



開始使用 GenAI

NetApp
October 06, 2025

目錄

開始使用	1
GenAI 連接器快速入門	1
GenAI 連接器需求	1
基本 GenAI 需求	2
NetApp Connector for Amazon Q Business 的要求	2
識別要新增至連接器的資料來源	3
資料來源的最大數量	3
資料來源的位置	3
支援的傳輸協定	3
支援的資料來源檔案格式	3
部署 GenAI 基礎架構	4
基礎架構詳細資料	4
部署 GenAI 基礎架構	5

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 Workload Factory 上建立帳戶"](#)並使用以下任一方式登入 ["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 Workload Factory 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。

工作負載工廠登入和帳戶

您需要 ["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"](#)

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
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

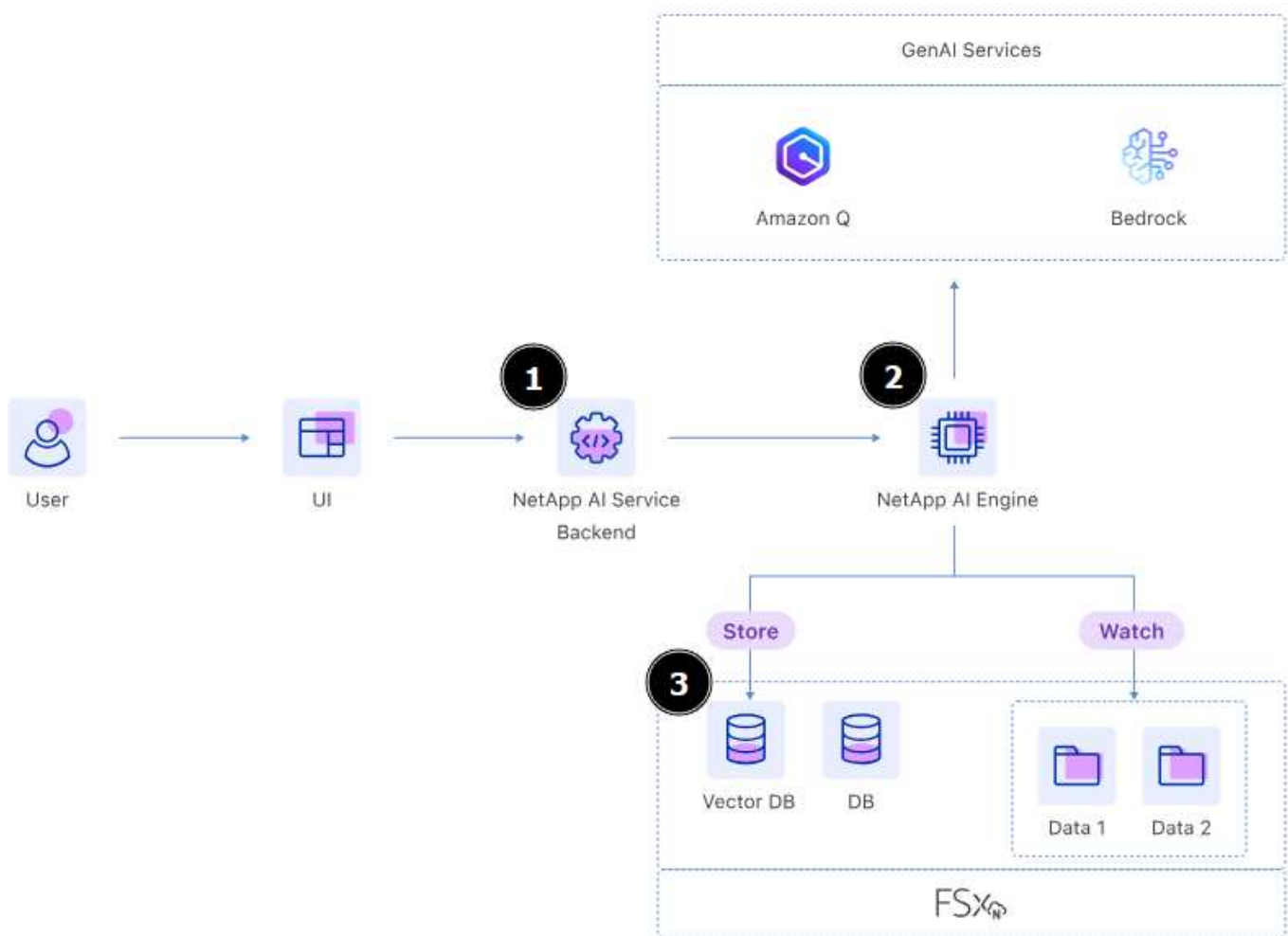
適用於 ONTAP 檔案系統的 FSX 可為您的 GenAI 系統提供儲存設備。

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。

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。