



了解 NetApp Workload Factory for GenAI GenAI

NetApp
October 06, 2025

This PDF was generated from <https://docs.netapp.com/zh-tw/workload-genai/general/ai-workloads-overview.html> on October 06, 2025. Always check docs.netapp.com for the latest.

目錄

了解NetApp Workload Factory for GenAI	1
了解NetApp Workload Factory for GenAI	1
NetApp Workload Factory for GenAI 是什麼？	1
使用 GenAI 建立泛用 AI 應用程式的優點	1
GenAI 的運作方式	2
NetApp Workload Factory for GenAI 如何協助建立生成式 AI 應用程式	2
使用NetApp Workload Factory 的工具	3
成本	3
授權	3
區域	3
NetApp GenAI 引擎的元件	4

了解NetApp Workload Factory for GenAI

了解NetApp Workload Factory for GenAI

NetApp Workload Factory for GenAI 讓您能夠將Amazon FSx for NetApp ONTAP檔案系統與 GenAI 基礎模型整合。這為您的 AI 資料集提供了高效能儲存以及豐富的保護、安全和成本優化功能。

NetApp Workload Factory for GenAI 是什麼？

NetApp Workload Factory for GenAI 讓您能夠將Amazon FSx for NetApp ONTAP上的企業資料來源與生成式 AI 應用程式結合使用。利用檢索增強生成 (RAG)，您可以快速將資料來源連接到可透過 Amazon Bedrock 或 Amazon Q Business 獲得的基礎模型，以開發生成式 AI 驅動的應用程式，例如虛擬助理、問答聊天機器人、文件摘要、內容創建等。

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。

有關 Workload Factory 的更多信息，請參閱 "[工作負載工廠概述](#)"。

使用 GenAI 建立泛用 AI 應用程式的優點

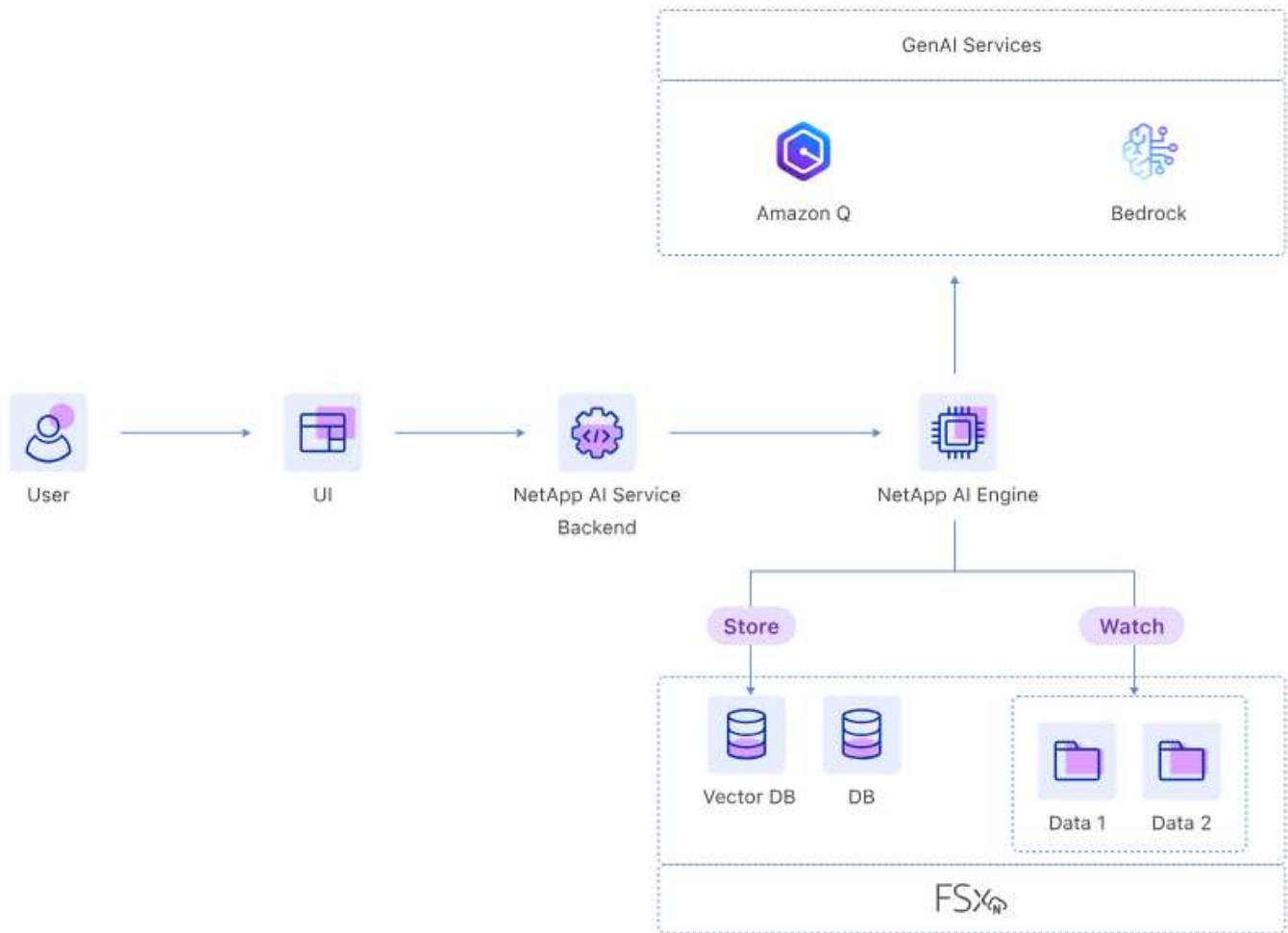
NetApp Workload Factory for GenAI 簡化了使用檢索增強生成 (RAG) 建立生成式 AI 應用程式所需的基礎架構的部署過程。具體來說，GenAI 具有以下優勢：

- 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。



NetApp Workload Factory for GenAI 如何協助建立生成式 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 NetApp Workload Factory 的工具

您可以將 NetApp Workload Factory 與以下工具一起使用：

- **Workload Factory** 控制台：Workload Factory 控制台提供您的應用程式和專案的視覺化、整體視圖。
- * NetApp 控制台*：NetApp 控制台提供混合介面體驗，以便您可以將 Workload Factory 與其他 NetApp 資料服務一起使用。
- 問我：使用問我 AI 助理來提問並了解有關 Workload Factory 的更多信息，而無需離開 Workload Factory 控制台。從 Workload Factory 幫助選單中存取「問我」。
- **CloudShell CLI**：Workload Factory 包含 CloudShell CLI，可透過基於瀏覽器的單一 CLI 跨帳號管理和操作 AWS 和 NetApp 環境。從 Workload Factory 控制台頂部欄存取 CloudShell。
- **REST API**：使用 Workload Factory REST API 部署和管理您的 FSx for ONTAP 檔案系統和其他 AWS 資源。
- **CloudFormation**：使用 AWS CloudFormation 程式碼執行您在 Workload Factory 控制台中定義的操作，以從您的 AWS 帳戶中的 CloudFormation 堆疊對 AWS 和第三方資源進行建模、配置和管理。
- **Terraform NetApp Workload Factory** 提供者：使用 Terraform 建置和管理在 Workload Factory 控制台中產生的基礎架構工作流程。

成本

使用 Workload Factory 的 GenAI 功能無需任何成本。

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

無需 NetApp 的特殊授權即可使用 Workload Factory 的 AI 功能。

區域

所有支援 FSx for ONTAP 的商業區域均支援 Workload Factory。["看看支援的亞馬遜區域。"](#)

以下 AWS 區域不受支援：

- 中國地區
- GovCloud（美國）區域

- 秘密雲
- 絕密雲

NetApp GenAI 引擎的元件

當您部署 GenAI 基礎架構時，Workload Factory 會為 GenAI 引擎建立一個 EC2 執行個體。它還為此實例建立 IAM 角色、安全群組和私有端點。您可能希望了解有關 Workload Factory 在您的 AWS 環境中建立的這些元件的更多詳細資訊。

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": [
        "iam:CreateRole",
        "iam:CreatePolicy",
        "iam:AttachRolePolicy",
        "iam:PassRole"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "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:<id>:kbId": "*"
        }
    },
    "Action": "fsx>DeleteVolume",
    "Resource": [
        "arn:aws:fsx:*:*:volume/*/*",
        "arn:aws:fsx:*:*:backup/*"
    ],
    "Effect": "Allow"
},
{
    "Condition": {
        "StringLike": {
            "aws:ResourceTag/netapp:wlmai:<id>:qConnectorId": "*"

```

```

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

```

```

        "bedrock:PutModelInvocationLoggingConfiguration"
    ],
    "Resource": "*",
    "Effect": "Allow"
},
{
    "Action": [
        "ec2messages:GetMessages",
        "ec2messages:GetEndpoint",
        "ec2messages:AcknowledgeMessage",
        "ec2messages:DeleteMessage",
        "ec2messages:FailMessage",
        "ec2messages:SendReply"
    ],
    "Resource": "*",
    "Effect": "Allow"
},
{
    "Action": [
        "qbusiness:ListWebExperiences",
        "qbusiness:ListApplications",
        "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:/aws/bedrock*",
        "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 還沒有它們，Workload Factory 會為 GenAI 引擎 EC2 執行個體建立私有端點，以便它可以與下列 AWS 服務通訊：

- 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

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。