Opções e arquiteturas de geração aumentada de recuperação em AWS

AWS Orientação prescritiva

AWS Orientação prescritiva: Opções e arquiteturas de geração aumentada de recuperação em AWS

Table of Contents

Opções e arquiteturas de geração aumentada de recuperação em AWS

Mithil Shah, Rajeev Muralidhar e Natacha Fort, da Amazon Web Services

Outubro de 2024 (histórico do documento)

A IA generativa se refere a um subconjunto de modelos de IA que podem criar novos conteúdos e artefatos, como imagens, vídeos, texto e áudio, a partir de uma simples solicitação de texto. Os modelos generativos de IA são treinados em grandes quantidades de dados que abrangem uma ampla variedade de assuntos e tarefas. Isso permite que eles demonstrem uma versatilidade notável na execução de várias tarefas, mesmo aquelas para as quais não foram explicitamente treinados. Devido à capacidade de um único modelo de realizar várias tarefas, esses modelos geralmente são chamados de modelos básicos (FMs).

Uma das aplicações notáveis dos modelos generativos de IA é sua proficiência em responder perguntas. No entanto, existem desafios específicos que surgem quando esses modelos são usados para responder perguntas com base em documentos personalizados. Documentos personalizados podem incluir informações proprietárias, sites internos, documentação interna, Confluence páginas, SharePoint páginas e outros. Uma opção é usar o Retrieval Augmented Generation (RAG). Com o RAG, o modelo básico faz referência a uma fonte de dados confiável que está fora de suas fontes de dados de treinamento (como seus documentos personalizados) antes de gerar uma resposta.

Este guia descreve as diferentes opções generativas de IA que estão disponíveis para responder perguntas da documentação personalizada, incluindo sistemas de geração aumentada de recuperação (RAG). Ele também fornece uma visão geral da criação de sistemas RAG na Amazon Web Services (AWS). Ao analisar as opções e arquiteturas do RAG, você pode escolher entre serviços totalmente gerenciados AWS e arquiteturas RAG personalizadas.

Público-alvo

O público-alvo deste guia são arquitetos e gerentes generativos de IA que desejam criar uma solução RAG, analisar as arquiteturas disponíveis e entender as vantagens e desvantagens de cada opção.

Objetivos

Este guia ajuda você a:

- Entenda as opções generativas de IA disponíveis para responder perguntas de documentos personalizados
- Analise as opções de arquitetura para sistemas RAG em AWS
- Entenda as vantagens e desvantagens de cada opção de RAG
- Escolha uma arquitetura RAG para seu ambiente AWS

Opções generativas de IA para consultar documentos personalizados

As organizações geralmente têm várias fontes de dados estruturados e não estruturados. Este guia se concentra em como você pode usar a IA generativa para responder perguntas de dados não estruturados.

Os dados não estruturados em sua organização podem vir de várias fontes. Podem ser arquivos de texto PDFs, wikis internos, documentos técnicos, sites públicos, bases de conhecimento ou outros. Se você quiser um modelo básico que possa responder perguntas sobre dados não estruturados, as seguintes opções estão disponíveis:

- Treine um novo modelo básico usando seus documentos personalizados e outros dados de treinamento
- Ajuste um modelo básico existente usando dados de seus documentos personalizados
- Use o aprendizado contextual para passar um documento para o modelo básico ao fazer uma pergunta
- Use uma abordagem de geração aumentada de recuperação (RAG)

Treinar do zero um novo modelo básico que inclua seus dados personalizados é uma tarefa ambiciosa. Algumas empresas fizeram isso com sucesso, como Bloomberg com seu BloombergGPTmodelo. Outro exemplo é o EXAONEmodelo multimodal deLG AI Research, que foi treinado usando 600 bilhões de obras de arte e 250 milhões de imagens de alta resolução, acompanhadas de texto. De acordo com The Cost of AI: Should You Build or Buy Your Foundation Model (LinkedIn), um modelo semelhante Meta Llama 2 custa cerca de USD $4,8 milhões para ser treinado. Existem dois pré-requisitos principais para treinar um modelo do zero: acesso a recursos (financeiros, técnicos, de tempo) e um claro retorno sobre o investimento. Se isso não parecer adequado, a próxima opção é ajustar um modelo de fundação existente.

O ajuste fino de um modelo existente envolve pegar um modelo, como Amazon Titan, Mistral ou Llama, e depois adaptar o modelo aos seus dados personalizados. Existem várias técnicas de ajuste fino, a maioria das quais envolve a modificação de apenas alguns parâmetros em vez de modificar todos os parâmetros do modelo. Isso é chamado de ajuste fino com eficiência de parâmetros. Há dois métodos principais de ajuste fino:

- O ajuste fino supervisionado usa dados rotulados e ajuda você a treinar o modelo para um novo tipo de tarefa. Por exemplo, se você quiser gerar um relatório com base em um formulário PDF, talvez seja necessário ensinar ao modelo como fazer isso fornecendo exemplos suficientes.
- O ajuste fino não supervisionado é independente da tarefa e adapta o modelo básico aos seus próprios dados. Ele treina o modelo para entender o contexto de seus documentos. Em seguida, o modelo ajustado cria conteúdo, como um relatório, usando um estilo mais personalizado para sua organização.

No entanto, o ajuste fino pode não ser ideal para casos de uso de perguntas e respostas. Para obter mais informações, consulte Comparação do RAG e ajuste fino neste guia.

Ao fazer uma pergunta, você pode transmitir a um documento o modelo básico e usar o aprendizado contextual do modelo para retornar as respostas do documento. Essa opção é adequada para consultas ad hoc de um único documento. No entanto, essa solução não funciona bem para consultar vários documentos ou para consultar sistemas e aplicativos, como o Microsoft SharePoint ou o Atlassian Confluence.

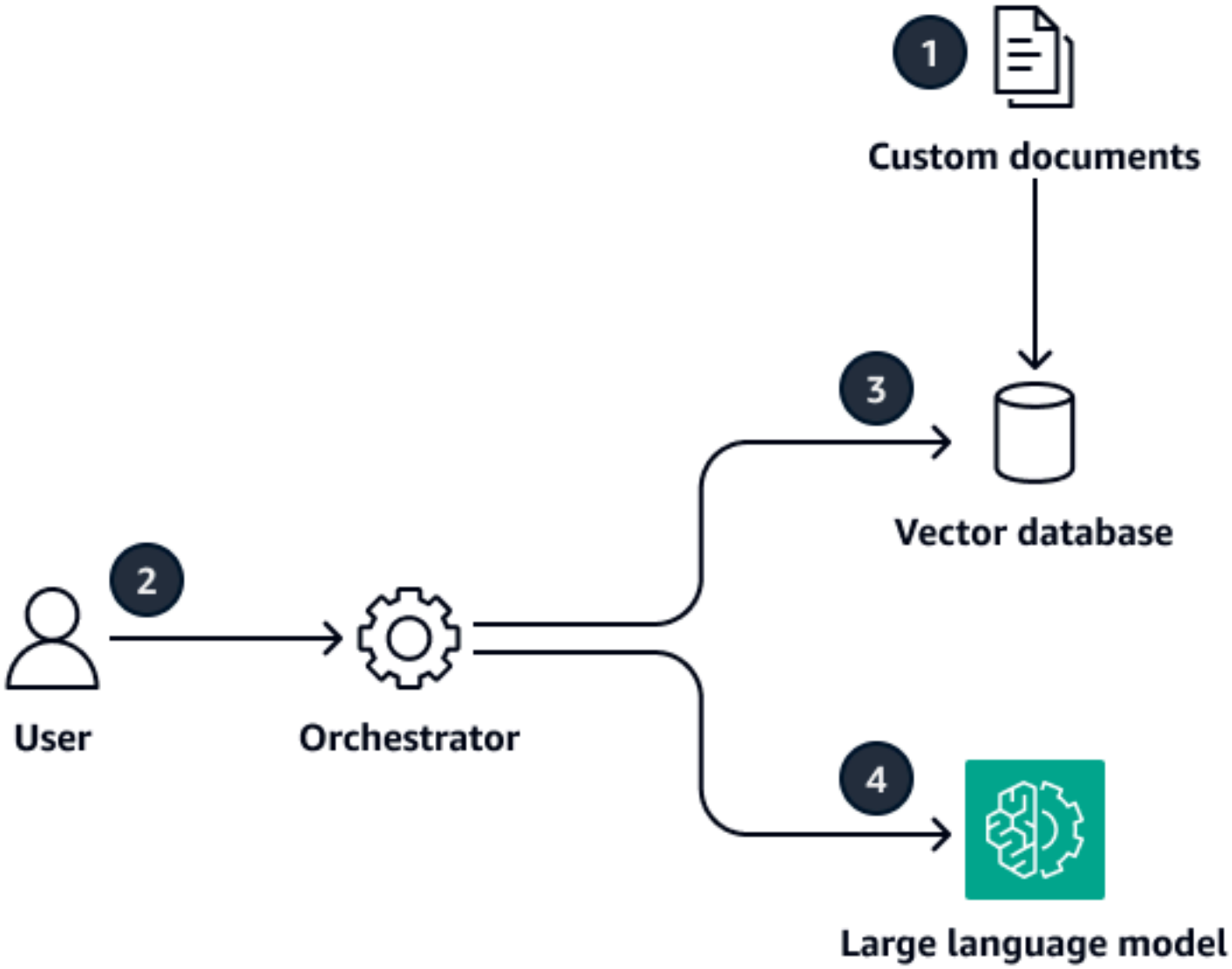
A opção final é usar o RAG. Com o RAG, o modelo básico faz referência aos seus documentos personalizados antes de gerar uma resposta. O RAG estende os recursos do modelo para a base de conhecimento interna da sua organização, tudo sem a necessidade de retreinar o modelo. É uma abordagem econômica para melhorar a saída do modelo para que ela permaneça relevante, precisa e útil em vários contextos.

Tópicos nesta seção:

- Compreendendo a geração aumentada de recuperação
- Comparando a geração aumentada de recuperação e o ajuste fino
- Casos de uso da Geração Aumentada de Recuperação

Compreendendo a geração aumentada de recuperação

A Geração Aumentada de Recuperação (RAG) é uma técnica usada para ampliar um modelo de linguagem grande (LLM) com dados externos, como documentos internos de uma empresa. Isso fornece ao modelo o contexto necessário para produzir resultados precisos e úteis para seu caso de uso específico. O RAG é uma abordagem pragmática e eficaz para uso LLMs em uma empresa. O diagrama a seguir mostra uma visão geral de alto nível de como uma abordagem RAG funciona.



De um modo geral, o processo RAG consiste em quatro etapas. A primeira etapa é feita uma vez e as outras três etapas são executadas quantas vezes forem necessárias:

1. Você cria incorporações para ingerir os documentos internos em um banco de dados vetorial. As incorporações são representações numéricas do texto nos documentos que capturam o significado semântico ou contextual dos dados. Um banco de dados vetoriais é essencialmente um banco de dados dessas incorporações e às vezes é chamado de armazenamento vetorial ou índice vetorial. Essa etapa requer limpeza, formatação e fragmentação de dados, mas essa é uma atividade única e inicial.
2. Um humano envia uma consulta em linguagem natural.

3. Um orquestrador realiza uma pesquisa por similaridade no banco de dados vetoriais e recupera os dados relevantes. O orquestrador adiciona os dados recuperados (também conhecidos como contexto) ao prompt que contém a consulta.
4. O orquestrador envia a consulta e o contexto para o LLM. O LLM gera uma resposta à consulta usando o contexto adicional.

Do ponto de vista do usuário, o RAG parece interagir com qualquer LLM. No entanto, o sistema sabe muito mais sobre o conteúdo em questão e fornece respostas ajustadas à base de conhecimento da organização.

Para obter mais informações sobre como uma abordagem RAG funciona, consulte O que é RAG no AWS site.

Componentes dos sistemas RAG em nível de produção

A criação de um sistema RAG em nível de produção exige pensar em vários aspectos diferentes do fluxo de trabalho do RAG. Conceitualmente, um fluxo de trabalho RAG em nível de produção requer os seguintes recursos e componentes, independentemente da implementação específica:

- Conectores — Eles conectam diferentes fontes de dados corporativos ao banco de dados vetoriais. Exemplos de fontes de dados estruturadas incluem bancos de dados transacionais e analíticos. Exemplos de fontes de dados não estruturadas incluem armazenamentos de objetos, bases de código e plataformas de software como serviço (SaaS). Cada fonte de dados pode exigir padrões de conectividade, licenças e configurações diferentes.
- Processamento de dados — Os dados vêm em várias formas PDFs, como imagens digitalizadas, documentos, apresentações e Microsoft SharePoint arquivos. Você deve usar técnicas de processamento de dados para extrair, processar e preparar os dados para indexação.
- Incorporações — Para realizar uma pesquisa de relevância, você deve converter seus documentos e consultas do usuário em um formato compatível. Ao usar modelos de linguagem de incorporação, você converte os documentos em representação numérica. Essas são essencialmente entradas para o modelo de fundação subjacente.
- Banco de dados vetorial — O banco de dados vetoriais é um índice das incorporações, do texto associado e dos metadados. O índice é otimizado para pesquisa e recuperação.
- Recuperador — Para a consulta do usuário, o recuperador busca o contexto relevante do banco de dados vetoriais e classifica as respostas com base nos requisitos de negócios.

- Modelo de fundação — O modelo básico para um sistema RAG é normalmente um LLM. Ao processar o contexto e a solicitação, o modelo básico gera e formata uma resposta para o usuário.
- Guardrails — Os guardrails são projetados para garantir que a consulta, o contexto imediato e recuperado e a resposta do LLM sejam precisos, responsáveis, éticos e livres de alucinações e preconceitos.
- Orquestrador — O orquestrador é responsável por programar e gerenciar o fluxo de trabalho. end-to-end
- Experiência do usuário — Normalmente, o usuário interage com uma interface de bate-papo conversacional que tem recursos avançados, incluindo a exibição do histórico do bate-papo e a coleta de feedback do usuário sobre as respostas.
- Gerenciamento de identidade e usuários — É fundamental controlar o acesso do usuário ao aplicativo com granularidade fina. No Nuvem AWS, as políticas, funções e permissões geralmente são gerenciadas por meio do AWS Identity and Access Management (IAM).

Claramente, há uma quantidade significativa de trabalho para planejar, desenvolver, lançar e gerenciar um sistema RAG. Serviços totalmente gerenciados, como o Amazon Bedrock ou o Amazon Q Business, podem ajudá-lo a gerenciar parte do trabalho pesado indiferenciado. No entanto, arquiteturas RAG personalizadas podem fornecer mais controle sobre os componentes, como o recuperador ou o banco de dados vetoriais.

Comparando a geração aumentada de recuperação e o ajuste fino

A tabela a seguir descreve as vantagens e desvantagens das abordagens de ajuste fino e baseadas em RAG.

Abordagem	Vantagens	Desvantagens
Ajuste fino	• Se um modelo aperfeiço ado for treinado usando a abordagem não supervisi onada, ele poderá criar conteúdo que correspon da melhor ao estilo da sua organização.	• O ajuste fino pode levar de algumas horas a dias, dependendo do tamanho do modelo. Portanto, não é uma boa solução se seus documentos personali zados forem alterados com frequência.

Abordagem	Vantagens	Desvantagens
	• Um modelo ajustado e treinado em dados proprietá rios ou regulatórios pode ajudar sua organização a seguir padrões de conformid ade e dados internos ou específicos do setor.	• O ajuste fino requer uma compreensão de técnicas, como adaptação de baixa classificação (LoRa) e ajuste fino com eficiência de parâmetros (PEFT). O ajuste fino pode exigir um cientista de dados. • O ajuste fino pode não estar disponível para todos os modelos. • Modelos ajustados não fornecem uma referência à fonte em suas respostas. • Pode haver um risco maior de alucinação ao usar um modelo ajustado para responder perguntas.

Abordagem	Vantagens	Desvantagens
RAG	• O RAG permite que você crie um sistema de resposta a perguntas para seus documentos personalizados sem precisar fazer ajustes. • O RAG pode incorporar os documentos mais recentes em alguns minutos. • AWS oferece soluções RAG totalmente gerenciadas. Portanto, nenhum cientista de dados ou conhecimento especializado em aprendiza do de máquina é necessári o. • Em sua resposta, um modelo RAG fornece uma referência à fonte de informações. • Como o RAG usa o contexto da pesquisa vetorial como base de sua resposta gerada, há um risco reduzido de alucinaçã o.	• O RAG não funciona bem ao resumir informações de documentos inteiros.

Se você precisar criar uma solução de resposta a perguntas que faça referência aos seus documentos personalizados, recomendamos que você comece com uma abordagem baseada em RAG. Use o ajuste fino se precisar que o modelo execute tarefas adicionais, como resumo.

Você pode combinar as abordagens de ajuste fino e RAG em um único modelo. Nesse caso, a arquitetura RAG não muda, mas o LLM que gera a resposta também é ajustado com os documentos personalizados. Isso combina o melhor dos dois mundos e pode ser a solução ideal para seu caso

de uso. Para obter mais informações sobre como combinar o ajuste fino supervisionado com o RAG, consulte a pesquisa RAFT: Adaptando o modelo de linguagem ao RAG específico do domínio, do. University of California, Berkeley

Casos de uso da Geração Aumentada de Recuperação

Veja a seguir casos de uso comuns para usar uma abordagem RAG:

- Mecanismos de pesquisa — Os mecanismos de pesquisa habilitados para RAG podem fornecer trechos mais precisos e up-to-date destacados em seus resultados de pesquisa.
- Sistemas de resposta a perguntas — O RAG pode melhorar a qualidade das respostas nos sistemas de resposta a perguntas. O modelo baseado em recuperação usa a pesquisa por similaridade para encontrar passagens ou documentos relevantes que contenham a resposta. Em seguida, ele gera uma resposta concisa e relevante com base nessas informações.
- Varejo ou comércio eletrônico — O RAG pode aprimorar a experiência do usuário no comércio eletrônico, fornecendo recomendações de produtos mais relevantes e personalizadas. Ao recuperar e incorporar informações sobre as preferências do usuário e detalhes do produto, o RAG pode gerar recomendações mais precisas e úteis para os clientes.
- Industrial ou manufatura — Na manufatura, o RAG ajuda você a acessar rapidamente informações críticas, como as operações da fábrica. Também pode ajudar nos processos de tomada de decisão, solução de problemas e inovação organizacional. Para fabricantes que operam dentro de estruturas regulatórias rigorosas, a RAG pode recuperar rapidamente os regulamentos atualizados e os padrões de conformidade de fontes internas e externas, como padrões do setor ou agências reguladoras.
- Saúde — A RAG tem potencial no setor de saúde, onde o acesso a informações precisas e oportunas é crucial. Ao recuperar e incorporar conhecimento médico relevante de fontes externas, o RAG pode fornecer respostas mais precisas e contextuais em aplicativos de saúde. Esses aplicativos aumentam as informações acessíveis por um médico humano, que, em última análise, faz a ligação e não o modelo.
- Legal — O RAG pode ser aplicado poderosamente em cenários legais, como fusões e aquisições, em que documentos jurídicos complexos fornecem contexto para consultas. Isso pode ajudar os profissionais jurídicos a lidar rapidamente com questões regulatórias complexas.

Opções de geração aumentada de recuperação totalmente gerenciadas em AWS

Para gerenciar fluxos de trabalho de Retrieval Augmented Generation (RAG) AWS, você pode usar pipelines RAG personalizados ou usar alguns dos recursos de serviços totalmente gerenciados que oferece. AWS Como incluem muitos dos principais componentes de um sistema baseado em RAG, os serviços totalmente gerenciados podem ajudá-lo a gerenciar parte do trabalho pesado indiferenciado. No entanto, esses serviços oferecem menos oportunidades de personalização.

O totalmente gerenciado Serviços da AWS usa conectores para ingerir dados de fontes de dados externas, como sites, Atlassian Confluence ou Microsoft. SharePoint As fontes de dados suportadas variam de acordo com AWS service (Serviço da AWS).

Esta seção explora as seguintes opções totalmente gerenciadas para criar fluxos de trabalho do RAG em: AWS

- Bases de Conhecimento para Amazon Bedrock
- Amazon Q Business
- Amazon SageMaker AI Canvas

Para obter mais informações sobre como escolher entre essas opções, consulte Escolhendo uma opção de geração aumentada de recuperação em AWS este guia.

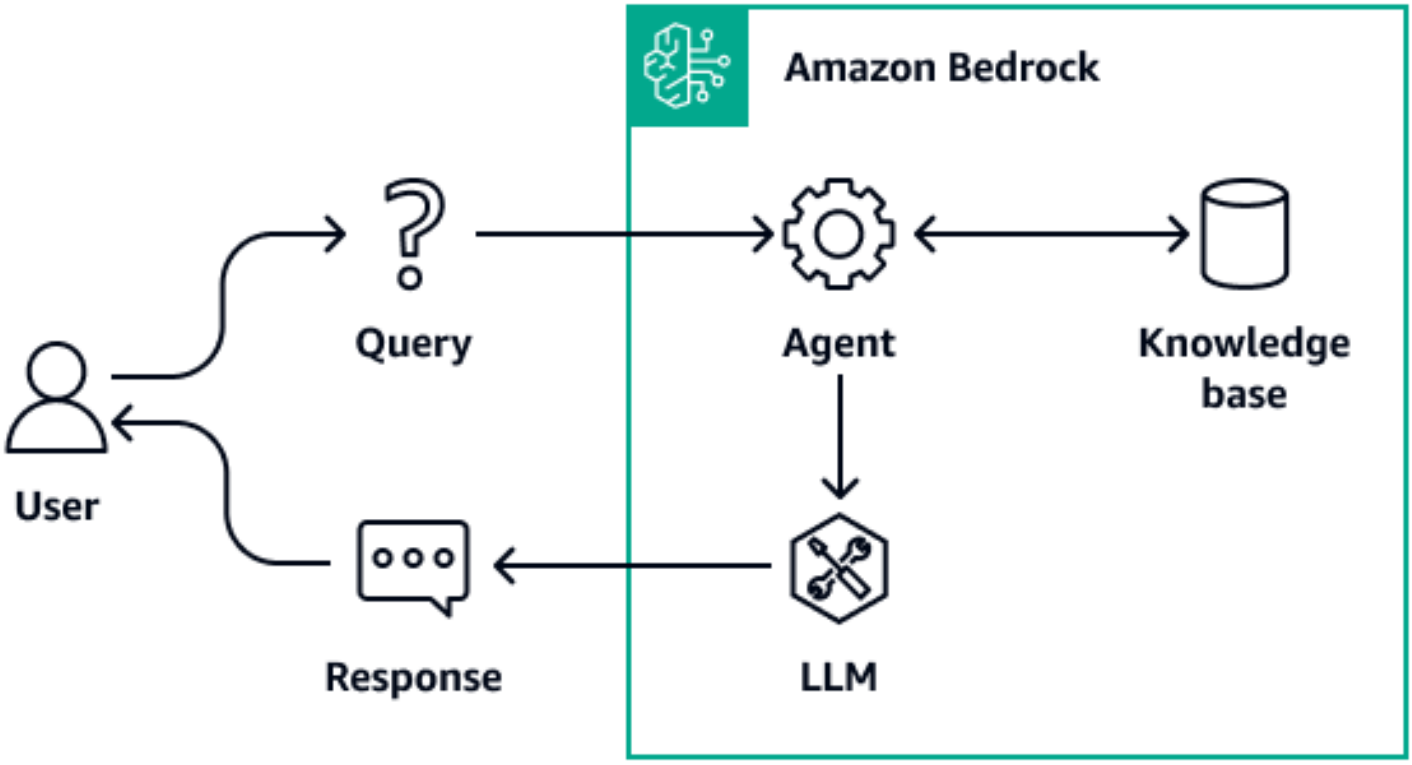
Bases de Conhecimento para Amazon Bedrock

O Amazon Bedrock é um serviço totalmente gerenciado que disponibiliza modelos básicos de alto desempenho (FMs) das principais startups de IA e da Amazon para seu uso por meio de uma API unificada. As bases de conhecimento são um recurso do Amazon Bedrock que ajuda você a implementar todo o fluxo de trabalho do RAG, desde a ingestão até a recuperação e o aumento imediato. Não há necessidade de criar integrações personalizadas com fontes de dados ou gerenciar fluxos de dados. O gerenciamento do contexto da sessão é incorporado para que seu aplicativo generativo de IA possa suportar prontamente conversas em vários turnos.

Depois de especificar a localização dos seus dados, as bases de conhecimento do Amazon Bedrock buscam internamente os documentos, os fragmentam em blocos de texto, convertem o texto

em incorporações e, em seguida, armazenam as incorporações no banco de dados vetorial de sua escolha. O Amazon Bedrock gerencia e atualiza as incorporações, mantendo o banco de dados vetoriais sincronizado com os dados. Para obter mais informações sobre como as bases de conhecimento funcionam, consulte Como funcionam as bases de conhecimento Amazon Bedrock.

Se você adicionar bases de conhecimento a um agente do Amazon Bedrock, o agente identificará a base de conhecimento apropriada com base na entrada do usuário. O agente recupera as informações relevantes e as adiciona ao prompt de entrada. O prompt atualizado fornece ao modelo mais informações de contexto para gerar uma resposta. Para melhorar a transparência e minimizar as alucinações, as informações recuperadas da base de conhecimento podem ser rastreadas até sua fonte.



O Amazon Bedrock oferece suporte aos dois seguintes APIs para RAG:

- RetrieveAndGenerate— Você pode usar essa API para consultar sua base de conhecimento e gerar respostas a partir das informações que ela recupera. Internamente, o Amazon Bedrock converte as consultas em incorporações, consulta a base de conhecimento, aumenta a solicitação com os resultados da pesquisa como informações de contexto e retorna a resposta gerada pelo LLM. O Amazon Bedrock também gerencia a memória de curto prazo da conversa para fornecer resultados mais contextuais.

- Recuperar — Você pode usar essa API para consultar sua base de conhecimento com informações recuperadas diretamente da base de conhecimento. Você pode usar as informações retornadas dessa API para processar o texto recuperado, avaliar sua relevância ou desenvolver um fluxo de trabalho separado para geração de respostas. Internamente, o Amazon Bedrock converte as consultas em incorporações, pesquisa na base de conhecimento e retorna os resultados relevantes. Você pode criar fluxos de trabalho adicionais com base nos resultados da pesquisa. Por exemplo, você pode usar o LangChain`AmazonKnowledgeBasesRetriever`plug-in para integrar fluxos de trabalho do RAG em aplicativos generativos de IA.

Para exemplos de padrões arquitetônicos e step-by-step instruções de uso do APIs, consulte O Knowledge Bases agora oferece uma experiência RAG totalmente gerenciada no Amazon Bedrock (postagem no AWS blog). Para obter mais informações sobre como usar a `RetrieveAndGenerate` API para criar um fluxo de trabalho RAG para um aplicativo inteligente baseado em bate-papo, consulte Criar um aplicativo de chatbot contextual usando o Amazon Bedrock Knowledge Bases (postagem no blog).AWS

Fontes de dados para bases de conhecimento

É possível conectar os dados proprietários a uma base de conhecimento. Depois de configurar um conector de fonte de dados, você pode sincronizar ou manter seus dados atualizados com sua base de conhecimento e disponibilizá-los para consulta. As bases de conhecimento do Amazon Bedrock oferecem suporte a conexões com as seguintes fontes de dados:

- Amazon Simple Storage Service (Amazon S3) — Você pode conectar um bucket do Amazon S3 a uma base de conhecimento do Amazon Bedrock usando o console ou a API. A base de conhecimento ingere e indexa os arquivos no bucket. Esse tipo de fonte de dados oferece suporte aos seguintes recursos:
  - Campos de metadados do documento — Você pode incluir um arquivo separado para especificar os metadados dos arquivos no bucket do Amazon S3. Em seguida, você pode usar esses campos de metadados para filtrar e melhorar a relevância das respostas.
  - Filtros de inclusão ou exclusão — você pode incluir ou excluir determinados conteúdos durante o rastreamento.
  - Sincronização incremental — As alterações do conteúdo são monitoradas e somente o conteúdo que foi alterado desde a última sincronização é rastreado.

- Confluence— Você pode conectar uma Atlassian Confluence instância a uma base de conhecimento do Amazon Bedrock usando o console ou a API. Esse tipo de fonte de dados oferece suporte aos seguintes recursos:
  - Detecção automática dos campos do documento principal — Os campos de metadados são detectados e rastreados automaticamente. Você pode usar esses campos para filtragem.
  - Filtros de conteúdo de inclusão ou exclusão — Você pode incluir ou excluir determinados conteúdos usando um prefixo ou um padrão de expressão regular no espaço, título da página, título do blog, comentário, nome do anexo ou extensão.
  - Sincronização incremental - As alterações do conteúdo são monitoradas e somente o conteúdo que foi alterado desde a última sincronização é rastreado.
  - OAuth Autenticação 2.0, autenticação com token de Confluence API — As credenciais de autenticação são armazenadas em AWS Secrets Manager.
- Microsoft SharePoint— Você pode conectar uma SharePoint instância a uma base de conhecimento usando o console ou a API. Esse tipo de fonte de dados oferece suporte aos seguintes recursos:
  - Detecção automática dos campos do documento principal — Os campos de metadados são detectados e rastreados automaticamente. Você pode usar esses campos para filtragem.
  - Filtros de conteúdo de inclusão ou exclusão — Você pode incluir ou excluir determinado conteúdo usando um prefixo ou um padrão de expressão regular no título da página principal, no nome do evento e no nome do arquivo (incluindo sua extensão).
  - Sincronização incremental - As alterações do conteúdo são monitoradas e somente o conteúdo que foi alterado desde a última sincronização é rastreado.
  - OAuth Autenticação 2.0 — As credenciais de autenticação são armazenadas em AWS Secrets Manager.
- Salesforce— Você pode conectar uma Salesforce instância a uma base de conhecimento usando o console ou a API. Esse tipo de fonte de dados oferece suporte aos seguintes recursos:
  - Detecção automática dos campos do documento principal — Os campos de metadados são detectados e rastreados automaticamente. Você pode usar esses campos para filtragem.
  - Filtros de conteúdo de inclusão ou exclusão — Você pode incluir ou excluir determinado conteúdo usando um prefixo ou um padrão de expressão regular. Para obter uma lista dos tipos de conteúdo aos quais você pode aplicar filtros, consulte Filtros de inclusão/exclusão na documentação do Amazon Bedrock.
  - Sincronização incremental — As alterações do conteúdo são monitoradas e somente o conteúdo que foi alterado desde a última sincronização é rastreado.

  - OAuth Autenticação 2.0 — As credenciais de autenticação são armazenadas em AWS Secrets Manager.
- Web Crawler — Um Amazon Bedrock Web Crawler se conecta e rastreia o que você fornece. URLs Os seguintes atributos são compatíveis:
  - Selecione vários URLs para rastrear
  - Respeite as diretivas padrão do robots.txt, como e `Allow` `Disallow`
  - Exclua URLs que correspondam a um padrão
  - Limite a taxa de rastreamento
  - Na Amazon CloudWatch, veja o status de cada URL rastreado

Para obter mais informações sobre as fontes de dados que você pode conectar à sua base de conhecimento Amazon Bedrock, consulte Criar um conector de fonte de dados para sua base de conhecimento.

Bancos de dados vetoriais para bases de conhecimento

Ao configurar uma conexão entre a base de conhecimento e a fonte de dados, você deve configurar um banco de dados vetorial, também conhecido como armazenamento vetorial. Um banco de dados vetoriais é onde o Amazon Bedrock armazena, atualiza e gerencia as incorporações que representam seus dados. Cada fonte de dados oferece suporte a diferentes tipos de banco de dados vetoriais. Para determinar quais bancos de dados vetoriais estão disponíveis para sua fonte de dados, consulte os tipos de fonte de dados.

Se você preferir que o Amazon Bedrock crie automaticamente um banco de dados vetoriais no Amazon OpenSearch Serverless para você, você pode escolher essa opção ao criar a base de conhecimento. No entanto, você também pode optar por configurar seu próprio banco de dados vetoriais. Se você configurar seu próprio banco de dados vetoriais, consulte Pré-requisitos para seu próprio armazenamento de vetores para obter uma base de conhecimento. Cada tipo de banco de dados vetorial tem seus próprios pré-requisitos.

Dependendo do tipo de fonte de dados, as bases de conhecimento do Amazon Bedrock oferecem suporte aos seguintes bancos de dados vetoriais:

- Amazon sem OpenSearch servidor
- Amazon Aurora Edição Compatível com PostgreSQL
- Pinecone (documentação da Pinecone)

- Redis Enterprise Cloud (documentação da Redis)
- MongoDB Atlas (documentação da MongoDB)

Amazon Q Business

O Amazon Q Business é um assistente totalmente gerenciado com inteligência artificial generativa que você pode configurar para responder perguntas, fornecer resumos, gerar conteúdo e concluir tarefas com base nos dados da sua empresa. Ele permite que os usuários finais recebam respostas imediatas e com reconhecimento de permissões de fontes de dados corporativas com citações.

Recursos principais do

Os seguintes recursos do Amazon Q Business podem ajudar você a criar um aplicativo de IA generativa baseado em RAG de nível de produção:

- Conectores integrados — O Amazon Q Business suporta mais de 40 tipos de conectores, como conectores paraAdobe Experience Manager (AEM),Salesforce, e. Jira Microsoft SharePoint Para obter uma lista completa, consulte Conectores compatíveis. Se precisar de um conector que não seja suportado, você pode usar AppFlow a Amazon para extrair dados da sua fonte de dados para o Amazon Simple Storage Service (Amazon S3) e, em seguida, conectar o Amazon Q Business ao bucket do Amazon S3. Para obter uma lista completa das fontes de dados AppFlow suportadas pela Amazon, consulte Aplicativos compatíveis.
- Pipelines de indexação incorporados — O Amazon Q Business fornece um pipeline integrado para indexar dados em um banco de dados vetorial. Você pode usar uma AWS Lambda função para adicionar lógica de pré-processamento ao seu pipeline de indexação.
- Opções de índice — Você pode criar e provisionar um índice nativo no Amazon Q Business e usar um Amazon Q Business retriever para extrair dados desse índice. Como alternativa, você pode usar um índice pré-configurado do Amazon Kendra como recuperador. Para obter mais informações, consulte Criação de um recuperador para um aplicativo Amazon Q Business.
- Modelos básicos — O Amazon Q Business usa os modelos básicos que são compatíveis com o Amazon Bedrock. Para obter uma lista completa, consulte Modelos de fundação compatíveis no Amazon Bedrock.
- Plug-ins — O Amazon Q Business fornece a capacidade de usar plug-ins para integração com sistemas de destino, como uma forma automatizada de resumir as informações e a criação de tíquetes emJira. Depois de configurados, os plug-ins podem comportar ações de leitura e de

gravação que podem ajudar a otimizar a produtividade do usuário final. O Amazon Q Business oferece suporte a dois tipos de plug-ins: plug-ins integrados e plug-ins personalizados.

- Guardrails — O Amazon Q Business oferece suporte a controles globais e controles em nível de tópico. Por exemplo, esses controles podem detectar informações de identificação pessoal (PII), abuso ou informações confidenciais em avisos. Para obter mais informações, consulte Controles administrativos e proteções no Amazon Q Business.
- Gerenciamento de identidade — Com o Amazon Q Business, você pode gerenciar usuários e seu acesso ao aplicativo de IA generativa baseado em RAG. Para obter mais informações, consulte Gerenciamento de identidade e acesso para o Amazon Q Business. Além disso, os conectores Amazon Q Business indexam as informações da lista de controle de acesso (ACL) anexadas a um documento junto com o próprio documento. Em seguida, o Amazon Q Business armazena as informações de ACL que indexa no Amazon Q Business User Store para criar mapeamentos de usuários e grupos e filtrar respostas de bate-papo com base no acesso do usuário final aos documentos. Para obter mais informações, consulte Conceitos do conector de fonte de dados.
- Enriquecimento de documentos — O recurso de enriquecimento de documentos ajuda você a controlar quais documentos e atributos do documento são ingeridos em seu índice e também como são ingeridos. Isso pode ser feito por meio de duas abordagens:
  - Configurar operações básicas — Use operações básicas para adicionar, atualizar ou excluir atributos do documento de seus dados. Por exemplo, você pode limpar dados de PII optando por excluir quaisquer atributos do documento relacionados às PII.
  - Configurar funções Lambda — Use uma função Lambda pré-configurada para executar uma lógica de manipulação de atributos de documentos mais personalizada e avançada em seus dados. Por exemplo, os dados empresariais podem ser armazenados como imagens digitalizadas. Nesse caso, você pode usar uma função Lambda para executar o reconhecimento óptico de caracteres (OCR) nos documentos digitalizados para extrair texto deles. Depois, cada documento digitalizado é tratado como um documento de texto durante a ingestão. Por fim, durante o bate-papo, o Amazon Q fatorará os dados textuais extraídos dos documentos digitalizados ao gerar respostas.

  Ao implementar sua solução, você pode optar por combinar as duas abordagens de enriquecimento de documentos. Você pode usar operações básicas para fazer uma primeira análise de seus dados e depois usar uma função Lambda para operações mais complexas. Para obter mais informações, consulte Enriquecimento de documentos no Amazon Q Business.
- Integração — Depois de criar seu aplicativo Amazon Q Business, você pode integrá-lo a outros aplicativos, como Slack ouMicrosoft Teams. Por exemplo, consulte Implantar um Slack gateway

para o Amazon Q Business e Implantar um Microsoft Teams gateway para o Amazon Q Business (postagens AWS no blog).

Personalização do usuário final

O Amazon Q Business suporta o upload de documentos que podem não estar armazenados nas fontes de dados e no índice da sua organização. Os documentos enviados não são armazenados. Eles estão disponíveis para uso somente na conversa na qual os documentos são carregados. O Amazon Q Business oferece suporte a tipos específicos de documentos para upload. Para obter mais informações, consulte Carregar arquivos e conversar no Amazon Q Business.

O Amazon Q Business inclui um recurso de filtragem por atributo de documento. Tanto administradores quanto usuários finais podem usar esse recurso. Os administradores podem personalizar e controlar as respostas de bate-papo para usuários finais usando atributos. Por exemplo, se o tipo de fonte de dados for um atributo anexado aos documentos, poderá especificar as respostas de chat para que sejam geradas somente a partir de uma fonte de dados específica. Ou você pode permitir que os usuários finais restrinjam o escopo das respostas do chat usando os filtros de atributos que você selecionou.

Os usuários finais podem criar aplicativos Amazon Q leves e personalizados em seu ambiente mais amplo de aplicativos Amazon Q Business. Os aplicativos Amazon Q permitem a automação de tarefas para um domínio específico, como um aplicativo criado especificamente para a equipe de marketing.

Amazon SageMaker AI Canvas

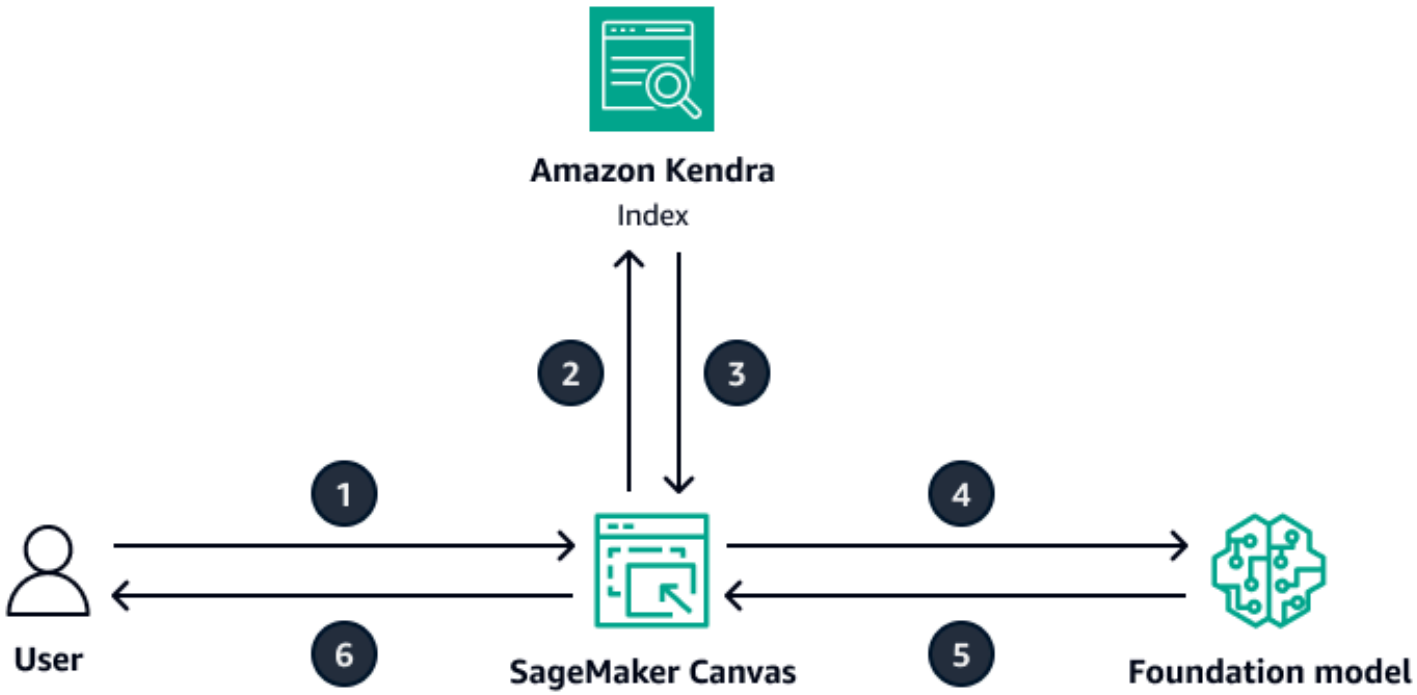
O Amazon SageMaker AI Canvas ajuda você a usar o aprendizado de máquina para gerar previsões sem precisar escrever nenhum código. Ele fornece uma interface visual sem código que permite preparar dados, criar e implantar modelos de ML, simplificando o ciclo de vida do end-to-end ML em um ambiente unificado. As complexidades da preparação de dados, desenvolvimento de modelos, detecção de viés, explicabilidade e monitoramento são resumidas por trás de uma interface intuitiva. Os usuários não precisam ser especialistas em SageMaker IA ou operações de aprendizado de máquina (MLOps) para desenvolver, operacionalizar e monitorar modelos com o SageMaker AI Canvas.

Com o SageMaker AI Canvas, a funcionalidade RAG é fornecida por meio de um recurso de consulta de documentos sem código. Você pode enriquecer a experiência de bate-papo no SageMaker AI

Canvas usando um índice da Amazon Kendra como a pesquisa corporativa subjacente. Para obter mais informações, consulte Extrair informações de documentos com a consulta de documentos.

Conectar o SageMaker AI Canvas ao índice Amazon Kendra requer uma configuração única. Como parte da configuração do domínio, um administrador de nuvem pode escolher um ou mais índices Kendra que o usuário pode consultar ao interagir com o Canvas. SageMaker Para obter instruções sobre como ativar o recurso de consulta de documentos, consulte Introdução ao uso do Amazon SageMaker AI Canvas.

SageMaker O AI Canvas gerencia a comunicação subjacente entre a Amazon Kendra e o modelo de fundação selecionado. Para obter mais informações sobre os modelos básicos que o SageMaker AI Canvas suporta, consulte Modelos básicos de IA generativos no SageMaker AI Canvas. O diagrama a seguir mostra como o recurso de consulta de documentos funciona depois que o administrador da nuvem conecta o SageMaker AI Canvas a um índice da Amazon Kendra.



O diagrama mostra o seguinte fluxo de trabalho:

1. O usuário inicia um novo bate-papo no SageMaker AI Canvas, ativa os documentos do Query, seleciona o índice de destino e, em seguida, envia uma pergunta.
2. SageMaker O AI Canvas usa a consulta para pesquisar dados relevantes no índice Amazon Kendra.
3. SageMaker O AI Canvas recupera os dados e suas fontes do índice Amazon Kendra.

4. SageMaker O AI Canvas atualiza a solicitação para incluir o contexto recuperado do índice Amazon Kendra e envia a solicitação para o modelo básico.
5. O modelo básico usa a pergunta original e o contexto recuperado para gerar uma resposta.
6. SageMaker O AI Canvas fornece a resposta gerada ao usuário. Ela inclui referências às fontes de dados, como documentos, que foram usadas para gerar a resposta.

Arquiteturas de geração aumentada de recuperação personalizada em AWS

A seção anterior descreve como usar um RAG (Geração Aumentada AWS service (Serviço da AWS) de Recuperação) totalmente gerenciado. No entanto, alguns casos de uso exigem mais controle sobre os componentes do sistema, como o recuperador ou o LLM (também chamado de gerador). Por exemplo, talvez você precise da flexibilidade de escolher seu próprio banco de dados vetorial ou acessar uma fonte de dados sem suporte. Para esses casos de uso, você pode criar uma arquitetura RAG personalizada.

Esta seção contém os seguintes tópicos:

- Recuperadores para fluxos de trabalho do RAG
- Geradores para fluxos de trabalho do RAG

Para obter mais informações sobre como escolher entre as opções de recuperador e gerador nesta seção, consulte Escolhendo uma opção de geração aumentada de recuperação em AWS este guia.

Recuperadores para fluxos de trabalho do RAG

Esta seção explica como criar um retriever. Você pode usar uma solução de pesquisa semântica totalmente gerenciada, como o Amazon Kendra, ou criar uma pesquisa semântica personalizada usando um banco de dados vetoriais. AWS

Antes de analisar as opções do recuperador, certifique-se de compreender as três etapas do processo de pesquisa vetorial:

1. Você separa os documentos que precisam ser indexados em partes menores. Isso é chamado de fragmentação.
2. Você usa um processo chamado incorporação para converter cada fragmento em um vetor matemático. Em seguida, você indexa cada vetor em um banco de dados vetoriais. A abordagem usada para indexar os documentos influencia a velocidade e a precisão da pesquisa. A abordagem de indexação depende do banco de dados vetoriais e das opções de configuração que ele fornece.
3. Você converte a consulta do usuário em um vetor usando o mesmo processo. O recuperador pesquisa no banco de dados vetoriais por vetores semelhantes ao vetor de consulta do usuário.

A similaridade é calculada usando métricas como distância euclidiana, distância do cosseno ou produto escalar.

Este guia descreve como usar os serviços a seguir Serviços da AWS ou de terceiros para criar uma camada de recuperação personalizada em AWS:

- Amazon Kendra
- OpenSearch Serviço Amazon
- Amazon Aurora PostgreSQL e pgvector
- Amazon Neptune Analytics
- Amazon MemoryDB
- Amazon DocumentDB
- Pinecone
- MongoDB Atlas
- Weaviate

Amazon Kendra

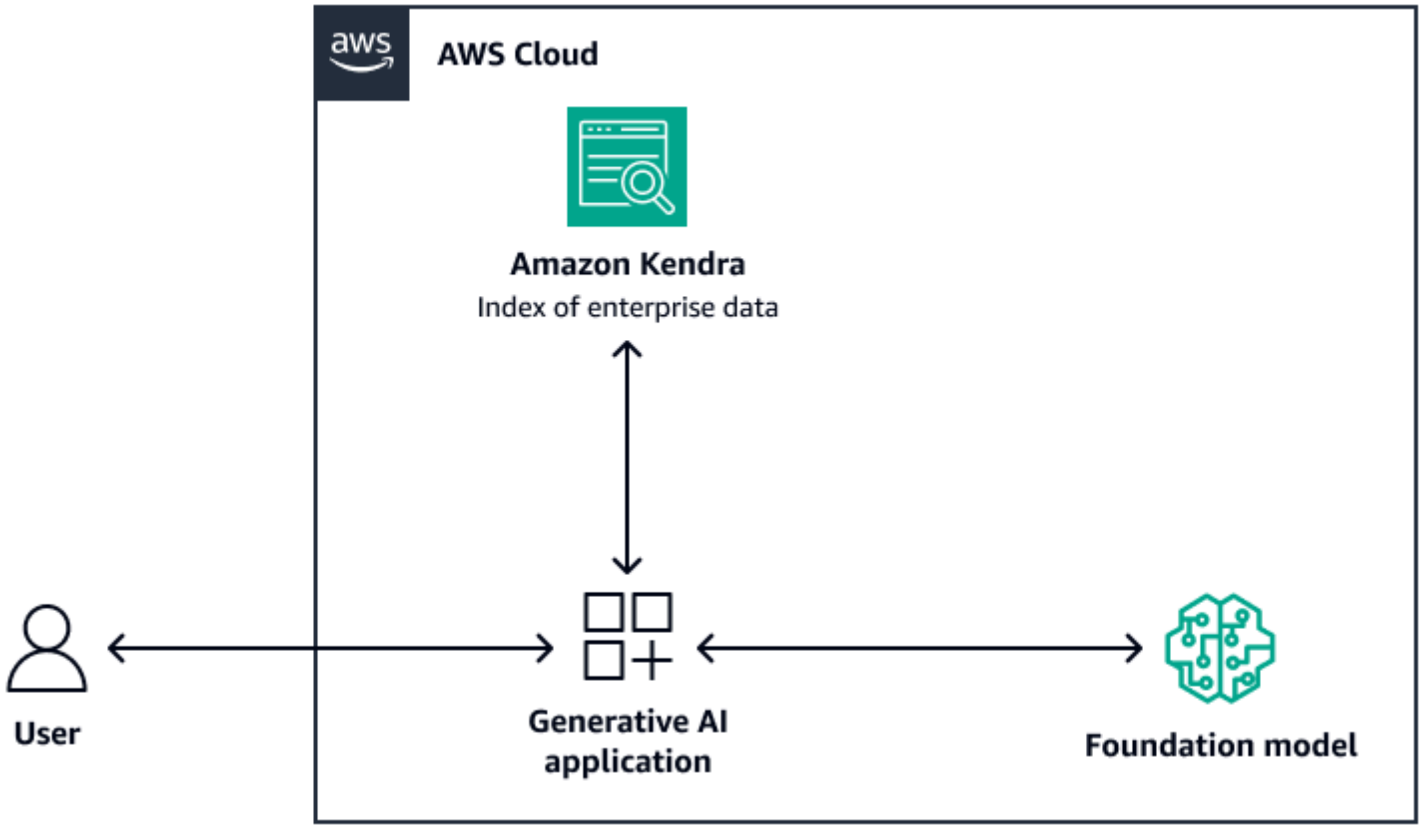
O Amazon Kendra é um serviço de pesquisa inteligente e totalmente gerenciado que usa processamento de linguagem natural e algoritmos avançados de aprendizado de máquina para retornar respostas específicas às perguntas de pesquisa de seus dados. O Amazon Kendra ajuda você a ingerir documentos diretamente de várias fontes e consultá-los depois de serem sincronizados com sucesso. O processo de sincronização cria a infraestrutura necessária para criar uma pesquisa vetorial no documento ingerido. Portanto, o Amazon Kendra não exige as três etapas tradicionais do processo de busca vetorial. Após a sincronização inicial, você pode usar um cronograma definido para lidar com a ingestão contínua.

A seguir estão as vantagens de usar o Amazon Kendra para RAG:

- Você não precisa manter um banco de dados vetoriais porque o Amazon Kendra gerencia todo o processo de pesquisa vetorial.
- O Amazon Kendra contém conectores pré-criados para fontes de dados populares, como bancos de dados, rastreadores de sites, buckets, instâncias e instâncias do Amazon S3. Microsoft SharePoint Atlassian Confluence Conectores desenvolvidos por AWS parceiros estão disponíveis, como conectores para e. Box GitLab

- O Amazon Kendra fornece filtragem de lista de controle de acesso (ACL) que retorna somente documentos aos quais o usuário final tem acesso.
- O Amazon Kendra pode impulsionar as respostas com base em metadados, como data ou repositório de origem.

A imagem a seguir mostra um exemplo de arquitetura que usa o Amazon Kendra como a camada de recuperação do sistema RAG. Para obter mais informações, consulte Crie rapidamente aplicativos de IA generativa de alta precisão em dados corporativos usando Amazon Kendra LangChain e grandes modelos de linguagem (postagem no blog).AWS



Para o modelo básico, você pode usar o Amazon Bedrock ou um LLM implantado por meio do Amazon AI. SageMaker JumpStart Você pode usar AWS Lambda with LangChainpara orquestrar o fluxo entre o usuário, a Amazon Kendra e o LLM. Para criar um sistema RAG que usa o Amazon LangChain Kendra LLMs e vários outros, consulte o repositório Amazon LangChain Kendra Extensions. GitHub

OpenSearch Serviço Amazon

O Amazon OpenSearch Service fornece algoritmos de ML integrados para pesquisas de k vizinhos mais próximos (k-NN) a fim de realizar uma pesquisa vetorial. OpenSearch O serviço também fornece um mecanismo vetorial para o Amazon EMR Serverless. Você pode usar esse mecanismo vetorial para criar um sistema RAG que tenha recursos de pesquisa e armazenamento vetorial escaláveis e de alto desempenho. Para obter mais informações sobre como criar um sistema RAG usando o OpenSearch Serverless, consulte Criar fluxos de trabalho RAG escaláveis e sem servidor com um mecanismo vetorial para os modelos Amazon Serverless e Amazon Bedrock Claude (OpenSearch postagem no blog).AWS

A seguir estão as vantagens de usar o OpenSearch Service para pesquisa vetorial:

- Ele fornece controle total sobre o banco de dados vetoriais, incluindo a criação de uma pesquisa vetorial escalável usando o OpenSearch Serverless.
- Ele fornece controle sobre a estratégia de fragmentação.
- Ele usa algoritmos de vizinho mais próximo aproximado (ANN) das bibliotecas Non-Metric Space Library (NMSLIB), Faiss e Apache Lucene para potencializar uma pesquisa k-NN. Você pode alterar o algoritmo com base no caso de uso. Para obter mais informações sobre as opções para personalizar a pesquisa vetorial por meio do OpenSearch Service, consulte Explicação sobre os recursos do banco de dados vetoriais do Amazon OpenSearch Service (postagem AWS no blog).
- OpenSearch O Serverless se integra às bases de conhecimento do Amazon Bedrock como um índice vetorial.

Amazon Aurora PostgreSQL e pgvector

A edição compatível com o Amazon Aurora PostgreSQL é um mecanismo de banco de dados relacional totalmente gerenciado que ajuda você a configurar, operar e escalar implantações do PostgreSQL. pgvector é uma extensão PostgreSQL de código aberto que fornece recursos de pesquisa por similaridade vetorial. Essa extensão está disponível tanto para o Aurora PostgreSQL compatível quanto para o Amazon Relational Database Service (Amazon RDS) para PostgreSQL. Para obter mais informações sobre como criar um sistema baseado em RAG que usa o Aurora PostgreSQL e o pgvector, consulte as seguintes postagens no blog: AWS

- Criando pesquisas com inteligência artificial no PostgreSQL usando Amazon AI e pgvector SageMaker

- Utilize o pgvector e o Amazon Aurora PostgreSQL para processamento de linguagem natural, chatbots e análise de sentimentos

Veja a seguir as vantagens de usar pgvector e Aurora PostgreSQL compatíveis:

- Ele suporta a pesquisa exata e aproximada do vizinho mais próximo. Ele também suporta as seguintes métricas de similaridade: distância L2, produto interno e distância do cosseno.
- Ele suporta arquivo invertido com compressão plana (IVFFlat) e indexação hierárquica de mundos pequenos navegáveis (HNSW).
- Você pode combinar a pesquisa vetorial com consultas sobre dados específicos do domínio que estão disponíveis na mesma instância do PostgreSQL.
- O Aurora compatível com PostgreSQL é otimizado e fornece armazenamento em cache em camadas. I/O Para cargas de trabalho que excedem a memória de instância disponível, o pgvector pode aumentar as consultas por segundo para pesquisa vetorial em até 8 vezes.

Amazon Neptune Analytics

O Amazon Neptune Analytics é um mecanismo de banco de dados gráfico otimizado para memória para análise. Ele oferece suporte a uma biblioteca de algoritmos analíticos gráficos otimizados, consultas gráficas de baixa latência e recursos de pesquisa vetorial em travessias gráficas. Ele também possui pesquisa de similaridade vetorial integrada. Ele fornece um ponto final para criar um gráfico, carregar dados, invocar consultas e realizar pesquisas de similaridade vetorial. Para obter mais informações sobre como criar um sistema baseado em RAG que usa o Neptune Analytics, consulte Usando gráficos de conhecimento para criar aplicativos GraphRag com o Amazon Bedrock e o Amazon Neptune (postagem do blog).AWS

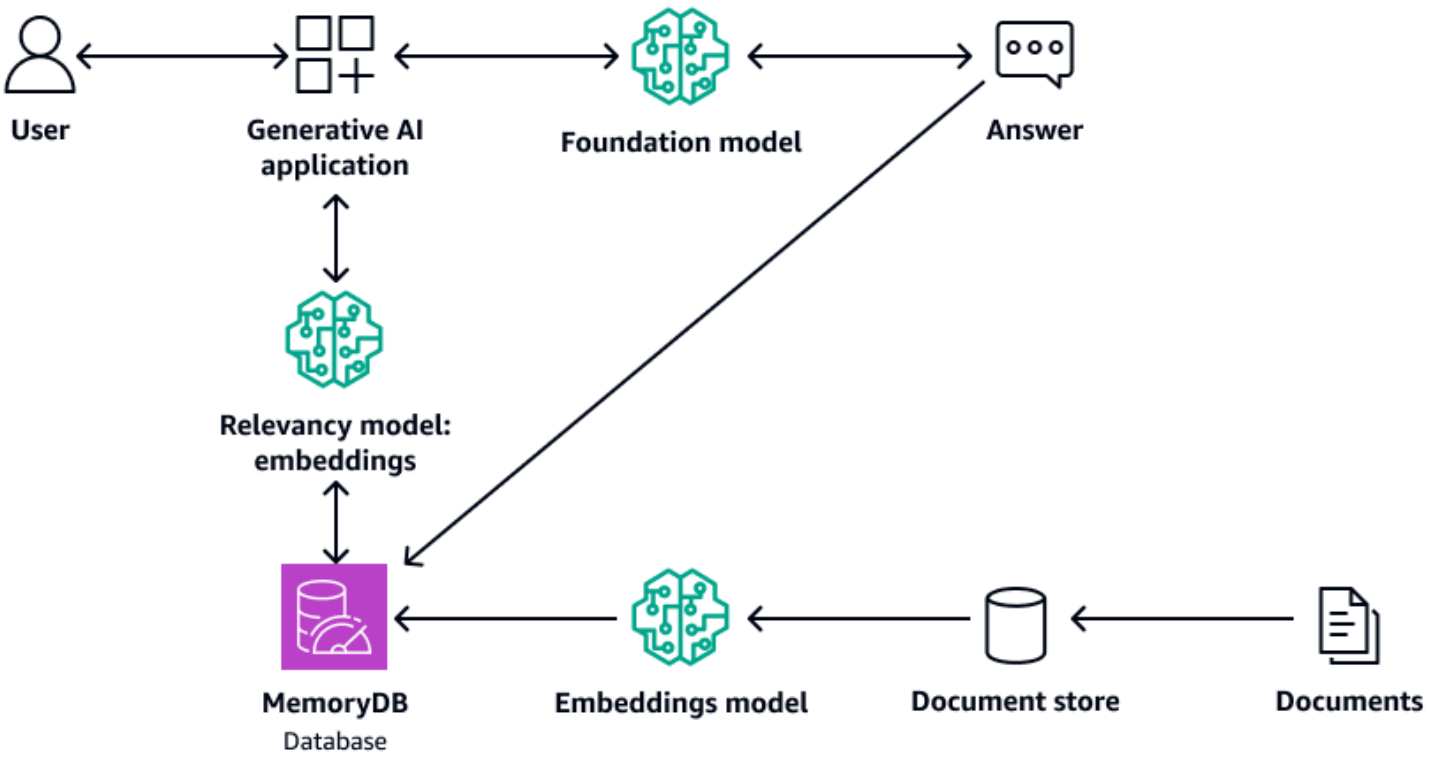
A seguir estão as vantagens de usar o Neptune Analytics:

- Você pode armazenar e pesquisar incorporações em consultas gráficas.
- Se você integrar o Neptune Analytics LangChain com, essa arquitetura oferece suporte a consultas gráficas em linguagem natural.
- Essa arquitetura armazena grandes conjuntos de dados gráficos na memória.

Amazon MemoryDB

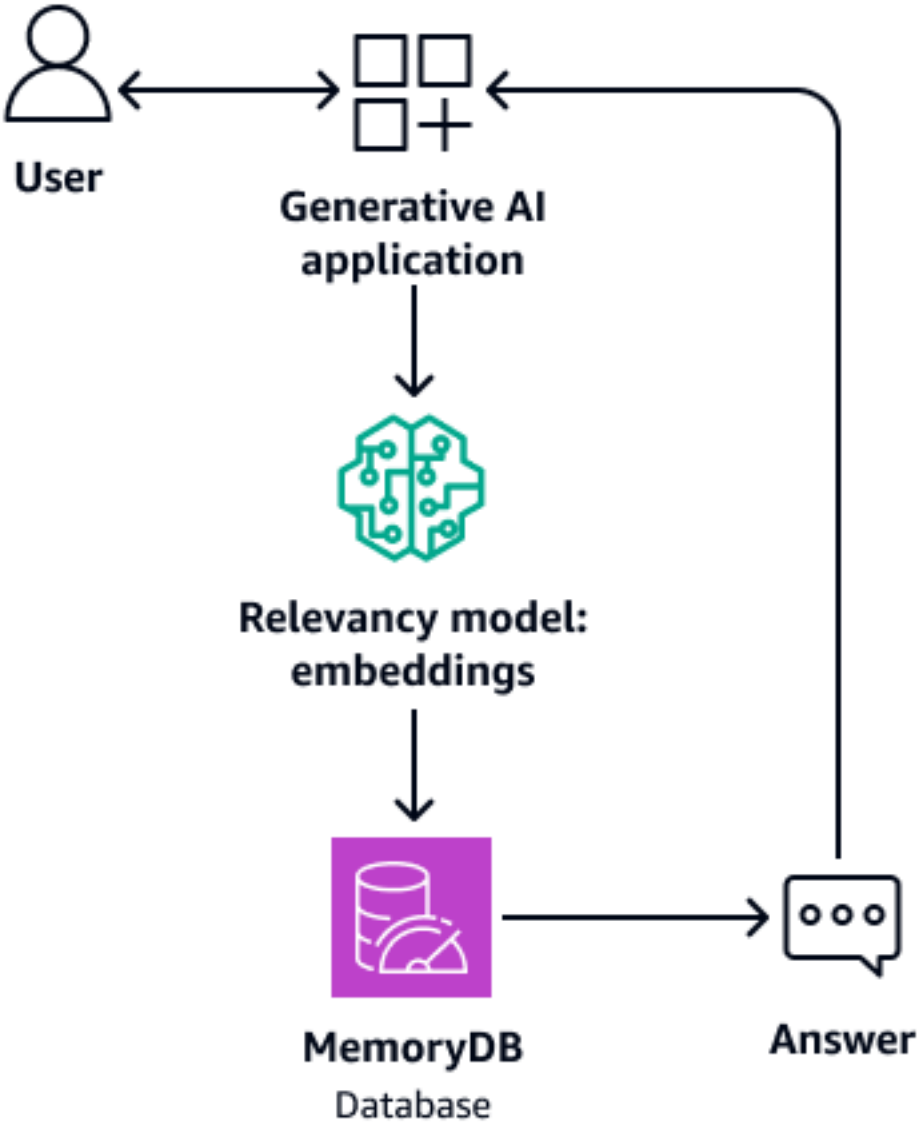
O Amazon MemoryDB é um serviço de banco de dados em memória durável que oferece desempenho ultrarrápido. Todos os seus dados são armazenados na memória, que suporta leitura em microssegundos, latência de gravação de um dígito em milissegundos e alta taxa de transferência. A pesquisa vetorial do MemoryDB amplia a funcionalidade do MemoryDB e pode ser usada em conjunto com a funcionalidade existente do MemoryDB. Para obter mais informações, consulte a Resposta de perguntas com o repositório LLM e RAG ativado. GitHub

O diagrama a seguir mostra um exemplo de arquitetura que usa o MemoryDB como banco de dados vetorial.



A seguir estão as vantagens de usar o MemoryDB:

- Ele suporta algoritmos de indexação Flat e HNSW. Para obter mais informações, consulte A pesquisa vetorial do Amazon MemoryDB agora está disponível ao público em geral no AWS blog de notícias.
- Ele também pode atuar como uma memória de buffer para o modelo básico. Isso significa que as perguntas respondidas anteriormente são recuperadas do buffer em vez de passarem pelo processo de recuperação e geração novamente. O diagrama a seguir mostra esse processo.



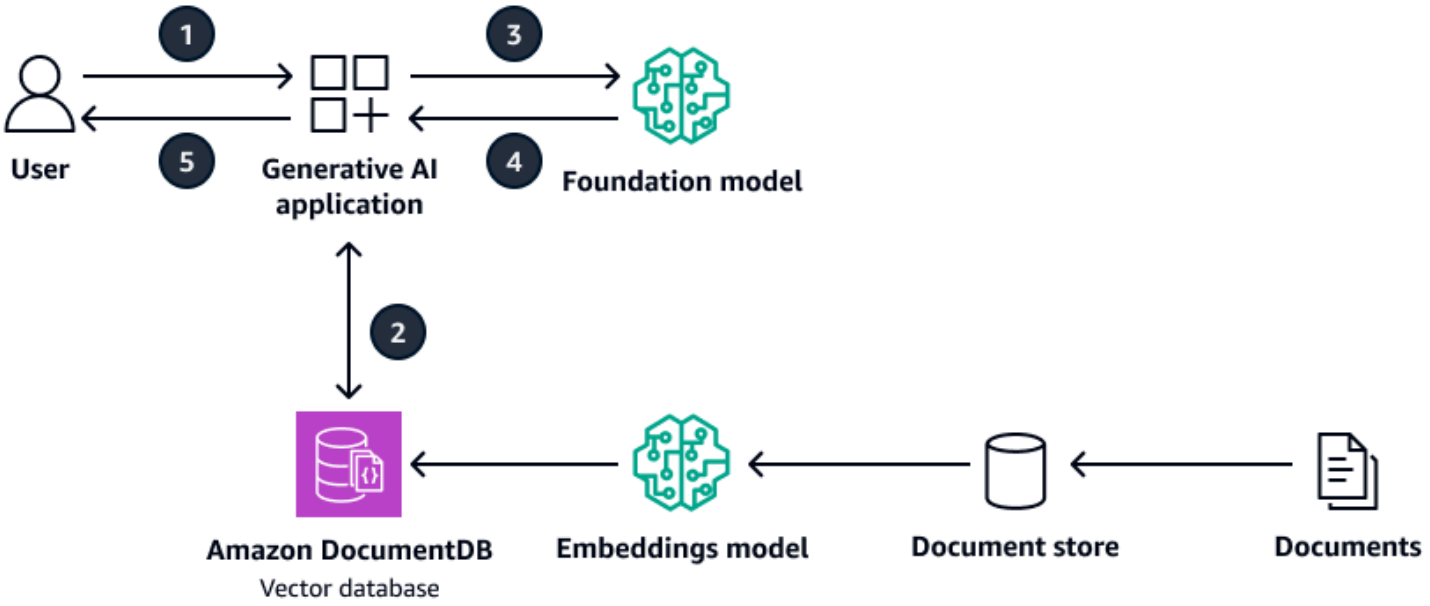
- Como usa um banco de dados na memória, essa arquitetura fornece um tempo de consulta de milissegundos de um dígito para a pesquisa semântica.
- Ele fornece até 33.000 consultas por segundo com 95— 99% de recall e 26.500 consultas por segundo com mais de 99% de recall. Para obter mais informações, consulte o vídeo AWS re:Invent 2023 - Pesquisa vetorial de latência ultrabaixa para Amazon MemoryDB em. YouTube

Amazon DocumentDB

O Amazon DocumentDB (compatível com MongoDB) é um serviço de banco de dados rápido, confiável e inteiramente gerenciado. Ele facilita a configuração, a operação e a escalabilidade de bancos MongoDB de dados compatíveis na nuvem. A pesquisa vetorial do Amazon DocumentDB

combina a flexibilidade e a rica capacidade de consulta de um banco de dados de documentos baseado em JSON com o poder da pesquisa vetorial. Para obter mais informações, consulte a Resposta de perguntas com o repositório LLM e RAG ativado. GitHub

O diagrama a seguir mostra um exemplo de arquitetura que usa o Amazon DocumentDB como banco de dados vetoriais.



O diagrama mostra o seguinte fluxo de trabalho:

1. O usuário envia uma consulta para o aplicativo generativo de IA.
2. O aplicativo generativo de IA realiza uma pesquisa por similaridade no banco de dados vetorial Amazon DocumentDB e recupera os extratos relevantes do documento.
3. O aplicativo generativo de IA atualiza a consulta do usuário com o contexto recuperado e envia a solicitação ao modelo básico de destino.
4. O modelo básico usa o contexto para gerar uma resposta à pergunta do usuário e retorna a resposta.
5. O aplicativo generativo de IA retorna a resposta ao usuário.

A seguir estão as vantagens de usar o Amazon DocumentDB:

- Ele suporta tanto o HNSW quanto os métodos de IVFFlat indexação.
- Ele suporta até 2.000 dimensões nos dados vetoriais e suporta as métricas de distância do produto euclidiano, cosseno e ponto.

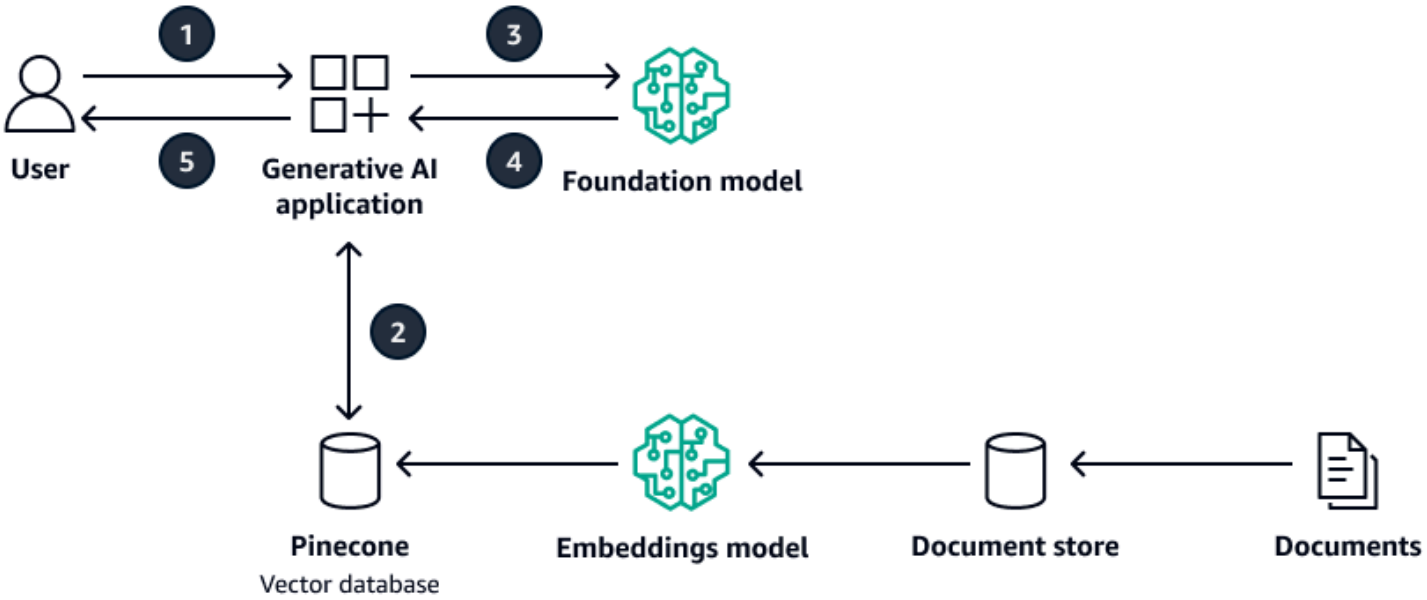
- Ele fornece tempos de resposta em milissegundos.

Pinecone

Pineconeé um banco de dados vetorial totalmente gerenciado que ajuda você a adicionar a pesquisa vetorial aos aplicativos de produção. Ele está disponível por meio do AWS Marketplace. O faturamento é baseado no uso, e as cobranças são calculadas multiplicando o preço do pod pela contagem do pod. Para obter mais informações sobre como criar um sistema baseado em RAG que usaPinecone, consulte as seguintes postagens no AWS blog:

- Mitigue alucinações por meio de RAG usando banco de dados Pinecone vetoriais e Llama-2 da Amazon AI SageMaker JumpStart
- Use o Amazon SageMaker AI Studio para criar uma solução de resposta a perguntas RAG com o Llama 2,LangChain, e Pinecone para uma rápida experimentação

O diagrama a seguir mostra uma arquitetura de exemplo usada Pinecone como banco de dados vetoriais.



O diagrama mostra o seguinte fluxo de trabalho:

1. O usuário envia uma consulta para o aplicativo generativo de IA.
2. O aplicativo generativo de IA realiza uma pesquisa por similaridade no banco de dados Pinecone vetorial e recupera os extratos relevantes do documento.

3. O aplicativo generativo de IA atualiza a consulta do usuário com o contexto recuperado e envia a solicitação ao modelo básico de destino.
4. O modelo básico usa o contexto para gerar uma resposta à pergunta do usuário e retorna a resposta.
5. O aplicativo generativo de IA retorna a resposta ao usuário.

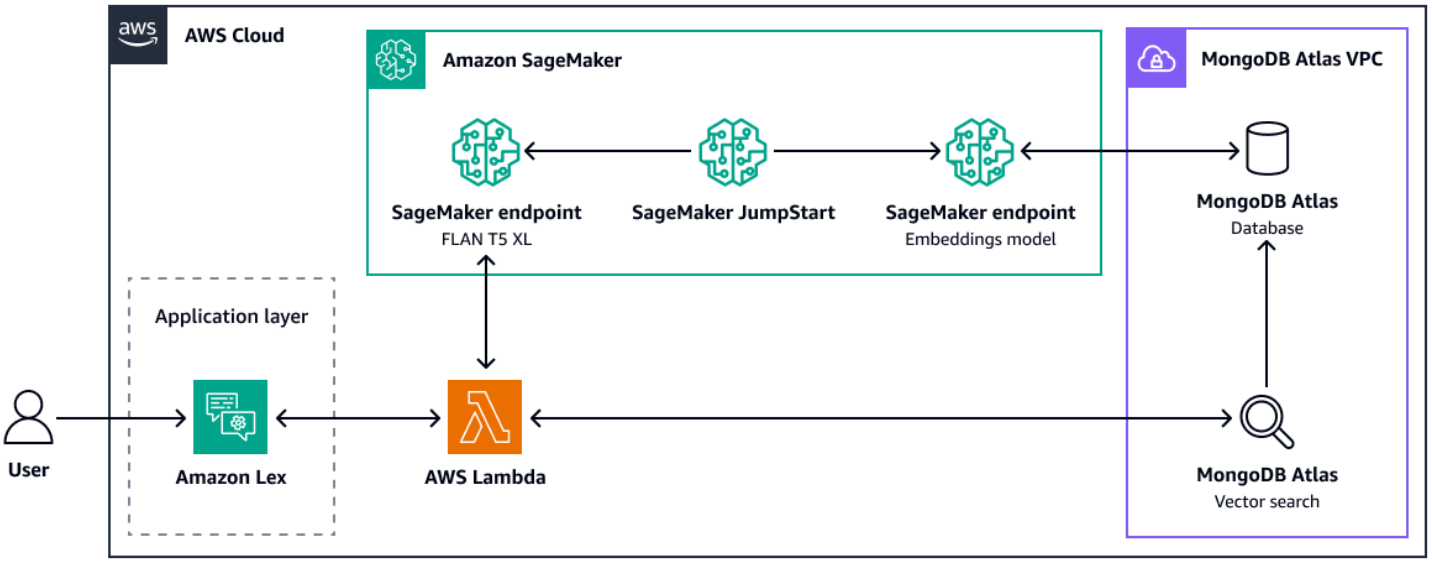
A seguir estão as vantagens de usarPinecone:

- É um banco de dados vetorial totalmente gerenciado e elimina a sobrecarga de gerenciar sua própria infraestrutura.
- Ele fornece os recursos adicionais de filtragem, atualizações dinâmicas de índices e aumento de palavras-chave (pesquisa híbrida).

MongoDB Atlas

MongoDB Atlasé um banco de dados em nuvem totalmente gerenciado que lida com toda a complexidade da implantação e do gerenciamento de suas implantações no. AWS Você pode usar a Pesquisa vetorial MongoDB Atlas para armazenar incorporações vetoriais em seu MongoDB banco de dados. As bases de conhecimento do Amazon Bedrock oferecem suporte MongoDB Atlas para armazenamento vetorial. Para obter mais informações, consulte Get Started with the Amazon Bedrock Knowledge Base Integration na MongoDB documentação.

Para obter mais informações sobre como usar a pesquisa MongoDB Atlas vetorial para RAG, consulte Retrieval-Augmented Generation with, LangChain Amazon SageMaker AI e MongoDB Atlas Semantic Search (JumpStartpostagem no blog).AWS O diagrama a seguir mostra a arquitetura da solução detalhada nesta postagem do blog.



A seguir estão as vantagens de usar a pesquisa MongoDB Atlas vetorial:

- Você pode usar sua implementação existente do MongoDB Atlas para armazenar e pesquisar incorporações vetoriais.
- Você pode usar a API de MongoDB consulta para consultar as incorporações vetoriais.
- Você pode escalar de forma independente a pesquisa vetorial e o banco de dados.
- As incorporações vetoriais são armazenadas perto dos dados de origem (documentos), o que melhora o desempenho da indexação.

Weaviate

Weaviateé um popular banco de dados vetorial de código aberto e baixa latência que oferece suporte a tipos de mídia multimodais, como texto e imagens. O banco de dados armazena objetos e vetores, o que combina pesquisa vetorial com filtragem estruturada. Para obter mais informações sobre como usar o Weaviate Amazon Bedrock para criar um fluxo de trabalho RAG, consulte Crie soluções de IA generativa prontas para empresas com modelos básicos Cohere no Amazon Bedrock Weaviate e banco de dados vetoriais em (postagem do blog). AWS MarketplaceAWS

A seguir estão as vantagens de usarWeaviate:

- É de código aberto e apoiado por uma comunidade forte.
- Ele foi criado para pesquisa híbrida (vetores e palavras-chave).

- Você pode implantá-lo AWS como uma oferta gerenciada de software como serviço (SaaS) ou como um cluster Kubernetes.

Geradores para fluxos de trabalho do RAG

Modelos de linguagem grandes (LLMs) são modelos de aprendizado profundo muito grandes que são pré-treinados em grandes quantidades de dados. Eles são incrivelmente flexíveis. LLMs pode realizar tarefas variadas, como responder perguntas, resumir documentos, traduzir idiomas e completar frases. Eles têm o potencial de interromper a criação de conteúdo e a forma como as pessoas usam mecanismos de pesquisa e assistentes virtuais. Embora não seja perfeito, LLMs demonstre uma capacidade notável de fazer previsões com base em um prompt ou número de entradas relativamente pequeno.

LLMs são um componente essencial de uma solução RAG. Para arquiteturas RAG personalizadas, há duas Serviços da AWS que servem como opções principais:

- O Amazon Bedrock é um serviço totalmente gerenciado que disponibiliza as principais empresas LLMs de IA e a Amazon para seu uso por meio de uma API unificada.
- O Amazon SageMaker AI JumpStart é um hub de ML que oferece modelos básicos, algoritmos integrados e soluções de ML pré-criadas. Com a SageMaker IA JumpStart, você pode acessar modelos pré-treinados, incluindo modelos básicos. Você também pode usar seus próprios dados para ajustar os modelos pré-treinados.

Amazon Bedrock

O Amazon Bedrock oferece modelos líderes do setor daAnthropic,,,, Stability AI Meta CohereAI21 Labs, Mistral AI e da Amazon. Para obter uma lista completa, consulte Modelos de fundação compatíveis no Amazon Bedrock. O Amazon Bedrock também permite que você personalize modelos com seus próprios dados.

Você pode avaliar o desempenho do modelo para determinar quais são os mais adequados para seu caso de uso do RAG. Você pode testar os modelos mais recentes e também para ver quais recursos oferecem os melhores resultados e pelo melhor preço. O modelo Anthropic Claude Sonnet é uma escolha comum para aplicações RAG porque se destaca em uma ampla variedade de tarefas e fornece um alto grau de confiabilidade e previsibilidade.

SageMaker AI JumpStart

SageMaker JumpStart A IA fornece modelos pré-treinados de código aberto para uma ampla variedade de tipos de problemas. Você pode treinar e ajustar esses modelos de forma incremental antes da implantação. Você pode acessar modelos pré-treinados, modelos de soluções e exemplos por meio da página JumpStart inicial de SageMaker IA no Amazon SageMaker AI Studio ou usar o SDK SageMaker AI Python.

SageMaker JumpStart A IA oferece modelos state-of-the-art básicos para casos de uso, como criação de conteúdo, geração de código, resposta a perguntas, redação, resumo, classificação, recuperação de informações e muito mais. Use modelos JumpStart básicos para criar suas próprias soluções generativas de IA e integrar soluções personalizadas com recursos adicionais de SageMaker IA. Para obter mais informações, consulte Introdução à Amazon SageMaker AI JumpStart.

SageMaker A IA JumpStart integra e mantém modelos básicos disponíveis publicamente para você acessar, personalizar e integrar em seus ciclos de vida de ML. Para obter mais informações, consulte Modelos de fundação disponíveis publicamente. SageMaker A IA JumpStart também inclui modelos básicos proprietários de fornecedores terceirizados. Para obter mais informações, consulte Modelos de fundação proprietários.

Escolhendo uma opção de geração aumentada de recuperação em AWS

As seções Opções de RAG totalmente gerenciadas e Arquiteturas de RAG personalizadas deste guia descrevem várias abordagens para criar uma solução de pesquisa baseada em RAG em. AWS Esta seção descreve como selecionar entre essas opções com base no seu caso de uso. Em algumas situações, mais de uma opção pode funcionar. Nesse cenário, a escolha depende da facilidade de implementação, das habilidades disponíveis em sua organização e das políticas e padrões de sua empresa.

Recomendamos que você considere as opções de RAG totalmente gerenciadas e personalizadas na sequência a seguir e escolha a primeira opção adequada ao seu caso de uso:

1. Use o Amazon Q Business, a menos que:
   - Este serviço não está disponível no seu Região da AWS, e seus dados não podem ser movidos para uma região onde estejam disponíveis
   - Você tem um motivo específico para personalizar o fluxo de trabalho do RAG
   - Você deseja usar um banco de dados vetorial existente ou um LLM específico
2. Use bases de conhecimento para o Amazon Bedrock, a menos que:
   - Você tem um banco de dados vetorial que não é suportado
   - Você tem um motivo específico para personalizar o fluxo de trabalho do RAG
3. Combine o Amazon Kendra com o gerador de sua escolha, a menos que:
   - Você deseja escolher seu próprio banco de dados vetoriais
   - Você deseja personalizar a estratégia de fragmentação
4. Se você quiser ter mais controle sobre o recuperador e quiser selecionar seu próprio banco de dados vetoriais:
   - Se você não tem um banco de dados vetorial existente e não precisa de consultas gráficas ou de baixa latência, considere usar o Amazon OpenSearch Service.
   - Se você já tem um banco de dados PostgreSQL vetorial, considere usar a opção Amazon Aurora PostgreSQL and. pgvector
   - Se você precisar de baixa latência, considere uma opção na memória, como Amazon MemoryDB ou Amazon DocumentDB.

- Se você quiser combinar a pesquisa vetorial com uma consulta gráfica, considere o Amazon Neptune Analytics.
- Se você já estiver usando um banco de dados vetorial de terceiros ou encontrar um benefício específico em um PineconeMongoDB Atlas, considere, Weaviatee.

5. Se você quiser escolher um LLM:
   - Se você usa o Amazon Q Business, não pode escolher o LLM.
   - Se você usa o Amazon Bedrock, pode escolher um dos modelos de fundação compatíveis.
   - Se você usa o Amazon Kendra ou um banco de dados vetorial personalizado, pode usar um dos geradores descritos neste guia ou usar um LLM personalizado.

   > **Note**
   >
   > Você também pode usar seus documentos personalizados para ajustar um LLM existente para aumentar a precisão de suas respostas. Para obter mais informações, consulte Comparando o RAG e o ajuste fino neste guia.

6. Se você tem uma implementação existente do Amazon SageMaker AI Canvas que deseja usar ou se quiser comparar respostas RAG de diferentes LLMs, considere o Amazon SageMaker AI Canvas.

Conclusão

Este guia descreve as várias opções para criar um sistema Retrieval Augmented Generation (RAG) em. AWS Você pode começar com serviços totalmente gerenciados, como as bases de conhecimento Amazon Q Business e Amazon Bedrock. Se quiser ter mais controle sobre o fluxo de trabalho do RAG, você pode escolher um recuperador personalizado. Como gerador, você pode usar uma API para chamar um LLM compatível no Amazon Bedrock ou pode implantar seu próprio LLM usando o Amazon AI. SageMaker JumpStart Analise as recomendações em Escolha de uma opção de RAG para determinar qual opção é mais adequada para seu caso de uso. Depois de selecionar a melhor opção para seu caso de uso, use as referências fornecidas neste guia para começar a criar seu aplicativo baseado em RAG.

Histórico do documento

A tabela a seguir descreve alterações significativas feitas neste guia. Se desejar receber notificações sobre futuras atualizações, inscreva-se em um feed RSS.

Alteração	Descrição	Data
Publicação inicial	—	28 de outubro de 2024

AWS Glossário de orientação prescritiva

A seguir estão os termos comumente usados em estratégias, guias e padrões fornecidos pela Orientação AWS Prescritiva. Para sugerir entradas, use o link Fornecer feedback no final do glossário.

Números

7 Rs

Sete estratégias comuns de migração para mover aplicações para a nuvem. Essas estratégias baseiam-se nos 5 Rs identificados pela Gartner em 2011 e consistem em:

- Refactor/re-architect — mova um aplicativo e modifique sua arquitetura aproveitando ao máximo os recursos nativos da nuvem para melhorar a agilidade, o desempenho e a escalabilidade. Isso normalmente envolve a portabilidade do sistema operacional e do banco de dados. Exemplo: migre seu banco de dados Oracle local para a Amazon PostgreSQL-Compatible Aurora Edition.
- Redefinir a plataforma (mover e redefinir [mover e redefinir (lift-and-reshape]):: mova uma aplicação para a nuvem e introduza algum nível de otimização a fim de aproveitar os recursos da nuvem. Exemplo: migrar seu banco de dados Oracle on-premises para o Amazon Relational Database Service (Amazon RDS) para Oracle na Nuvem AWS.
- Recomprar (drop and shop): mude para um produto diferente, normalmente migrando de uma licença tradicional para um modelo SaaS. Exemplo: Migre seu sistema de gerenciamento de relacionamento com o cliente (CRM) para o. Salesforce.com
- Redefinir a hospedagem (mover sem alterações [lift-and-shift]): mover uma aplicação para a nuvem sem fazer nenhuma alteração a fim de aproveitar os recursos da nuvem. Exemplo: migrar seu banco de dados Oracle on-premises para o Oracle em uma instância do EC2 na Nuvem AWS.
- Realocar (mover o hipervisor sem alterações [hypervisor-level lift-and-shift]): mover a infraestrutura para a nuvem sem comprar novo hardware, reescrever aplicações ou modificar suas operações existentes. Você migra servidores de uma plataforma on-premises para um serviço de nuvem para a mesma plataforma. Exemplo: Migrar um Microsoft Hyper-V aplicativo para o. AWS
- Reter (revisitar): mantenha as aplicações em seu ambiente de origem. Isso pode incluir aplicações que exigem grande refatoração, e você deseja adiar esse trabalho para um

momento posterior, e aplicações antigas que você deseja manter porque não há justificativa comercial para migrá-las.

- Retirar: desative ou remova aplicações que não são mais necessárias em seu ambiente de origem.

A

A2A () Agent-to-Agent

Um protocolo com estado para colaboração entre agentes, apoiando a delegação de tarefas e a transferência de estados.

ABAC

Consulte controle de acesso baseado em atributo.

serviços abstraídos

Veja serviços gerenciados.

ACID

Veja atomicidade, consistência, isolamento, durabilidade.

migração ativa-ativa

Um método de migração de banco de dados no qual os bancos de dados de origem e de destino são mantidos em sincronia (por meio de uma ferramenta de replicação bidirecional ou operações de gravação dupla), e ambos os bancos de dados lidam com transações de aplicações conectadas durante a migração. Esse método oferece suporte à migração em lotes pequenos e controlados, em vez de exigir uma substituição única. É mais flexível, mas exige mais trabalho do que a migração ativa-passiva.

migração ativa-passiva

Um método de migração de banco de dados em que os bancos de dados de origem e de destino são mantidos em sincronia, mas somente o banco de dados de origem manipula as transações das aplicações conectadas, enquanto os dados são replicados no banco de dados de destino. O banco de dados de destino não aceita nenhuma transação durante a migração.

Agente

Um sistema de IA que pode raciocinar, planejar e realizar ações de forma autônoma usando ferramentas para atingir metas.

Agente Ops

Práticas operacionais para criar, testar, implantar e executar agentes de IA na produção em grande escala.

AGGREGATE FUNCTION

Uma função SQL que opera em um grupo de linhas e calcula um único valor de retorno para o grupo. Exemplos de funções agregadas incluem `SUM` e `MAX`.

AI

Veja inteligência artificial.

AIOps

Veja operações de inteligência artificial.

anonimização

O processo de excluir permanentemente informações pessoais em um conjunto de dados. A anonimização pode ajudar a proteger a privacidade pessoal. Dados anônimos não são mais considerados dados pessoais.

antipadrões

Uma solução frequentemente usada para um problema recorrente em que a solução é contraproducente, ineficaz ou menos eficaz do que uma alternativa.

controle de aplicações

Uma abordagem de segurança que permite o uso somente de aplicações aprovadas para ajudar a proteger um sistema contra malware.

portfólio de aplicações

Uma coleção de informações detalhadas sobre cada aplicação usada por uma organização, incluindo o custo para criar e manter a aplicação e seu valor comercial. Essas informações são fundamentais para o processo de descoberta e análise de portfólio e ajudam a identificar e priorizar as aplicações a serem migradas, modernizadas e otimizadas.

inteligência artificial (IA)

O campo da ciência da computação que se dedica ao uso de tecnologias de computação para desempenhar funções cognitivas normalmente associadas aos humanos, como aprender, resolver problemas e reconhecer padrões. Para obter mais informações, consulte O que é inteligência artificial?

operações de inteligência artificial (AIOps)

O processo de usar técnicas de machine learning para resolver problemas operacionais, reduzir incidentes operacionais e intervenção humana e aumentar a qualidade do serviço. Para obter mais informações sobre como as AIOps são usadas na estratégia de migração para a AWS , consulte o guia de integração de operações.

criptografia assimétrica

Um algoritmo de criptografia que usa um par de chaves, uma chave pública para criptografia e uma chave privada para descriptografia. É possível compartilhar a chave pública porque ela não é usada na descriptografia, mas o acesso à chave privada deve ser altamente restrito.

atomicidade, consistência, isolamento, durabilidade (ACID)

Um conjunto de propriedades de software que garantem a validade dos dados e a confiabilidade operacional de um banco de dados, mesmo no caso de erros, falhas de energia ou outros problemas.

controle de acesso por atributo (ABAC)

A prática de criar permissões minuciosas com base nos atributos do usuário, como departamento, cargo e nome da equipe. Para obter mais informações, consulte ABAC AWS na documentação AWS Identity and Access Management (IAM).

fonte de dados autorizada

Um local onde você armazena a versão principal dos dados, que é considerada a fonte de informações mais confiável. Você pode copiar dados da fonte de dados autorizada para outros locais com o objetivo de processar ou modificar os dados, como anonimizá-los, redigi-los ou pseudonimizá-los.

Zona de disponibilidade

Um local distinto dentro de um Região da AWS que está isolado de falhas em outras zonas de disponibilidade e fornece conectividade de rede barata e de baixa latência a outras zonas de disponibilidade na mesma região.

AWS Estrutura de adoção da nuvem (AWS CAF)

Uma estrutura de diretrizes e melhores práticas AWS para ajudar as organizações a desenvolver um plano eficiente e eficaz para migrar com sucesso para a nuvem. AWS O CAF organiza a orientação em seis áreas de foco chamadas perspectivas: negócios, pessoas, governança, plataforma, segurança e operações. As perspectivas de negócios, pessoas e governança têm

como foco habilidades e processos de negócios; as perspectivas de plataforma, segurança e operações concentram-se em habilidades e processos técnicos. Por exemplo, a perspectiva das pessoas tem como alvo as partes interessadas que lidam com recursos humanos (RH), funções de pessoal e gerenciamento de pessoal. Nessa perspectiva, o AWS CAF fornece orientação para desenvolvimento, treinamento e comunicação de pessoas para ajudar a preparar a organização para a adoção bem-sucedida da nuvem. Para obter mais informações, consulte o site da AWS CAF e o whitepaper da AWS CAF.

AWS Estrutura de qualificação da carga de trabalho (AWS WQF)

Uma ferramenta que avalia as cargas de trabalho de migração do banco de dados, recomenda estratégias de migração e fornece estimativas de trabalho. AWS O WQF está incluído com AWS Schema Conversion Tool ()AWS SCT. Ela analisa esquemas de banco de dados e objetos de código, código de aplicações, dependências e características de performance, além de fornecer relatórios de avaliação.

B

bot malicioso

Um bot destinado a causar disrupção ou danos a indivíduos ou organizações.

BCP

Veja planejamento de continuidade de negócios

gráfico de comportamento

Uma visualização unificada e interativa do comportamento e das interações de recursos ao longo do tempo. É possível usar um gráfico de comportamento com o Amazon Detective para examinar tentativas de login malsucedidas, chamadas de API suspeitas e ações similares. Para obter mais informações, consulte Dados em um gráfico de comportamento na documentação do Detective.

sistema big-endian

Um sistema que armazena o byte mais significativo antes. Veja também endianness.

classificação binária

Um processo que prevê um resultado binário (uma de duas classes possíveis). Por exemplo, seu modelo de ML pode precisar prever problemas como “Este e-mail é ou não é spam?” ou “Este produto é um livro ou um carro?”

filtro de bloom

Uma estrutura de dados probabilística e eficiente em termos de memória que é usada para testar se um elemento é membro de um conjunto.

blue/green implantação

Uma estratégia de implantação em que você cria dois ambientes separados, mas idênticos. Você executa a versão atual da aplicação em um ambiente (azul) e a nova versão da aplicação no outro ambiente (verde). Essa estratégia ajuda você a reverter rapidamente com o mínimo de impacto.

bot

Uma aplicação de software que executa tarefas automatizadas na internet e simula a atividade ou interação humana. Alguns bots são úteis ou benéficos, como crawlers da web que indexam informações na internet. Outros bots, conhecidos como bots maliciosos, têm como objetivo causar disrupção ou danos a indivíduos ou organizações.

botnet

Redes de bots infectadas por malware e sob o controle de uma única parte, conhecidas como bot herder ou operador de bots. Os botnets são o mecanismo mais conhecido para escalar bots e seu impacto.

ramo

Uma área contida de um repositório de código. A primeira ramificação criada em um repositório é a ramificação principal. Você pode criar uma nova ramificação a partir de uma ramificação existente e, em seguida, desenvolver recursos ou corrigir bugs na nova ramificação. Uma ramificação que você cria para gerar um recurso é comumente chamada de ramificação de recurso. Quando o recurso estiver pronto para lançamento, você mesclará a ramificação do recurso de volta com a ramificação principal. Para obter mais informações, consulte Sobre filiais (GitHub documentação).

Acesso de emergência

Em circunstâncias excepcionais e por meio de um processo aprovado, um meio rápido para um usuário obter acesso a um Conta da AWS que ele normalmente não tem permissão para acessar. Para obter mais informações, consulte o indicador Implementar procedimentos de quebra de vidros na AWS Well-Architected orientação.

estratégia brownfield

A infraestrutura existente em seu ambiente. Ao adotar uma estratégia brownfield para uma arquitetura de sistema, você desenvolve a arquitetura de acordo com as restrições dos sistemas e da infraestrutura atuais. Se estiver expandindo a infraestrutura existente, poderá combinar as estratégias brownfield e greenfield.

cache do buffer

A área da memória em que os dados acessados com mais frequência são armazenados.

capacidade de negócios

O que uma empresa faz para gerar valor (por exemplo, vendas, atendimento ao cliente ou marketing). As arquiteturas de microsserviços e as decisões de desenvolvimento podem ser orientadas por recursos de negócios. Para obter mais informações, consulte a seção Organizados de acordo com as capacidades de negócios do whitepaper Executar microsserviços conteinerizados na AWS.

planejamento de continuidade de negócios (BCP)

Um plano que aborda o impacto potencial de um evento disruptivo, como uma migração em grande escala, nas operações e permite que uma empresa retome as operações rapidamente.

C

CAF

Veja AWS Cloud Adoption Framework.

implantação canário

O lançamento lento e incremental de uma versão para usuários finais. Quando estiver confiante, você implanta a nova versão e substitui a versão atual por completo.

CCoE

Veja Centro de Excelência da Nuvem.

CDC

Veja captura de dados de alteração.

captura de dados de alterações (CDC)

O processo de rastrear alterações em uma fonte de dados, como uma tabela de banco de dados, e registrar metadados sobre a alteração. É possível usar o CDC para várias finalidades, como auditar ou replicar alterações em um sistema de destino para manter a sincronização.

engenharia do caos

Introduzir intencionalmente falhas ou eventos disruptivos para testar a resiliência de um sistema. Você pode usar AWS Fault Injection Service (AWS FIS) para realizar experimentos que estressam suas AWS cargas de trabalho e avaliar sua resposta.

CI/CD

Veja integração e entrega contínuas.

classificação

Um processo de categorização que ajuda a gerar previsões. Os modelos de ML para problemas de classificação predizem um valor discreto. Os valores discretos são sempre diferentes uns dos outros. Por exemplo, um modelo pode precisar avaliar se há ou não um carro em uma imagem.

Desenvolvedor cidadão

Um usuário corporativo que cria aplicativos de IA usando plataformas sem code/low código sem habilidades técnicas especializadas.

criptografia no lado do cliente

Criptografia de dados localmente, antes que o alvo os AWS service (Serviço da AWS) receba.

Centro de Excelência da Nuvem (CCoE)

Uma equipe multidisciplinar que impulsiona os esforços de adoção da nuvem em toda a organização, incluindo o desenvolvimento de práticas recomendas de nuvem, a mobilização de recursos, o estabelecimento de cronogramas de migração e a liderança da organização em transformações em grande escala. Para obter mais informações, consulte as postagens do CCoE no blog de estratégia Nuvem AWS corporativa.

computação em nuvem

A tecnologia de nuvem normalmente usada para armazenamento de dados remoto e gerenciamento de dispositivos de IoT. A computação em nuvem é normalmente conectada à tecnologia de computação de borda.

modelo operacional em nuvem

Em uma organização de TI, o modelo operacional usado para criar, amadurecer e otimizar um ou mais ambientes de nuvem. Para obter mais informações, consulte Criar seu modelo operacional de nuvem.

estágios de adoção da nuvem

As quatro fases pelas quais as organizações normalmente passam ao migrar para a Nuvem AWS:

- Projeto: executar alguns projetos relacionados à nuvem para fins de prova de conceito e aprendizado
- Fundação: realizar investimentos fundamentais para escalar sua adoção da nuvem (por exemplo, criar uma zona de pouso, definir um CCoE, estabelecer um modelo de operações)
- Migração: migrar aplicações individuais
- Re-invention — Otimizando produtos e serviços e inovando na nuvem

Esses estágios foram definidos por Stephen Orban na postagem do blog The Journey Toward Cloud-First & the Stages of Adoption no blog Nuvem AWS Enterprise Strategy. Para obter informações sobre como eles se relacionam com a estratégia de AWS migração, consulte o guia de preparação para migração.

CMDB

Veja banco de dados de gerenciamento de configuração.

repositório de código

Um local onde o código-fonte e outros ativos, como documentação, amostras e scripts, são armazenados e atualizados por meio de processos de controle de versão. Os repositórios de nuvem comuns incluem o GitHub ou o Bitbucket Cloud. Cada versão do código é chamada de ramificação. Em uma estrutura de microsserviços, cada repositório é dedicado a uma única peça de funcionalidade. Um único CI/CD pipeline pode usar vários repositórios.

cache frio

Um cache de buffer que está vazio, não está bem preenchido ou contém dados obsoletos ou irrelevantes. Isso afeta a performance porque a instância do banco de dados deve ler da memória principal ou do disco, um processo que é mais lento do que a leitura do cache do buffer.

dados frios

Dados que raramente são acessados e geralmente são históricos. Ao consultar esse tipo de dados, consultas lentas geralmente são aceitáveis. Mover esses dados para níveis ou classes de armazenamento de baixo desempenho e menos caros pode reduzir os custos.

visão computacional (CV)

Um campo de IA que usa machine learning para analisar e extrair informações de formatos visuais, como vídeos e imagens digitais. Por exemplo, a Amazon SageMaker AI fornece algoritmos de processamento de imagem para CV.

desvio de configuração

Em uma workload, uma alteração de configuração em relação ao estado esperado. Isso pode fazer com que a workload se torne incompatível e, normalmente, é gradual e não intencional.

banco de dados de gerenciamento de configuração (CMDB)

Um repositório que armazena e gerencia informações sobre um banco de dados e seu ambiente de TI, incluindo componentes de hardware e software e suas configurações. Normalmente, os dados de um CMDB são usados no estágio de descoberta e análise do portfólio da migração.

pacote de conformidade

Uma coleção de AWS Config regras e ações de remediação que você pode montar para personalizar suas verificações de conformidade e segurança. Você pode implantar um pacote de conformidade como uma entidade única em uma Conta da AWS região ou em uma organização usando um modelo YAML. Para obter mais informações, consulte Pacotes de conformidade na documentação. AWS Config

integração contínua e entrega contínua (CI/CD)

O processo de automatizar os estágios de origem, criação, teste, preparação e produção do processo de lançamento do software. CI/CD é comumente descrito como um pipeline. CI/CD pode ajudá-lo a automatizar processos, melhorar a produtividade, melhorar a qualidade do código e entregar com mais rapidez. Para obter mais informações, consulte Benefícios da entrega contínua. CD também pode significar implantação contínua. Para obter mais informações, consulte Entrega contínua versus implantação contínua.

CV

Veja visão computacional.

D

dados em repouso

Dados estacionários em sua rede, por exemplo, dados que estão em um armazenamento.

classificação de dados

Um processo para identificar e categorizar os dados em sua rede com base em criticalidade e confidencialidade. É um componente crítico de qualquer estratégia de gerenciamento de riscos de segurança cibernética, pois ajuda a determinar os controles adequados de proteção e retenção para os dados. A classificação de dados é um componente do pilar de segurança na AWS Well-Architected Estrutura. Para obter mais informações, consulte Classificação de dados.

desvio de dados

Uma variação significativa entre os dados de produção e os dados usados para treinar um modelo de ML ou uma alteração significativa nos dados de entrada ao longo do tempo. O desvio de dados pode reduzir a qualidade geral, a precisão e a imparcialidade das previsões do modelo de ML.

dados em trânsito

Dados que estão se movendo ativamente pela sua rede, como entre os recursos da rede.

data mesh

Um framework de arquitetura que fornece propriedade de dados distribuída e descentralizada com gerenciamento e governança centralizados.

minimização de dados

O princípio de coletar e processar apenas os dados estritamente necessários. Praticar a minimização de dados no Nuvem AWS pode reduzir os riscos de privacidade, os custos e a pegada de carbono de sua análise.

perímetro de dados

Um conjunto de proteções preventivas em seu AWS ambiente que ajudam a garantir que somente identidades confiáveis acessem recursos confiáveis das redes esperadas. Para obter mais informações, consulte Construindo um perímetro de dados em. AWS

pré-processamento de dados

A transformação de dados brutos em um formato que seja facilmente analisado por seu modelo de ML. O pré-processamento de dados pode significar a remoção de determinadas colunas ou linhas e o tratamento de valores ausentes, inconsistentes ou duplicados.

proveniência dos dados

O processo de rastrear a origem e o histórico dos dados ao longo de seu ciclo de vida, por exemplo, como os dados foram gerados, transmitidos e armazenados.

titular dos dados

Um indivíduo cujos dados estão sendo coletados e processados.

data warehouse

Um sistema de gerenciamento de dados compatível com business intelligence, como analytics. Os data warehouses geralmente contêm grandes quantidades de dados históricos e geralmente são usados para consultas e análises.

linguagem de definição de dados (DDL)

Instruções ou comandos para criar ou modificar a estrutura de tabelas e objetos em um banco de dados.

linguagem de manipulação de dados (DML)

Instruções ou comandos para modificar (inserir, atualizar e excluir) informações em um banco de dados.

DDL

Veja linguagem de definição de banco de dados.

deep ensemble

A combinação de vários modelos de aprendizado profundo para gerar previsões. Os deep ensembles podem ser usados para produzir uma previsão mais precisa ou para estimar a incerteza nas previsões.

Aprendizado profundo

Um subcampo do ML que usa várias camadas de redes neurais artificiais para identificar o mapeamento entre os dados de entrada e as variáveis-alvo de interesse.

defesa completa

Uma abordagem de segurança da informação na qual uma série de mecanismos e controles de segurança são cuidadosamente distribuídos por toda a rede de computadores para proteger a confidencialidade, a integridade e a disponibilidade da rede e dos dados nela contidos. Ao adotar essa estratégia AWS, você adiciona vários controles em diferentes camadas da AWS Organizations estrutura para ajudar a proteger os recursos. Por exemplo, uma abordagem de defesa aprofundada pode combinar autenticação multifatorial, segmentação de rede e criptografia.

administrador delegado

Em AWS Organizations, um serviço compatível pode registrar uma conta de AWS membro para administrar as contas da organização e gerenciar as permissões desse serviço. Essa conta é chamada de administrador delegado para esse serviço Para obter mais informações e uma lista de serviços compatíveis, consulte Serviços que funcionam com o AWS Organizations na documentação do AWS Organizations .

implantação

O processo de criar uma aplicação, novos recursos ou correções de código disponíveis no ambiente de destino. A implantação envolve a implementação de mudanças em uma base de código e, em seguida, a criação e execução dessa base de código nos ambientes da aplicação

ambiente de desenvolvimento

Veja ambiente.

controle detectivo

Um controle de segurança projetado para detectar, registrar e alertar após a ocorrência de um evento. Esses controles são uma segunda linha de defesa, alertando você sobre eventos de segurança que contornaram os controles preventivos em vigor. Para obter mais informações, consulte Controles detectivos em Como implementar controles de segurança na AWS.

mapeamento do fluxo de valor de desenvolvimento (DVSM)

Um processo usado para identificar e priorizar restrições que afetam negativamente a velocidade e a qualidade em um ciclo de vida de desenvolvimento de software. O DVSM estende o processo de mapeamento do fluxo de valor originalmente projetado para práticas de manufatura enxuta. Ele se concentra nas etapas e equipes necessárias para criar e movimentar valor por meio do processo de desenvolvimento de software.

gêmeo digital

Uma representação virtual de um sistema real, como um prédio, fábrica, equipamento industrial ou linha de produção. Os gêmeos digitais oferecem suporte à manutenção preditiva, ao monitoramento remoto e à otimização da produção.

tabela de dimensões

Em um esquema em estrela, uma tabela menor que contém atributos de dados sobre dados quantitativos em uma tabela de fatos. Os atributos da tabela de dimensões geralmente são campos de texto ou números discretos que se comportam como texto. Esses atributos normalmente são usados para restringir consultas, filtrar e rotular conjuntos de resultados.

desastre

Um evento que impede que uma workload ou sistema cumpra seus objetivos de negócios em seu local principal de implantação. Esses eventos podem ser desastres naturais, falhas técnicas ou o resultado de ações humanas, como configuração incorreta não intencional ou ataque de malware.

Recuperação de desastres (RD)

A estratégia e o processo que você usa para minimizar o tempo de inatividade e a perda de dados causados por um disastre. Para obter mais informações, consulte Recuperação de desastres de cargas de trabalho em AWS: Recuperação na nuvem na AWS Well-Architected estrutura.

DML

Veja linguagem de manipulação de banco de dados.

design orientado por domínio

Uma abordagem ao desenvolvimento de um sistema de software complexo conectando seus componentes aos domínios em evolução, ou principais metas de negócios, atendidos por cada componente. Esse conceito foi introduzido por Eric Evans em seu livro Domain-Driven Design: Tackling Complexity in the Heart of Software (Boston: Addison-Wesley Professional, 2003). Para obter informações sobre como você pode usar o design orientado por domínio com o padrão strangler fig, consulte Modernizando os serviços web legados da Microsoft ASP.NET (ASMX) de forma incremental usando contêineres e o Amazon API Gateway.

DR

Veja recuperação de desastres.

Detecção da oscilação

Rastreamento de desvios de uma configuração de linha de base. Por exemplo, você pode usar AWS CloudFormation para detectar desvios nos recursos do sistema ou AWS Control Tower para detectar mudanças em seu landing zone que possam afetar a conformidade com os requisitos de governança.

DVSM

Veja mapeamento do fluxo de valor de desenvolvimento.

E

EDA

Veja análise exploratória de dados.

EDI

Veja intercâmbio eletrônico de dados.

computação de borda

A tecnologia que aumenta o poder computacional de dispositivos inteligentes nas bordas de uma rede de IoT. Quando comparada com a computação em nuvem, a computação de borda pode reduzir a latência da comunicação e melhorar o tempo de resposta.

intercâmbio eletrônico de dados (EDI)

A troca automatizada de documentos comerciais entre organizações. Para obter mais informações, consulte O que é EDI (Intercâmbio eletrônico de dados)?.

criptografia

Um processo de computação que transforma dados de texto simples, legíveis por humanos, em texto cifrado.

chave de criptografia

Uma sequência criptográfica de bits aleatórios que é gerada por um algoritmo de criptografia. As chaves podem variar em tamanho, e cada chave foi projetada para ser imprevisível e exclusiva.

endianismo

A ordem na qual os bytes são armazenados na memória do computador. Big-endian os sistemas armazenam primeiro o byte mais significativo. Little-endian os sistemas armazenam primeiro o byte menos significativo.

endpoint

Veja endpoint de serviço.

serviço de endpoint

Um serviço que pode ser hospedado em uma nuvem privada virtual (VPC) para ser compartilhado com outros usuários. Você pode criar um serviço de endpoint com AWS PrivateLink e conceder permissões a outros diretores Contas da AWS ou a AWS Identity and Access Management (IAM). Essas contas ou entidades principais podem se conectar ao serviço de endpoint de maneira privada criando endpoints da VPC de interface. Para obter mais informações, consulte Criar um serviço de endpoint na documentação do Amazon Virtual Private Cloud (Amazon VPC).

planejamento de recursos empresariais (ERP)

Um sistema que automatiza e gerencia os principais processos de negócios (como contabilidade, MES e gerenciamento de projetos) para uma empresa.

criptografia envelopada

O processo de criptografar uma chave de criptografia com outra chave de criptografia. Para obter mais informações, consulte Criptografia de envelope na documentação AWS Key Management Service (AWS KMS).

ambiente

Uma instância de uma aplicação em execução. Estes são tipos comuns de ambientes na computação em nuvem:

- ambiente de desenvolvimento: uma instância de uma aplicação em execução que está disponível somente para a equipe principal responsável pela manutenção da aplicação. Ambientes de desenvolvimento são usados para testar mudanças antes de promovê-las para ambientes superiores. Esse tipo de ambiente às vezes é chamado de ambiente de teste.
- ambientes inferiores: todos os ambientes de desenvolvimento para uma aplicação, como aqueles usados para compilações e testes iniciais.
- ambiente de produção: uma instância de uma aplicação em execução que os usuários finais podem acessar. Em um CI/CD pipeline, o ambiente de produção é o último ambiente de implantação.

- ambientes superiores: todos os ambientes que podem ser acessados por usuários que não sejam a equipe principal de desenvolvimento. Isso pode incluir um ambiente de produção, ambientes de pré-produção e ambientes para testes de aceitação do usuário.

epic

Em metodologias ágeis, categorias funcionais que ajudam a organizar e priorizar seu trabalho. Os epics fornecem uma descrição de alto nível dos requisitos e das tarefas de implementação. Por exemplo, os épicos de segurança AWS da CAF incluem gerenciamento de identidade e acesso, controles de detetive, segurança de infraestrutura, proteção de dados e resposta a incidentes. Para obter mais informações sobre epics na estratégia de migração da AWS , consulte o guia de implementação do programa.

ERP

Veja planejamento de recursos empresariais.

análise exploratória de dados (EDA)

O processo de analisar um conjunto de dados para entender suas principais características. Você coleta ou agrega dados e, em seguida, realiza investigações iniciais para encontrar padrões, detectar anomalias e verificar suposições. O EDA é realizado por meio do cálculo de estatísticas resumidas e da criação de visualizações de dados.

F

tabela de fatos

A tabela central em um esquema em estrela. Ela armazena dados quantitativos sobre as operações comerciais. Normalmente, uma tabela de fatos contém dois tipos de colunas: as que contêm medidas e as que contêm uma chave externa para uma tabela de dimensões.

Antecipar-se à falha

Uma filosofia que usa testes frequentes e incrementais para reduzir o ciclo de vida do desenvolvimento. É uma parte essencial de uma abordagem ágil.

delimitação de isolamento contra falhas

No Nuvem AWS, um limite, como uma zona de disponibilidade, Região da AWS um plano de controle ou um plano de dados, que limita o efeito de uma falha e ajuda a melhorar a resiliência das cargas de trabalho. Para obter mais informações, consulte AWS Fault Isolation Boundaries.

ramificação de recursos

Veja ramificação.

recursos

Os dados de entrada usados para fazer uma previsão. Por exemplo, em um contexto de manufatura, os recursos podem ser imagens capturadas periodicamente na linha de fabricação.

importância do recurso

O quanto um recurso é importante para as previsões de um modelo. Isso geralmente é expresso como uma pontuação numérica que pode ser calculada por meio de várias técnicas, como Shapley Additive Explanations (SHAP) e gradientes integrados. Para obter mais informações, consulte Interpretabilidade do modelo de aprendizado de máquina com AWS.

transformação de recursos

O processo de otimizar dados para o processo de ML, incluindo enriquecer dados com fontes adicionais, escalar valores ou extrair vários conjuntos de informações de um único campo de dados. Isso permite que o modelo de ML se beneficie dos dados. Por exemplo, se a data “2021-05-27 00:15:37” for dividida em “2021”, “maio”, “quinta” e “15”, isso poderá ajudar o algoritmo de aprendizado a aprender padrões diferenciados associados a diferentes componentes de dados.

prompt few shot

Fornecer a um LLM um pequeno número de exemplos que demonstram a tarefa e o resultado desejado antes de solicitar que ele execute uma tarefa semelhante. Essa técnica é uma aplicação do aprendizado contextual, em que os modelos aprendem com exemplos (fotos) incorporados aos prompts. Few-shot a solicitação pode ser eficaz para tarefas que exigem formatação, raciocínio ou conhecimento de domínio específicos. Veja também prompts zero-shot.

FGAC

Veja controle de acesso refinado.

Controle de acesso refinado (FGAC)

O uso de várias condições para permitir ou negar uma solicitação de acesso.

migração flash-cut

Um método de migração de banco de dados que usa replicação contínua de dados via captura de dados de alteração para migrar os dados no menor tempo possível, em vez de usar uma abordagem em fases. O objetivo é reduzir ao mínimo o tempo de inatividade.

FM

Veja modelo de base.

modelo de base (FM)

Uma grande rede neural de aprendizado profundo que treina em grandes conjuntos de dados generalizados e não rotulados. Os FMs são capazes de realizar uma ampla variedade de tarefas gerais, como entender a linguagem, gerar texto e imagens e conversar em linguagem natural. Para obter mais informações, consulte O que são modelos de base?.

Gateway FM

Um intermediário centralizado que controla e normaliza o acesso aos modelos de fundação. Também conhecido como gateway LLM.

G

IA generativa

Um subconjunto de modelos de IA que foram treinados em grandes quantidades de dados e que podem usar um simples prompt de texto para criar novos artefatos e conteúdo, como imagens, vídeos, texto e áudio. Para obter mais informações, consulte O que é IA generativa?.

bloqueio geográfico

Veja restrições geográficas.

restrições geográficas (bloqueio geográfico)

Na Amazon CloudFront, uma opção para impedir que usuários em países específicos acessem distribuições de conteúdo. É possível usar uma lista de permissões ou uma lista de bloqueios para especificar países aprovados e banidos. Para obter mais informações, consulte Restringir a distribuição geográfica do seu conteúdo na CloudFront documentação.

Fluxo de trabalho do GitFlow

Uma abordagem na qual ambientes inferiores e superiores usam ramificações diferentes em um repositório de código-fonte. O fluxo de trabalho do Gitflow é considerado legado, e o fluxo de trabalho trunk-based é a abordagem moderna e preferencial.

golden image

Um snapshot de um sistema ou software usado como modelo para implantar novas instâncias desse sistema ou software. Por exemplo, na manufatura, uma golden image pode ser usada para

provisionar software em vários dispositivos e ajudar a melhorar a velocidade, a escalabilidade e a produtividade nas operações de fabricação de dispositivos.

estratégia greenfield

A ausência de infraestrutura existente em um novo ambiente. Ao adotar uma estratégia greenfield para uma arquitetura de sistema, é possível selecionar todas as novas tecnologias sem a restrição da compatibilidade com a infraestrutura existente, também conhecida como brownfield. Se estiver expandindo a infraestrutura existente, poderá combinar as estratégias brownfield e greenfield.

barreira de proteção

Uma regra de alto nível que ajuda a gerenciar recursos, políticas e conformidade em todas as unidades organizacionais (UOs). Barreiras de proteção preventivas impõem políticas para garantir o alinhamento a padrões de conformidade. Elas são implementadas usando políticas de controle de serviço e limites de permissões do IAM. Barreiras de proteção detectivas detectam violações de políticas e problemas de conformidade e geram alertas para remediação. Eles são implementados usando AWS Config, AWS Security Hub CSPM, Amazon GuardDuty AWS Trusted Advisor, Amazon Inspector e verificações personalizadas AWS Lambda .

grades de proteção (IA)

Mecanismos de segurança que filtram, validam e restringem as entradas e saídas dos agentes para ajudar a garantir um comportamento de IA responsável e seguro.

H

HA

Veja alta disponibilidade.

migração heterogênea de bancos de dados

Migrar seu banco de dados de origem para um banco de dados de destino que usa um mecanismo de banco de dados diferente (por exemplo, Oracle para Amazon Aurora). A migração heterogênea geralmente faz parte de um esforço de redefinição da arquitetura, e converter o esquema pode ser uma tarefa complexa. O AWS fornece o AWS SCT para ajudar nas conversões de esquemas.

alta disponibilidade (HA)

A capacidade de uma workload operar continuamente, sem intervenção, em caso de desafios ou desastres. Os sistemas AH são projetados para realizar o failover automático, oferecer consistentemente desempenho de alta qualidade e lidar com diferentes cargas e falhas com impacto mínimo no desempenho.

modernização de historiador

Uma abordagem usada para modernizar e atualizar os sistemas de tecnologia operacional (OT) para melhor atender às necessidades do setor de manufatura. Um historiador é um tipo de banco de dados usado para coletar e armazenar dados de várias fontes em uma fábrica.

dados de hold-out

Uma parte dos dados históricos rotulados que são retidos de um conjunto de dados usado para treinar um modelo de machine learning. Você pode usar dados de hold-out para avaliar a performance do modelo comparando as predições do modelo com os dados de retenção.

humano no circuito (HiTL)

Um padrão de fluxo de trabalho em que a execução do agente é pausada para análise e aprovação humana em pontos críticos de decisão.

migração homogênea de bancos de dados

Migrar seu banco de dados de origem para um banco de dados de destino que compartilha o mesmo mecanismo de banco de dados (por exemplo, Microsoft SQL Server para Amazon RDS para SQL Server). A migração homogênea geralmente faz parte de um esforço de redefinição da hospedagem ou da plataforma. É possível usar utilitários de banco de dados nativos para migrar o esquema.

dados quentes

Dados acessados com frequência, como dados em tempo real ou dados translacionais recentes. Esses dados normalmente exigem uma camada ou classe de armazenamento de alto desempenho para fornecer respostas rápidas às consultas.

hotfix

Uma correção urgente para um problema crítico em um ambiente de produção. Devido à sua urgência, um hotfix geralmente é feito fora do fluxo de trabalho típico de uma DevOps versão.

período de hipercuidados

Imediatamente após a substituição, o período em que uma equipe de migração gerencia e monitora as aplicações migradas na nuvem para resolver quaisquer problemas. Normalmente, a duração desse período é de 1 a 4 dias. No final do período de hipercuidados, a equipe de migração normalmente transfere a responsabilidade pelas aplicações para a equipe de operações de nuvem.

eu

IaC

Veja infraestrutura como código.

Política baseada em identidade

Uma política anexada a um ou mais diretores do IAM que define suas permissões no Nuvem AWS ambiente.

aplicação ociosa

Uma aplicação que tem um uso médio de CPU e memória entre 5 e 20% em um período de 90 dias. Em um projeto de migração, é comum retirar essas aplicações ou retê-las on-premises.

IIoT

Veja Internet das Coisas Industrial.

infraestrutura imutável

Um modelo que implanta uma nova infraestrutura para workloads de produção em vez de atualizar, aplicar patches ou modificar a infraestrutura existente. Infraestruturas imutáveis são inerentemente mais consistentes, confiáveis e preditivas do que infraestruturas mutáveis. Para obter mais informações, consulte as melhores práticas de implantação usando infraestrutura imutável na AWS Well-Architected Estrutura.

VPC de entrada (admissão)

Em uma arquitetura de AWS várias contas, uma VPC que aceita, inspeciona e roteia conexões de rede de fora de um aplicativo. A Arquitetura de referência de segurança da AWS recomenda configurar sua conta de rede com VPCs de entrada, saída e inspeção para proteger a interface bidirecional entre a aplicação e a Internet em geral.

migração incremental

Uma estratégia de substituição na qual você migra a aplicação em pequenas partes, em vez de realizar uma única substituição completa. Por exemplo, é possível mover inicialmente apenas alguns microsserviços ou usuários para o novo sistema. Depois de verificar se tudo está funcionando corretamente, mova os microsserviços ou usuários adicionais de forma incremental até poder descomissionar seu sistema herdado. Essa estratégia reduz os riscos associados a migrações de grande porte.

Indústria 4.0

Um termo que foi introduzido por Klaus Schwab em 2016 para se referir à modernização dos processos de fabricação por meio de avanços na conectividade, dados em tempo real, automação, análise e. AI/ML

infraestrutura

Todos os recursos e ativos contidos no ambiente de uma aplicação.

Infraestrutura como código (IaC)

O processo de provisionamento e gerenciamento da infraestrutura de uma aplicação por meio de um conjunto de arquivos de configuração. A IaC foi projetada para ajudar você a centralizar o gerenciamento da infraestrutura, padronizar recursos e escalar rapidamente para que novos ambientes sejam reproduzíveis, confiáveis e consistentes.

Internet das Coisas Industrial (IIoT)

O uso de sensores e dispositivos conectados à Internet nos setores industriais, como manufatura, energia, automotivo, saúde, ciências biológicas e agricultura. Para obter mais informações, consulte Construir uma estratégia de transformação digital para a Internet das Coisas Industrial (IIoT).

VPC de inspeção

Em uma arquitetura de AWS várias contas, uma VPC centralizada que gerencia as inspeções do tráfego de rede entre VPCs (na mesma ou em diferentes Regiões da AWS), a Internet e as redes locais. A Arquitetura de referência de segurança da AWS recomenda configurar sua conta de rede com VPCs de entrada, saída e inspeção para proteger a interface bidirecional entre a aplicação e a Internet em geral.

Internet das coisas (IoT)

A rede de objetos físicos conectados com sensores ou processadores incorporados que se comunicam com outros dispositivos e sistemas pela Internet ou por uma rede de comunicação local. Para obter mais informações, consulte O que é IoT?

interpretabilidade

Uma característica de um modelo de machine learning que descreve o grau em que um ser humano pode entender como as previsões do modelo dependem de suas entradas. Para obter mais informações, consulte Interpretabilidade do modelo de aprendizado de máquina com AWS.

IoT

Veja Internet das Coisas.

Biblioteca de informações de TI (ITIL)

Um conjunto de práticas recomendadas para fornecer serviços de TI e alinhar esses serviços a requisitos de negócios. A ITIL fornece a base para o ITSM.

Gerenciamento de serviços de TI (ITSM)

Atividades associadas a design, implementação, gerenciamento e suporte de serviços de TI para uma organização. Para obter informações sobre a integração de operações em nuvem com ferramentas de ITSM, consulte o guia de integração de operações.

ITIL

Veja biblioteca de informações de TI.

ITSM

Veja gerenciamento de serviços de TI.

L

controle de acesso baseado em etiqueta (LBAC)

Uma implementação do controle de acesso obrigatório (MAC) em que os usuários e os dados em si recebem explicitamente um valor de etiqueta de segurança. A interseção entre a etiqueta de segurança do usuário e a etiqueta de segurança dos dados determina quais linhas e colunas podem ser vistas pelo usuário.

zona de pouso

Uma landing zone é um AWS ambiente bem arquitetado, com várias contas, escalável e seguro. Um ponto a partir do qual suas organizações podem iniciar e implantar rapidamente workloads e aplicações com confiança em seu ambiente de segurança e infraestrutura. Para obter mais informações sobre zonas de pouso, consulte Configurar um ambiente da AWS com várias contas seguro e escalável.

grande modelo de linguagem (LLM)

Um modelo de IA de aprendizado profundo pré-treinado em uma grande quantidade de dados. Um LLM pode realizar várias tarefas, como responder a perguntas, resumir documentos, traduzir texto para outros idiomas e completar frases. Para obter mais informações, consulte O que é grande modelo de linguagem (LLM)?.

migração de grande porte

Uma migração de 300 servidores ou mais.

LBAC

Veja controle de acesso baseado em rótulo.

privilégio mínimo

A prática recomendada de segurança de conceder as permissões mínimas necessárias para executar uma tarefa. Para obter mais informações, consulte Aplicar permissões de privilégios mínimos na documentação do IAM.

mover sem alterações (lift-and-shift)

Veja 7 Rs.

sistema little-endian

Um sistema que armazena o byte menos significativo antes. Veja também endianness.

LLM

Veja grande modelo de linguagem.

ambientes inferiores

Veja ambiente.

M

machine learning (ML)

Um tipo de inteligência artificial que usa algoritmos e técnicas para reconhecimento e aprendizado de padrões. O ML analisa e aprende com dados gravados, por exemplo, dados da Internet das Coisas (IoT), para gerar um modelo estatístico baseado em padrões. Para obter mais informações, consulte Machine learning.

ramificação principal

Veja ramificação.

Malware

Software projetado para comprometer a segurança ou a privacidade do computador. O malware pode interromper os sistemas do computador, vazar informações sensíveis ou obter acesso não autorizado. Exemplos de malware incluem vírus, worms, ransomware, cavalos de Troia, spyware e keyloggers.

Serviços gerenciados

Serviços da AWS para o qual AWS opera a camada de infraestrutura, o sistema operacional e as plataformas, e você acessa os endpoints para armazenar e recuperar dados. O Amazon Simple Storage Service (Amazon S3) e o Amazon DynamoDB são exemplos de serviços gerenciados. Eles também são conhecidos como serviços abstraídos.

sistema de execução de manufatura (MES)

Um sistema de software para rastrear, monitorar, documentar e controlar processos de produção que convertem matérias-primas em produtos acabados no chão de fábrica.

MAP

Veja Programa de Aceleração da Migração.

MCP

Consulte Protocolo de contexto do modelo.

Protocolo de contexto para modelos (MCP)

Um protocolo sem estado para comunicação entre agentes e ferramentas.

Servidor MCP

Um serviço que expõe uma ou mais ferramentas por meio do Model Context Protocol.

mecanismo

Um processo completo em que você cria uma ferramenta, impulsiona a adoção da ferramenta e, em seguida, inspeciona os resultados para fazer ajustes. Um mecanismo é um ciclo que se reforça e se aprimora à medida que opera. Para obter mais informações, consulte Criação de mecanismos na AWS Well-Architected estrutura.

conta de membro

Todos, Contas da AWS exceto a conta de gerenciamento, que fazem parte de uma organização em AWS Organizations. Uma conta só pode ser membro de uma organização de cada vez.

MES

Veja sistema de execução de manufatura.

Transporte de Telemetria de Enfileiramento de Mensagens (MQTT)

Um protocolo de comunicação leve, máquina a máquina (M2M), baseado no padrão, para dispositivos de IoT com recursos publish/subscribelimitados.

microsserviço

Um serviço pequeno e independente que se comunica por meio de APIs bem definidas e normalmente pertence a equipes pequenas e autônomas. Por exemplo, um sistema de seguradora pode incluir microsserviços que mapeiam as capacidades comerciais, como vendas ou marketing, ou subdomínios, como compras, reclamações ou análises. Os benefícios dos microsserviços incluem agilidade, escalabilidade flexível, fácil implantação, código reutilizável e resiliência. Para obter mais informações, consulte Integração de microsserviços usando serviços sem AWS servidor.

arquitetura de microsserviços

Uma abordagem à criação de aplicações com componentes independentes que executam cada processo de aplicação como um microsserviço. Esses microsserviços se comunicam por meio de uma interface bem definida usando APIs leves. Cada microsserviço nessa arquitetura pode ser atualizado, implantado e escalado para atender à demanda por funções específicas de uma aplicação. Para obter mais informações, consulte Implementação de microsserviços em. AWS

Programa de Aceleração da Migração (MAP)

Um AWS programa que fornece suporte de consultoria, treinamento e serviços para ajudar as organizações a criar uma base operacional sólida para migrar para a nuvem e ajudar a

compensar o custo inicial das migrações. O MAP inclui uma metodologia de migração para executar migrações legadas de forma metódica e um conjunto de ferramentas para automatizar e acelerar cenários comuns de migração.

migração em escala

O processo de mover a maior parte do portfólio de aplicações para a nuvem em ondas, com mais aplicações sendo movidas em um ritmo mais rápido a cada onda. Essa fase usa as práticas recomendadas e lições aprendidas nas fases anteriores para implementar uma fábrica de migração de equipes, ferramentas e processos para agilizar a migração de workloads por meio de automação e entrega ágeis. Esta é a terceira fase da estratégia de migração para a AWS.

fábrica de migração

Cross-functional equipes que simplificam a migração de cargas de trabalho por meio de abordagens automatizadas e ágeis. As equipes da fábrica de migração geralmente incluem operações, analistas e proprietários de negócios, engenheiros de migração, desenvolvedores e DevOps profissionais que trabalham em sprints. Entre 20 e 50% de um portfólio de aplicações corporativas consiste em padrões repetidos que podem ser otimizados por meio de uma abordagem de fábrica. Para obter mais informações, consulte discussão sobre fábricas de migração e o guia do Cloud Migration Factory neste conjunto de conteúdo.

metadados de migração

As informações sobre a aplicação e o servidor necessárias para concluir a migração. Cada padrão de migração exige um conjunto de metadados de migração diferente. Exemplos de metadados de migração incluem a sub-rede, o grupo de segurança e AWS a conta de destino.

padrão de migração

Uma tarefa de migração repetível que detalha a estratégia de migração, o destino da migração e a aplicação ou o serviço de migração usado. Exemplo: rehospede a migração para o Amazon EC2 AWS com o Application Migration Service.

Avaliação de Portfólio para Migração (MPA)

Uma ferramenta on-line que fornece informações para validar o caso de negócios para migrar para a Nuvem AWS. O MPA fornece avaliação detalhada do portfólio (dimensionamento correto do servidor, preços, comparações de TCO, análise de custos de migração), bem como planejamento de migração (análise e coleta de dados de aplicações, agrupamento de aplicações, priorização de migração e planejamento de ondas). A ferramenta MPA (requer login) está disponível gratuitamente para todos os AWS consultores e consultores parceiros da APN.

Avaliação de Preparação para Migração (MRA)

O processo de obter insights sobre o status de prontidão de uma organização para a nuvem, identificar pontos fortes e fracos e criar um plano de ação para fechar as lacunas identificadas, usando o CAF. AWS Para mais informações, consulte o guia de preparação para migração. A MRA é a primeira fase da estratégia de migração para a AWS.

estratégia de migração

A abordagem usada para migrar uma workload para a Nuvem AWS. Para obter mais informações, veja a entrada 7 Rs neste glossário e consulte Mobilize sua organização para acelerar migrações em grande escala.

ML

Veja machine learning.

modernização

Transformar uma aplicação desatualizada (herdada ou monolítica) e sua infraestrutura em um sistema ágil, elástico e altamente disponível na nuvem para reduzir custos, ganhar eficiência e aproveitar as inovações. Para obter mais informações, consulte Strategy for modernizing applications in the Nuvem AWS.

avaliação de preparação para modernização

Uma avaliação que ajuda a determinar a preparação para modernização das aplicações de uma organização. Ela identifica benefícios, riscos e dependências e determina o quão bem a organização pode acomodar o estado futuro dessas aplicações. O resultado da avaliação é um esquema da arquitetura de destino, um roteiro que detalha as fases de desenvolvimento e os marcos do processo de modernização e um plano de ação para abordar as lacunas identificadas. Para obter mais informações, consulte Evaluating modernization readiness for applications in the Nuvem AWS.

aplicações monolíticas (monólitos)

Aplicações que são executadas como um único serviço com processos fortemente acoplados. As aplicações monolíticas apresentam várias desvantagens. Se um recurso da aplicação apresentar um aumento na demanda, toda a arquitetura deverá ser escalada. Adicionar ou melhorar os recursos de uma aplicação monolítica também se torna mais complexo quando a base de código cresce. Para resolver esses problemas, é possível criar uma arquitetura de microsserviços. Para obter mais informações, consulte Decompor monólitos em microsserviços.

MPA

Veja Avaliação do Portfólio para Migração.

MQTT

Veja Transporte de Telemetria de Enfileiramento de Mensagens.

classificação multiclasse

Um processo que ajuda a gerar previsões para várias classes (prevendo um ou mais de dois resultados). Por exemplo, um modelo de ML pode perguntar “Este produto é um livro, um carro ou um telefone?” ou “Qual categoria de produtos é mais interessante para este cliente?”

infraestrutura mutável

Um modelo que atualiza e modifica a infraestrutura existente para workloads de produção. Para melhorar a consistência, confiabilidade e previsibilidade, a AWS Well-Architected Estrutura recomenda o uso de infraestrutura imutável como uma prática recomendada.

O

OAC

Veja controle de acesso de origem.

OAI

Veja identidade de acesso de origem.

OCM

Veja gerenciamento de alterações organizacionais.

migração offline

Um método de migração no qual a workload de origem é desativada durante o processo de migração. Esse método envolve tempo de inatividade prolongado e geralmente é usado para workloads pequenas e não críticas.

OI

Veja integração de operações.

Ola

Veja acordo de nível operacional.

migração online

Um método de migração no qual a workload de origem é copiada para o sistema de destino sem ser colocada offline. As aplicações conectadas à workload podem continuar funcionando durante a migração. Esse método envolve um tempo de inatividade nulo ou mínimo e normalmente é usado para workloads essenciais para a produção.

OPC-UA

Veja Open Process Communications - Unified Architecture.

Comunicação de processo aberto - Arquitetura unificada (OPC-UA)

Um protocolo de comunicação máquina a máquina (M2M) para automação industrial. OPC-UA fornece um padrão de interoperabilidade com esquemas de criptografia, autenticação e autorização de dados.

acordo de nível operacional (OLA)

Um acordo que esclarece o que os grupos funcionais de TI prometem oferecer uns aos outros para apoiar um acordo de serviço (SLA).

análise de prontidão operacional (ORR)

Uma lista de verificação de perguntas e práticas recomendadas associadas que ajudam você a entender, avaliar, prevenir ou reduzir o escopo de incidentes e possíveis falhas. Para obter mais informações, consulte Operational Readiness Reviews (ORR) na AWS Well-Architected Estrutura.

tecnologia operacional (TO)

Sistemas de hardware e software que trabalham com o ambiente físico para controlar operações, equipamentos e infraestrutura industriais. Na manufatura, a integração dos sistemas de tecnologia da informação (TI) e tecnologia operacional (TO) é o foco principal das transformações da Indústria 4.0.

integração de operações (OI)

O processo de modernização das operações na nuvem, que envolve planejamento de preparação, automação e integração. Para obter mais informações, consulte o guia de integração de operações.

trilha organizacional

Uma trilha criada por ela AWS CloudTrail registra todos os eventos de todos Contas da AWS em uma organização em AWS Organizations. Essa trilha é criada em cada Conta da AWS que faz parte da organização e monitora a atividade em cada conta. Para obter mais informações, consulte Criação de uma trilha para uma organização na CloudTrail documentação.

gerenciamento de alterações organizacionais (OCM)

Uma estrutura para gerenciar grandes transformações de negócios disruptivas de uma perspectiva de pessoas, cultura e liderança. O OCM ajuda as organizações a se prepararem e fazerem a transição para novos sistemas e estratégias, acelerando a adoção de alterações, abordando questões de transição e promovendo mudanças culturais e organizacionais. Na estratégia de AWS migração, essa estrutura é chamada de aceleração de pessoas, devido à velocidade de mudança necessária nos projetos de adoção da nuvem. Para obter mais informações, consulte o guia do OCM.

controle de acesso de origem (OAC)

Em CloudFront, uma opção aprimorada para restringir o acesso para proteger seu conteúdo do Amazon Simple Storage Service (Amazon S3). O OAC oferece suporte a todos os buckets do S3 Regiões da AWS, à criptografia do lado do servidor com AWS KMS (SSE-KMS) e à dinâmica `PUT` e `DELETE` às solicitações ao bucket do S3.

Identidade do acesso de origem (OAI)

Em CloudFront, uma opção para restringir o acesso para proteger seu conteúdo do Amazon S3. Quando você usa o OAI, CloudFront cria um principal com o qual o Amazon S3 pode se autenticar. Os diretores autenticados podem acessar o conteúdo em um bucket do S3 somente por meio de uma distribuição específica. CloudFront Veja também OAC, que fornece um controle de acesso mais granular e aprimorado.

ORR

Veja análise de prontidão operacional.

OT

Veja tecnologia operacional.

VPC de saída (egresso)

Em uma arquitetura de AWS várias contas, uma VPC que gerencia conexões de rede que são iniciadas de dentro de um aplicativo. A Arquitetura de referência de segurança da AWS

recomenda configurar sua conta de rede com VPCs de entrada, saída e inspeção para proteger a interface bidirecional entre a aplicação e a Internet em geral.

P

limite de permissões

Uma política de gerenciamento do IAM anexada a entidades principais do IAM para definir as permissões máximas que o usuário ou perfil podem ter. Para obter mais informações, consulte Limites de permissões na documentação do IAM.

Informações de identificação pessoal (PII)

Informações que, quando visualizadas diretamente ou combinadas com outros dados relacionados, podem ser usadas para inferir razoavelmente a identidade de um indivíduo. Exemplos de PII incluem nomes, endereços e informações de contato.

PII

Veja informações de identificação pessoal.

manual

Um conjunto de etapas predefinidas que capturam o trabalho associado às migrações, como a entrega das principais funções operacionais na nuvem. Um manual pode assumir a forma de scripts, runbooks automatizados ou um resumo dos processos ou etapas necessários para operar seu ambiente modernizado.

PLC

Veja controlador lógico programável.

PLM

Veja gerenciamento do ciclo de vida do produto.

política

Um objeto que pode definir permissões (veja política baseada em identidade), especificar condições de acesso (veja políta baseada em recurso) ou definir as permissões máximas para todas as contas em uma organização no AWS Organizations (veja política de controle de serviços).

persistência poliglota

Escolher de forma independente a tecnologia de armazenamento de dados de um microsserviço com base em padrões de acesso a dados e outros requisitos. Se seus microsserviços tiverem a mesma tecnologia de armazenamento de dados, eles poderão enfrentar desafios de implementação ou apresentar baixa performance. Os microsserviços serão implementados com mais facilidade e alcançarão performance e escalabilidade melhores se usarem o armazenamento de dados mais bem adaptado às suas necessidades.

avaliação do portfólio

Um processo de descobrir, analisar e priorizar o portfólio de aplicações para planejar a migração. Para obter mais informações, consulte Avaliar a preparação para a migração.

predicado

Uma condição de consulta que retorna `true` ou `false`, normalmente localizada em uma cláusula `WHERE`.

pushdown de predicados

Uma técnica de otimização de consultas de banco de dados que filtra os dados na consulta antes da transferência. Isso reduz a quantidade de dados que devem ser recuperados e processados do banco de dados relacional e melhora a performance das consultas.

controle preventivo

Um controle de segurança projetado para evitar que um evento ocorra. Esses controles são a primeira linha de defesa para ajudar a evitar acesso não autorizado ou alterações indesejadas em sua rede. Para obter mais informações, consulte Controles preventivos em Como implementar controles de segurança na AWS.

principal (entidade principal)

Uma entidade AWS que pode realizar ações e acessar recursos. Essa entidade geralmente é um usuário raiz para um Conta da AWS, uma função do IAM ou um usuário. Para obter mais informações, consulte Entidade principal em Termos e conceitos de perfis na documentação do IAM.

Privacidade por design

Uma abordagem em engenharia de sistemas que leva em consideração a privacidade em todo o processo de desenvolvimento.

zonas hospedadas privadas

Um contêiner que armazena informações sobre como você quer que o Amazon Route 53 responda a consultas ao DNS para um domínio e seus subdomínios dentro de uma ou mais VPCs. Para obter mais informações, consulte Como trabalhar com zonas hospedadas privadas na documentação do Route 53.

controle proativo

Um controle de segurança desenvolvido para evitar a implantação de recursos não conformes. Esses controles verificam os recursos antes de serem provisionados. Se o recurso não estiver em conformidade com o controle, ele não será provisionado. Para obter mais informações, consulte o guia de referência de controles na AWS Control Tower documentação e consulte Controles proativos em Implementação de controles de segurança em AWS.

gerenciamento do ciclo de vida do produto (PLM)

O gerenciamento de dados e processos de um produto em todo o seu ciclo de vida, desde a concepção, o desenvolvimento e o lançamento, passando pelo crescimento e maturidade, até o declínio e a remoção.

ambiente de produção

Veja ambiente.

controlador lógico programável (PLC)

Na manufatura, um computador altamente confiável e adaptável que monitora as máquinas e automatiza os processos de fabricação.

encadeamento de prompts

Uso da saída de um prompt do LLM como entrada para o próximo prompt para gerar respostas melhores. Essa técnica é usada para dividir uma tarefa complexa em subtarefas, ou para refinar ou expandir iterativamente uma resposta preliminar. Isso ajuda a melhorar a precisão e a relevância das respostas de um modelo e permite resultados mais granulares e personalizados.

pseudonimização

O processo de substituir identificadores pessoais em um conjunto de dados por valores de espaço reservado. A pseudonimização pode ajudar a proteger a privacidade pessoal. Os dados pseudonimizados ainda são considerados dados pessoais.

publish/subscribe (pub/sub)

Um padrão que permite comunicações assíncronas entre microsserviços para melhorar a escalabilidade e a capacidade de resposta. Por exemplo, em um MES baseado em microsserviços, um microsserviço pode publicar mensagens de eventos em um canal em que outros microsserviços possam assinar. O sistema pode adicionar novos microsserviços sem alterar o serviço de publicação.

Q

plano de consulta

Uma série de etapas, como instruções, usadas para acessar os dados em um sistema de banco de dados relacional SQL.

regressão de planos de consultas

Quando um otimizador de serviço de banco de dados escolhe um plano menos adequado do que escolhia antes de uma determinada alteração no ambiente de banco de dados ocorrer. Isso pode ser causado por alterações em estatísticas, restrições, configurações do ambiente, associações de parâmetros de consulta e atualizações do mecanismo de banco de dados.

R

Matriz RACI

Veja responsável, aprovador, consultado, informado (RACI).

RAG

Veja geração aumentada via recuperação.

ransomware

Um software mal-intencionado desenvolvido para bloquear o acesso a um sistema ou dados de computador até que um pagamento seja feito.

Matriz RASCI

Veja responsável, aprovador, consultado, informado (RACI).

RCAC

Veja controle de acesso por linha e coluna.

réplica de leitura

Uma cópia de um banco de dados usada somente para leitura. É possível encaminhar consultas para a réplica de leitura e reduzir a carga no banco de dados principal.

Redefinir arquitetura

Veja 7 Rs.

objetivo de ponto de recuperação (RPO).

O máximo período de tempo aceitável desde o último ponto de recuperação de dados. Isso determina o que é considerado uma perda aceitável de dados entre o último ponto de recuperação e a interrupção do serviço.

objetivo de tempo de recuperação (RTO)

O máximo atraso aceitável entre a interrupção e a restauração do serviço.

refatorar

Veja 7 Rs.

Região

Uma coleção de AWS recursos em uma área geográfica. Cada um Região da AWS é isolado e independente dos outros para fornecer tolerância a falhas, estabilidade e resiliência. Para obter informações, consulte Specify which Regiões da AWS your account can use.

regressão

Uma técnica de ML que prevê um valor numérico. Por exemplo, para resolver o problema de “Por qual preço esta casa será vendida?” um modelo de ML pode usar um modelo de regressão linear para prever o preço de venda de uma casa com base em fatos conhecidos sobre a casa (por exemplo, a metragem quadrada).

redefinir a hospedagem

Veja 7 Rs.

versão

Em um processo de implantação, o ato de promover mudanças em um ambiente de produção.

realocar

Veja 7 Rs.

redefinir a plataforma

Veja 7 Rs.

recomprar

Veja 7 Rs.

resiliência

A capacidade de uma aplicação de resistir ou se recuperar de interrupções. Alta disponibilidade e recuperação de desastres são considerações comuns ao planejar a resiliência na Nuvem AWS. Para obter mais informações, consulte Nuvem AWS Resilience.

política baseada em recurso

Uma política associada a um recurso, como um bucket do Amazon S3, um endpoint ou uma chave de criptografia. Esse tipo de política especifica quais entidades principais têm acesso permitido, ações válidas e quaisquer outras condições que devem ser atendidas.

matriz responsável, accountable, consultada, informada (RACI)

Uma matriz que define as funções e responsabilidades de todas as partes envolvidas nas atividades de migração e nas operações de nuvem. O nome da matriz é derivado dos tipos de responsabilidade definidos na matriz: responsável (R), responsabilizável (A), consultado (C) e informado (I). O tipo de suporte (S) é opcional. Se você incluir suporte, a matriz será chamada de matriz RASCI e, se excluir, será chamada de matriz RACI.

controle responsivo

Um controle de segurança desenvolvido para conduzir a remediação de eventos adversos ou desvios em relação à linha de base de segurança. Para obter mais informações, consulte Controles responsivos em Como implementar controles de segurança na AWS.

reter

Veja 7 Rs.

Retirada

Veja 7 Rs.

Geração Aumentada de Recuperação (RAG)

Uma tecnologia de IA generativa em que um LLM faz referência a uma fonte de dados autorizada que está fora de suas fontes de dados de treinamento antes de gerar uma resposta. Por exemplo, um modelo RAG pode realizar uma pesquisa semântica na base de conhecimento ou nos dados personalizados de uma organização. Para obter mais informações, consulte O que é RAG (geração aumentada via recuperação)?.

alternância

O processo de atualizar periodicamente um segredo para dificultar o acesso de um invasor às credenciais.

controle de acesso por linha e coluna (RCAC)

O uso de expressões SQL básicas e flexíveis que tenham regras de acesso definidas. O RCAC consiste em permissões de linha e máscaras de coluna.

RPO

Veja objetivo de ponto de recuperação.

RTO

Veja objetivo de tempo de recuperação.

runbook

Um conjunto de procedimentos manuais ou automatizados necessários para realizar uma tarefa específica. Eles são normalmente criados para agilizar operações ou procedimentos repetitivos com altas taxas de erro.

S

SAML 2.0

Um padrão aberto que muitos provedores de identidade (IdPs) usam. Esse recurso permite o login único federado (SSO), para que os usuários possam fazer login Console de gerenciamento da AWS ou chamar as operações da AWS API sem que você precise criar um usuário no IAM para todos em sua organização. Para obter mais informações sobre a federação baseada em SAML 2.0, consulte Sobre a federação baseada em SAML 2.0 na documentação do IAM.

SCADA

Veja controle de supervisão e aquisição de dados.

SCP

Veja política de controle de serviço.

secret

Em AWS Secrets Manager, informações confidenciais ou restritas, como uma senha ou credenciais de usuário, que você armazena de forma criptografada. Consiste no valor secreto e em seus metadados. O valor secreto pode ser binário, uma única string ou várias strings. Para obter mais informações, consulte What's in a Secrets Manager secret? na documentação do Secrets Manager.

segurança desde a concepção

Uma abordagem em engenharia de sistemas que leva em consideração a segurança em todo o processo de desenvolvimento.

controle de segurança

Uma barreira de proteção técnica ou administrativa que impede, detecta ou reduz a capacidade de uma ameaça explorar uma vulnerabilidade de segurança. Existem quatro tipos primários de controles de segurança: preventivos, detectivos, responsivos e proativos.

hardening da segurança

O processo de reduzir a superfície de ataque para torná-la mais resistente a ataques. Isso pode incluir ações como remover recursos que não são mais necessários, implementar a prática recomendada de segurança de conceder privilégios mínimos ou desativar recursos desnecessários em arquivos de configuração.

sistema de gerenciamento de eventos e informações de segurança (SIEM)

Ferramentas e serviços que combinam sistemas de gerenciamento de informações de segurança (SIM) e gerenciamento de eventos de segurança (SEM). Um sistema SIEM coleta, monitora e analisa dados de servidores, redes, dispositivos e outras fontes para detectar ameaças e violações de segurança e gerar alertas.

automação de resposta de segurança

Uma ação predefinida e programada projetada para responder ou remediar automaticamente um evento de segurança. Essas automações servem como controles de segurança responsivos ou detectivos que ajudam você a implementar as melhores práticas AWS de segurança. Exemplos de ações de resposta automatizada incluem a modificação de um grupo de segurança da VPC, a aplicação de patches em uma instância do Amazon EC2 ou a alternância de credenciais.

Criptografia do lado do servidor

Criptografia dos dados em seu destino, por AWS service (Serviço da AWS) quem os recebe.

política de controle de serviços (SCP)

Uma política que fornece controle centralizado sobre as permissões de todas as contas em uma organização no AWS Organizations. As SCPs definem barreiras de proteção ou estabelecem limites para as ações que um administrador pode delegar a usuários ou perfis. É possível usar SCPs como listas de permissão ou de negação para especificar quais serviços ou ações são permitidos ou proibidos. Para obter mais informações, consulte Políticas de controle de serviço na AWS Organizations documentação.

service endpoint (endpoint de serviço)

O URL do ponto de entrada para um AWS service (Serviço da AWS). Você pode usar o endpoint para se conectar programaticamente ao serviço de destino. Para obter mais informações, consulte Endpoints do AWS service (Serviço da AWS) na Referência geral da AWS.

acordo de serviço (SLA)

Um acordo que esclarece o que uma equipe de TI promete fornecer aos clientes, como tempo de atividade e performance do serviço.

indicador de nível de serviço (SLI)

Uma avaliação de um aspecto de performance de um serviço, como taxa de erro, disponibilidade ou throughput.

objetivo de nível de serviço (SLO)

Uma métrica alvo que representa a integridade de um serviço, conforme avaliado por um indicador de nível de serviço.

modelo de responsabilidade compartilhada

Um modelo que descreve a responsabilidade com a qual você compartilha AWS pela segurança e conformidade na nuvem. AWS é responsável pela segurança da nuvem, enquanto você é responsável pela segurança na nuvem. Para obter mais informações, consulte o Modelo de responsabilidade compartilhada.

Inteligência artificial sombria

Aplicativos de IA não autorizados criados ou usados fora dos canais controlados dentro de uma organização.

SIEM

Veja sistema de gerenciamento de eventos e informações de segurança.

ponto único de falha (SPOF)

Uma falha em um único componente crítico de uma aplicação que pode interromper o sistema.

SLA

Veja cacordo de serviço.

SLI

Veja indicador de nível de serviço.

SLO

Veja objetivo de nível de serviço.

modelo dividir e semear

Um padrão para escalar e acelerar projetos de modernização. À medida que novos recursos e lançamentos de produtos são definidos, a equipe principal se divide para criar novas equipes de produtos. Isso ajuda a escalar os recursos e os serviços da sua organização, melhora a produtividade do desenvolvedor e possibilita inovações rápidas. Para obter mais informações, consulte Phased approach to modernizing applications in the Nuvem AWS.

SPOF

Veja ponto único de falha.

esquema em estrela

Uma estrutura organizacional de banco de dados que usa uma grande tabela de fatos para armazenar dados transacionais ou medidos e usa uma ou mais tabelas dimensionais menores para armazenar atributos de dados. Essa estrutura foi projetada para ser usada em um data warehouse ou para fins de inteligência comercial.

padrão strangler fig

Uma abordagem à modernização de sistemas monolíticos que consiste em reescrever e substituir incrementalmente a funcionalidade do sistema até que o sistema herdado possa ser desativado. Esse padrão usa a analogia de uma videira que cresce e se torna uma árvore estabelecida e, eventualmente, supera e substitui sua hospedeira. O padrão foi apresentado por Martin Fowler

como forma de gerenciar riscos ao reescrever sistemas monolíticos. Para ver um exemplo de como aplicar esse padrão, consulte Modernizando os serviços web legados da Microsoft ASP.NET (ASMX) de forma incremental usando contêineres e o Amazon API Gateway.

sub-rede

Um intervalo de endereços IP na VPC. Cada sub-rede fica alocada em uma única zona de disponibilidade.

controle supervisório e aquisição de dados (SCADA)

Na manufatura, um sistema que usa hardware e software para monitorar ativos físicos e operações de produção.

symmetric encryption (criptografia simétrica)

Um algoritmo de criptografia que usa a mesma chave para criptografar e descriptografar dados.

testes sintéticos

Testar um sistema de forma que simule as interações do usuário para detectar possíveis problemas ou monitorar a performance. Você pode usar o Amazon CloudWatch Synthetics para criar esses testes.

prompt do sistema

Uma técnica para fornecer contexto, instruções ou orientações a um LLM a fim de direcionar seu comportamento. Os prompts do sistema ajudam a definir o contexto e a estabelecer regras para interações com os usuários.

T

tags

Key-value pares que atuam como metadados para organizar seus AWS recursos. As tags podem ajudar você a gerenciar, identificar, organizar, pesquisar e filtrar recursos da . Para obter mais informações, consulte Marcar seus recursos do AWS.

variável-alvo

O valor que você está tentando prever no ML supervisionado. Ela também é conhecida como variável de resultado. Por exemplo, em uma configuração de fabricação, a variável-alvo pode ser um defeito do produto.

lista de tarefas

Uma ferramenta usada para monitorar o progresso por meio de um runbook. Uma lista de tarefas contém uma visão geral do runbook e uma lista de tarefas gerais a serem concluídas. Para cada tarefa geral, ela inclui o tempo estimado necessário, o proprietário e o progresso.

ambiente de teste

Veja ambiente.

treinamento

O processo de fornecer dados para que seu modelo de ML aprenda. Os dados de treinamento devem conter a resposta correta. O algoritmo de aprendizado descobre padrões nos dados de treinamento que mapeiam os atributos dos dados de entrada no destino (a resposta que você deseja prever). Ele gera um modelo de ML que captura esses padrões. Você pode usar o modelo de ML para obter previsões de novos dados cujo destino você não conhece.

ferramenta

Uma função ou API que um agente pode invocar para realizar operações em sistemas externos.

gateway de trânsito

Um hub de trânsito de rede que pode ser usado para interconectar as VPCs e as redes on-premises. Para obter mais informações, consulte O que é um gateway de trânsito na AWS Transit Gateway documentação.

fluxo de trabalho baseado em troncos

Uma abordagem na qual os desenvolvedores criam e testam recursos localmente em uma ramificação de recursos e, em seguida, mesclam essas alterações na ramificação principal. A ramificação principal é então criada para os ambientes de desenvolvimento, pré-produção e produção, sequencialmente.

Acesso confiável

Conceder permissões a um serviço que você especifica para realizar tarefas em sua organização AWS Organizations e em suas contas em seu nome. O serviço confiável cria um perfil vinculado ao serviço em cada conta, quando esse perfil é necessário, para realizar tarefas de gerenciamento para você. Para obter mais informações, consulte Usando AWS Organizations com outros AWS serviços na AWS Organizations documentação.

tuning (ajustar)

Alterar aspectos do processo de treinamento para melhorar a precisão do modelo de ML. Por exemplo, você pode treinar o modelo de ML gerando um conjunto de rótulos, adicionando rótulos e repetindo essas etapas várias vezes em configurações diferentes para otimizar o modelo.

equipe de duas pizzas

Uma pequena DevOps equipe que você pode alimentar com duas pizzas. Uma equipe de duas pizzas garante a melhor oportunidade possível de colaboração no desenvolvimento de software.

U

incerteza

Um conceito que se refere a informações imprecisas, incompletas ou desconhecidas que podem minar a confiabilidade dos modelos preditivos de ML. Há dois tipos de incertezas: a incerteza epistêmica é causada por dados limitados e incompletos, enquanto a incerteza aleatória é causada pelo ruído e pela aleatoriedade inerentes aos dados.

tarefas indiferenciadas

Também conhecido como trabalho pesado, trabalho necessário para criar e operar um aplicativo, mas que não fornece valor direto ao usuário final nem oferece vantagem competitiva. Exemplos de tarefas indiferenciadas incluem aquisição, manutenção e planejamento de capacidade.

ambientes superiores

Veja ambiente.

V

aspiração

Uma operação de manutenção de banco de dados que envolve limpeza após atualizações incrementais para recuperar armazenamento e melhorar a performance.

controle de versões

Processos e ferramentas que rastreiam mudanças, como alterações no código-fonte em um repositório.

emparelhamento de VPC

Uma conexão entre duas VPCs que permite rotear tráfego usando endereços IP privados. Para ter mais informações, consulte O que é emparelhamento de VPC? na documentação da Amazon VPC.

Vulnerabilidade

Uma falha de software ou hardware que compromete a segurança do sistema.

W

cache quente

Um cache de buffer que contém dados atuais e relevantes que são acessados com frequência. A instância do banco de dados pode ler do cache do buffer, o que é mais rápido do que ler da memória principal ou do disco.

dados mornos

Dados acessados raramente. Ao consultar esse tipo de dados, consultas moderadamente lentas geralmente são aceitáveis.

função de janela

Uma função SQL que executa um cálculo em um grupo de linhas que se relacionam de alguma forma com o registro atual. As funções de janela são úteis para processar tarefas, como calcular uma média móvel ou acessar o valor das linhas com base na posição relativa da linha atual.

workload

Uma coleção de códigos e recursos que geram valor empresarial, como uma aplicação voltada para o cliente ou um processo de backend.

workstreams

Grupos funcionais em um projeto de migração que são responsáveis por um conjunto específico de tarefas. Cada workstream é independente, mas oferece suporte aos outros workstreams do projeto. Por exemplo, o workstream de portfólio é responsável por priorizar aplicações, planejar ondas e coletar metadados de migração. O workstream de portfólio entrega esses ativos ao workstream de migração, que então migra os servidores e as aplicações.

WORM

Veja gravação única e várias leituras.

WQF

Veja AWS Workload Qualification Framework.

gravação única e várias leituras (WORM)

Um modelo de armazenamento que grava dados uma única vez e evita que os dados sejam excluídos ou modificados. Os usuários autorizados podem ler os dados quantas vezes forem necessárias, mas não podem alterá-los. Essa infraestrutura de armazenamento de dados é considerada imutável.

Z

exploração de dia zero

Um ataque, normalmente malware, que tira proveito de uma vulnerabilidade zero-day.

vulnerabilidade de dia zero

Uma falha ou vulnerabilidade não mitigada em um sistema de produção. Os agentes de ameaças podem usar esse tipo de vulnerabilidade para atacar o sistema. Os desenvolvedores frequentemente ficam cientes da vulnerabilidade como resultado do ataque.

prompt zero shot

Fornecer a um LLM instruções para realizar uma tarefa, mas sem exemplos (shots) que possam ajudar a orientá-lo. O LLM deve usar seu conhecimento pré-treinado para lidar com a tarefa. A eficácia dos prompts zero-shot depende da complexidade da tarefa e da qualidade do prompt. Veja também prompts few-shot.

aplicação zumbi

Uma aplicação que tem um uso médio de CPU e memória inferior a 5%. Em um projeto de migração, é comum retirar essas aplicações.

As traduções são geradas por tradução automática. Em caso de conflito entre o conteúdo da tradução e da versão original em inglês, a versão em inglês prevalecerá.