



为 RAG 用例选择 AWS 矢量数据库

AWS 规范性指导



AWS 规范性指导: 为 RAG 用例选择 AWS 矢量数据库

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

Amazon 的商标和商业外观不得用于任何非 Amazon 的商品或服务，也不得以任何可能引起客户混淆、贬低或诋毁 Amazon 的方式使用。所有非 Amazon 拥有的其他商标均为各自所有者的财产，这些所有者可能附属于 Amazon、与 Amazon 有关联或由 Amazon 赞助，也可能不是如此。

Table of Contents

简介	1
目标受众	1
向量概述	2
矢量数据库概述	4
矢量数据库选项	6
单个矢量数据库选项	6
Amazon Kendra	6
亚马逊 OpenSearch 服务	7
带有 pgvector 的 Amazon RDS for PostgreSQL	7
Amazon DocumentDB	8
Amazon MemoryDB	8
Amazon Neptune Analytics	9
Amazon S3 Vectors	10
托管服务选项	10
选择正确的矢量数据库	11
矢量数据库比较	13
单个矢量数据库	13
托管服务 — Amazon Bedrock 知识库	15
在个人选项和托管选项之间进行选择	17
成本比较和注意事项	19
矢量数据库用例	22
使用 Amazon Kendra 进行知识管理	22
使用 OpenSearch 无服务器进行实时分析	22
后续步骤和资源	24
资源	24
AWS 博客文章	24
AWS 服务文档	25
其他 AWS 资源	25
其他 资源	25
文档历史记录	26
.....	xxvii

为 RAG 用例选择 AWS 矢量数据库

亚马逊 Web Services 的 Mayuri Shinde、Ivan Cui 和 Anand Bukkapatnam Tirumala

2026 年 3 月 ([文件历史记录](#))

对于实施生成式 AI 应用程序的组织来说，矢量数据库变得越来越重要。这些数据库存储和管理向量，这些向量是数据的数字表示形式，能够以捕捉文本、图像和其他内容的含义和关系的方式处理文本、图像和其他内容。

在组织探索矢量数据库选项 AWS 时，他们需要了解不同解决方案的功能、权衡和最佳实践。本指南可帮助您比较常用的矢量存储，AWS 并就哪些选项最适合您的特定需求或[用例](#)做出明智的决定。无论您是在实现检索增强生成 (RAG)、构建推荐系统还是开发其他 AI 应用程序，本指南都提供了一个框架来帮助您评估和选择矢量数据库解决方案。

目标受众

本指南适用于担任以下角色的人员：

- 数据科学家和机器学习 (ML) 工程师，他们使用矢量数据库存储和检索 ML 模型的高维数据。
- 设计和实现数据管道的数据工程师，其中包括用于存储和处理高维数据的矢量数据库。
- MLOps 使用矢量数据库作为 ML 管道的一部分来存储和提供模型输出或中间表示的工程师。
- 将矢量数据库集成到需要相似度搜索或推荐系统的应用程序中的软件工程师。
- DevOps 负责在生产环境中部署和维护矢量数据库的工程师，确保可扩展性和可靠性。
- 使用矢量数据库存储和分析嵌入或特征向量的大型数据集的人工智能研究人员。
- 需要了解矢量数据库的功能和局限性，以便就产品功能和架构做出明智决策的人工智能产品经理。

向量概述

向量是帮助机器理解 and 处理数据的数值表示。在生成式人工智能中，它们有两个关键用途：

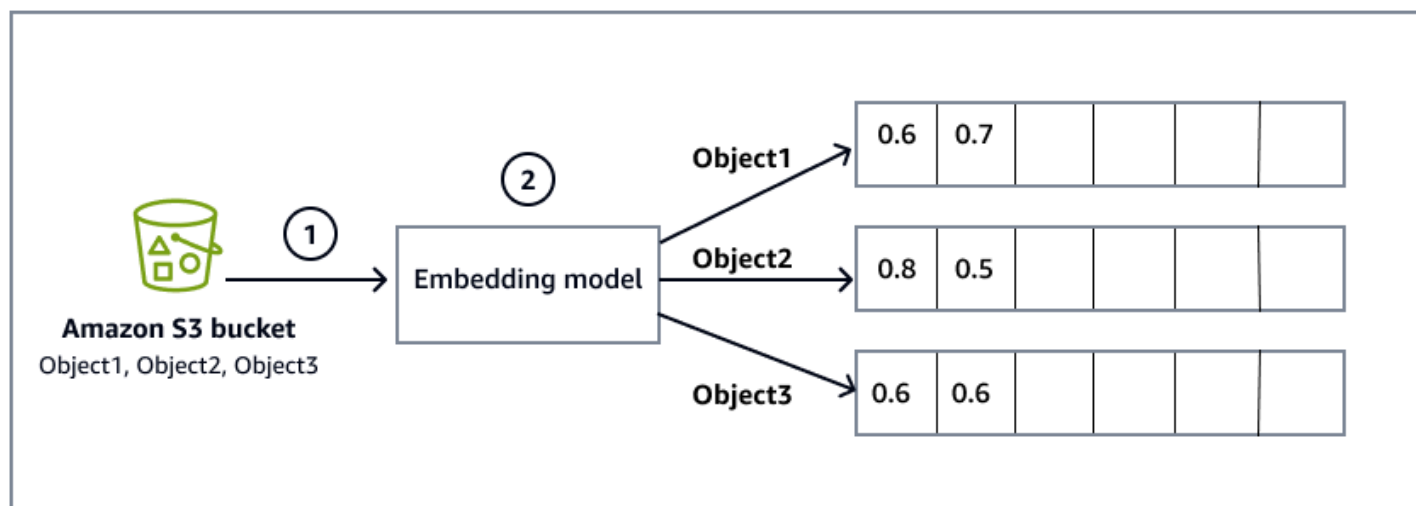
- 表示以压缩形式捕获数据结构的潜在空间
- 为数据（例如单词、句子和图像）创建嵌入式

诸如 [Word2Vec GloVe](#) 和 [Amazon Titan 文本嵌入](#) 之类的嵌入模型通过称为嵌入的过程将数据转换为向量。这些嵌入模型可以执行以下操作：

- 从上下文中学习，将单词表示为向量
- 在向量空间中将相似的单词放得更近
- 使机器能够在连续的空间中处理数据

下图简要概述了嵌入过程：

1. A [Amazon Simple Storage Service \(Amazon S3\)](#) 存储桶包含文件，这些文件是系统从中读取和处理信息的数据源。Amazon S3 存储桶是在配置 [Amazon Bedrock](#) 知识库时指定的，其中还包括将[数据与知识库同步](#)。
2. 嵌入模型将 Amazon S3 存储桶中对象文件中的原始数据转换为矢量嵌入。例如，Object1 被转换为 $[0.6, 0.7, \dots]$ 表示其在多维空间中的内容的向量。



单词嵌入对于自然语言处理 (NLP) 至关重要，因为它们具有以下作用：

- 捕捉单词之间的语义关系
- 启用生成与上下文相关的文本
- 支持大型语言模型 (LLM) 以生成类似人类的响应

矢量数据库概述

矢量数据库是一种专门的系统，可以有效地存储和查询高维向量。这些数据库是检索增强生成 (RAG) 应用程序的基础。

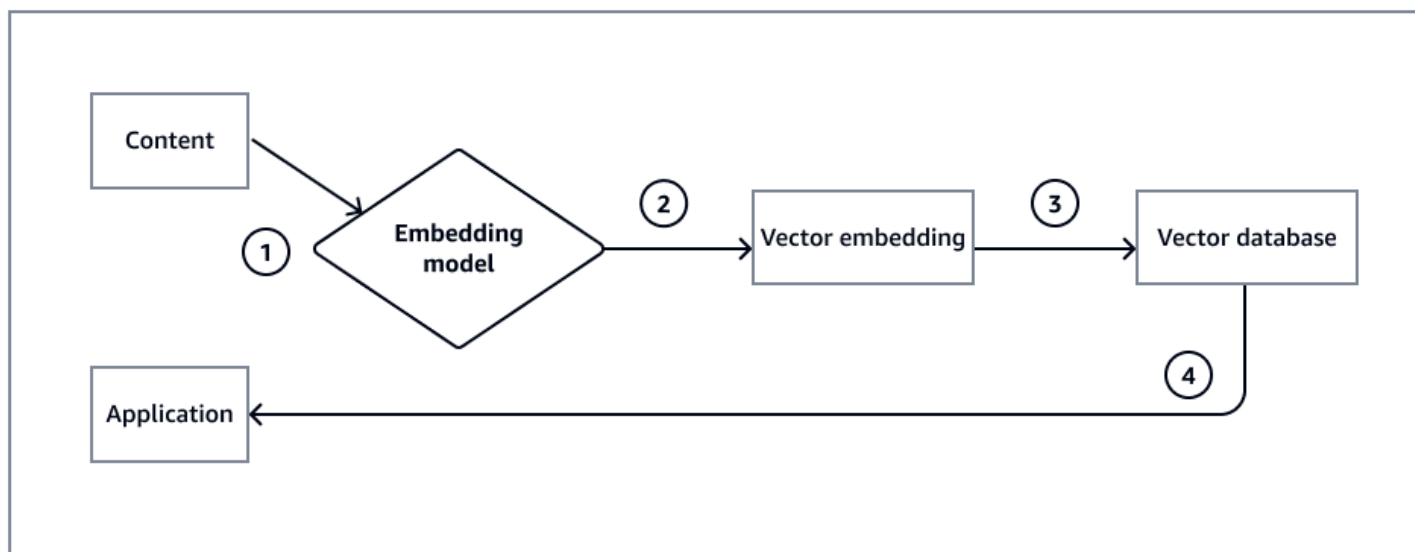
矢量数据库通过以下方式处理数据转换和存储：

- 使用嵌入模型将对象（例如音频、图像和文本文件）转换为向量。
- 向量以专门的数据格式存储。
- 矢量数据库支持快速相似度搜索。

与传统数据库相比，矢量数据库具有多项关键优势，因此特别适合应对现代数据挑战。它们专门针对向量运算进行了优化，可以高效处理高维数据。他们还专门研究传统数据库难以应对的相似度搜索。除了这些核心功能外，矢量数据库还旨在满足机器学习和生成式人工智能应用不断变化的需求。他们擅长大规模矢量存储，并使用分布式计算来平衡多个节点之间的工作负载。随着数据量的增长，这提供了可扩展性和性能。

下图显示了 RAG 的实现：

1. 文档、PDF 或文本文件等内容作为原始数据输入到嵌入模型中进行处理。
2. 嵌入模型将原始数据转换为数字向量，这些向量表示内容的语义含义。
3. 生成的向量嵌入存储在矢量数据库中，该数据库针对高维向量的存储和检索进行了优化。
4. 应用程序现在可以查询矢量数据库，以响应语义搜索和内容推荐等用例。



为RAG解决方案选择不合适的矢量数据库可能会导致重大的困难和局限性，包括：

- 查询性能不佳
- 可扩展性瓶颈
- 数据摄取挑战
- 缺少高级功能，例如筛选和排名
- 与其他系统的集成困难
- 持久性和耐久性问题
- 有多个用户的环境中的并发性和一致性问题
- 更高的许可成本或供应商锁定
- 社区支持和资源有限
- 潜在的安全和合规风险

矢量数据库选项

AWS 提供各种矢量数据库解决方案，以支持生成式 AI 应用中的不同用例和要求。这些选项可以大致分为单独的数据库服务和托管服务产品，每种服务都有不同的特征和优势。对于希望在保持最佳性能、可扩展性和成本效益的同时有效实施矢量搜索功能的组织来说，了解这些选项至关重要。

有关矢量数据库解决方案的更多信息，请参阅以下各节：

- [单个矢量数据库选项](#)
- [托管服务选项](#)
- [选择正确的矢量数据库](#)

单个矢量数据库选项

[上的单个矢量数据库选项](#) AWS 包括[亚马逊 Kendra](#)、[亚马逊 OpenSearch 服务](#)、带有 [pgvector](#) 的 [PostgreSQL](#) 的[亚马逊 RDS](#)、[亚马逊 MemoryDB](#)、[亚马逊 DocumentDB](#)、[亚马逊 Neptune Analytics](#) 和 [Amazon S3 Vector](#)。（[pgvector](#) 是一个开源扩展，它增加了存储和搜索 ML 生成的矢量嵌入的功能。）这些解决方案提供了不同的矢量搜索方法，允许组织根据其现有基础架构、技术要求和特定[用例](#)进行选择。

Amazon Kendra

Amazon Kendra 是一项企业级智能搜索服务，它使用自然语言处理和高级机器学习算法，从您的数据中返回搜索问题的具体答案。Amazon Kendra 简化了搜索功能的实现，使其成为生成式 AI 应用程序的有效后端解决方案。

Amazon Kendra 的其他主要功能包括：

- 与 [40 多个数据源](#) 的原生连接
- 内置数据准备功能
- 无需深厚的技术专业知识即可快速设置

亚马逊 Kendra 的好处包括以下几点

- 自动数据处理（分块、摄取、检索）
- 强大的自定义选项：

- [分面搜索](#)
- [搜索分析](#)
- [调整搜索相关性](#)
- 通过以下方式进行简单的编程访问 [适用于 Python \(Boto3\) 的 AWS SDK](#)

有关更多信息，请参阅亚马逊 Kendra 文档中的[亚马逊 Kendra 的好处](#)。

亚马逊 OpenSearch 服务

Amazon OpenSearch Service 是一项托管服务，可帮助您在部署、操作和扩展 OpenSearch 服务集群 AWS Cloud。

OpenSearch 服务的核心功能包括以下内容：

- 开源搜索和分析引擎
- 分布式架构
- 实时数据处理

使用 OpenSearch 服务的一些优势包括：

- 横向可扩展性
- RESTful API 支持
- 处理结构化和非结构化数据
- 实时数据分析
- 适用于各种部署规模

有关更多信息，请参阅 [OpenSearch 服务文档中的 Amazon OpenSearch 服务的功能](#)。

带有 pgvector 的 Amazon RDS for PostgreSQL

带有 pgvector 的 Amazon RDS for PostgreSQL 将托管关系数据库服务与 PostgreSQL AWS 的矢量处理扩展插件相结合。这种组合使组织能够在维护 Amazon RDS 的同时存储和查询高维向量。该解决方案特别适用于需要实时矢量操作而无需管理数据库基础设施开销的生成式 AI 应用程序。

带有 pgvector 的 Amazon RDS for PostgreSQL 的主要优势包括：

- 高可用性
- 自动失效转移
- 性价比高 (pay-per-use)
- 内置监控
- 实时矢量数据集成

有关更多信息，请参阅 Amazon RDS 文档[中的 Amazon RDS 的优势](#)。

Amazon DocumentDB

亚马逊 DocumentDB (兼容 MongoDB) 是一个文档数据库，在 5.0 及更高版本中提供原生矢量搜索功能。它将基于 JSON 的文档存储的灵活性与矢量搜索相结合，支持分层可导航小世界 (HNSW) 和 Inverted File Flat () 索引方法。IVFFlat

Amazon DocumentDB 的核心功能包括以下内容：

- 存储和索引最多 2,000 个维度的向量 (不使用索引最多可达 16,000 个维度)
- 向量相似度搜索的毫秒响应时间
- Support 支持欧几里得、余弦和点积距离度量
- 与现有兼容 MongoDB 的应用程序无缝集成

在以下情况下使用 Amazon DocumentDB：

- 适用于已经在使用 MongoDB 且需要矢量搜索功能的应用程序
- 适用于需要灵活的文档数据结构和语义搜索的用例
- 适用于既需要传统文档查询又需要向量相似度搜索的场景
- 适用于提供产品推荐、个性化、聊天助手和欺诈检测的应用程序

有关更多信息，请参阅亚马逊 Document [DB 文档中的向量搜索](#)亚马逊 DocumentDB。

Amazon MemoryDB

Amazon MemoryDB 是一款与 Redis 兼容的内存数据库，其矢量搜索性能是目前流行的矢量数据库中最快的。AWS 它提供亚毫秒级的查询延迟和多可用区域的持久性。

MemoryDB 的核心功能包括以下内容：

- 将应用程序数据和数百万个向量存储在单个数据库中
- 个位数毫秒的查询和更新响应时间
- 以最快的性能实现最高的召回率 AWS
- Support 每个向量最多支持 32,768 个维度
- 实时语义搜索和缓存功能

在以下情况下使用 MemoryDB：

- 适用于需要超低延迟（低于 10 毫秒）的实时应用程序
- 适用于每天处理数百万个请求的高吞吐量工作负载
- 适用于实时推荐引擎、语义缓存和异常检测等用例
- 适用于同时需要内存数据存储和矢量搜索功能的应用程序

有关更多信息，请参阅 MemoryDB 文档中的[矢量搜索](#)。

Amazon Neptune Analytics

Amazon Neptune Analytics 是一款提供原生矢量搜索功能的图形分析引擎，非常适合图形检索增强生成 (GraphRag) 用例。它将向量相似度搜索与图遍历和算法相结合。

Neptune Analytics 的核心功能包括以下内容：

- 在几秒钟内分析数百亿个关系
- 将矢量搜索与图形算法（路径查找、群落检测、中心性）相结合
- 支持具有拓扑知识的 GraphRag 应用程序
- 比现有图形分析解决方案快 80 倍
- 与 Amazon Bedrock 集成，实现完全托管 GraphRag

在以下情况下使用 Neptune Analytics：

- 适用于需要带有矢量嵌入的知识图的 GraphRag 应用程序
- 适用于需要遍历复杂关系以及向量相似度的用例
- 适用于需要通过关系上下文进行可解释的 AI 响应的应用程序
- 适用于客户 360 度视图、欺诈检测网络和知识发现等场景

有关更多信息，请参阅[亚马逊 Neptune Analytics 文档](#)。

Amazon S3 Vectors

Amazon S3 Vectors 是第一个 AWS 具有原生矢量存储和查询功能的云对象存储。它为需要大规模规模的人工智能应用程序提供专门构建的、经过成本优化的矢量存储。

Amazon S3 Vectors 的核心功能包括以下内容：

- 每个索引最多可存储 20 亿个向量，每个向量桶最多支持 10,000 个索引
- 低于 100 毫秒的查询延迟，针对长期存储和不频繁访问模式进行了优化
- 与专业矢量数据库相比，向量运算的成本最多可降低 90%
- 无服务器架构，具有自动扩展功能和 99.999999999% (11 9 秒) 的耐久性

在以下情况下使用 Amazon S3 矢量图：

- 适用于需要以最低成本存储数十亿个向量的应用程序
- 适用于允许亚秒级查询延迟 (100 毫秒或更长) 而不是低于 10 毫秒的工作负载
- 用于长期保存载体和存档用例
- 适用于检索模式不频繁的 RAG 应用程序
- 适用于优先考虑存储经济性而不是超低延迟的组织

Amazon S3 Vectors 与亚马逊 Bedrock 知识库原生集成，在分层架构中与亚马逊 OpenSearch 服务配合得很好。您可以将 Amazon S3 矢量用于冷存储，也可以使用 OpenSearch 服务进行热查询。

有关更多信息，请参阅 [Amazon S3 文档中的使用 S3 向量和矢量存储桶](#)。

托管服务选项

Amazon Bedrock 知识库代表了实现矢量数据库的 AWS 完全托管方法。该服务在存储选项方面的灵活性，再加上其自动化管理功能，对于希望在不管理复杂基础架构的情况下实施 RAG 的组织来说特别有价值。

借助 Amazon Bedrock 知识库，您可以创建、维护和查询知识库，从而使用 RAG 增强基础模型。该服务通过管理整个数据摄取、矢量化和检索管道，简化了实施 RAG 的复杂过程。

Amazon Bedrock 知识库的主要优势包括：

- 简化了数据处理
 - 自动数据摄取和分块
 - 内置从多种文件格式提取文本
 - 托管向量嵌入生成
 - 自动提取元数据并编制索引
- 简化了 RAG 的实现
 - 预先配置的检索策略
 - 自动优化上下文窗口
 - 内置相关性调整
 - 开箱即用的语义搜索功能
- 安全和治理
 - 集成 AWS Identity and Access Management (IAM) 控件
 - 静态和传输中的数据加密
 - VPC 支持
 - 使用审核日志 AWS CloudTrail

Amazon Bedrock 知识库支持多种[矢量存储选项](#)，包括：

- 带有 pgvector 的亚马逊 Aurora PostgreSQL
- Amazon Neptune Analytics
- Amazon EMR Serverless
- Amazon S3 Vectors
- Pinecone
- Redis 企业云

此托管服务可处理自动数据摄取、矢量化和检索。这简化了 RAG 的实现。

有关每个支持的矢量存储的详细信息，请参阅 [Amazon Bedrock 知识库文档](#)。

选择正确的矢量数据库

根据以下关键决策因素选择您的矢量数据库：

选择正确的矢量数据库

- 如果您需要具有矢量搜索功能的 MongoDB 兼容文档数据库，请选择 Amazon DocumentDB。当您的应用程序使用 MongoDB 并且您想在管理单独的向量基础设施的情况下添加语义搜索功能时，这是理想的选择。
- 如果您需要超低延迟的实时应用程序，请选择 Amazon MemoryDB。这提供了最快的矢量搜索性能，响应 AWS 时间为亚毫秒。它非常适合实时推荐引擎和高吞吐量应用程序。
- 如果您需要通过矢量搜索进行基于图表的知识表达，请选择 Amazon Neptune Analytics。这最适合需要遍历复杂关系并在向量搜索的同时执行基于图形的查询的 GraphRag 应用程序，从而提供可解释的 AI 响应。
- 如果您需要将关系查询与向量搜索相结合，请选择带有 pgvector 的 Amazon Aurora PostgreSQL。当您的应用程序需要在同一个数据库中同时进行传统 SQL 操作和向量相似度搜索时，此选项非常理想。
- 如果您需要延迟低于 10 毫秒的高吞吐量查询，请选择 Amazon OpenSearch 服务。它擅长处理高频查询和实时应用程序，并包括最近的 GPU 加速改进。
- 如果您需要经济高效地存储数十亿个向量，请选择 Amazon S3 向量图。此选项最多可节省 90% 的成本，非常适合具有不频繁检索模式（查询间隔几分钟到几小时）且可容忍低于 100 毫秒延迟的应用程序。
- 如果您需要在进行矢量搜索的同时进行全文搜索，请选择 Amazon OpenSearch 服务。此选项在单个平台上将强大的全文搜索功能与向量搜索相结合。

矢量数据库比较

AWS 提供了多种实现矢量搜索功能的方法，从单个矢量数据库到 Amazon Bedrock 知识库，后者是一项完全托管的服务。在评估这些选项时，组织必须考虑各个方面，包括架构、可扩展性、集成能力、性能特征和安全功能。

单个矢量数据库

下表概述了几种 AWS 单独的矢量数据库解决方案的主要功能，重点介绍其架构、扩展能力、数据源集成和性能特征。

功能	Amazon Kendra	亚马逊 OpenSearch 服务	PostgreSQL 版本亚马逊 RDS pgvector	Amazon DocumentDB	Amazon MemoryDB	Amazon Neptune Analytics	Amazon S3 Vectors
主要使用案例	企业搜索和 RAG	分布式搜索和分析	支持向量的关系数据库	使用矢量搜索功能的文档数据库	实时内存中矢量搜索	使用矢量搜索进行图形分析	成本优化的矢量存储
架构	完全托管	分布式集群	关系数据库	以文档为导向	内存数据库	图形分析引擎	无服务器对象存储
数据模型	基于文档	JSON 文档	关系表	JSON 文档	带有 JSON 的键值	房产图	对象存储
向量尺寸	自动管理	最多 16,000	可配置	最多 2,000 (已编入索引) ； 16,000 (未编入索引)	最多 32,768	可配置	最多 4,096

索引方法	自动	哈哈哈，试管婴儿	新南威尔士州，IVFFlat	新南威尔士州，IVFFlat	HNSW	原生图形和向量	自动
距离指标	自动	余弦、欧几里得、点积	余弦、欧几里得、内积	余弦、欧几里得、点积	余弦、欧几里得、内积	余弦、欧几里得	余弦、欧几里得
查询延迟	亚秒级	小于 10 毫秒（ GPU 加速 ）	10-100 ms	毫秒	亚毫秒级	亚秒级	低于 100 毫秒
缩放模型	自动	水平（ 添加节点 ）	垂直副本和只读副本	水平（ 添加实例 ）	垂直和复制品	自动	自动（ 无服务器 ）
最大向量	托管式	十亿（ 取决于集群 ）	数百万（ 取决于实例 ）	每个收藏数百万美元	每个数据库数百万美元	数十亿	每个索引 20 亿个；每个存储桶 10,000 个索引
吞吐量	高	非常高（ 数千个 QPS ）	中	高	非常高（ 每天有数百万个请求 ）	高	中（ 针对不频繁的查询进行了优化 ）
数据持久性	99.999999 999%（ 11 个 9 ）	可通过副本进行配置	99.99%（ 多可用区 ）	99.99%（ 多可用区 ）	99.99%（ 多可用区 ）	99.99%	99.999999 999%（ 11 个 9 ）
一致性模型	最终	最终（ 可配置 ）	强（ 酸 ）	最终	很强	很强	很强

附加功能	40 个或更多数据连接器，NLP	全文搜索、分析、仪表板	SQL 查询、ACID 事务	MongoDB API 兼容性	Redis API 兼容性、缓存	图形算法、遍历	Amazon S3 集成、生命周期策略
定价模式	按查询和存储付费	实例小时数和存储空间	实例小时数和存储空间	实例小时数和存储空间	实例小时数和存储空间	容量单位和存储空间	存储、查询和数据传输
成本优化	基于用量	预留实例，自动缩放	预留实例，Aurora 无服务器	预留实例	预留实例	自动扩缩	与专业相比，最多可节省 90% DBs
适用于	只需最少的设置即可进行企业搜索	高吞吐量、低延迟查询	混合 SQL 和矢量工作负载	兼容 MongoDB 的应用程序需要向量	实时、超低延迟的应用程序	GraphRag 和知识图表	经济实惠的长期存储
理想的查询模式	频繁的企业搜索	高频实时查询	混合 SQL 和向量查询	使用语义搜索进行文档查询	每天有数百万个请求	使用矢量搜索进行图形遍历	不频繁的查询（几分钟到几小时）
设置复杂性	低（完全托管）	中（群集配置）	中等（扩展设置）	中（群集配置）	中（群集配置）	低（完全托管）	低（无服务器）
需要团队专业知识	Minimal	OpenSearch 或者弹性搜索	PostgreSQL、SQL	MongoDB	Redis	图形数据库	亚马逊 S3，基本矢量概念

托管服务 — Amazon Bedrock 知识库

Amazon Bedrock 知识库提供了具有多个矢量存储选项的完全托管解决方案。下表比较了这些存储选项。

功能	Aurora Postgre pgvect SQLwith	Neptune 分 析	OpenSearc h 无服务器 服务	亚马逊 S3 矢量图	Pinecone	RedisEnte rprise 云
主要使用案例	带有向量 RAG 的关系数据库	GraphRag 的基于图形的 矢量搜索	知识管理 RAG	成本优化的 矢量 RAG	高性能矢量 搜索	内存中的矢 量搜索
架构	完全托管的 关系	完全托管的 图形分析	完全托管的 无服务器	无服务器对 象存储	完全托管的 混合云	完全托管的 内存中
数据模型	关系表	房产图	JSON 文档	对象存储	专门构建的 矢量图	带向量的键 值
矢量存储	通过 pgvector 扩 展	原生图形向 量	通过 OpenSearc h 发动机	亚马逊 S3 原生矢量存 储	原生矢量数 据库	内存中的向 量
Amazon Bedrock 集 成	Native	Native	Native	Native	Native	Native
自动摄取	是 (通过 Amazon Bedrock)	是 (通过 Amazon Bedrock)	是 (通过 Amazon Bedrock)	是 (通过 Amazon Bedrock)	是 (通过 Amazon Bedrock)	是 (通过 Amazon Bedrock)
自动矢量化	是 (通过 Amazon Bedrock)	是 (通过 Amazon Bedrock)	是 (通过 Amazon Bedrock)	是 (通过 Amazon Bedrock)	是 (通过 Amazon Bedrock)	是 (通过 Amazon Bedrock)
扩展	自动扩展 (Aurora 无服务器)	自动缩放图 表	自动无服务 器	自动 (数十 亿个向量)	自动缩放 pod	自动扩展集 群
查询性能	关系或向量 为高	图形向量的 值很高	高	中等 (延迟 100 毫秒或 更长)	非常高	非常高

最大向量	数百万 (取决于实例)	数十亿	数十亿	每个指数 20 亿美元	数十亿	数百万 (依赖内存)
附加功能	SQL 查询、ACID 事务	图形算法、遍历	全文搜索、分析	Amazon S3 生命周期，分层	元数据筛选、命名空间	Redis 数据结构，缓存
成本优化	中等 (Aurora 无服务器)	中等 (容量单位)	高 (无服务器， pay-per-use)	非常高 (最多可节省 90%)	中等 (基于 Pod 的定价)	低 (内存溢价)
适用于	混合 SQL/ vector 工作负载	互联知识图表	带矢量搜索的全文	长期、不经常访问的向量	大规模实时矢量搜索	超低延迟需求
理想的查询模式	混合 SQL 和向量查询	使用向量进行图遍历	使用分析进行频繁搜索	不频繁检索 (几分钟到几小时)	高频实时查询	每秒数百万个请求
使用 Amazon Bedrock 进行设置	简单 (由 Amazon Bedrock 管理)	简单 (由 Amazon Bedrock 管理)	简单 (由 Amazon Bedrock 管理)	简单 (由 Amazon Bedrock 管理)	简单 (由 Amazon Bedrock 管理)	简单 (由 Amazon Bedrock 管理)
数据驻留	AWS 区域	AWS 区域	AWS 区域	AWS 区域	多云 (AWS 和其他)	多云 (AWS 和其他)
定价模式	实例小时数和存储空间	容量单位和存储空间	计算和存储 (无服务器)	存储、查询和传输	Pod 使用时间和存储空间	节点使用时间和存储空间

在个人选项和托管选项之间进行选择

考虑	选择单个向量 DB	选择 Amazon Bedrock 知识库 (托管)
----	-----------	------------------------------

RAG 的实现	你想完全控制 RAG 管道	你想要完全托管的 RAG，只需最少的设置
自定义	您需要自定义检索逻辑和预处理	标准 RAG 图案可满足您的需求
现有基础架构	您已经部署了数据库	您正在重新开始，或者想要简化管理
团队专业知识	您的团队拥有数据库管理专业知识	您更愿意关注应用程序逻辑，而不是基础架构
集成复杂性	您需要与现有系统进行深度集成	你想快速与 Amazon Bedrock 模型集成
运营开销	您可以管理数据库操作	你 AWS 想处理操作
成本结构	您更喜欢直接数据库定价	你更喜欢统一的 Amazon Bedrock 定价
是时候上市了	你有时间进行自定义实现	你需要快速部署

成本比较和注意事项

了解不同矢量数据库选项的成本结构对于做出明智的实施决策至关重要。下表概述了各种矢量数据库解决方案的一些关键成本注意事项，包括单个数据库和托管服务。每个选项都有不同的定价因素，这些因素可能会影响总拥有成本 (TCO)，从 pay-as-you-go 模型到基础架构和运营成本。

矢量数据库	成本模型	成本考虑因素
Amazon Kendra	根据查询情况按使用量付费	费用可能因查询数量和索引的数据量而异。数据存储和数据传输可能会收取额外费用。有关更多信息，请参阅 Amazon Kendra 定价 。
亚马逊 OpenSearch 服务	即用即付，基于实例小时数和存储空间	费用包括实例使用时间、存储（Amazon EBS 卷）、数据传输和可选 UltraWarm 存储。预留实例最多可节省 30%。GPU 加速实例为矢量工作负载提供了更好的性价比。有关更多信息，请参阅 Amazon OpenSearch 服务定价 。
开源 OpenSearch	开源，没有直接成本（您无需付费即可下载或使用该软件，也没有许可费用）	成本包括基础架构（例如服务器和存储）和运营成本（例如维护和监控）。Organizations 需要为管理和维护基础架构的人员编制预算。
带有 pgvector 的 Amazon RDS for PostgreSQL	即用即付，根据使用情况而定	成本包括数据库实例类型、存储、数据传输和备份。超出 AWS 免费套餐的数据传输、实例类型和存储可能会收取额外费用。有关更多信息，请参阅 Amazon RDS 定价 。

Amazon DocumentDB	即用即付，基于实例小时数和存储空间	费用包括实例小时数、存储空间（每月 GB）、I/O 请求、备份存储和数据传输。弹性集群支持动态扩展。预留实例可用于成本优化。有关更多信息，请参阅 Amazon DocumentDB 定价 。
Amazon MemoryDB	即用即付，基于节点小时数和数据存储空间	成本包括节点小时数（每种节点类型）、数据存储（GB 小时）、快照存储和数据传输。预留节点最多可节省 55%。针对高吞吐量、低延迟的工作负载进行了优化。有关更多信息，请参阅 Amazon MemoryDB 定价 。
Amazon Neptune Analytics	即用即付，基于容量单位	成本包括 Neptune 容量单位 (NCUs)、存储（GB / 月）和数据传输。根据工作负载自动扩展，无需预先承诺。至少 NCUs 需要 128 个。有关更多信息，请参阅 亚马逊 Neptune 定价 。
Amazon S3 Vectors	根据存储空间和请求量按使用量付费	成本包括存储（GB / 月）、PUT 和 GET 请求、向量索引管理和数据传输。与专业矢量数据库相比，最多可节省 90% 的成本。Amazon S3 智能分层和生命周期策略可用于进一步优化。有关更多信息，请参阅 Amazon S3 定价 。

Amazon 基岩知识库

即用即付，根据使用情况而定

费用可能因知识库和其他服务（例如 Amazon OpenSearch Serverless）的使用情况而异。数据存储、数据传输和其他功能可能会收取额外费用。有关定价的更多信息，请参阅[Amazon OpenSearch 服务定价](#)。

矢量数据库用例

以下示例重点介绍了如何有效地使用不同的矢量数据库选项来增强知识管理、提高运营效率和提供更好的业务成果。这些用例说明了本指南前面讨论的矢量数据库解决方案的实际应用，并提供了对它们在现实世界中的性能和优势的见解。

使用 Amazon Kendra 进行知识管理

客户问题 — 日本最大的总承包商之一面临着经验人员减少的问题。公司需要一种方法来有效地将经验丰富的人员的知识和技能传授给年轻一代。他们需要一种解决方案来收集和传播复杂的建筑工程知识和过去的经验。

AWS 解决方案 — 为了解决这个问题，客户求助于了 Amazon Kendra，这是一款可以快速准确地处理其内部知识库并允许自然语言查询的人工智能解决方案。借助 Amazon Kendra，员工现在可以更快地找到所需的信息，从而提高工作效率，促进经验丰富的员工向年轻员工传授知识。

影响力 — 通过实施由 Amazon Kendra 提供支持的生成式人工智能聊天机器人，该公司创建了一个统一的知识平台。聊天机器人允许员工快速访问建筑工程方面的技术知识和过去的经验。该解决方案显著提高了组织内部知识转移和决策流程的效率，有助于确保宝贵的专业知识得到保留并便于所有员工使用。此解决方案的成本可能会有所不同，具体取决于您的使用情况和配置。有关详细的成本估算，请参阅[AWS 定价计算器](#)。有关矢量数据库成本估算，请参阅本指南的[成本比较和注意事项](#)部分或参阅[Amazon Kendra](#) 定价。

有关其他客户使用案例的信息，请参阅 [Amazon Kendra](#) 客户。

使用 OpenSearch 无服务器进行实时分析

客户问题 — 一家领先的金融服务提供商面临着管理庞大的数据生态系统的挑战。它每年处理3亿份授权和900亿笔交易，累积的数据约为1.1 PB (PB)。现有系统为 300,000 名需要访问 6,000 多份报告的用户提供服务，需要进行现代化改造，以提供全球一致性并实现实时决策。

AWS 解决方案 — 解决方案架构使用 Amazon Bedrock 提供的基础模型 (包括 Anthropic、Sonnet 3、Sonnet 3.5 和 Haiku) 进行自然语言处理。客户之所以选择 OpenSearch Serverless 作为矢量数据库，是因为它具有卓越的可扩展性和高效处理海量数据的能力。这种架构实现了复杂查询的无缝处理和动态报告的生成。

影响 — 该实施无需手动生成 100 多个商业智能仪表板，从而将工作效率提高了 50%。用户现在可以通过自然语言查询生成报告，响应时间在 20-40 秒之间。此解决方案的成本可能会有所不同，具体取

决于您的使用情况和配置。有关详细的成本估算，请参阅[AWS 定价计算器](#)。有关矢量数据库成本估算，请参阅本指南的[成本比较和注意事项](#)部分或参阅 [Amazon S OpenSearch service 定价](#)。有关其他客户使用案例的信息，请参阅 [Amazon OpenSearch Serverless](#)。

后续步骤和资源

阅读本指南后，请考虑采取以下行动，从理解转变为实施：

1. 评估您当前的需求：
 - 评估您现有的数据库基础架构和专业知识。
 - 记录您的特定矢量搜索要求。
 - 定义您的绩效、扩展和成本目标。
2. 选择以下选项之一来测试矢量数据库选项：
 - 选项 1：使用您首选的矢量数据库解决方案设置概念验证。
 - 选项 2：在 Amazon Bedrock 知识库中使用示例数据集进行实验。试试 Amazon Bedrock 知识库的快速创建体验。有关示例，请参阅 Aurora 文档中的为[亚马逊 Bedrock 快速创建 Aurora PostgreSQL 知识库](#)。
3. 查看其他[资源](#)。
4. 获取专家帮助：
 - 请联系您的 AWS 账户 团队或 AWS 解决方案架构师获取实施指导。
 - [与专门从事矢量数据库的 AWS 合作伙伴合作](#)。
5. 规划您的生产部署：
 - 如果从现有数据库迁移，请制定迁移策略。
 - 为您选择的解决方案制定扩展计划。
 - 设计您的监控和维护程序。

资源

以下资源可以帮助您选择矢量数据库。

AWS 博客文章

- [使用 Amazon Bedrock 知识库快速创建和亚马逊 Aurora Serverless 加速生成式 AI 应用程序的开发](#)
- [亚马逊 OpenSearch 服务的矢量数据库功能详解](#)
- [使用 Amazon Bedrock 知识库深入研究矢量数据存储](#)
- [利用 pgvector 和 Amazon Aurora PostgreSQL 进行自然语言处理、聊天机器人和情感分析](#)

AWS 服务文档

- [选择 AWS 数据库服务](#)
- [Amazon Bedrock 知识库是如何运作的](#)
- [Neptune Analytics 文档](#)
- [亚马逊 Web Services 概述：数据库](#)
- [使用 Aurora PostgreSQL 作为亚马逊 Bedrock 的知识库](#)
- [使用 Amazon Aurora PostgreSQL](#)
- [Amazon DocumentDB](#)
- [Amazon MemoryDB](#)
- [Amazon S3 Vectors](#)

其他 AWS 资源

- [Amazon 基岩知识库](#)
- [矢量数据库和嵌入式](#)
- [用于生成式 AI 应用的矢量数据库](#)
- [Machine Learning 中的嵌入式是什么？](#)

其他 资源

- [关于 PostgreSQL](#)
- [pgvector 文档](#)
- [Pinecone 作为 Amazon Bedrock 的知识库](#)
- [Redis 企业云已开启 AWS](#)

文档历史记录

下表描述了本指南“为 RAG 用例选择 AWS 矢量数据库”的重大更改。如果您希望收到有关未来更新的通知，可以订阅 [RSS 源](#)。

变更	说明	日期
已添加 AWS 服务	我们添加了有关使用亚马逊 DocumentDB、Amazon MemoryDB、Amazon S3 Vectors 和亚马逊 Neptune Analytics 的信息。	2026年3月30日
初次发布	—	2025 年 3 月 6 日

本文属于机器翻译版本。若本译文内容与英语原文存在差异，则一律以英文原文为准。