



選擇 RAG 使用案例的 AWS 向量資料庫

AWS 方案指引



AWS 方案指引: 選擇 RAG 使用案例的 AWS 向量資料庫

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

Amazon 的商標和商業外觀不得用於任何非 Amazon 的產品或服務，也不能以任何可能造成客戶混淆、任何貶低或使 Amazon 名譽受損的方式使用 Amazon 的商標和商業外觀。所有其他非 Amazon 擁有的商標均為其各自擁有者的財產，這些擁有者可能附屬於 Amazon，或與 Amazon 有合作關係，亦或受到 Amazon 贊助。

Table of Contents

簡介	1
目標對象	1
向量概觀	2
向量資料庫概觀	4
向量資料庫選項	6
個別向量資料庫選項	6
Amazon Kendra	6
Amazon OpenSearch Service	7
Amazon RDS for PostgreSQL 搭配 pgvector	7
Amazon DocumentDB	8
Amazon MemoryDB	8
Amazon Neptune Analytics	9
Amazon S3 Vectors	10
受管服務選項	10
選擇正確的向量資料庫	12
向量資料庫比較	13
個別向量資料庫	13
受管服務 – Amazon Bedrock 知識庫	15
在個別選項和受管選項之間進行選擇	17
成本比較和考量事項	19
向量資料庫使用案例	22
Amazon Kendra 的知識管理	22
使用 OpenSearch Serverless 進行即時分析	22
後續步驟和資源	24
Resources	24
AWS 部落格文章	24
AWS 服務文件	25
其他 AWS 資源	25
其他資源	25
文件歷史紀錄	26
.....	xxvii

選擇 RAG 使用案例的 AWS 向量資料庫

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

2026 年 3 月 ([文件歷史記錄](#))

對於實作生成式 AI 應用程式的組織而言，向量資料庫變得越來越重要。這些資料庫會存放和管理向量，這些向量是資料的數值表示法，能夠以擷取其意義和關係的方式處理文字、影像和其他內容。

當組織探索向量資料庫選項時 AWS，他們需要了解不同解決方案的功能、權衡和最佳實務。本指南可協助您比較上常用的向量存放區，AWS 並針對哪些選項最適合您的特定需求或使用[案例](#)做出明智的決策。無論您是實作擷取增強生成 (RAG)、建置建議系統，還是開發其他 AI 應用程式，本指南都會提供架構，協助您評估和選擇向量資料庫解決方案。

目標對象

本指南適用於擔任下列角色的人員：

- 使用向量資料庫來存放和擷取 ML 模型之高維度資料的資料科學家和機器學習 (ML) 工程師。
- 設計和實作包含向量資料庫的資料管道的資料工程師，用於存放和處理高維度資料。
- 在 ML 管道中使用向量資料庫來存放和提供模型輸出或中繼表示法的 MLOps 工程師。
- 將向量資料庫整合到需要相似性搜尋或建議系統的應用程式的軟體工程師。
- 負責在生產環境中部署和維護向量資料庫的 DevOps 工程師，確保可擴展性和可靠性。
- 使用向量資料庫來存放和分析內嵌或特徵向量大型資料集的 AI 研究人員。
- 需要了解向量資料庫的功能和限制，才能對產品功能和架構做出明智決策的 AI 產品管理員。

向量概觀

向量是數值表示法，可協助機器了解和處理資料。在生成式 AI 中，它們提供兩個主要目的：

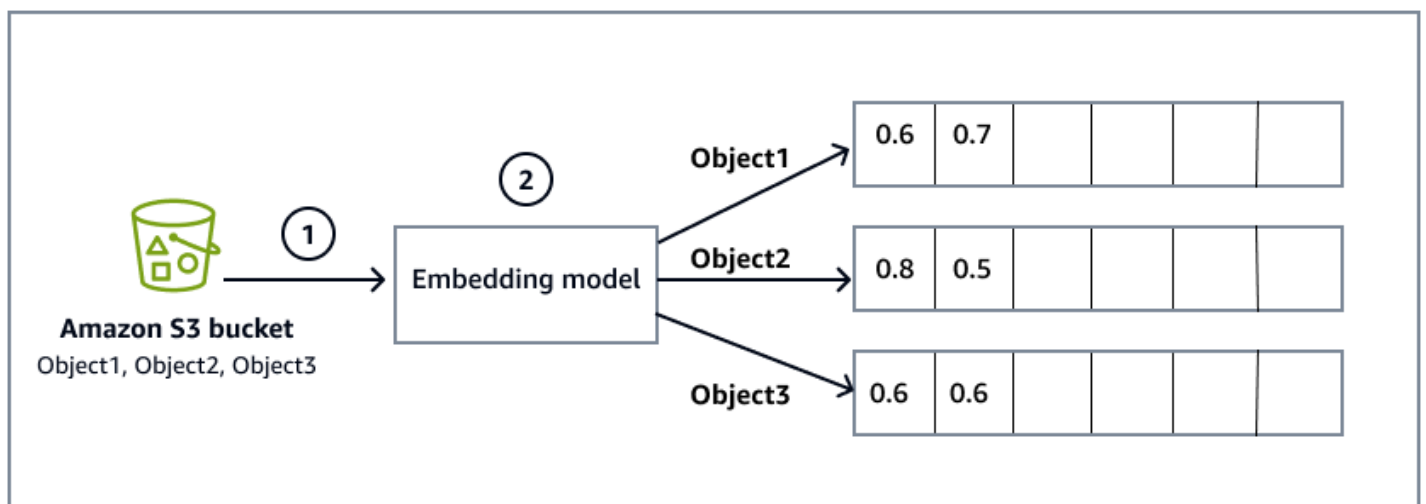
- 代表以壓縮形式擷取資料結構的隱含空間
- 建立資料的內嵌，例如單字、句子和影像

內嵌 [Word2Vec](#)、[GloVe](#) 和 [Amazon Titan Text Embeddings](#) 等模型會透過稱為內嵌的程序將資料轉換為向量。這些內嵌模型可以執行下列動作：

- 從內容中學習，以將單字表示為向量
- 在向量空間中將類似的字詞放在更接近的位置
- 讓機器能夠處理連續空間中的資料

下圖提供內嵌程序的高階概觀：

1. [Amazon Simple Storage Service \(Amazon S3\)](#) 儲存貯體包含的檔案是系統會從中讀取和處理資訊的資料來源。Amazon S3 儲存貯體是在 [Amazon Bedrock](#) 知識庫組態期間指定，其中包括[與知識庫同步資料](#)。
2. 內嵌模型會將 Amazon S3 儲存貯體中物件檔案的原始資料轉換為向量內嵌。例如，Object1 會轉換為向量[0.6, 0.7, ...]，代表其在多維空間中的內容。



文字內嵌對於自然語言處理 (NLP) 至關重要，因為它們會執行下列動作：

- 擷取字詞之間的語意關係
- 啟用產生內容相關的文字
- 支援大型語言模型 (LLMs) 來產生類似人類的回應

向量資料庫概觀

向量資料庫是一種特殊化系統，可有效率地存放和查詢高維度向量。這些資料庫是擷取增強產生 (RAG) 應用程式的基礎。

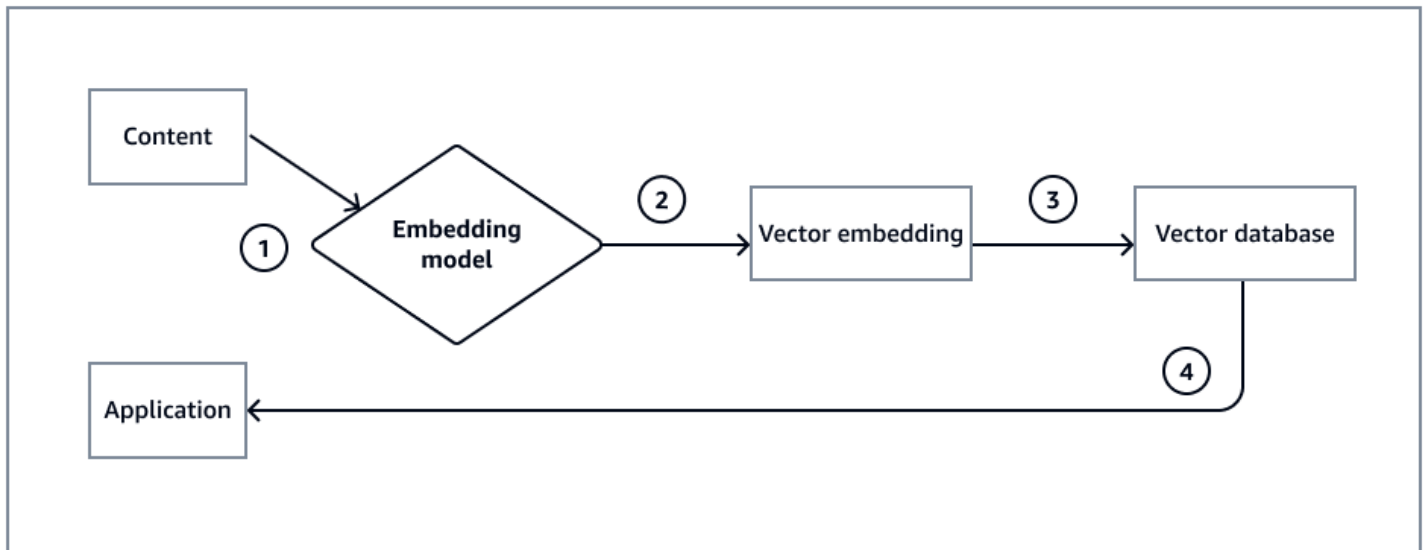
向量資料庫會以下列方式處理資料轉換和儲存：

- 物件（例如音訊、影像和文字檔案）會使用內嵌模型轉換為向量。
- 向量會以特殊的資料格式儲存。
- 向量資料庫可啟用快速相似性搜尋。

相較於傳統資料庫，向量資料庫提供多種主要優點，使其特別適合現代資料挑戰。它們專門針對向量操作進行最佳化，並有效率地處理高維度資料。他們也專精於傳統資料庫難以處理的相似性搜尋。除了這些核心功能之外，向量資料庫也是為了滿足 ML 和生成式 AI 應用程式不斷變化的需求而建置。它們擅長大規模向量儲存，並使用分散式運算來平衡跨多個節點的工作負載。隨著資料量的增長，這可提供可擴展性和效能。

下圖顯示 RAG 實作：

1. 文件、PDFs 或文字檔案等內容會以原始資料的形式饋送至內嵌模型進行處理。
2. 內嵌模型會將原始資料轉換為數值向量，代表內容的語意意義。
3. 產生的向量內嵌會存放在向量資料庫中，該資料庫中已針對高維度向量的儲存和擷取進行最佳化。
4. 應用程式現在可以查詢向量資料庫，以回應語意搜尋和內容建議等使用案例。



為 RAG 解決方案選擇不適當的向量資料庫可能會導致重大困難和限制，包括下列項目：

- 查詢效能不佳
- 可擴展性瓶頸
- 資料擷取挑戰
- 缺乏進階功能，例如篩選和排名
- 與其他系統的整合困難
- 持久性和持久性問題
- 多使用者環境中的並行和一致性問題
- 更高的授權成本或廠商鎖定
- 有限的社群支援和資源
- 潛在的安全與合規風險

向量資料庫選項

AWS 提供各種向量資料庫解決方案，以支援生成式 AI 應用程式中的不同使用案例和需求。這些選項可以廣泛分類為個別資料庫服務和受管服務產品，每個服務都有不同的特性和優勢。了解這些選項對於希望有效實作向量搜尋功能，同時維持最佳效能、可擴展性和成本效益的組織至關重要。

如需向量資料庫解決方案的詳細資訊，請參閱下列各節：

- [個別向量資料庫選項](#)
- [受管服務選項](#)
- [選擇正確的向量資料庫](#)

個別向量資料庫選項

上的個別向量資料庫選項 AWS 包括 [Amazon Kendra](#)、[Amazon OpenSearch Service](#)、[Amazon RDS for PostgreSQL](#) 搭配 pgvector、[Amazon MemoryDB](#)、[Amazon DocumentDB](#)、[Amazon Neptune Analytics](#) 和 [Amazon S3 Vector](#)。(開放原始碼延伸模組，pgvector 新增了儲存和搜尋 ML 產生的向量內嵌功能。) 這些解決方案提供不同的向量搜尋方法，允許組織根據其現有的基礎設施、技術需求和特定[使用案例](#)進行選擇。

Amazon Kendra

Amazon Kendra 是一種企業級智慧型搜尋服務，使用自然語言處理和進階機器學習演算法傳回從資料搜尋問題的特定答案。Amazon Kendra 簡化了搜尋功能的實作，使其成為生成式 AI 應用程式的有效後端解決方案。

Amazon Kendra 的其他主要功能包括下列項目：

- 原生連線至超過 [40 個資料來源](#)
- 內建資料準備功能
- 不需要深度技術專業知識的快速設定

Amazon Kendra 的優點包括下列項目

- 自動化資料處理（區塊、擷取、擷取）
- 強大的自訂選項：

- [面向搜尋](#)
- [搜尋分析](#)
- [調整搜尋相關性](#)
- 透過 進行簡單的程式設計存取 [適用於 Python \(Boto3\) 的 AWS SDK](#)

如需詳細資訊，請參閱 [Amazon Kendra 文件中的 Amazon Kendra 優點](#)。

Amazon OpenSearch Service

Amazon OpenSearch Service 是一種受管服務，可協助您在 中部署、操作和擴展 OpenSearch Service 叢集 AWS 雲端。

OpenSearch Service 的核心功能包括下列項目：

- 開放原始碼搜尋和分析引擎
- 分散式架構
- 即時資料處理

使用 OpenSearch Service 的一些優點包括：

- 水平可擴展性
- RESTful API 支援
- 處理結構化和非結構化資料
- 即時資料分析
- 適合各種部署大小

如需詳細資訊，請參閱 [OpenSearch Service 文件中的 Amazon](#) OpenSearch Service 的功能。

Amazon RDS for PostgreSQL 搭配 pgvector

Amazon RDS for PostgreSQL 與 [pgvector](#) 結合 AWS 受管關聯式資料庫服務與 PostgreSQL 的向量處理延伸。此組合可讓組織儲存和查詢高維度向量，同時維護 Amazon RDS。此解決方案特別適合需要即時向量操作的生成式 AI 應用程式，而不需要管理資料庫基礎設施的額外負荷。

Amazon RDS for PostgreSQL 搭配 pgvector 的主要優點如下：

- 高可用性
- 自動容錯移轉
- 經濟實惠 (pay-per-use)
- 內建監控
- 即時向量資料整合

如需詳細資訊，請參閱 [Amazon RDS 文件中的 Amazon RDS 的優點](#)。

Amazon DocumentDB

Amazon DocumentDB（與 MongoDB 相容）是一種文件資料庫，可在 5.0 版和更新版本中提供原生向量搜尋功能。它結合了 JSON 型文件儲存與向量搜尋的靈活性，同時支援階層式可導覽小型世界 (HNSW) 和反轉檔案平面 (IVFFlat) 索引方法。

Amazon DocumentDB 的核心功能包括下列項目：

- 儲存和索引向量最多 2,000 個維度（最多 16,000 個維度，無需索引）
- 向量相似性搜尋的毫秒回應時間
- 支援 euclidean、cosine 和 dot 產品距離指標
- 與現有的 MongoDB 相容應用程式無縫整合

在下列情況中使用 Amazon DocumentDB：

- 對於已經在使用 MongoDB APIs 且需要向量搜尋功能的應用程式
- 對於需要彈性文件資料結構結合語意搜尋的使用案例
- 對於需要傳統文件查詢和向量相似性搜尋的案例
- 對於提供產品建議、個人化、聊天助理和詐騙偵測的應用程式

如需詳細資訊，請參閱 [Amazon DocumentDB 文件中的 Amazon DocumentDB 向量搜尋](#)。Amazon DocumentDB

Amazon MemoryDB

Amazon MemoryDB 是與 Redis 相容的記憶體內資料庫，可在熱門向量資料庫之間提供最快的向量搜尋效能 AWS。它提供毫秒以下的查詢延遲，具有多可用區域耐久性。

MemoryDB 的核心功能包括下列項目：

- 將應用程式資料和數百萬個向量儲存在單一資料庫中
- 單一位數毫秒查詢和更新回應時間
- 上效能最快的最高召回率 AWS
- 每個向量最多支援 32,768 個維度
- 即時語意搜尋和快取功能

在下列情況中使用 MemoryDB：

- 對於需要超低延遲 (sub-10ms) 的即時應用程式
- 適用於每天有數百萬個請求的高輸送量工作負載
- 對於即時建議引擎、語意快取和異常偵測等使用案例
- 對於需要記憶體內資料存放區和向量搜尋功能的應用程式

如需詳細資訊，請參閱 MemoryDB 文件中的[向量搜尋](#)。

Amazon Neptune Analytics

Amazon Neptune Analytics 是一種圖形分析引擎，可提供原生向量搜尋功能，因此非常適合圖形擷取增強生成 (GraphRAG) 使用案例。它結合了向量相似性搜尋與圖形周遊和演算法。

Neptune Analytics 的核心功能包括下列項目：

- 在幾秒內分析數十億個關係
- 結合向量搜尋與圖形演算法（路徑調查結果、社群偵測、集中性）
- 支援具有拓撲知識的 GraphRAG 應用程式
- 比現有的圖形分析解決方案快上 80 倍
- 與 Amazon Bedrock 整合，以實現全受管 GraphRAG

在下列情況中使用 Neptune Analytics：

- 對於需要具有向量內嵌的知識圖表的 GraphRAG 應用程式
- 對於需要與向量相似性周遊複雜關係的使用案例

- 對於需要具有關係內容的可解釋 AI 回應的應用程式
- 對於客戶 360 檢視、詐騙偵測網路和知識探索等案例

如需詳細資訊，請參閱 [Amazon Neptune Analytics 文件](#)。

Amazon S3 Vectors

Amazon S3 Vectors 是中第一個 AWS 具有原生向量儲存和查詢功能的雲端物件存放區。它為需要大規模擴展的 AI 應用程式提供專用、成本最佳化的向量儲存。

Amazon S3 Vectors 的核心功能包括下列項目：

- 每個索引最多可儲存 20 億個向量，每個向量儲存貯體最多可支援 10,000 個索引
- Sub-100 毫秒的查詢延遲，針對長期儲存和不常存取模式進行最佳化
- 與專用向量資料庫相比，向量操作的成本降低高達 90%
- 具有自動擴展和 99.99999999% (11 個 9s) 耐用性的無伺服器架構

在下列情況中使用 Amazon S3 向量：

- 對於需要以最低成本儲存數十億個向量的應用程式
- 對於容忍低於秒查詢延遲 (100 毫秒或以上) 而非低於 10 毫秒的工作負載
- 對於長期向量保留和封存使用案例
- 對於不常擷取模式的 RAG 應用程式
- 對於將儲存經濟性優先於超低延遲的組織

Amazon S3 Vectors 原生與 Amazon Bedrock 知識庫整合，並與 Amazon OpenSearch Service 在分層架構中運作良好。您可以使用 Amazon S3 Vectors 進行冷儲存，並使用 OpenSearch Service 進行熱查詢。

如需詳細資訊，請參閱 Amazon [S3 文件中的使用 S3 向量和向量儲存貯體](#)。Amazon S3

受管服務選項

Amazon Bedrock 知識庫代表向量資料庫實作的 AWS 全受管方法。此服務在儲存選項的彈性，結合其自動化管理功能，對於尋求實作 RAG 而管理複雜基礎設施的組織來說特別有價值。

透過 Amazon Bedrock 知識庫，您可以建立、維護和查詢使用 RAG 增強基礎模型的知識庫。此服務可透過管理整個資料擷取、向量化和擷取管道，簡化實作 RAG 的複雜程序。

Amazon Bedrock 知識庫的主要優點包括：

- 簡化的資料處理
 - 自動資料擷取和區塊
 - 從多個檔案格式擷取內建文字
 - 受管向量內嵌產生
 - 自動中繼資料擷取和索引
- 簡化 RAG 實作
 - 預先設定的擷取策略
 - 自動內容視窗最佳化
 - 內建關聯性調校
 - 現成的語意搜尋功能
- 安全與管控
 - Integrated AWS Identity and Access Management (IAM) 控制項
 - 靜態和傳輸中的資料加密
 - VPC 支援
 - 使用 稽核記錄 AWS CloudTrail

Amazon Bedrock 知識庫支援多個[向量存放區選項](#)，包括：

- Amazon Aurora PostgreSQL 搭配 pgvector
- Amazon Neptune Analytics
- Amazon EMR Serverless
- Amazon S3 Vectors
- Pinecone
- Redis Enterprise Cloud

此受管服務會處理自動化資料擷取、向量化和擷取。這可簡化 RAG 實作。

如需每個支援的向量存放區的詳細資訊，請參閱 [Amazon Bedrock 知識庫文件](#)。

選擇正確的向量資料庫

根據這些關鍵決策因素選取向量資料庫：

- 如果您需要具有向量搜尋的 MongoDB 相容文件資料庫 – 選擇 Amazon DocumentDB。當您的應用程式使用 MongoDB APIs，而且您想要新增語意搜尋功能而不管管理個別向量基礎設施時，這是理想的選擇。
- 如果您需要超低延遲的即時應用程式 – 選擇 Amazon MemoryDB。這可在 AWS 具有低於毫秒回應時間的 上提供最快的向量搜尋效能。它非常適合即時推薦引擎和高輸送量應用程式。
- 如果您需要以圖形為基礎的知識表示法搭配向量搜尋 – 選擇 Amazon Neptune Analytics。這最適合需要周遊複雜關係和執行圖形型查詢以及向量搜尋的 GraphRAG 應用程式，以提供可解釋的 AI 回應。
- 如果您需要結合關聯式查詢與向量搜尋 – 選擇 Amazon Aurora PostgreSQL 與 pgvector。當您的應用程式同時需要相同資料庫中的傳統 SQL 操作和向量相似性搜尋時，此選項是理想的選擇。
- 如果您需要延遲低於 10 毫秒的高輸送量查詢 – 選擇 Amazon OpenSearch Service。它擅長處理高頻率查詢和即時應用程式，並包括最近的 GPU 加速改進。
- 如果您需要以符合成本效益的方式存放數十億個向量 – 選擇 Amazon S3 向量。此選項可節省高達 90% 的成本，非常適合不常擷取模式（查詢之間的分鐘到小時）可容忍低於 100 毫秒延遲的應用程式。
- 如果您需要搭配向量搜尋進行全文搜尋 – 選擇 Amazon OpenSearch Service。此選項在單一平台中結合了強大的全文搜尋功能與向量搜尋。

向量資料庫比較

AWS 提供實作向量搜尋功能的多種方法，從個別向量資料庫到 Amazon Bedrock 知識庫，這是全受管服務。評估這些選項時，組織必須考慮各種層面，包括架構、可擴展性、整合功能、效能特性和安全功能。

個別向量資料庫

下表提供數個 AWS 個別向量資料庫解決方案的主要功能概觀，著重於其架構、擴展功能、資料來源整合和效能特性。

功能	Amazon Kendra	Amazon OpenSearch Service	Amazon RDS for PostgreSQL with pgvector	Amazon DocumentDB	Amazon MemoryDB	Amazon Neptune Analytics	Amazon S3 Vectors
主要使用案例	企業搜尋和 RAG	分散式搜尋和分析	支援向量的關係式資料庫	具有向量搜尋的文件資料庫	即時記憶體內向量搜尋	使用向量搜尋進行圖形分析	成本最佳化向量儲存
Architecture	全受管	分散式叢集	關聯式資料庫	文件導向	記憶體資料庫	圖形分析引擎	無伺服器物件儲存
資料模型	以文件為基礎的	JSON 文件	關聯式資料表	JSON 文件	使用 JSON 的鍵值	屬性圖表	物件儲存體
向量維度	自動管理	高達 16,000	可設定	最多 2,000 個 (索引) ; 16,000 個 (未索引)	高達 32,768	可設定	高達 4,096

索引方法	自動	HNSW、IVFHNSW、IVFFHNSW、IVFFHNSW lat lat				原生圖形 和向量	自動
距離指標	自動	Cosine、Euclidean、點產品	Cosine、Euclidean、內部產品	Cosine、Euclidean、點產品	Cosine、Euclidean、內部產品	Cosine、Euclidean	Cosine、Euclidean
查詢延遲	次秒	Sub-10 毫秒 (GPU 加速)	10-100 毫秒	毫秒	亞毫秒	次秒	Sub-100 毫秒
擴展模型	自動	水平 (新增節點)	垂直和僅供讀取複本	水平 (新增執行個體)	垂直和複本	自動	自動 (無伺服器)
向量上限	受管	十億元 (取決於叢集)	百萬 (執行個體相依)	每個集合的百萬	每個資料庫的百萬	十億	每個索引 20 億 ; 每個儲存貯體 10,000 個索引
輸送量	高	極高 (數千個 QPS)	中	高	非常高 (每天數百萬個請求)	高	中 (針對不常查詢進行最佳化)
資料耐用性	99.999999 999% (11 9s)	可使用複本設定	99.99% (異地同步備份)	99.99% (異地同步備份)	99.99% (異地同步備份)	99.99%	99.999999 999% (11 9s)
一致性模型	最終	最終 (可設定)	強 (ACID)	最終	強大	強大	強大

其他功能	40 個或更多資料連接器，NLP	全文搜尋、分析、儀表板	SQL 查詢、ACID 交易	MongoDB API 相容性	Redis API 相容性、快取	圖形演算法、周遊	Amazon S3 整合、生命週期政策
定價方式	按查詢和儲存付費	執行個體時數和儲存	執行個體時數和儲存	執行個體時數和儲存	執行個體時數和儲存	容量單位和儲存體	儲存、查詢和資料傳輸
成本最佳化	以用量為基礎的	預留執行個體、自動擴展	預留執行個體，Aurora Serverless	預留執行個體	預留執行個體	自動調整規模	與專用 DBs 相比，節省高達 90%
最適合	以最少的設定進行企業搜尋	高輸送量、低延遲查詢	混合 SQL 和向量工作負載	需要向量的 MongoDB 相容應用程式	即時、極低延遲的應用程式	GraphRAG 和知識圖表	長期、經濟實惠的儲存體
理想的查詢模式	頻繁的企業搜尋	高頻率即時查詢	混合 SQL 和向量查詢	使用語意搜尋記錄查詢	每天數百萬個請求	使用向量搜尋繪製周遊圖形	不常查詢（分鐘到小時）
設定複雜性	低（完全受管）	中（叢集組態）	中（延伸設定）	中（叢集組態）	中（叢集組態）	低（完全受管）	低（無伺服器）
所需的團隊專業知識	極小	OpenSearch 或 Elasticsearch	PostgreSQL、SQL	MongoDB	Redis	圖形資料庫	Amazon S3，基本向量概念

受管服務 – Amazon Bedrock 知識庫

Amazon Bedrock 知識庫提供具有多個向量儲存選項的全受管解決方案。下表會比較這些儲存選項。

功能	Aurora PostgreSQL with pgvector	Neptune 分析	OpenSearch Service Serverless	Amazon S3 向量	Pinecone	RedisEnterprise 雲端
主要使用案例	具有向量 RAG 的關係式資料庫	GraphRAG 的圖形型向量搜尋	知識管理 RAG	成本最佳化向量 RAG	高效能向量搜尋	記憶體內向量搜尋
Architecture	全受管關聯式	全受管圖形分析	全受管無伺服器	無伺服器物件儲存	全受管混合雲端	完全受管的記憶體內
資料模型	關聯式資料表	屬性圖表	JSON 文件	物件儲存體	專用向量	具有向量的鍵值
向量儲存	透過 pgvector 延伸模組	原生圖形向量	透過 OpenSearch 引擎	原生 Amazon S3 向量儲存	原生向量資料庫	記憶體內向量
Amazon Bedrock 整合	原生	原生	原生	原生	原生	原生
自動擷取	是 (透過 Amazon Bedrock)	是 (透過 Amazon Bedrock)	是 (透過 Amazon Bedrock)	是 (透過 Amazon Bedrock)	是 (透過 Amazon Bedrock)	是 (透過 Amazon Bedrock)
自動向量化	是 (透過 Amazon Bedrock)	是 (透過 Amazon Bedrock)	是 (透過 Amazon Bedrock)	是 (透過 Amazon Bedrock)	是 (透過 Amazon Bedrock)	是 (透過 Amazon Bedrock)
擴展	自動擴展 (Aurora Serverless)	自動圖形擴展	自動無伺服器	自動 (十億個向量)	自動擴展 Pod	自動擴展叢集
查詢效能	關聯式或向量高	圖形向量高	高	中等 (100 毫秒或更多延遲)	非常高	非常高

向量上限	百萬（執行個體相依）	十億	十億	每個索引 20 億個	十億	百萬（記憶體相依）
其他功能	SQL 查詢、ACID 交易	圖形演算法、周遊	全文搜尋、分析	Amazon S3 生命週期、分層	中繼資料篩選、命名空間	Redis 資料結構、快取
成本最佳化	中等 (Aurora Serverless)	中等（容量單位）	高（無伺服器、pay-per-use）	非常高 （節省高達 90%）	中等（以 Pod 為基礎的定價）	低（記憶體內溢價）
最適合	混合 SQL/ 向量工作負載	連線的知識圖表	具有向量搜尋的全文	長期、不常存取的向量	大規模即時向量搜尋	超低延遲需求
理想的查詢模式	混合 SQL 和向量查詢	使用向量繪製周遊	經常使用分析進行搜尋	不常擷取（分鐘到小時）	高頻率即時查詢	每秒數百萬個請求
使用 Amazon Bedrock 設定	簡單（由 Amazon Bedrock 管理）	簡單（由 Amazon Bedrock 管理）	簡單（由 Amazon Bedrock 管理）	簡單（由 Amazon Bedrock 管理）	簡單（由 Amazon Bedrock 管理）	簡單（由 Amazon Bedrock 管理）
資料落地	AWS 區域	AWS 區域	AWS 區域	AWS 區域	多雲端 (AWS 及其他)	多雲端 (AWS 及其他)
定價方式	執行個體時數和儲存體	容量單位和儲存體	運算和儲存（無伺服器）	儲存、查詢和傳輸	Pod 時數和儲存	節點時數和儲存

在個別選項和受管選項之間進行選擇

考量事項	選擇個別向量資料庫	選擇 Amazon Bedrock 知識庫 （受管）
------	-----------	-------------------------------

RAG 實作	您想要完全控制 RAG 管道	您想要設定最少的全受管 RAG
自訂	您需要自訂擷取邏輯和預先處理	標準 RAG 模式符合您的需求
現有基礎設施	您已部署資料庫	您開始全新或想要簡化的管理
團隊專業知識	您的團隊具備資料庫管理專業知識	您偏好專注於應用程式邏輯，而不是基礎設施
整合複雜性	您需要與現有系統深度整合	您想要快速整合 Amazon Bedrock 模型
營運開銷	您可以管理資料庫操作	AWS 您想要處理 操作
成本結構	您偏好直接資料庫定價	您偏好統一的 Amazon Bedrock 定價
上市時間	您有時間進行自訂實作	您需要快速部署

成本比較和考量事項

了解不同向量資料庫選項的成本結構對於做出明智的實作決策至關重要。下表概述各種向量資料庫解決方案的一些關鍵成本考量，包括個別資料庫和受管服務。每個選項都有不同的定價因素，可能影響總體擁有成本 (TCO)，從pay-as-you-go模型到基礎設施和營運成本。

向量資料庫	成本模型	成本考量
Amazon Kendra	根據查詢，隨需付費	成本可能會根據查詢數量和資料索引數量而有所不同。資料儲存和資料傳輸可能會收取額外費用。如需詳細資訊，請參閱 Amazon Kendra 定價 。
Amazon OpenSearch Service	根據執行個體時數和儲存體，隨需付費	成本包括執行個體時數、儲存 (Amazon EBS 磁碟區)、資料傳輸和選用的 UltraWarm 儲存。預留執行個體最多可節省 30%。GPU 加速執行個體可為向量工作負載提供更好的價格效能。如需詳細資訊，請參閱 Amazon OpenSearch Service 定價 。
開放原始碼 OpenSearch	開放原始碼，無需直接成本 (您不需要付費即可下載或使用軟體，也無需任何授權成本)	成本包括基礎設施 (例如伺服器和儲存) 和營運成本 (例如維護和監控)。組織需要為人員編列預算，才能管理和維護基礎設施。
Amazon RDS for PostgreSQL 搭配 pgvector	根據用量，隨需付費	成本包括資料庫執行個體類型、儲存體、資料傳輸和備份。資料傳輸、執行個體類型和超出 AWS 免費方案的儲存體可能需支付額外費用。如需詳細資訊，請參閱 Amazon RDS 定價 。

Amazon DocumentDB	根據執行個體時數和儲存體， 隨需付費	成本包括執行個體時數、儲存 (GB 月)、I/O 請求、備份儲存和資料傳輸。彈性叢集可啟用動態擴展。可用於成本最佳化的預留執行個體。如需詳細資訊，請參閱 Amazon DocumentDB 定價 。
Amazon MemoryDB	根據節點時數和資料儲存， 隨需付費	成本包括節點時數（每個節點類型）、資料儲存 (GB 小時)、快照儲存和資料傳輸。預留節點最多可節省 55%。針對高輸送量、低延遲工作負載進行最佳化。如需詳細資訊，請參閱 Amazon MemoryDB 定價 。
Amazon Neptune Analytics	根據容量單位，隨需付費	成本包括 Neptune 容量單位 (NCUs)、儲存 (GB 月) 和資料傳輸。根據沒有預先承諾的工作負載自動擴展。至少需要 128 NCUs。如需詳細資訊，請參閱 Amazon Neptune 定價 。
Amazon S3 Vectors	根據儲存體和請求，隨需付費	成本包括儲存 (GB 月)、PUT 和 GET 請求、向量索引管理和資料傳輸。與專用向量資料庫相比，可節省高達 90% 的成本。Amazon S3 Intelligent-Tiering 和生命週期政策可用於其他最佳化。如需詳細資訊，請參閱 Simple Storage Service (Amazon S3) 定價 。

Amazon Bedrock 知識庫

根據用量，隨需付費

成本可能會因知識庫的使用量和 Amazon OpenSearch Serverless 等其他服務而有所不同。資料儲存、資料傳輸和其他功能可能需支付額外費用。如需定價的詳細資訊，請參閱 [Amazon OpenSearch Service 定價](#)。

向量資料庫使用案例

下列範例強調如何有效地使用不同的向量資料庫選項來增強知識管理、提高營運效率，以及提供更好的業務成果。這些使用案例說明本指南稍早討論的向量資料庫解決方案的實際應用程式，並提供其實際效能和優點的洞見。

Amazon Kendra 的知識管理

客戶問題 – 日本最大的一般承包商之一面臨經驗豐富的人員下降。公司需要一種方法來有效地將體驗人員的知識和技能轉移到較年輕的世代。他們需要解決方案來擷取和傳播複雜的建構工程知識和過去的經驗。

AWS 解決方案 – 為了解決此問題，客戶轉向 Amazon Kendra，這是一種 AI 解決方案，可以快速準確地處理其內部知識庫並允許自然語言查詢。使用 Amazon Kendra，員工現在可以更快地找到他們所需的資訊，提高生產力並促進從經驗豐富的人員到較年幼員工的知識轉移。

影響 – 透過實作由 Amazon Kendra 提供支援的生成式 AI 聊天機器人，該公司建立了統一的知識平台。聊天機器人可讓員工快速存取技術知識和過去的建構工程經驗。此解決方案已大幅提升組織內知識轉移和決策程序的效率，有助於確保所有員工都能輕鬆獲得寶貴的專業知識。此解決方案的成本可能會因您的用量和組態而有所不同。如需詳細成本估算，請參閱 [AWS 定價計算工具](#)。如需向量資料庫成本估算，請參閱本指南的[成本比較和考量](#)一節，或參閱 [Amazon Kendra 定價](#)。

如需其他客戶使用案例的資訊，請參閱 [Amazon Kendra 客戶](#)。

使用 OpenSearch Serverless 進行即時分析

客戶問題 – 一家領先的金融服務供應商面臨管理巨型資料生態系統的挑戰。它每年處理 3 億個授權和 900 億筆交易，累積到大約 11 PB (PB) 的資料。現有系統為需要存取超過 6,000 個報告的 300,000 名使用者提供服務，需要現代化以提供全球一致性並啟用即時決策。

AWS 解決方案 – 解決方案架構使用透過 Amazon Bedrock (包括 Anthropic、Sonnet 3、Sonnet 3.5 和 Haiku) 提供的基礎模型進行自然語言處理。客戶選擇 OpenSearch Serverless 作為向量資料庫，因其卓越的可擴展性和有效處理大量資料量的能力。此架構可無縫處理複雜的查詢和產生動態報告。

影響 – 實作透過消除手動產生超過 100 個商業智慧儀表板的需求，將生產力提高 50%。使用者現在可以透過回應時間介於 20-40 秒之間的自然語言查詢產生報告。此解決方案的成本可能會因您的用量和組態而有所不同。如需詳細成本估算，請參閱 [AWS 定價計算工具](#)。如需向量資料庫成本估算，請參閱

本指南的[成本比較和考量](#)一節，或參閱 [Amazon OpenSearch Service 定價](#)。如需其他客戶使用案例的資訊，請參閱 [Amazon OpenSearch Serverless](#)。

後續步驟和資源

檢閱本指南後，請考慮下列動作，以便從理解移至實作：

1. 評估您目前的需求：
 - 評估您現有的資料庫基礎設施和專業知識。
 - 記錄您的特定向量搜尋需求。
 - 定義您的效能、擴展和成本目標。
2. 選擇下列其中一個選項來測試向量資料庫選項：
 - 選項 1：使用您偏好的向量資料庫解決方案設定概念驗證。
 - 選項 2：在 Amazon Bedrock 知識庫中實驗範例資料集。嘗試 Amazon Bedrock 知識庫的快速建立體驗。如需範例，請參閱 [Aurora 文件中的快速建立 Amazon Bedrock 的 Aurora PostgreSQL 知識庫](#)。
3. 檢閱其他[資源](#)。
4. 取得專家協助：
 - 如需實作指引，請聯絡您的 AWS 帳戶 團隊或 AWS 解決方案架構師。
 - 與專精於向量資料庫的[AWS 合作夥伴互動](#)。
5. 規劃您的生產部署：
 - 如果從現有的資料庫移動，請建立遷移策略。
 - 為您選擇的解決方案制定擴展計劃。
 - 設計您的監控和維護程序。

Resources

下列資源可協助您選擇向量資料庫。

AWS 部落格文章

- [使用 Amazon Bedrock 知識庫快速建立和 Amazon Aurora Serverless 加速生成式 AI 應用程式開發](#)
- [Amazon OpenSearch Service 的向量資料庫功能說明](#)
- [使用 Amazon Bedrock 知識庫深入探索向量資料存放區](#)
- [利用 pgvector 和 Amazon Aurora PostgreSQL 進行自然語言處理、聊天機器人和情緒分析](#)

AWS 服務文件

- [選擇 AWS 資料庫服務](#)
- [Amazon Bedrock 知識庫的運作方式](#)
- [Neptune Analytics 文件](#)
- [Amazon Web Services 概觀：資料庫](#)
- [使用 Aurora PostgreSQL 做為 Amazon Bedrock 的知識庫](#)
- [使用 Amazon Aurora PostgreSQL](#)
- [Amazon DocumentDB](#)
- [Amazon MemoryDB](#)
- [Amazon S3 Vectors](#)

其他 AWS 資源

- [Amazon Bedrock 知識庫](#)
- [向量資料庫和內嵌](#)
- [生成式 AI 應用程式的向量資料庫](#)
- [什麼是Machine Learning中的內嵌？](#)

其他資源

- [關於 PostgreSQL](#)
- [pgvector 文件](#)
- [Pinecone 作為 Amazon Bedrock 的知識庫](#)
- [Redis Enterprise Cloud on AWS](#)

文件歷史紀錄

下表說明本指南的重大變更，選擇 RAG 使用案例的 AWS 向量資料庫。如果您想收到有關未來更新的通知，可以訂閱 [RSS 摘要](#)。

變更	描述	日期
已新增 AWS 服務	我們新增了有關使用 Amazon DocumentDB、Amazon MemoryDB、Amazon S3 Vectors 和 Amazon Neptune Analytics 的資訊。	2026 年 3 月 30 日
初次出版	—	2025 年 3 月 6 日

本文為英文版的機器翻譯版本，如內容有任何歧義或不一致之處，概以英文版為準。