



开始使用 GenAI

NetApp
October 06, 2025

目录

- 开始使用 1
 - GenAI连接器快速入门 1
 - GenAI连接器要求 1
 - 基本GenAI要求 1
 - NetApp Connector for Amazon Q Business 的要求 2
- 确定要添加到连接器的数据源 2
 - 数据源的最大数量 3
 - 数据源的位置 3
 - 支持的协议 3
 - 支持的数据源文件格式 3
- 部署GenAI基础架构 4
 - 基础架构详细信息 4
 - 部署GenAI基础架构 5

开始使用

GenAI连接器快速入门

开始使用您组织在 Amazon FSx for NetApp ONTAP 文件系统上的数据创建 NetApp Connector for Amazon Q Business。创建连接器后、最终用户可以访问亚马逊Q业务助理、获得以组织为中心的问题解答。

1

登录工作负载工厂

您需要 ["在 Workload Factory 上创建一个帐户"](#)并使用以下任一方式登录 ["控制台体验"](#)。

2

设置您的环境以满足GenAI要求

您需要AWS凭据来部署AWS基础架构、部署和发现FSx for ONTAP文件系统、要在连接器中集成的数据源列表、对Amazon Q Business应用程序的访问等。

["详细了解GenAI要求"](#)(英文)

3

确定包含数据源的FSx for ONTAP文件系统

您要集成到连接器中的数据源可以位于一个FSx for ONTAP文件系统上、也可以位于多个FSx for ONTAP文件系统上。如果这些系统位于不同的VPC中、则必须可在同一网络中访问这些系统、或者必须为这些VPC建立对等关系、并使用与AI引擎相同的区域和AWS帐户。

["了解如何识别数据源"](#)(英文)

4

部署GenAI基础架构

启动基础架构部署向导、在AWS环境中部署GenAI基础架构。此过程会为NetApp GenAI引擎部署一个EC2实例、并在FSx for ONTAP文件系统上部署一个卷以包含NetApp AI引擎数据库。卷用于存储有关连接器的信息。

["了解如何部署GenAI基础架构"](#)(英文)

下一步行动

现在、您可以为Amazon Q Business创建一个连接器、以便为最终用户提供以组织为中心的响应。

GenAI连接器要求

在为 Amazon Q Business 创建NetApp连接器之前，请确保 Workload Factory 和 AWS 已正确设置。

基本GenAI要求

GenAI具有您的环境在开始使用之前需要满足的一般要求。

工作负载工厂登录和帐户

您需要 ["在 Workload Factory 上创建一个帐户"](#) 并使用以下任一方式登录 ["控制台体验"](#)。

AWS 凭据和权限

您需要将具有读/写权限的 AWS 凭证添加到 Workload Factory，这意味着您将以读/写模式使用 Workload Factory 进行 GenAI。

目前不支持_基本_模式和_只读_模式权限。

设置凭据时、选择如下所示的权限后、您将拥有管理FSx for ONTAP文件系统以及部署和管理GenAI EC2实例以及知识库和聊天机器人所需的其他AWS资源的完全访问权限。

["了解如何将 AWS 凭证添加到 Workload Factory"](#)

NetApp Connector for Amazon Q Business 的要求

确保您的环境满足Amazon Q Business连接器的以下特定要求。

Amazon Q Business应用程序

您需要创建Amazon Q Business应用程序或使用现有应用程序。

- 确保应用程序位于您的一个AWS区域中。
- 确保应用程序已安装 ["已创建索引"](#)。
- 确保应用程序未处于故障状态。

FSx for ONTAP文件系统

对于ONTAP文件系统、至少需要一个FSx：

- NetApp GenAI引擎将使用一个文件系统(如果不存在、则会创建一个文件系统)来存储有关连接器的信息。

此FSx for ONTAP文件系统必须使用FlexVol卷。不支持 FlexGroup 卷。

- 一个或多个文件系统将包含要添加到连接器的数据源。

一个FSx for ONTAP文件系统可同时用于这两种用途、也可以使用多个FSx for ONTAP文件系统。

- 您需要了解AWS FSx for ONTAP文件系统所在的AWS区域、VPC和子网。
- 您需要考虑要应用于此部署中的AWS资源的标记键/值对(可选)。
- 您需要知道密钥对信息、以便安全地连接到NetApp AI引擎实例。

["了解如何部署和管理FSx for ONTAP文件系统"](#)

确定要添加到连接器的数据源

确定或创建将集成到连接器中的FSx for ONTAP文件系统上的文档(数据源)。通过这些数据源、Amazon Q Business可以根据与您的组织相关的数据为用户查询提供准确且个性化的答案。

数据源的最大数量

支持的最大数据源数为10。

数据源的位置

数据源可以存储在Amazon FSx for NetApp ONTAP文件系统上的SMB共享或NFS导出上的单个卷或卷内的文件夹中。数据源还可以存储在NetApp SnapMirror数据保护关系中适用于NetApp ONTAP的Amazon FSx卷上。

您不能选择卷或文件夹中的单个文档、因此、应确保包含数据源的每个卷或文件夹不包含不应与知识库集成的无关文档。

您可以将多个数据源添加到每个连接器中、但这些数据源都需要驻留在可从AWS帐户访问的FSx for ONTAP文件系统上。

每个数据源的最大文件大小为50 MB。

支持的协议

连接器支持使用NFS或SMB/CCIFS协议的卷中的数据。选择使用SMB协议存储的文件时、您需要输入Active Directory信息、以便连接器可以访问这些卷上的文件。这包括Active Directory域、IP地址、用户名和密码。

将数据源存储在通过SMB访问的共享(文件或目录)上时、只有有权访问该共享的聊天机器人用户或组才能访问该数据。启用此"权限感知功能"后、AI系统会将auth0中的用户电子邮件与允许查看或使用SMB共享上文件的用户进行比较。聊天机器人将根据用户对嵌入文件的权限提供答案。

例如、如果您已将10个文件(数据源)集成到连接器中、并且其中2个文件是包含受限信息的人力资源文件、则只有经过身份验证可以访问这2个文件的聊天机器人用户才会收到来自包含这些文件中数据的聊天机器人的响应。



将数据源添加到Amazon Q Business Connector时、只有用户权限才会应用于数据源文件。未应用组权限。



如果数据源中的某个文件缺少文本(例如无文本图像)、Amazon Q Business不会对其编制索引、而是在Amazon CloudWatch日志中记录一个条目、指出缺少文本。

支持的数据源文件格式

NetApp Connector for Amazon Q Business 目前支持以下数据源文件格式。

文件格式	扩展
逗号分隔值文件	.csv
JSON和JSONP	.json
降价	.MD
Microsoft Word	.docx
纯文本	.txt
可移植文档格式	.pdf

文件格式	扩展
Microsoft PowerPoint	.PPT或.pptx
超文本标记语言	.html
可扩展标记语言	.xml
。	..x等等
Microsoft Excel	.xls
富文本格式	.rtf

部署GenAI基础架构

您需要先在环境中部署适用于RAG框架的GenAI基础架构、然后才能为组织构建FSx for ONTAP知识库、连接器和应用程序。主要基础架构组件包括Amazon Brock服务、NetApp GenAI引擎的虚拟机实例以及适用于ONTAP文件系统的FSx。

部署的基础架构可以支持多个知识库、聊天机器人和连接器、因此您通常只需要执行一次此任务。

基础架构详细信息

您的GenAI部署必须位于启用了Amazon Brock的AWS地区。 ["查看受支持区域的列表"](#)

基础架构由以下组件组成。

亚马逊基岩服务

Amazon Brock是一项完全托管的服务、支持您通过单个API使用领先AI公司的基础模型(FMS)。它还提供构建安全生成型AI应用程序所需的功能。

["了解有关Amazon Brock的更多信息"](#)

Amazon Q Business

Amazon Q基于Amazon BROBON构建、可提供一个完全托管的生成型AI助手、您可以使用它来回答问题并根据数据源中的信息生成内容。

["详细了解Amazon Q Business"](#)

NetApp GenAI引擎的虚拟机

在此过程中会部署NetApp GenAI引擎。它提供了从数据源中加载数据、然后将该数据写入向量数据库的处理能力。

FSx for ONTAP文件系统

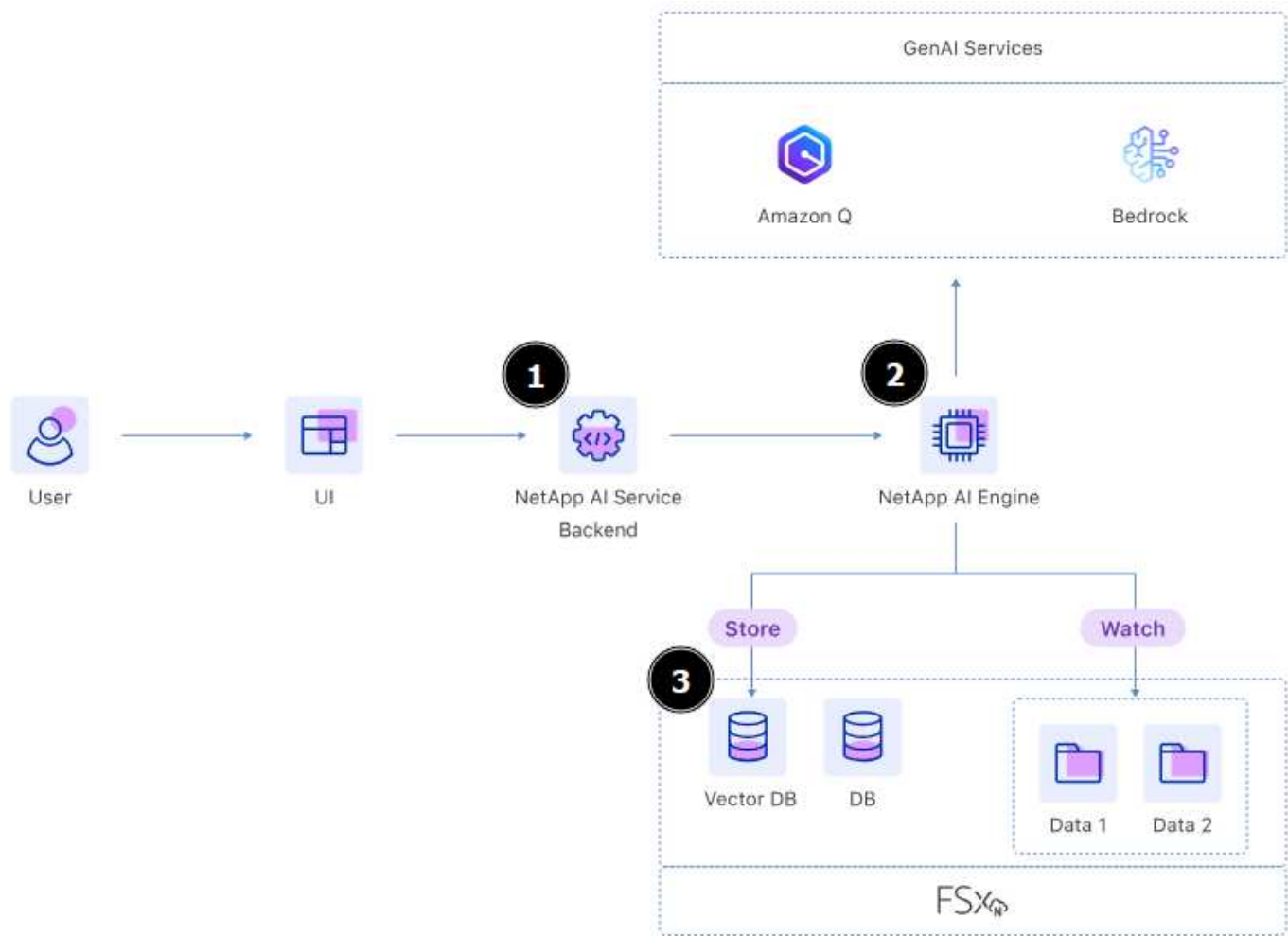
FSx for ONTAP文件系统为您的GenAI系统提供存储。

此时将部署一个卷、其中包含用于存储基础模型根据数据源生成的数据的向量数据库。

您要集成到知识库中的数据源可以位于同一个FSx for ONTAP文件系统上、也可以位于不同的系统上。

NetApp GenAI引擎可监控这两个卷并与之交互。

下图显示了GenAI基础架构。编号为1、2和3的部件将在此过程中展开。在开始部署之前、其他要素必须已准备就绪。



部署GenAI基础架构

您需要输入AWS凭据并选择FSx for ONTAP文件系统、以部署检索增强型生成(RAG)基础架构。

开始之前

在开始此过程之前、根据您选择的内容、确保您的环境满足知识库或连接器的要求。

- "知识库要求"
- "连接器要求"

步骤

1. 使用以下方式之一登录 Workload Factory"控制台体验"。
2. 在AI Workloads (AI工作负载)磁贴中、选择*部署和管理*。
3. 查看基础架构图并选择"*下一步"。
4. 完成*AWS设置*部分中的项目：
 - a. **AWS凭据**：选择或添加可提供部署AWS资源权限的AWS凭据。

- b. 位置：选择AWS区域、VPC和子网。

GenAI部署必须位于启用了Amazon Brock的AWS区域中。 ["查看受支持区域的列表"](#)

5. 完成*Infrastructure settings (基础架构设置)*部分中的项目：

- a. 标签：输入您想要应用于此部署的所有 AWS 资源的任何标签键/值对。这些标签在 AWS 管理控制台和 Workload Factory 内的基础设施信息区域中可见，可以帮助您跟踪 Workload Factory 资源。

6. 完成*连接*部分：

- a. 密钥对：选择一个可安全连接到NetApp GenAI引擎实例的密钥对。

7. 完成*AI引擎*部分：

- a. 实例名称：可选地，选择*定义实例名称*并输入 AI 引擎实例的自定义名称。实例名称出现在 AWS 管理控制台和 Workload Factory 内的基础设施信息区域，可以帮助您跟踪 Workload Factory 资源。

8. 选择*Deploy开始部署。



如果部署失败并显示凭据错误、您可以通过选择错误消息中的超链接来获取更多信息。您可以查看缺少或阻止的权限列表、以及GenAI工作负载部署GenAI基础架构所需的权限列表。

结果

Workload Factory 开始部署聊天机器人基础设施。此过程最多可能需要 10 分钟。

在部署过程中、将设置以下各项：

- 网络与专用端点一起设置。
- 此时将创建IAM角色、实例配置文件和安全组。
- 此时将部署GenAI引擎的虚拟机实例。
- Amazon Brock配置为使用带有前缀的日志组将日志发送到Amazon CloudWatch日志 `/aws/bedrock/`。
- GenAI 引擎配置为使用名为 `/netapp/wlmai/<tenancyAccountId>/randomId`，在哪里 `<tenancyAccountId>` 是 ["NetApp控制台帐户 ID"](#)对于当前用户。

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。