



了解**NetApp Workload Factory for GenAI** GenAI

NetApp
October 06, 2025

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

目录

- 了解NetApp Workload Factory for GenAI 1
 - 了解NetApp Workload Factory for GenAI 1
 - NetApp Workload Factory for GenAI 是什么? 1
 - 使用GenAI创建生成性AI应用程序的优势 1
 - GenAI的工作原理 2
 - NetApp Workload Factory for GenAI 如何帮助构建生成式 AI 应用程序 2
 - 使用NetApp Workload Factory 的工具 3
 - 成本 3
 - 许可 3
 - 区域 3
 - NetApp GenAI引擎的组件 4

了解NetApp Workload Factory for GenAI

了解NetApp Workload Factory for GenAI

NetApp Workload Factory for GenAI 使您能够将Amazon FSx for NetApp ONTAP文件系统与 GenAI 基础模型集成。这为您的 AI 数据集提供了高性能存储以及丰富的保护、安全和成本优化功能。

NetApp Workload Factory for GenAI 是什么？

NetApp Workload Factory for GenAI 使您能够将Amazon FSx for NetApp ONTAP上的企业数据源与生成式 AI 应用程序结合使用。利用检索增强生成 (RAG)，您可以快速将数据源连接到可通过 Amazon Bedrock 或 Amazon Q Business 获得的基础模型，以开发生成式 AI 驱动的应用程序，例如虚拟助手、问答聊天机器人、文档摘要、内容创建等。

将生成式AI与组织数据结合使用、您可以利用自己的知识和专业知识、而不仅仅是依靠模型基于模型所训练的公共数据的智能。使用RAG自定义模型可确保对组织特定问题做出准确且相关的响应、从而提高使用生成性AI的应用程序用户的工作效率。

开发一款专为企业数据定制的GenAI应用程序、让您可以利用自己的知识和专业知识。这种自定义功能可确保准确、相关地回答组织特定的问题、从而提高所有用户的满意度和工作效率。

如果“[创建知识库](#)”，GenAI将从数据源中获取数据，将向量化结果存储在数据库中，并使您能够完全控制如何使用所获取的数据来回答查询。这种方法需要更多的初始配置、但允许您为不同的结果选择不同的聊天模式。如果是“[为 Amazon Q Business 定义 NetApp 连接器](#)”，则数据源中的数据将由Amazon Q Business读取并存储在索引中。这种方法所需的初始配置较少、但对结果的控制较少。

有关 Workload Factory 的更多信息，请参阅 [“工作负载工厂概述”](#)。

使用GenAI创建生成性AI应用程序的优势

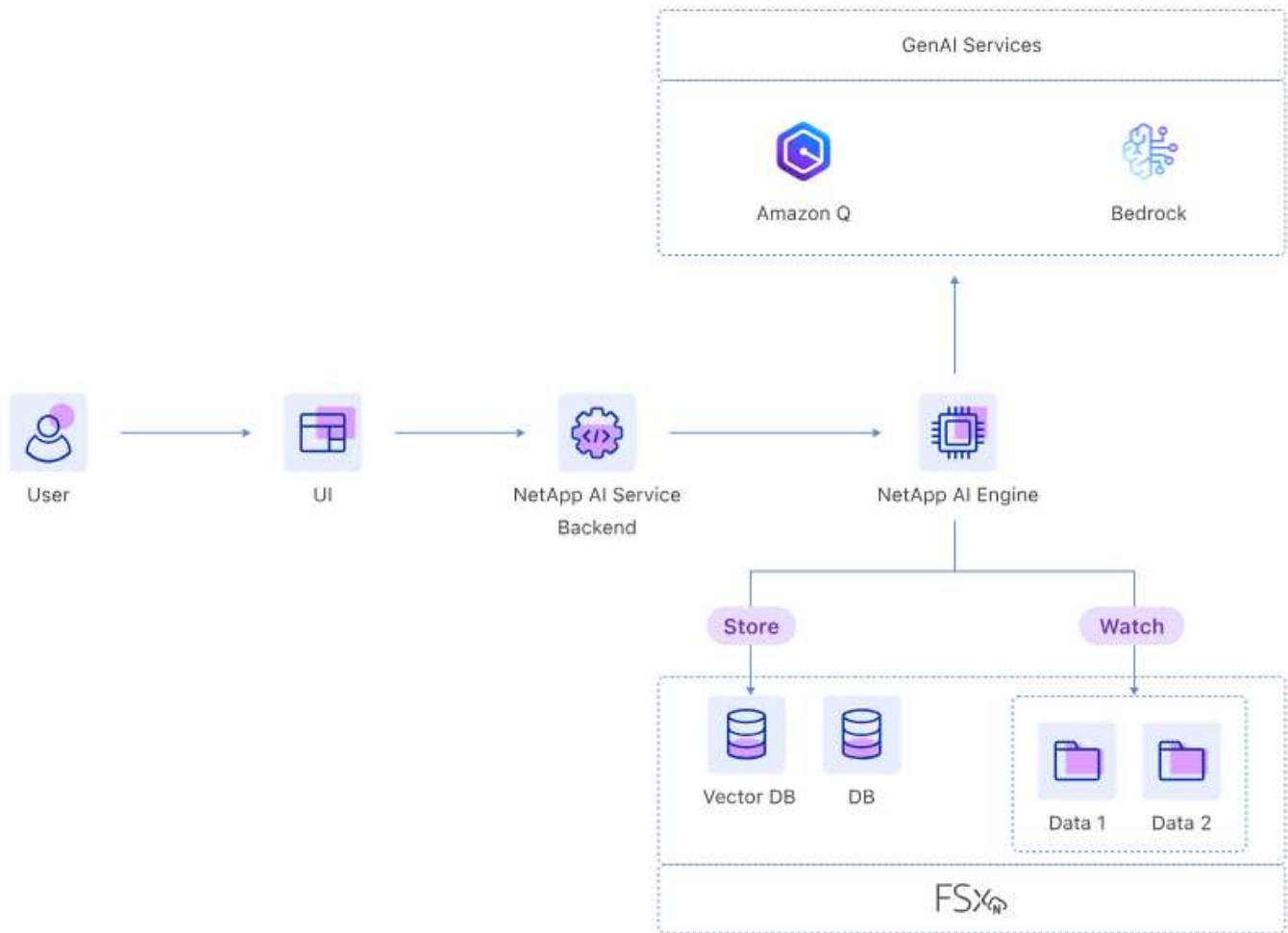
NetApp Workload Factory for GenAI 简化了使用检索增强生成 (RAG) 构建生成式 AI 应用程序所需的基础架构的部署过程。具体来说，GenAI 具有以下优势：

- IT管理员和开发人员无需深入了解数据基础架构、基础架构和语言模型、即可利用GenAI提供的自动化功能加快应用程序开发速度。数据管理员和开发人员可以轻松快速地创建企业知识库、以嵌入贵组织的非结构化数据、供生成式AI应用程序使用。
- 通过在知识库中嵌入的文件中保留用户权限来增强安全性、以确保维护数据安全和隐私。可以开发一个应用程序(例如聊天机器人)、以便仅为经过身份验证的用户提供基于用户有权访问的数据的答案。
- 在您的AWS客户帐户中保护企业数据的隐私和安全、您的组织数据绝不会对外公开。
- 使用开源框架(如LangChain)加快GenAI应用程序(如问答聊天机器人)的开发速度、利用GenAI API配置和管理知识库和连接器、与知识库聊天以及存储和检索聊天历史记录。
- 通过在适用于NetApp ONTAP文件系统的FSx上部署生成式AI数据基础架构并利用ONTAP功能(例如、高可用性、用于本地数据保护和恢复的快照、用于灾难恢复的SnapMirror以及用于备份数据基础架构的SnapVault)、改善数据保护和可用性状况。
- 利用重复数据删除、数据压缩和数据缩减、数据分层和精简配置等ONTAP数据效率功能、降低生成式AI数据基础架构的整体存储成本。

- 利用GenAI提供的混合搜索和重新排序功能、从数据中获得高质量的结果。混合搜索与重新排名相结合、可显著提高搜索结果的相关性。这些功能可通过Amazon AWS获得、并且取决于区域。

GenAI的工作原理

GenAI使用您组织的私有数据来补充模型的智能(基于模型所训练的数据)、为您组织中的用户提出的问题提供自定义的答案。您首先部署RAG框架所需的基础架构、然后使用您组织的数据源和基础模型(可通过Amazon Brock或Amazon Q Business获得)构建知识库或定义连接器、最后将应用程序(例如问答聊天机器人)连接到知识库或连接器。



NetApp Workload Factory for GenAI 如何帮助构建生成式 AI 应用程序

GenAI可通过以下方式帮助使用RAG构建生成性AI应用程序：

- 部署用于检索增强生成(RAG)框架的所需基础架构、以便与FSx上适用于ONTAP文件系统的数据源以及Amazon Brock或Amazon Q Business结合使用。该基础架构包括用于管理数据的NetApp GenAI引擎实例、嵌入式矢量数据库(LanceDB)以及Vector数据库FSx for ONTAP文件系统上的存储。
- 帮助将数据源连接到Amazon Brock或Amazon Q Business提供的嵌入和语言模型、以嵌入数据源并检索用户查询的响应。数据源及其型号及其配置以FSx for ONTAP知识库的形式提供。
- 将源数据内嵌到知识库或连接器中、以在SMB共享上嵌入源文件、并在适用于ONTAP文件系统的FSx上嵌入NFS导出、同时存储SMB共享中文件的文件权限。

- 根据知识库中的内容自动构建对话开场问题。
- 为数据管理员提供了一个聊天模拟器、用于测试与知识库的聊天。
- 提供一个简单的连接器接口、便于您将GenAI与Amazon Q Business连接起来、快速轻松地利用此AI助手的功能。

使用NetApp Workload Factory 的工具

您可以将NetApp Workload Factory 与以下工具一起使用：

- **Workload Factory** 控制台：Workload Factory 控制台提供您的应用程序和项目的可视化、整体视图。
- * NetApp控制台*： NetApp控制台提供混合界面体验，以便您可以将 Workload Factory 与其他NetApp数据服务一起使用。
- 问我：使用问我 AI 助手来提问并了解有关 Workload Factory 的更多信息，而无需离开 Workload Factory 控制台。从 Workload Factory 帮助菜单访问“问我”。
- **CloudShell CLI**：Workload Factory 包含一个 CloudShell CLI，可通过基于浏览器的单个 CLI 跨账户管理和操作 AWS 和NetApp环境。从 Workload Factory 控制台顶部栏访问 CloudShell。
- **REST API**：使用 Workload Factory REST API 部署和管理您的 FSx for ONTAP文件系统和其他 AWS 资源。
- **CloudFormation**：使用 AWS CloudFormation 代码执行您在 Workload Factory 控制台中定义的操作，以从您的 AWS 账户中的 CloudFormation 堆栈对 AWS 和第三方资源进行建模、配置和管理。
- **Terraform NetApp Workload Factory** 提供商：使用 Terraform 构建和管理在 Workload Factory 控制台中生成的基础架构工作流。

成本

使用 Workload Factory 的 GenAI 功能无需任何成本。

但是、要支持生成性AI基础架构、您需要为部署的AWS资源付费。例如、您需要为Amazon Brock或Amazon Q Business支付AWS费用、为ONTAP文件系统和存储容量支付FSx费用以及为GenAI引擎EC2实例支付费用。

某些多模式操作(如扫描图像以查找文本信息)可能会占用更多资源、因此成本会更高。某些配置操作(例如更改数据库的设置)可能会导致重新扫描数据源、数据源扫描的成本也会更高。

许可

无需NetApp的特殊许可即可使用 Workload Factory 的 AI 功能。

区域

所有支持 FSx for ONTAP 的商业区域均支持 Workload Factory。"[查看支持的亚马逊区域。](#)"

以下 AWS 区域不受支持：

- 中国地区
- GovCloud（美国）区域
- 秘密云

- 绝密云

NetApp GenAI引擎的组件

当您部署 GenAI 基础设施时，Workload Factory 会为 GenAI 引擎创建一个 EC2 实例。它还为此实例创建 IAM 角色、安全组和私有端点。您可能希望了解有关 Workload Factory 在您的 AWS 环境中创建的这些组件的更多详细信息。

EC2实例类型

m5.large

IAM角色

GenAI引擎实例需要具有向Amazon Brock上的嵌入模型发送数据区块以及与NetApp AI服务后端进行通信的权限。IAM角色包括以下权限：

```

{
  "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"
}
]
}

```

安全组

出站规则对所有流量开放、而入站规则完全关闭。

专用端点

如果目标 VPC 还没有它们，Workload Factory 会为 GenAI 引擎 EC2 实例创建私有端点，以便它可以与以下 AWS 服务通信：

- 亚马逊基岩
 - 基岩
 - 基岩-运行时
 - 基岩-代理-运行时
- Amazon ECR 容器注册表(ECR)
 - API
 - Docker
- AWS Systems Manager (SSM)

- SSM
 - ec2messages
 - Ssmmessages
- 适用于 NetApp ONTAP 的 Amazon FSX
- Amazon CloudWatch

版权信息

版权所有 © 2025 NetApp, Inc.。保留所有权利。中国印刷。未经版权所有者事先书面许可，本文档中受版权保护的任何部分不得以任何形式或通过任何手段（图片、电子或机械方式，包括影印、录音、录像或存储在电子检索系统中）进行复制。

从受版权保护的 NetApp 资料派生的软件受以下许可和免责声明的约束：

本软件由 NetApp 按“原样”提供，不含任何明示或暗示担保，包括但不限于适销性以及针对特定用途的适用性的隐含担保，特此声明不承担任何责任。在任何情况下，对于因使用本软件而以任何方式造成的任何直接性、间接性、偶然性、特殊性、惩罚性或后果性损失（包括但不限于购买替代商品或服务；使用、数据或利润方面的损失；或者业务中断），无论原因如何以及基于何种责任理论，无论出于合同、严格责任或侵权行为（包括疏忽或其他行为），NetApp 均不承担责任，即使已被告知存在上述损失的可能性。

NetApp 保留在不另行通知的情况下随时对本文档所述的任何产品进行更改的权利。除非 NetApp 以书面形式明确同意，否则 NetApp 不承担因使用本文档所述产品而产生的任何责任或义务。使用或购买本产品不表示获得 NetApp 的任何专利权、商标权或任何其他知识产权许可。

本手册中描述的产品可能受一项或多项美国专利、外国专利或正在申请的专利的保护。

有限权利说明：政府使用、复制或公开本文档受 DFARS 252.227-7013（2014 年 2 月）和 FAR 52.227-19（2007 年 12 月）中“技术数据权利 — 非商用”条款第 (b)(3) 条规定的限制条件的约束。

本文档中所含数据与商业产品和/或商业服务（定义见 FAR 2.101）相关，属于 NetApp, Inc. 的专有信息。根据本协议提供的所有 NetApp 技术数据和计算机软件具有商业性质，并完全由私人出资开发。美国政府对这些数据的使用权具有非排他性、全球性、受限且不可撤销的许可，该许可既不可转让，也不可再许可，但仅限在与交付数据所依据的美国政府合同有关且受合同支持的情况下使用。除本文档规定的情形外，未经 NetApp, Inc. 事先书面批准，不得使用、披露、复制、修改、操作或显示这些数据。美国政府对国防部的授权仅限于 DFARS 的第 252.227-7015(b)（2014 年 2 月）条款中明确的权利。

商标信息

NetApp、NetApp 标识和 <http://www.netapp.com/TM> 上所列的商标是 NetApp, Inc. 的商标。其他公司和产品名称可能是其各自所有者的商标。