RAG ユースケース用の AWS ベクトルデータベースの選択

AWS 規範ガイダンス

AWS 規範ガイダンス: RAG ユースケース用の AWS ベクトルデータベースの選択

Table of Contents

RAG ユースケース用の AWS ベクトルデータベースの選択

Mayuri Shinde、Ivan Cui、Anand Bukkapatnam Tirumala、Amazon Web Services

2026 年 3 月 (ドキュメント履歴)

ベクトルデータベースは、生成 AI アプリケーションを実装する組織にとってますます重要になっています。これらのデータベースはベクトルを保存および管理します。ベクトルは、意味や関係をキャプチャする方法でテキスト、画像、その他のコンテンツを処理できるデータの数値表現です。

組織がベクトルデータベースのオプションを検討する際には AWS、さまざまなソリューションの機能、トレードオフ、ベストプラクティスを理解する必要があります。このガイドは、 で一般的に使用されるベクトルストアを比較 AWS し、特定のニーズやユースケースに最適なオプションについて情報に基づいた意思決定を行うのに役立ちます。このガイドでは、検索拡張生成 (RAG) の実装、レコメンデーションシステムの構築、その他の AI アプリケーションの開発など、ベクトルデータベースソリューションの評価と選択に役立つフレームワークを提供します。

対象者

このガイドは、以下の役割を担うユーザーを対象としています。

- ベクトルデータベースを使用して ML モデルの高次元データを保存および取得するデータサイエンティストと機械学習 (ML) エンジニア。
- 高次元データを保存および処理するためのベクトルデータベースを含むデータパイプラインを設計および実装するデータエンジニア。
- ML パイプラインの一部としてベクトルデータベースを使用してモデル出力または中間表現を保存および提供する MLOps エンジニア。
- ベクトルデータベースを類似度検索またはレコメンデーションシステムを必要とするアプリケーションに統合するソフトウェアエンジニア。
- ベクトルデータベースを本番環境にデプロイして維持し、スケーラビリティと信頼性を確保する DevOps エンジニア。
- ベクトルデータベースを使用して埋め込みまたは特徴ベクトルの大規模なデータセットを保存および分析する AI 研究者。
- 製品の特徴とアーキテクチャについて情報に基づいた意思決定を行うために、ベクトルデータベースの機能と制限を理解する必要がある AI 製品マネージャー。

ベクトルの概要

ベクトルは、マシンがデータを理解して処理するのに役立つ数値表現です。生成 AI では、2 つの主要な目的があります。

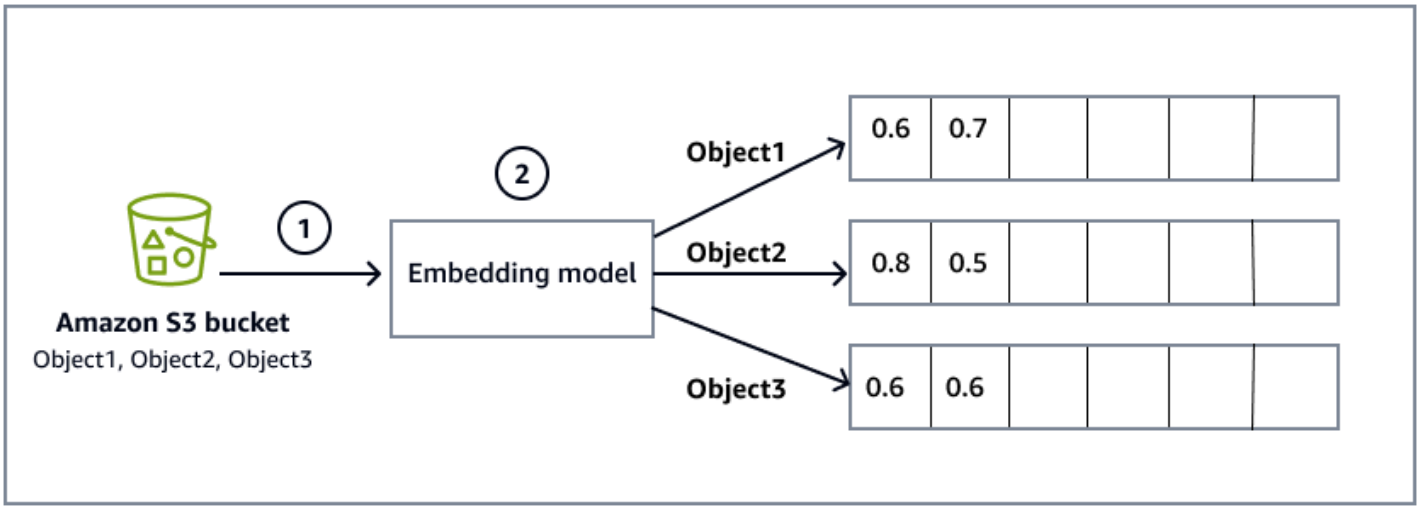
- データ構造を圧縮形式でキャプチャする潜在スペースを表す
- 単語、文、画像などのデータの埋め込みの作成

Word2Vec、GloVe、Amazon Titan Text Embeddings などの埋め込みモデルは、埋め込みと呼ばれるプロセスを通じてデータをベクトルに変換します。これらの埋め込みモデルでは、以下を実行できます。

- コンテキストから学び、単語をベクトルとして表現する
- ベクトルスペースに同様の単語を近づける
- マシンが連続スペース内のデータを処理できるようにする

次の図は、埋め込みプロセスの概要を示しています。

1. Amazon Simple Storage Service (Amazon S3) バケットには、システムが情報を読み取って処理するデータソースであるファイルが含まれています。Amazon S3 バケットは、Amazon Bedrock ナレッジベース設定中に指定されます。これには、ナレッジベースとのデータの同期も含まれます。
2. 埋め込みモデルは、Amazon S3 バケット内のオブジェクトファイルからの raw データをベクトル埋め込みに変換します。たとえば、 `Object1`は多次元空間内のコンテンツ`[0.6, 0.7, ...]`を表すベクトルに変換されます。



自然言語処理 (NLP) では、以下を行うため、単語埋め込みが不可欠です。

- 単語間のセマンティック関係をキャプチャする
- コンテキストに関連するテキストの生成を有効にする
- 大規模言語モデル (LLMs) を強化して人間のようなレスポンスを生成する

ベクトルデータベースの概要

ベクトルデータベースは、高次元ベクトルを効率的に保存およびクエリする特殊なシステムです。これらのデータベースは、取得拡張生成 (RAG) アプリケーションの基本です。

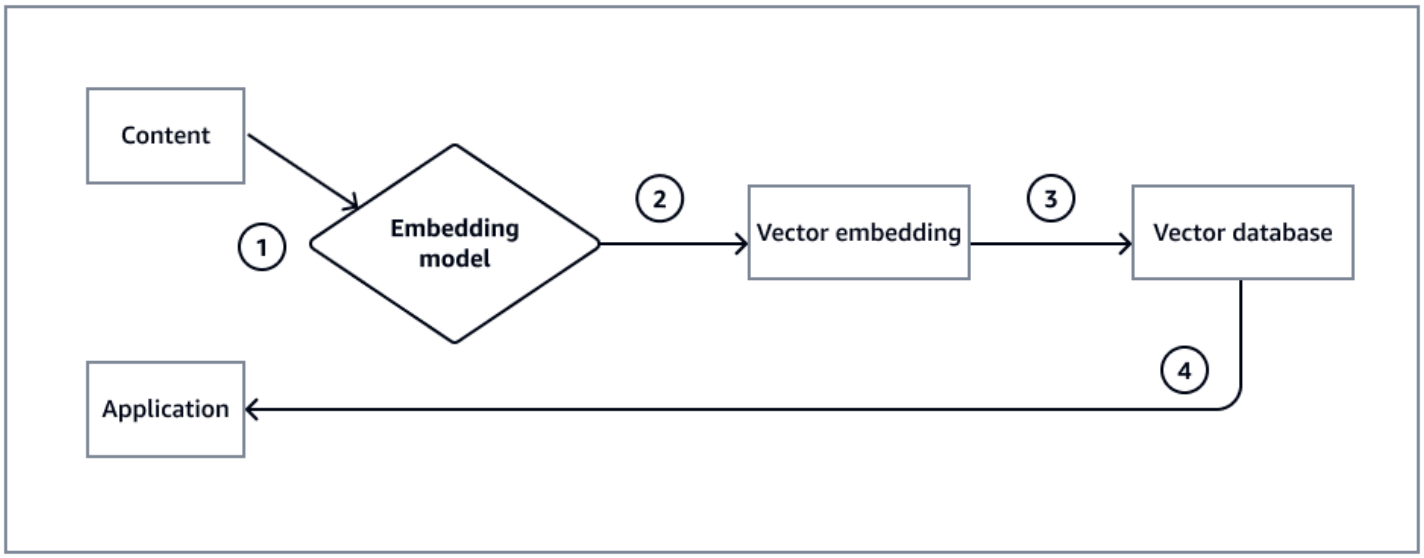
ベクトルデータベースは、次の方法でデータ変換とストレージを処理します。

- オブジェクト (オーディオ、イメージ、テキストファイルなど) は、埋め込みモデルを使用してベクトルに変換されます。
- ベクトルは特殊なデータ形式で保存されます。
- ベクトルデータベースは、迅速な類似度検索を可能にします。

ベクトルデータベースには、従来のデータベースよりもいくつかの重要な利点があるため、最新のデータ課題に特に適しています。ベクトルオペレーションに特化して最適化されており、高次元データを効率的に処理します。また、従来のデータベースが苦労する類似度検索にも特化しています。これらのコア機能を超えて、ベクトルデータベースは、ML および生成 AI アプリケーションの進化する需要を満たすように構築されています。大規模なベクトルストレージに優れ、分散コンピューティングを使用して複数のノード間でワークロードのバランスを取ります。これにより、データボリュームの増加に応じてスケーラビリティとパフォーマンスが提供されます。

次の図は、RAG の実装を示しています。

1. ドキュメント、PDFs、テキストファイルなどのコンテンツは、処理用の raw データとして埋め込みモデルにフィードされます。
2. 埋め込みモデルは未加工データを数値ベクトルに変換し、コンテンツの意味的意味を表します。
3. 生成されたベクトル埋め込みは、高次元ベクトルの保存と取得に最適化されたベクトルデータベースに保存されます。
4. アプリケーションは、セマンティック検索やコンテンツのレコメンデーションなどのユースケースに応じてベクトルデータベースをクエリできるようになりました。



RAG ソリューションに不適切なベクトルデータベースを選択すると、次のような大きな問題や制限が発生する可能性があります。

- クエリパフォーマンスの低下
- スケーラビリティのボトルネック
- データインジェストの課題
- フィルタリングやランキングなどの高度な機能がない
- 他のシステムとの統合の問題
- 永続性と耐久性に関する懸念
- 複数のユーザーを持つ環境での同時実行と整合性の問題
- ライセンスコストの増加またはベンダーのロックイン
- コミュニティのサポートとリソースの制限
- セキュリティとコンプライアンスに関する潜在的なリスク

ベクトルデータベースオプション

AWS は、生成 AI アプリケーションのさまざまなユースケースと要件をサポートするさまざまなベクトルデータベースソリューションを提供します。これらのオプションは、個々のデータベースサービスとマネージドサービスに広く分類でき、それぞれに異なる特性と利点があります。これらのオプションを理解することは、最適なパフォーマンス、スケーラビリティ、コスト効率を維持しながら、ベクトル検索機能を効果的に実装することを検討している組織にとって不可欠です。

ベクトルデータベースソリューションの詳細については、以下のセクションを参照してください。

- 個々のベクトルデータベースオプション
- マネージドサービスオプション
- 適切なベクトルデータベースの選択

個々のベクトルデータベースオプション

の個々のベクトルデータベースオプション AWS には、Amazon Kendra、Amazon OpenSearch Service、pgvector を使用した Amazon RDS for PostgreSQL、Amazon MemoryDB、Amazon DocumentDB、Amazon Neptune Analytics、Amazon S3 Vector などがあります。(オープンソースの拡張機能である pgvector は、ML が生成したベクトル埋め込みを保存および検索する機能を追加します)。これらのソリューションはベクトル検索にさまざまなアプローチを提供し、組織は既存のインフラストラクチャ、技術要件、特定のユースケースに基づいて選択できます。

Amazon Kendra

Amazon Kendra は、自然言語処理と高度な機械学習アルゴリズムを使用して、データからの検索質問に対する特定の回答を返すエンタープライズグレードのインテリジェント検索サービスです。Amazon Kendra は検索機能の実装を簡素化し、生成 AI アプリケーションのための効果的なバックエンドソリューションです。

Amazon Kendra のその他の主要な機能は次のとおりです。

- 40 を超えるデータソースへのネイティブ接続
- 組み込みのデータ準備機能
- 技術的な深い専門知識を必要としない高速セットアップ

Amazon Kendra の利点は次のとおりです。

- 自動データ処理 (チャンキング、取り込み、取り出し)
- 強力なカスタマイズオプション:
  - ファセット検索
  - 検索分析
  - 検索の関連性の調整
- を介したシンプルなプログラムによるアクセス AWS SDK for Python (Boto3)

詳細については、Amazon Kendra ドキュメントの「Amazon Kendra の利点」を参照してください。

Amazon OpenSearch Service

Amazon OpenSearch Service は、 で OpenSearch Service クラスターをデプロイ、運用、スケーリングするのに役立つマネージドサービスです AWS クラウド。

OpenSearch Service の主な機能は次のとおりです。

- オープンソースの検索および分析エンジン
- 分散アーキテクチャ
- リアルタイムデータ処理

OpenSearch Service を使用する利点には、次のようなものがあります。

- 水平スケーラビリティ
- RESTful API のサポート
- 構造化データと非構造化データを処理します
- リアルタイムデータ分析
- さまざまなデプロイサイズに適しています

詳細については、OpenSearch Service ドキュメントの「Amazon OpenSearch Service の機能」を参照してください。

pgvector を使用した Amazon RDS for PostgreSQL

Amazon RDS for PostgreSQL with pgvector は、 AWS マネージドリレーショナルデータベースサービスと PostgreSQL のベクトル処理拡張機能を組み合わせます。この組み合わせにより、組織は Amazon RDS を維持しながら高次元ベクトルを保存およびクエリできます。このソリューション

は、データベースインフラストラクチャを管理するオーバーヘッドなしでリアルタイムのベクトル操作を必要とする生成 AI アプリケーションに特に適しています。

pgvector を使用した Amazon RDS for PostgreSQL の主な利点は次のとおりです。

- 高可用性
- 自動フェイルオーバー
- コスト効率 (pay-per-use制)
- 組み込みモニタリング
- リアルタイムベクトルデータ統合

詳細については、Amazon RDS ドキュメントの「Amazon RDS の利点」を参照してください。

Amazon DocumentDB

Amazon DocumentDB (MongoDB 互換) は、バージョン 5.0 以降のネイティブベクトル検索機能を提供するドキュメントデータベースです。JSON ベースのドキュメントストレージの柔軟性とベクトル検索を組み合わせることで、階層ナビゲーション可能なスモールワールド (HNSW) と Inverted File Flat (IVFFlat) の両方のインデックス作成方法をサポートしています。

Amazon DocumentDB の主な機能は次のとおりです。

- 最大 2,000 個のディメンション (インデックス作成なしで最大 16,000 個のディメンション) のベクトルを保存およびインデックス作成
- ベクトル類似度検索のミリ秒応答時間
- euclidean、cosine、dot の製品距離メトリクスのサポート
- 既存の MongoDB 互換アプリケーションとのシームレスな統合

以下の状況で Amazon DocumentDB を使用します。

- MongoDB API APIsしていて、ベクトル検索機能が必要なアプリケーションの場合
- セマンティック検索と組み合わせた柔軟なドキュメントデータ構造を必要とするユースケースの場合
- 従来のドキュメントクエリとベクトル類似度検索の両方を必要とするシナリオの場合
- 製品のレコメンデーション、パーソナライゼーション、チャットアシスタント、不正検出を提供するアプリケーション向け

詳細については、Amazon DocumentDB ドキュメントの「Amazon DocumentDB のベクトル検索」を参照してください。 Amazon DocumentDB

Amazon MemoryDB

Amazon MemoryDB は Redis 互換のインメモリデータベースで、一般的なベクトルデータベースの中で最速のベクトル検索パフォーマンスを提供します AWS。マルチアベイラビリティーゾーンの耐久性を備えたミリ秒未満のクエリレイテンシーを提供します。

MemoryDB の主な機能は次のとおりです。

- アプリケーションデータと数百万のベクトルを 1 つのデータベースに保存
- 1 桁ミリ秒のクエリと更新の応答時間
- で最高の再現率と最速のパフォーマンス AWS
- ベクトルあたり最大 32,768 ディメンションのサポート
- リアルタイムのセマンティック検索およびキャッシュ機能

次の状況で MemoryDB を使用します。

- 超低レイテンシー (10 ミリ秒未満) を必要とするリアルタイムアプリケーションの場合
- 1 日あたり数百万リクエストの高スループットワークロードの場合
- リアルタイムレコメンデーションエンジン、セマンティックキャッシュ、異常検出などのユースケース向け
- インメモリデータストアとベクトル検索機能の両方を必要とするアプリケーションの場合

詳細については、MemoryDB ドキュメントの「ベクトル検索」を参照してください。 MemoryDB

Amazon Neptune Analytics

Amazon Neptune Analytics は、ネイティブベクトル検索機能を提供するグラフ分析エンジンであり、Graph Retrieval Augmented Generation (GraphRAG) のユースケースに最適です。ベクトル類似度検索とグラフトラバーサルおよびアルゴリズムを組み合わせます。

Neptune Analytics の主な機能は次のとおりです。

- 数秒で何十億もの関係を分析する
- ベクトル検索とグラフアルゴリズムを組み合わせる (パス検出、コミュニティ検出、中心性)

- トポロジに関する知識を持つ GraphRAG アプリケーションのサポート
- 既存のグラフ分析ソリューションよりも最大 80 倍高速
- フルマネージド GraphRAG 用の Amazon Bedrock との統合

Neptune Analytics は、以下の状況で使用します。

- ベクトル埋め込みを含むナレッジグラフを必要とする GraphRAG アプリケーションの場合
- ベクトルの類似性とともに複雑な関係をトラバースする必要があるユースケースの場合
- 関係コンテキストで説明可能な AI レスポンスを必要とするアプリケーションの場合
- Customer 360 ビュー、不正検出ネットワーク、ナレッジ検出などのシナリオの場合

詳細については、Amazon Neptune Analytics のドキュメントを参照してください。

Amazon S3 Vectors

Amazon S3 Vectors は、 AWS ネイティブベクトルストレージとクエリ機能を備えた の最初のクラウドオブジェクトストアです。大規模な規模を必要とする AI アプリケーション向けに、専用のコスト最適化ベクトルストレージを提供します。

Amazon S3 Vectors の主な機能は次のとおりです。

- ベクトルバケットあたり最大 10,000 個のインデックスをサポートする、インデックスあたり最大 20 億個のベクトルのストレージ
- 長期ストレージと低頻度アクセスパターン用に最適化された Sub-100 ミリ秒未満のクエリレイテンシー
- 特殊なベクトルデータベースと比較して、ベクトルオペレーションのコストを最大 90% 削減
- 自動スケーリングと 99.999999999% (11 9s) の耐久性を備えたサーバーレスアーキテクチャ

以下の状況では Amazon S3 Vectors を使用します。

- 最小限のコストで数十億のベクトルのストレージを必要とするアプリケーション向け
- 10 ミリ秒未満ではなく 1 秒未満のクエリレイテンシー (100 ミリ秒以上) を許容するワークロードの場合
- 長期ベクトル保持とアーカイブのユースケースの場合
- 取得パターンの頻度が低い RAG アプリケーションの場合

- 超低レイテンシーよりもストレージの経済性を優先する組織向け

Amazon S3 Vectors は Amazon Bedrock ナレッジベースとネイティブに統合され、Amazon OpenSearch Service の階層型アーキテクチャでうまく機能します。コールドストレージには Amazon S3 Vectors を使用し、ホットクエリには OpenSearch Service を使用できます。

詳細については、Amazon S3 ドキュメントの「S3 ベクトルとベクトルバケットの使用」を参照してください。 Amazon S3

マネージドサービスオプション

Amazon Bedrock ナレッジベースは、ベクトルデータベースの実装に対する AWS フルマネージド型のアプローチを表します。このサービスのストレージオプションの柔軟性と自動管理機能は、複雑なインフラストラクチャを管理せずに RAG を実装しようとしている組織にとって特に役立ちます。

Amazon Bedrock ナレッジベースを使用すると、RAG を使用して基盤モデルを強化するナレッジベースを作成、保守、クエリできます。このサービスは、データの取り込み、ベクトル化、取得パイプライン全体を管理することで、RAG を実装する複雑なプロセスを簡素化します。

Amazon Bedrock ナレッジベースの主な利点は次のとおりです。

- データ処理の簡素化
  - データの自動取り込みとチャンキング
  - 複数のファイル形式からの組み込みテキスト抽出
  - マネージドベクトル埋め込みの生成
  - メタデータの自動抽出とインデックス作成
- 効率的な RAG 実装
  - 事前設定された取得戦略
  - 自動コンテキストウィンドウの最適化
  - 組み込みの関連性チューニング
  - すぐに使えるセマンティック検索機能
- セキュリティとガバナンス
  - 統合 AWS Identity and Access Management (IAM) コントロール
  - 保管中および転送中のデータ暗号化
  - VPC サポート

- を使用した監査ログ記録 AWS CloudTrail

Amazon Bedrock ナレッジベースは、以下を含む複数のベクトルストアオプションをサポートしています。

- pgvector を使用した Amazon Aurora PostgreSQL
- Amazon Neptune Analytics
- Amazon EMR Serverless
- Amazon S3 Vectors
- Pinecone
- Redis エンタープライズクラウド

このマネージドサービスは、自動データ取り込み、ベクトル化、取得を処理します。これにより、RAG の実装が簡素化されます。

サポートされている各ベクトルストアの詳細については、Amazon Bedrock ナレッジベースのドキュメントを参照してください。

適切なベクトルデータベースの選択

これらの主要な決定要因に基づいてベクトルデータベースを選択します。

- ベクトル検索で MongoDB 互換ドキュメントデータベースが必要な場合は、Amazon DocumentDB を選択します。これは、アプリケーションが MongoDB APIs を使用していて、個別のベクトルインフラストラクチャを管理せずにセマンティック検索機能を追加する場合に最適です。
- リアルタイムアプリケーションに超低レイテンシーが必要な場合は、Amazon MemoryDB を選択します。これにより、ミリ秒未満の応答時間 AWS で で最速のベクトル検索パフォーマンスが得られます。リアルタイムのレコメンデーションエンジンや高スループットアプリケーションに最適です。
- ベクトル検索でグラフベースのナレッジ表現が必要な場合は、Amazon Neptune Analytics を選択します。これは、複雑な関係を横断し、ベクトル検索とともにグラフベースのクエリを実行し、説明可能な AI レスポンスを提供する必要がある GraphRAG アプリケーションに最適です。
- リレーショナルクエリとベクトル検索を組み合わせる必要がある場合 – Amazon Aurora PostgreSQL と pgvector を選択します。このオプションは、アプリケーションが同じデータベース内で従来の SQL オペレーションとベクトル類似度検索の両方を必要とする場合に最適です。

- レイテンシーが 10 ミリ秒未満の高スループットクエリが必要な場合は、Amazon OpenSearch Service を選択します。高頻度クエリとリアルタイムアプリケーションの処理に優れており、最近の GPU アクセラレーションの改善が含まれています。
- 数十億のベクトルを費用対効果の高い方法で保存する必要がある場合 – Amazon S3 Vectors を選択します。このオプションは最大 90% のコスト削減を実現し、100 ミリ秒未満のレイテンシーを許容できる取得パターン (クエリ間の数分から数時間) が少ないアプリケーションに最適です。
- ベクトル検索とともに全文検索が必要な場合は、Amazon OpenSearch Service を選択します。このオプションは、強力な全文検索機能とベクトル検索を単一のプラットフォームで組み合わせます。

ベクトルデータベースの比較

AWS は、個々のベクトルデータベースからフルマネージドサービスである Amazon Bedrock ナレッジベースまで、ベクトル検索機能を実装するための複数のアプローチを提供します。これらのオプションを評価する際、組織はアーキテクチャ、スケーラビリティ、統合機能、パフォーマンス特性、セキュリティ機能など、さまざまな側面を考慮する必要があります。

個々のベクトルデータベース

次の表は、アーキテクチャ、スケーリング機能、データソース統合、パフォーマンス特性に焦点を当てた、個々の AWS ベクトルデータベースソリューションの主な機能の概要を示しています。

機能	Amazon Kendra	Amazon OpenSearch Service	pgvector を使用した Amazon RDS for PostgreSQL with	Amazon DocumentDB	Amazon MemoryDB	Amazon Neptune Analytics	Amazon S3 Vectors
主なユースケース	エンタープライズ検索と RAG	分散検索と分析	ベクトルをサポートするリレーショナル DB	ベクトル検索を使用したドキュメント DB	リアルタイムインメモリベクトル検索	ベクトル検索によるグラフ分析	コスト最適化ベクトルストレージ
アーキテクチャ	フルマネージド型	分散クラスター	リレーショナルデータベース	ドキュメント指向	インメモリデータベース	グラフ分析エンジン	サーバーレスオブジェクトストレージ
データモデル	ドキュメントベース	JSON ドキュメント	リレーショナルテーブル	JSON ドキュメント	JSON を使用したキーと値	プロパティグラフ	オブジェクトストレージ

ベクトルディメンション	自動管理	最大 16,000	設定可能	最大 2,000 (インデックス付き)、16,000 (インデックスなし)	最大 32,768	設定可能	最大 4,096
インデックス作成方法	自動	HNSW、IVF	HNSW、IVFlat	HNSW、IVFlat	HNSW	ネイティブグラフとベクトル	自動
距離メトリクス	自動	コサイン、ユークリッド、ドット製品	コサイン、ユークリッド、内部製品	コサイン、ユークリッド、ドット製品	コサイン、ユークリッド、内部製品	コサイン、ユークリッド語	コサイン、ユークリッド語
クエリレイテンシー	1 秒未満	Sub-10 (GPU アクセラレーション)	10～100 ミリ秒	ミリ秒	ミリ秒未満	1 秒未満	Sub-100
スケーリングモデル	自動	水平 (ノードの追加)	垂直レプリカとリードレプリカ	水平 (インスタンスの追加)	垂直およびレプリカ	自動	自動 (サーバーレス)

最大ベクトル	マネージド	数十億 (クラスター依存)	百万 (インスタンス依存)	コレクションあたりの数百万	データベースあたり 100 万	数十億	インデックスあたり 20 億、バケットあたり 10,000 インデックス
スループット	高	非常に高い (数千の QPS)	中	高	非常に高い (1 日あたり数百万件のリクエスト)	高	中 (低頻度クエリに最適化)
データ耐久性	99.999999999% (11 9s)	レプリカで設定可能	99.99% (マルチ AZ)	99.99% (マルチ AZ)	99.99% (マルチ AZ)	99.99%	99.999999999% (11 9s)
整合性モデル	結果	結果 (設定可能)	強力 (ACID)	結果	強力	強力	強力
その他の機能	40 個以上のデータコネクタ、NLP	全文検索、分析、ダッシュボード	SQL クエリ、ACID トランザクション	MongoDB API の互換性	Redis API の互換性、キャッシュ	グラフアルゴリズム、トラバーサル	Amazon S3 統合、ライフサイクルポリシー
料金モデル	クエリとストレージあたりの支払い	インスタンス時間とストレージ	インスタンス時間とストレージ	インスタンス時間とストレージ	インスタンス時間とストレージ	キャパシティーユニットとストレージ	ストレージ、クエリ、データ転送

コスト最適化	使用状況ベース	リザーブドインスタンス、自動スケーリング	リザーブドインスタンス、Aurora Serverless	リザーブドインスタンス	リザーブドインスタンス	Auto-scaling	特殊なDBs節約
次の用途に適しています	最小限のセットアップでエンタープライズ検索	高スループット、低レイテンシーのクエリ	ハイブリッド SQL およびベクトルワークロード	ベクトルを必要とする MongoDB 互換アプリケーション	超低レイテンシーのリアルタイムアプリケーション	GraphRAG グラフとナレッジグラフ	長期で費用対効果の高いストレージ
理想的なクエリパターン	頻繁なエンタープライズ検索	高頻度リアルタイムクエリ	SQL クエリとベクトルクエリの混在	セマンティック検索を使用したドキュメントクエリ	1 日あたり数百万のリクエスト	ベクトル検索を使用したグラフトラバーサル	低頻度のクエリ (分 ~ 時間)
セットアップの複雑さ	低 (フルマネージド)	中 (クラスター設定)	中 (拡張設定)	中 (クラスター設定)	中 (クラスター設定)	低 (フルマネージド)	低 (サーバーレス)
チームの専門知識が必要	Minimal	OpenSearch または Elasticsearch	PostgreSQL、SQL	MongoDB	Redis	グラフデータベース	Amazon S3、基本的なベクトルの概念

マネージドサービス – Amazon Bedrock ナレッジベース

Amazon Bedrock ナレッジベースは、複数のベクトルストレージオプションを備えたフルマネージドソリューションを提供します。次の表は、これらのストレージオプションを比較したものです。

機能	pgvector PostgreSQ Lwith	Neptune 分析	OpenSearch Service サーバーレス	Amazon S3 ベクトル	Pinecone	RedisEnterprise クラウド
主なユースケース	ベクトル RAG を使用したリレーショナル DB	GraphRAG のグラフベースのベクトル検索	ナレッジ管理 RAG	コスト最適化ベクトル RAG	高性能ベクトル検索	インメモリベクトル検索
アーキテクチャ	フルマネージドリレーショナル	完全マネージド型グラフ分析	フルマネージドサーバーレス	サーバーレスオブジェクトストレージ	フルマネージドハイブリッドクラウド	完全マネージド型インメモリ
データモデル	リレーショナルテーブル	プロパティグラフ	JSON ドキュメント	オブジェクトストレージ	専用ベクトル	ベクトルを含むキー値
ベクトルストレージ	pgvector 拡張機能経由	ネイティブグラフベクトル	OpenSearch エンジン経由	ネイティブ Amazon S3 ベクトルストレージ	ネイティブベクトルデータベース	インメモリベクトル
Amazon Bedrock の統合	ネイティブ	ネイティブ	ネイティブ	ネイティブ	ネイティブ	ネイティブ
自動取り込み	はい (Amazon Bedrock 経由)	はい (Amazon Bedrock 経由)	はい (Amazon Bedrock 経由)	はい (Amazon Bedrock 経由)	はい (Amazon Bedrock 経由)	はい (Amazon Bedrock 経由)
自動ベクトル化	はい (Amazon Bedrock 経由)	はい (Amazon Bedrock 経由)	はい (Amazon Bedrock 経由)	はい (Amazon Bedrock 経由)	はい (Amazon Bedrock 経由)	はい (Amazon Bedrock 経由)

スケーリング	Auto Scaling (Aurora サーバーレス)	自動グラフスケーリング	自動サーバーレス	自動 (数十億のベクトル)	Auto Scaling ポッド	Auto Scaling クラスター
クエリパフォーマンス	リレーショナルまたはベクトルの場合は高	グラフベクトルの高	高	中 (100 ミリ秒以上のレイテンシー)	非常に高い	非常に高い
最大ベクトル	百万 (インスタンス依存)	数十億	数十億	インデックスあたり 20 億	数十億	百万 (メモリ依存)
その他の機能	SQL クエリ、ACID トランザクション	グラフアルゴリズム、トラバーサル	全文検索、分析	Amazon S3 ライフサイクル、階層化	メタデータフィルタリング、名前空間	Redis データ構造、キャッシュ
コスト最適化	中程度 (Aurora サーバーレス)	中程度 (キャパシティーユニット)	高 (サーバーレス、pay-per-use制)	非常に高い (最大 90% の節約)	中程度 (ポッドベースの料金)	低 (インメモリプレミアム)
次の用途に適しています	ハイブリッド SQL/ベクトルワークロード	接続されたナレッジグラフ	ベクトル検索を使用した全文	長期、低頻度のアクセスベクトル	大規模なリアルタイムベクトル検索	超低レイテンシーのニーズ
理想的なクエリパターン	SQL クエリとベクトルクエリの混在	ベクトルを使用したグラフトラバーサル	分析による頻繁な検索	取得頻度が低い (分から時間)	高頻度リアルタイムクエリ	1 秒あたり数百万のリクエスト

Amazon Bedrock でのセットアップ	Simple (Amazon Bedrock によって管理)	Simple (Amazon Bedrock によって管理)	Simple (Amazon Bedrock によって管理)	Simple (Amazon Bedrock によって管理)	Simple (Amazon Bedrock によって管理)	Simple (Amazon Bedrock によって管理)
データレジデンシー	AWS リージョン	AWS リージョン	AWS リージョン	AWS リージョン	マルチクラウド (AWS その他)	マルチクラウド (AWS その他)
料金モデル	インスタンス時間とストレージ	キャパシティーユニットとストレージ	コンピューティングとストレージ (サーバーレス)	ストレージ、クエリ、転送	ポッド時間とストレージ	ノード時間とストレージ

個々のオプションとマネージドオプションの選択

考慮事項	個々のベクトル DB を選択する	Amazon Bedrock ナレッジベースを選択する (マネージド)
RAG の実装	RAG パイプラインを完全に制御したい	最小限のセットアップでフルマネージド RAG が必要
カスタマイズ	カスタム取得ロジックと前処理が必要です	ニーズを満たす標準 RAG パターン
既存のインフラストラクチャ	データベースが既にデプロイされている	新しく開始する場合、または管理を簡素化したい場合
チームに関する専門知識	チームにデータベース管理の専門知識がある	インフラストラクチャではなくアプリケーションロジックに集中したい
統合の複雑さ	既存のシステムとの深い統合が必要	Amazon Bedrock モデルとの迅速な統合が必要

運用オーバーヘッド	データベースオペレーションを管理できます	オペレーション AWS を処理する場合
コスト構造	データベースの直接料金が望ましい	Amazon Bedrock の統一された料金が望ましい
市場投入までの時間	カスタム実装の時間がある	迅速なデプロイが必要

コストの比較と考慮事項

さまざまなベクトルデータベースオプションのコスト構造を理解することは、情報に基づいた実装の決定を行う上で不可欠です。次の表は、個々のデータベースとマネージドサービスの両方を含む、さまざまなベクトルデータベースソリューションのコストに関する重要な考慮事項の概要を示しています。各オプションには、pay-as-you-goモデルからインフラストラクチャや運用コストまで、総所有コスト (TCO) に影響を与える可能性のある個別の料金要因があります。

ベクトルデータベース	コストモデル	コストに関する考慮事項
Amazon Kendra	クエリに基づいて、従量課金制で支払う	コストは、クエリの数とインデックス化されたデータの量によって異なります。データストレージとデータ転送には追加料金が適用される場合があります。詳細については、「Amazon Kendra の料金」を参照してください。
Amazon OpenSearch Service	インスタンスの時間とストレージに基づいて、従量制料金を支払う	コストには、インスタンス時間、ストレージ (Amazon EBS ボリューム) 、データ転送、オプションの UltraWarm ストレージが含まれます。リザーブドインスタンスでは、最大 30% の節約が可能です。GPU アクセラレーションインスタンスは、ベクトルワークロードの価格パフォーマンスを向上させます。詳細については、「Amazon OpenSearch Service の料金」を参照してください。
オープンソースの OpenSearch	オープンソース、直接コストなし (ソフトウェアをダウンロードまたは使用するために	コストには、インフラストラクチャ (サーバーやストレージなど) と運用コスト (メンテナ

	料金を支払う必要はなく、ライセンスコストもかかりません)	ンスやモニタリングなど) が含まれます。組織は、人員がインフラストラクチャを管理および維持するための予算を立てる必要があります。
pgvector を使用した Amazon RDS for PostgreSQL	使用量に基づいて従量制で支払う	コストには、データベースインスタンスタイプ、ストレージ、データ転送、バックアップが含まれます。 AWS 無料利用枠を超えるデータ転送、インスタンスタイプ、ストレージには追加料金が適用される場合があります。詳細については、「Amazon RDS の料金」を参照してください。
Amazon DocumentDB	インスタンスの時間とストレージに基づいて、従量制料金を支払う	コストには、インスタンス時間、ストレージ (GB/月) 、I/O リクエスト、バックアップストレージ、データ転送が含まれます。Elastic クラスターは動的スケーリングを有効にします。コスト最適化に使用できるリザーブドインスタンス。詳細については、Amazon DocumentDB の料金 を参照してください。

Amazon MemoryDB	ノード時間とデータストレージに基づいて、従量制料金を支払う	コストには、ノード時間 (ノードタイプごと) 、データストレージ (GB-時間) 、スナップショットストレージ、データ転送が含まれます。リザーブドノードでは、最大 55% の節約が可能です。高スループット、低レイテンシーのワークロード向けに最適化されています。詳細については、「Amazon MemoryDB の料金」を参照してください。
Amazon Neptune Analytics	キャパシティーユニットに基づく従量制料金	コストには、Neptune キャパシティーユニット (NCUsストレージ (GB/月) 、データ転送が含まれます。事前コミットメントのないワークロードに基づく自動スケーリング。最小 128 NCUsが必要です。詳細については、Amazon Neptune の料金」を参照してください。

Amazon S3 Vectors	ストレージとリクエストに基づいて、従量制料金を支払う	コストには、ストレージ (GB/月) 、PUT および GET リクエスト、ベクトルインデックス管理、データ転送が含まれます。特殊なベクトルデータベースと比較して、最大 90% のコスト削減を実現します。Amazon S3 Intelligent-Tiering およびライフサイクルポリシーは、追加の最適化に使用できます。詳細については、「Amazon S3 の料金」を参照してください。
Amazon Bedrock ナレッジベース	使用量に基づいて従量制で支払う	コストは、ナレッジベースの使用や Amazon OpenSearch Serverless などの追加サービスによって異なります。データストレージ、データ転送、および追加機能には追加料金が適用される場合があります。料金の詳細については、「Amazon OpenSearch Service の料金」を参照してください。

ベクトルデータベースのユースケース

次の例では、さまざまなベクトルデータベースオプションを効果的に使用して、ナレッジ管理の強化、運用効率の向上、ビジネス成果の向上を実現する方法について説明します。これらのユースケースは、このガイドの前半で説明したベクトルデータベースソリューションの実用的なアプリケーションを示し、実際のパフォーマンスと利点に関するインサイトを提供します。

Amazon Kendra によるナレッジ管理

顧客問題 — 日本最大の一般請負業者の 1 人が、経験豊富な人材の減少に直面していました。同社は、経験担当者の知識とスキルを若年世代に効率的に伝達する方法を必要としていました。複雑な建設エンジニアリングの知識と過去の経験を捉えて広めるソリューションが必要でした。

AWS ソリューション – この問題に対処するために、顧客は内部ナレッジベースを迅速かつ正確に処理し、自然言語クエリを許可する AI ソリューションである Amazon Kendra を利用しました。Amazon Kendra を使用すると、従業員は必要な情報をより迅速に見つけることができるようになり、生産性が向上し、経験豊富なスタッフから若年スタッフへの知識移転が容易になります。

影響 — Amazon Kendra を搭載した生成 AI チャットボットを実装することで、同社は統合されたナレッジプラットフォームを作成しました。チャットボットを使用すると、従業員は技術的な知識や建設エンジニアリングに関する過去の経験にすばやくアクセスできます。このソリューションにより、組織内の知識移転と意思決定プロセスの効率が大幅に向上し、貴重な専門知識が保持され、すべての従業員が簡単にアクセスできるようになります。このソリューションのコストは、使用状況と設定によって異なる場合があります。詳細なコスト見積もりについては、「」を参照してくださいAWS 料金見積りツール。ベクトルデータベースのコスト見積もりについては、このガイドの「コストの比較と考慮事項」セクションまたは「Amazon Kendra の料金」を参照してください。

その他のお客様のユースケースについては、「Amazon Kendra のお客様」を参照してください。

OpenSearch Serverless によるリアルタイム分析

顧客問題 — ある主要な金融サービスプロバイダーが、膨大なデータエコシステムを管理するという課題に直面しました。年間 3 億件の認可と 900 億件のトランザクションを処理し、約 1.1 ペタバイト (PB) のデータに蓄積されています。6,000 を超えるレポートへのアクセスを必要とする 300,000 人のユーザーに対応する既存のシステムでは、グローバルな一貫性を提供し、リアルタイムの意思決定を可能にするためにモダナイゼーションが必要でした。

AWS ソリューション – ソリューションアーキテクチャでは、自然言語処理に Amazon Bedrock (Anthropic、Sonnet 3、Sonnet 3.5、Haiku を含む) を通じて利用可能な基盤モデルを使用しました。お客様は、膨大なデータボリュームを効率的に処理するための優れたスケーラビリティと機能のために、ベクトルデータベースとして OpenSearch Serverless を選択しました。このアーキテクチャにより、複雑なクエリのシームレスな処理と動的なレポート生成が可能になりました。

影響 – この実装では、100 を超えるビジネスインテリジェンスダッシュボードを手動で生成する必要がなくなるため、生産性が 50% 向上しました。ユーザーは自然言語クエリを通じて 20～40 秒の応答時間でレポートを生成できるようになりました。このソリューションのコストは、使用状況と設定によって異なる場合があります。詳細なコスト見積もりについては、「」を参照してくださいAWS 料金見積りツール。ベクトルデータベースのコスト見積もりについては、このガイドの「コストの比較と考慮事項」セクションまたは「Amazon OpenSearch Service の料金」を参照してください。その他のお客様のユースケースについては、「Amazon OpenSearch Serverless」を参照してください。

次のステップとリソース

このガイドを確認したら、理解から実装に移行するために以下のアクションを検討してください。

1. 現在のニーズを評価します。
   - 既存のデータベースインフラストラクチャと専門知識を評価します。
   - 特定のベクトル検索要件を文書化します。
   - パフォーマンス、スケーリング、コスト目標を定義します。
2. ベクトルデータベースオプションをテストするには、次のいずれかのオプションを選択します。
   - オプション 1: 任意のベクトルデータベースソリューションを使用して概念実証を設定します。
   - オプション 2: Amazon Bedrock ナレッジベースでサンプルデータセットを試す。Amazon Bedrock ナレッジベースのクイック作成エクスペリエンスをお試しください。例については、Aurora ドキュメントの「Amazon Bedrock 用の Aurora PostgreSQL ナレッジベースのクイック作成」を参照してください。
3. 追加のリソースを確認します。
4. エキスパートのサポートを受ける:
   - 実装ガイダンスについては、 AWS アカウント チームまたは AWS ソリューションアーキテクトにお問い合わせください。
   - ベクトルデータベースを専門とする AWS パートナーと連携します。
5. 本番稼働用デプロイを計画します。
   - 既存のデータベースから移行する場合は、移行戦略を作成します。
   - 選択したソリューションのスケーリングプランを作成します。
   - モニタリングおよびメンテナンス手順を設計します。

リソース

以下のリソースは、ベクトルデータベースの選択に役立ちます。

AWS ブログ投稿

- Amazon Bedrock ナレッジベースのクイック作成と Amazon Aurora Serverless を使用して生成 AI アプリケーション開発を加速する
- Amazon OpenSearch Service のベクトルデータベース機能の説明

- Amazon Bedrock ナレッジベースを使用してベクトルデータストアを詳しく調べる
- pgvector と Amazon Aurora PostgreSQL を活用して自然言語処理、チャットボット、感情分析を行う

AWS サービスドキュメント

- AWS データベースサービスの選択
- Amazon Bedrock ナレッジベースの仕組み
- Neptune Analytics ドキュメント
- アマゾン ウェブ サービスの概要: データベース
- Amazon Bedrock のナレッジベースとしての Aurora PostgreSQL の使用
- Amazon Aurora PostgreSQL の操作
- Amazon DocumentDB
- Amazon MemoryDB
- Amazon S3 Vectors

その他の AWS リソース

- Amazon Bedrock ナレッジベース
- ベクトルデータベースと埋め込み
- 生成 AI アプリケーション用のベクトルデータベース
- Machine Learningの埋め込みとは

その他のリソース

- PostgreSQL について
- pgvector ドキュメント
- Amazon Bedrock のナレッジベースとしての Pinecone
- での Redis Enterprise Cloud AWS

ドキュメント履歴

次の表に、RAG ユースケースの AWS ベクトルデータベースを選択するという、このガイドの重要な変更点を示します。今後の更新に関する通知を受け取る場合は、RSS フィードをサブスクライブできます。

変更	説明	日付
追加済み AWS のサービス	Amazon DocumentDB、Amazon MemoryDB、Amazon S3 Vectors、Amazon Neptune Analytics の使用に関する情報を追加しました。	2026 年 3 月 30 日
初版発行	—	2025 年 3 月 6 日

翻訳は機械翻訳により提供されています。提供された翻訳内容と英語版の間で齟齬、不一致または矛盾がある場合、英語版が優先します。