使用 GenAI 為 Amazon Bedrock 建立知識庫

GenAI

NetApp
October 06, 2025

目錄

使用 GenAI 為 Amazon Bedrock 建立知識庫

開始使用

GenAI 知識庫快速入門

開始使用貴組織在 Amazon FSX for NetApp ONTAP 檔案系統上的資料，建立知識庫或 Amazon Q Business Connector 。例如 chatbot 之類的應用程式將存取此知識庫或 Connector ，為終端使用者提供以組織為中心的回應。


登入工作負載工廠

你需要 "在 Workload Factory 上建立帳戶"並使用以下任一方式登入 "主控台體驗"。


設定您的環境以符合 **GenAI** 需求

您需要 AWS 認證才能部署 AWS 基礎架構，部署並探索 ONTAP 檔案系統的 FSX ，您想要整合到知識庫或連接器中的資料來源清單，存取 Amazon bedrock AI 服務或 Amazon Q Business 應用程式等。

"深入瞭解 GenAI 需求"。


識別包含資料來源的 **ONTAP** 檔案系統的 **FSX**

您將整合至知識庫的資料來源可以位於單一適用於 ONTAP 檔案系統的 FSX 、或是多個適用於 ONTAP 檔案系統的 FSX 上。如果這些系統位於不同的 VPC ，則必須可在同一個網路中存取，或是必須使用與 AI 引擎相同的區域和 AWS 帳戶來連接 VPC 。

"瞭解如何識別資料來源"。


部署 **GenAI** 基礎架構

啟動基礎架構部署精靈，在 AWS 環境中部署 GenAI 基礎架構。此程序會部署 NetApp GenAI 引擎的 EC2 執行個體、以及 ONTAP 檔案系統的 FSX 上的磁碟區、以包含 NetApp AI 引擎資料庫。該 Volume 用於儲存知識庫所使用的向量資料庫。

"瞭解如何部署知識庫基礎架構"。

下一步

您現在可以建立知識庫、為終端使用者提供以組織為中心的回應。

GenAI 知識庫需求

在建置知識庫之前，請確保 Workload Factory 和 AWS 已正確設定。這包括擁有您的 AWS 登入憑證、已部署的 FSx for ONTAP檔案系統（其中包含您想要整合到知識庫中的資料來源）、存取 Amazon Bedrock AI 服務等等。

基本 GenAI 需求

GenAI 的一般需求是您開始使用前所需的環境。

工作負載工廠登入和帳戶

你需要 "在 Workload Factory 上建立帳戶"並使用以下任一方式登入 "主控台體驗"。

AWS 認證與權限

您需要將具有讀取/寫入權限的 AWS 憑證新增至 Workload Factory，這表示您將以讀取/寫入模式使用 Workload Factory 進行 GenAI。

目前不支援 _ 基本 _ 模式和 _ 唯讀 _ 模式權限。

設定認證時、如下所示選取權限、可讓您完全存取以管理 ONTAP 檔案系統的 FSX 、並部署及管理知識庫和 chatbot 所需的 GenAI EC2 執行個體和其他 AWS 資源。

"了解如何將 AWS 憑證新增至 Workload Factory"

GenAI 知識庫需求

如果您計畫使用知識庫，請確保您的環境符合下列需求。

Amazon bedrock

Amazon 基礎架構可讓您使用基礎模型、並提供建置泛用 AI 應用程式的功能。

在開始使用NetApp Workload Factory for GenAI 之前，您必須設定 Amazon Bedrock。您的 GenAI 部署必須位於啟用了 Amazon Bedrock 的 AWS 區域。

- "AWS 文件：設定 Amazon bedrock"
- "AWS 文件： Amazon bedrock 知識庫支援的區域和模型"

GenAI 預設會重新排列搜尋結果，以改善結果相關性。為獲得最佳結果，請確保 Amazon 基礎模型組態包括存取重新排名模型，例如 Cohere Rerank 或 Amazon Rerank （如果您所在地區有）。

內嵌模型

建立知識庫之前、您必須先啟用您計畫使用的內嵌模型。支援下列內嵌模型：

- Titan 嵌入式 G1 - 文字
- Titan 內嵌文字 v2
- Titan Multic形式 嵌入式 G1
- 內嵌英文
- 內嵌多國語言

"深入瞭解 Amazon Titan"

聊天模式

您必須先啟用您計畫使用的基礎聊天模式、才能建立知識庫。由於各 AWS 地區的機型支援不盡相同、請參閱 "AWS 文件" 、確認您可以在計畫部署知識庫的地區使用哪些機型。

GenAI 支援 Anthropic ， Amazon ， Mistral AI ， Meta ， Jamba 和 Cohere 等多種機型。

深入瞭解如何在 Amazon Bedrock 中使用這些模型：

- "Anthropic 在 Amazon bedrock 的 Claude"
- "在 Amazon Bedrock 主控台開始使用 Amazon Nova"
- "Mistral AI 機型"
- "Amazon Titan 文字模型"
- "中繼 Llama 機型"
- "Jamba 機型"
- "Cohere Command 模型"

適用於 **ONTAP** 檔案系統的 **FSX**

您至少需要一個適用於 ONTAP 檔案系統的 FSX ：

- NetApp GenAI 引擎將使用一個檔案系統（或建立一個檔案系統，如果不存在）來儲存知識庫所使用的向量資料庫。

  此適用於 ONTAP 檔案系統的 FSX 必須使用 FlexVol Volume 。不支援支援的支援。FlexGroup

- 一或多個檔案系統將包含您要整合至知識庫的資料來源。

  一個適用於 ONTAP 檔案系統的 FSX 可同時用於上述兩種用途、或者您可以將多個 FSX 用於 ONTAP 檔案系統。

- 您需要知道 AWS 區域、 VPC 和子網路、這些都是 AWS FSX for ONTAP 檔案系統所在的位置。檔案系統必須位於啟用 Amazon bedrock 的 AWS 區域。
- 您需要考慮要套用至屬於此部署一部分的 AWS 資源的標記金鑰 / 值配對（選用）。
- 您必須知道金鑰配對資訊、才能安全地連線至 NetApp AI 引擎執行個體。

"瞭解如何部署及管理適用於 ONTAP 檔案系統的 FSX"

識別要新增至知識庫或連接器的資料來源

識別或建立位於您的 FSX for ONTAP 檔案系統上的文件（資料來源）、這些文件將整合到您的知識庫中。這些資料來源可讓知識庫根據與組織相關的資料、為使用者查詢提供準確且個人化的答案。

資料來源的最大數量

支援的資料來源數量上限為 10 個。

資料來源的位置

資料來源可以儲存在單一磁碟區中、或儲存在磁碟區內的資料夾中、儲存在適用於 NetApp ONTAP 檔案系統的 Amazon FSX 上的 SMB 共用區或 NFS 匯出區中。資料來源也可以儲存在 Amazon FSX 上、以供 NetApp SnapMirror 資料保護關係中的 NetApp ONTAP 磁碟區使用。

您無法在磁碟區或資料夾中選取個別文件、因此您應確保包含資料來源的每個磁碟區或資料夾中、都不包含不應與知識庫整合的額外文件。

您可以將多個資料來源新增至每個知識庫、但這些資料來源都必須位於可從 AWS 帳戶存取的 ONTAP 檔案系統的 FSX 上。

每個資料來源的檔案大小上限為 50 MB 。

支援的傳輸協定

知識庫可支援使用 NFS 或 SMB/CIFS 通訊協定之磁碟區的資料。選取使用 SMB 傳輸協定儲存的檔案時、您必須輸入 Active Directory 資訊、知識庫才能存取這些磁碟區上的檔案。其中包括 Active Directory 網域、 IP 位址、使用者名稱和密碼。

在透過 SMB 存取的共用（檔案或目錄）上儲存資料來源時、資料只能由具有存取該共用權限的聊天機器人使用者或群組存取。啟用此「權限感知功能」時、 AI 系統會將驗證 0 中的使用者電子郵件與允許檢視或使用 SMB 共用上檔案的使用者進行比較。chatbot 將根據內嵌檔案的使用者權限提供答案。

例如、如果您將 10 個檔案（資料來源）整合到知識庫、其中 2 個檔案是包含受限資訊的人力資源檔案、則只有通過驗證以存取這 2 個檔案的聊天機器人程式使用者、才會收到來自聊天機器人程式的回應、其中包含來自這些檔案的資料。

支援的資料來源檔案格式

Workload Factory GenAI 知識庫目前支援以下資料來源檔案格式。

檔案格式	擴充
Apache Parquetfed腳 註：免責聲明 [將結構化資料檔案擷取至知識庫時，不支援資料 guardrails 功能。]	硬地板
以逗號分隔的值檔案說明：免責聲明 []	.csv
圖形交換格式	.gif
JPEG	.jpg or.jpeg
JSON 與 JSONP 註腳：免責聲明 []	json
請下標	.md
Microsoft Word	.doc 或 .docx
純文字	.txt
可攜式文件格式	.pdf
可攜式網路圖形	.png-
WebP 映像	webp

部署 **GenAI** 基礎架構

您必須先在環境中部署適用於 RAG 架構的 GenAI 基礎架構，才能為組織建置適用於 ONTAP 知識庫，連接器和應用程式的 FSX 。主要基礎架構元件為 Amazon 基礎架構服務、 NetApp GenAI 引擎的虛擬機器執行個體、以及 ONTAP 檔案系統的 FSX 。

部署的基礎架構可支援多個知識庫，閒聊機器人程式和連接器，因此您通常只需要執行一次此工作。

基礎架構詳細資料

您的 GenAI 部署必須位於啟用 Amazon 基礎的 AWS 區域。 "檢視支援區域的清單"

基礎架構包含下列元件。

Amazon bedrock 服務

Amazon bedrock 是一項完全託管的服務、可讓您透過單一 API 使用來自頂尖 AI 公司的基礎模型（ FMS ）。它也提供您建置安全泛用 AI 應用程式所需的功能。

"深入瞭解 Amazon bedrock"

Amazon Q Business

Amazon Q 以 Amazon 為基礎，提供完全託管的泛型 AI 助理，可用於回答問題，並根據資料來源的資訊產生內容。

"深入瞭解 Amazon Q Business"

NetApp GenAI 引擎的虛擬機器

NetApp GenAI 引擎會在此程序中部署。它提供從資料來源擷取資料的處理能力、然後將該資料寫入向量資料庫。

適用於 **ONTAP** 檔案系統的 **FSX**

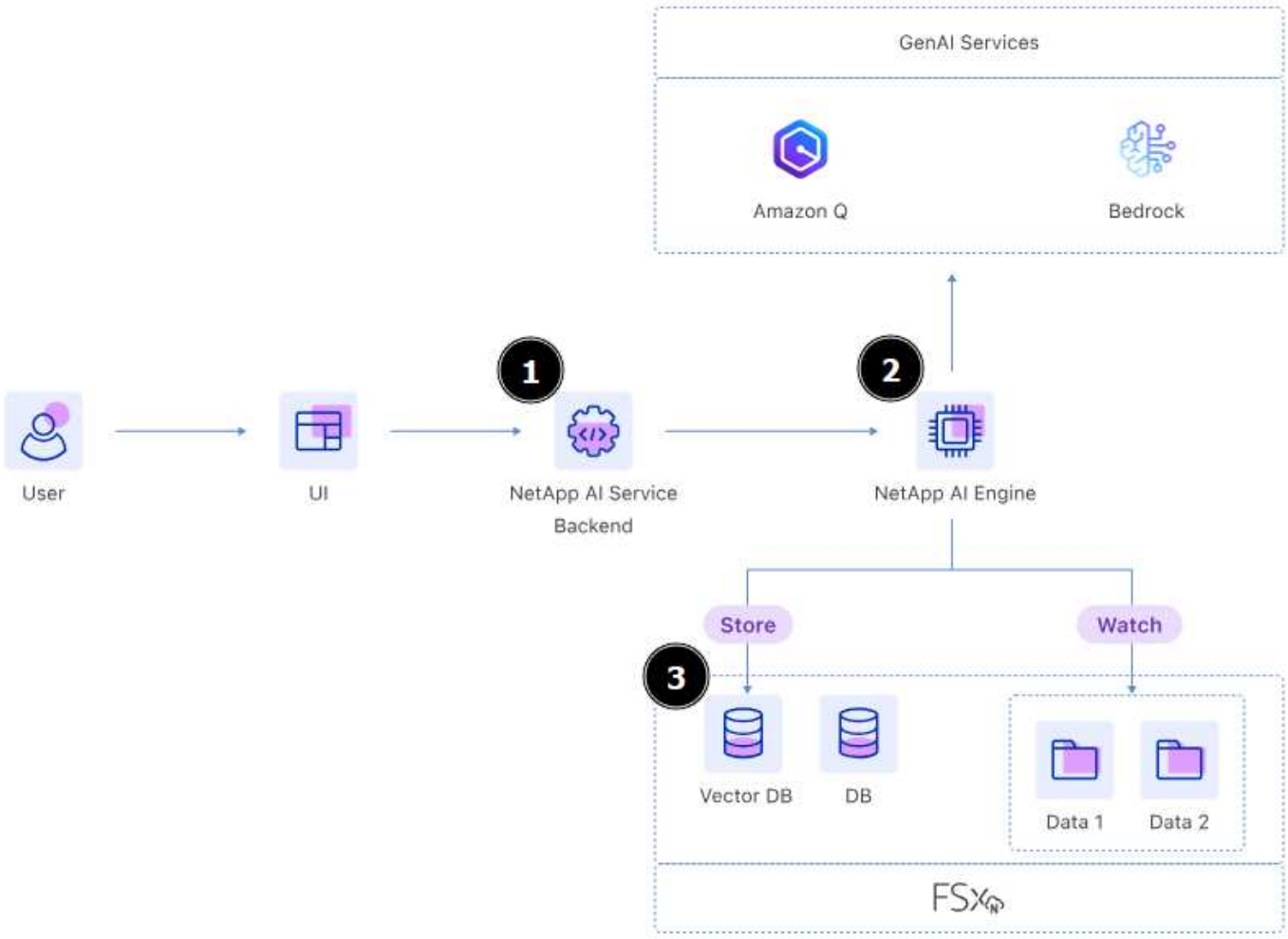
適用於 ONTAP 檔案系統的 FSX 可為您的 GenAI 系統提供儲存設備。

部署的單一 Volume 會包含向量資料庫、其中會儲存基礎模型根據您的資料來源所產生的資料。

您將整合至知識庫的資料來源可以位於 ONTAP 檔案系統的相同 FSX 或不同系統上。

NetApp GenAI 引擎會監控並與這兩個磁碟區互動。

下圖顯示 GenAI 基礎架構。編號 1 、 2 和 3 的元件會在此程序中部署。開始部署之前、必須先具備其他元素。



部署 **GenAI** 基礎架構

您需要輸入 AWS 認證資料，然後選取適用於 ONTAP 檔案系統的 FSX ，以部署擷取擴增產生（ RAG ）基礎架構。

開始之前

開始此程序之前，請確定您的環境符合知識庫或連接器的需求，視您選擇的類型而定。

- "知識庫需求"
- "連接器需求"

步驟

1. 使用以下方式之一登入 Workload Factory"主控台體驗" 。
2. 在 AI 工作負載方塊中、選取 * 部署與管理 * 。
3. 檢閱基礎架構圖表、然後選取 * 下一步 * 。
4. 完成「 * AWS 設定 * 」區段中的項目：
   a. * AWS 認證 * ：選取或新增 AWS 認證、以提供部署 AWS 資源的權限。
   b. * 位置 * ：選取 AWS 區域、 VPC 和子網路。

      GenAI 部署必須位於啟用 Amazon 基礎的 AWS 區域。 "檢視支援區域的清單"

5. 完成 * 基礎架構設定 * 區段中的項目：
   a. 標籤：輸入您想要套用於此部署的所有 AWS 資源的任何標籤鍵/值對。這些標籤在 AWS 管理控制台和 Workload Factory 內的基礎架構資訊區域中可見，可協助您追蹤 Workload Factory 資源。
6. 完成 **Connectivity** 部分：
   a. * 金鑰配對 * ：選取金鑰配對、讓您安全地連線至 NetApp GenAI 引擎執行個體。
7. 完成「 *AI 引擎 * 」一節：
   a. 實例名稱：可選地，選擇*定義實例名稱*並輸入 AI 引擎實例的自訂名稱。實例名稱出現在 AWS 管理控制台和 Workload Factory 內的基礎架構資訊區域，可協助您追蹤 Workload Factory 資源。
8. 選擇 * 部署 * 開始部署。

> ⓘ 如果部署失敗並出現認證錯誤，您可以選取錯誤訊息中的超連結，以取得進一步的錯誤詳細資料。您可以查看遺失或封鎖的權限清單，以及 GenAI 工作負載需要的權限清單，以便部署 GenAI 基礎架構。

結果

Workload Factory 開始部署聊天機器人基礎架構。此過程可能需要長達 10 分鐘。

在部署過程中、會設定下列項目：

- 網路與私有端點一起設定。
- 隨即建立 IAM 角色、執行個體設定檔和安全性群組。
- 已部署 GenAI 引擎的虛擬機器執行個體。
- Amazon bedrock 已設定為使用前置字元的記錄群組、將記錄傳送至 Amazon CloudWatch 記錄檔 `/aws/bedrock/`。
- GenAI 引擎配置為使用名為 `/netapp/wlmai/<tenancyAccountId>/randomId`， 在哪裡 `<tenancyAccountID>`是 "NetApp控制台帳戶 ID"對於當前用戶。

建立 GenAI 知識庫

部署 AI 基礎架構並確定將從 FSx for ONTAP資料儲存整合到知識庫中的資料來源後，您就可以使用 Workload Factory 建置知識庫了。作為此步驟的一部分，您還將定義 AI 特性並建立對話開場白。

請確保您的環境符合 "需求"for 知識庫，然後再繼續。

關於這項工作

知識庫有兩種資料整合模式： _ 公開模式 _ 和 _ 企業模式 _ 。

公共模式

您可以使用知識庫、而無需整合組織的資料來源。在這種情況下、與知識庫整合的應用程式只會提供來自網際網路上公開資訊的結果。這稱為 _public 模式 _ 整合。

企業模式

在大多數情況下、您會想要將組織的資料來源整合到知識庫中。這稱為 _ 企業模式 _ 整合、因為它能提供企業的知識。

您組織的資料來源可能包含個人識別資訊 (PII)。為了保護這些敏感訊息，您可以在建立和配置知識庫時啟用_資料護欄_。由NetApp資料分類提供支援的資料護欄可識別和屏蔽 PII，使其無法存取和復原。

"了解NetApp資料分類" 。

NetApp Workload Factory for GenAI 不會屏蔽敏感的個人資訊 (SPii)。參考"敏感個人資料的類型"有關此類數據的更多資訊。

數據護欄可以隨時啟用或停用。如果您切換資料護欄啟用，Workload Factory 將從頭開始掃描整個知識庫，這會產生成本。

建立及設定知識庫

知識庫定義了您想要用來建立知識庫的特性、例如基礎 AI 模型和內嵌格式。

步驟

1. 使用以下方式之一登入 Workload Factory"主控台體驗" 。
2. 在 AI 工作負載方塊中、選取 * 部署與管理 * 。
3. 從知識庫和連接器選單中，選擇*新建*下拉選單並選擇* NetApp GenAI 知識庫 for Bedrock*。
4. 在建立NetApp GenAI 知識庫頁面上，設定知識庫設定：

知識庫詳細資訊

1. * 名稱 * ：輸入您要用於知識庫的名稱。
2. * 說明 * ：輸入知識庫的詳細說明。
3. **Bedrock**：選擇您的 AWS 帳號可使用 Amazon Bedrock 的區域。

攝取

1. 嵌入模型：
   - 選擇一個嵌入模型用於知識庫。嵌入模型定義如何將資料轉換為知識庫的向量嵌入。 Workload Factory 支援以下模型：
   - Titan 嵌入式 G1 - 文字
   - Titan 內嵌文字 v2
   - Titan Multic形式 嵌入式 G1
   - 內嵌英文
   - 內嵌多國語言

     請注意、您必須已啟用 Amazon bedrock 的內嵌模型。

"深入瞭解 Amazon Titan"

- 如果適用，請選擇與所選嵌入模型的配置相符的推理類型。

2. 資料護欄：選擇是否要啟用或停用資料護欄。"了解由NetApp資料分類提供支援的資料護欄" 。

   必須符合下列先決條件，才能啟用資料欄。

   - 需要服務帳戶才能與NetApp資料分類通訊。您必須在NetApp控制台租賃帳戶上擁有「組織管理員」角色才能建立服務帳戶。具有組織管理員角色的成員可以完成組織中的所有操作。"了解如何在NetApp控制台中為成員新增角色"
   - AI 引擎必須能夠存取"NetApp控制台 API 端點"。
   - 您需要按照"NetApp資料分類文檔"：
     i. 建立控制台代理
     ii. 確保您的環境符合先決條件
     iii. 部署NetApp資料分類

   擷取 CSV ， JSON ， JSONP 或 Parquet 等結構化資料檔案時，不支援資料欄功能。

聊天和檢索設定

1. 聊天模型：
   - 從 Amazon Bedrock 整合的各種聊天模型中進行選擇。請注意，您必須已經啟用來自 Amazon Bedrock 的聊天模型。
   - 如果適用，請選擇與所選模型的配置相符的推理類型。
2. 聊天設定：
   - 為聊天機器人選擇一個溫度來配置回應的隨機性和創造性。較低的溫度會導致更可預測的反應，而較高的溫度會導致更多樣化的反應。
   - 選擇最大響應長度來配置響應的詳細程度。回應長度越長，使用的回應令牌就越多，並且會產生更高的成本。
3. 思考模式：啟用思考模式後，聊天機器人將花費更多時間來處理查詢，結果通常會更準確。當您啟用思考模式時，您可以控制在產生結果時使用多少個推理標記。使用更多的推理標記可以獲得更準確的回應，但可能會產生更高的成本。
4. 重新排名：啟用或停用重新排名，這可以提高查詢結果的相關性和品質。選擇標準聊天模型或專門的重新排名模型用於重新排名。僅當您所在地區可用時才會顯示 Reranker 模型選項。選擇與所選模型的配置相符的推理類型。
5. * 對話開場白 * ：選擇是否要提供最多四個對話啟動器提示、讓與使用此知識庫的聊天機器人程式互動的使用者看到。建議您啟用此設定。

   如果您啟動交談啟動器、預設會選取「自動模式」。只有在您將資料來源新增至知識庫之後、才能啟用「手動模式」。"瞭解如何修改知識庫設定"。

儲存定義

1. **FSx for ONTAP**檔案系統：當您定義新的知識庫時，Workload Factory 會建立一個新的Amazon FSx for

NetApp ONTAP磁碟區來儲存它。選擇將在其中建立新磁碟區的現有檔案系統名稱和 SVM（也稱為儲存 VM）。

2. 快照策略：從 Workload Factory 儲存清單中定義的現有策略清單中選擇快照策略。知識庫的定期快照將根據您選擇的快照策略以一定頻率自動建立。
3. **S3** 儲存桶：如果聊天機器人查詢結果包含結構化數據，GenAI 可以將結果儲存在 S3 儲存桶中。若要使用此功能，請啟用*啟動 S3 儲存桶*設定並從清單中選擇與您的帳戶關聯的 S3 儲存桶。當這些結果儲存在 S3 儲存桶中時，您可以使用聊天會話中的下載連結下載它們。

   如果您需要的快照原則不存在、您可以 "建立快照原則"在包含該 Volume 的儲存 VM 上執行。

4. 選取 * 建立知識庫 * 、將知識庫新增至 GenAI 。

   建立知識庫時會出現進度指標。

   建立知識庫之後、您可以選擇將資料來源新增至新的知識庫、或在不新增資料來源的情況下結束程序。建議您選擇 * 新增資料來源 * 、然後立即新增一或多個資料來源。

將資料來源新增至知識庫

您可以新增一或多個資料來源、以便將組織的資料填入知識庫。

關於這項工作

支援的資料來源數量上限為 10 個。

步驟

1. 選擇*新增資料來源*後，選擇要新增的資料來源類型：
   - 新增 FSx for ONTAP 檔案系統（使用現有 FSx for ONTAP 磁碟區中的檔案）
   - 新增檔案系統（使用來自通用 SMB 或 NFS 共享的檔案）

新增 **FSx for ONTAP** 檔案系統

1. * 選取檔案系統 * ：選取資料來源檔案所在的 ONTAP 檔案系統的 FSX ，然後選取 * 下一步 * 。
2. * 選取磁碟區 * ：選取資料來源檔案所在的磁碟區、然後選取 * 下一步 * 。

   選取使用 SMB 傳輸協定儲存的檔案時、您需要輸入 Active Directory 資訊、其中包括網域、 IP 位址、使用者名稱和密碼。

3. * 選取資料來源 * ：根據您儲存檔案的位置選取資料來源位置。這可以是整個磁碟區、或只是磁碟區中的特定資料夾或子資料夾、然後選取 * 下一步 * 。
4. * 組態 * ：設定資料來源如何從檔案中擷取資訊，以及其包含在掃描中的檔案：

   ◦ * 定義資料來源 * ：在 * 區塊策略 * 區段中，定義當資料來源與知識庫整合時， GenAI 引擎如何將資料來源內容分割成區塊。您可以選擇下列其中一個策略：

   - * 多重句子區塊 * ：將資料來源中的資訊組織成句子定義的區塊。您可以選擇每個區塊中包含多少句話（最多 100 句）。
   - * 重疊區塊 * ：將資料來源中的資訊組織成字元定義區塊、以重疊鄰近區塊。您可以選擇每個區塊的字元大小、以及每個區塊與相鄰區塊重疊的量。您可以設定 50 到 3000 個字元之間的區塊大小、以及介於 1 到 99% 之間的重疊百分比。

     選擇高重疊百分比可大幅增加儲存需求、只需稍微改善擷取準確度。

   ◦ * 檔案篩選 * ：設定掃描中包含哪些檔案：

   - 在「 * 檔案類型支援 * 」區段中，選擇要包含所有類型的檔案，或選擇要包含在資料來源掃描中的個別檔案類型。

     如果您包含圖像或 PDF 文件， NetApp Workload Factory for GenAI 會解析圖像中的文字（包括 PDF 文件中的圖像），這會產生更高的成本。

   當包含影像的文字資料時，當掃描的文字資料從您的環境傳送至 AWS 時， GenAI 無法從影像中遮罩個人識別資訊（ PII ）。然而，一旦儲存資料， GenAI 資料庫就會隱藏所有 PII 。

   您選擇在掃描中包含影像檔案，與知識庫聊天模式有關。如果您在掃描中包含影像檔案，則聊天模式必須支援影像。如果在此選取映像檔案類型，您就無法將知識庫切換至不支援映像檔案的聊天模式。

   ◦ 在 * 檔案修改時間篩選器 * 區段中，選擇根據檔案的修改時間來啟用或停用檔案的包含。如果啟用修改時間篩選，請從清單中選取日期範圍。

   如果您根據修改日期範圍來包含檔案，只要日期範圍不滿足（檔案尚未在您指定的日期範圍內修改），檔案就會排除在定期掃描之外，而且資料來源也不會包含這些檔案。

5. 在 * 權限感知 * 區段中，只有當您選取的資料來源位於使用 SMB 通訊協定的磁碟區上時，才能使用此區段，您可以啟用或停用權限感知回應：

   ◦ * 已啟用 * ：存取此知識庫的聊天機器人程式使用者只能從其存取的資料來源取得查詢回應。

   ◦ * 停用 * ：聊天機器人程式的使用者將會使用所有整合式資料來源的內容接收回應。

6. 選取 * 新增 * 將此資料來源新增至您的知識庫。

新增通用 **NFS** 檔案系統

1. 選擇檔案系統：輸入資料來源檔案所在的檔案系統主機的 IP 位址或 FQDN，選擇網路共用的 NFS 協議，然後選擇*下一步*。
2. * 選取資料來源 * ：根據您儲存檔案的位置選取資料來源位置。這可以是整個磁碟區、或只是磁碟區中的特定資料夾或子資料夾、然後選取 * 下一步 * 。

   在某些情況下，您可能需要手動輸入 NFS 匯出名稱，然後選擇「擷取目錄」以顯示可用目錄。您可以選擇整個匯出，或僅選擇匯出中的特定資料夾。

3. * 組態 * ：設定資料來源如何從檔案中擷取資訊，以及其包含在掃描中的檔案：

   ◦ * 定義資料來源 * ：在 * 區塊策略 * 區段中，定義當資料來源與知識庫整合時， GenAI 引擎如何將資料來源內容分割成區塊。您可以選擇下列其中一個策略：

   - * 多重句子區塊 * ：將資料來源中的資訊組織成句子定義的區塊。您可以選擇每個區塊中包含多少句話（最多 100 句）。
   - * 重疊區塊 * ：將資料來源中的資訊組織成字元定義區塊、以重疊鄰近區塊。您可以選擇每個區塊的字元大小、以及每個區塊與相鄰區塊重疊的量。您可以設定 50 到 3000 個字元之間的區塊大小、以及介於 1 到 99% 之間的重疊百分比。

     選擇高重疊百分比可大幅增加儲存需求、只需稍微改善擷取準確度。

   ◦ * 檔案篩選 * ：設定掃描中包含哪些檔案：

   - 在「 * 檔案類型支援 * 」區段中，選擇要包含所有類型的檔案，或選擇要包含在資料來源掃描中的個別檔案類型。

     如果您包含圖像或 PDF 文件， NetApp Workload Factory for GenAI 會解析圖像中的文字（包括 PDF 文件中的圖像），這會產生更高的成本。

   當包含影像的文字資料時，當掃描的文字資料從您的環境傳送至 AWS 時， GenAI 無法從影像中遮罩個人識別資訊（ PII ）。然而，一旦儲存資料， GenAI 資料庫就會隱藏所有 PII 。

   您選擇在掃描中包含影像檔案，與知識庫聊天模式有關。如果您在掃描中包含影像檔案，則聊天模式必須支援影像。如果在此選取映像檔案類型，您就無法將知識庫切換至不支援映像檔案的聊天模式。

   ◦ 在 * 檔案修改時間篩選器 * 區段中，選擇根據檔案的修改時間來啟用或停用檔案的包含。如果啟用修改時間篩選，請從清單中選取日期範圍。

   如果您根據修改日期範圍來包含檔案，只要日期範圍不滿足（檔案尚未在您指定的日期範圍內修改），檔案就會排除在定期掃描之外，而且資料來源也不會包含這些檔案。

4. 選擇*新增資料來源*將此資料來源新增至您的知識庫。

新增通用 **SMB** 檔案系統

1. 選擇檔案系統：
   a. 輸入資料來源檔案所在的檔案系統主機的 IP 位址或 FQDN。
   b. 為網路共享選擇 SMB 協定。
   c. 輸入 Active Directory 訊息，包括網域、IP 位址、使用者名稱和密碼。
   d. 選擇*下一步*。
2. * 選取資料來源 * ：根據您儲存檔案的位置選取資料來源位置。這可以是整個磁碟區、或只是磁碟區中的特定資料夾或子資料夾、然後選取 * 下一步 * 。

   在某些情況下，您可能需要手動輸入 SMB 共享名稱，然後選擇「檢索目錄」以顯示可用目錄。您可以選擇整個共享，或僅選擇共享中的特定資料夾。

3. * 組態 * ：設定資料來源如何從檔案中擷取資訊，以及其包含在掃描中的檔案：
   - * 定義資料來源 * ：在 * 區塊策略 * 區段中，定義當資料來源與知識庫整合時， GenAI 引擎如何將資料來源內容分割成區塊。您可以選擇下列其中一個策略：
     - * 多重句子區塊 * ：將資料來源中的資訊組織成句子定義的區塊。您可以選擇每個區塊中包含多少句話（最多 100 句）。
     - * 重疊區塊 * ：將資料來源中的資訊組織成字元定義區塊、以重疊鄰近區塊。您可以選擇每個區塊的字元大小、以及每個區塊與相鄰區塊重疊的量。您可以設定 50 到 3000 個字元之間的區塊大小、以及介於 1 到 99% 之間的重疊百分比。

       選擇高重疊百分比可大幅增加儲存需求、只需稍微改善擷取準確度。

   - 權限感知：啟用或停用權限感知回應：
     - * 已啟用 * ：存取此知識庫的聊天機器人程式使用者只能從其存取的資料來源取得查詢回應。
     - * 停用 * ：聊天機器人程式的使用者將會使用所有整合式資料來源的內容接收回應。
   - * 檔案篩選 * ：設定掃描中包含哪些檔案：
     - 在「 * 檔案類型支援 * 」區段中，選擇要包含所有類型的檔案，或選擇要包含在資料來源掃描中的個別檔案類型。

       如果您包含圖像或 PDF 文件， NetApp Workload Factory for GenAI 會解析圖像中的文字（包括 PDF 文件中的圖像），這會產生更高的成本。

     當包含影像的文字資料時，當掃描的文字資料從您的環境傳送至 AWS 時， GenAI 無法從影像中遮罩個人識別資訊（ PII ）。然而，一旦儲存資料， GenAI 資料庫就會隱藏所有 PII 。

     您選擇在掃描中包含影像檔案，與知識庫聊天模式有關。如果您在掃描中包含影像檔案，則聊天模式必須支援影像。如果在此選取映像檔案類型，您就無法將知識庫切換至不支援映像檔案的聊天模式。

   - 在 * 檔案修改時間篩選器 * 區段中，選擇根據檔案的修改時間來啟用或停用檔案的包含。如果啟用修改時間篩選，請從清單中選取日期範圍。

如果您根據修改日期範圍來包含檔案，只要日期範圍不滿足（檔案尚未在您指定的日期範圍內修改），檔案就會排除在定期掃描之外，而且資料來源也不會包含這些檔案。

4. 選擇*新增資料來源*將此資料來源新增至您的知識庫。

結果

資料來源開始內嵌到您的知識庫中。資料來源完全內嵌時、狀態會從「內嵌」變更為「內嵌」。

將單一資料來源新增至知識庫之後、您可以在聊天機器人程式模擬器視窗中進行本機測試、並在將聊天機器人程式提供給使用者之前進行任何必要的變更。您也可以依照相同步驟、將其他資料來源新增至知識庫。

測試 GenAI 知識庫

建立知識庫之後、您就可以使用 chatbot 模擬器在本機進行測試、並在您透過 chatbot 應用程式將知識庫提供給使用者之前進行任何必要的變更。

關於這項工作

您可以測試知識庫、確保其效能如預期、也可以自訂此知識庫的聊天機器人使用者預設可使用的對話啟動器。chatbot 模擬器會針對內嵌於知識庫中的所有資料來源執行。

您可以在聊天機器人模擬器中與內嵌資料來源聊天、以測試知識庫。請注意、在本機測試知識庫時、 GenAI 向量資料庫不會擷取任何互動或深入分析。

在為使用者在應用程式中部署知識庫之前，您將在 Workload Factory 中執行大部分測試。如果您需要變更資料來源或聊天機器人操作，則需要在發布知識庫之前進行變更。

您可以調整聊天機器人模擬器視窗的大小並重新設定標題，並將問題和回應複製到剪貼簿。

測試您的聊天機器人程式時、需要執行的一些工作包括：

- 輸入大量與組織相關的問題、以確保答案符合預期。
- 自訂您想要在預設情況下在 chatbot 應用程式中供使用者使用的交談啟動器。
- 請確定在聊天機器人程式答案底部提供的歸屬內容包含正確的參考資料。

步驟

1. 從「知識庫」清查頁面中、選取您要測試的知識庫。

   chatbot 模擬器會出現在右窗格中。如果已定義、也會顯示現有的對話啟動器。

2. 在 chatbot 項目欄位中、輸入提示或問題、然後選取 以查看您的 chatbot 如何回應您的組織知識。

- 您可以在回應下方展開 * 來源 * 清單，以查看用來產生答案的來源。這會提供用來產生答案的檔案清單。您可以將游標移到檔案名稱上方，以檢視並複製從每個檔案和磁碟區路徑所使用的資料區塊到每個檔案。
- 如果答案中包含表格，您可以對每列中的資料進行排序，然後將每個表格複製到剪貼簿。
- 如果答案結果包含結構化數據，而知識庫啟用了 **S3 Bucket** 功能，GenAI 會將結果儲存在 S3 Bucket 中。您可以使用聊天會話中的*下載結果*連結從儲存桶中下載結果。

3. 如果您需要更新任何資料來源、以便知識庫提供更集中的答案、請立即進行這些變更、然後重新測試知識庫。

啟動 GenAI 知識庫的外部驗證

啟動知識庫的驗證、以便在使用 API 端點將知識庫與 chatbot 應用程式整合時、需要權杖驗證和 ACL 。當您啟動驗證時、您可以設定 JSON Web Token 的設定、用於從 chatbot 用戶端向知識庫提出 API 要求。

步驟

1. 使用以下方式之一登入 Workload Factory"主控台體驗" 。
2. 在 AI 工作負載方塊中、選取 * 部署與管理 * 。
3. 從「知識庫」清查頁面中、選取您要啟動驗證的知識庫。
4. 選取 ••• 並選取 * 管理知識庫 * 。
5. 選取 * 動作 * 功能表、然後選取 * 管理驗證設定 * 。
6. 設定驗證：
   a. 選取 * 啟動驗證設定 * 。
   b. 提供必要資訊。提供範例、但您應該從驗證供應商處取得這些欄位的值：
      - * 演算法 * ：驗證提供者使用的簽署演算法。
      - * 目標對象 * （選用）：包含目標權杖收件者的字串（有時為 URL ）。
      - * 發卡行 * ：識別發出權杖之提供者的字串。

        例如、 Amazon Cognito 使用具有下列格式的發卡行字串：

        ```
        https://cognito-idp-<region>.amazonaws.com/<UserPoolID>
        ```

        其中 `<region>` 是包含使用者集區的 AWS 區域、 `<UserPoolID>` 也是您的使用者集區 ID 。您可以使用下列命令擷取使用者集區 ID ：

        ```
        aws cognito-idp list-user-pools --max-results=60 --output=table
        ```

      - **JWKS URI** ：提供驗證此 Token 簽章所需之公開金鑰的 URI 字串。

例如、 Amazon Cognito 使用以下格式的 JWKS URI 字串：

```
https://cognito-idp.<region>.amazonaws.com/<userPoolId>/.well-known/jwks.json
```

+ 其中 `<region>` 是包含使用者集區的 AWS 區域、 `<UserPoolID>` 也是您的使用者集區 ID 。您可以使用下列命令擷取使用者集區 ID ：

```
aws cognito-idp list-user-pools --max-results=60 --output=table
```

7. 選擇*保存*。

結果

知識庫的驗證現已啟用、您可以使用 API 端點與知識庫互動、並將知識庫與 chatbot 應用程式整合。

發佈 GenAI 知識庫，並檢視獨特的端點

在您在本機建立並測試知識庫之後、您可以發佈知識庫、以便將其整合至能夠讓使用者查詢知識庫的聊天機器人應用程式。

關於這項工作

發布知識庫使您能夠在聊天應用程式中使用它。發布操作觸發 Workload Factory API 產生並發布唯一端點。發布後，知識庫可供聊天應用程式訪問，並且 API 端點已準備好進行整合。

您發佈的每個知識庫都有獨特的端點。

步驟

1. 使用以下方式之一登入 Workload Factory"主控台體驗" 。
2. 在 AI 工作負載方塊中、選取 * 部署與管理 * 。
3. 從「知識庫」清查頁面中、選取您要發佈的知識庫。
4. 選取 ••• 並選取 * 管理知識庫 * 。

   此頁面會顯示發佈的狀態、資料來源的內嵌狀態、內嵌模式、以及所有內嵌資料來源的清單。

5. 選取 * 動作 * 功能表、然後選取 * 發佈 * 。

   Workload Factory 發布知識庫。在知識庫的詳細資訊頁面上，狀態從*未發布*變更為*已發布*。

   您現在可以取得知識庫專屬端點的詳細資料。

6. 在「已發佈」狀態旁邊、選取 * 檢視 * 。

   顯示有關如何使用 Workload Factory API 存取知識庫的詳細資訊。

7. 從「 * 檢視發佈的資訊 * 」對話方塊中、複製可用於將知識庫與應用程式整合的 API 端點。

若要深入瞭解 API 端點、請移至 "API 文件" 、然後選取「 * AI > 外部 * 」。

您必須先從驗證提供者取得使用者權杖、才能使用這些端點。

結果

您現在擁有已發佈的知識庫和獨特的端點、可用來將知識庫與 chatbot 應用程式整合。

使用 GenAI 外部範例 chatbot 應用程式

配置、啟動和發布知識庫後，外部應用程式開發人員可以配置和運行NetApp提供的開源範例聊天機器人應用程序，以便與您的知識庫進行交互，並了解如何使用 Workload Factory API 創建自己的生成式 AI 應用程式。

步驟

1. "建立知識庫"。
2. "啟動驗證" 以瞭解您建立的知識庫。

   這可讓知識庫驗證 API 要求、並在使用 API 端點時要求權杖驗證和 ACL 。

   與此知識庫整合的外部聊天應用程式需要使用您在知識庫驗證設定中所設定的相同驗證提供者（發卡行）。

3. "發佈知識庫" 啟用外部應用程式的 API 存取。

   知識庫發佈後、即可從外部存取 API 端點、您可以將知識庫與外部聊天應用程式（例如、 chatbot 應用程式範例）整合。

4. 從下載範例 chatbot 應用程式套件 "GitHub"。
5. 請依照套件中的 README 檔案中的指示、安裝並執行 chatbot 應用程式。
6. 瀏覽至 "http://localhost:9091" 以登入應用程式。

   此時會出現 chatbot 應用程式範例。

深入瞭解

"工作負載工廠 API 文檔"

建立 RAG 型 GenAI 應用程式

在您建立知識庫並測試自己的聊天機器人程式之後、您就可以開始設定應用程式、讓使用者能夠查詢聊天機器人程式。

"瞭解如何在適用於 ONTAP 的 FSX 上建立以 RAG 為基礎的 AI 應用程式"

GenAI 的後續功能

現在您已經使用企業資料建立知識庫、並將其部署給使用者、您可以管理知識庫、資料來源和 RAG 基礎架構、包括 ONTAP 檔案系統的 FSX 。

管理知識庫元件時、您可以執行的一些工作包括：

- 更新資料來源的內容、或新增資料來源、並將這些變更與您的知識庫和聊天機器人同步。
- 管理資料來源設定、包括區塊策略和權限認知（ SMB 檔案存取）。
- 管理您的知識庫設定、包括聊天模式和對話啟動器。
- 進行變更後、請取消發佈知識庫或重新發佈知識庫。
- 備份並保護適用於 ONTAP 檔案系統的 FSX 上的重要資料、確保您的知識庫資料和其他基礎架構元件隨時可用。

有關管理 FSx for ONTAP檔案系統的信息，請訪問 "Amazon FSx for NetApp ONTAP的 Workload Factory 文檔" 查看您可以使用的備份和保護功能。