



BlueXP Workload Factory for GenAI 文件 GenAI

NetApp
June 30, 2025

This PDF was generated from <https://docs.netapp.com/zh-tw/workload-genai/index.html> on June 30, 2025. Always check docs.netapp.com for the latest.

目錄

BlueXP Workload Factory for GenAI 文件	1
版本資訊	2
GenAI BlueXP 工作負載工廠的新功能	2
2025 年 6 月 29 日	2
2025 年 6 月 03 日	2
2025 年 04 月 5 日	2
2025 年 02 月 2 日	3
2025 年 1 月 05 日	4
2024 年 12 月 01 日	5
2024 年 11 月 3 日	5
2024 年 9 月 29 日	5
2024 年 9 月 1 日	6
2024 年 8 月 4 日	6
2024 年 7 月 7 日	6
瞭解 GenAI 的 BlueXP 工作負載工廠	7
瞭解 GenAI 的 BlueXP 工作負載工廠	7
什麼是 GenAI 的 BlueXP 工作負載工廠？	7
使用 GenAI 建立泛用 AI 應用程式的優點	7
GenAI 的運作方式	8
GenAI 的 BlueXP 工作負載工廠如何協助建置泛用 AI 應用程式	8
使用工作負載工廠的工具	9
成本	9
授權	9
NetApp GenAI 引擎的元件	9
使用 GenAI 為 Amazon Bedrock 建立知識庫	16
開始使用	16
GenAI 知識庫快速入門	16
GenAI 知識庫需求	16
識別要新增至知識庫或連接器的資料來源	18
部署 GenAI 基礎架構	19
建立 GenAI 知識庫	22
建立及設定知識庫	23
將資料來源新增至知識庫	24
測試 GenAI 知識庫	28
啟動 GenAI 知識庫的外部驗證	29
發佈 GenAI 知識庫，並檢視獨特的端點	30
使用 GenAI 外部範例 chatbot 應用程式	31
深入瞭解	31
建立 RAG 型 GenAI 應用程式	31

GenAI 的後續功能	31
使用 GenAI 為 Amazon Q Business 建立連接器	33
開始使用	33
GenAI 連接器快速入門	33
GenAI 連接器需求	33
識別要新增至連接器的資料來源	35
部署 GenAI 基礎架構	36
為 Amazon Q Business 建立 NetApp 連接器	38
定義連接器	39
將資料來源新增至連接器	40
管理與監控	42
管理 GenAI 基礎架構	42
檢視基礎架構的相關資訊	42
移除基礎架構	42
管理 GenAI 知識庫	43
檢視知識庫的相關資訊	43
編輯知識庫	43
利用快照保護知識庫	44
將其他資料來源新增至知識庫	46
將資料來源與知識庫同步	50
建立知識庫之前、請先評估聊天模式	50
解除發佈您的知識庫	51
刪除知識庫	51
管理 Amazon Q Business 連接器	52
檢視連接器的相關資訊	52
編輯連接器	52
新增其他資料來源至連接器	53
將資料來源與連接器同步	57
刪除連接器	57
管理 GenAI 資料來源	58
檢視資料來源的相關資訊	58
編輯資料來源設定	58
更新現有資料來源的內容	59
刪除資料來源	59
使用 BlueXP 工作負載工廠中的 Tracker 監控工作負載操作	59
追蹤及監控作業	60
檢視 API 要求	61
重試失敗的作業	61
編輯並重試失敗的作業	61
知識與支援	62
註冊以取得 GenAI BlueXP 工作負載工廠的支援	62

支援登錄總覽	62
註冊您的帳戶以取得 NetApp 支援	62
GenAI 疑難排解	64
常見問題與解決方案	64
取得 GenAI BlueXP 工作負載工廠的協助	67
取得適用於 ONTAP 的 FSX 支援	67
使用自我支援選項	67
利用 NetApp 支援建立案例	67
管理支援案例（預覽）	70
GenAI 法律聲明的 BlueXP 工作負載工廠	73
版權	73
商標	73
專利	73
隱私權政策	73
開放原始碼	73

BlueXP Workload Factory for GenAI 文件

版本資訊

GenAI BlueXP 工作負載工廠的新功能

瞭解工作負載工廠的通用 AI 工作負載功能有哪些新功能。

2025 年 6 月 29 日

支援通用 **NFS/SMB** 檔案系統上託管的資料來源

現在您可以從通用 SMB 或 NFS 共用中新增資料來源。這使您能夠包含儲存在 Amazon FSx for NetApp ONTAP 以外的檔案系統所託管的磁碟區上的檔案。

["在知識庫中新增資料來源"](#)

["新增資料來源"](#)

2025 年 6 月 03 日

Tracker 可用於監控及追蹤作業

GenAI 現已提供 Tracker 監控功能。您可以使用 Tracker 監控和追蹤待處理、正在進行和已完成操作的進度和狀態，查看操作任務和子任務的詳細信息，診斷任何問題或故障，編輯失敗操作的參數，以及重試失敗的操作。

["使用 BlueXP 工作負載工廠中的 Tracker 監控工作負載操作"](#)

為知識庫選擇重新排序模型

現在，您可以透過選擇與知識庫配合使用的特定重排序模型來提高重排序查詢結果的相關性。GenAI 支援 Cohere Rerank 和 Amazon Rerank 模型。

["建立 GenAI 知識庫"](#)

2025 年 04 月 5 日

支援 **NetApp Connector for Amazon Q Business**

此版本的 GenAI 引入了對 NetApp Connector for Amazon Q Business 的支持，使您能夠為 Amazon Q Business 建立連接器。與為 Amazon Bedrock 建立 GenAI 知識庫相比，較少的初始組態，可快速輕鬆地利用 Amazon Q Business AI 助理。

["為 Amazon Q Business 建立 NetApp 連接器"](#)

增強的聊天模式支援

GenAI 現在支援下列額外的聊天模式以供知識庫使用：

- ["Mistral AI 機型"](#)

- ["Amazon Titan 文字模型"](#)
- ["中繼 Llama 機型"](#)
- ["Jamba 1.5 機型"](#)
- ["Cohere Command 模型"](#)
- ["Deepseek 機型"](#)

GenAI 支援 Amazon Bedrock 支援的每個供應商機型：["Amazon 基礎架構支援的基礎模型"](#)

["建立 GenAI 知識庫"](#)

更新的權限術語

工作負載工廠使用者介面和文件現在使用“只讀”來指讀取權限，使用“讀取/寫入”來指稱自動化權限。= 2025 年 3 月 02 日

嵌入式聊天機器人程式增強功能

您現在可以直接將問題和回應複製到剪貼簿，調整聊天視窗的大小，以及變更其標題。此外，聊天回應現在可以包含表格，這些表格也可以複製。

["測試 GenAI 知識庫"](#)

聊天回應引文支援

聊天回應現在包含的引文會列出用來產生回應的檔案和資料區塊。

["測試 GenAI 知識庫"](#)

增強的檔案類型支援

此版本的 GenAI 提供增強的檔案支援：

- 聊天模式提供更好的 CSV 支援。這可在從 CSV 檔案查詢資料時，提供更有用的回應。
- GenAI 現在可以從資料來源擷取 Apache Parquet 檔案。
- GenAI 現在支援擷取包含影像的 Microsoft Word DOCX 檔案。系統會掃描 DOCX 文件中內嵌的影像，並在回應知識庫查詢時，納入內嵌影像的文字洞見。

["支援的資料來源檔案格式"](#)

2025 年 02 月 2 日

支援 **Amazon Nova** 基礎模式

GenAI 現在支援 Amazon Nova 基礎模式。支援 Amazon Nova Micro，Amazon Nova Lite 和 Amazon Nova Pro。

["GenAI 需求"](#)

資料來源的檔案類型篩選

GenAI 現在支援在新增資料來源時，選取要納入資料來源掃描的特定檔案類型。

["將資料來源新增至知識庫"](#)

資料來源的檔案修改日期篩選

GenAI 現在支援篩選檔案，可在新增資料來源時，依修改日期納入資料來源掃描。您可以為包含的檔案選擇修改日期範圍。

["將資料來源新增至知識庫"](#)

支援影像檔案，並加強支援 PDF 檔案

GenAI 現在支援強化對知識庫查詢的回應，從影像和圖表說明以及文件文字中獲得深入見解，進而提供更豐富，更高品質的答案。GenAI 現在可以掃描 PDF 檔案中的影像檔案和影像（也稱為多模式檔案支援）。如果您選擇掃描影像或 PDF 檔案，則會將影像中的文字（包括內嵌於 PDF 文件中的影像）掃描至資料來源，而掃描結果中的洞見則會包含在知識庫查詢的回應中。

["將資料來源新增至知識庫"](#)

混合式搜尋與重新部署支援

GenAI 現在可以使用混合式搜尋並重新排列結果的順序，大幅提升搜尋結果的相關性和準確度。混合式搜尋結合了傳統關鍵字型搜尋的優勢，以及進階密集向量型語義搜尋技術。標準關鍵字搜尋結果會以接近的相符項目和語言差異加以擴充，以增強相關性。然後 GenAI 使用進階重新排名模型（例如 Cohere Rerank 和 Amazon Rerank）進一步精簡這些結果，並傳回最相關的結果。這項功能適用於新建立的知識庫。

["瞭解 GenAI 的 BlueXP 工作負載工廠"](#)

2025 年 1 月 05 日

自訂快照名稱

您現在可以為隨機操作快照提供快照名稱。

["利用快照保護知識庫"](#)

自訂 AI 引擎執行個體名稱

您現在可以在部署期間為 AI 引擎執行個體提供自訂名稱。

["部署 GenAI 基礎架構"](#)

重建毀損或遺失的 GenAI 基礎架構

如果您的 AI 引擎執行個體毀損或遭到刪除，您可以讓工作負載在出廠時重新建置。工作負載原廠會在重建完成後，自動將您的知識庫重新附加到基礎架構，以便準備好使用。

["疑難排解"](#)

2024 年 12 月 01 日

從快照複製知識庫

GenAI 的 BlueXP 工作負載工廠現在支援從快照複製知識庫。這可快速恢復知識庫，並利用現有資料來源建立新的知識庫，有助於資料恢復與開發。

["複製知識庫"](#)

內部部署 ONTAP 叢集探索與複寫

探索內部部署的 ONTAP 叢集資料，並將其複寫至適用於 ONTAP 檔案系統的 FSX，以便用於豐富 AI 知識庫。所有內部部署探索和複寫工作流程都可以從儲存設備詳細目錄中的全新 * 內部部署 ONTAP (*) 標籤中進行。

["探索內部部署 ONTAP 的叢集"](#)

2024 年 11 月 3 日

使用資料欄遮罩個人識別資訊

泛用型 AI 工作負載採用 BlueXP 分類技術、引進資料 guardrails 功能。資料安全欄功能可識別及遮罩個人識別資訊 (PII)、協助您維持法規遵循、並強化敏感組織資料的安全性。

["建立 GenAI 知識庫"](#)

["瞭解 BlueXP 分類"](#)

2024 年 9 月 29 日

快照與還原支援知識庫磁碟區

您現在可以取得知識庫的時間點複本、以保護您的泛用 AI 工作負載資料。這可讓您保護資料、避免意外遺失或測試知識庫設定的變更。您可以隨時還原舊版的知識庫磁碟區。

["拍攝知識庫磁碟區的快照"](#)

["還原知識庫磁碟區的快照"](#)

暫停排程掃描

您現在可以暫停排程的資料來源掃描。根據預設、泛化 AI 工作負載會每天掃描每個資料來源、以便將新資料擷取到每個知識庫中。如果您不想擷取最新的變更（例如在測試期間或還原快照時）、您可以暫停排程的掃描、並隨時繼續進行。

["管理知識庫"](#)

知識庫現在支援資料保護磁碟區

選取知識庫磁碟區時、您現在可以選擇屬於 NetApp SnapMirror 複寫關係一部分的資料保護磁碟區。這可讓您將知識庫儲存在已受 SnapMirror 複寫保護的磁碟區上。

"識別要整合到知識庫中的資料來源"

2024 年 9 月 1 日

其他的分塊策略

泛用 AI 工作負載現在支援資料來源的多重句子區塊和重疊區塊。

每個知識庫的專用磁碟區

現在、泛用 AI 工作負載會為每個新知識庫建立專屬的 Amazon FSX for NetApp ONTAP Volume、為每個知識庫啟用個別的快照原則、並針對故障和資料中毒提供更好的保護。

2024 年 8 月 4 日

Amazon CloudWatch 記錄整合

現在、泛用 AI 工作負載已與 Amazon CloudWatch 記錄整合、可讓您監控泛用 AI 工作負載記錄檔。

例如：**chatbot** 應用程式

NetApp 工作負載工廠 GenAI 範例應用程式可讓您在網路型的聊天應用程式中直接與 IT 互動，從已發佈的 NetApp 工作負載工廠知識庫測試驗證和擷取。

2024 年 7 月 7 日

GenAI 工作負載工廠的初始版本

初始版本包含開發知識庫的功能、可藉由內嵌組織資料來自訂。使用者可透過聊天機器人應用程式存取知識庫。這項功能可確保針對組織特定問題做出準確且相關的回應、提高所有使用者的滿意度與生產力。

瞭解 GenAI 的 BlueXP 工作負載工廠

瞭解 GenAI 的 BlueXP 工作負載工廠

BlueXP Workload Factory for GenAI 可讓您將適用於 NetApp ONTAP 檔案系統的 Amazon FSX 與 GenAI 基礎模型整合。這可為您提供高效能儲存設備、並針對 AI 資料集提供豐富的保護、安全性和成本最佳化功能。

什麼是 GenAI 的 BlueXP 工作負載工廠？

BlueXP Workload Factory for GenAI 可讓您在 Amazon FSX for NetApp ONTAP 上使用企業資料來源，搭配泛用 AI 應用程式。利用擷取擴增世代（RAG），您可以快速將資料來源連接至 Amazon Bedrock 或 Amazon Q Business 提供的基礎模型，以開發具泛用性的 AI 驅動應用程式，例如虛擬助理，問答聊天機器人程式，文件摘要，內容建立等

將泛用 AI 與組織資料搭配使用、可讓您充分運用自己的知識和專業技能、而不只是仰賴模型所訓練的公開資料所提供的情報。使用 RAG 自訂模型可確保正確且相關地回應組織特定的問題、並使用泛化 AI 來提高應用程式使用者的生產力與效率。

開發專為貴組織資料量身打造的 GenAI 應用程式、可讓您充分運用自己的知識與專業技能。這項自訂功能可確保準確且相關地回應組織特定問題、提高所有使用者的滿意度與生產力。

如果您"[建立知識庫](#)"是，GenAI 會從資料來源擷取資料，將向量化的結果儲存在資料庫中，並讓您完全掌控如何使用擷取的資料來回答查詢。此方法需要更多初始設定，但可讓您針對不同的結果選擇不同的聊天模式。如果您"[為 Amazon Q Business 定義 NetApp 連接器](#)"是，Amazon Q Business 會擷取資料來源的資料，並儲存在索引中。這種方法需要較少的初始組態，但對結果的控制較少。

如需工作負載工廠的詳細資訊，請 "[工作負載原廠總覽](#)"參閱。

使用 GenAI 建立泛用 AI 應用程式的優點

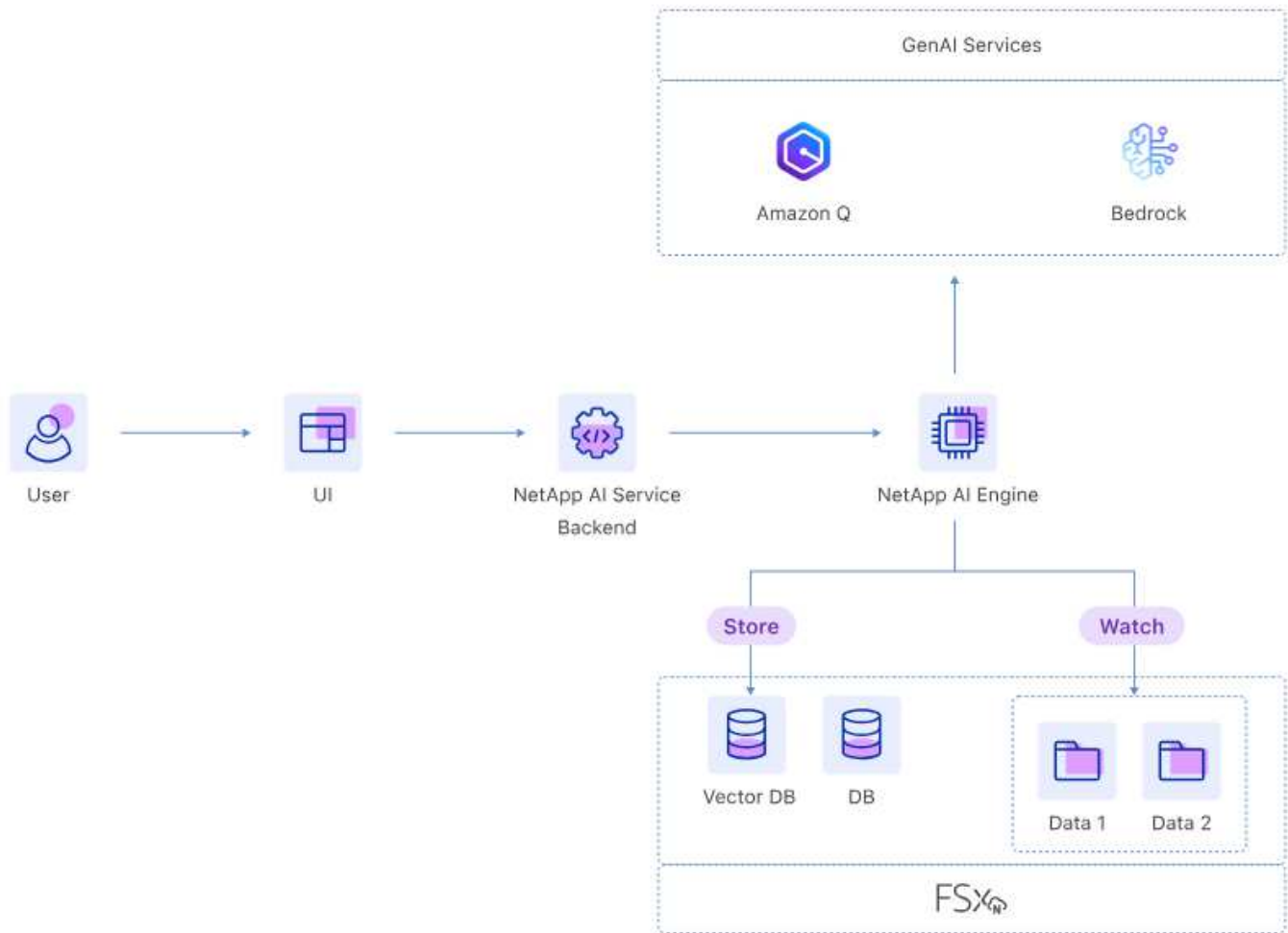
BlueXP Workload Factory for GenAI 使用擷取擴充世代（RAG），簡化部署基礎架構所需的程序，以建置泛用 AI 應用程式。具體而言、GenAI 提供下列優點：

- 無需深入瞭解資料基礎架構、基礎和語言模型、IT 管理員和開發人員就能利用 GenAI 提供的自動化功能來加速應用程式開發。資料管理員和開發人員可以輕鬆快速地建立企業知識庫、將組織的非結構化資料內嵌在其中、供泛化 AI 應用程式使用。
- 保留內嵌於知識庫之檔案的使用者權限、以確保資料安全性與隱私權得以維持、進而提升安全性。開發應用程式（例如 chatbot）時、只能根據使用者存取的資料、提供經過驗證的使用者答案。
- 將企業資料保密、安全地存放在 AWS 客戶帳戶內、讓您的組織資料永遠不會暴露在外部。
- 使用開放原始碼架構（例如 LangChain）加速開發 GenAI 應用程式（例如 Q&A 聊天機器人），利用 GenAI API 來配置和管理知識庫和連接器，與知識庫聊天，以及儲存和擷取聊天記錄。
- 透過在適用於 NetApp ONTAP 檔案系統的 FSX 上部署泛用 AI 資料基礎架構、並利用 ONTAP 功能（例如高可用度、用於本機資料保護與恢復的快照、用於災難恢復的 SnapMirror、以及用於備份資料基礎架構的 SnapVault）來改善資料保護與可用度。
- 利用 ONTAP 資料效率功能（例如重複資料刪除、壓縮和壓縮、資料分層和精簡配置）、降低通用 AI 資料基礎架構的整體儲存成本。

- 利用 GenAI 提供的混合式搜尋和重新排名功能，從您的資料中獲得高品質結果。混合式搜尋結合重新排名，可大幅提升搜尋結果的相關性。這些功能可透過 Amazon AWS 取得，且因地區而定。

GenAI 的運作方式

GenAI 使用貴組織的私有資料來輔助模型的智慧（根據訓練的資料）、針對貴組織中的使用者所提出的問題提供自訂答案。您先部署 RAG 架構所需的基礎架構，然後建立知識庫，或使用貴組織透過 Amazon Bedrock 或 Amazon Q Business 提供的資料來源和基礎模型來定義連接器，然後將應用程式（例如 Q&A 聊天機器人程式）連接到知識庫或連接器。



GenAI 的 BlueXP 工作負載工廠如何協助建置泛用 AI 應用程式

GenAI 以下列方式協助使用 RAG 建置泛型 AI 應用程式：

- 部署所需的基礎架構，以便擷取擴增型世代（RAG）架構，以處理適用於 ONTAP 檔案系統和 Amazon Bedrock 或 Amazon Q Business 的 FSX 資料來源。基礎架構包括用於管理資料的 NetApp GenAI Engine 執行個體、內嵌向量資料庫（LanceDB）、以及用於向量資料庫的 FSX for ONTAP 檔案系統上的儲存設備。
- 協助將資料來源連接至 Amazon bedrock 或 Amazon Q Business 提供的嵌入資料來源和語言模型，以擷取使用者查詢的回應。資料來源及模型及其組態會顯示為適用於 ONTAP 知識庫的 FSX。
- 將來源資料擷取至知識庫或連接器，以便將來源檔案內嵌於 SMB 共用上，並將 NFS 匯出至 ONTAP 檔案系統的 FSX，同時儲存 SMB 共用中檔案的檔案權限。

- 根據知識庫中的內容，自動建立對話入門問題。
- 為資料管理員提供聊天模擬器，以測試與知識庫的聊天。
- 提供簡單的連接器介面，讓您可以快速輕鬆地使用此 AI 助理的功能，將 GenAI 與 Amazon Q Business 連線。

使用工作負載工廠的工具

您可以搭配下列工具使用 BlueXP 工作負載工廠：

- *** 工作負載原廠主控台 ***：工作負載原廠主控台可提供應用程式與專案的視覺化整體檢視。
- *** BlueXP console***：BlueXP 主控台提供混合式介面體驗，讓您可以將 BlueXP 工作負載工廠與其他 BlueXP 服務一起使用。
- *** 詢問我 ***：使用「詢問我 AI」助理來詢問問題，並深入瞭解工作負載工廠，而無需離開工作負載工廠的網路 UI。從工作負載原廠說明功能表存取 Ask me。
- *** CloudShell CL1***：工作負載工廠包含 CloudShell CLI，可透過單一瀏覽器型 CLI，跨帳戶管理及操作 AWS 和 NetApp 環境。從工作負載原廠主控台的頂端列存取 CloudShell。
- *** REST API***：使用工作負載原廠 REST API 來部署及管理適用於 ONTAP 檔案系統和其他 AWS 資源的 FSX。
- *** CloudForgius***：使用 AWS CloudForgation 程式碼來執行您在工作負載工廠主控台中定義的動作，以便在 AWS 帳戶中建構，配置及管理 CloudForgation 堆疊中的 AWS 和第三方資源。
- *** Terraform BlueXP 工作負載工廠供應商 ***：使用 Terraform 建置及管理在工作負載工廠主控台產生的基礎架構工作流程。

成本

使用工作負載工廠的 GenAI 功能不需花費任何成本。

不過，您必須支付所部署的 AWS 資源、才能支援泛用 AI 基礎架構。例如，您將為 Amazon Bedrock 或 Amazon Q Business，ONTAP 檔案系統和儲存容量的 FSX，以及 GenAI 引擎 EC2 執行個體支付 AWS 費用。

某些多模式作業（例如掃描影像以取得文字資訊）可能會使用更多資源，因此會產生較高的成本。某些組態作業（例如變更知識庫的設定）可能會導致重新掃描資料來源，而且資料來源掃描也會產生較高的成本。

授權

NetApp 不需要特殊授權即可使用工作負載工廠的 AI 功能。

NetApp GenAI 引擎的元件

當您部署 GenAI 基礎架構時，工作負載工廠會為 GenAI 引擎建立一個 EC2 執行個體。它也會為此執行個體建立 IAM 角色、安全性群組和私有端點。您可能想要瞭解更多有關工作負載工廠在 AWS 環境中建立的這些元件的詳細資料。

EC2 執行個體類型

m5.large

IAM角色

GenAI 引擎執行個體需要權限、才能將資料區塊傳送至 Amazon 基礎上的嵌入模型、以及與 NetApp AI 服務後端通訊。IAM 角色包括下列權限：

```

{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Action": [
        "ssm:DescribeDocument",
        "ssm:DescribeAssociation",
        "ssm:GetDeployablePatchSnapshotForInstance",
        "ssm:GetManifest",
        "ssm:ListInstanceAssociations",
        "ssm:ListAssociations",
        "ssm:PutInventory",
        "ssm:PutComplianceItems",
        "ssm:PutConfigurePackageResult",
        "ssm:UpdateAssociationStatus",
        "ssm:UpdateInstanceAssociationStatus",
        "ssm:UpdateInstanceInformation",
        "ssmmessages:CreateControlChannel",
        "ssmmessages:CreateDataChannel",
        "ssmmessages:OpenControlChannel",
        "ssmmessages:OpenDataChannel"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "ssm:GetParameter"
      ],
      "Resource": "arn:aws:ssm:*:*:parameter/netapp/wlmai/*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "fsx:DescribeVolumes",
        "fsx:DescribeStorageVirtualMachines",
        "fsx:DescribeFileSystems"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [

```

```

        "fsx:TagResource",
        "fsx:ListTagsForResource"
    ],
    "Resource": [
        "arn:aws:fsx:*:*:storage-virtual-machine/*/*",
        "arn:aws:fsx:*:*:volume/*/*"
    ],
    "Effect": "Allow"
},
{
    "Action": [
        "fsx:CreateVolume"
    ],
    "Resource": [
        "arn:aws:fsx:*:*:volume/*/*",
        "arn:aws:fsx:*:*:storage-virtual-machine/*/*"
    ],
    "Effect": "Allow"
},
{
    "Condition": {
        "StringLike": {
            "aws:ResourceTag/netapp:wlmai: <ai-engine-id>:kbId": "*"
        }
    },
    "Action": "fsx:DeleteVolume",
    "Resource": [
        "arn:aws:fsx:*:*:volume/*/*",
        "arn:aws:fsx:*:*:backup/*"
    ],
    "Effect": "Allow"
},
{
    "Condition": {
        "StringLike": {
            "aws:ResourceTag/netapp:wlmai: <ai-engine-id>:qConnectorId": "*"
        }
    },
    "Action": "fsx:DeleteVolume",
    "Resource": [
        "arn:aws:fsx:*:*:volume/*/*",
        "arn:aws:fsx:*:*:backup/*"
    ],
    "Effect": "Allow"
},

```

```

{
  "Condition": {
    "StringLike": {
      "aws:ResourceTag/netapp:wlmai: <ai-engine-id>": "*"
    }
  },
  "Action": "fsx:UntagResource",
  "Resource": "arn:aws:fsx:*:*:storage-virtual-machine/*/*",
  "Effect": "Allow"
},
{
  "Condition": {
    "StringLike": {
      "aws:ResourceTag/netapp:wlmai: <ai-engine-id>:kbId": "*"
    }
  },
  "Action": "fsx:UntagResource",
  "Resource": "arn:aws:fsx:*:*:volume/*/*",
  "Effect": "Allow"
},
{
  "Condition": {
    "StringLike": {
      "aws:ResourceTag/netapp:wlmai: <ai-engine-id>:qConnectorId": "*"
    }
  },
  "Action": "fsx:UntagResource",
  "Resource": "arn:aws:fsx:*:*:volume/*/*",
  "Effect": "Allow"
},
{
  "Action": [
    "bedrock:InvokeModel",
    "bedrock:Rerank",
    "bedrock:GetFoundationModel",
    "bedrock:GetInferenceProfile"
  ],
  "Resource": "*",
  "Effect": "Allow"
},
{
  "Action": [
    "ec2messages:GetMessages",
    "ec2messages:GetEndpoint",
    "ec2messages:AcknowledgeMessage",

```

```

        "ec2messages:DeleteMessage",
        "ec2messages:FailMessage",
        "ec2messages:SendReply"
    ],
    "Resource": "*",
    "Effect": "Allow"
},
{
    "Action": [
        "qbusiness:ListWebExperiences",
        "qbusiness:GetApplication",
        "qbusiness:CreateDataSource",
        "qbusiness:DeleteDataSource",
        "qbusiness:ListIndices",
        "qbusiness:StartDataSourceSyncJob",
        "qbusiness:StopDataSourceSyncJob",
        "qbusiness:ListDataSourceSyncJobs",
        "qbusiness:BatchPutDocument",
        "qbusiness:BatchDeleteDocument"
    ],
    "Resource": "*",
    "Effect": "Allow"
},
{
    "Action": [
        "logs:DescribeLogGroups"
    ],
    "Resource": "*",
    "Effect": "Allow"
},
{
    "Action": [
        "logs:DescribeLogStreams",
        "logs:PutLogEvents",
        "logs:CreateLogStream",
        "logs:CreateLogGroup"
    ],
    "Resource": [
        "arn:aws:logs:*:*:log-group:/netapp/wlmai/*:log-stream:*",
        "arn:aws:logs:*:*:log-group:/netapp/wlmai/*"
    ],
    "Effect": "Allow"
},
{
    "Action": [
        "s3:GetObject",

```

```

        "s3:PutObject"
    ],
    "Resource": "*",
    "Effect": "Allow"
  },
  {
    "Action": [
      "kms:Decrypt",
      "kms:GenerateDataKey"
    ],
    "Resource": "*",
    "Effect": "Allow"
  }
]
}

```

安全性群組

傳出規則會對所有流量開放、而傳入規則則會完全關閉。

私有端點

如果目標 VPC 尚未擁有這些端點，則工作負載工廠會為 GenAI 引擎 EC2 執行個體建立私有端點，以便與下列 AWS 服務進行通訊：

- Amazon bedrock
 - Bedrock
 - Bedrock-runtime
 - bedrock-agent-runtime
- Amazon Elastic Container 登錄 (ECR)
 - API
 - Docker
- AWS Systems Manager (SSM)
 - SSM)
 - ec2messages
 - ssmmessages
- Amazon FSX for NetApp ONTAP 產品
- Amazon CloudWatch

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。

1

登入工作負載工廠

您需要 ["在工作負載工廠設定帳戶"](#) 使用其中一項登入 ["主控台體驗"](#)。

2

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"](#)。

3

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"](#)。

4

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

在您建置知識庫之前，請先確定工作負載工廠和 AWS 已正確設定。這包括擁有 AWS 登入認證、部署的 ONTAP 檔案系統 FSX、其中包含您想要整合至知識庫的資料來源、存取 Amazon bedrock AI 服務等。

基本 GenAI 需求

GenAI 的一般需求是您開始使用前所需的環境。

工作負載工廠登入與帳戶

您需要 ["在工作負載工廠設定帳戶"](#) 使用其中一項登入 ["主控台體驗"](#)。

AWS 認證與權限

您需要將 AWS 憑證新增至具有讀取/寫入權限的工作負載工廠，這表示您將在 GenAI 的讀取/寫入模式下使用工作負載工廠。

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。

["瞭解如何將 AWS 認證新增至工作負載工廠"](#)

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 GenAI 的 BlueXP 工作負載工廠之前，您必須先設定 Amazon 基礎。您的 GenAI 部署必須位於啟用 Amazon 基礎的 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

工作負載工廠 GenAI 知識庫目前支援下列資料來源檔案格式。

檔案格式	擴充
Apache Parquet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

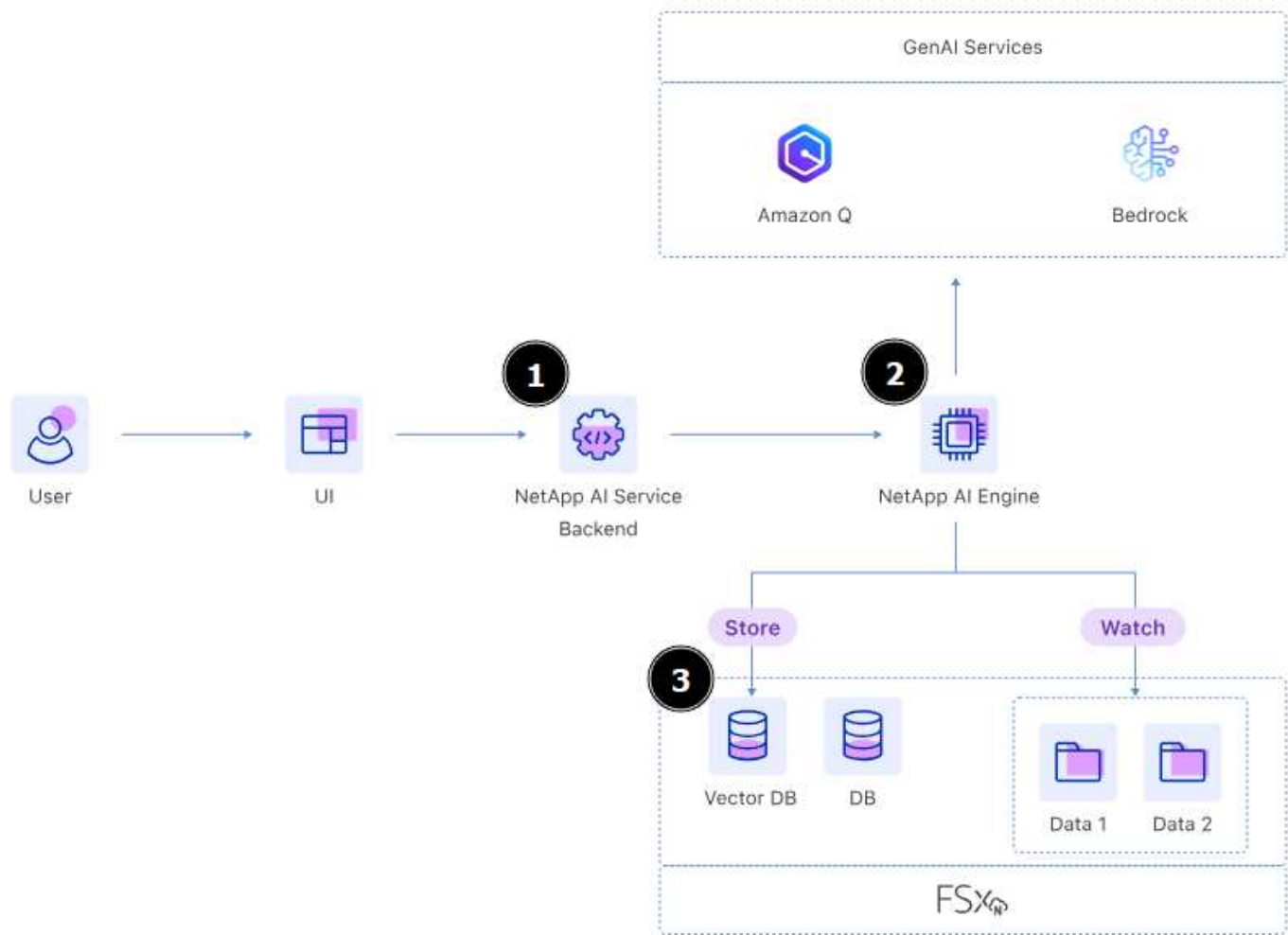
適用於 ONTAP 檔案系統的 FSX 可為您的 GenAI 系統提供儲存設備。

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. 使用其中一項登入工作負載工廠["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. * 標記 *：輸入您要套用至此部署中所有 AWS 資源的任何標記金鑰 / 值配對。這些標籤可在 AWS 管理主控台和工作負載工廠內的基礎架構資訊區域中看到，並可協助您追蹤工作負載原廠資源。

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. * 執行個體名稱 *：選擇性地選取 * 定義執行個體名稱 *，然後輸入 AI 引擎執行個體的自訂名稱。執行個體名稱會出現在 AWS 管理主控台，以及工作負載工廠內的基礎架構資訊區域，可協助您追蹤工作負載原廠資源。

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

工作負載工廠開始部署 chatbot 基礎架構。此程序最多可能需要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 引擎已設定為使用名稱為的記錄群組、將記錄傳送至 Amazon CloudWatch 記錄檔 /netapp/wlmai/<tenancyAccountId>/randomId、其中 <tenancyAccountId> 是 "BlueXP 帳戶 ID" 目前使用者的。

建立 GenAI 知識庫

在您部署 AI 基礎架構，並從適用於 ONTAP 資料存放區的 FSX 中找出要整合到知識庫的資料來源之後，您就可以使用工作負載工廠來建置知識庫。在本步驟中、您也將定義 AI 特性、並建立對話前人。

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。

貴組織的資料來源可能包含個人識別資訊（PII）。為了保護這項敏感資訊、您可以在建立和設定知識庫時啟用 [_data guardadrails_](#)。由 BlueXP 分類技術支援的資料欄可識別及遮罩 PII、使其無法存取且無法擷取。

["瞭解 BlueXP 分類"](#)。



GenAI 的 BlueXP 工作負載工廠不會遮罩敏感的個人資訊（SPii）。如需此類型資料的詳細資訊，請參閱["敏感個人資料的類型"](#)。



資料欄可隨時啟用或停用。如果您切換資料防護功能，工作負載工廠會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. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 在 AI 工作負載方塊中、選取 [* 部署與管理 *](#)。
3. 從「知識庫與連接器」標籤中，選取「[* 建立新 *](#)」下拉式清單，然後選擇 [* NetApp GenAI 知識庫作為基礎 *](#)。
4. 在「定義知識庫」頁面上、設定知識庫設定：
 - a. [* 名稱 *](#)：輸入您要用於知識庫的名稱。
 - b. [* 說明 *](#)：輸入知識庫的詳細說明。
 - c. [* 內嵌模型 *](#)：內嵌模型定義如何將資料轉換成知識庫的向量內嵌。工作負載工廠支援下列機型：
 - 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"](#)

- d. [* 聊天模式 *](#)：從 Amazon bedrock 整合的各種聊天模式中選擇。請注意、您必須已啟用 Amazon Bedrock 的聊天模式。
- e. 重新排名：啟用或停用重新排名功能，此功能可提高查詢結果的相關性和品質。您可以選擇標準聊天模型或專用的重新排名模型進行重新排名。只有您所在地區支援的重新排名模型選項才會顯示。

- f. * 資料欄 * : 選擇是否要啟用或停用資料欄。["瞭解採用 BlueXP 分類技術的資料欄"](#)。

必須符合下列先決條件，才能啟用資料欄。

- 必須有服務帳戶才能與 BlueXP 分類通訊。您必須在 BlueXP 租賃帳戶上擁有 `_ 組織管理員 _` 角色、才能建立服務帳戶。擁有組織管理員角色的成員可以在 BlueXP 中完成所有動作。["瞭解如何在 BlueXP 中新增角色至成員"](#)
- AI 引擎必須能夠存取["BlueXP API 端點"](#)。
- 您需要執行下列步驟、如所述["BlueXP 分類文件"](#)：
 - A. 建立 BlueXP Connector
 - B. 確保您的環境符合先決條件
 - C. 部署 BlueXP 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。

- g. * 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"](#)。

- h. * 適用於 ONTAP 檔案系統 * 的 FSX : 當您定義新的知識庫時，工作負載工廠會為 NetApp ONTAP Volume 建立新的 Amazon FSX 以儲存它。選擇要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)。
- i. * Snapshot 原則 * : 從工作負載原廠儲存設備庫中定義的現有原則清單中選擇快照原則。知識庫的週期性快照將根據您選擇的快照原則、以頻率自動建立。

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。

5. 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。

如果您包含影像或 PDF 檔案，則 GenAI 的 BlueXP 工作負載工廠會剖析影像中的文字（包括 PDF 文件中的影像），這會產生較高的成本。

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。

如果您包含影像或 PDF 檔案，則 GenAI 的 BlueXP 工作負載工廠會剖析影像中的文字（包括 PDF 文件中的影像），這會產生較高的成本。

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：

- 輸入資料來源檔案所在的檔案系統主機的 IP 位址或 FQDN。
- 為網路共享選擇 SMB 協定。
- 輸入 Active Directory 訊息，包括網域、IP 位址、使用者名稱和密碼。
- 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。

如果您包含影像或 PDF 檔案，則 GenAI 的 BlueXP 工作負載工廠會剖析影像中的文字（包括 PDF 文件中的影像），這會產生較高的成本。

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。

您將在工作負載工廠內執行大部分的測試，然後再將知識庫部署到使用者適用的應用程式中。如果您需要變更資料來源或 chatbot 作業、您現在就想要在發佈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。
- 如果回應中包含表格，您可以排序每一欄中的資料，並將每個表格複製到剪貼簿。

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. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從「知識庫」清查頁面中、選取您要啟動驗證的知識庫。
3. 選取  並選取 * 管理知識庫 *。
4. 選取 * 動作 * 功能表、然後選取 * 管理驗證設定 *。
5. 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
```

6. 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

發佈知識庫可讓您在聊天應用程式中使用知識庫。發佈動作會觸發工作負載原廠 API 來產生及發佈唯一的端點。發佈之後、即可存取線上對談應用程式的知識庫、API 端點也已準備好進行整合。

您發佈的每個知識庫都有獨特的端點。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從「知識庫」清查頁面中、選取您要發佈的知識庫。
3. 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。

4. 選取 * 動作 * 功能表、然後選取 * 發佈 *。

工作負載工廠會發佈知識庫。在知識庫的詳細資料頁面上、狀態會從 * 未發佈 * 變更為 * 已發佈 *。

您現在可以取得知識庫專屬端點的詳細資料。

5. 在「已發佈」狀態旁邊、選取 * 檢視 *。

畫面會顯示如何使用工作負載原廠 API 存取知識庫的詳細資料。

6. 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

在您設定，啟動及發佈知識庫之後，外部應用程式開發人員可以設定及執行 NetApp 提供的開放原始碼範例 chatbot 應用程式，與您的知識庫互動，並瞭解如何使用工作負載工廠 API 來建立自己的泛化 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"](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

["工作負載原廠 API 文件"](#)

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。

如需管理適用於 ONTAP 檔案系統的 FSX 的相關資訊、請移至 ["Amazon FSX for NetApp ONTAP 工作負載工廠文件"](#) 以檢視您可以使用的備份與保護功能。

使用 GenAI 為 Amazon Q Business 建立連接器

開始使用

GenAI 連接器快速入門

開始使用您組織位於 Amazon FSx for NetApp ONTAP 檔案系統上的資料建立 NetApp Connector for Amazon Q Business。建立連接器之後，終端使用者就可以存取 Amazon Q Business Assistant，以針對組織的方式回應他們的問題。

1

登入工作負載工廠

您需要 ["在工作負載工廠設定帳戶"](#) 使用其中一項登入 ["主控台體驗"](#)。

2

設定您的環境以符合 GenAI 需求

您將需要 AWS 認證來部署 AWS 基礎架構，部署並探索 ONTAP 檔案系統的 FSX，您想要整合至連接器的資料來源清單，存取 Amazon Q Business 應用程式等。

["深入瞭解 GenAI 需求"](#)。

3

識別包含資料來源的 ONTAP 檔案系統的 FSX

您將整合至連接器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"](#)。

4

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。磁碟區用於儲存有關連接器的資訊。

["瞭解如何部署 GenAI 基礎架構"](#)。

下一步

您現在可以為 Amazon Q Business 建立連接器，為終端使用者提供以組織為中心的回應。

GenAI 連接器需求

在為 Amazon Q Business 建立 NetApp 連接器之前，請確保工作負載工廠和 AWS 已正確設定。

基本 GenAI 需求

GenAI 的一般需求是您開始使用前所需的環境。

工作負載工廠登入與帳戶

您需要 ["在工作負載工廠設定帳戶"](#) 使用其中一項登入 ["主控台體驗"](#)。

AWS 認證與權限

您需要將 AWS 憑證新增至具有讀取/寫入權限的工作負載工廠，這表示您將在 GenAI 的讀取/寫入模式下使用工作負載工廠。

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。

["瞭解如何將 AWS 認證新增至工作負載工廠"](#)

NetApp Connector for Amazon Q Business 的要求

確保您的環境符合 Amazon Q Business 連接器的下列特定要求。

Amazon Q Business 應用程式

您必須建立 Amazon Q Business 應用程式，或使用現有的應用程式。

- 確保應用程式存在於您的 AWS 區域之一。
- 確保您擁有 ["已建立索引"](#) 應用程式。
- 確保應用程式未處於故障狀態。

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））來儲存有關連接器的資訊。

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新增至連接器的資料來源。

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。
- 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"](#)

識別要新增至連接器的資料來源

識別或建立要整合到連接器中的適用於 ONTAP 檔案系統的 FSX 文件（資料來源）。這些資料來源可讓 Amazon Q Business 根據與貴組織相關的資料，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連接器，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

連接器支援來自使用 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，您需要輸入 Active Directory 資訊，以便連接器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，如果您已將 10 個檔案（資料來源）整合至您的 Connector，而其中 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，才會收到包含這些檔案資料的聊天機器人程式回應。



當您將資料來源新增至 Amazon Q Business Connector 時，只有使用者權限會套用至資料來源檔案。未套用群組權限。



如果資料來源中的檔案缺少文字（例如，無文字影像），Amazon Q Business 不會對其進行索引，而是會在 Amazon CloudWatch 記錄中記錄一個項目，指出沒有文字。

支援的資料來源檔案格式

NetApp Connector for Amazon Q Business 目前支援以下資料來源檔案格式。

檔案格式	擴充
以逗號分隔的值檔案	.CSV

檔案格式	擴充
JSON 和 JSONP	json
請下標	.md
Microsoft Word	.docx
純文字	.txt
可攜式文件格式	.pdf
Microsoft PowerPoint	.ppt 或 .pptx
超文字標記語言	.html
可延伸標記語言	.xml
xslt	.xslt
Microsoft Excel	xls
RTF 格式	rtf

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

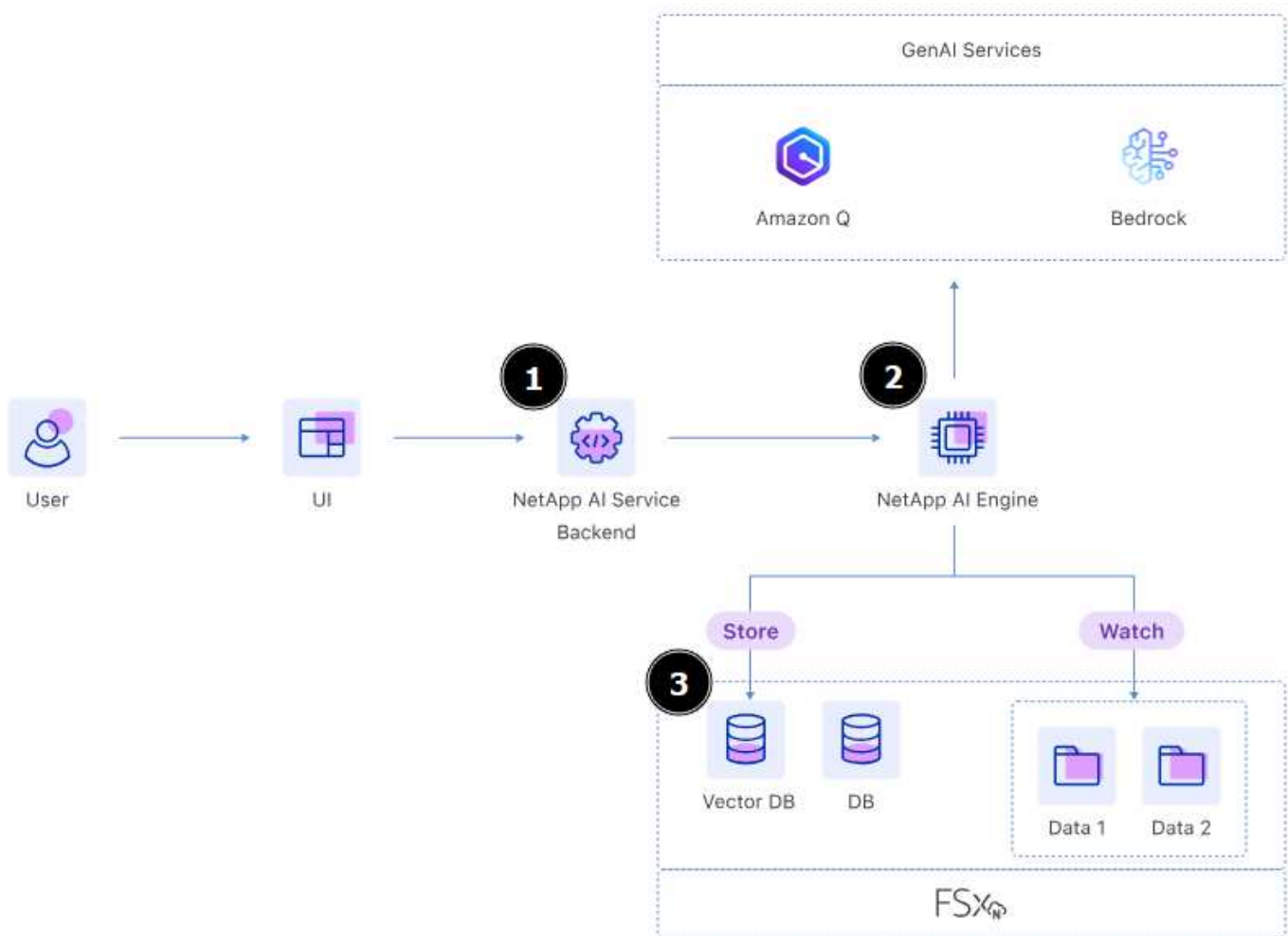
適用於 ONTAP 檔案系統的 FSX 可為您的 GenAI 系統提供儲存設備。

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. 使用其中一項登入工作負載工廠["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. * 標記 *：輸入您要套用至此部署中所有 AWS 資源的任何標記金鑰 / 值配對。這些標籤可在 AWS 管理主控台和工作負載工廠內的基礎架構資訊區域中看到，並可協助您追蹤工作負載原廠資源。
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. * 執行個體名稱 *：選擇性地選取 * 定義執行個體名稱 *，然後輸入 AI 引擎執行個體的自訂名稱。執行個體名稱會出現在 AWS 管理主控台，以及工作負載工廠內的基礎架構資訊區域，可協助您追蹤工作負載原廠資源。
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

工作負載工廠開始部署 chatbot 基礎架構。此程序最多可能需要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 引擎已設定為使用名稱的記錄群組、將記錄傳送至 Amazon CloudWatch 記錄檔 /netapp/wlmai/<tenancyAccountId>/randomId、其中 <tenancyAccountId> 是 ["BlueXP 帳戶 ID"](#) 目前使用者的。

為 Amazon Q Business 建立 NetApp 連接器

部署 AI 基礎架構並確定將從 FSx for ONTAP 資料儲存區使用的資料來源後，您就可以為

Amazon Q Business 定義 NetApp 連接器了。

繼續之前，請確保您的環境符合["需求"](#)適用於 Amazon Q Business 的。

關於這項工作

貴組織的資料來源可能包含個人識別資訊（PII）。若要保護這項敏感資訊，您可以在定義連接器時啟用 `_data guardrails`。由 BlueXP 分類技術支援的資料欄可識別及遮罩 PII、使其無法存取且無法擷取。

["瞭解 BlueXP 分類"](#)。



GenAI 的 BlueXP 工作負載工廠不會遮罩敏感的個人資訊（SPII）。如需此類型資料的詳細資訊，請參閱["敏感個人資料的類型"](#)。



資料欄可隨時啟用或停用。如果您切換資料防護功能，工作負載工廠會從頭開始掃描整個資料來源，這可能會產生成本。

定義連接器

為 Amazon Q Business 建立 NetApp 連接器。此連接器支援 GenAI 和 Amazon Q Business 之間的 API 和資料來源通訊。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 在 AI 工作負載方塊中、選取 `* 部署與管理 *`。
3. 從「知識庫與連接器」標籤中，選取「`* 建立新的 *`」下拉式清單，然後選擇「`* Amazon Q Business Connector*`」。
4. 在「定義連接器」頁面上，設定連接器設定：
 - a. `* 名稱 *`：輸入您要用於連接器的名稱。
 - b. `* 說明 *`：輸入連接器的詳細說明。
 - c. `* Amazon Q*`：您想要整合的 Amazon Q Business 執行個體的地區和應用程式名稱。
 - d. `* 資料欄 *`：選擇是否要啟用或停用資料欄。["瞭解採用 BlueXP 分類技術的資料欄"](#)。

必須符合下列先決條件，才能啟用資料欄。

- 必須有服務帳戶才能與 BlueXP 分類通訊。您必須在 BlueXP 租賃帳戶上擁有 `_ 組織管理員 _` 角色、才能建立服務帳戶。擁有組織管理員角色的成員可以在 BlueXP 中完成所有動作。["瞭解如何在 BlueXP 中新增角色至成員"](#)
- AI 引擎必須能夠存取["BlueXP API 端點"](#)。
- 您需要執行下列步驟、如所述["BlueXP 分類文件"](#)：
 - A. 建立 BlueXP Connector
 - B. 確保您的環境符合先決條件
 - C. 部署 BlueXP 分類



當您啟用資料欄功能時，GenAI 只會擷取純文字（不包括內嵌影像或媒體文字）並遮罩任何私有或敏感資料，以處理 .txt、.md、.csv、.doc 和 .pdf 檔案。所有其他檔案類型都會正常處理，而不會遮罩私有或敏感資料。

- e. **FSx for ONTAP** 檔案系統：當您為 Amazon Q Business 定義新的 NetApp 連接器時，工作負載工廠會建立一個新的 Amazon FSx for NetApp ONTAP 磁碟區來儲存連接器資訊。選擇要在其中建立新磁碟區的現有檔案系統和 SVM（也稱為儲存 VM）。
- f. * **Snapshot 原則** *：從工作負載原廠儲存設備庫存中定義的現有原則清單中選擇快照原則。GenAI 會根據您選擇的快照原則，自動建立磁碟區的週期性快照，以頻率儲存連接器資訊。

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。

- 5. 選取 * **建立連接器** * 以整合 Amazon Q Business 與 GenAI。

建立連接器時會出現進度指示器。

建立連接器之後，您可以選擇將資料來源新增至連接器，以便 Amazon Q Business 擷取您的資料並將其新增至索引。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。

將資料來源新增至連接器

您可以新增一或多個資料來源，以便在 Amazon Q Business 索引中填入組織的資料。

關於這項工作

- 支援的資料來源數量上限為 10 個。
- 如需 Amazon Q Business 索引的特定服務限制，請參閱 ["Amazon Q 商業文件"](#)。

步驟

1. 選擇***新增資料來源***後，將出現***選擇檔案系統***頁面。
2. * **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** *。
3. * **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。

4. * **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** *。
5. * **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：
 - * **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。
 - 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。

6. 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 *：存取此連接器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。



Amazon Q Business Connector 資料來源不支援 Active Directory 群組權限。

7. 選取 * 新增 *，將此資料來源新增至 Amazon Q Business Connector。

結果

資料來源內嵌於 Amazon Q Business 索引中。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 Connector 之後，您可以在 Amazon Q Business chatbot 環境中測試，並在將服務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連接器。

管理與監控

管理 GenAI 基礎架構

您可以檢視已部署 GenAI RAG 基礎架構的詳細資料、或是在不再需要時移除 chatbot 基礎架構。

檢視基礎架構的相關資訊

您可以檢視有關 chatbot 基礎架構的資訊。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從工作負載原廠導覽功能表中，選取 * AI *。
3. 選擇 **Infrastructure** 選項卡。
4. 檢視基礎架構的相關資訊、其中包含下列元件的詳細資料：
 - AWS 設定
 - 基礎架構設定
 - AI 引擎
 - 向量資料庫

移除基礎架構

如果您不再需要為一或多個聊天機器人程式部署的聊天機器人程式基礎架構，您可以將其從工作負載工廠移除。



已部署在此基礎架構上的所有聊天機器人程式都將停用、所有聊天記錄都將刪除。

此作業只會從工作負載工廠移除 AI 基礎架構的連結，並不會從 AWS 移除所有元件。您需要從 AWS 手動刪除下列基礎架構元件：

- VM 執行個體
- 私有端點
- 適用於 ONTAP 檔案系統的 FSX 上包含 AI 資料庫的 Volume
- IAM 角色
- 原則
- 安全性群組

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從工作負載原廠導覽功能表中，選取 * AI *。
3. 選擇 **Infrastructure** 選項卡。

4. 選取  並選取 * 移除 chatbot 基礎架構 * 。
5. 確認您要刪除基礎架構、然後選取 * 移除 * 。

結果

chatbot 基礎架構元件會從工作負載工廠移除。

管理 GenAI 知識庫

建立知識庫之後、您可以檢視知識庫詳細資料、修改知識庫、整合其他資料來源、或刪除知識庫。

檢視知識庫的相關資訊

您可以檢視有關知識庫設定和整合資料來源的資訊。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從工作負載原廠導覽功能表中，選取 * AI * 。
3. 選取您要檢視的知識庫。

如果已定義、目前正在使用的對話啟動器會顯示在右窗格中。

4. 若要檢視知識庫詳細資料、請選取  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、所有內嵌資料來源的清單等等。

「* 動作 *」功能表可讓您在需要進行任何變更時、管理知識庫。

編輯知識庫

您可以變更某些設定來更新知識庫、也可以新增或移除資料來源。

每次您從知識庫新增、修改或移除資料來源時、都必須同步資料來源、以便將其重新索引至知識庫。同步是漸進式的、因此 Amazon bedrock 只會處理您的 FSX for ONTAP Volume 中自上次同步以來已新增、修改或刪除的物件。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從「知識庫」清查頁面中、選取您要更新的知識庫。
3. 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、所有內嵌資料來源的清單等等。

4. 選取 * 動作 * 功能表、然後選取 * 編輯知識庫 * 。
5. 在編輯知識庫頁面中，您可以變更知識庫名稱、描述、嵌入模型、聊天模型、功能啟用，選擇是自動或手動建立對話啟動器，以及包含知識庫的磁碟區所使用的快照策略。

如果您使用「手動」模式來啟動對話、也可以在此變更對話的開始者。



每次知識庫掃描、包括內嵌、成本。如果在建立知識庫之後啟用資料 guardrails、則會再次掃描知識庫並產生成本。同樣地，如果您變更聊天模式， GenAI 也會重新掃描相關的資料來源（產生成本）。

6. 變更後請選擇 * 儲存 * 。

利用快照保護知識庫

您可以拍攝及還原知識庫磁碟區的快照、以保護知識庫資料。您可以隨時從快照還原至舊版知識庫。

快照可以比備份更快、更有效率地儲存、並可讓您使用不同的保護原則來保護每個知識庫。以下是一些快照可能有用的案例：

- 意外資料遺失或毀損
- 從錯誤的資料擷取到知識庫中恢復
- 測試不同的資料來源或分塊策略、並在測試完成時快速還原

拍攝知識庫磁碟區的快照

您可以手動擷取知識庫磁碟區的快照、以儲存知識庫的狀態。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從「知識庫」清查頁面中、選取您要保護的知識庫。
3. 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、所有內嵌資料來源的清單等等。

4. 選取 * 動作 * 功能表、然後選取 * 快照 > 建立新的快照 * 。
5. （可選）選擇 * 定義快照名稱 * 並輸入快照的自定義名稱。

定義自訂名稱可協助您在未來需要還原快照時，更有效地判斷快照的內容。

6. 選擇 * Create （建立） 。

隨即建立知識庫的快照。

還原知識庫磁碟區的快照

您可以隨時還原知識庫磁碟區的手動或排程快照。



如果儲存在磁碟區上的資料庫毀損或已刪除、則無法使用泛用 AI 工作負載 UI 還原快照。因應措施是、您可以使用託管磁碟區的 ONTAP 叢集上的來還原快照 ["CLI ONTAP"](#) 。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從「知識庫」清查頁面中、選取您要還原的知識庫。
3. 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、所有內嵌資料來源的清單等等。

4. 選取 * 「動作」 * 功能表、然後選取 * 「快照」 > 「還原快照」 * 。

快照選擇對話方塊隨即出現、您可以在其中查看已為此知識庫建立的快照清單。

5. (可選) 如果希望在恢復快照後繼續排程和目前執行的資料來源掃描、請取消選取 * 「恢復快照後暫停執行和排程掃描」選項。

此選項預設為啟用、可確保在知識庫處於部分還原狀態時不會進行掃描、或掃描不會以舊資料更新新還原的知識庫。

6. 從清單中選取您要還原的快照。
7. 選擇*還原*。

複製知識庫

您可以從知識庫快照建立新的知識庫。如果原始知識庫毀損或遺失，這項功能就很實用。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從「知識庫」清查頁面中、選取您要還原的知識庫。
3. 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、所有內嵌資料來源的清單等等。

4. 選取 * 動作 * 功能表，然後選取 * 快照 > 複製知識庫 * 。

此時將出現 Clone（克隆）對話框。

5. 或者，如果您想要在複製快照之後繼續排程及目前執行的資料來源掃描，請取消選取 * 在複製快照後暫停執行和排程掃描 * 選項。

此選項預設為啟用、可確保在知識庫處於部分還原狀態時不會進行掃描、或掃描不會以舊資料更新新還原的知識庫。

6. 從清單中選取您要複製的快照。
7. 選擇*繼續*。
8. 輸入新知識庫的名稱。
9. 為新知識庫選擇檔案系統 SVM 和 Volume 名稱。
10. 選擇* Clone（克隆）*。

將其他資料來源新增至知識庫

您可以在知識庫中內嵌其他資料來源、以便將其他組織資料填入其中。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從「知識庫」清查頁面中、選取您要新增資料來源的知識庫。
3. 選取  並選取 * 新增資料來源 * 。
4. 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。

如果您包含影像或 PDF 檔案，則 GenAI 的 BlueXP 工作負載工廠會剖析影像中的文字（包括 PDF 文件中的影像），這會產生較高的成本。

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。

如果您包含影像或 PDF 檔案，則 GenAI 的 BlueXP 工作負載工廠會剖析影像中的文字（包括 PDF 文件中的影像），這會產生較高的成本。

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。

如果您包含影像或 PDF 檔案，則 GenAI 的 BlueXP 工作負載工廠會剖析影像中的文字（包括 PDF 文件中的影像），這會產生較高的成本。

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

資料來源已整合至您的知識庫。

將資料來源與知識庫同步

資料來源會每天自動與相關知識庫同步一次、以便在聊天機器人程式中反映任何資料來源變更。如果您變更任何資料來源、而且想要立即同步處理資料、您可以執行隨需同步。

同步是漸進式的、所以 Amazon bedrock 只會處理資料來源中自上次同步以來已新增、修改或刪除的物件。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從「知識庫」清查頁面中、選取您要同步的知識庫。
3. 選取  並選取 * 管理知識庫 *。
4. 選取 * 動作 * 功能表、然後選取 * 立即掃描 *。

您會看到一則訊息、指出正在掃描資料來源、以及掃描完成後的最後一則訊息。

結果

知識庫會與附加的資料來源同步、任何作用中的聊天機器人程式都會從您的資料來源開始使用最新資訊。

暫停或繼續排程的同步處理

如果您想要暫停或恢復資料來源的下一次同步（掃描），您可以隨時進行。如果您要變更資料來源，而不想在變更期間進行同步處理，則可能需要暫停下一次排程的同步處理。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從「知識庫與連接器」標籤中，選取您要暫停或恢復掃描的知識庫。
3. 選取  並選取 * 管理知識庫 *。
4. 選取 * 動作 * 功能表，然後選取 * 掃描 > 暫停排程掃描 * 或 * 掃描 > 恢復排程掃描 *。

您會看到訊息，表示下一次排程掃描已暫停或恢復。

建立知識庫之前、請先評估聊天模式

您可以在建立知識庫之前評估可用的基礎聊天模型、以便查看哪些模型最適合您的實作。由於各 AWS 地區的機型支援不盡相同、請參閱 ["此 AWS 文件頁面"](#)、確認您可以在計畫部署知識庫的地區使用哪些機型。



只有在知識庫庫存頁面中沒有知識庫時、才能使用此功能。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 您可以在「知識庫」的庫存頁面上、選擇交談模式的選項、就在交談機器人的頁面右側。
3. 從清單中選取聊天模式、然後在提示區域中輸入一組問題、以查看聊天機器人如何回應。
4. 請嘗試多種模式、看看哪種模式最適合您的實作。

結果

建立知識庫時、請使用該聊天模式。

解除發佈您的知識庫

當您發佈知識庫以將其與某個 chatbot 應用程式整合之後、如果您想要停用 chatbot 應用程式來存取知識庫、就可以將其解除發佈。

解除發佈知識庫會停止任何聊天應用程式的運作。可存取知識庫的唯一 API 端點會停用。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從「知識庫」清查頁面中、選取您要解除發佈的知識庫。
3. 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。

4. 選取 * 動作 * 功能表、然後選取 * 解除發佈 *。

結果

知識庫已停用、無法再由 chatbot 應用程式存取。

刪除知識庫

如果您不再需要知識庫、可以將其刪除。當您刪除知識庫時，知識庫會從工作負載原廠移除，並刪除內含知識庫的磁碟區。任何使用知識庫的應用程式或聊天機器人程式都將停止運作。刪除知識庫是不可還原的。

刪除知識庫時、您也應該解除知識庫與任何與其相關的代理程式的關聯、以便完全刪除與知識庫相關的所有資源。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從「知識庫」清查頁面中、選取您要刪除的知識庫。
3. 選取  並選取 * 管理知識庫 *。
4. 選取 * 動作 * 功能表、然後選取 * 刪除知識庫 *。
5. 在刪除知識庫對話方塊中、確認您要刪除它、然後選取 * 刪除 *。

結果

知識庫會從工作負載工廠移除，並刪除其相關的磁碟區。

管理 Amazon Q Business 連接器

為 Amazon Q Business 建立連接器之後，您可以檢視連接器詳細資料，修改連接器，整合其他資料來源，或刪除連接器。

檢視連接器的相關資訊

您可以檢視連接器和整合資料來源的設定資訊。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從工作負載原廠導覽功能表中，選取 * AI * 。
3. 選取您要檢視的連接器。
4. 若要檢視連接器詳細資料，請選取 [...](#) 並選取 * 管理連接器 * 。

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

「* 動作 *」功能表可讓您在需要進行任何變更時，管理連接器。

編輯連接器

您可以變更某些設定來更新連接器，也可以新增或移除資料來源。

每當您從連接器新增，修改或移除資料來源時，GenAI 都需要將資料來源資訊傳送至 Amazon Q Business，以便重新建立索引。同步是漸進式的，因此 Amazon Q Business 只會處理您的 FSX for ONTAP Volume 中自上次同步後已新增，修改或刪除的物件。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從「知識庫與連接器」庫存頁面中，選取您要更新的連接器。
3. 選取 [...](#) 並選取 * 管理連接器 * 。

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

4. 選取 * 動作 * 功能表，然後選取 * 編輯連接器 * 。
5. 在「編輯連接器」頁面中，您可以變更連接器名稱，說明，內嵌模型，啟用資料欄，以及用於包含連接器之磁碟區的快照原則。



每次資料來源掃描（包括內嵌）都會產生成本。如果您在建立連接器之後啟用資料 guardrails，則會再次掃描資料來源並產生成本。

6. 變更後請選擇 * 儲存 * 。

新增其他資料來源至連接器

您可以在連接器中內嵌其他資料來源，以便將其他組織資料填入連接器中。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從「知識庫與連接器」庫存頁面中，選取您要新增資料來源的連接器。
3. 選取  並選取 * 新增資料來源 * 。
4. 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。

如果您包含影像或 PDF 檔案，則 GenAI 的 BlueXP 工作負載工廠會剖析影像中的文字（包括 PDF 文件中的影像），這會產生較高的成本。

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。

如果您包含影像或 PDF 檔案，則 GenAI 的 BlueXP 工作負載工廠會剖析影像中的文字（包括 PDF 文件中的影像），這會產生較高的成本。

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。

如果您包含影像或 PDF 檔案，則 GenAI 的 BlueXP 工作負載工廠會剖析影像中的文字（包括 PDF 文件中的影像），這會產生較高的成本。

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

資料來源已整合至您的連接器。

將資料來源與連接器同步

資料來源會每天自動與相關的連接器同步一次，以便在 Amazon Q Business 中反映任何資料來源變更。如果您變更任何資料來源，而且想要立即同步（掃描）資料，您可以執行隨需同步。

同步是漸進式的，因此 Amazon Q Business 只會處理資料來源中自上次同步以來已新增，修改或刪除的物件。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從「知識庫與連接器」標籤中，選取您要同步的連接器。
3. 選取  並選取 * 管理連接器 *。
4. 選取 * 動作 * 功能表、然後選取 * 立即掃描 *。

您會看到一則訊息、指出正在掃描資料來源、以及掃描完成後的最後一則訊息。

結果

連接器會與附加的資料來源同步，Amazon Q Business 會開始使用您資料來源的最新資訊。

暫停或繼續排程的同步處理

如果您想要暫停或恢復資料來源的下次同步（掃描），您可以隨時進行。如果您要變更資料來源，而不想在變更期間進行同步處理，則可能需要暫停下次排程的同步處理。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從連接器庫存頁面中，選取您要暫停或恢復掃描的連接器。
3. 選取  並選取 * 管理連接器 *。
4. 選取 * 動作 * 功能表，然後選取 * 掃描 > 暫停排程掃描 * 或 * 掃描 > 恢復排程掃描 *。

您會看到訊息，表示下次排程掃描已暫停或恢復。

刪除連接器

如果您不再需要連接器，可以將其刪除。刪除連接器時，它會從工作負載工廠移除，並刪除包含連接器的磁碟區。刪除連接器是不可還原的。

刪除連接器時，您也應該解除連接器與任何與其相關的代理程式的關聯，以便完全刪除與連接器相關的所有資源。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從「知識庫與連接器」庫存頁面中，選取您要刪除的連接器。
3. 選取  並選取 * 管理連接器 *。
4. 選取 * 動作 * 功能表，然後選取 * 刪除連接器 *。
5. 在刪除連接器對話方塊中，確認您要刪除它，然後選取 * 刪除 *。

結果

連接器會從工作負載工廠移除，並刪除其相關的磁碟區。

管理 GenAI 資料來源

使用適用於 ONTAP 檔案系統的 FSX 資料來源建立知識庫或連接器之後，您可以檢視資料來源詳細資料，更新或變更資料來源內容，編輯資料來源設定，或刪除資料來源。

檢視資料來源的相關資訊

您可以檢視資料來源內容的相關資訊，也可以使用知識庫或連接器來檢視其內嵌狀態。由於資料來源與知識庫或連接器相關聯，因此您必須先選擇知識庫或連接器，才能檢視資料來源詳細資料。

步驟

1. 使用其中一項登入工作負載工廠 ["主控台體驗"](#)。
2. 從工作負載原廠導覽功能表中，選取 * AI *。
3. 選取資料來源所在的知識庫或連接器，然後選取 * 管理知識庫 * 或 * 管理連接  器 *。

頁面底部會列出相關的資料來源。

4. 選取以檢視每個資料來源的詳細資訊、例如 ONTAP 檔案系統的 FSX、磁碟區和資料來源所在的路徑、以展開每一列 。

它也會列出內嵌資訊，以及該資料來源目前是否內嵌於知識庫或連接器中。

編輯資料來源設定

您可以編輯與知識庫或連接器整合的資料來源相關資訊。在新增資料來源之後，大部分的資訊都是固定的，但您可以變更某些組態（例如分塊定義或權限感知）。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從知識庫庫存頁面中、選取資料來源所在的知識庫、然後選取  並選取 * 管理知識庫 *。

頁面底部會列出屬於此知識庫的資料來源。

3. 在您要編輯的資料來源列中、選取  並選取 * 編輯資料來源 * 。
4. 在「編輯資料來源」頁面中、選取  以展開區塊定義列。
5. 更新區塊策略和組態的設定、以及權限認知（SMB 磁碟區）、然後選取 * 儲存 * 。

結果

資料來源設定會更新、AI 系統會同步資料來源、以便將其重新索引至知識庫。

更新現有資料來源的內容

您可以隨時變更資料來源的內容、以新增或更新組織資料。如果此資料來源正在知識庫中積極使用、您必須同步資料來源、以便將其重新索引至知識庫。同步是漸進式的、因此 Amazon bedrock 只會處理您的 FSX for ONTAP Volume 中自上次同步以來已新增、修改或刪除的物件。

資料來源會每天自動與知識庫同步一次、以便在聊天機器人程式中反映任何資料來源變更。如果您對資料來源進行變更，並且想要立即同步處理資料，您可以 ["執行隨需同步"](#)。

刪除資料來源

如果您不再需要資料來源作為知識庫的一部分、可以將其刪除。

步驟

1. 使用其中一項登入工作負載工廠["主控台體驗"](#)。
2. 從知識庫庫存頁面中、選取資料來源所在的知識庫、然後選取  並選取 * 管理知識庫 * 。

頁面底部會列出屬於此知識庫的資料來源。

3. 在您要刪除的資料來源列中、選取  並選取 * 刪除資料來源 * 。
4. 在刪除資料來源對話方塊中、確認您要刪除資料來源、然後選取 * 確認 * 。

結果

資料來源會從知識庫中移除、AI 系統會從知識庫中移除此資料來源的索引資訊。來自該資料來源的任何資訊將不再提供給使用知識庫的聊天機器人程式。

使用 BlueXP 工作負載工廠中的 Tracker 監控工作負載操作

使用 BlueXP 工作負載工廠中的 Tracker，監控及追蹤工作負載作業的執行，並監控工作進度。

關於這項工作

工作負載工廠提供 Tracker 這項監控功能，讓您能夠監控及追蹤工作負載作業的進度和狀態，檢閱作業工作和子任務的詳細資料，以及診斷任何問題或故障。

Tracker 中有幾個動作可供使用。您可以依時間範圍（過去 24 小時，7 天，14 天或 30 天），工作負載，狀態和使用者來篩選工作；使用搜尋功能尋找工作；並將工作表下載為 CSV 檔案。您可以隨時重新整理 Tracker，快速重試失敗的作業或編輯失敗作業的參數，然後再試一次。

Tracker 支援兩種監控層級，視作業而定。每項工作（例如檔案系統部署）都會顯示工作說明，狀態，開始時間

，工作持續時間，使用者，區域， Proxy 資源，工作 ID 及所有相關的子工作。您可以檢視 API 回應，瞭解作業期間發生的情況。

以範例追蹤工作層級

- 層級 1（工作）：追蹤檔案系統部署。
- 層級 2（子工作）：追蹤與檔案系統部署相關的子工作。

作業狀態

Tracker 中的作業狀態如下：_ 進行中 _，_ 成功 _ 及 _ 失敗 _。

操作頻率

作業頻率是根據工作類型和工作排程而定。

事件保留

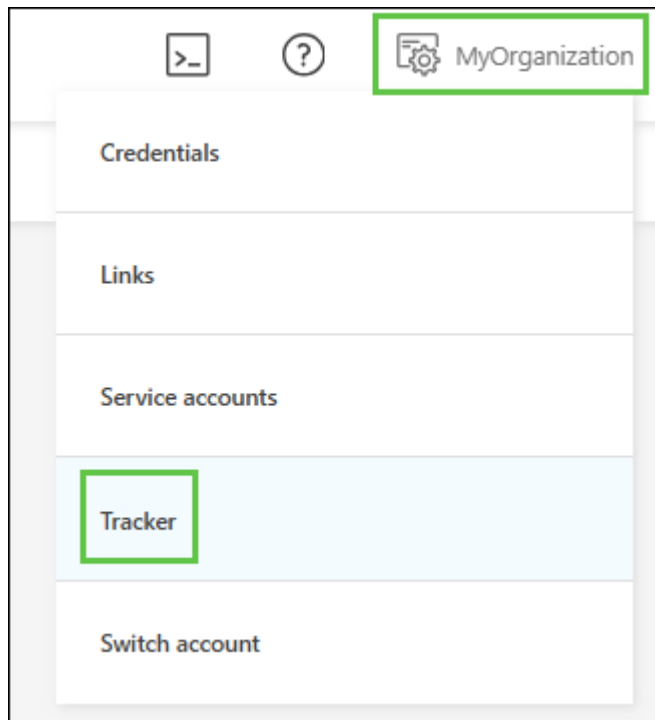
事件會保留在使用者介面上 30 天。

追蹤及監控作業

使用 Tracker 追蹤和監控 BlueXP 中的作業。

步驟

1. 使用其中一項登"主控台體驗"入。
2. 在工作負載中，選取帳戶設定功能表，然後選取 * Tracker *。



3. 在 Tracker 標籤中，使用篩選器或搜尋來縮小工作結果範圍。您也可以下載工作報告。

檢視 API 要求

在 Codebox 中檢視 Tracker 工作的 API 要求。

步驟

1. 在 Tracker 中，選取工作。
2. 選取三點功能表，然後選取 * 檢視 API 要求 *。

重試失敗的作業

在 Tracker 中重試失敗的作業。您也可以複製失敗作業的錯誤訊息。



您最多可以重試 10 次失敗的作業。

步驟

1. 在 Tracker 中，選取失敗的作業。
2. 選擇三點選單，然後選擇 * 重試 *。

結果

作業將重新啟動。

編輯並重試失敗的作業

編輯失敗作業的參數，然後在 Tracker 外重試作業。

步驟

1. 在 Tracker 中，選取失敗的作業。
2. 選取三點選單，然後選取 * 編輯並重試 *。

系統會將您重新導向至作業頁面，您可以在其中編輯參數，然後重試作業。

結果

作業將重新啟動。前往 Tracker 以檢視作業狀態。

知識與支援

註冊以取得 **GenAI BlueXP** 工作負載工廠的支援

需要註冊支援，才能獲得 BlueXP 工作負載工廠及其儲存解決方案與服務的專屬技術支援。您必須向 BlueXP 主控台註冊以取得支援，該主控台是獨立於工作負載工廠的網路型主控台。

註冊支援並不會啟用雲端供應商檔案服務的 NetApp 支援。如需雲端供應商檔案服務，其基礎架構或任何使用服務的解決方案的相關技術支援，請參閱該產品工作負載原廠文件中的「取得說明」。

["Amazon FSX for ONTAP Sf"](#)

支援登錄總覽

註冊您的帳戶 ID 支援訂閱（您的 20 位數 960xxxxxxxx 序號位於 BlueXP 的「支援資源」頁面）、即為您的單一支援訂閱 ID。每個 BlueXP 帳戶層級的支援訂閱都必須註冊。

註冊可提供開啟支援服務單和自動產生個案等功能。如下列所述、將 NetApp 支援網站（NSS）帳戶新增至 BlueXP 即可完成登錄。

註冊您的帳戶以取得 **NetApp** 支援

若要註冊以取得支援並啟動支援授權、您帳戶中的一位使用者必須將 NetApp 支援網站帳戶與其 BlueXP 登入建立關聯。您如何註冊 NetApp 支援取決於您是否已擁有 NetApp 支援網站一個 NetApp（NSS）帳戶。

現有的客戶、擁有一個新服務客戶帳戶

如果您是擁有 NSS 帳戶的 NetApp 客戶、您只需透過 BlueXP 註冊即可獲得支援。

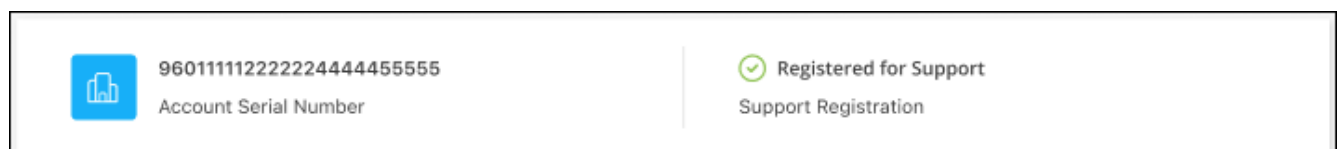
步驟

1. 在工作負載原廠主控台的右上角，選取 * 說明 > 支援 *。

選取此選項會開啟 BlueXP 主控台新的瀏覽器索引標籤、並載入支援儀表板。

2. 在 BlueXP 主控台的右上角、選取「設定」圖示、然後選取 * 認證 *。
3. 選取 * 使用者認證 *。
4. 選取 * 新增 NSS 認證 *、然後遵循 NetApp 支援網站（NSS）驗證提示。
5. 若要確認註冊程序是否成功、請選取「說明」圖示、然後選取 * 「支援 *」。

「* 資源 *」頁面應顯示您的帳戶已註冊以取得支援。



請注意、如果其他 BlueXP 使用者尚未將 NetApp 支援網站 帳戶與 BlueXP 登入建立關聯、則不會看到此相同的支援登錄狀態。不過、這並不表示您的 BlueXP 帳戶尚未註冊支援。只要帳戶中有一位使用者已遵循這些步驟、您的帳戶就已登錄。

現有客戶、但無NSS帳戶

如果您是現有的 NetApp 客戶、擁有現有的授權和序號、但沒有 NSS_ 帳戶、則需要建立一個 NSS 帳戶、並將其與您的 BlueXP 登入建立關聯。

步驟

1. 完成建立 NetApp 支援網站帳戶 "[《使用者登錄表》NetApp 支援網站](#)"
 - a. 請務必選擇適當的使用者層級、通常為* NetApp客戶/終端使用者*。
 - b. 請務必複製上述序號欄位使用的BlueXP帳戶序號（960xxxx）。這將加速帳戶處理。
2. 完成下的步驟、將新的 NSS 帳戶與 BlueXP 登入建立關聯 [\[現有的客戶、擁有一個新服務客戶帳戶\]](#)。

NetApp全新推出

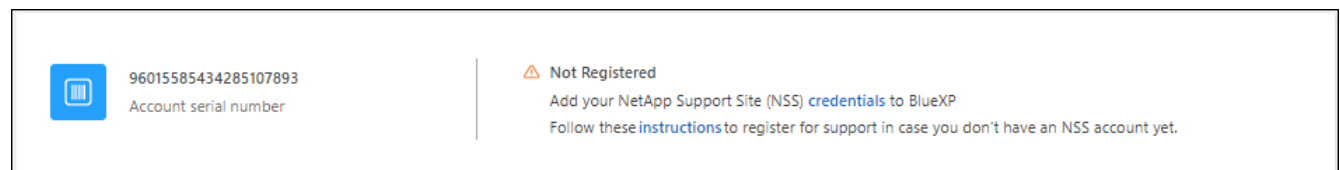
如果您是NetApp的新客戶、而且您沒有新的NSS帳戶、請依照下列每個步驟操作。

步驟

1. 在工作負載原廠主控台的右上角，選取 * 說明 > 支援 *。

選取此選項會開啟 BlueXP 主控台新的瀏覽器索引標籤、並載入支援儀表板。

2. 從「支援資源」頁面找到您的帳戶 ID 序號。



3. 瀏覽 "[NetApp的支援註冊網站](#)" 並選取 * 我不是 NetApp 的註冊客戶 *。
4. 填寫必填欄位（紅色星號）。
5. 在*產品系列*欄位中、選取* Cloud Manager*、然後選取適用的帳單供應商。
6. 複製上述步驟2的帳戶序號、完成安全性檢查、然後確認您已閱讀NetApp的全球資料隱私權政策。

系統會立即將電子郵件傳送至提供的信箱、以完成此安全交易。如果驗證電子郵件在幾分鐘內未送達、請務必檢查您的垃圾郵件資料夾。

7. 確認電子郵件中的行動。

確認將您的申請提交給NetApp、並建議您建立NetApp 支援網站 一個申請表。

8. 完成建立 NetApp 支援網站帳戶 "[《使用者登錄表》NetApp 支援網站](#)"
 - a. 請務必選擇適當的使用者層級、通常為* NetApp客戶/終端使用者*。
 - b. 請務必複製上述序號欄位使用的帳戶序號（960xxxx）。這將加速帳戶處理。

完成後

在此過程中、NetApp應與您聯絡。這是新使用者的一次性就職練習。

擁有 NetApp 支援網站帳戶後、請完成下的步驟、將帳戶與 BlueXP 登入建立關聯 [\[現有的客戶、擁有一個新服務客戶帳戶\]](#)。

GenAI 疑難排解

瞭解如何解決您可能遇到的一些常見問題。

常見問題與解決方案

如果您有上述其中一個問題、可以使用「因應措施」欄中的步驟來嘗試解決。

區域	問題	原因	因應措施
部署	部署失敗、因為磁碟區已經存在。	GenAI 的 BlueXP 工作負載工廠需要在部署過程中建立新的 Volume，但已使用您指定的名稱存在一個 Volume。	指定新磁碟區使用的唯一名稱、然後再次嘗試部署。
部署	由於 GenAI 的 BlueXP 工作負載工廠無法裝載磁碟區，因此部署失敗。	NetApp ONTAP 的 FSX 所需的一或多個輸入連接埠會關閉或篩選。	開啟下列輸入連接埠：

| 傳輸協定 | 連接埠 | 目的

| 所有 ICMP | 全部 | Ping 執行個體

| HTTPS | 443 | 從Connector存取fsxadmin管理LIF、將API呼叫傳送至FSX

| SSH | 22 | SSH 存取叢集管理 LIF 的 IP 位址或節點管理 LIF

| TCP | 111 | 遠端程序需要 NFS

| TCP | 139 | CIFS 的 NetBios 服務工作階段

| TCP | 161-162 | 簡單的網路管理傳輸協定

| TCP | 445 | Microsoft SMB/CIFS over TCP 搭配 NetBios 架構

| TCP | 635 | NFS 掛載

| TCP | 749 | Kerberos

| TCP | 2049 | NFS 伺服器精靈

| TCP | 3260 | 透過 iSCSI 資料 LIF 存取 iSCSI

| TCP | 4045 | NFS 鎖定精靈

| TCP | 4046 | NFS 的網路狀態監控

| TCP | 10000 | 使用 NDMP 備份

| TCP | 11104 | 管理 SnapMirror 的叢集間通訊工作階段

| TCP | 11105 | 使用叢集間生命體進行 SnapMirror 資料傳輸

| UDP | 111 | 遠端程序需要 NFS

| UDP | 161-162 | 簡單的網路管理傳輸協定

| UDP | 635 | NFS 掛載

| UDP | 2049 | NFS 伺服器精靈

| UDP | 4045 | NFS 鎖定精靈

| UDP | 4046 | NFS 的網路狀態監控

| UDP | 4049 | NFS rquotad 傳輸協定

維護	AI 引擎無法啟動，您會在 * 知識庫 * 頁面上看到「AI engine instance error」（AI 引擎執行個體錯誤）錯誤。	AI 引擎執行個體毀損或不存在。	選擇 * 重建 * 按鈕。GenAI 的 BlueXP 工作負載工廠會重建基礎架構，並顯示重建進度。完成後，您的知識庫會重新連線至重建的基礎架構，並顯示知識庫清單。
維護	AI 引擎無法啟動，您會在 * 知識庫 * 頁面上看到錯誤「GenAI 引擎執行個體已停止」。	AI 引擎執行個體未執行。	使用 AWS 管理主控台或 AWS CLI 啟動 AI 引擎執行個體。
維護	AI 引擎無法啟動，您會在 * 知識庫 * 頁面上看到錯誤「GenAI 引擎伺服器沒有回應」。	AI 引擎執行個體沒有回應。	請使用下列還原步驟： 步驟 1. 修改 GenAI 引擎執行個體安全性群組，以啟用對 GenAI 引擎執行個體的 SSH 存取。 2. 使用 SSH 登入執行個體。 3. 執行下列命令： <pre>docker- compose up</pre>

維護	GenAI 的 BlueXP 工作負載工廠所使用的後端 Docker 執行個體無法啟動。	磁碟區已刪除、EC2 執行個體已重新啟動。	請使用下列還原步驟： 步驟 1. 在適用於 NetApp ONTAP 的 FSX 上建立新的 Volume。例如、Volume 名稱可以是 netapp_ai /netapp_ai、而 Volume 路徑可以是。 2. SSH 至 Amazon EC2 執行個體。 3. 列出磁碟區： <pre>docker volume list</pre> 4. 移除舊 Volume： <pre>docker volume rm ec2- user_persistent_folder</pre> 5. `docker-compose.yml` 使用文字編輯器開啟檔案。 6. 在 `volumes` 區段中、將裝置路徑變更為新的 Volume 路徑。例如：
----	--	-----------------------	---

維護	GenAI 的 BlueXP 工作負載工廠所使用的後端 Docker 執行個體無法啟動。	根磁碟區已刪除。	使用名稱和路徑建立 Volume 、然後從 Amazon EC2 重新啟動後端 Docker 執行個體。
維護	GenAI 的 BlueXP 工作負載工廠所使用的後端 Docker 執行個體無法啟動。	根磁碟區已刪除。	使用名稱和路徑建立 Volume 、然後從 Amazon EC2 重新啟動後端 Docker 執行個體。

取得 GenAI BlueXP 工作負載工廠的協助

NetApp 以多種方式支援 BlueXP 工作負載工廠及其雲端服務。我們全年無休提供豐富的免費自助支援選項、例如知識庫 (KB) 文章和社群論壇。您的支援註冊包括透過網路票證提供遠端技術支援。

取得適用於 ONTAP 的 FSX 支援

如需與適用於 ONTAP 的 FSX，其基礎架構或使用服務的任何解決方案相關的技術支援，請參閱該產品工作負載原廠文件中的「取得說明」。

["Amazon FSX for ONTAP Sf"](#)

若要獲得工作負載工廠及其儲存解決方案與服務的專屬技術支援、請使用下列支援選項。

使用自我支援選項

這些選項可供免費使用、一天24小時、一週7天：

- 文件
您目前正在檢視的工作負載原廠文件。
- ["知識庫"](#)
搜尋工作負載原廠知識庫，找出有助於疑難排解問題的文章。
- ["社群"](#)
加入工作負載工廠社群，追蹤後續討論或建立新討論。

利用NetApp支援建立案例

除了上述的自我支援選項、您也可以在啟動支援之後、與NetApp支援專家合作解決任何問題。

開始之前

若要使用 * 建立案例 * 功能，您必須先註冊以取得支援。將您的 NetApp 支援網站認證與工作負載原廠登入建立關聯。["瞭解如何註冊以取得支援"](#)。

步驟

```
'nfs'
  o:
    "addr=svm-0d6673008aaca12bc3.\
    fs-0d673008aaca12bc3.\
    fs-us-east-1.amazonaws.com,no\
    lock,soft,rw"
  device:
    ':/netapp_ai'
    ' # Path to new volume'
```

1. 在工作負載原廠主控台的右上角，選取 * 說明 > 支援 * 。

選取此選項會開啟 BlueXP 主控台新的瀏覽器索引標籤、並載入支援儀表板。

2. 在「資源」頁面上、選擇「技術支援」下的其中一個可用選項：

- a. 如果您想與電話上的某人通話、請選取 * 致電 * 。您將會被導向netapp.com上的頁面、其中列出您可以撥打的電話號碼。

- b. 選擇 * 建立案例 * 、與 NetApp 支援專家一起開啟 Ticket ：

- * 服務 * ：選擇 * 工作負載工廠 * 。
- 案例優先順序：選擇案例的優先順序、可以是低、中、高或嚴重。

若要深入瞭解這些優先順序、請將滑鼠游標暫留在欄位名稱旁的資訊圖示上。

- 問題說明：提供問題的詳細說明、包括任何適用的錯誤訊息或您執行的疑難排解步驟。
- 其他電子郵件地址：如果您想讓其他人知道此問題、請輸入其他電子郵件地址。
- * 附件（選填） * ：上傳最多五個附件、一次上傳一個。

每個檔案的附件上限為 25 MB 。支援下列副檔名：txt 、 log 、 pdf 、 jpg/jpeg 、 rtf 、 doc/dox 、 xls/xlsx 和 csv 。

ntapitdemo 

NetApp Support Site Account

Service

Select ▼

Working Enviroment

Select ▼

Case Priority 

Low - General guidance ▼

Issue Description

Provide detailed description of problem, applicable error messages and troubleshooting steps taken.

Additional Email Addresses (Optional) 

Type here

Attachment (Optional)

No files selected

 Upload 

完成後

您的支援案例編號會出現快顯視窗。NetApp支援專家將會審查您的案例、並盡快回覆您。

如需支援案例的記錄、您可以選取 * 設定 > 時間軸 *、然後尋找名為「建立支援案例」的動作。最右側的按鈕可讓您展開動作以查看詳細資料。

嘗試建立案例時、可能會遇到下列錯誤訊息：

"您無權針對所選服務建立案例"

此錯誤可能表示、與該帳戶相關聯的NSS帳戶及記錄公司與BlueXP帳戶序號的記錄公司不同（例如960xxxx）或工作環境序號。您可以使用下列其中一個選項尋求協助：

- 使用產品內對談
- 請至提交非技術案例 <https://mysupport.netapp.com/site/help>

管理支援案例（預覽）

您可以直接從BlueXP檢視及管理作用中和已解決的支援案例。您可以管理與您的NSS帳戶和貴公司相關的個案。

案例管理可透過預覽取得。我們計畫改善這項體驗、並在即將推出的版本中加入增強功能。請使用產品內建聊天功能、向我們傳送意見反應。

請注意下列事項：

- 頁面頂端的案例管理儀表板提供兩種檢視：
 - 左側檢視顯示您所提供的使用者nssc帳戶在過去3個月內開啟的個案總數。
 - 右側檢視顯示過去3個月內、貴公司層級根據您的使用者nssc帳戶所開啟的個案總數。表格中的結果會反映您所選檢視的相關個案。
- 您可以新增或移除感興趣的欄、也可以篩選優先順序和狀態等欄的內容。其他欄則只提供排序功能。如需詳細資料、請參閱下列步驟。
- 在個別案例層級、我們提供更新案例附註或關閉尚未處於「已結案」或「待結案」狀態的案例的功能。

步驟

1. 在工作負載原廠主控台的右上角，選取 * 說明 > 支援 * 。

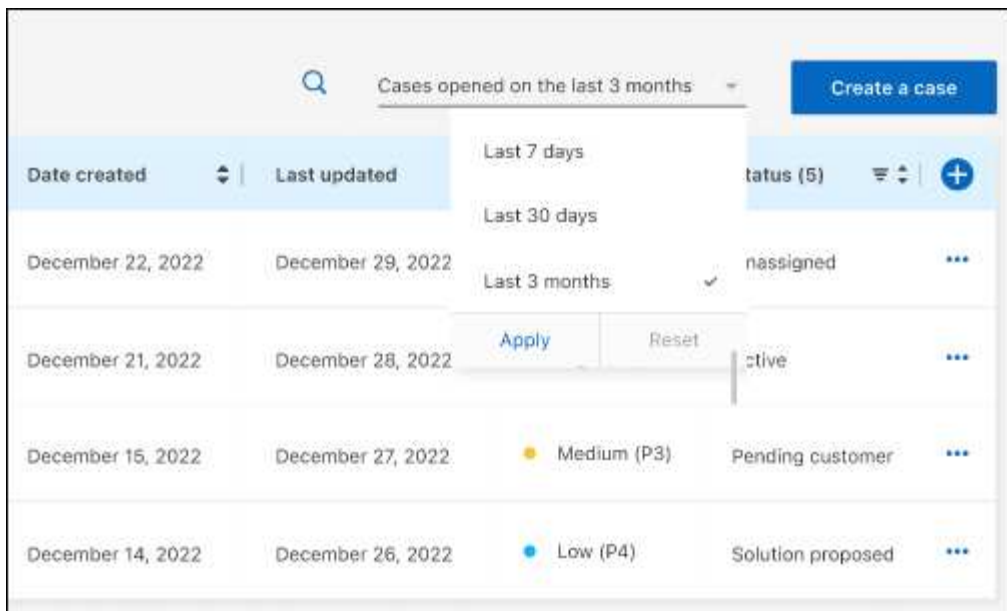
選取此選項會開啟 BlueXP 主控台新的瀏覽器索引標籤、並載入支援儀表板。

2. 選取 * 個案管理 * 、如果出現提示、請將您的 NSS 帳戶新增至 BlueXP 。

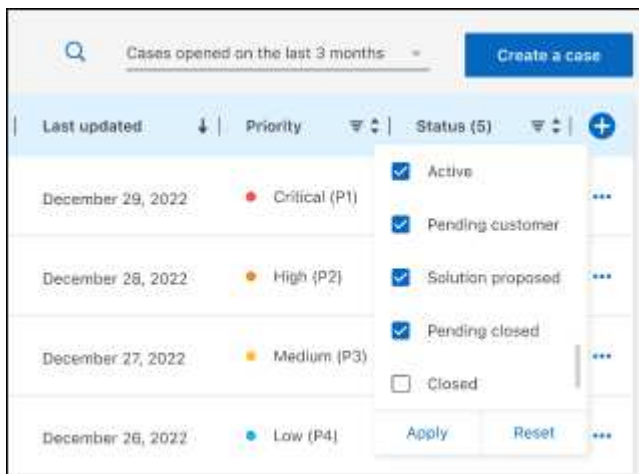
「個案管理」頁面會顯示與您的BlueXP使用者帳戶相關聯的與NSS帳戶相關的未決個案。這是顯示在「* nssmanagement *」頁面頂端的相同nss.帳戶。

3. （可選）修改表格中顯示的資訊：

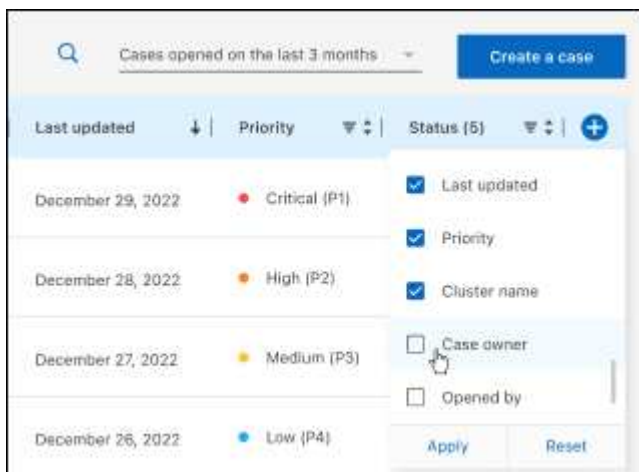
- 在 * 組織案例 * 下、選取 * 檢視 * 以檢視與貴公司相關的所有案例。
- 選擇確切的日期範圍或選擇不同的時間範圍、以修改日期範圍。



。篩選欄的內容。



。選取要顯示的欄、然後選擇要顯示的欄、即可變更表格中  顯示的欄。

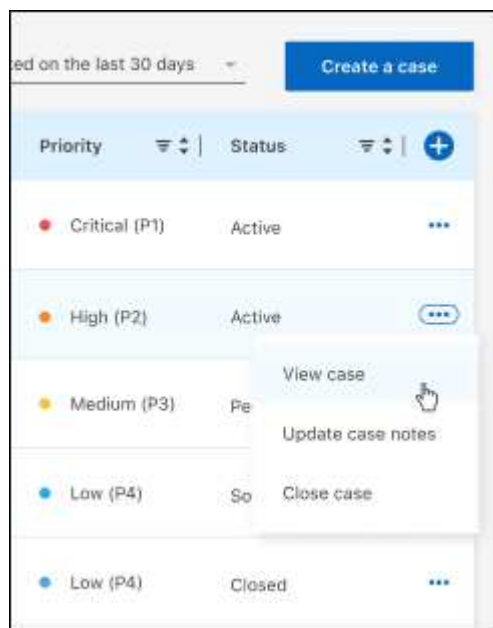


4. 選取並選取其中一個可用選項、以管理現有案例 ... :

- 檢視案例：檢視特定案例的完整詳細資料。
- * 更新案例附註 *：提供問題的其他詳細資料、或選擇 * 上傳檔案 * 最多附加五個檔案。

每個檔案的附件上限為 25 MB。支援下列副檔名：txt、log、pdf、jpg/jpeg、rtf、doc/dox、xls/xlsx 和 csv。

- * 結案案例 *：提供結案原因的詳細資料、並選取 * 結案案例 *。



GenAI 法律聲明的 BlueXP 工作負載工廠

法律聲明提供版權聲明、商標、專利等存取權限。

版權

["https://www.netapp.com/company/legal/copyright/"](https://www.netapp.com/company/legal/copyright/)

商標

NetApp、NetApp 標誌及 NetApp 商標頁面上列出的標章均為 NetApp、Inc. 的商標。其他公司與產品名稱可能為其各自所有者的商標。

["https://www.netapp.com/company/legal/trademarks/"](https://www.netapp.com/company/legal/trademarks/)

專利

如需最新的 NetApp 擁有專利清單、請參閱：

<https://www.netapp.com/pdf.html?item=/media/11887-patentspage.pdf>

隱私權政策

["https://www.netapp.com/company/legal/privacy-policy/"](https://www.netapp.com/company/legal/privacy-policy/)

開放原始碼

通知檔案提供有關 NetApp 軟體所使用之協力廠商版權與授權的資訊。

["BlueXP 工作負載工廠"](#)

版權資訊

Copyright © 2025 NetApp, Inc. 版權所有。台灣印製。非經版權所有人事先書面同意，不得將本受版權保護文件的任何部分以任何形式或任何方法（圖形、電子或機械）重製，包括影印、錄影、錄音或儲存至電子檢索系統中。

由 NetApp 版權資料衍伸之軟體必須遵守下列授權和免責聲明：

此軟體以 NETAPP「原樣」提供，不含任何明示或暗示的擔保，包括但不限於有關適售性或特定目的適用性之擔保，特此聲明。於任何情況下，就任何已造成或基於任何理論上責任之直接性、間接性、附隨性、特殊性、懲罰性或衍生性損害（包括但不限於替代商品或服務之採購；使用、資料或利潤上的損失；或企業營運中斷），無論是在使用此軟體時以任何方式所產生的契約、嚴格責任或侵權行為（包括疏忽或其他）等方面，NetApp 概不負責，即使已被告知有前述損害存在之可能性亦然。

NetApp 保留隨時變本文所述之任何產品的權利，恕不另行通知。NetApp 不承擔因使用本文所述之產品而產生的責任或義務，除非明確經過 NetApp 書面同意。使用或購買此產品並不會在依據任何專利權、商標權或任何其他 NetApp 智慧財產權的情況下轉讓授權。

本手冊所述之產品受到一項（含）以上的美國專利、國外專利或申請中專利所保障。

有限權利說明：政府機關的使用、複製或公開揭露須受 DFARS 252.227-7013（2014 年 2 月）和 FAR 52.227-19（2007 年 12 月）中的「技術資料權利 - 非商業項目」條款 (b)(3) 小段所述之限制。

此處所含屬於商業產品和 / 或商業服務（如 FAR 2.101 所定義）的資料均為 NetApp, Inc. 所有。根據本協議提供的所有 NetApp 技術資料和電腦軟體皆屬於商業性質，並且完全由私人出資開發。美國政府對於該資料具有非專屬、非轉讓、非轉授權、全球性、有限且不可撤銷的使用權限，僅限於美國政府為傳輸此資料所訂合約所允許之範圍，並基於履行該合約之目的方可使用。除非本文另有規定，否則未經 NetApp Inc. 事前書面許可，不得逕行使用、揭露、重製、修改、履行或展示該資料。美國政府授予國防部之許可權利，僅適用於 DFARS 條款 252.227-7015(b)（2014 年 2 月）所述權利。

商標資訊

NETAPP、NETAPP 標誌及 <http://www.netapp.com/TM> 所列之標章均為 NetApp, Inc. 的商標。文中所涉及的所有其他公司或產品名稱，均為其各自所有者的商標，不得侵犯。