Escolhendo um banco de dados AWS vetoriais para casos de uso do RAG

AWS Recomendações

AWS Recomendações: Escolhendo um banco de dados AWS vetoriais para casos de uso do RAG

Table of Contents

Escolhendo um banco de dados AWS vetoriais para casos de uso do RAG

Mayuri Shinde, Ivan Cui e Anand Bukkapatnam Tirumala, da Amazon Web Services

Março de 2026 (histórico do documento)

Os bancos de dados vetoriais estão se tornando cada vez mais importantes para organizações que implementam aplicativos generativos de IA. Esses bancos de dados armazenam e gerenciam vetores, que são representações numéricas de dados que permitem o processamento de texto, imagens e outros conteúdos de forma a capturar seu significado e seus relacionamentos.

À medida que as organizações exploram as opções de banco de dados vetoriais AWS, elas precisam entender os recursos, as vantagens e as melhores práticas de diferentes soluções. Este guia ajuda você a comparar lojas de vetores comumente usadas AWS e a tomar decisões informadas sobre quais opções melhor atendem às suas necessidades específicas ou ao seu caso de uso. Se você está implementando Retrieval Augmented Generation (RAG), criando sistemas de recomendação ou desenvolvendo outros aplicativos de IA, este guia fornece uma estrutura para ajudá-lo a avaliar e escolher uma solução de banco de dados vetorial.

Público-alvo

Este guia é destinado a pessoas nas seguintes funções:

- Cientistas de dados e engenheiros de aprendizado de máquina (ML) que usam bancos de dados vetoriais para armazenar e recuperar dados de alta dimensão para modelos de ML.
- Engenheiros de dados que projetam e implementam pipelines de dados que incluem bancos de dados vetoriais para armazenar e processar dados de alta dimensão.
- MLOps engenheiros que usam bancos de dados vetoriais como parte do pipeline de ML para armazenar e fornecer saídas de modelos ou representações intermediárias.
- Engenheiros de software que integram bancos de dados vetoriais em aplicativos que exigem sistemas de recomendação ou pesquisa por similaridade.
- DevOps engenheiros responsáveis pela implantação e manutenção de bancos de dados vetoriais em ambientes de produção, garantindo escalabilidade e confiabilidade.
- Pesquisadores de IA que usam bancos de dados vetoriais para armazenar e analisar grandes conjuntos de dados de incorporações ou vetores de recursos.

- Gerentes de produtos de IA que precisam entender os recursos e as limitações dos bancos de dados vetoriais para tomar decisões informadas sobre os recursos e a arquitetura do produto.

Visão geral dos vetores

Os vetores são representações numéricas que ajudam as máquinas a entender e processar dados. Na IA generativa, eles servem a dois propósitos principais:

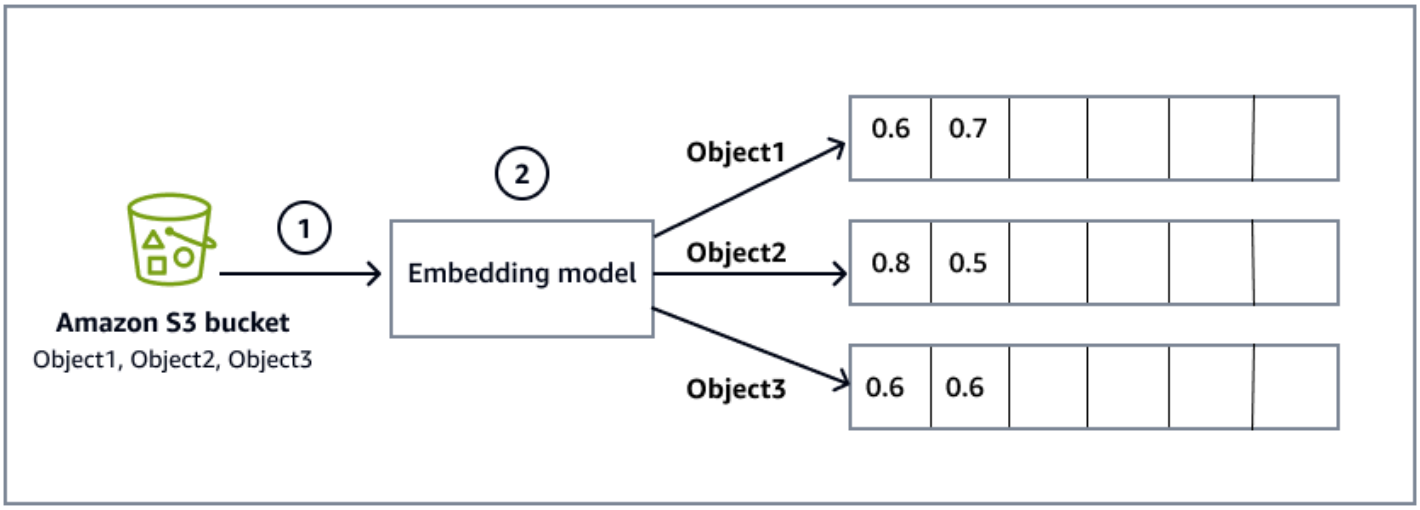
- Representando espaços latentes que capturam a estrutura de dados em formato comprimido
- Criação de incorporações para dados, como palavras, frases e imagens

Modelos de incorporação como Word2Vec GloVee Amazon Titan Text Embeddings convertem dados em vetores por meio de um processo chamado incorporação. Esses modelos de incorporação podem fazer o seguinte:

- Aprenda com o contexto para representar palavras como vetores
- Coloque palavras similares mais próximas umas das outras no espaço vetorial
- Permita que as máquinas processem dados em um espaço contínuo

O diagrama a seguir fornece uma visão geral de alto nível do processo de incorporação:

1. Um bucket do Amazon Simple Storage Service (Amazon S3) contém arquivos que são as fontes de dados a partir das quais o sistema lerá e processará informações. O bucket do Amazon S3 é especificado durante a configuração da base de conhecimento Amazon Bedrock, que também inclui a sincronização de dados com a base de conhecimento.
2. O modelo de incorporação converte os dados brutos dos arquivos de objetos no bucket do Amazon S3 em incorporações vetoriais. Por exemplo, `Object1` é convertido em um vetor `[0.6, 0.7, ...]` que representa seu conteúdo em um espaço multidimensional.



As incorporações de palavras são cruciais para o processamento de linguagem natural (PNL) porque fazem o seguinte:

- Capture relações semânticas entre palavras
- Permita a geração de texto contextualmente relevante
- Potencialize modelos de linguagem grande (LLMs) para produzir respostas semelhantes às humanas

Visão geral dos bancos de dados vetoriais

Um banco de dados vetoriais é um sistema especializado que armazena e consulta vetores de alta dimensão com eficiência. Esses bancos de dados são fundamentais para aplicativos de Retrieval Augmented Generation (RAG).

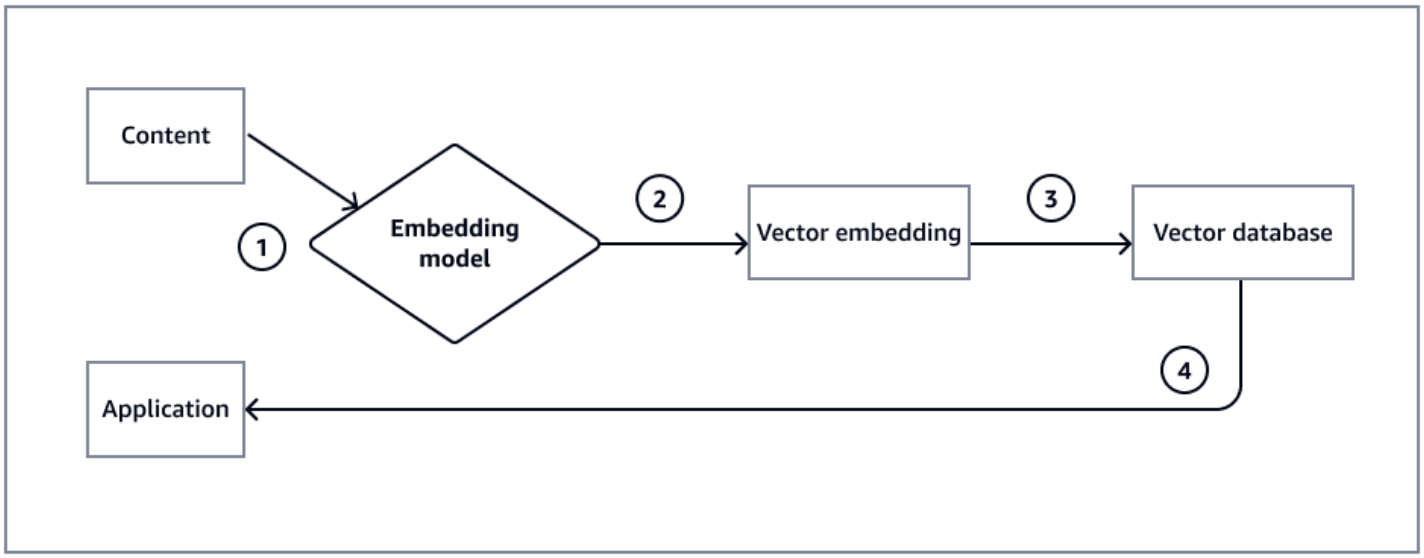
Os bancos de dados vetoriais lidam com a conversão e o armazenamento de dados das seguintes maneiras:

- Objetos (como arquivos de áudio, imagens e texto) são convertidos em vetores usando modelos de incorporação.
- Os vetores são armazenados em formatos de dados especializados.
- Os bancos de dados vetoriais permitem pesquisas rápidas por similaridade.

Os bancos de dados vetoriais oferecem várias vantagens importantes em relação aos bancos de dados tradicionais, o que os torna particularmente adequados para os desafios de dados modernos. Eles são especificamente otimizados para operações vetoriais e lidam com dados de alta dimensão com eficiência. Eles também se especializam em pesquisas por similaridade que os bancos de dados tradicionais enfrentam dificuldades. Além desses recursos principais, os bancos de dados vetoriais são criados para atender às crescentes demandas dos aplicativos de ML e IA generativa. Eles se destacam no armazenamento vetorial em grande escala e usam a computação distribuída para equilibrar as cargas de trabalho em vários nós. Isso fornece escalabilidade e desempenho à medida que os volumes de dados aumentam.

O diagrama a seguir mostra uma implementação do RAG:

1. O conteúdo, como documentos, PDFs ou arquivos de texto, é inserido no modelo de incorporação como dados brutos para processamento.
2. O modelo de incorporação transforma os dados brutos em vetores numéricos, que representam o significado semântico do conteúdo.
3. As incorporações vetoriais geradas são armazenadas em um banco de dados vetoriais otimizado para o armazenamento e a recuperação de vetores de alta dimensão.
4. Agora, os aplicativos podem consultar o banco de dados vetoriais em resposta a casos de uso, como pesquisa semântica e recomendação de conteúdo.



A escolha de um banco de dados vetorial inadequado para uma solução RAG pode levar a dificuldades e limitações significativas, incluindo as seguintes:

- Baixo desempenho da consulta
- Gargalos de escalabilidade
- Desafios de ingestão de dados
- Falta de recursos avançados, como filtragem e classificação
- Dificuldades de integração com outros sistemas
- Preocupações com persistência e durabilidade
- Problemas de simultaneidade e consistência em ambientes com vários usuários
- Custos de licenciamento mais altos ou dependência de fornecedor
- Suporte e recursos comunitários limitados
- Riscos potenciais de segurança e conformidade

Opções de banco de dados vetoriais

AWS oferece uma ampla variedade de soluções de banco de dados vetoriais para dar suporte a diferentes casos de uso e requisitos em aplicativos generativos de IA. Essas opções podem ser amplamente categorizadas em serviços de banco de dados individuais e ofertas de serviços gerenciados, cada uma com características e vantagens distintas. Compreender essas opções é crucial para organizações que desejam implementar recursos de pesquisa vetorial de forma eficaz, mantendo o desempenho, a escalabilidade e a eficiência de custos ideais.

Para obter mais informações sobre soluções de banco de dados vetoriais, consulte as seções a seguir:

- Opções individuais de banco de dados vetoriais
- Opção de serviço gerenciado
- Escolhendo o banco de dados vetorial correto

Opções individuais de banco de dados vetoriais

As opções individuais de banco de dados vetoriais AWS incluem Amazon Kendra, AmazonService, OpenSearch AmazonRDS para PostgreSQL com pgvector, AmazonMemoryDB, Amazon DocumentDB, Amazon Neptune Analytics e Amazon S3 Vector. (Uma extensão de código aberto, o pgvector adiciona a capacidade de armazenar e pesquisar incorporações vetoriais geradas por ML.) Essas soluções oferecem abordagens diferentes para a pesquisa vetorial, permitindo que as organizações escolham com base em sua infraestrutura existente, requisitos técnicos e casos de uso específicos.

Amazon Kendra

O Amazon Kendra é um serviço de pesquisa inteligente de nível corporativo 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 simplifica a implementação da funcionalidade de pesquisa, tornando-o uma solução de back-end eficaz para aplicativos generativos de IA.

Outros recursos importantes do Amazon Kendra incluem o seguinte:

- Conexões nativas com mais de 40 fontes de dados
- Recursos integrados de preparação de dados
- Configuração rápida que não requer profundo conhecimento técnico

Os benefícios do Amazon Kendra incluem o seguinte

- Processamento automatizado de dados (fragmentação, ingestão, recuperação)
- Opções poderosas de personalização:
  - Pesquisa de facetas
  - Análise de pesquisa
  - Ajustando a relevância da pesquisa
- Acesso programático simples por meio do AWS SDK para Python (Boto3)

Para obter mais informações, consulte Benefícios do Amazon Kendra na documentação do Amazon Kendra.

OpenSearch Serviço Amazon

O Amazon OpenSearch Service é um serviço gerenciado que ajuda você a implantar, operar e escalar clusters de OpenSearch serviços no Nuvem AWS.

Os principais recursos do OpenSearch Serviço incluem o seguinte:

- Mecanismo de pesquisa e análise de código aberto
- Arquitetura distribuída
- Processamento de dados em tempo real

Algumas vantagens de usar o OpenSearch Serviço incluem o seguinte:

- Escalabilidade horizontal
- RESTful Suporte à API
- Lida com dados estruturados e não estruturados
- Análise de dados em tempo real
- Adequado para vários tamanhos de implantação

Para obter mais informações, consulte Recursos do Amazon OpenSearch Service na documentação do OpenSearch serviço.

Amazon RDS para PostgreSQL com pgvector

O Amazon RDS for PostgreSQL com pgvector AWS combina o serviço de banco de dados relacional gerenciado com a extensão de processamento vetorial do PostgreSQL. Essa combinação permite que as organizações armazenem e consultem vetores de alta dimensão enquanto mantêm o Amazon RDS. A solução é particularmente adequada para aplicativos generativos de IA que exigem operações vetoriais em tempo real sem a sobrecarga de gerenciar a infraestrutura do banco de dados.

Os principais benefícios do Amazon RDS for PostgreSQL com pgvector incluem o seguinte:

- Alta disponibilidade
- Failover automático
- Econômico () pay-per-use
- Monitoramento integrado
- Integração de dados vetoriais em tempo real

Para obter mais informações, consulte Vantagens do Amazon RDS na documentação do Amazon RDS.

Amazon DocumentDB

O Amazon DocumentDB (com compatibilidade com o MongoDB) é um banco de dados de documentos que oferece recursos de pesquisa vetorial nativa na versão 5.0 e posterior. Ele combina a flexibilidade do armazenamento de documentos baseado em JSON com a pesquisa vetorial, suportando os métodos de indexação hierárquica navegável de pequeno mundo (HNSW) e Inverted File Flat (). IVFFlat

Os principais recursos do Amazon DocumentDB incluem o seguinte:

- Armazene e indexe vetores de até 2.000 dimensões (até 16.000 dimensões sem indexação)
- Tempos de resposta em milissegundos para pesquisas de similaridade vetorial
- Support para métricas de distância de produtos euclidianos, cossenos e pontos
- Integração perfeita com aplicativos existentes compatíveis com MongoDB

Use o Amazon DocumentDB nas seguintes situações:

- Para aplicativos que já estão usando o APIs MongoDB e que precisam de recursos de pesquisa vetorial
- Para casos de uso que exigem estruturas flexíveis de dados de documentos combinadas com pesquisa semântica
- Para cenários que precisam tanto de consultas tradicionais de documentos quanto de pesquisas por similaridade vetorial
- Para aplicativos que fornecem recomendações de produtos, personalização, assistentes de bate-papo e detecção de fraudes

Para obter mais informações, consulte Pesquisa vetorial para Amazon DocumentDB na documentação do Amazon DocumentDB.

Amazon MemoryDB

O Amazon MemoryDB é um banco de dados em memória compatível com Redis que oferece o desempenho de pesquisa vetorial mais rápido entre os bancos de dados vetoriais mais populares. AWS Ele fornece latências de consulta inferiores a um milissegundo com durabilidade de zona de disponibilidade múltipla.

Os principais recursos do MemoryDB incluem o seguinte:

- Armazene dados de aplicativos e milhões de vetores em um único banco de dados
- Tempos de resposta de consulta e atualização de um dígito em milissegundos
- As maiores taxas de recall com o desempenho mais rápido em AWS
- Support para até 32.768 dimensões por vetor
- Recursos de pesquisa semântica e armazenamento em cache em tempo real

Use o MemoryDB nas seguintes situações:

- Para aplicativos em tempo real que exigem latência ultrabaixa (menos de 10 ms)
- Para cargas de trabalho de alto rendimento com milhões de solicitações por dia
- Para casos de uso como mecanismos de recomendação em tempo real, cache semântico e detecção de anomalias

- Para aplicativos que precisam de recursos de armazenamento de dados na memória e pesquisa vetorial

Para obter mais informações, consulte Pesquisa vetorial na documentação do MemoryDB.

Amazon Neptune Analytics

O Amazon Neptune Analytics é um mecanismo de análise gráfica que oferece recursos de pesquisa vetorial nativa, tornando-o ideal para casos de uso da Geração Aumentada de Recuperação de Gráficos (GraphRag). Ele combina pesquisa de similaridade vetorial com percursos gráficos e algoritmos.

Os principais recursos do Neptune Analytics incluem o seguinte:

- Analise dezenas de bilhões de relacionamentos em segundos
- Combine a pesquisa vetorial com algoritmos gráficos (localização de caminhos, detecção de comunidades, centralidade)
- Support para aplicativos GraphRag com conhecimento topológico
- Até 80 vezes mais rápido do que as soluções analíticas gráficas existentes
- Integração com o Amazon Bedrock para GraphRag totalmente gerenciado

Use o Neptune Analytics nas seguintes situações:

- Para aplicativos GraphRag que exigem gráficos de conhecimento com incorporações vetoriais
- Para casos de uso que exigem a travessia de relacionamentos complexos ao lado da similaridade vetorial
- Para aplicativos que exigem respostas explicáveis de IA com contexto de relacionamento
- Para cenários como visualizações 360 de clientes, redes de detecção de fraudes e descoberta de conhecimento

Para obter mais informações, consulte a documentação do Amazon Neptune Analytics.

Amazon S3 Vectors

O Amazon S3 Vectors é o primeiro armazenamento de objetos em nuvem AWS com armazenamento vetorial nativo e recursos de consulta. Ele fornece armazenamento vetorial personalizado e com custo otimizado para aplicativos de IA que exigem grande escala.

Os principais recursos dos Amazon S3 Vectors incluem o seguinte:

- Armazenamento para até 2 bilhões de vetores por índice com suporte para até 10.000 índices por compartimento vetorial
- Latência de consulta abaixo de 100 ms otimizada para armazenamento de longo prazo e padrões de acesso pouco frequentes
- Redução de custos de até 90% para operações vetoriais em comparação com bancos de dados vetoriais especializados
- Arquitetura sem servidor com escalabilidade automática e durabilidade de 99,999999999% (11 9s)

Use os vetores do Amazon S3 nas seguintes situações:

- Para aplicativos que exigem armazenamento de bilhões de vetores a um custo mínimo
- Para cargas de trabalho que toleram latência de consulta inferior a um segundo (100 ms ou mais) em vez de menos de 10 ms
- Para casos de uso de arquivamento e retenção de vetores a longo prazo
- Para aplicações RAG com padrões de recuperação pouco frequentes
- Para organizações que priorizam a economia de armazenamento em vez da latência ultrabaixa

O Amazon S3 Vectors se integra nativamente às bases de conhecimento do Amazon Bedrock e funciona bem em arquiteturas hierárquicas com o Amazon Service. OpenSearch Você pode usar o Amazon S3 Vectors para armazenamento a frio e usar o OpenSearch Service para consultas dinâmicas.

Para obter mais informações, consulte Como trabalhar com vetores e buckets vetoriais do S3 na documentação do Amazon S3.

Opção de serviço gerenciado

O Amazon Bedrock Knowledge Bases representa a abordagem AWS totalmente gerenciada para a implementação de bancos de dados vetoriais. A flexibilidade do serviço nas opções de armazenamento, combinada com seus recursos de gerenciamento automatizado, o torna particularmente valioso para organizações que buscam implementar o RAG sem gerenciar uma infraestrutura complexa.

Com o Amazon Bedrock Knowledge Bases, você pode criar, manter e consultar bases de conhecimento que aprimoram seus modelos básicos usando o RAG. Esse serviço simplifica o processo complexo de implementação do RAG gerenciando todo o pipeline de ingestão, vetorização e recuperação de dados.

Os principais benefícios das bases de conhecimento Amazon Bedrock incluem o seguinte:

- Processamento de dados simplificado
  - Ingestão e fragmentação automáticas de dados
  - Extração de texto integrada de vários formatos de arquivo
  - Geração gerenciada de incorporações vetoriais
  - Extração e indexação automáticas de metadados
- Implementação simplificada do RAG
  - Estratégias de recuperação pré-configuradas
  - Otimização automática da janela de contexto
  - Ajuste de relevância integrado
  - Recursos de pesquisa semântica prontos para uso
- Segurança e governança
  - Controles integrados AWS Identity and Access Management (IAM)
  - Criptografia de dados em repouso e em trânsito
  - Suporte à VPC
  - Registro de auditoria com AWS CloudTrail

O Amazon Bedrock Knowledge Bases oferece suporte a várias opções de armazenamento de vetores, incluindo:

- Amazon Aurora PostgreSQL com pgvector

- Amazon Neptune Analytics
- Amazon EMR Sem Servidor
- Amazon S3 Vectors
- Pinha
- Nuvem empresarial Redis

Esse serviço gerenciado gerencia a ingestão, a vetorização e a recuperação automatizadas de dados. Isso simplifica as implementações do RAG.

Para obter informações detalhadas sobre cada loja de vetores compatível, consulte a documentação do Amazon Bedrock Knowledge Bases.

Escolhendo o banco de dados vetorial correto

Selecione seu banco de dados vetoriais com base nesses principais fatores de decisão:

- Se você precisar de um banco de dados de documentos compatível com MongoDB com pesquisa vetorial, escolha Amazon DocumentDB. Isso é ideal quando seu aplicativo usa o APIs MongoDB e você deseja adicionar recursos de pesquisa semântica sem gerenciar uma infraestrutura vetorial separada.
- Se você precisar de latência ultrabaixa para aplicativos em tempo real, escolha o Amazon MemoryDB. Isso fornece o desempenho de pesquisa vetorial mais rápido AWS com tempos de resposta abaixo de um milissegundo. É ideal para mecanismos de recomendação em tempo real e aplicações de alto rendimento.
- Se você precisar de representações de conhecimento baseadas em gráficos com pesquisa vetorial, escolha o Amazon Neptune Analytics. Isso é melhor para aplicativos GraphRag que precisam atravessar relacionamentos complexos e realizar consultas baseadas em gráficos junto com pesquisas vetoriais, fornecendo respostas de IA explicáveis.
- Se você precisar combinar consultas relacionais com pesquisa vetorial, escolha Amazon Aurora PostgreSQL com pgvector. Essa opção é ideal quando seu aplicativo exige operações SQL tradicionais e pesquisas de similaridade vetorial no mesmo banco de dados.
- Se você precisar de consultas de alto rendimento com latência inferior a 10 ms, escolha o Amazon Service. OpenSearch Ele se destaca no tratamento de consultas de alta frequência e aplicativos em tempo real e inclui melhorias recentes na aceleração da GPU.

- Se você precisar armazenar bilhões de vetores de forma econômica, escolha Amazon S3 Vectors. Essa opção oferece até 90% de economia de custos e é ideal para aplicativos com padrões de recuperação pouco frequentes (minutos a horas entre consultas) que podem tolerar latência abaixo de 100 ms.
- Se você precisar de uma pesquisa de texto completo junto com a pesquisa vetorial, escolha o Amazon OpenSearch Service. Essa opção combina recursos poderosos de pesquisa de texto completo com pesquisa vetorial em uma única plataforma.

Comparação de bancos de dados vetoriais

AWS fornece várias abordagens para implementar recursos de pesquisa vetorial, desde bancos de dados vetoriais individuais até o Amazon Bedrock Knowledge Bases, que é um serviço totalmente gerenciado. Ao avaliar essas opções, as organizações devem considerar vários aspectos, incluindo arquitetura, escalabilidade, recursos de integração, características de desempenho e recursos de segurança.

Bancos de dados vetoriais individuais

A tabela a seguir fornece uma visão geral dos principais recursos de várias soluções AWS individuais de banco de dados vetoriais, com foco em suas arquiteturas, recursos de escalabilidade, integrações de fontes de dados e características de desempenho.

Recurso	Amazon Kendra	OpenSearch Serviço Amazon	Amazon RDS para SQLwith Postgre pgvector	Amazon DocumentDB	Amazon MemoryDB	Amazon Neptune Analytics	Amazon S3 Vectors
Caso de uso principal	Pesquisa corporativa e RAG	Pesquisa e análise distribuídas	Banco de dados relacional com suporte vetorial	Banco de dados de documentos com pesquisa vetorial	Pesquisa vetorial na memória em tempo real	Análise gráfica com pesquisa vetorial	Armazenamento vetorial com custo otimizado
Arquitetura	Totalmente gerenciado	Cluster distribuído	Banco de dados relacional	Orientado a documentos	Banco de dados na memória	Mecanismo de análise gráfica	Armazenamento de objetos sem servidor

Modelo de dados	Baseado em documentos	Documentos JSON	Tabelas relacionais	Documentos JSON	Valor-chave com JSON	Gráfico de propriedades	Armazenamento de objetos
Dimensões vetoriais	Gerenciado automaticamente	Até 16.000	Configurável	Até 2.000 (indexado); 16.000 (não indexado)	Até 32.768	Configurável	Até 4.096
Métodos de indexação	Automatico	NSW, FERTILIZAÇÃO IN VITRO	HNSW, IVFFlat	HNSW, IVFFlat	HNSW	Gráfico e vetor nativos	Automatico
Métricas de distância	Automatico	Cosseno, euclidiano, produto pontilhado	Cosseno, euclidiano, produto interno	Cosseno, euclidiano, produto pontilhado	Cosseno, euclidiano, produto interno	Cosseno, Euclidiano	Cosseno, Euclidiano
Latência da consulta	Inferior a um segundo	Menos de 10 ms (acelerado por GPU)	10-100 ms	Milissegundo	Submilissegundo	Inferior a um segundo	Menos de 100 ms
Modelo de escalabilidade	Automatico	Horizontal (adicionar nós)	Réplicas verticais e de leitura	Horizontal (adicionar instâncias)	Vertical e réplicas	Automatico	Automático (sem servidor)

Vetores máximos	Gerenciados	Bilhões (dependente do cluster)	Milhões (dependendo da instância)	Milhões por coleção	Milhões por banco de dados	Bilhões	2 bilhões por índice; 10.000 índices por bucket
Throughput	Alto	Muito alto (milhares de QPS)	Médio	Alto	Muito alto (milhões de solicitações por dia)	Alto	Médio (otimizado para consultas pouco frequentes)
Durabilidade de dados	99.999999999% (11 9s)	Configurável com réplicas	99,99% (Multi-AZ)	99,99% (Multi-AZ)	99,99% (Multi-AZ)	99,99%	99.999999999% (11 9s)
Modelo de consistência	Eventual	Eventual (configurável)	Forte (ÁCIDO)	Eventual	Forte	Forte	Forte
Recursos adicionais	40 ou mais conectores de dados, NLP	Pesquisa de texto completo, análises e painéis	Consultas SQL, transações ACID	Compatibilidade com a API MongoDB	Compatibilidade com a API Redis, armazenamento em cache	Algoritmos gráficos, travessias	Integração com o Amazon S3, políticas de ciclo de vida

Modelo de definição de preços	Pague por consulta e armazenamento	Horas de instância e armazenamento	Horas de instância e armazenamento	Horas de instância e armazenamento	Horas de instância e armazenamento	Unidades de capacidade e armazenamento	Armazenamento, consultas e transferência de dados
Otimização de custos	Baseado no uso	Instâncias reservadas, auto-scaling	Instâncias reservadas, Aurora sem servidor	Instâncias reservadas	Instâncias reservadas	Ajuste de escala automático	Economia de até 90% em comparação com produtos especializados DBs
Melhor para	Pesquisa corporativa com configuração mínima	Consultas de alto rendimento e baixa latência	SQL híbrido e cargas de trabalho vetoriais	Aplicativos compatíveis com MongoDB que precisam de vetores	Aplicativos em tempo real com latência ultrabaixa	GraphRag e gráficos de conhecimento	Armazenamento econômico e de longo prazo
Padrão de consulta ideal	Pesquisas corporativas frequentes	Consultas em tempo real de alta frequência	Consultas SQL e vetoriais mistas	Consultas de documentos com pesquisa semântica	Milhões de solicitações por dia	Travessias gráficas com pesquisa vetorial	Consultas pouco frequentes (minutos a horas)

Complexidade da configuração	Baixo (totalmente gerenciado)	Médio (configuração de cluster)	Médio (configuração de extensão)	Médio (configuração de cluster)	Médio (configuração de cluster)	Baixo (totalmente gerenciado)	Baixo (sem servidor)
É necessária a experiência da equipe	Mínimo	OpenSearch ou Elasticsearch	PostgreSQL, SQL	MongoDB	Redis	Bancos de dados gráficos	Amazon S3, conceitos vetoriais básicos

Serviço gerenciado — Bases de conhecimento Amazon Bedrock

O Amazon Bedrock Knowledge Bases fornece uma solução totalmente gerenciada com várias opções de armazenamento vetorial. A tabela a seguir compara essas opções de armazenamento.

Recurso	Vetor de imagem Aurora Postgre SQLwith	Análise do Neptune	OpenSearch Serviço sem servidor	Vetores do Amazon S3	Pinha	RedisEnterprise Nuvem
Caso de uso principal	Banco de dados relacional com RAG vetorial	Pesquisa vetorial baseada em gráficos para GraphRag	Gestão do conhecimento RAG	RAG vetorial com custo otimizado	Pesquisa vetorial de alto desempenho	Pesquisa vetorial na memória
Arquitetura	Relacional totalmente gerenciado	Análise gráfica totalmente gerenciada	Totalmente gerenciado sem servidor	Armazenamento de objetos sem servidor	Nuvem híbrida totalmente gerenciada	Totalmente gerenciado na memória

Modelo de dados	Tabelas relacionais	Gráfico de propriedades	Documentos JSON	Armazenamento de objetos	Vetores criados especificamente	Valor-chave com vetores
Armazenamento vetorial	Por meio da extensão pgvector	Vetores gráficos nativos	Através do OpenSearch motor	Armazenamento vetorial nativo do Amazon S3	Banco de dados de vetores nativos	Vetores na memória
Integração do Amazon Bedrock	Nativo	Nativo	Nativo	Nativo	Nativo	Nativo
Ingestão automática	Sim (via Amazon Bedrock)	Sim (via Amazon Bedrock)	Sim (via Amazon Bedrock)	Sim (via Amazon Bedrock)	Sim (via Amazon Bedrock)	Sim (via Amazon Bedrock)
Vetorização automática	Sim (via Amazon Bedrock)	Sim (via Amazon Bedrock)	Sim (via Amazon Bedrock)	Sim (via Amazon Bedrock)	Sim (via Amazon Bedrock)	Sim (via Amazon Bedrock)
Escalabilidade	Escalonamento automático (Aurora sem servidor)	Escalonamento automático de gráficos	Automático sem servidor	Automático (bilhões de vetores)	Pods de escalonamento automático	Clusters de escalabilidade automática
Performance da consulta	Alto para relacional ou vetorial	Alto para vetores gráficos	Alto	Médio (100 ms ou mais de latência)	Muito alto	Muito alto
Vetores máximos	Milhões (dependendo da instância)	Bilhões	Bilhões	2 bilhões por índice	Bilhões	Milhões (dependendo da memória)

Recursos adicionais	Consultas SQL, transações ACID	Algoritmos gráficos, travessias	Pesquisa de texto completo, análise	Ciclo de vida, hierarquização do Amazon S3	Filtragem de metadados, namespaces	Estruturas de dados Redis, armazenamento em cache
Otimização de custos	Moderado (Aurora Serverless)	Moderado (unidades de capacidade)	Alto (sem servidor,) pay-per-use	Muito alto (economia de até 90%)	Moderado (preços baseados em cápsulas)	Baixo (premium em memória)
Melhor para	Cargas de SQL/vector trabalho híbridas	Gráficos de conhecimento conectados	Texto completo com pesquisa vetorial	Vetores de acesso pouco frequente e de longo prazo	Pesquisa vetorial em tempo real em grande escala	Necessidades de latência ultrabaixa
Padrão de consulta ideal	Consultas SQL e vetoriais mistas	Travessias gráficas com vetores	Pesquisas frequentes com análises	Recuperação pouco frequente (minutos a horas)	Consultas em tempo real de alta frequência	Milhões de solicitações por segundo
Configuração com o Amazon Bedrock	Simple (gerenciado pela Amazon Bedrock)	Simple (gerenciado pela Amazon Bedrock)	Simple (gerenciado pela Amazon Bedrock)	Simple (gerenciado pela Amazon Bedrock)	Simple (gerenciado pela Amazon Bedrock)	Simple (gerenciado pela Amazon Bedrock)
Residência de dados	Regiões da AWS	Regiões da AWS	Regiões da AWS	Regiões da AWS	Multinuvem (AWS e outras)	Multinuvem (AWS e outras)

Modelo de definição de preços	Horas de instância e armazenamento	Unidades de capacidade e armazenamento	Computação e armazenamento (sem servidor)	Armazenamento, consultas e transferência	Horário de funcionamento e armazenamento da cápsula	Horas e armazenamento dos nós

Escolha entre opções individuais e gerenciadas

Consideração	Escolha um banco de dados vetorial individual	Escolha as bases de conhecimento Amazon Bedrock (gerenciadas)
Implementação do RAG	Você quer controle total sobre o pipeline RAG	Você quer um RAG totalmente gerenciado com configuração mínima
Personalização	Você precisa de lógica de recuperação e pré-processamento personalizados	Os padrões RAG padrão atendem às suas necessidades
Infraestrutura existente	Você já tem o banco de dados implantado	Você está começando do zero ou quer um gerenciamento simplificado
Experiência da equipe	Sua equipe tem experiência em administração de banco de dados	Você prefere se concentrar na lógica do aplicativo, não na infraestrutura
Complexidade de integração	Você precisa de uma integração profunda com os sistemas existentes	Você quer uma integração rápida com os modelos Amazon Bedrock
Sobrecarga operacional	Você pode gerenciar as operações do banco de dados	Você quer AWS lidar com as operações

Estrutura de custos	Você prefere preços diretos do banco de dados	Você prefere preços unificados do Amazon Bedrock
Hora de comercializar	Você tem tempo para uma implementação personalizada	Você precisa de uma implantação rápida

Comparações e considerações de custos

Compreender a estrutura de custos das diferentes opções de banco de dados vetoriais é essencial para tomar decisões informadas de implementação. A tabela a seguir descreve algumas das principais considerações de custo de várias soluções de banco de dados vetoriais, incluindo bancos de dados individuais e serviços gerenciados. Cada opção tem fatores de preço distintos que podem afetar o custo total de propriedade (TCO), desde pay-as-you-go modelos até custos operacionais e de infraestrutura.

Banco de dados vetoriais	Modelo de custo	Considerações sobre custos
Amazon Kendra	Pague conforme o uso, com base em consultas	Os custos podem variar com base no número de consultas e na quantidade de dados indexados. Taxas adicionais podem ser aplicadas para armazenamento e transferência de dados. Para obter mais informações, consulte os preços do Amazon Kendra.
OpenSearch Serviço Amazon	Pague conforme o uso, com base nas horas da instância e no armazenamento	Os custos incluem horas de instância, armazenamento (volumes do Amazon EBS), transferência de dados e UltraWarm armazenamento opcional. As instâncias reservadas podem oferecer até 30% de economia. As instâncias aceleradas por GPU oferecem melhor relação preço-desempenho para cargas de trabalho vetoriais. Para obter mais informações, consulte os preços OpenSearch do Amazon Service.

Código aberto OpenSearch	Código aberto, sem custo direto (você não precisa pagar para baixar ou usar o software e não há custos de licença)	Os custos incluem infraestrutura (como servidores e armazenamento) e custos operacionais (como manutenção e monitoramento). As organizações precisam fazer um orçamento para que o pessoal gerencie e mantenha a infraestrutura.
Amazon RDS para PostgreSQL com pgvector	Pague conforme o uso, com base no uso	Os custos incluem tipos de instância de banco de dados, armazenamento, transferência de dados e backups. Cobranças adicionais podem ser aplicadas para transferência de dados, tipos de instância e armazenamento além do nível AWS gratuito. Para obter mais informações, consulte Definição de preço do Amazon RDS.
Amazon DocumentDB	Pague conforme o uso, com base nas horas da instância e no armazenamento	Os custos incluem horas de instância, armazenamento (GB por mês), I/O solicitações, armazenamento de backup e transferência de dados. Clusters elásticos permitem escalabilidade dinâmica. Instâncias reservadas disponíveis para otimização de custos. Para obter mais informações, consulte Preços do Amazon DocumentDB.

Amazon MemoryDB	Pague conforme o uso, com base nas horas dos nós e no armazenamento de dados	Os custos incluem horas de nós (por tipo de nó), armazenamento de dados (GB-hora), armazenamento de instantâneos e transferência de dados. Os nós reservados podem oferecer até 55% de economia. Otimizado para cargas de trabalho de alto rendimento e baixa latência. Para obter mais informações, consulte os preços do Amazon MemoryDB.
Amazon Neptune Analytics	Pague conforme o uso, com base nas unidades de capacidade	Os custos incluem Neptune Capacity Units NCUs (), armazenamento (GB/mês) e transferência de dados. Dimensionamento automático com base na carga de trabalho sem compromissos iniciais. É NCUs necessário um mínimo de 128. Para obter mais informações, consulte os preços do Amazon Neptune.

Amazon S3 Vectors	Pague conforme o uso, com base no armazenamento e nas solicitações	Os custos incluem armazenamento (GB/mês), solicitações PUT e GET, gerenciamento de índices vetoriais e transferência de dados. Oferece até 90% de economia de custos em comparação com bancos de dados vetoriais especializados. Políticas de hierarquização inteligente e ciclo de vida do Amazon S3 disponíveis para otimização adicional. Para obter mais informações, consulte Preço do Amazon S3.
Bases de conhecimento do Amazon Bedrock	Pague conforme o uso, com base no uso	Os custos podem variar com base no uso da base de conhecimento e de serviços adicionais, como o Amazon OpenSearch Serverless. Cobranças adicionais podem ser aplicadas para armazenamento de dados, transferência de dados e recursos adicionais. Para obter mais informações sobre preços, consulte Preços do Amazon OpenSearch Service.

Casos de uso de banco de dados vet

Os exemplos a seguir destacam como diferentes opções de banco de dados vetoriais podem ser usadas de forma eficaz para aprimorar o gerenciamento do conhecimento, melhorar a eficiência operacional e oferecer melhores resultados comerciais. Esses casos de uso ilustram as aplicações práticas das soluções de banco de dados vetoriais discutidas anteriormente neste guia e fornecem informações sobre seu desempenho e benefícios no mundo real.

Gestão do conhecimento com a Amazon Kendra

Problema do cliente — Uma das maiores empreiteiras gerais do Japão estava enfrentando um declínio no número de funcionários experientes. A empresa precisava de uma maneira eficiente de transferir o conhecimento e as habilidades do pessoal experiente para a geração mais jovem. Eles precisavam de uma solução para capturar e disseminar conhecimentos complexos de engenharia de construção e experiências passadas.

AWS solução — Para resolver esse problema, o cliente recorreu ao Amazon Kendra, uma solução de IA que poderia lidar com sua base de conhecimento interna de forma rápida e precisa e permitir consultas em linguagem natural. Com o Amazon Kendra, os funcionários agora podem encontrar as informações de que precisam com muito mais rapidez, melhorando a produtividade e facilitando a transferência de conhecimento de funcionários experientes para funcionários mais jovens.

Impacto — Ao implementar um chatbot generativo de IA desenvolvido pela Amazon Kendra, a empresa criou uma plataforma de conhecimento unificada. O chatbot permite que os funcionários acessem rapidamente o conhecimento técnico e as experiências anteriores em engenharia de construção. Essa solução melhorou significativamente a eficiência dos processos de transferência de conhecimento e tomada de decisão dentro da organização, ajudando a garantir que conhecimentos valiosos sejam preservados e facilmente acessíveis a todos os funcionários. O custo dessa solução pode variar dependendo do uso e da configuração. Para obter uma estimativa de custo detalhada, consulte AWS Calculadora de Preços. Para uma estimativa de custo do banco de dados vetoriais, consulte a seção Comparação de custos e considerações deste guia ou consulte a definição de preço do Amazon Kendra.

Para obter informações sobre outros casos de uso de clientes, consulte Clientes da Amazon Kendra.

Análise em tempo real com OpenSearch Serverless

Problema do cliente — Um provedor líder de serviços financeiros enfrentou o desafio de gerenciar um enorme ecossistema de dados. Ela processou 300 milhões de autorizações e 90 bilhões de transações anualmente, acumulando aproximadamente 1,1 petabytes (PB) de dados. O sistema existente, que atende a 300.000 usuários que precisavam de acesso a mais de 6.000 relatórios, precisava de modernização para fornecer consistência global e permitir a tomada de decisões em tempo real.

AWS solução — A arquitetura da solução usou modelos básicos disponíveis no Amazon Bedrock (incluindo Anthropic, Sonnet 3, Sonnet 3.5 e Haiku) para processamento de linguagem natural. O cliente escolheu o OpenSearch Serverless como banco de dados vetorial por sua escalabilidade superior e capacidade de lidar com o enorme volume de dados com eficiência. Essa arquitetura permitiu o processamento contínuo de consultas complexas e a geração dinâmica de relatórios.

Impacto — A implementação alcançou um aumento de 50% na produtividade ao eliminar a necessidade de geração manual de mais de 100 painéis de business intelligence. Agora, os usuários podem gerar relatórios por meio de consultas em linguagem natural com tempos de resposta entre 20 e 40 segundos. O custo dessa solução pode variar dependendo do uso e da configuração. Para obter uma estimativa de custo detalhada, consulte AWS Calculadora de Preços. Para uma estimativa de custo do banco de dados vetoriais, consulte a seção Comparação de custos e considerações deste guia ou consulte os preços do Amazon OpenSearch Service. Para obter informações sobre outros casos de uso de clientes, consulte Amazon OpenSearch Serverless.

Próximas etapas e recursos

Depois de analisar este guia, considere as seguintes ações para passar da compreensão à implementação:

1. Avalie suas necessidades atuais:
   - Avalie sua infraestrutura de banco de dados e sua experiência existentes.
   - Documente seus requisitos específicos de pesquisa vetorial.
   - Defina suas metas de desempenho, escalabilidade e custo.
2. Escolha uma das opções a seguir para testar as opções do banco de dados vetoriais:
   - Opção 1: configure uma prova de conceito usando sua solução de banco de dados vetorial preferida.
   - Opção 2: Experimente com conjuntos de dados de amostra nas bases de conhecimento Amazon Bedrock. Experimente a experiência de criação rápida de uma base de conhecimento Amazon Bedrock. Para ver um exemplo, consulte Criação rápida de uma base de conhecimento do Aurora PostgreSQL para o Amazon Bedrock na documentação do Aurora.
3. Analise os recursos adicionais.
4. Obtenha ajuda especializada:
   - Entre em contato com sua Conta da AWS equipe ou com os arquitetos de AWS soluções para obter orientação de implementação.
   - Interaja com AWS parceiros especializados em bancos de dados vetoriais.
5. Planeje sua implantação de produção:
   - Crie uma estratégia de migração se estiver migrando dos bancos de dados existentes.
   - Desenvolva um plano de escalabilidade para a solução escolhida.
   - Projete seus procedimentos de monitoramento e manutenção.

Recursos

Os recursos a seguir podem ajudá-lo a escolher um banco de dados vetoriais.

AWS postagens no blog

- Acelere o desenvolvimento de aplicativos de IA generativa com o Amazon Bedrock Knowledge Bases Quick Create e o Amazon Aurora Serverless
- Explicação dos recursos de banco de dados vetoriais do Amazon OpenSearch Service
- Mergulhe profundamente nos armazenamentos de dados vetoriais usando as bases de conhecimento Amazon Bedrock
- Utilize o pgvector e o Amazon Aurora PostgreSQL para processamento de linguagem natural, chatbots e análise de sentimentos

AWS documentação de serviço

- Escolhendo um serviço AWS de banco de dados
- Como funcionam as bases de conhecimento do Amazon Bedrock
- Documentação do Neptune Analytics
- Visão geral da Amazon Web Services: bancos de dados
- Usando o Aurora PostgreSQL como base de conhecimento para o Amazon Bedrock
- Trabalho com Amazon Aurora PostgreSQL
- Amazon DocumentDB
- Amazon MemoryDB
- Amazon S3 Vectors

Outros AWS recursos

- Bases de conhecimento do Amazon Bedrock
- Bancos de dados vetoriais e incorporações
- Bancos de dados vetoriais para aplicativos generativos de IA
- O que são incorporações no Machine Learning?

Outros recursos da

- Sobre o PostgreSQL

- Documentação do pgvector
- Pinecone como base de conhecimento para o Amazon Bedrock
- Redis Enterprise Cloud ativado AWS

Histórico do documento

A tabela a seguir descreve mudanças significativas neste guia, Escolhendo um banco de dados AWS vetoriais para casos de uso do RAG. Se desejar receber notificações sobre futuras atualizações, inscreva-se em um feed RSS.

Alteração	Descrição	Data
Adicionado Serviços da AWS	Adicionamos informações sobre o uso do Amazon DocumentDB, Amazon MemoryDB, Amazon S3 Vectors e Amazon Neptune Analytics.	30 de março de 2026
Publicação inicial	—	6 de março de 2025

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.