在 上擷取增強產生選項和架構 AWS

AWS 方案指引

AWS 方案指引: 在 上擷取增強產生選項和架構 AWS

Table of Contents

簡介 1
目標對象 1
目標 1
生成式 AI 選項 2
了解 RAG 3
元件 5
比較 RAG 和微調 6
RAG 的使用案例 7
全受管 RAG 選項 8
Amazon Bedrock 的知識庫 8
資料來源 10
向量資料庫 11
Amazon Q Business 12
主要功能 12
最終使用者自訂 13
Amazon SageMaker AI Canvas 13
自訂 RAG 架構 16
檢索器 16
Amazon Kendra 17
Amazon OpenSearch Service 18
Amazon Aurora PostgreSQL 和 pgvector 19
Amazon Neptune Analytics 19
Amazon MemoryDB 20
Amazon DocumentDB 21
Pinecone 22
MongoDB Atlas 24
Weaviate 24
產生器 25
Amazon Bedrock 25
SageMaker AI JumpStart 25
選擇 RAG 選項 27
結論 29
文件歷史紀錄 30
詞彙表 31

..... 31
A 31
B 34
C 35
D 38
E 41
F 43
G 44
H 45
I 46
L 48
M 49
O 53
P 55
Q 57
R 57
S 60
T 63
U 64
V 64
W 65
Z 66
..... lxvii

在 上擷取增強產生選項和架構 AWS

Mithil Shah、Rajeev Muralidhar 和 Natacha Fort，Amazon Web Services

2024 年 10 月 (文件歷史記錄)

生成式 AI 是指 AI 模型的子集，可以從簡單的文字提示建立新的內容和成品，例如影像、影片、文字和音訊。生成式 AI 模型會根據包含各種主題和任務的大量資料進行訓練。這讓他們能夠在執行各種任務時展現顯著的多樣性，即使是尚未明確訓練的任務也一樣。由於單一模型能夠執行多個任務，這些模型通常稱為基礎模型 (FMs)。

生成式 AI 模型的其中一個值得注意的應用程式是其回答問題的熟練度。不過，當這些模型用於根據自訂文件回答問題時，會發生特定挑戰。自訂文件可以包含專屬資訊、內部網站、內部文件、Confluence頁面、SharePoint頁面等。其中一個選項是使用擷取增強產生 (RAG)。使用 RAG，基礎模型會參考訓練資料來源 （例如您的自訂文件） 以外的授權資料來源，再產生回應。

本指南說明可從自訂文件回答問題的不同生成式 AI 選項，包括擷取增強生成 (RAG) 系統。它還提供在 Amazon Web Services () 上建置 RAG 系統的概觀AWS。透過檢閱 RAG 選項和架構，您可以在 AWS 和自訂 RAG 架構上選擇全受管服務。

目標對象

本指南的目標受眾是想要建置 RAG 解決方案、檢閱可用架構，以及了解每個選項優點和缺點的生成式 AI 架構師和經理。

目標

本指南可協助您執行以下操作：

- 了解可用於回答自訂文件中問題的生成式 AI 選項
- 在 上檢閱 RAG 系統的架構選項 AWS
- 了解每個 RAG 選項的優點和缺點
- 為您的 AWS 環境選擇 RAG 架構

用於查詢自訂文件的生成式 AI 選項

組織通常具有各種結構化和非結構化資料來源。本指南著重於如何使用生成式 AI 來回答非結構化資料中的問題。

組織中的非結構化資料可能來自各種來源。這些可能是 PDFs、文字檔案、內部 Wiki、技術文件、面向公眾的網站、知識庫或其他。如果您希望基礎模型可以回答有關非結構化資料的問題，可使用下列選項：

- 使用您的自訂文件和其他訓練資料來訓練新的基礎模型
- 使用自訂文件中的資料微調現有的基礎模型
- 當您提出問題時，使用內容內學習將文件傳遞至基礎模型
- 使用擷取增強產生 (RAG) 方法

從頭開始訓練包含自訂資料的新基礎模型是一項有野心的任務。少數幾家公司已成功完成，例如Bloomberg使用其BloombergGPT模型。另一個範例是 的多模態EXAONE模型LG AI Research，其訓練方式是使用 6，000 億件美工和 2.5 億個高解析度影像，並附有文字。根據 AI 成本：您應該建置或購買您的基礎模型 (LinkedIn)，此模型類似於要訓練MetaLlama 2的 48 0 萬美金成本。從頭開始訓練模型有兩個主要先決條件：存取資源（財務、技術、時間）和明確的投資回報。如果這看起來不適合，則下一個選項是微調現有的基礎模型。

微調現有模型需要採用模型，例如 Amazon Titan、Mistral 或 Llama 模型，然後根據自訂資料調整模型。微調有多種技術，其中大部分只涉及修改幾個參數，而不是修改模型中的所有參數。這稱為參數效率微調。進行微調的主要方法有兩種：

- 監督式微調會使用標記的資料，並協助您訓練新任務類型的模型。例如，如果您想要根據 PDF 表單產生報告，則您可能需要提供足夠的範例來教導模型如何執行此操作。
- 非監督式微調與任務無關，可根據您自己的資料調整基礎模型。它會訓練模型以了解文件的內容。然後，微調的模型會使用更自訂您組織的樣式來建立內容，例如報告。

不過，微調可能不適用於問答使用案例。如需詳細資訊，請參閱本指南中的比較 RAG 和微調。

當您提出問題時，您可以傳遞文件基礎模型，並使用模型的內容內學習來傳回文件的答案。此選項適用於單一文件的臨機操作查詢。不過，此解決方案不適用於查詢多個文件或查詢系統和應用程式，例如 Microsoft SharePoint 或 Atlassian Confluence。

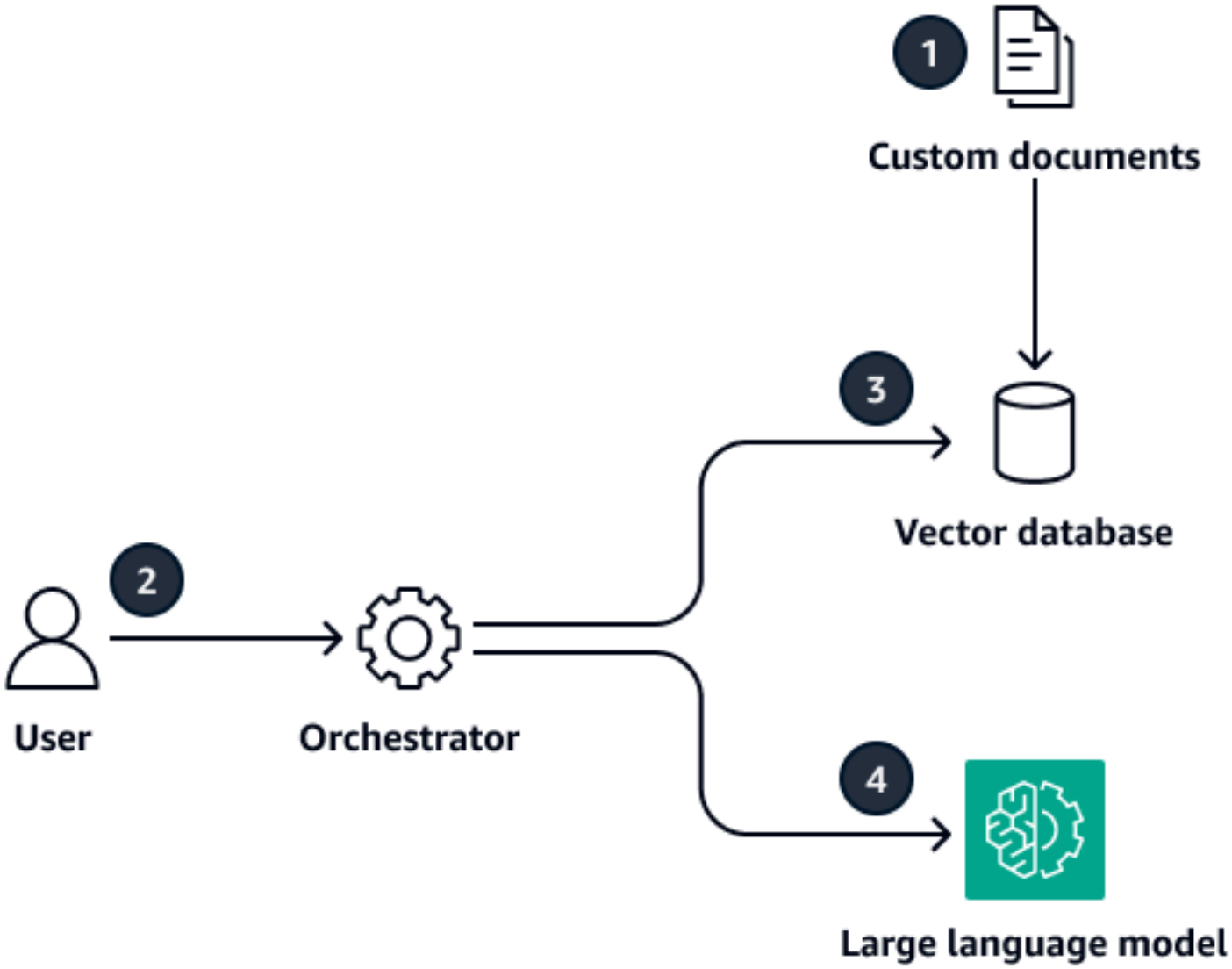
最後一個選項是使用 RAG。透過 RAG，基礎模型會在產生回應之前參考您的自訂文件。RAG 會將模型的功能擴展到組織的內部知識庫，而不需要重新訓練模型。這是一種經濟實惠的方法，可改善模型輸出，以便在各種環境中保持相關性、準確性和實用性。

本節主題：

- 了解擷取增強生成
- 比較擷取增強生成和微調
- 擷取增強生成的使用案例

了解擷取增強生成

擷取增強生成 (RAG) 是一種技術，用於使用外部資料增強大型語言模型 (LLM)，例如公司的內部文件。這可提供模型所需的內容，以針對特定使用案例產生準確且實用的輸出。RAG 是一種在企業中使用 LLMs 的實用且有效方法。下圖顯示 RAG 方法運作方式的高階概觀。



大致而言，RAG 程序是四個步驟。第一個步驟會完成一次，其他三個步驟會視需要執行任意次數：

1. 您可以建立內嵌，將內部文件擷取至向量資料庫。內嵌是文件中文字的數值表示法，可擷取資料的語意或內容意義。向量資料庫基本上是這些內嵌項目的資料庫，有時稱為向量存放區或向量索引。此步驟需要資料清理、格式化和區塊化，但這是一次性的前期活動。
2. 人類以自然語言提交查詢。
3. 協調器會在向量資料庫中執行相似性搜尋，並擷取相關資料。協調器會將擷取的資料 （也稱為內容) 新增至包含查詢的提示。
4. 協調器會將查詢和內容傳送至 LLM。LLM 會使用其他內容產生查詢的回應。

從使用者的角度來看，RAG 看起來像與任何 LLM 互動。不過，系統更了解有問題的內容，並提供根據組織的知識庫進行微調的答案。

如需 RAG 方法運作方式的詳細資訊，請參閱網站上的 AWS 什麼是 RAG。

生產層級 RAG 系統的元件

建置生產層級的 RAG 系統需要考慮 RAG 工作流程的幾個不同層面。概念上，生產層級的 RAG 工作流程需要下列功能和元件，無論特定實作為何：

- 連接器 — 這些會將不同的企業資料來源與向量資料庫連線。結構化資料來源的範例包括交易和分析資料庫。非結構化資料來源的範例包括物件存放區、程式碼基底和軟體即服務 (SaaS) 平台。每個資料來源可能需要不同的連線模式、授權和組態。
- 資料處理 — 資料有許多形狀和形式，例如 PDFs、掃描的影像、文件、簡報和Microsoft SharePoint檔案。您必須使用資料處理技術來擷取、處理和準備要編製索引的資料。
- 內嵌 - 若要執行關聯式搜尋，您必須將文件和使用者查詢轉換為相容的格式。透過使用內嵌語言模型，您可以將文件轉換為數值表示法。這些基本上是基礎基礎模型的輸入。
- 向量資料庫 — 向量資料庫是內嵌項目、相關文字和中繼資料的索引。索引已針對搜尋和擷取進行最佳化。
- 擷取器 — 對於使用者查詢，擷取器會從向量資料庫中擷取相關內容，並根據業務需求對回應進行排名。
- 基礎模型 — RAG 系統的基礎模型通常是 LLM。透過處理內容和提示，基礎模型會為使用者產生並格式化回應。
- 護欄 — 護欄旨在確保查詢、提示、擷取的內容和 LLM 回應準確、負責任、符合道德，且無幻覺和偏差。
- 協調程式 — 協調程式負責排程和管理end-to-end工作流程。
- 使用者體驗 — 通常，使用者與具有豐富功能的對話聊天界面互動，包括顯示聊天歷史記錄和收集使用者對回應的意見回饋。
- 身分和使用者管理 — 以精細程度控制使用者對應用程式的存取至關重要。在 中 AWS 雲端，政策、角色和許可通常是透過 AWS Identity and Access Management (IAM) 管理。

顯然，規劃、開發、發行和管理 RAG 系統的工作量很大。Amazon Bedrock 或 Amazon Q Business 等全受管服務可協助您管理一些未區分的繁重工作。不過，自訂 RAG 架構可以提供更多對元件的控制，例如擷取器或向量資料庫。

比較擷取增強生成和微調

下表說明微調和以 RAG 為基礎的方法的優點和缺點。

方法	優點	缺點
微調	• 如果使用非監督式方法訓練微調後的模型，則能夠建立更符合您組織風格的內容。 • 根據專屬或法規資料訓練的微調模型，可協助您的組織遵循內部或產業特定的資料和合規標準。	• 微調可能需要幾小時到幾天的時間，取決於模型的大小。因此，如果您的自訂文件經常變更，這不是理想的解決方案。 • 微調需要了解技術，例如低階適應 (LoRA) 和參數效率微調 (PEFT)。微調可能需要資料科學家。 • 微調可能不適用於所有模型。 • 微調的模型不會在其回應中提供來源的參考。 • 使用微調的模型回答問題時，可能會增加幻覺的風險。
RAG	• RAG 可讓您為自訂文件建置問答系統，而無需微調。 • RAG 可以在幾分鐘內納入最新的文件。 • AWS 提供全受管 RAG 解決方案。因此，不需要資料科學家或機器學習的專業知識。 • 在其回應中，RAG 模型會提供資訊來源的參考。 • 由於 RAG 使用向量搜尋的內容作為其產生答案的基	• 從整個文件摘要資訊時，RAG 無法正常運作。

方法	優點	缺點
	礎，因此幻覺的風險會降低。	

如果您需要建立參考自訂文件的問答解決方案，建議您從以 RAG 為基礎的方法開始。如果您需要模型執行其他任務，例如摘要，請使用微調。

您可以在單一模型中結合微調和 RAG 方法。在這種情況下，RAG 架構不會變更，但產生答案的 LLM 也會使用自訂文件進行微調。這結合了兩個領域的優點，而且可能是適合您使用案例的最佳解決方案。如需如何將監督式微調與 RAG 結合的詳細資訊，請參閱《》中的 RAFT：調整語言模型與網域特定 RAG 研究University of California, Berkeley。

擷取增強生成的使用案例

以下是使用 RAG 方法的常見使用案例：

- 搜尋引擎 – 啟用 RAG 的搜尋引擎可以在其搜尋結果中提供更準確和up-to-date特色程式碼片段。
- 問答系統 – RAG 可以改善問答系統中的回應品質。擷取型模型使用相似性搜尋來尋找包含答案的相關段落或文件。然後，它會根據該資訊產生簡潔且相關的回應。
- 零售或電子商務 – RAG 可以透過提供更相關且個人化的產品建議來增強電子商務中的使用者體驗。透過擷取和整合使用者偏好設定和產品詳細資訊的相關資訊，RAG 可以為客戶產生更準確且實用的建議。
- 工業或製造 – 在製造中，RAG 可協助您快速存取重要資訊，例如工廠工廠營運。它也可以協助決策過程、故障診斷和組織創新。對於在嚴格的法規架構內操作的製造商，RAG 可以快速從內部和外部來源擷取更新的法規和合規標準，例如從產業標準或監管機構。
- 醫療保健 – RAG 在醫療保健產業具有潛力，在醫療保健產業中存取準確且及時的資訊至關重要。透過從外部來源擷取和整合相關的醫學知識，RAG 可以在醫療保健應用程式中提供更準確且內容感知的回應。這類應用程式可擴增人類臨床醫生可存取的資訊，該臨床醫生最終會進行呼叫，而不是模型。
- 法律 – RAG 可以在合併和收購等法律案例中強而有力地套用，其中複雜的法律文件提供查詢的內容。這可協助法律專業人員快速解決複雜的法規問題。

上的全受管擷取增強產生選項 AWS

若要管理 上的擷取增強產生 (RAG) 工作流程 AWS，您可以使用自訂 RAG 管道，或使用 AWS 提供的一些全受管服務功能。由於它們包含以 RAG 為基礎的系統的許多核心元件，全受管服務可協助您管理一些未區分的繁重工作。不過，這些服務提供較少的自訂機會。

全受管 AWS 服務 使用連接器從外部資料來源擷取資料，例如網站、Atlassian Confluence 或 Microsoft SharePoint。支援的資料來源因 而異 AWS 服務。

本節探討下列在 上建置 RAG 工作流程的全受管選項 AWS：

- Amazon Bedrock 的知識庫
- Amazon Q Business
- Amazon SageMaker AI Canvas

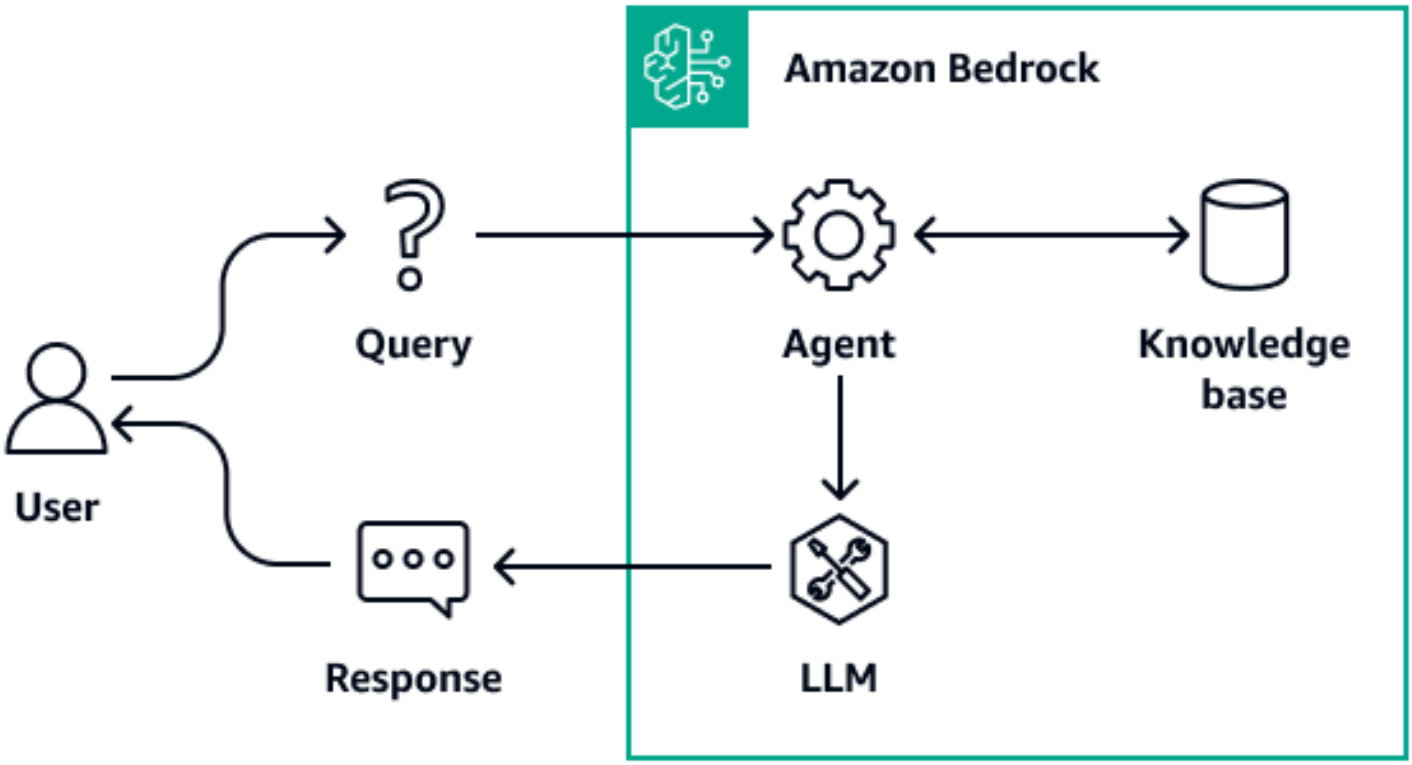
如需如何在這些選項之間進行選擇的詳細資訊，請參閱本指南在 上選擇擷取增強產生選項 AWS中的 。

Amazon Bedrock 的知識庫

Amazon Bedrock 是一項全受管服務，可讓您透過統一 API 使用來自領導級 AI 新創公司和 Amazon 的高效能基礎模型 (FMs)。知識庫是一種 Amazon Bedrock 功能，可協助您實作從擷取到擷取和提示擴增的整個 RAG 工作流程。您不需要建立與資料來源的自訂整合或管理資料流程。工作階段內容管理是內建的，因此您的生成式 AI 應用程式可以輕鬆支援多轉對話。

在您指定資料的位置後，Amazon Bedrock 的知識庫會在內部擷取文件、將其區塊化為文字區塊、將文字轉換為內嵌，然後將內嵌內容存放在您選擇的向量資料庫中。Amazon Bedrock 會管理和更新內嵌項目，使向量資料庫與資料保持同步。如需知識庫如何運作的詳細資訊，請參閱 Amazon Bedrock 知識庫如何運作。

如果您將知識庫新增至 Amazon Bedrock 代理程式，代理程式會根據使用者輸入來識別適當的知識庫。代理程式會擷取相關資訊，並將資訊新增至輸入提示。更新的提示會提供模型更多內容資訊來產生回應。為了提高透明度並將幻覺降至最低，從知識庫擷取的資訊可追蹤至其來源。



Amazon Bedrock 支援下列兩個適用於 RAG APIs：

- RetrieveAndGenerate – 您可以使用此 API 查詢您的知識庫，並從其擷取的資訊產生回應。在內部，Amazon Bedrock 會將查詢轉換為內嵌、查詢知識庫、使用搜尋結果增強提示做為內容資訊，並傳回 LLM 產生的回應。Amazon Bedrock 也會管理對話的短期記憶體，以提供更具體的結果。
- 擷取 – 您可以使用此 API，透過直接從知識庫擷取的資訊來查詢知識庫。您可以使用從此 API 傳回的資訊來處理擷取的文字、評估其相關性，或開發個別的工作流程來產生回應。在內部，Amazon Bedrock 會將查詢轉換為內嵌、搜尋知識庫，並傳回相關結果。您可以在搜尋結果之上建置其他工作流程。例如，您可以使用 LangChain `AmazonKnowledgeBasesRetriever` 外掛程式將 RAG 工作流程整合到生成式 AI 應用程式。

如需使用 APIs 的範例架構模式和step-by-step說明，請參閱 Amazon Bedrock 中的知識庫現在提供全受管 RAG 體驗 (AWS 部落格文章）。如需如何使用 `RetrieveAndGenerate` API 為智慧型聊天型應用程式建置 RAG 工作流程的詳細資訊，請參閱使用 Amazon Bedrock 知識庫建置情境式聊天機器人應用程式 (AWS 部落格文章）。

知識庫的資料來源

您可以將專屬資料連接到知識庫。設定資料來源連接器之後，您可以將資料與知識庫同步或保持最新狀態，並讓資料可供查詢。Amazon Bedrock 知識庫支援連線至下列資料來源：

- Amazon Simple Storage Service (Amazon S3) – 您可以使用 主控台或 API，將 Amazon S3 儲存貯體連線至 Amazon Bedrock 知識庫。知識庫會擷取和索引儲存貯體中的檔案。這種資料來源類型支援下列功能：
  - 文件中繼資料欄位 – 您可以包含個別的檔案，以指定 Amazon S3 儲存貯體中檔案的中繼資料。然後，您可以使用這些中繼資料欄位來篩選和改善回應的相關性。
  - 包含或排除篩選條件 – 您可以在爬取時包含或排除特定內容。
  - 增量同步 – 會追蹤內容變更，而且只會爬取自上次同步以來變更的內容。
- Confluence – 您可以使用 主控台或 API 將Atlassian Confluence執行個體連線至 Amazon Bedrock 知識庫。這種資料來源類型支援下列功能：
  - 自動偵測主要文件欄位 – 會自動偵測和編目中繼資料欄位。您可以使用這些欄位進行篩選。
  - 包含或排除內容篩選條件 – 您可以在空格、頁面標題、部落格標題、註解、附件名稱或延伸上使用字首或規則表達式模式，來包含或排除特定內容。
  - 累加式同步 - 追蹤內容變更，而且只會爬取自上次同步以來變更的內容。
  - OAuth 2.0 身分驗證，使用 Confluence API 字符進行身分驗證 – 身分驗證憑證會存放在其中 AWS Secrets Manager。
- Microsoft SharePoint – 您可以使用 主控台或 API 將SharePoint執行個體連線至知識庫。這種資料來源類型支援下列功能：
  - 自動偵測主要文件欄位 – 會自動偵測和爬取中繼資料欄位。您可以使用這些欄位進行篩選。
  - 包含或排除內容篩選條件 – 您可以在主頁面標題、事件名稱和檔案名稱（包括其副檔名）上使用字首或規則表達式模式來包含或排除特定內容。
  - 增量同步 - 追蹤內容變更，並且只會爬取自上次同步以來變更的內容。
  - OAuth 2.0 身分驗證 – 身分驗證憑證存放在其中 AWS Secrets Manager。
- Salesforce – 您可以使用 主控台或 API 將Salesforce執行個體連線至知識庫。這種資料來源類型支援下列功能：
  - 自動偵測主要文件欄位 – 會自動偵測和編目中繼資料欄位。您可以使用這些欄位進行篩選。
  - 包含或排除內容篩選條件 – 您可以使用字首或規則表達式模式來包含或排除特定內容。如需可套用篩選條件的內容類型清單，請參閱 Amazon Bedrock 文件中的包含/排除篩選條件。
  - 增量同步 – 會追蹤內容變更，而且只會爬取自上次同步以來變更的內容。

  - OAuth 2.0 身分驗證 – 身分驗證憑證存放在其中 AWS Secrets Manager。
- Web 爬蟲程式 – Amazon Bedrock Web 爬蟲程式會連接至您提供的 URLs並進行爬蟲。支援下列功能：
  - 選取要爬取URLs
  - 遵守標準 robots.txt 指令，例如 `Allow`和 `Disallow`
  - 排除符合模式URLs
  - 限制爬取的速率
  - 在 Amazon CloudWatch 中，檢視每個抓取 URL 的狀態

如需可連線至 Amazon Bedrock 知識庫之資料來源的詳細資訊，請參閱為您的知識庫建立資料來源連接器。

知識庫的向量資料庫

當您設定知識庫與資料來源之間的連線時，您必須設定向量資料庫，也稱為向量存放區。向量資料庫是 Amazon Bedrock 存放、更新和管理代表您資料的內嵌項目的地方。每個資料來源都支援不同類型的向量資料庫。若要判斷您的資料來源可使用哪些向量資料庫，請參閱資料來源類型。

如果您希望 Amazon Bedrock 自動為您在 Amazon OpenSearch Serverless 中建立向量資料庫，您可以在建立知識庫時選擇此選項。不過，您也可以選擇設定自己的向量資料庫。如果您設定自己的向量資料庫，請參閱您自己的向量存放區的先決條件以取得知識庫。每種類型的向量資料庫都有自己的先決條件。

根據您的資料來源類型，Amazon Bedrock 知識庫支援下列向量資料庫：

- Amazon OpenSearch Serverless
- Amazon Aurora PostgreSQL-Compatible Edition
- Pinecone (Pinecone 文件)
- Redis Enterprise Cloud (Redis 文件)
- MongoDB Atlas (MongoDB 文件)

Amazon Q Business

Amazon Q Business 是全受管、生成式 AI 支援的助理，您可以設定 來回答問題、提供摘要、產生內容，以及根據您的企業資料完成任務。它允許最終使用者從具有引文的企業資料來源接收立即的許可感知回應。

主要功能

Amazon Q Business 的下列功能可協助您建置生產級 RAG 型生成式 AI 應用程式：

- 內建連接器 – Amazon Q Business 支援超過 40 種類型的連接器，例如 Adobe Experience Manager (AEM)、Jira、 Salesforce和 的連接器Microsoft SharePoint。如需完整清單，請參閱支援的連接器。如果您需要不支援的連接器，您可以使用 Amazon AppFlow 將資料來源中的資料提取至 Amazon Simple Storage Service (Amazon S3)，然後將 Amazon Q Business 連線至 Amazon S3 儲存貯體。如需 Amazon AppFlow 支援的資料來源完整清單，請參閱支援的應用程式。
- 內建索引管道 – Amazon Q Business 提供內建管道，用於編製向量資料庫中資料的索引。您可以使用 AWS Lambda 函數為索引管道新增預先處理邏輯。
- 索引選項 – 您可以在 Amazon Q Business 中建立和佈建原生索引，並使用 Amazon Q Business 擷取器從該索引提取資料。或者，您可以使用預先設定的 Amazon Kendra 索引做為擷取器。如需詳細資訊，請參閱為 Amazon Q Business 應用程式建立擷取器。
- 基礎模型 – Amazon Q Business 使用 Amazon Bedrock 支援的基礎模型。如需完整清單，請參閱 Amazon Bedrock 中支援的基礎模型。
- 外掛程式 – Amazon Q Business 提供使用外掛程式與目標系統整合的功能，例如在 中摘要票證資訊和建立票證的自動化方式Jira。設定完成後，外掛程式可以支援讀取和寫入動作，協助您提高最終使用者的生產力。Amazon Q Business 支援兩種類型的外掛程式：內建外掛程式和自訂外掛程式。
- 護欄 – Amazon Q Business 支援全域控制和主題層級控制。例如，這些控制項可以在提示中偵測個人身分識別資訊 (PII)、濫用或敏感資訊。如需詳細資訊，請參閱 Amazon Q Business 中的管理員控制項和護欄。
- 身分管理 – 透過 Amazon Q Business，您可以管理使用者及其對 RAG 型生成式 AI 應用程式的存取。如需詳細資訊，請參閱 Amazon Q Business 的身分和存取管理。此外，Amazon Q Business 連接器索引存取控制清單 (ACL) 資訊會與文件本身一起連接至文件。然後，Amazon Q Business 會將索引的 ACL 資訊存放在 Amazon Q Business User Store 中，以建立使用者和群組映射，並根據最終使用者對文件的存取篩選聊天回應。如需詳細資訊，請參閱資料來源連接器概念。
- 文件擴充 – 文件擴充功能可協助您控制哪些文件和文件屬性會擷取到您的索引，以及如何擷取它們。這可以透過兩種方法完成：

- 設定基本操作 – 使用基本操作來新增、更新或刪除資料中的文件屬性。例如，您可以選擇刪除與 PII 相關的任何文件屬性，以清除 PII 資料。
- 設定 Lambda 函數 – 使用預先設定的 Lambda 函數，對資料執行更自訂的進階文件屬性操作邏輯。例如，您企業的部分資料可能以掃描影像形式儲存。在這種情況下，您可以使用 Lambda 函數在掃描的文件上執行光學字元辨識 (OCR)，從中擷取文字。後續在資料匯入期間，每份掃描文件都會視為文字文件處理。最後，在聊天期間，Amazon Q 會在產生回應時考量從掃描文件擷取的文字資料。

實作解決方案時，您可以選擇結合兩種文件擴充方法。您可以使用基本操作來執行資料的第一個剖析，然後使用 Lambda 函數進行更複雜的操作。如需詳細資訊，請參閱 Amazon Q Business 中的文件擴充。

- 整合 – 建立 Amazon Q Business 應用程式後，您可以將其整合到其他應用程式，例如 Slack或 Microsoft Teams。例如，請參閱部署 forAmazonSlack閘道和部署 Amazon Q Business 的Microsoft Teams閘道 (AWS 部落格文章）。

最終使用者自訂

Amazon Q Business 支援上傳可能不會存放在組織資料來源和索引中的文件。上傳的文件不會儲存。它們僅適用於上傳文件的對話。Amazon Q Business 支援用於上傳的特定文件類型。如需詳細資訊，請參閱在 Amazon Q Business 中上傳檔案和聊天。

Amazon Q Business 包含依文件屬性功能的篩選。管理員和最終使用者可以使用此功能。管理員可以使用 屬性自訂和控制最終使用者的聊天回應。例如，如果資料來源類型是連接至文件的屬性，您可以指定僅從特定資料來源產生聊天回應。或者，您可以使用您選取的屬性篩選條件，允許最終使用者限制聊天回應的範圍。

最終使用者可以在更廣泛的 Amazon Q Business 應用程式環境中建立輕量型專用 Amazon Q 應用程式。Amazon Q 應用程式允許特定網域的任務自動化，例如專為行銷團隊打造的應用程式。

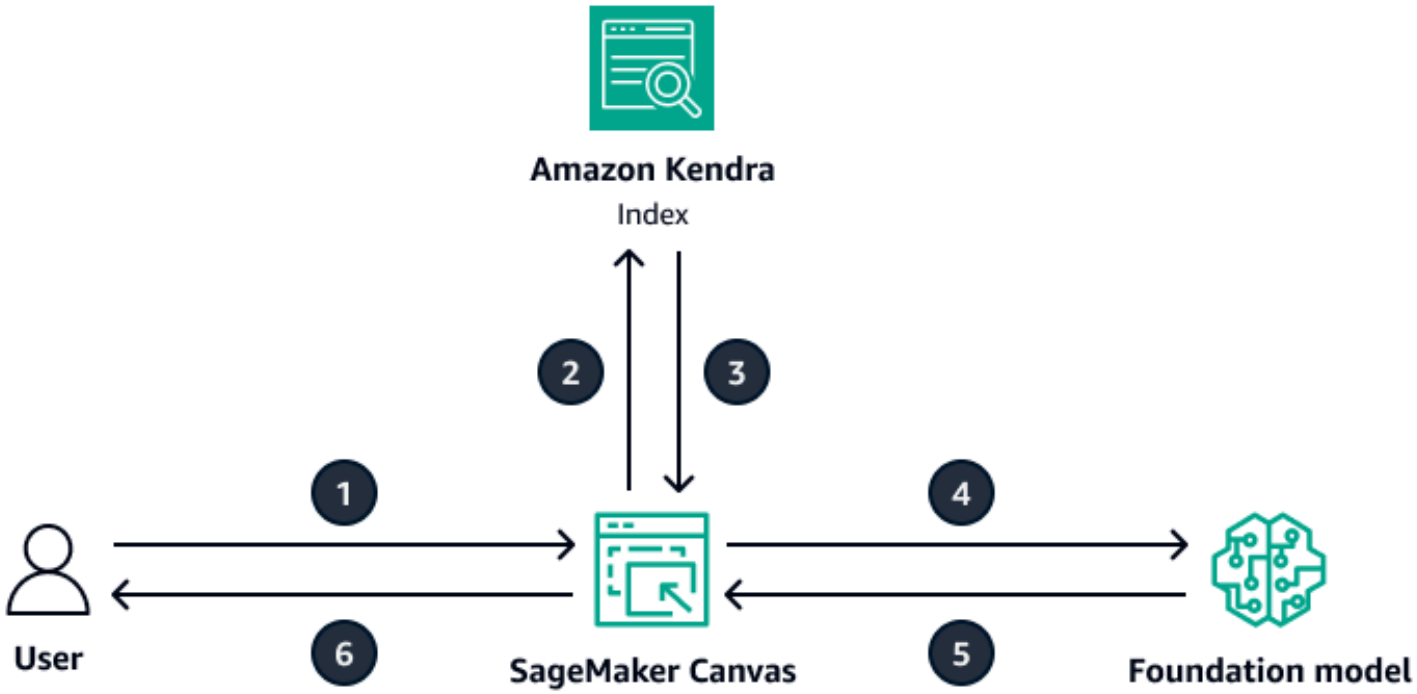
Amazon SageMaker AI Canvas

Amazon SageMaker AI Canvas 可協助您使用機器學習產生預測，而無需撰寫任何程式碼。它提供無程式碼視覺化界面，可讓您準備資料、建置和部署 ML 模型，簡化統一環境中end-to-end ML 生命週期。資料準備、模型開發、偏差偵測、可解釋性和監控的複雜性會在直覺式界面後方抽象化。使用者不需要是 SageMaker AI 或機器學習操作 (MLOps) 專家，即可使用 SageMaker AI Canvas 開發、操作和監控模型。

使用 SageMaker AI Canvas 時，RAG 功能是透過無程式碼的文件查詢功能提供。您可以使用 Amazon Kendra 索引作為基礎企業搜尋，豐富 SageMaker AI Canvas 中的聊天體驗。如需詳細資訊，請參閱使用文件查詢從文件擷取資訊。

將 SageMaker AI Canvas 連接到 Amazon Kendra 索引需要一次性設定。做為網域組態的一部分，雲端管理員可以選擇一或多個 Kendra 索引，供使用者在與 SageMaker Canvas 互動時查詢。如需如何啟用文件查詢功能的指示，請參閱開始使用 Amazon SageMaker AI Canvas。

SageMaker AI Canvas 會管理 Amazon Kendra 與所選基礎模型之間的基礎通訊。如需有關 SageMaker AI Canvas 支援的基礎模型的詳細資訊，請參閱 SageMaker AI Canvas 中的生成式 AI 基礎模型。下圖顯示雲端管理員將 SageMaker AI Canvas 連線至 Amazon Kendra 索引後，文件查詢功能的運作方式。



該圖顯示以下工作流程：

1. 使用者在 SageMaker AI Canvas 中開始新的聊天，開啟查詢文件，選取目標索引，然後提交問題。
2. SageMaker AI Canvas 使用查詢來搜尋 Amazon Kendra 索引以取得相關資料。
3. SageMaker AI Canvas 會從 Amazon Kendra 索引擷取資料及其來源。
4. SageMaker AI Canvas 會更新提示，以包含從 Amazon Kendra 索引擷取的內容，並將提示提交至基礎模型。
5. 基礎模型使用原始問題和擷取的內容來產生答案。

6. SageMaker AI Canvas 為使用者提供產生的答案。它包含對資料來源的參考，例如用來產生回應的文件。

上的自訂擷取增強產生架構 AWS

上一節說明如何使用完全受管 AWS 服務 的擷取增強生成 (RAG)。不過，某些使用案例需要對系統元件進行更多控制，例如擷取器或 LLM （ 也稱為產生器)。例如，您可能需要彈性來選擇自己的向量資料庫或存取不支援的資料來源。對於這些使用案例，您可以建置自訂 RAG 架構。

本節包含下列主題：

- RAG 工作流程的擷取器
- RAG 工作流程的產生器

如需本節中如何選擇擷取器和產生器選項的詳細資訊，請參閱本指南在 上選擇擷取增強產生選項 AWS中的 。

RAG 工作流程的擷取器

本節說明如何建置擷取器。您可以使用全受管語意搜尋解決方案，例如 Amazon Kendra，也可以使用 AWS 向量資料庫建置自訂語意搜尋。

在您檢閱擷取工具選項之前，請確定您了解向量搜尋程序的三個步驟：

1. 您可以將需要編製索引的文件分隔為較小的部分。這稱為區塊。
2. 您可以使用稱為內嵌的程序，將每個區塊轉換為數學向量。然後，您可以為向量資料庫中的每個向量編製索引。您用來為文件編製索引的方法會影響搜尋的速度和準確性。索引方法取決於向量資料庫及其提供的組態選項。
3. 您可以使用相同的程序，將使用者查詢轉換為向量。擷取器會搜尋向量資料庫，尋找與使用者查詢向量類似的向量。透過使用歐幾里得距離、餘弦距離或點積等指標來計算相似性。

本指南說明如何使用下列 AWS 服務 或第三方服務在 上建置自訂擷取層 AWS：

- Amazon Kendra
- Amazon OpenSearch Service
- Amazon Aurora PostgreSQL 和 pgvector
- Amazon Neptune Analytics
- Amazon MemoryDB

- Amazon DocumentDB
- Pinecone
- MongoDB Atlas
- Weaviate

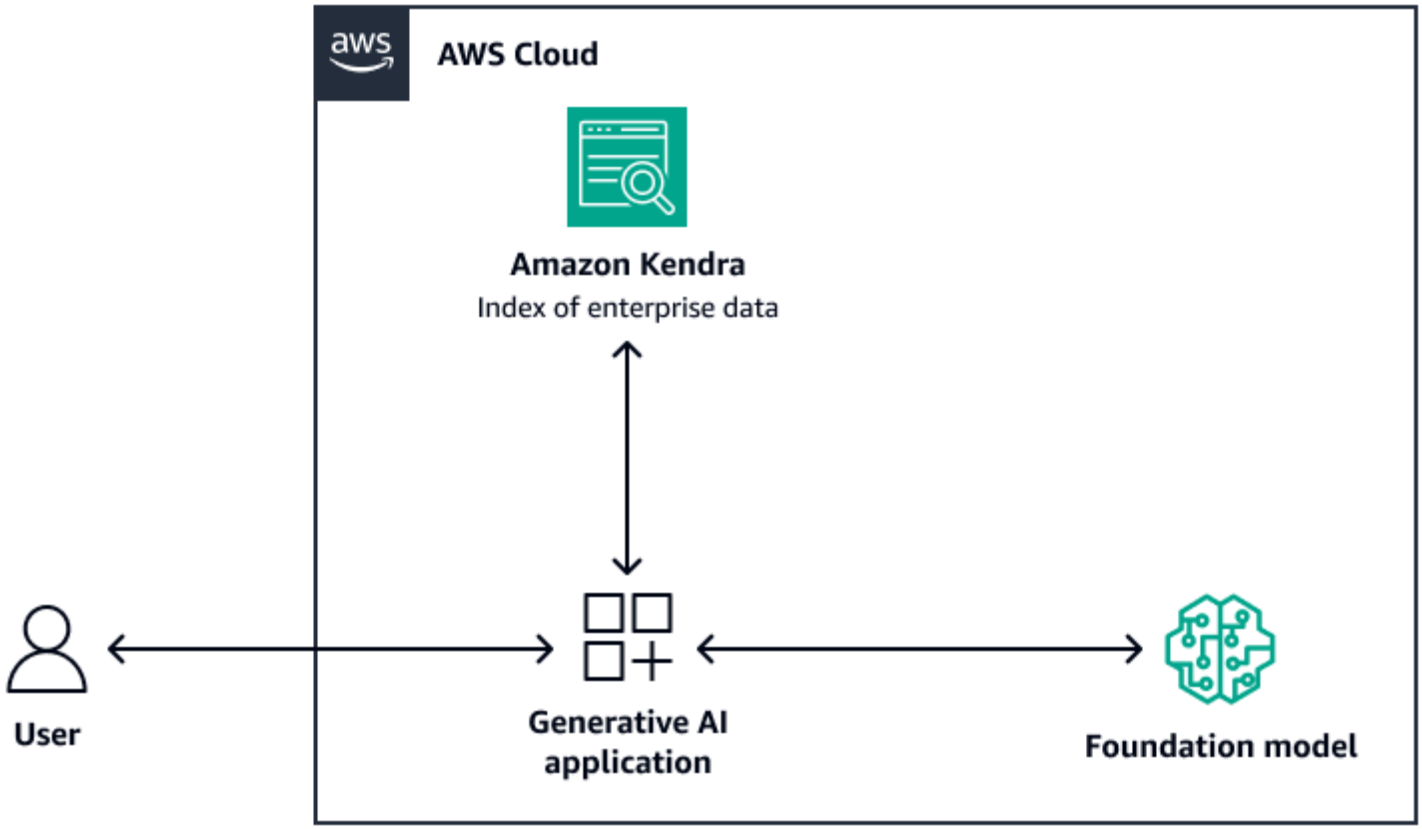
Amazon Kendra

Amazon Kendra 是一種全受管智慧型搜尋服務，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 可協助您直接從多個來源擷取文件，並在文件成功同步後查詢文件。同步程序會建立必要的基礎設施，以在擷取的文件上建立向量搜尋。因此，Amazon Kendra 不需要向量搜尋程序的傳統三個步驟。初始同步後，您可以使用定義的排程來處理持續的擷取。

以下是使用 Amazon Kendra for RAG 的優勢：

- 您不需要維護向量資料庫，因為 Amazon Kendra 會處理整個向量搜尋程序。
- Amazon Kendra 包含適用於熱門資料來源的預先建置連接器，例如資料庫、網站爬蟲程式、Amazon S3 儲存貯體、Microsoft SharePoint執行個體和Atlassian Confluence執行個體。 AWS 合作夥伴開發的連接器可供使用，例如 Box和 的連接器GitLab。
- Amazon Kendra 提供存取控制清單 (ACL) 篩選，只會傳回最終使用者可存取的文件。
- Amazon Kendra 可以根據中繼資料提升回應，例如日期或來源儲存庫。

下圖顯示使用 Amazon Kendra 做為 RAG 系統擷取層的範例架構。如需詳細資訊，請參閱使用 Amazon KendraLangChain、 和大型語言模型，在企業資料上快速建置高準確性的生成式 AI 應用程式 (AWS 部落格文章）。



對於基礎模型，您可以使用 Amazon Bedrock 或透過 Amazon SageMaker AI JumpStart 部署的 LLM。您可以使用 AWS Lambda 搭配 LangChain 來協調使用者、Amazon Kendra 和 LLM 之間的流程。若要建置使用 Amazon Kendra、 LangChain和各種 LLMs RAG 系統，請參閱 Amazon Kendra LangChain Extensions GitHub 儲存庫。

Amazon OpenSearch Service

Amazon OpenSearch Service 為 k 近鄰 (k-NN) 搜尋提供內建 ML 演算法，以執行向量搜尋。OpenSearch Service 也為 Amazon EMR Serverless 提供向量引擎。您可以使用此向量引擎來建置具有可擴展性和高效能向量儲存和搜尋功能的 RAG 系統。如需如何使用 OpenSearch Serverless 建置 RAG 系統的詳細資訊，請參閱使用 Amazon OpenSearch Serverless 和 Amazon Bedrock Claude 模型的向量引擎建置可擴展和無伺服器 RAG 工作流程 (AWS 部落格文章）。

以下是使用 OpenSearch Service 進行向量搜尋的優勢：

- 它提供向量資料庫的完整控制權，包括使用 OpenSearch Serverless 建置可擴展的向量搜尋。
- 它提供對區塊策略的控制。
- 它使用來自非指標空間程式庫 (NMSLIB)、Faiss 和 Apache Lucene 程式庫的近似近鄰 (ANN) 演算法來支援 k-NN 搜尋。 https://github.com/facebookresearch/faiss https://lucene.apache.org/您可以

根據使用案例變更演算法。如需透過 OpenSearch Service 自訂向量搜尋選項的詳細資訊，請參閱 Amazon OpenSearch Service 向量資料庫功能說明 (AWS 部落格文章）。

- OpenSearch Serverless 將 Amazon Bedrock 知識庫整合為向量索引。

Amazon Aurora PostgreSQL 和 pgvector

Amazon Aurora PostgreSQL 相容版本是全受管關聯式資料庫引擎，可協助您設定、操作和擴展 PostgreSQL 部署。pgvector 是開放原始碼 PostgreSQL 延伸模組，可提供向量相似性搜尋功能。此擴充功能適用於 Aurora PostgreSQL 相容和 Amazon Relational Database Service (Amazon RDS) for PostgreSQL。如需如何建置使用 Aurora PostgreSQL 相容和 pgvector 的 RAG 系統的詳細資訊，請參閱下列 AWS 部落格文章：

- 使用 Amazon SageMaker AI 和 pgvector 在 PostgreSQL 中建置 AI 支援的搜尋
- 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

以下是使用 pgvector 和 Aurora PostgreSQL 相容的優勢：

- 它支援精確且近似最近的鄰搜尋。它也支援下列相似性指標：L2 距離、內部產品和餘弦距離。
- 它支援具有平面壓縮 (IVFFlat) 和階層式可導覽小型世界 (HNSW) 索引的反轉檔案。 https://github.com/pgvector/pgvector#hnsw
- 您可以將向量搜尋與查詢結合到相同 PostgreSQL 執行個體中可用的特定網域資料。
- Aurora PostgreSQL 相容已針對 I/O 最佳化，並提供分層快取。對於超過可用執行個體記憶體的工作負載，pgvector 每秒最多可以將向量搜尋的查詢增加 8 次。

Amazon Neptune Analytics

Amazon Neptune Analytics 是用於分析的記憶體最佳化圖形資料庫引擎。它支援圖形周遊中最佳化圖形分析演算法、低延遲圖形查詢和向量搜尋功能的程式庫。它還具有內建向量相似性搜尋。它提供一個端點來建立圖形、載入資料、叫用查詢，以及執行向量相似性搜尋。如需如何建置使用 Neptune Analytics 的 RAG 型系統的詳細資訊，請參閱使用知識圖表搭配 Amazon Bedrock 和 Amazon Neptune 建置 GraphRAG 應用程式 (AWS 部落格文章）。

以下是使用 Neptune Analytics 的優點：

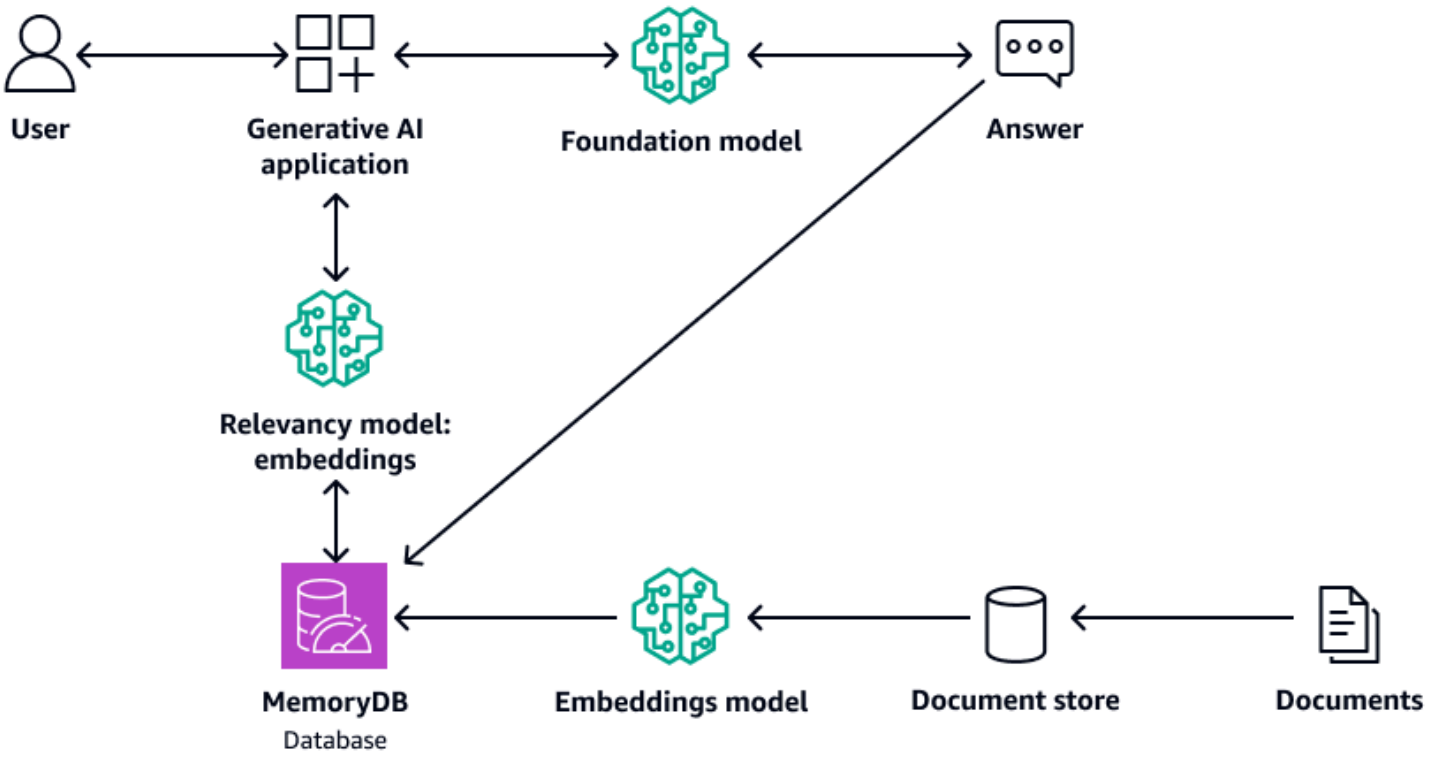
- 您可以在圖形查詢中存放和搜尋內嵌項目。
- 如果您將 Neptune Analytics 與 整合LangChain，則此架構支援自然語言圖形查詢。

- 此架構會將大型圖形資料集存放在記憶體中。

Amazon MemoryDB

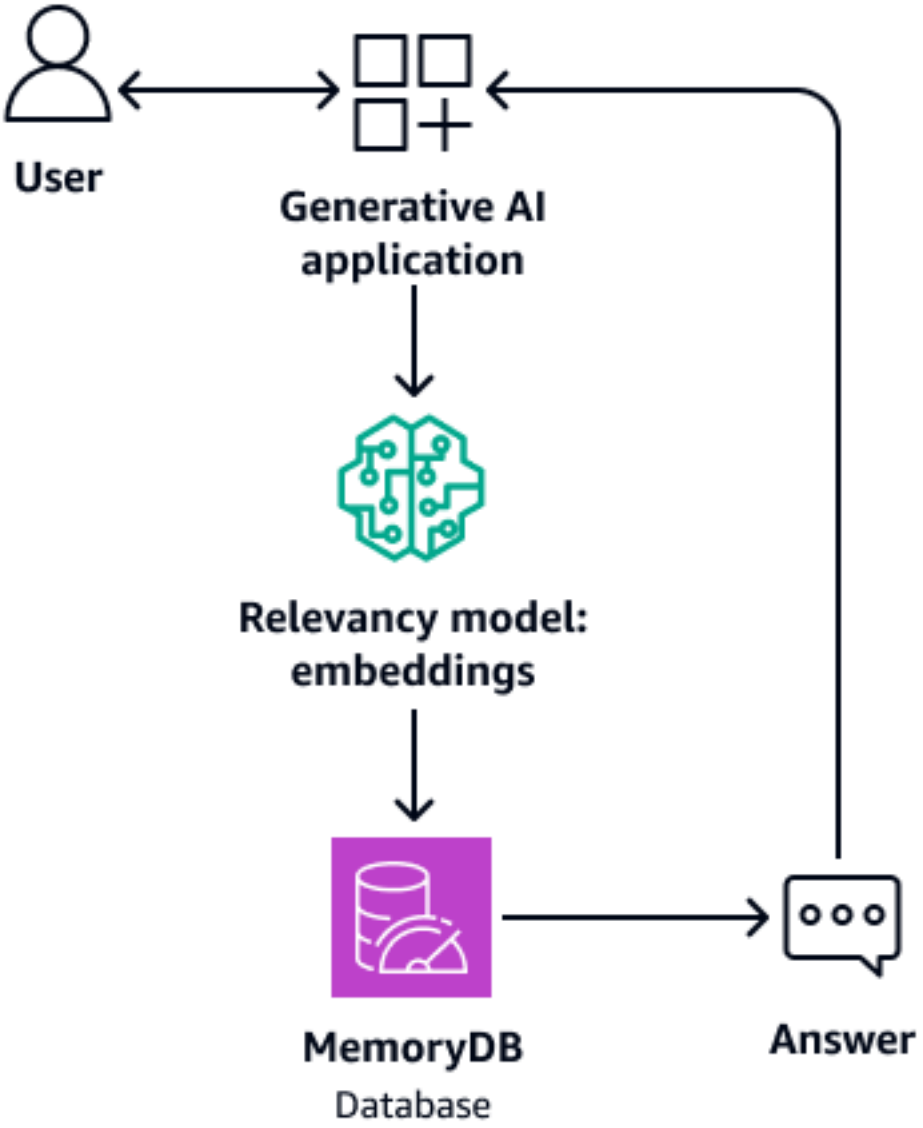
Amazon MemoryDB 是一種耐用的記憶體內資料庫服務，可提供超快速的效能。所有資料都存放在記憶體中，支援微秒讀取、單一位數毫秒寫入延遲和高輸送量。MemoryDB 的向量搜尋可擴展 MemoryDB 的功能，並可搭配現有的 MemoryDB 功能使用。如需詳細資訊，請參閱 GitHub 上的使用 LLM 和 RAG 儲存庫回答問題。

下圖顯示使用 MemoryDB 做為向量資料庫的範例架構。



以下是使用 MemoryDB 的優點：

- 它同時支援平面和 HNSW 索引演算法。如需詳細資訊，請參閱 AWS 新聞部落格上的 Amazon MemoryDB 向量搜尋現在已全面推出
- 它也可以充當基礎模型的緩衝記憶體。這表示從緩衝區擷取先前回答的問題，而不是再次進行擷取和產生程序。下圖顯示此程序。

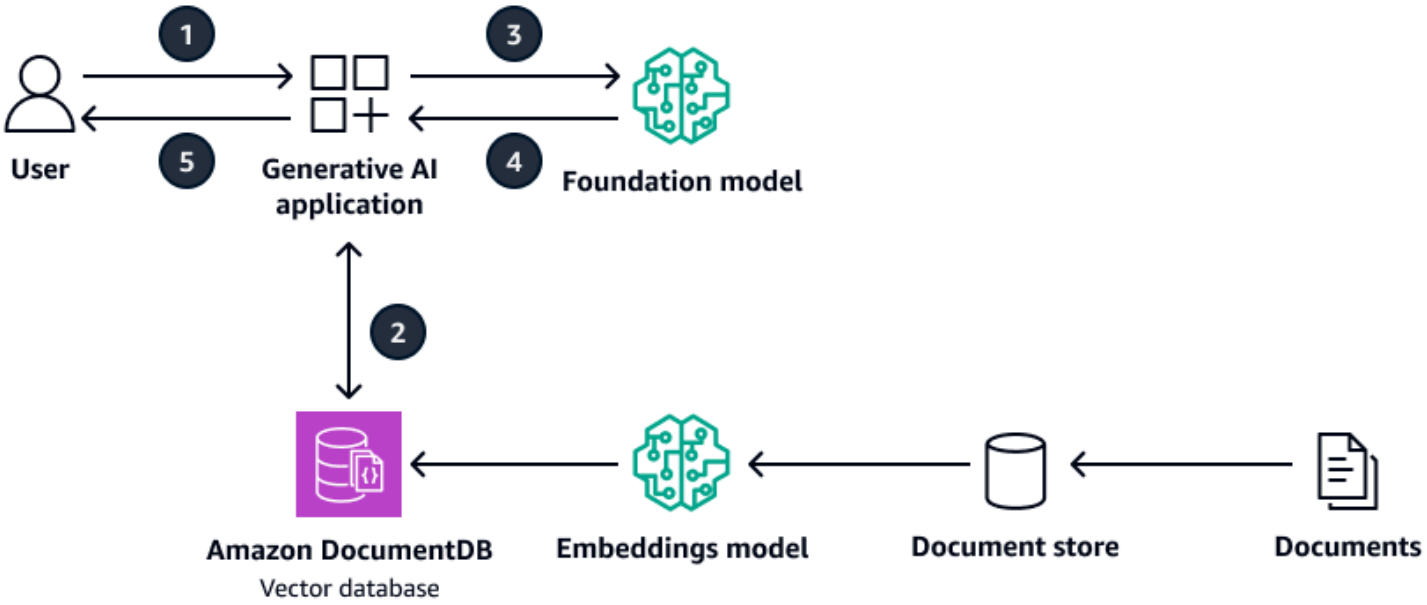


- 由於它使用記憶體內資料庫，因此此架構為語意搜尋提供單一位數毫秒的查詢時間。
- 它在 95–99% 的召回率中提供每秒最多 33，000 個查詢，以及在大於 99% 的召回率中提供每秒 26，500 個查詢。如需詳細資訊，請參閱 AWS re：Invent 2023 - 在 上搜尋 Amazon MemoryDB 影片的超低延遲向量YouTube。

Amazon DocumentDB

Amazon DocumentDB （與 MongoDB 相容） 是一種快速、可靠且全受管的資料庫服務。它可讓您輕鬆地在雲端中設定、操作和擴展MongoDB與 相容的資料庫。Amazon DocumentDB 的向量搜尋結合了 JSON 型文件資料庫的彈性和豐富的查詢功能，以及向量搜尋的強大功能。如需詳細資訊，請參閱 GitHub 上的使用 LLM 和 RAG 儲存庫回答問題。

下圖顯示使用 Amazon DocumentDB 做為向量資料庫的範例架構。



該圖顯示以下工作流程：

1. 使用者向生成式 AI 應用程式提交查詢。
2. 生成式 AI 應用程式會在 Amazon DocumentDB 向量資料庫中執行相似性搜尋，並擷取相關文件擷取。
3. 生成式 AI 應用程式會使用擷取的內容更新使用者查詢，並將提示提交至目標基礎模型。
4. 基礎模型使用內容來產生對使用者問題的回應，並傳回回應。
5. 生成式 AI 應用程式會將回應傳回給使用者。

以下是使用 Amazon DocumentDB 的優點：

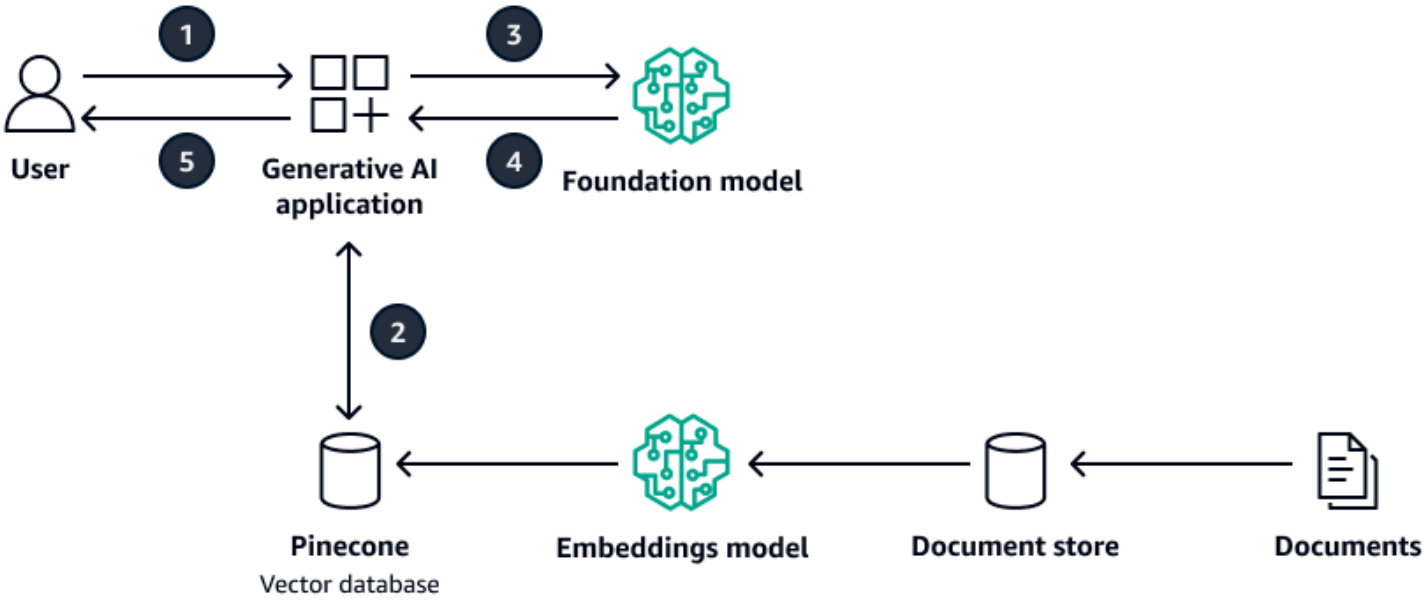
- 它同時支援 HNSW 和 IVFFlat 索引方法。
- 它在向量資料中最多支援 2，000 個維度，並支援 Euclidean、餘弦和點產品距離指標。
- 它提供毫秒的回應時間。

Pinecone

Pinecone 是全受管向量資料庫，可協助您將向量搜尋新增至生產應用程式。可透過 取得AWS Marketplace。帳單是根據用量，費用的計算方式是將 Pod 價格乘以 Pod 計數。如需如何建置使用 之 RAG 系統的詳細資訊Pinecone，請參閱下列 AWS 部落格文章：

- 使用Pinecone向量資料庫和來自 Amazon SageMaker AI JumpStart 的 Llama-2，透過 RAG 緩解幻覺
- 使用 Amazon SageMaker AI Studio 透過 Llama 2、 LangChain和 建置 RAG 問題回答解決方案 Pinecone，以進行快速實驗

下圖顯示使用 Pinecone做為向量資料庫的範例架構。



該圖顯示以下工作流程：

1. 使用者向生成式 AI 應用程式提交查詢。
2. 生成式 AI 應用程式會在Pinecone向量資料庫中執行相似性搜尋，並擷取相關文件擷取。
3. 生成式 AI 應用程式會使用擷取的內容更新使用者查詢，並將提示提交至目標基礎模型。
4. 基礎模型使用內容來產生對使用者問題的回應，並傳回回應。
5. 生成式 AI 應用程式會將回應傳回給使用者。

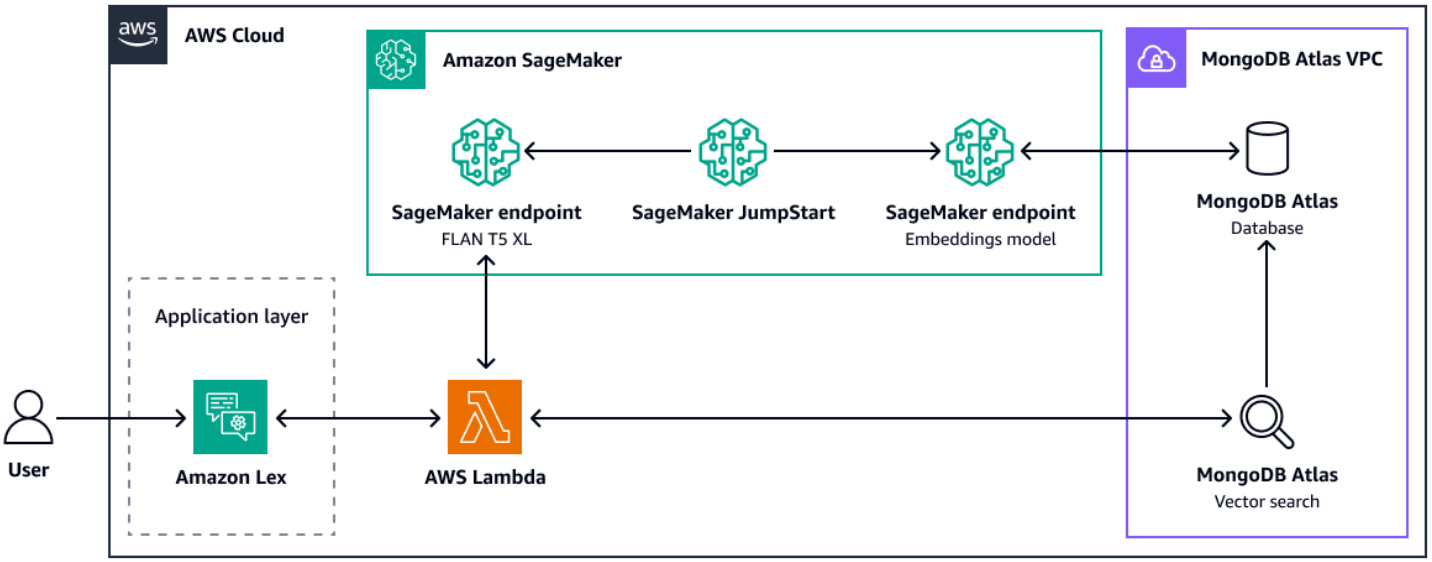
以下是使用 的優點Pinecone：

- 這是全受管向量資料庫，可減輕管理自有基礎設施的額外負荷。
- 它提供篩選、即時索引更新和關鍵字提升（混合式搜尋）的其他功能。

MongoDB Atlas

MongoDB Atlas 是一種全受管雲端資料庫，可處理部署和管理部署的所有複雜性 AWS。您可以使用的向量搜尋MongoDB Atlas，將向量內嵌儲存在MongoDB資料庫中。Amazon Bedrock 知識庫支援 MongoDB Atlas向量儲存。如需詳細資訊，請參閱 MongoDB 文件中的開始使用 Amazon Bedrock 知識庫整合。

如需如何使用MongoDB Atlas向量搜尋 RAG 的詳細資訊，請參閱使用 擷取增強生成 LangChain、Amazon SageMaker AI JumpStart 和語MongoDB Atlas意搜尋 (AWS 部落格文章）。下圖顯示本部落格文章中詳述的解決方案架構。



以下是使用MongoDB Atlas向量搜尋的優點：

- 您可以使用現有的 實作MongoDB Atlas來存放和搜尋向量內嵌。
- 您可以使用MongoDB查詢 API 來查詢向量內嵌。
- 您可以獨立擴展向量搜尋和資料庫。
- 向量內嵌存放在來源資料（文件）附近，可改善索引效能。

Weaviate

Weaviate 是熱門的開放原始碼、低延遲向量資料庫，支援多模式媒體類型，例如文字和影像。資料庫同時存放物件和向量，結合向量搜尋與結構化篩選。如需使用 Weaviate和 Amazon Bedrock 建置 RAG 工作流程的詳細資訊，請參閱 上的使用 Amazon Bedrock 中的 Cohere 基礎模型和Weaviate向量資料庫建置企業就緒生成式 AI 解決方案 AWS Marketplace (AWS 部落格文章）。

以下是使用 的優點Weaviate：

- 它是開放原始碼，並由強大的社群提供支援。
- 它專為混合式搜尋 （ 向量和關鍵字 ） 而建置。
- 您可以在 上將其部署 AWS 為受管軟體即服務 (SaaS) 方案或 Kubernetes 叢集。

RAG 工作流程的產生器

大型語言模型 LLMs) 是針對大量資料預先訓練的非常大型深度學習模型。它們非常靈活。LLMs可以執行各種任務，例如回答問題、摘要文件、翻譯語言和完成句子。他們可能會中斷內容建立，以及人們使用搜尋引擎和虛擬助理的方式。雖然不完美，但 LLMs展現了根據相對較小的提示或輸入數量進行預測的卓越能力。

LLMs是 RAG 解決方案的關鍵元件。對於自訂 RAG 架構，有兩個 AWS 服務 做為主要選項：

- Amazon Bedrock 是一項全受管服務，可讓您透過統一 API 使用來自領導 AI 公司和 Amazon 的 LLMs。
- Amazon SageMaker AI JumpStart 是一種 ML 中樞，提供基礎模型、內建演算法和預先建置的 ML 解決方案。使用 SageMaker AI JumpStart，您可以存取預先訓練的模型，包括基礎模型。您也可以使用自己的資料來微調預先訓練的模型。

Amazon Bedrock

Amazon Bedrock 提供來自 Anthropic、Stability AI、、MetaCohere、AI21 Labs、 Mistral AI和 Amazon 的業界領先模型。如需完整清單，請參閱 Amazon Bedrock 中支援的基礎模型。Amazon Bedrock 也可讓您使用自己的資料自訂模型。

您可以評估模型效能，以判斷哪些最適合您的 RAG 使用案例。您可以測試最新的模型，也可以測試哪些功能可提供最佳結果和最佳價格。Anthropic Claude Sonnet 模型是 RAG 應用程式常見的選擇，因為它擅長各種任務，並提供高度的可靠性和可預測性。

SageMaker AI JumpStart

SageMaker AI JumpStart 為各種問題類型提供預先訓練的開放原始碼模型。您可以在部署之前逐步訓練和微調這些模型。您可以透過 Amazon SageMaker AI Studio 中的 SageMaker AI JumpStart 登陸頁面或使用 SageMaker AI Python SDK，存取預先訓練的模型、解決方案範本和範例。 Amazon SageMaker

SageMaker AI JumpStart 為內容撰寫、程式碼產生、問題回答、複製、摘要、分類、資訊擷取等使用案例提供state-of-the-art基礎模型。使用 JumpStart 基礎模型建置您自己的生成式 AI 解決方案，並將自訂解決方案與其他 SageMaker AI 功能整合。如需詳細資訊，請參閱 Amazon SageMaker AI JumpStart 入門。

SageMaker AI JumpStart 加入並維護可公開取得的基礎模型，供您存取、自訂和整合到您的 ML 生命週期。如需詳細資訊，請參閱公開提供的基礎模型。SageMaker AI JumpStart 也包含第三方供應商的專屬基礎模型。如需詳細資訊，請參閱 專屬基礎模型。

在 上選擇擷取增強產生選項 AWS

本指南的全受管 RAG 選項和自訂 RAG 架構章節描述了在其中建置以 RAG 為基礎的搜尋解決方案的各種方法 AWS。本節說明如何根據您的使用案例在這些選項之間進行選擇。在某些情況下，可能會使用多個選項。在這種情況下，選擇取決於易於實作、組織中可用的技能，以及您公司的政策和標準。

我們建議您考慮以下順序中的全受管和自訂 RAG 選項，並選擇適合您使用案例的第一個選項：

1. 使用 Amazon Q Business，除非：
   - 您的 中無法使用此服務 AWS 區域，而且您的資料無法移至可使用該服務的區域
   - 您有特定原因可自訂 RAG 工作流程
   - 您想要使用現有的向量資料庫或特定的 LLM
2. 使用 Amazon Bedrock 的知識庫，除非：
   - 您有不支援的向量資料庫
   - 您有特定原因可自訂 RAG 工作流程
3. 將 Amazon Kendra 與您選擇的產生器結合，除非：
   - 您想要選擇自己的向量資料庫
   - 您想要自訂區塊化策略
4. 如果您想要對擷取器進行更多控制，並想要選取自己的向量資料庫：
   - 如果您沒有現有的向量資料庫，而且不需要低延遲或圖形查詢，請考慮使用 Amazon OpenSearch Service。
   - 如果您有現有的PostgreSQL向量資料庫，請考慮使用 Amazon Aurora PostgreSQL 和 pgvector 選項。
   - 如果您需要低延遲，請考慮記憶體內選項，例如 Amazon MemoryDB 或 Amazon DocumentDB。
   - 如果您想要將向量搜尋與圖形查詢結合，請考慮使用 Amazon Neptune Analytics。
   - 如果您已經在使用第三方向量資料庫或從中尋找特定利益，請考慮 Pinecone、 MongoDB Atlas和 Weaviate。
5. 如果您想要選擇 LLM：
   - 如果您使用 Amazon Q Business，則無法選擇 LLM。
   - 如果您使用 Amazon Bedrock，您可以選擇其中一個支援的基礎模型。
   - 如果您使用 Amazon Kendra 或自訂向量資料庫，您可以使用本指南所述的其中一個產生器，或使用自訂 LLM。

> **Note**
>
> 您也可以使用自訂文件來微調現有的 LLM，以提高其回應的準確性。如需詳細資訊，請參閱本指南中的 比較 RAG 和微調。

6. 如果您有想要使用的 Amazon SageMaker AI Canvas 現有實作，或想要比較不同 LLMs RAG 回應，請考慮 Amazon SageMaker AI Canvas。

結論

本指南說明在 上建置擷取增強產生 (RAG) 系統的各種選項 AWS。您可以從全受管服務開始，例如 Amazon Q Business 和 Amazon Bedrock 知識庫。如果您想要對 RAG 工作流程進行更多控制，您可以選擇自訂擷取器。對於產生器，您可以使用 API 在 Amazon Bedrock 中呼叫支援的 LLM，或使用 Amazon SageMaker AI JumpStart 部署您自己的 LLM。檢閱選擇 RAG 選項中的建議，以判斷哪個選項最適合您的使用案例。在您為使用案例選取最佳選項後，請使用本指南提供的參考，開始建置以 RAG 為基礎的應用程式。

文件歷史紀錄

下表描述了本指南的重大變更。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
初次出版	—	2024 年 10 月 28 日

AWS 規範性指引詞彙表

以下是 AWS Prescriptive Guidance 提供的策略、指南和模式中常用的術語。若要建議項目，請使用詞彙表末尾的提供意見回饋連結。

數字

7 R

將應用程式移至雲端的七種常見遷移策略。這些策略以 Gartner 在 2011 年確定的 5 R 為基礎，包括以下內容：

- 重構/重新架構 – 充分利用雲端原生功能來移動應用程式並修改其架構，以提高敏捷性、效能和可擴展性。這通常涉及移植作業系統和資料庫。範例：將您的現場部署 Oracle 資料庫 遷移至 Amazon Aurora PostgreSQL 相容版本。
- 平台轉換 (隨即重塑) – 將應用程式移至雲端，並引入一定程度的優化以利用雲端功能。範例：將內部部署 Oracle 資料庫 遷移至 中的 Amazon Relational Database Service (Amazon RDS) for Oracle AWS 雲端。
- 重新購買 (捨棄再購買) – 切換至不同的產品，通常從傳統授權移至 SaaS 模型。範例：將您的客戶關係管理 (CRM) 系統遷移至 Salesforce.com。
- 主機轉換 (隨即轉移) – 將應用程式移至雲端，而不進行任何變更以利用雲端功能。範例：將您的現場部署 Oracle 資料庫 遷移至 中 EC2 執行個體上的 Oracle AWS 雲端。
- 重新放置 (虛擬機器監視器等級隨即轉移) – 將基礎設施移至雲端，無需購買新硬體、重寫應用程式或修改現有操作。您可以將伺服器從內部部署平台遷移到相同平台的雲端服務。範例：將 Microsoft Hyper-V應用程式 遷移至 AWS。
- 保留 (重新檢視) – 將應用程式保留在來源環境中。其中可能包括需要重要重構的應用程式，且您希望將該工作延遲到以後，以及您想要保留的舊版應用程式，因為沒有業務理由來進行遷移。
- 淘汰 – 解除委任或移除來源環境中不再需要的應用程式。

A

ABAC

請參閱屬性型存取控制。

抽象服務

請參閱 受管服務。

ACID

請參閱原子性、一致性、隔離性、耐久性。

主動-主動式遷移

一種資料庫遷移方法，其中來源和目標資料庫保持同步 (透過使用雙向複寫工具或雙重寫入操作)，且兩個資料庫都在遷移期間處理來自連接應用程式的交易。此方法支援小型、受控制批次的遷移，而不需要一次性切換。它更靈活，但需要比主動-被動遷移更多的工作。

主動-被動式遷移

一種資料庫遷移方法，其中來源和目標資料庫會保持同步，但只有來源資料庫會在資料複寫至目標資料庫時處理來自連線應用程式的交易。目標資料庫在遷移期間不接受任何交易。

彙總函數

在一組資料列上運作的 SQL 函數，會計算群組的單一傳回值。彙總函數的範例包括 SUM和 MAX。

AI

請參閱人工智慧。

AIOps

請參閱人工智慧操作。

匿名化

在資料集中永久刪除個人資訊的程序。匿名化有助於保護個人隱私權。匿名資料不再被視為個人資料。

反模式

經常用於經常性問題的解決方案，其中解決方案具有反生產力、無效或比替代解決方案更有效。

應用程式控制

一種安全方法，僅允許使用核准的應用程式，以協助保護系統免受惡意軟體攻擊。

應用程式組合

有關組織使用的每個應用程式的詳細資訊的集合，包括建置和維護應用程式的成本及其商業價值。此資訊是產品組合探索和分析程序的關鍵，有助於識別要遷移、現代化和優化的應用程式並排定其優先順序。

人工智慧 (AI)

電腦科學領域，致力於使用運算技術來執行通常與人類相關的認知功能，例如學習、解決問題和識別模式。如需詳細資訊，請參閱什麼是人工智慧？

人工智慧操作 (AIOps)

使用機器學習技術解決操作問題、減少操作事件和人工干預以及提高服務品質的程序。如需有關如何在 AWS 遷移策略中使用 AIOps 的詳細資訊，請參閱操作整合指南。

非對稱加密

一種加密演算法，它使用一對金鑰：一個用於加密的公有金鑰和一個用於解密的私有金鑰。您可以共用公有金鑰，因為它不用於解密，但對私有金鑰存取應受到高度限制。

原子性、一致性、隔離性、耐久性 (ACID)

一組軟體屬性，即使在出現錯誤、電源故障或其他問題的情況下，也能確保資料庫的資料有效性和操作可靠性。

屬性型存取控制 (ABAC)

根據使用者屬性 (例如部門、工作職責和團隊名稱) 建立精細許可的實務。如需詳細資訊，請參閱《 AWS Identity and Access Management (IAM) 文件》中的 ABAC for AWS。

授權資料來源

您存放主要版本資料的位置，被視為最可靠的資訊來源。您可以將授權資料來源中的資料複製到其他位置，以處理或修改資料，例如匿名、修訂或假名化資料。

可用區域

中的不同位置 AWS 區域 ，可隔離其他可用區域中的故障，並提供相同區域中其他可用區域的低成本、低延遲網路連線能力。

AWS 雲端採用架構 (AWS CAF)

的指導方針和最佳實務架構 AWS ，可協助組織制定高效且有效的計劃，以成功地移至雲端。 AWS CAF 將指導方針組織到六個重點領域：業務、人員、治理、平台、安全和營運。業務、人員和控管層面著重於業務技能和程序；平台、安全和操作層面著重於技術技能和程序。例如，人員層面針對處理人力資源 (HR)、人員配備功能和人員管理的利害關係人。為此， AWS CAF 為人員開發、訓練和通訊提供指引，協助組織做好成功採用雲端的準備。如需詳細資訊，請參閱 AWS CAF 網站和 AWS CAF 白皮書。

AWS 工作負載資格架構 (AWS WQF)

一種工具，可評估資料庫遷移工作負載、建議遷移策略，並提供工作預估值。 AWS WQF 隨附於 AWS Schema Conversion Tool (AWS SCT)。它會分析資料庫結構描述和程式碼物件、應用程式程式碼、相依性和效能特性，並提供評估報告。

B

錯誤的機器人

旨在中斷或傷害個人或組織的機器人。

BCP

請參閱業務持續性規劃。

行為圖

資源行為的統一互動式檢視，以及一段時間後的互動。您可以將行為圖與 Amazon Detective 搭配使用來檢查失敗的登入嘗試、可疑的 API 呼叫和類似動作。如需詳細資訊，請參閱偵測文件中的行為圖中的資料。

大端序系統

首先儲存最高有效位元組的系統。另請參閱 Endianness。

二進制分類

預測二進制結果的過程 (兩個可能的類別之一)。例如，ML 模型可能需要預測諸如「此電子郵件是否是垃圾郵件？」等問題 或「產品是書還是汽車？」

Bloom 篩選條件

一種機率性、記憶體高效的資料結構，用於測試元素是否為集的成員。

藍/綠部署

一種部署策略，您可以在其中建立兩個不同但相同的環境。您可以在一個環境 （藍色） 中執行目前的應用程式版本，並在另一個環境 （綠色） 中執行新的應用程式版本。此策略可協助您快速復原，並將影響降至最低。

機器人

透過網際網路執行自動化任務並模擬人類活動或互動的軟體應用程式。有些機器人有用或有益，例如在網際網路上編製資訊索引的 Web 爬蟲程式。某些其他機器人稱為惡意機器人，旨在中斷或傷害個人或組織。

殭屍網路

受到惡意軟體感染且受單一方控制之機器人的網路，稱為機器人繼承器或機器人運算子。殭屍網路是擴展機器人及其影響的最佳已知機制。

分支

程式碼儲存庫包含的區域。儲存庫中建立的第一個分支是主要分支。您可以從現有分支建立新分支，然後在新分支中開發功能或修正錯誤。您建立用來建立功能的分支通常稱為功能分支。當準備好發佈功能時，可以將功能分支合併回主要分支。如需詳細資訊，請參閱關於分支 (GitHub 文件)。

碎片存取

在特殊情況下，以及透過核准的程序，讓使用者快速取得他們通常無權存取 AWS 帳戶 之 的存取權。如需詳細資訊，請參閱 Well-Architected 指南中的 AWS 實作打破玻璃程序指標。

棕地策略

環境中的現有基礎設施。對系統架構採用棕地策略時，可以根據目前系統和基礎設施的限制來設計架構。如果正在擴展現有基礎設施，則可能會混合棕地和綠地策略。

緩衝快取

儲存最常存取資料的記憶體區域。

業務能力

業務如何創造價值 (例如，銷售、客戶服務或營銷)。業務能力可驅動微服務架構和開發決策。如需詳細資訊，請參閱在 AWS上執行容器化微服務白皮書的圍繞業務能力進行組織部分。

業務連續性規劃 (BCP)

一種解決破壞性事件 (如大規模遷移) 對營運的潛在影響並使業務能夠快速恢復營運的計畫。

C

CAF

請參閱AWS 雲端採用架構。

Canary 部署

版本對最終使用者的緩慢和增量版本。當您有信心時，您可以部署新版本並完全取代目前的版本。

CCoE

請參閱 Cloud Center of Excellence。

CDC

請參閱變更資料擷取。

變更資料擷取 (CDC)

追蹤對資料來源 (例如資料庫表格) 的變更並記錄有關變更的中繼資料的程序。您可以將 CDC 用於各種用途，例如稽核或複寫目標系統中的變更以保持同步。

混沌工程

故意引入故障或破壞性事件，以測試系統的彈性。您可以使用 AWS Fault Injection Service (AWS FIS) 來執行實驗，為您的 AWS 工作負載帶來壓力，並評估其回應。

CI/CD

請參閱持續整合和持續交付。

分類

有助於產生預測的分類程序。用於分類問題的 ML 模型可預測離散值。離散值永遠彼此不同。例如，模型可能需要評估影像中是否有汽車。

用戶端加密

在目標 AWS 服務 接收資料之前，在本機加密資料。

雲端卓越中心 (CCoE)

一個多學科團隊，可推動整個組織的雲端採用工作，包括開發雲端最佳實務、調動資源、制定遷移時間表以及領導組織進行大規模轉型。如需詳細資訊，請參閱 AWS 雲端 企業策略部落格上的 CCoE 文章。

雲端運算

通常用於遠端資料儲存和 IoT 裝置管理的雲端技術。雲端運算通常連接到邊緣運算技術。

雲端操作模型

在 IT 組織中，用於建置、成熟和最佳化一或多個雲端環境的操作模型。如需詳細資訊，請參閱建置您的雲端操作模型。

採用雲端階段

組織在遷移至 時通常會經歷的四個階段 AWS 雲端：

- 專案 – 執行一些與雲端相關的專案以進行概念驗證和學習用途
- 基礎 – 進行基礎投資以擴展雲端採用 (例如，建立登陸區域、定義 CCoE、建立營運模型)

- 遷移 – 遷移個別應用程式
- 重塑 – 優化產品和服務，並在雲端中創新

這些階段由 Stephen Orban 在部落格文章 The Journey Toward Cloud-First 和企業策略部落格上的採用階段中定義。 AWS 雲端 如需有關它們如何與 AWS 遷移策略關聯的資訊，請參閱遷移整備指南。

CMDB

請參閱組態管理資料庫。

程式碼儲存庫

透過版本控制程序來儲存及更新原始程式碼和其他資產 (例如文件、範例和指令碼) 的位置。常見的雲端儲存庫包括 GitHub或 Bitbucket Cloud。程式碼的每個版本都稱為分支。在微服務結構中，每個儲存庫都專用於單個功能。單一 CI/CD 管道可以使用多個儲存庫。

冷快取

一種緩衝快取，它是空的、未填充的，或者包含過時或不相關的資料。這會影響效能，因為資料庫執行個體必須從主記憶體或磁碟讀取，這比從緩衝快取讀取更慢。

冷資料

很少存取且通常是歷史資料的資料。查詢這類資料時，通常可接受慢查詢。將此資料移至效能較低且成本較低的儲存層或類別，可以降低成本。

電腦視覺 (CV)

使用機器學習從數位影像和影片等視覺化格式分析和擷取資訊的 AI 欄位。例如，Amazon SageMaker AI 提供 CV 的影像處理演算法。

組態偏離

對於工作負載，組態會從預期狀態變更。這可能會導致工作負載變得不合規，而且通常是漸進和無意的。

組態管理資料庫 (CMDB)

儲存和管理有關資料庫及其 IT 環境的資訊的儲存庫，同時包括硬體和軟體元件及其組態。您通常在遷移的產品組合探索和分析階段使用 CMDB 中的資料。

一致性套件

您可以組合的 AWS Config 規則和修補動作集合，以自訂您的合規和安全檢查。您可以使用 YAML 範本，將一致性套件部署為 AWS 帳戶 和 區域中或整個組織的單一實體。如需詳細資訊，請參閱 AWS Config 文件中的一致性套件。

持續整合和持續交付 (CI/CD)

自動化軟體發行程序的來源、建置、測試、暫存和生產階段的程序。CI/CD 通常被描述為管道。CI/CD 可協助您將程序自動化、提升生產力、改善程式碼品質以及加快交付速度。如需詳細資訊，請參閱持續交付的優點。CD 也可表示持續部署。如需詳細資訊，請參閱持續交付與持續部署。

CV

請參閱電腦視覺。

D

靜態資料

網路中靜止的資料，例如儲存中的資料。

資料分類

根據重要性和敏感性來識別和分類網路資料的程序。它是所有網路安全風險管理策略的關鍵組成部分，因為它可以協助您確定適當的資料保護和保留控制。資料分類是 AWS Well-Architected Framework 中安全支柱的元件。如需詳細資訊，請參閱資料分類。

資料偏離

生產資料與用於訓練 ML 模型的資料之間有意義的變化，或輸入資料隨時間有意義的變更。資料偏離可以降低 ML 模型預測的整體品質、準確性和公平性。

傳輸中的資料

在您的網路中主動移動的資料，例如在網路資源之間移動。

資料網格

架構架構，提供分散式、分散式資料擁有權與集中式管理。

資料最小化

僅收集和處理嚴格必要資料的原則。在 中實作資料最小化 AWS 雲端 可以降低隱私權風險、成本和分析碳足跡。

資料周邊

AWS 環境中的一組預防性防護機制，可協助確保只有信任的身分才能從預期的網路存取信任的資源。如需詳細資訊，請參閱在 上建置資料周邊 AWS。

資料預先處理

將原始資料轉換成 ML 模型可輕鬆剖析的格式。預處理資料可能意味著移除某些欄或列，並解決遺失、不一致或重複的值。

資料來源

在整個生命週期中追蹤資料的原始伺服器和歷史記錄的程序，例如資料的產生、傳輸和儲存方式。

資料主體

正在收集和處理資料的個人。

資料倉儲

支援商業智慧的資料管理系統，例如 分析。資料倉儲通常包含大量歷史資料，通常用於查詢和分析。

資料庫定義語言 (DDL)

用於建立或修改資料庫中資料表和物件之結構的陳述式或命令。

資料庫處理語言 (DML)

用於修改 (插入、更新和刪除) 資料庫中資訊的陳述式或命令。

DDL

請參閱資料庫定義語言。

深度整體

結合多個深度學習模型進行預測。可以使用深度整體來獲得更準確的預測或估計預測中的不確定性。

深度學習

一個機器學習子領域，它使用多層人工神經網路來識別感興趣的輸入資料與目標變數之間的對應關係。

深度防禦

這是一種資訊安全方法，其中一系列的安全機制和控制項會在整個電腦網路中精心分層，以保護網路和其中資料的機密性、完整性和可用性。當您在 上採用此策略時 AWS，您可以在 AWS Organizations 結構的不同層新增多個控制項，以協助保護資源。例如，defense-in-depth方法可能會結合多重要素驗證、網路分割和加密。

委派的管理員

在 中 AWS Organizations，相容的服務可以註冊 AWS 成員帳戶，以管理組織的帳戶和管理該服務的許可。此帳戶稱為該服務的委派管理員。如需詳細資訊和相容服務清單，請參閱 AWS Organizations 文件中的可搭配 AWS Organizations運作的服務。

deployment

在目標環境中提供應用程式、新功能或程式碼修正的程序。部署涉及在程式碼庫中實作變更，然後在應用程式環境中建置和執行該程式碼庫。

開發環境

請參閱 環境。

偵測性控制

一種安全控制，用於在事件發生後偵測、記錄和提醒。這些控制是第二道防線，提醒您注意繞過現有預防性控制的安全事件。如需詳細資訊，請參閱在 AWS上實作安全控制中的偵測性控制。

開發值串流映射 (DVSM)

一種程序，用於識別對軟體開發生命週期中的速度和品質造成負面影響的限制並排定優先順序。DVSM 擴展了原本專為精簡製造實務設計的價值串流映射程序。它著重於透過軟體開發程序建立和移動價值所需的步驟和團隊。

數位分身

真實世界系統的虛擬呈現，例如建築物、工廠、工業設備或生產線。數位分身支援預測性維護、遠端監控和生產最佳化。

維度資料表

在星星結構描述中，較小的資料表包含有關事實資料表中量化資料的資料屬性。維度資料表屬性通常是文字欄位或離散數字，其行為類似於文字。這些屬性通常用於查詢限制、篩選和結果集標記。

災難

防止工作負載或系統在其主要部署位置中實現其業務目標的事件。這些事件可能是自然災難、技術故障或人為動作的結果，例如意外設定錯誤或惡意軟體攻擊。

災難復原 (DR)

您用來將災難造成的停機時間和資料遺失降至最低的策略和程序。如需詳細資訊，請參閱 AWS Well-Architected Framework 中的 上工作負載災難復原 AWS：雲端中的復原。

DML

請參閱資料庫處理語言。

領域驅動的設計

一種開發複雜軟體系統的方法，它會將其元件與每個元件所服務的不斷發展的領域或核心業務目標相關聯。Eric Evans 在其著作 Domain-Driven Design: Tackling Complexity in the Heart of Software (Boston: Addison-Wesley Professional, 2003) 中介紹了這一概念。如需有關如何將領域驅動的設計與 strangler fig 模式搭配使用的資訊，請參閱使用容器和 Amazon API Gateway 逐步現代化舊版 Microsoft ASP.NET (ASMX) Web 服務。

DR

請參閱災難復原。

偏離偵測

追蹤與基準組態的偏差。例如，您可以使用 AWS CloudFormation 來偵測系統資源中的偏離，也可以使用 AWS Control Tower 來偵測登陸區域中可能影響控管要求合規性的變更。 https://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/using-cfn-stack-drift.html

DVSM

請參閱開發值串流映射。

E

EDA

請參閱探索性資料分析。

EDI

請參閱電子資料交換。

邊緣運算

提升 IoT 網路邊緣智慧型裝置運算能力的技術。與雲端運算相比，邊緣運算可以減少通訊延遲並改善回應時間。

電子資料交換 (EDI)

在組織之間自動交換商業文件。如需詳細資訊，請參閱什麼是電子資料交換。

加密

一種運算程序，可將人類可讀取的純文字資料轉換為加密文字。

加密金鑰

由加密演算法產生的隨機位元的加密字串。金鑰長度可能有所不同，每個金鑰的設計都是不可預測且唯一的。

端序

位元組在電腦記憶體中的儲存順序。大端序系統首先儲存最高有效位元組。小端序系統首先儲存最低有效位元組。

端點

請參閱 服務端點。

端點服務

您可以在虛擬私有雲端 (VPC) 中託管以與其他使用者共用的服務。您可以使用 建立端點服務，AWS PrivateLink 並將許可授予其他 AWS 帳戶 或 AWS Identity and Access Management (IAM) 委託人。這些帳戶或主體可以透過建立介面 VPC 端點私下連接至您的端點服務。如需詳細資訊，請參閱 Amazon Virtual Private Cloud (Amazon VPC) 文件中的建立端點服務。

企業資源規劃 (ERP)

一種系統，可自動化和管理企業的關鍵業務流程 （例如會計、MES 和專案管理）。

信封加密

使用另一個加密金鑰對某個加密金鑰進行加密的程序。如需詳細資訊，請參閱 AWS Key Management Service (AWS KMS) 文件中的信封加密。

環境

執行中應用程式的執行個體。以下是雲端運算中常見的環境類型：

- 開發環境 – 執行中應用程式的執行個體，只有負責維護應用程式的核心團隊才能使用。開發環境用來測試變更，然後再將開發環境提升到較高的環境。此類型的環境有時稱為測試環境。
- 較低的環境 – 應用程式的所有開發環境，例如用於初始建置和測試的開發環境。
- 生產環境 – 最終使用者可以存取的執行中應用程式的執行個體。在 CI/CD 管道中，生產環境是最後一個部署環境。
- 較高的環境 – 核心開發團隊以外的使用者可存取的所有環境。這可能包括生產環境、生產前環境以及用於使用者接受度測試的環境。

epic

在敏捷方法中，有助於組織工作並排定工作優先順序的功能類別。epic 提供要求和實作任務的高層級描述。例如， AWS CAF 安全概念包括身分和存取管理、偵測控制、基礎設施安全、資料保護和事件回應。如需有關 AWS 遷移策略中的 Epic 的詳細資訊，請參閱計畫實作指南。

ERP

請參閱企業資源規劃。

探索性資料分析 (EDA)

分析資料集以了解其主要特性的過程。您收集或彙總資料，然後執行初步調查以尋找模式、偵測異常並檢查假設。透過計算摘要統計並建立資料可視化來執行 EDA。

F

事實資料表

星狀結構描述中的中央資料表。它存放有關業務操作的量化資料。一般而言，事實資料表包含兩種類型的資料欄：包含度量的資料，以及包含維度資料表外部索引鍵的資料欄。

快速失敗

一種使用頻繁和增量測試來縮短開發生命週期的理念。這是敏捷方法的關鍵部分。

故障隔離界限

在 中 AWS 雲端，像是可用區域 AWS 區域、控制平面或資料平面等邊界會限制故障的影響，並有助於改善工作負載的彈性。如需詳細資訊，請參閱AWS 故障隔離界限。

功能分支

請參閱分支。

特徵

用來進行預測的輸入資料。例如，在製造環境中，特徵可能是定期從製造生產線擷取的影像。

功能重要性

特徵對於模型的預測有多重要。這通常表示為可以透過各種技術來計算的數值得分，例如 Shapley Additive Explanations (SHAP) 和積分梯度。如需詳細資訊，請參閱 的機器學習模型可解譯性 AWS。

特徵轉換

優化 ML 程序的資料，包括使用其他來源豐富資料、調整值、或從單一資料欄位擷取多組資訊。這可讓 ML 模型從資料中受益。例如，如果將「2021-05-27 00:15:37」日期劃分為「2021」、「五月」、「週四」和「15」，則可以協助學習演算法學習與不同資料元件相關聯的細微模式。

少量擷取提示

在要求 LLM 執行類似的任務之前，提供少量示範任務和所需輸出的範例。此技術是內容內學習的應用程式，其中模型會從內嵌在提示中的範例 (快照) 中學習。對於需要特定格式、推理或網域知識的任務，少量的提示可以有效。另請參閱零鏡頭提示。

FGAC

請參閱精細存取控制。

精細存取控制 (FGAC)

使用多個條件來允許或拒絕存取請求。

閃切遷移

一種資料庫遷移方法，透過變更資料擷取使用連續資料複寫，以盡可能在最短的時間內遷移資料，而不是使用分階段方法。目標是將停機時間降至最低。

FM

請參閱基礎模型。

基礎模型 (FM)

大型深度學習神經網路，已在廣義和未標記資料的大量資料集上進行訓練。FMs 能夠執行各種一般任務，例如了解語言、產生文字和影像，以及以自然語言交談。如需詳細資訊，請參閱什麼是基礎模型。

G

生成式 AI

已針對大量資料進行訓練的 AI 模型子集，可使用簡單的文字提示建立新的內容和成品，例如影像、影片、文字和音訊。如需詳細資訊，請參閱什麼是生成式 AI。

地理封鎖

請參閱地理限制。

地理限制 (地理封鎖)

Amazon CloudFront 中的選項，可防止特定國家/地區的使用者存取內容分發。您可以使用允許清單或封鎖清單來指定核准和禁止的國家/地區。如需詳細資訊，請參閱 CloudFront 文件中的限制內容的地理分佈。

Gitflow 工作流程

這是一種方法，其中較低和較高環境在原始碼儲存庫中使用不同分支。Gitflow 工作流程被視為舊版，而以幹線為基礎的工作流程是現代、偏好的方法。

黃金影像

系統或軟體的快照，做為部署該系統或軟體新執行個體的範本。例如，在製造中，黃金映像可用於在多個裝置上佈建軟體，並有助於提高裝置製造操作的速度、可擴展性和生產力。

綠地策略

新環境中缺乏現有基礎設施。對系統架構採用綠地策略時，可以選擇所有新技術，而不會限制與現有基礎設施的相容性，也稱為棕地。如果正在擴展現有基礎設施，則可能會混合棕地和綠地策略。

防護機制

有助於跨組織單位 (OU) 來管控資源、政策和合規的高層級規則。預防性防護機制會強制執行政策，以確保符合合規標準。透過使用服務控制政策和 IAM 許可界限來將其實作。偵測性防護機制可偵測政策違規和合規問題，並產生提醒以便修正。它們是透過使用 AWS Config AWS Security Hub CSPM、Amazon GuardDuty、、Amazon Inspector AWS Trusted Advisor和自訂 AWS Lambda 檢查來實作。

H

HA

請參閱高可用性。

異質資料庫遷移

將來源資料庫遷移至使用不同資料庫引擎的目標資料庫 (例如，Oracle 至 Amazon Aurora)。異質遷移通常是重新架構工作的一部分，而轉換結構描述可能是一項複雜任務。AWS 提供有助於結構描述轉換的 AWS SCT。

高可用性 (HA)

在遇到挑戰或災難時，工作負載能夠在不介入的情況下持續運作。HA 系統的設計目的是自動容錯移轉、持續提供高品質的效能，並處理不同的負載和故障，並將效能影響降至最低。

歷史現代化

一種方法，用於現代化和升級操作技術 (OT) 系統，以更好地滿足製造業的需求。歷史資料是一種資料庫，用於從工廠中的各種來源收集和存放資料。

保留資料

從用於訓練機器學習模型的資料集中保留的部分歷史標記資料。您可以使用保留資料，透過比較模型預測與保留資料來評估模型效能。

異質資料庫遷移

將您的來源資料庫遷移至共用相同資料庫引擎的目標資料庫 (例如，Microsoft SQL Server 至 Amazon RDS for SQL Server)。同質遷移通常是主機轉換或平台轉換工作的一部分。您可以使用原生資料庫公用程式來遷移結構描述。

熱資料

經常存取的資料，例如即時資料或最近的轉譯資料。此資料通常需要高效能儲存層或類別，才能提供快速的查詢回應。

修補程序

緊急修正生產環境中的關鍵問題。由於其緊迫性，通常會在典型 DevOps 發行工作流程之外執行修補程式。

超級護理期間

在切換後，遷移團隊在雲端管理和監控遷移的應用程式以解決任何問題的時段。通常，此期間的長度為 1-4 天。在超級護理期間結束時，遷移團隊通常會將應用程式的責任轉移給雲端營運團隊。

I

IaC

將基礎設施視為程式碼。

身分型政策

連接至一或多個 IAM 主體的政策，可定義其在 AWS 雲端 環境中的許可。

閒置應用程式

90 天期間 CPU 和記憶體平均使用率在 5% 至 20% 之間的應用程式。在遷移專案中，通常會淘汰這些應用程式或將其保留在內部部署。

IIoT

請參閱工業物聯網。

不可變的基礎設施

為生產工作負載部署新基礎設施的模型，而不是更新、修補或修改現有的基礎設施。不可變基礎設施本質上比可變基礎設施更一致、可靠且可預測。如需詳細資訊，請參閱 AWS Well-Architected Framework 中的使用不可變基礎設施的部署最佳實務。

傳入 (輸入) VPC

在 AWS 多帳戶架構中，接受、檢查和路由來自應用程式外部之網路連線的 VPC。AWS 安全參考架構建議您使用傳入、傳出和檢查 VPC 來設定網路帳戶，以保護應用程式與更廣泛的網際網路之間的雙向介面。

增量遷移

一種切換策略，您可以在其中將應用程式分成小部分遷移，而不是執行單一、完整的切換。例如，您最初可能只將一些微服務或使用者移至新系統。確認所有項目都正常運作之後，您可以逐步移動其他微服務或使用者，直到可以解除委任舊式系統。此策略可降低與大型遷移關聯的風險。

工業 4.0

2016 年 Klaus Schwab 推出的術語，透過連線能力、即時資料、自動化、分析和 AI/ML 的進展，指製造程序的現代化。

基礎設施

應用程式環境中包含的所有資源和資產。

基礎設施即程式碼 (IaC)

透過一組組態檔案來佈建和管理應用程式基礎設施的程序。IaC 旨在協助您集中管理基礎設施，標準化資源並快速擴展，以便新環境可重複、可靠且一致。

工業物聯網 (IIoT)

在製造業、能源、汽車、醫療保健、生命科學和農業等產業領域使用網際網路連線的感測器和裝置。如需詳細資訊，請參閱建立工業物聯網 (IIoT) 數位轉型策略。

檢查 VPC

在 AWS 多帳戶架構中，集中式 VPC，可管理 VPCs （在相同或不同的 中 AWS 區域)、網際網路和內部部署網路之間的網路流量檢查。AWS 安全參考架構建議您使用傳入、傳出和檢查 VPC 來設定網路帳戶，以保護應用程式與更廣泛的網際網路之間的雙向介面。

物聯網 (IoT)

具有內嵌式感測器或處理器的相連實體物體網路，其透過網際網路或本地通訊網路與其他裝置和系統進行通訊。如需詳細資訊，請參閱什麼是 IoT？

可解釋性

機器學習模型的一個特徵，描述了人類能夠理解模型的預測如何依賴於其輸入的程度。如需詳細資訊，請參閱 的機器學習模型可解譯性 AWS。

IoT

請參閱物聯網。

IT 資訊庫 (ITIL)

一組用於交付 IT 服務並使這些服務與業務需求保持一致的最佳實務。ITIL 為 ITSM 提供了基礎。

IT 服務管理 (ITSM)

與組織的設計、實作、管理和支援 IT 服務關聯的活動。如需有關將雲端操作與 ITSM 工具整合的資訊，請參閱操作整合指南。

ITIL

請參閱 IT 資訊庫。

ITSM

請參閱 IT 服務管理。

L

標籤型存取控制 (LBAC)

強制存取控制 (MAC) 的實作，其中使用者和資料本身都會獲得明確指派的安全標籤值。使用者安全標籤和資料安全標籤之間的交集會決定使用者可以看到哪些資料列和資料欄。

登陸區域

登陸區域是架構良好的多帳戶 AWS 環境，可擴展且安全。這是一個起點，您的組織可以從此起點快速啟動和部署工作負載與應用程式，並對其安全和基礎設施環境充滿信心。如需有關登陸區域的詳細資訊，請參閱設定安全且可擴展的多帳戶 AWS 環境。

大型語言模型 (LLM)

預先訓練大量資料的深度學習 AI 模型。LLM 可以執行多個任務，例如回答問題、摘要文件、將文字翻譯成其他語言，以及完成句子。如需詳細資訊，請參閱什麼是 LLMs。

大型遷移

遷移 300 部或更多伺服器。

LBAC

請參閱標籤型存取控制。

最低權限

授予執行任務所需之最低許可的安全最佳實務。如需詳細資訊，請參閱 IAM 文件中的套用最低權限許可。

隨即轉移

請參閱 7 個 R。

小端序系統

首先儲存最低有效位元組的系統。另請參閱 Endianness。

LLM

請參閱大型語言模型。

較低的環境

請參閱 環境。

M

機器學習 (ML)

一種使用演算法和技術進行模式識別和學習的人工智慧。機器學習會進行分析並從記錄的資料 (例如物聯網 (IoT) 資料) 中學習，以根據模式產生統計模型。如需詳細資訊，請參閱機器學習。

主要分支

請參閱分支。

惡意軟體

旨在危及電腦安全或隱私權的軟體。惡意軟體可能會中斷電腦系統、洩露敏感資訊，或取得未經授權的存取。惡意軟體的範例包括病毒、蠕蟲、勒索軟體、特洛伊木馬、間諜軟體和鍵盤記錄器。

受管服務

AWS 服務 會 AWS 操作基礎設施層、作業系統和平台，而您會存取端點來存放和擷取資料。Amazon Simple Storage Service (Amazon S3) 和 Amazon DynamoDB 是受管服務的範例。這些也稱為抽象服務。

製造執行系統 (MES)

一種軟體系統，用於追蹤、監控、記錄和控制生產程序，將原物料轉換為現場成品。

MAP

請參閱遷移加速計劃。

機制

建立工具、推動工具採用，然後檢查結果以進行調整的完整程序。機制是在操作時強化和改善自身的循環。如需詳細資訊，請參閱 AWS Well-Architected Framework 中的建置機制。

成員帳戶

屬於組織一部分的管理帳戶 AWS 帳戶 以外的所有 AWS Organizations。帳戶一次只能是一個組織的成員。

製造執行系統

請參閱製造執行系統。

訊息佇列遙測傳輸 (MQTT)

根據發佈/訂閱模式的輕量型machine-to-machine(M2M) 通訊協定，適用於資源受限的 IoT 裝置。

微服務

一種小型的獨立服務，它可透過定義明確的 API 進行通訊，通常由小型獨立團隊擁有。例如，保險系統可能包含對應至業務能力 (例如銷售或行銷) 或子領域 (例如購買、索賠或分析) 的微服務。微服務的優點包括靈活性、彈性擴展、輕鬆部署、可重複使用的程式碼和適應力。如需詳細資訊，請參閱使用無 AWS 伺服器服務整合微服務。

微服務架構

一種使用獨立元件來建置應用程式的方法，這些元件會以微服務形式執行每個應用程式程序。這些微服務會使用輕量型 API，透過明確定義的介面進行通訊。此架構中的每個微服務都可以進行

更新、部署和擴展，以滿足應用程式特定功能的需求。如需詳細資訊，請參閱在 上實作微服務 AWS。

Migration Acceleration Program (MAP)

一種 AWS 計畫，提供諮詢支援、訓練和服務，協助組織建立強大的營運基礎，以移至雲端，並協助抵銷遷移的初始成本。MAP 包括用於有條不紊地執行舊式遷移的遷移方法以及一組用於自動化和加速常見遷移案例的工具。

大規模遷移

將大部分應用程式組合依波次移至雲端的程序，在每個波次中，都會以更快的速度移動更多應用程式。此階段使用從早期階段學到的最佳實務和經驗教訓來實作團隊、工具和流程的遷移工廠，以透過自動化和敏捷交付簡化工作負載的遷移。這是 AWS 遷移策略的第三階段。

遷移工廠

可透過自動化、敏捷的方法簡化工作負載遷移的跨職能團隊。遷移工廠團隊通常包括營運、業務分析師和擁有者、遷移工程師、開發人員以及從事 Sprint 工作的 DevOps 專業人員。20% 至 50% 之間的企業應用程式組合包含可透過工廠方法優化的重複模式。如需詳細資訊，請參閱此內容集中的遷移工廠的討論和雲端遷移工廠指南。

遷移中繼資料

有關完成遷移所需的應用程式和伺服器的資訊。每種遷移模式都需要一組不同的遷移中繼資料。遷移中繼資料的範例包括目標子網路、安全群組和 AWS 帳戶。

遷移模式

可重複的遷移任務，詳細描述遷移策略、遷移目的地以及所使用的遷移應用程式或服務。範例：使用 AWS Application Migration Service 重新託管遷移至 Amazon EC2。

遷移組合評定 (MPA)

線上工具，提供驗證商業案例以遷移至 的資訊 AWS 雲端。MPA 提供詳細的組合評定 (伺服器適當規模、定價、總體擁有成本比較、遷移成本分析) 以及遷移規劃 (應用程式資料分析和資料收集、應用程式分組、遷移優先順序，以及波次規劃)。MPA 工具 （需要登入） 可供所有 AWS 顧問和 APN 合作夥伴顧問免費使用。

遷移準備程度評定 (MRA)

使用 AWS CAF 取得組織雲端整備狀態的洞見、識別優缺點，以及建立行動計劃以消除已識別差距的程序。如需詳細資訊，請參閱遷移準備程度指南。MRA 是 AWS 遷移策略的第一階段。

遷移策略

用來將工作負載遷移至 的方法 AWS 雲端。如需詳細資訊，請參閱此詞彙表中的 7 個 Rs 項目，並請參閱動員您的組織以加速大規模遷移。

機器學習 (ML)

請參閱機器學習。

現代化

將過時的 (舊版或單一) 應用程式及其基礎架構轉換為雲端中靈活、富有彈性且高度可用的系統，以降低成本、提高效率並充分利用創新。如需詳細資訊，請參閱《》中的現代化應用程式的策略 AWS 雲端。

現代化準備程度評定

這項評估可協助判斷組織應用程式的現代化準備程度；識別優點、風險和相依性；並確定組織能夠在多大程度上支援這些應用程式的未來狀態。評定的結果就是目標架構的藍圖、詳細說明現代化程序的開發階段和里程碑的路線圖、以及解決已發現的差距之行動計畫。如需詳細資訊，請參閱《》中的評估應用程式的現代化準備 AWS 雲端程度。

單一應用程式 (單一)

透過緊密結合的程序作為單一服務執行的應用程式。單一應用程式有幾個缺點。如果一個應用程式功能遇到需求激增，則必須擴展整個架構。當程式碼庫增長時，新增或改進單一應用程式的功能也會變得更加複雜。若要解決這些問題，可以使用微服務架構。如需詳細資訊，請參閱將單一體系分解為微服務。

MPA

請參閱遷移產品組合評估。

MQTT

請參閱訊息佇列遙測傳輸。

多類別分類

一個有助於產生多類別預測的過程 (預測兩個以上的結果之一)。例如，機器學習模型可能會詢問「此產品是書籍、汽車還是電話？」 或者「這個客戶對哪種產品類別最感興趣？」

可變基礎設施

更新和修改生產工作負載現有基礎設施的模型。為了提高一致性、可靠性和可預測性， AWS Well-Architected Framework 建議使用不可變的基礎設施作為最佳實務。

O

OAC

請參閱原始存取控制。

OAI

請參閱原始存取身分。

OCM

請參閱組織變更管理。

離線遷移

一種遷移方法，可在遷移過程中刪除來源工作負載。此方法涉及延長停機時間，通常用於小型非關鍵工作負載。

OI

請參閱 操作整合。

OLA

請參閱操作層級協議。

線上遷移

一種遷移方法，無需離線即可將來源工作負載複製到目標系統。連接至工作負載的應用程式可在遷移期間繼續運作。此方法涉及零至最短停機時間，通常用於關鍵的生產工作負載。

OPC-UA

請參閱開放程序通訊 - 統一架構。

開放程序通訊 - 統一架構 (OPC-UA)

用於工業自動化machine-to-machine(M2M) 通訊協定。OPC-UA 提供資料加密、身分驗證和授權機制的互通性標準。

操作水準協議 (OLA)

一份協議，闡明 IT 職能群組承諾向彼此提供的內容，以支援服務水準協議 (SLA)。

操作整備審查 (ORR)

問題及相關最佳實務的檢查清單，可協助您了解、評估、預防或減少事件和可能失敗的範圍。如需詳細資訊，請參閱 AWS Well-Architected Framework 中的操作準備度審查 (ORR)。

操作技術 (OT)

使用實體環境控制工業操作、設備和基礎設施的硬體和軟體系統。在製造中，OT 和資訊技術 (IT) 系統的整合是工業 4.0 轉型的關鍵重點。

操作整合 (OI)

在雲端中將操作現代化的程序，其中包括準備程度規劃、自動化和整合。如需詳細資訊，請參閱操作整合指南。

組織追蹤

由 建立的線索 AWS CloudTrail 會記錄 AWS 帳戶 組織中所有 的所有事件 AWS Organizations。在屬於組織的每個 AWS 帳戶 中建立此追蹤，它會跟蹤每個帳戶中的活動。如需詳細資訊，請參閱 CloudTrail 文件中的建立組織追蹤。

組織變更管理 (OCM)

用於從人員、文化和領導力層面管理重大、顛覆性業務轉型的架構。OCM 透過加速變更採用、解決過渡問題，以及推動文化和組織變更，協助組織為新系統和策略做好準備，並轉移至新系統和策略。在 AWS 遷移策略中，此架構稱為人員加速，因為雲端採用專案所需的變更速度。如需詳細資訊，請參閱 OCM 指南。

原始存取控制 (OAC)

CloudFront 中的增強型選項，用於限制存取以保護 Amazon Simple Storage Service (Amazon S3) 內容。OAC 支援所有 S3 儲存貯體中的所有伺服器端加密 AWS KMS (SSE-KMS) AWS 區域，以及對 S3 儲存貯體的動態PUT和DELETE請求。

原始存取身分 (OAI)

CloudFront 中的一個選項，用於限制存取以保護 Amazon S3 內容。當您使用 OAI 時，CloudFront 會建立一個可供 Amazon S3 進行驗證的主體。經驗證的主體只能透過特定 CloudFront 分發來存取 S3 儲存貯體中的內容。另請參閱 OAC，它可提供更精細且增強的存取控制。

ORR

請參閱操作整備審核。

OT

請參閱操作技術。

傳出 (輸出) VPC

在 AWS 多帳戶架構中，處理從應用程式內啟動之網路連線的 VPC。AWS 安全參考架構建議您使用傳入、傳出和檢查 VPC 來設定網路帳戶，以保護應用程式與更廣泛的網際網路之間的雙向介面。

P

許可界限

附接至 IAM 主體的 IAM 管理政策，可設定使用者或角色擁有的最大許可。如需詳細資訊，請參閱 IAM 文件中的許可界限。

個人身分識別資訊 (PII)

當直接檢視或與其他相關資料配對時，可用來合理推斷個人身分的資訊。PII 的範例包括名稱、地址和聯絡資訊。

PII

請參閱個人身分識別資訊。

手冊

一組預先定義的步驟，可擷取與遷移關聯的工作，例如在雲端中提供核心操作功能。手冊可以採用指令碼、自動化執行手冊或操作現代化環境所需的程序或步驟摘要的形式。

PLC

請參閱可程式設計邏輯控制器。

PLM

請參閱產品生命週期管理。

政策

可定義許可的物件（請參閱身分型政策）、指定存取條件（請參閱資源型政策），或定義組織中所有帳戶的最大許可 AWS Organizations（請參閱服務控制政策）。

混合持久性

根據資料存取模式和其他需求，獨立選擇微服務的資料儲存技術。如果您的微服務具有相同的資料儲存技術，則其可能會遇到實作挑戰或效能不佳。如果微服務使用最適合其需求的資料儲存，則可以更輕鬆地實作並達到更好的效能和可擴展性。

組合評定

探索、分析應用程式組合並排定其優先順序以規劃遷移的程序。如需詳細資訊，請參閱評估遷移準備程度。

述詞

傳回 `true`或 的查詢條件`false`，通常位於 `WHERE`子句中。

述詞下推

一種資料庫查詢最佳化技術，可在傳輸前篩選查詢中的資料。這可減少必須從關聯式資料庫擷取和處理的資料量，並改善查詢效能。

預防性控制

旨在防止事件發生的安全控制。這些控制是第一道防線，可協助防止對網路的未經授權存取或不必要變更。如需詳細資訊，請參閱在 AWS上實作安全控制中的預防性控制。

委託人

中可執行動作和存取資源 AWS 的實體。此實體通常是 AWS 帳戶、IAM 角色或使用者的根使用者。如需詳細資訊，請參閱 IAM 文件中角色術語和概念中的主體。

設計隱私權

透過整個開發程序將隱私權納入考量的系統工程方法。

私有託管區域

一種容器，它包含有關您希望 Amazon Route 53 如何回應一個或多個 VPC 內的域及其子域之 DNS 查詢的資訊。如需詳細資訊，請參閱 Route 53 文件中的使用私有託管區域。

主動控制

旨在防止部署不合規資源的安全控制。這些控制項會在佈建資源之前對其進行掃描。如果資源不符合控制項，則不會佈建。如需詳細資訊，請參閱 AWS Control Tower 文件中的控制項參考指南，並參閱實作安全控制項中的主動控制項。 AWS

產品生命週期管理 (PLM)

管理產品整個生命週期的資料和程序，從設計、開發和啟動，到成長和成熟，再到拒絕和移除。

生產環境

請參閱 環境。

可程式設計邏輯控制器 (PLC)

在製造中，高度可靠、可調整的電腦，可監控機器並自動化製造程序。

提示鏈結

使用一個 LLM 提示的輸出做為下一個提示的輸入，以產生更好的回應。此技術用於將複雜任務分解為子任務，或反覆精簡或展開初步回應。它有助於提高模型回應的準確性和相關性，並允許更精細、個人化的結果。

擬匿名化

將資料集中的個人識別符取代為預留位置值的程序。假名化有助於保護個人隱私權。假名化資料仍被視為個人資料。

發佈/訂閱 (pub/sub)

一種模式，可啟用微服務之間的非同步通訊，以改善可擴展性和回應能力。例如，在微服務型 MES 中，微服務可以將事件訊息發佈到其他微服務可訂閱的頻道。系統可以新增新的微服務，而無需變更發佈服務。

Q

查詢計劃

一系列步驟，如指示，用於存取 SQL 關聯式資料庫系統中的資料。

查詢計劃迴歸

在資料庫服務優化工具選擇的計畫比對資料庫環境進行指定的變更之前的計畫不太理想時。這可能因為對統計資料、限制條件、環境設定、查詢參數繫結的變更以及資料庫引擎的更新所導致。

R

RACI 矩陣

請參閱負責、負責、諮詢、告知 (RACI)。

RAG

請參閱擷取增強生成。

勒索軟體

一種惡意軟體，旨在阻止對計算機系統或資料的存取，直到付款為止。

RASCI 矩陣

請參閱負責、負責、諮詢、告知 (RACI)。

RCAC

請參閱資料列和資料欄存取控制。

僅供讀取複本

用於唯讀用途的資料庫複本。您可以將查詢路由至僅供讀取複本以減少主資料庫的負載。

重新架構師

請參閱 7 個 R。

復原點目標 (RPO)

自上次資料復原點以來可接受的時間上限。這會決定最後一個復原點與服務中斷之間可接受的資料遺失。

復原時間目標 (RTO)

服務中斷與服務還原之間的可接受延遲上限。

重構

請參閱 7 個 R。

區域

地理區域中的 AWS 資源集合。每個 AWS 區域 都獨立於其他 ，以提供容錯能力、穩定性和彈性。如需詳細資訊，請參閱指定 AWS 區域 您的帳戶可以使用哪些。

迴歸

預測數值的 ML 技術。例如，為了解決「這房子會賣什麼價格？」的問題 ML 模型可以使用線性迴歸模型，根據已知的房屋事實 (例如，平方英尺) 來預測房屋的銷售價格。

重新託管

請參閱 7 個 R。

版本

在部署程序中，它是將變更提升至生產環境的動作。

重新放置

請參閱 7 Rs。

Replatform

請參閱 7 Rs。

回購

請參閱 7 Rs。

彈性

應用程式抵禦中斷或從中斷中復原的能力。在 中規劃彈性時，高可用性和災難復原是常見的考量 AWS 雲端。如需詳細資訊，請參閱AWS 雲端 彈性。

資源型政策

附接至資源的政策，例如 Amazon S3 儲存貯體、端點或加密金鑰。這種類型的政策會指定允許存取哪些主體、支援的動作以及必須滿足的任何其他條件。

負責者、當責者、事先諮詢者和事後告知者 (RACI) 矩陣

矩陣，定義所有涉及遷移活動和雲端操作之各方的角色和責任。矩陣名稱衍生自矩陣中定義的責任類型：負責人 (R)、責任 (A)、已諮詢 (C) 和知情 (I)。支援 (S) 類型為選用。如果您包含支援，則矩陣稱為 RASCI 矩陣，如果您排除它，則稱為 RACI 矩陣。

回應性控制

一種安全控制，旨在驅動不良事件或偏離安全基準的補救措施。如需詳細資訊，請參閱在 AWS上實作安全控制中的回應性控制。

保留

請參閱 7 個 R。

淘汰

請參閱 7 個 R。

檢索增強生成 (RAG)

一種生成式 AI 技術，其中 LLM 會在產生回應之前參考訓練資料來源以外的授權資料來源。例如，RAG 模型可能會對組織的知識庫或自訂資料執行語意搜尋。如需詳細資訊，請參閱什麼是 RAG。

輪換

定期更新秘密的程序，讓攻擊者更難存取登入資料。

資料列和資料欄存取控制 (RCAC)

使用已定義存取規則的基本、彈性 SQL 表達式。RCAC 包含資料列許可和資料欄遮罩。

RPO

請參閱復原點目標。

RTO

請參閱復原時間目標。

執行手冊

執行特定任務所需的一組手動或自動程序。這些通常是為了簡化重複性操作或錯誤率較高的程序而建置。

S

SAML 2.0

許多身分提供者 (IdP) 使用的開放標準。此功能會啟用聯合單一登入 (SSO)，讓使用者可以登入 AWS 管理主控台 或呼叫 AWS API 操作，而不必為您組織中的每個人在 IAM 中建立使用者。如需有關以 SAML 2.0 為基礎的聯合詳細資訊，請參閱 IAM 文件中的關於以 SAML 2.0 為基礎的聯合。

SCADA

請參閱監督控制和資料擷取。

SCP

請參閱服務控制政策。

秘密

您以加密形式存放的 AWS Secrets Manager機密或限制資訊，例如密碼或使用者登入資料。它由秘密值及其中繼資料組成。秘密值可以是二進位、單一字串或多個字串。如需詳細資訊，請參閱 Secrets Manager 文件中的 Secrets Manager 秘密中的什麼內容？。

依設計的安全性

透過整個開發程序將安全性納入考量的系統工程方法。

安全控制

一種技術或管理防護機制，它可預防、偵測或降低威脅行為者利用安全漏洞的能力。安全控制有四種主要類型：預防性、偵測性、回應性和主動性。

安全強化

減少受攻擊面以使其更能抵抗攻擊的過程。這可能包括一些動作，例如移除不再需要的資源、實作授予最低權限的安全最佳實務、或停用組態檔案中不必要的功能。

安全資訊與事件管理 (SIEM) 系統

結合安全資訊管理 (SIM) 和安全事件管理 (SEM) 系統的工具與服務。SIEM 系統會收集、監控和分析來自伺服器、網路、裝置和其他來源的資料，以偵測威脅和安全漏洞，並產生提醒。

安全回應自動化

預先定義和程式設計的動作，旨在自動回應或修復安全事件。這些自動化可做為偵測或回應式安全控制，協助您實作 AWS 安全最佳實務。自動化回應動作的範例包括修改 VPC 安全群組、修補 Amazon EC2 執行個體或輪換登入資料。

伺服器端加密

由接收資料的 AWS 服務 在其目的地加密資料。

服務控制政策 (SCP)

為 AWS Organizations中的組織的所有帳戶提供集中控制許可的政策。SCP 會定義防護機制或設定管理員可委派給使用者或角色的動作限制。您可以使用 SCP 作為允許清單或拒絕清單，以指定允許或禁止哪些服務或動作。如需詳細資訊，請參閱 AWS Organizations 文件中的服務控制政策。

服務端點

的進入點 URL AWS 服務。您可以使用端點，透過程式設計方式連接至目標服務。如需詳細資訊，請參閱 AWS 一般參考 中的 AWS 服務 端點。

服務水準協議 (SLA)

一份協議，闡明 IT 團隊承諾向客戶提供的服務，例如服務正常執行時間和效能。

服務層級指標 (SLI)

服務效能方面的測量，例如其錯誤率、可用性或輸送量。

服務層級目標 (SLO)

代表服務運作狀態的目標指標，由服務層級指標測量。

共同責任模式

描述您與 共同 AWS 承擔雲端安全與合規責任的模型。 AWS 負責雲端的安全，而 負責雲端的安全。如需詳細資訊，請參閱共同責任模式。

SIEM

請參閱安全資訊和事件管理系統。

單一故障點 (SPOF)

應用程式的單一關鍵元件故障，可能會中斷系統。

SLA

請參閱服務層級協議。

SLI

請參閱服務層級指標。

SLO

請參閱服務層級目標。

先拆分後播種模型

擴展和加速現代化專案的模式。定義新功能和產品版本時，核心團隊會進行拆分以建立新的產品團隊。這有助於擴展組織的能力和服務，提高開發人員生產力，並支援快速創新。如需詳細資訊，請參閱 中的階段式應用程式現代化方法 AWS 雲端。

SPOF

請參閱單一故障點。

星狀結構描述

使用一個大型事實資料表來存放交易或測量資料的資料庫組織結構，並使用一或多個較小的維度資料表來存放資料屬性。此結構旨在用於資料倉儲或商業智慧用途。

Strangler Fig 模式

一種現代化單一系統的方法，它會逐步重寫和取代系統功能，直到舊式系統停止使用為止。此模式源自無花果藤，它長成一棵馴化樹並最終戰勝且取代了其宿主。該模式由 Martin Fowler 引入，作為重寫單一系統時管理風險的方式。如需有關如何套用此模式的範例，請參閱使用容器和 Amazon API Gateway 逐步現代化舊版 Microsoft ASP.NET (ASMX) Web 服務。

子網

您 VPC 中的 IP 地址範圍。子網必須位於單一可用區域。

監控控制和資料擷取 (SCADA)

在製造中，使用硬體和軟體來監控實體資產和生產操作的系統。

對稱加密

使用相同金鑰來加密及解密資料的加密演算法。

合成測試

以模擬使用者互動的方式測試系統，以偵測潛在問題或監控效能。您可以使用 Amazon CloudWatch Synthetics 來建立這些測試。

系統提示

一種向 LLM 提供內容、指示或指導方針以指示其行為的技術。系統提示有助於設定內容，並建立與使用者互動的規則。

T

標籤

做為中繼資料以組織 AWS 資源的鍵/值對。標籤可協助您管理、識別、組織、搜尋及篩選資源。如需詳細資訊，請參閱標記您的 AWS 資源。

目標變數

您嘗試在受監督的 ML 中預測的值。這也被稱為結果變數。例如，在製造設定中，目標變數可能是產品瑕疵。

任務清單

用於透過執行手冊追蹤進度的工具。任務清單包含執行手冊的概觀以及要完成的一般任務清單。對於每個一般任務，它包括所需的預估時間量、擁有者和進度。

測試環境

請參閱 環境。

訓練

為 ML 模型提供資料以供學習。訓練資料必須包含正確答案。學習演算法會在訓練資料中尋找將輸入資料屬性映射至目標的模式 (您想要預測的答案)。它會輸出擷取這些模式的 ML 模型。可以使用 ML 模型，來預測您不知道的目標新資料。

傳輸閘道

可以用於互連 VPC 和內部部署網路的網路傳輸中樞。如需詳細資訊，請參閱 AWS Transit Gateway 文件中的什麼是傳輸閘道。

主幹型工作流程

這是一種方法，開發人員可在功能分支中本地建置和測試功能，然後將這些變更合併到主要分支中。然後，主要分支會依序建置到開發環境、生產前環境和生產環境中。

受信任的存取權

將許可授予您指定的服務，以代表您在組織中 AWS Organizations 及其帳戶中執行任務。受信任的服務會在需要該角色時，在每個帳戶中建立服務連結角色，以便為您執行管理工作。如需詳細資訊，請參閱 文件中的 AWS Organizations 搭配使用 AWS Organizations 與其他 AWS 服務。

調校

變更訓練程序的各個層面，以提高 ML 模型的準確性。例如，可以透過產生標籤集、新增標籤、然後在不同的設定下多次重複這些步驟來訓練 ML 模型，以優化模型。

雙比薩團隊

兩個比薩就能吃飽的小型 DevOps 團隊。雙披薩團隊規模可確保軟體開發中的最佳協作。

U

不確定性

這是一個概念，指的是不精確、不完整或未知的資訊，其可能會破壞預測性 ML 模型的可靠性。有兩種類型的不確定性：認知不確定性是由有限的、不完整的資料引起的，而隨機不確定性是由資料中固有的噪聲和隨機性引起的。如需詳細資訊，請參閱量化深度學習系統的不確定性指南。

未區分的任務

也稱為繁重工作，這是建立和操作應用程式的必要工作，但不為最終使用者提供直接價值或提供競爭優勢。未區分任務的範例包括採購、維護和容量規劃。

較高的環境

請參閱 環境。

V

清空

一種資料庫維護操作，涉及增量更新後的清理工作，以回收儲存並提升效能。

版本控制

追蹤變更的程序和工具，例如儲存庫中原始程式碼的變更。

VPC 對等互連

兩個 VPC 之間的連線，可讓您使用私有 IP 地址路由流量。如需詳細資訊，請參閱 Amazon VPC 文件中的什麼是 VPC 對等互連。

漏洞

危害系統安全性的軟體或硬體瑕疵。

W

暖快取

包含經常存取的目前相關資料的緩衝快取。資料庫執行個體可以從緩衝快取讀取，這比從主記憶體或磁碟讀取更快。

暖資料

不常存取的資料。查詢這類資料時，通常可接受中等速度的查詢。

視窗函數

SQL 函數，對與目前記錄在某種程度上相關的資料列群組執行計算。視窗函數適用於處理任務，例如根據目前資料列的相對位置計算移動平均值或存取資料列的值。

工作負載

提供商業價值的資源和程式碼集合，例如面向客戶的應用程式或後端流程。

工作串流

遷移專案中負責一組特定任務的功能群組。每個工作串流都是獨立的，但支援專案中的其他工作串流。例如，組合工作串流負責排定應用程式、波次規劃和收集遷移中繼資料的優先順序。組合工作串流將這些資產交付至遷移工作串流，然後再遷移伺服器和應用程式。

WORM

請參閱寫入一次，讀取許多。

WQF

請參閱AWS 工作負載資格架構。

寫入一次，讀取許多 (WORM)

儲存模型，可一次性寫入資料，並防止刪除或修改資料。授權使用者可以視需要多次讀取資料，但無法變更資料。此資料儲存基礎設施被視為不可變。

Z

零時差入侵

利用零時差漏洞的攻擊，通常是惡意軟體。

零時差漏洞

生產系統中未緩解的缺陷或漏洞。威脅行為者可以使用這種類型的漏洞來攻擊系統。開發人員經常因為攻擊而意識到漏洞。

零鏡頭提示

提供 LLM 執行任務的指示，但沒有可協助引導任務的範例 (快照)。LLM 必須使用其預先訓練的知識來處理任務。零鏡頭提示的有效性取決於任務的複雜性和提示的品質。另請參閱少量擷取提示。

殭屍應用程式

CPU 和記憶體平均使用率低於 5% 的應用程式。在遷移專案中，通常會淘汰這些應用程式。

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。