Opções e arquiteturas de geração aumentada de recuperação em AWS

AWS Recomendações

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

AWS Recomendações: Opções e arquiteturas de geração aumentada de recuperação em AWS

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

As marcas comerciais e imagens de marcas da Amazon não podem ser usadas no contexto de nenhum produto ou serviço que não seja da Amazon, nem de qualquer maneira que possa gerar confusão entre os clientes ou que deprecie ou desprestigie a Amazon. Todas as outras marcas comerciais que não pertencem à Amazon pertencem a seus respectivos proprietários, que podem ou não ser afiliados, patrocinados pela Amazon ou ter conexão com ela.

Table of Contents

Introdução 1
Público-alvo 1
Objetivos 2
Opções generativas de IA 3
Entendendo o RAG 4
Componentes 6
Comparando o RAG e o ajuste fino 7
Casos de uso do RAG 10
Opções de RAG totalmente gerenciadas 11
Bases de Conhecimento para Amazon Bedrock 11
Fontes de dados 13
bancos de dados vetoriais 15
Amazon Q Business 16
Recursos principais 16
End-user personalização 18
Amazon SageMaker AI Canvas 18
Arquiteturas RAG personalizadas 21
Recuperadores 21
Amazon Kendra 22
OpenSearch Serviço Amazon 24
Amazon Aurora PostgreSQL e pgvector 24
Amazon Neptune Analytics 25
Amazon MemoryDB 25
Amazon DocumentDB 27
Pinha 29
Atlas do MongoDB 30
Nós aviamos 31
Geradores 32
Amazon Bedrock 32
SageMaker AI JumpStart 33
Escolhendo uma opção RAG 34
Conclusão 36
Histórico do documento 37
..... xxxviii

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 PDFs, arquivos de texto, 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
- Fine-tune 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.

Fine-tuning 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.

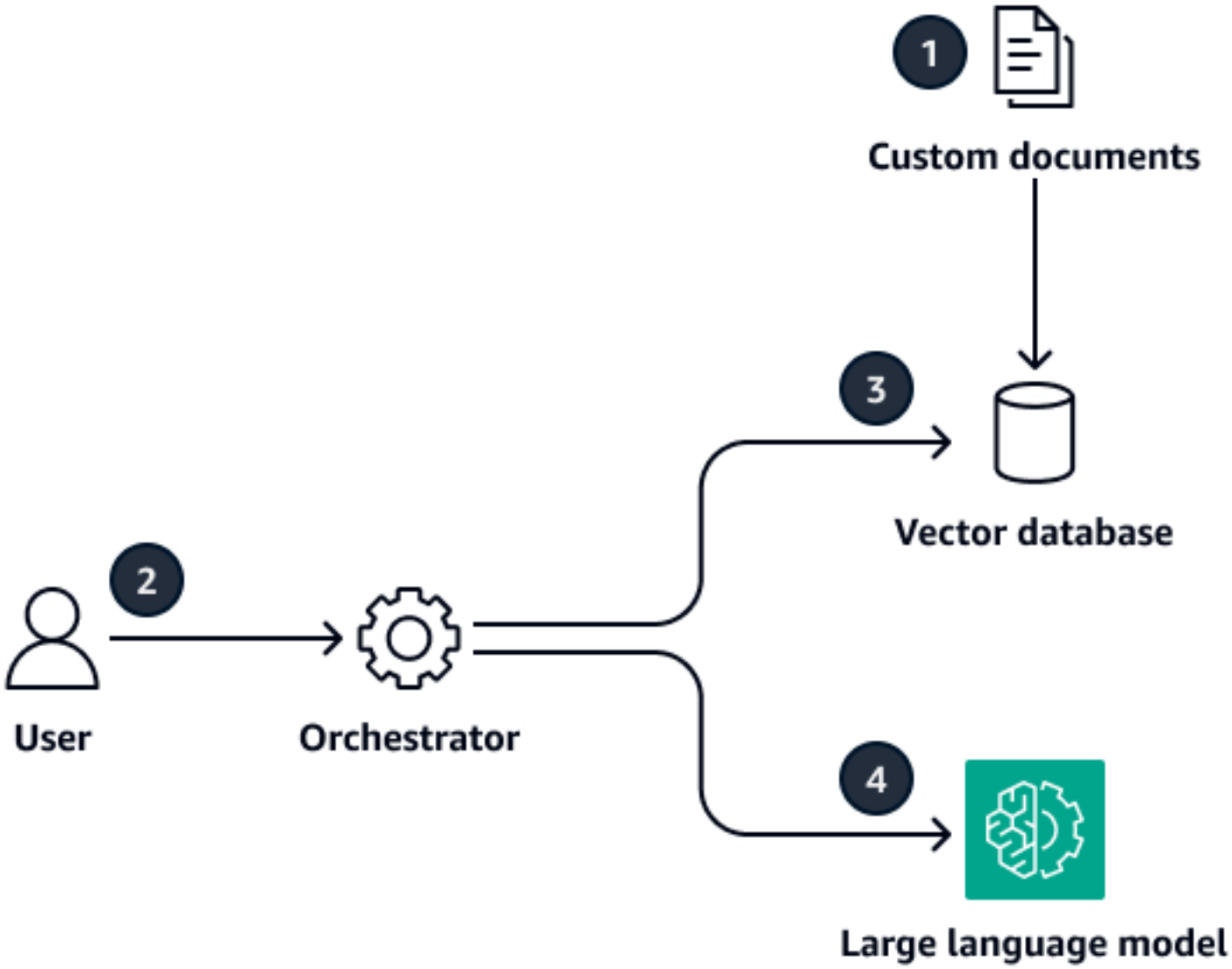
A opção final é usar o RAG. Com o RAG, o modelo básico faz referência aos seus documentos personalizados antes de gerar uma resposta. O RAG estende os recursos do modelo para a base de conhecimento interna da sua organização, tudo sem a necessidade de retreinar o modelo. É uma abordagem econômica para 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 o uso de 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, como PDFs, imagens digitalizadas, documentos, apresentações e arquivos. Microsoft SharePoint 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 de ponta a ponta.
- 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 do ajuste fino e RAG-based das abordagens.

Abordagem	Vantagens	Desvantagens
Fine-tuning	• Se um modelo aperfeiçoado for treinado usando a abordagem não supervisionada, ele poderá criar conteúdo que corresponda melhor ao estilo da sua organização.	• Fine-tuning pode levar de algumas horas a dias, dependendo do tamanho do modelo. Portanto, não é uma boa solução se seus documentos personalizados 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.	• Fine-tuning requer uma compreensão de técnicas, como adaptação de baixa classificação (LoRa) e ajuste fino com eficiênci a de parâmetros (PEFT). Fine-tuning pode exigir um cientista de dados. • Fine-tuning pode não estar disponível para todos os modelos. • Fine-tuned os modelos 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 aprendizado de máquina é necessário. • 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 alucinação.	• 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. RAG-based 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 RAG-enabled pesquisa — os mecanismos de pesquisa podem fornecer trechos em destaque mais precisos e atualizados em seus resultados de pesquisa.
- Question-answering sistemas — O RAG pode melhorar a qualidade das respostas em 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 RAG-based sistema, 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.

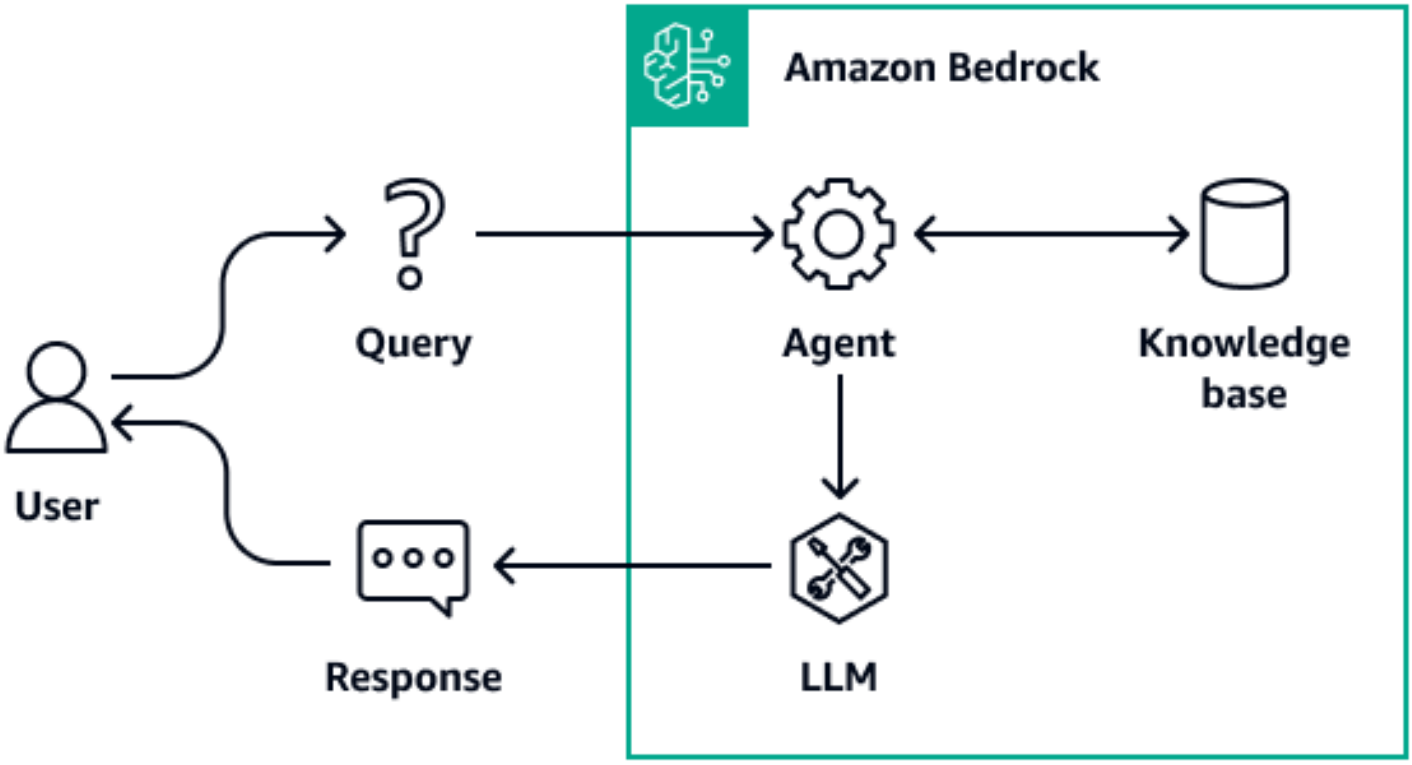
Bases de Conhecimento para Amazon Bedrock

O Amazon Bedrock é um serviço totalmente gerenciado que disponibiliza modelos de base (FMs) de alta performance, provenientes de startups líderes em IA e da própria Amazon, 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 às duas APIs a seguir 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. LLM-generated 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 de arquitetura e instruções passo a passo sobre o uso das APIs, consulte O Knowledge Bases agora oferece uma experiência RAG totalmente gerenciada no Amazon Bedrock AWS (postagem do 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, no título da página, no título do blog, no comentário, no nome do anexo ou na extensão.
  - Sincronização incremental - As alterações do conteúdo são monitoradas e somente o conteúdo que foi alterado desde a última sincronização é rastreado.
  - Autenticação OAuth 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.
  - Autenticação OAuth 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 Inclusion/exclusion filtros 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.

  - Autenticação OAuth 2.0 — As credenciais de autenticação são armazenadas em. AWS Secrets Manager
- Web Crawler — Um Amazon Bedrock Web Crawler se conecta e rastreia os URLs que você fornece. Os seguintes atributos são compatíveis:
  - Selecionar vários URLs para crawling.
  - Respeite as diretivas padrão do robots.txt, como e `Allow` `Disallow`
  - Excluir 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 do 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
- Edição Amazon Aurora PostgreSQL-Compatible
- 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

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

- Built-in conectores — O Amazon Q Business oferece suporte a 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.
- Built-in pipelines de indexação — 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 RAG-based generativo de IA. 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).

End-user personalização

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

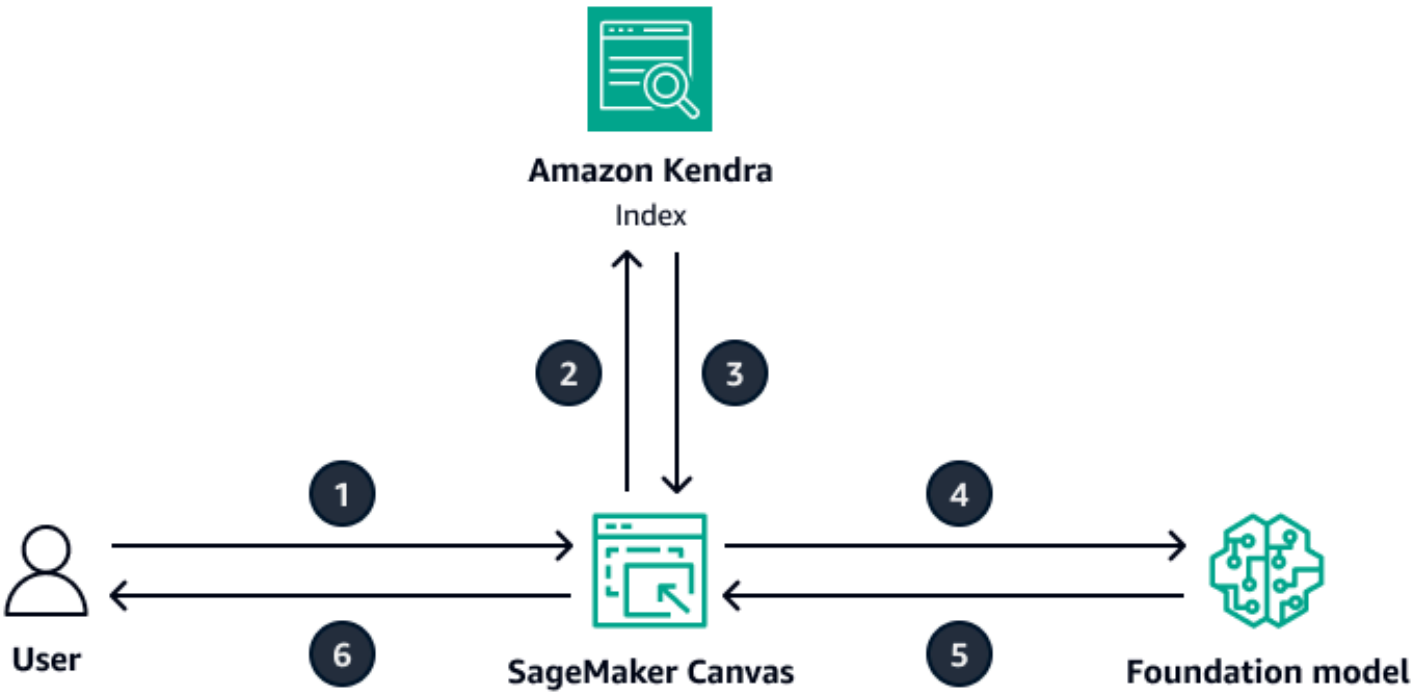
O Amazon SageMaker AI Canvas ajuda você a usar o aprendizado de máquina para gerar previsões sem precisar escrever nenhum código. Ele fornece uma interface visual sem código que permite que você prepare dados, crie e implante modelos de ML, simplificando o ciclo de vida de ML de ponta a ponta em um ambiente 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 AI Canvas. SageMaker

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
- Pinha
- Atlas do MongoDB
- Nós aviamos

Amazon Kendra

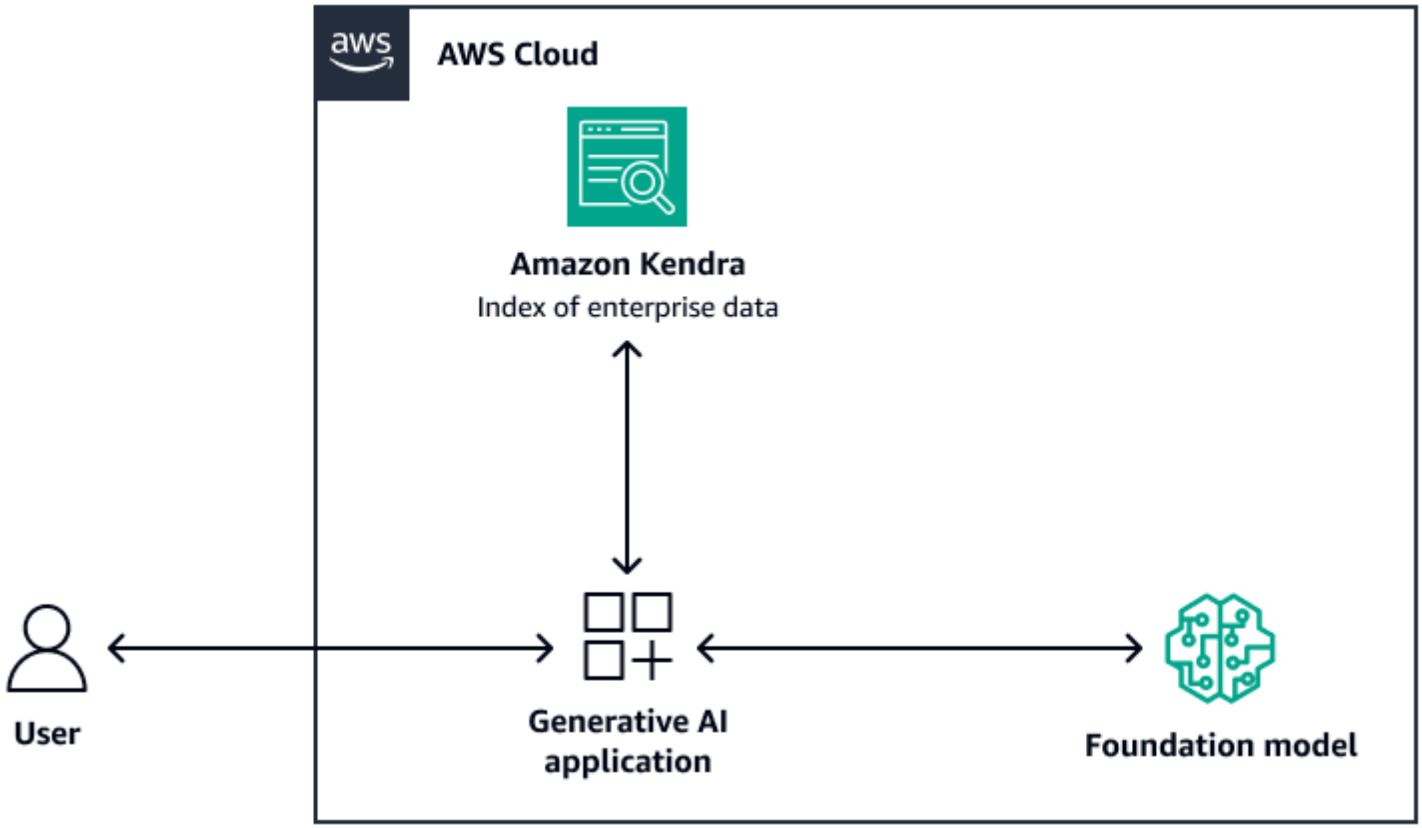
O Amazon Kendra é um serviço de pesquisa inteligente e totalmente gerenciado que usa processamento de linguagem natural e algoritmos avançados de aprendizado de máquina para retornar respostas específicas às perguntas de pesquisa de seus dados. O Amazon Kendra ajuda você a ingerir documentos diretamente 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 e vários LLMs, consulte o repositório Amazon Kendra Extensions. LangChain 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 alimentar 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

O Amazon Aurora PostgreSQL-Compatible Edition 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-Compatible quanto para o Amazon Relational Database Service (Amazon RDS) para PostgreSQL. Para obter mais informações sobre como criar um RAG-based sistema que usa o Aurora PostgreSQL-Compatible e o pgvector, consulte as seguintes postagens no blog: AWS

- Criando AI-powered pesquisa no PostgreSQL usando SageMaker Amazon AI e pgvector
- Utilize o pgvector e o Amazon Aurora PostgreSQL para processamento de linguagem natural, chatbots e análise de sentimentos

A seguir estão as vantagens de usar o pgvector e o Aurora: PostgreSQL-Compatible

- 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 PostgreSQL-Compatible é otimizado I/O e fornece armazenamento em cache em camadas. 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 de gráficos. 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 RAG-based sistema 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:

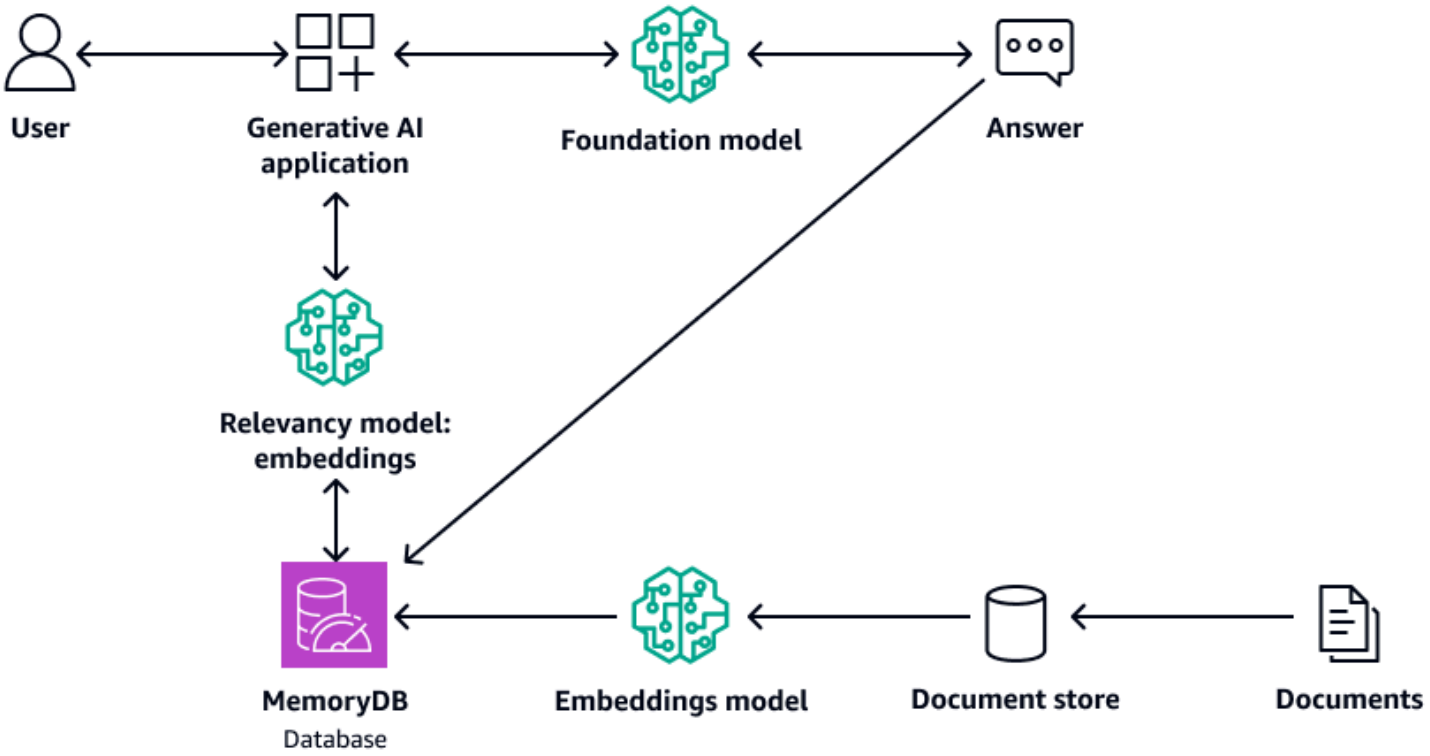
- Você pode armazenar e pesquisar incorporações em consultas gráficas.
- Se você integrar o Neptune Analytics LangChain com, essa arquitetura oferece suporte a consultas gráficas em linguagem natural.
- Essa arquitetura armazena grandes conjuntos de dados gráficos na memória.

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

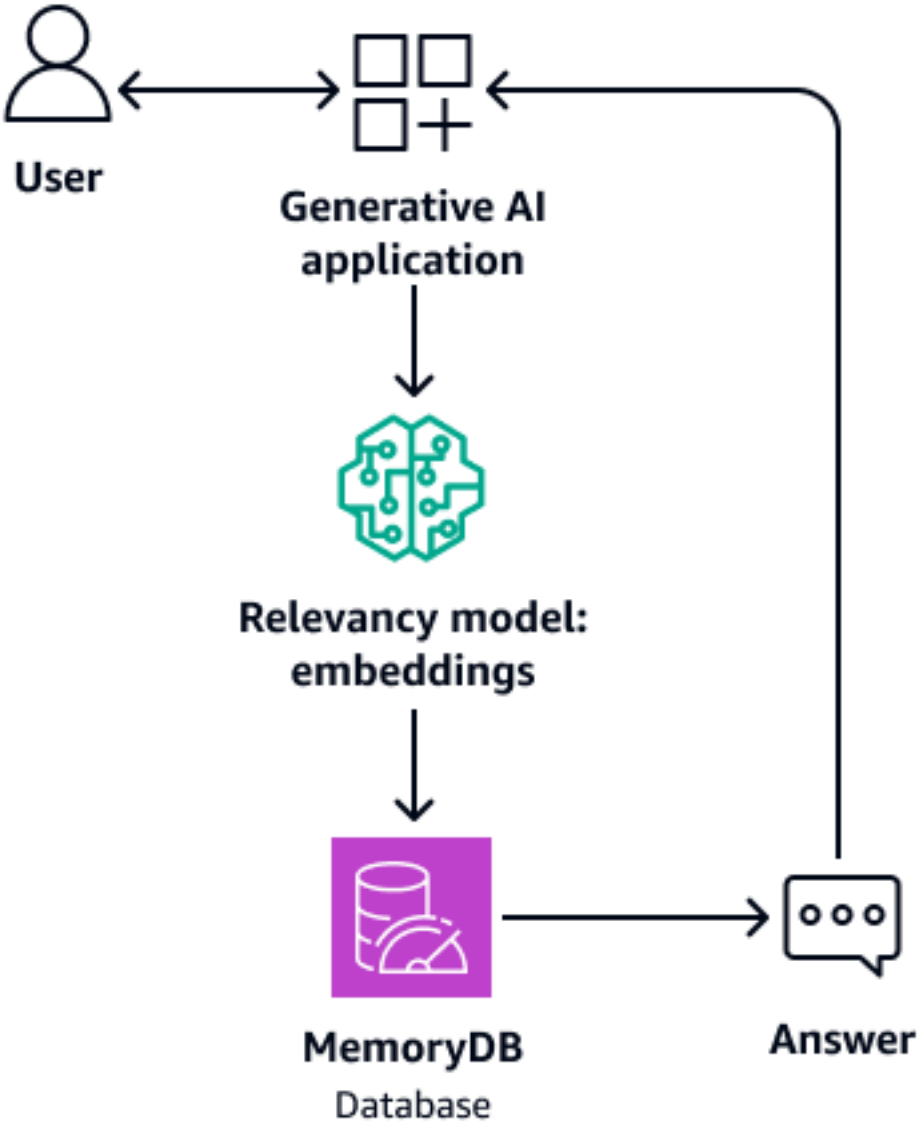
usada em conjunto com a funcionalidade existente do MemoryDB. Para obter mais informações, consulte a Resposta de perguntas com o repositório LLM e RAG ativado. GitHub

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.



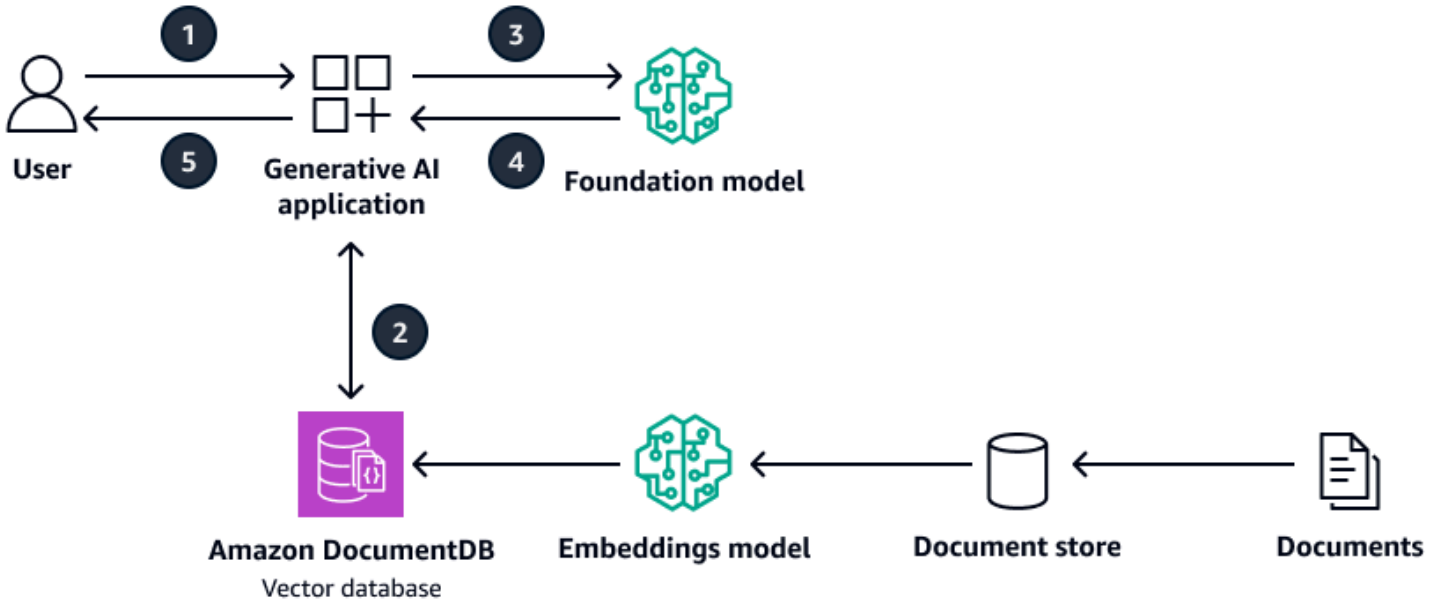
- Como usa um banco de dados na memória, essa arquitetura fornece um tempo de consulta de milissegundos de um dígito para a pesquisa semântica.
- Ele fornece até 33.000 consultas por segundo com 95— 99% de recall e 26.500 consultas por segundo com mais de 99% de recall. Para obter mais informações, consulte o vídeo AWS re:Invent 2023 - pesquisa vetorial de Ultra-low latência 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 JSON-based documentos 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 os métodos de indexação HNSW e IVFFlat.
- Ele suporta até 2.000 dimensões nos dados vetoriais e suporta as métricas de distância do produto euclidiano, cosseno e ponto.

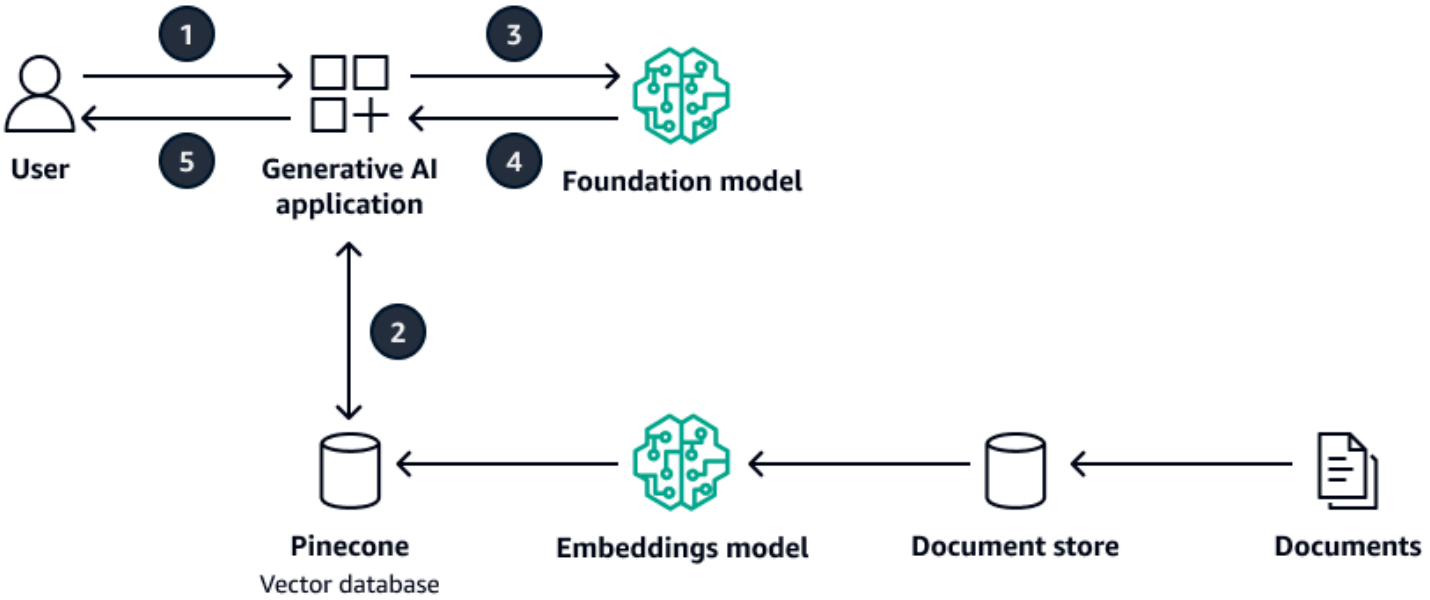
- Ele fornece tempos de resposta em milissegundos.

Pinha

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 RAG-based sistema que usaPinecone, consulte as seguintes postagens AWS no blog:

- Mitigue alucinações por meio do RAG usando Pinecone banco de dados vetoriais e da Amazon AI Llama-2 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.

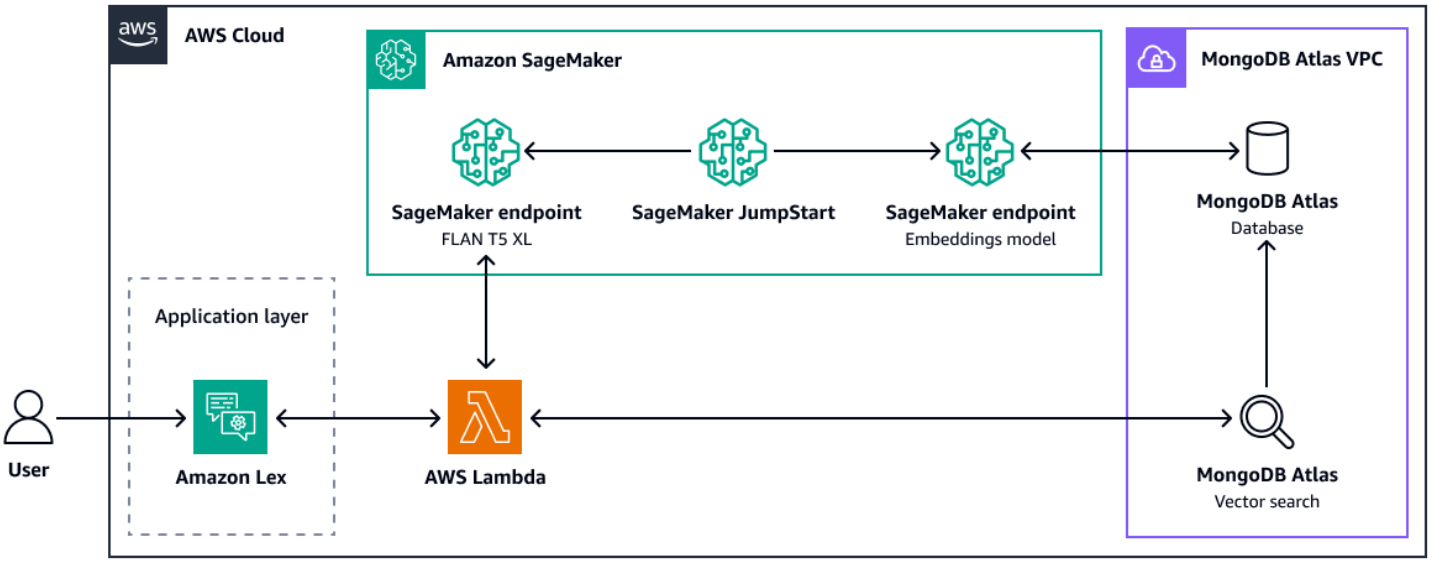
A seguir estão as vantagens de usarPinecone:

- 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).

Atlas do MongoDB

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 withLangChain, Amazon SageMaker AI JumpStart e MongoDB Atlas Semantic Search (postagem AWS no blog). 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.

Nós aviamos

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

Os modelos de linguagem grande (LLMs) são modelos de aprendizado profundo muito grandes que são pré-treinados em grandes quantidades de dados. Eles são incrivelmente flexíveis. Os LLMs podem 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 sejam perfeitos, os LLMs demonstram uma capacidade notável de fazer previsões com base em um prompt ou número de entradas relativamente pequeno.

Os LLMs são um componente crítico 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 LLMs das principais empresas de IA e da 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 e 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 os modelos pré-treinados, os modelos de solução e os 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 básicos de última geração para casos de uso, como redação 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

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.