendizaje encuentra patrones en los

datos de entrenamiento que asignan los atributos de los datos de entrada al destino (la respuesta que desea predecir). Genera un modelo de ML que captura estos patrones. Luego, el modelo de ML se puede utilizar para obtener predicciones sobre datos nuevos para los que no se conoce el destino.

herramienta

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

puerta de enlace de tránsito

Centro de tránsito de red que puede utilizar para interconectar las VPC y las redes en las instalaciones. Para obtener más información, consulte Qué es una pasarela de tránsito en la AWS Transit Gateway documentación.

flujo de trabajo basado en enlaces troncales

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

acceso de confianza

Otorgar permisos a un servicio que especifique para realizar tareas en su organización AWS Organizations y en sus cuentas en su nombre. El servicio de confianza crea un rol vinculado al servicio en cada cuenta, cuando ese rol es necesario, para realizar las tareas de administración por usted. Para obtener más información, consulte AWS Organizations Utilización con otros AWS servicios en la AWS Organizations documentación.

ajuste

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

equipo de dos pizzas

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

U

incertidumbre

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

tareas indiferenciadas

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

entornos superiores

Consulte entorno.

V

succión

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

control de versión

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

Emparejamiento de VPC

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

vulnerabilidad

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

W

caché caliente

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

datos templados

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

función de ventana

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

carga de trabajo

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

flujo de trabajo

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

WORM

Consulte escritura única y lectura múltiple.

WQF

Consulte AWS Workload Qualification Framework.

escritura única y lectura múltiple (WORM)

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

Z

ataque de día cero

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

vulnerabilidad de día cero

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

peticiones desde cero

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

aplicación zombi

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

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