



使用**GenAI**为**Amazon Brock**构建知识库 GenAI

NetApp
October 06, 2025

目录

- 使用GenAI为Amazon Brock构建知识库 1
 - 开始使用 1
 - GenAI知识库快速入门 1
 - GenAI知识库要求 1
 - 确定要添加到知识库或连接器的数据源 3
 - 部署GenAI基础架构 4
 - 创建GenAI知识库 7
 - 创建并配置知识库 8
 - 向知识库添加数据源 10
 - 测试GenAI知识库 14
 - 为GenAI知识库激活外部身份验证 15
 - 发布GenAI知识库并查看唯一端点 16
 - 使用GenAI外部示例聊天机器人应用程序 17
 - 了解更多信息。 17
 - 创建基于RAG的GenAI应用程序 17
 - 您接下来可以使用GenAI做什么 18

使用GenAI为Amazon Brock构建知识库

开始使用

GenAI知识库快速入门

开始使用Amazon FSx for NetApp ONTAP文件系统上的组织数据创建知识库或Amazon Q Business Connector。聊天机器人等应用程序将访问此知识库或连接器、以便为最终用户提供以组织为中心的响应。

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、要集成到知识库或连接器中的数据源列表、访问Amazon Brock AI服务或Amazon Q Business应用程序等。

["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

您将集成到知识库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。该卷用于存储知识库使用的矢量数据库。

["了解如何部署知识库基础架构"](#)(英文)

下一步行动

现在、您可以构建一个知识库、以便为最终用户提供以组织为中心的响应。

GenAI知识库要求

在构建知识库之前，请确保 Workload Factory 和 AWS 已正确设置。这包括拥有您的 AWS 登录凭证、已部署的 FSx for ONTAP文件系统（其中包含您想要集成到知识库中的数据源）、访问 Amazon Bedrock AI 服务等等。

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 `_Basic_` 模式和 `_read-only_` 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"](#)

GenAI知识库要求

如果您计划使用知识库、请确保您的环境满足以下要求。

亚马逊基岩

Amazon Brock使您能够使用基础模型、并提供构建生成性AI应用程序的功能。

在开始使用NetApp Workload Factory for GenAI 之前，您必须设置 Amazon Bedrock。您的 GenAI 部署必须位于启用了 Amazon Bedrock 的 AWS 区域。

- ["AWS文档：设置Amazon Brock"](#)
- ["AWS文档：Amazon Brock知识库支持的区域和型号"](#)

默认情况下、GenAI会重新排列搜索结果、以提高结果相关性。为了获得最佳结果、请确保您的Amazon基岩基础模型配置包含对重新排序模型的访问权限、例如Cohere Rerank或Amazon Rerank (如果您所在地区有)。

嵌入模型

在创建知识库之前、您必须启用计划使用的嵌入模型。支持以下嵌入模型：

- T人 嵌入G1 -文本
- T人 嵌入文本v2
- 第1代多模态嵌入
- 嵌入英语
- 嵌入多语言

["详细了解Amazon大力孙"](#)

聊天模式

在创建知识库之前、您必须启用计划使用的基本聊天模式。由于型号支持因AWS地区而异、请参见 ["AWS文](#)

档" 以验证您可以在计划部署知识库的地区使用哪些型号。

GenAI支持各种型号、包括Anthropic、Amazon、Mistral AI、Meta、JAMBA和cohere。

详细了解如何在Amazon Brock中使用这些模型：

- ["anthpic的克洛德在亚马逊基岩"](#)
- ["在Amazon B基石 控制台中开始使用Amazon Nova"](#)
- ["Mistral AI型号"](#)
- ["Amazon T人 文本模型"](#)
- ["Meta Llama模型"](#)
- ["JAMBA型号"](#)
- ["Cohere命令模型"](#)

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、则创建一个文件系统)来存储知识库使用的矢量数据库。

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集成到知识库中的数据源。

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。文件系统必须位于已启用Amazon Brock的AWS区域中。
- 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"](#)

确定要添加到知识库或连接器的数据源

确定或创建FSx for ONTAP文件系统上要集成到知识库中的文档(数据源)。通过这些数据源、知识库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知识库中、但这些数据源都需要驻留在可从AWS帐户访问的适用于ONTAP文件系统的FSx上。

每个数据源的最大文件大小为50 MB。

支持的协议

知识库支持使用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、以便知识库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)集成到知识库中、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。

支持的数据源文件格式

Workload Factory GenAI 知识库目前支持以下数据源文件格式。

文件格式	扩展
Apache Parquet脚 注：免责声明[将结构化数据文件纳入知识库时不支持数据防护功能。]	镶木地板
逗号分隔值文件注释：免责声明[]	.csv
图形交换格式	gif
JPEG	.jpg or.jpeg
JSON和JSON脚注：免责声明[]	.json
降价	.MD
Microsoft Word	doc或.docx
纯文本	.txt
可移植文档格式	.pdf
便携式网络图形	png
WebP映像	.webp

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

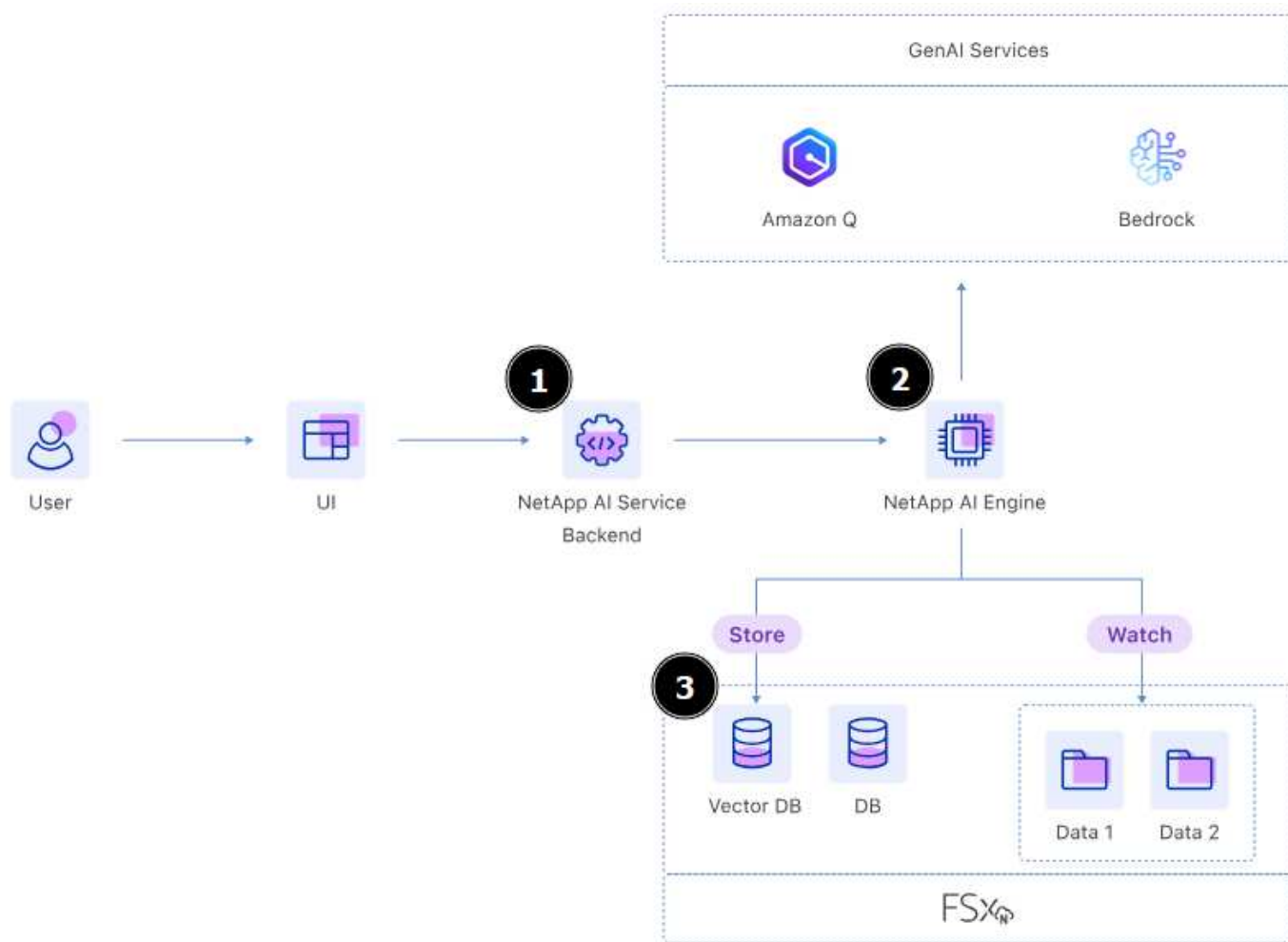
FSx for ONTAP文件系统为您的GenAI系统提供存储。

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、根据您的选择的内容、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错误详细信息。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。

创建GenAI知识库

部署 AI 基础架构并确定将从 FSx for ONTAP数据存储集成到知识库中的数据源后，您就可以使用 Workload Factory 构建知识库了。作为此步骤的一部分，您还将定义 AI 特性并创建对话开场白。

在继续操作之前、请确保您的环境符合[要求](#)知识库的要求。

关于此任务

知识库有两种数据集成方式-`_public mode/`和`_Enterprise mode/`。

公共模式

无需集成企业的数据源、即可使用知识库。在这种情况下、与知识库集成的应用程序将仅提供互联网上公开提供的信息的结果。这称为`_public mode_integration`。

企业模式

在大多数情况下、您需要将组织中的数据源集成到知识库中。这种方式称为`_Enterprise mode_integration`、因为它可以提供企业的知识。

您组织的数据源可能包含个人信息 (PII)。为了保护这些敏感信息，您可以在创建和配置知识库时启用_数据护栏_。由NetApp数据分类提供支持的数据护栏可识别和屏蔽 PII，使其无法访问和恢复。

["了解NetApp数据分类"](#)。



NetApp Workload Factory for GenAI 不会屏蔽敏感的个人信息 (SPii)。参考["敏感个人数据的类型"](#)有关此类数据的更多信息。



数据护栏可以随时启用或禁用。如果您切换数据护栏启用，Workload Factory 将从头开始扫描整个知识库，这会产生成本。

创建并配置知识库

知识库定义了一些特征、例如、要用来创建知识库的基础AI模型和嵌入格式。

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. 从知识库和连接器菜单中，选择*新建*下拉菜单并选择* NetApp GenAI 知识库 for Bedrock*。
4. 在创建NetApp GenAI 知识库页面上，配置知识库设置：

知识库详细信息

1. **Name**：输入要用于知识库的名称。
2. **Description**：输入知识库的详细说明。
3. **Bedrock**：选择您的 AWS 账户可使用 Amazon Bedrock 的区域。

摄入

1. 嵌入模型：
 - 选择一个嵌入模型用于知识库。嵌入模型定义如何将数据转换为知识库的向量嵌入。 Workload Factory 支持以下模型：
 - T人 嵌入G1 -文本
 - T人 嵌入文本v2
 - 第1代多模态嵌入
 - 嵌入英语
 - 嵌入多语言

请注意、您必须已启用Amazon Brock中的嵌入模型。

["详细了解Amazon大力孙"](#)

- 如果适用，请选择与所选嵌入模型的配置相匹配的推理类型。
2. 数据护栏：选择是否要启用或禁用数据护栏。["了解由NetApp数据分类提供支持的数据护栏"](#)。

要启用数据防护、必须满足以下前提条件。

- 需要服务帐户才能与NetApp数据分类进行通信。您必须在NetApp控制台租赁帐户上拥有“组织管理员”角色才能创建服务帐户。具有组织管理员角色的成员可以完成组织中的所有操作。["了解如何在NetApp控制台中为成员添加角色"](#)
- AI 引擎必须能够访问["NetApp控制台 API 端点"](#)。
- 您需要按照["NetApp数据分类文档"](#):
 - i. 创建控制台代理
 - ii. 确保您的环境满足前提条件
 - iii. 部署NetApp数据分类



在导出CSV、JSON、JSONP或Parquet等结构化数据文件时、不支持数据防护功能。

聊天和检索设置

1. 聊天模型：

- 从 Amazon Bedrock 中集成的各种聊天模型中进行选择。请注意，您必须已经启用来自 Amazon Bedrock 的聊天模型。
- 如果适用，请选择与所选模型的配置相匹配的推理类型。

2. 聊天设置：

- 为聊天机器人选择一个温度来配置响应的随机性和创造性。较低的温度会导致更可预测的响应，而较高的温度会导致更加多样化的响应。
- 选择最大响应长度来配置响应的详细程度。响应长度越长，使用的响应令牌就越多，并且会产生更高的成本。

3. 思考模式：启用思考模式后，聊天机器人将花费更多时间来处理查询，并且结果通常会更准确。当您启用思考模式时，您可以控制在生成结果时使用多少个推理标记。使用更多的推理标记可以得到更准确的响应，但可能会产生更高的成本。

4. 重新排名：启用或禁用重新排名，这可以提高查询结果的相关性和质量。选择标准聊天模型或专门的重新排名模型用于重新排名。仅当您所在地区可用时才会显示 Reranker 模型选项。选择与所选模型的配置相匹配的推理类型。

5. 对话开场白：选择是否要提供最多四个对话开场提示、这些提示将显示给与使用此知识库的聊天机器人进行交互的用户。建议您启用此设置。

如果您激活了启动对话器、则默认情况下会选中“自动模式”。只有在将数据源添加到知识库后、才能启用“手动模式”。["了解如何修改知识库设置"](#)(英文)

存储定义

1. **FSx for ONTAP**文件系统：当您定义新的知识库时，Workload Factory 会创建一个新的Amazon FSx for NetApp ONTAP卷来存储它。选择将在其中创建新卷的现有文件系统名称和 SVM（也称为存储 VM）。
2. 快照策略：从 Workload Factory 存储清单中定义的现有策略列表选择一个快照策略。知识库的定期快照将根据您选择的快照策略以一定频率自动创建。
3. **S3 存储桶**：如果聊天机器人查询结果包含结构化数据，GenAI 可以将结果存储在 S3 存储桶中。要使用此功能，请启用*激活 S3 存储桶*设置并从列表中选择与您的帐户关联的 S3 存储桶。当这些结果存储在 S3 存

储桶中时，您可以使用聊天会话中的下载链接下载它们。

如果不存在所需的快照策略、则可以 ["创建快照策略"](#)在包含此卷的Storage VM上执行此操作。

4. 选择*创建知识库*将知识库添加到GenAI中。

创建知识库时、将显示进度指示器。

创建知识库后、您可以选择向新知识库添加数据源、也可以选择结束此过程而不添加数据源。建议您选择*添加数据源*，然后立即添加一个或多个数据源。

向知识库添加数据源

您可以添加一个或多个数据源、以便使用组织的数据填充知识库。

关于此任务

支持的最大数据源数为10。

步骤

1. 选择*添加数据源*后，选择要添加的数据源类型：
 - 添加 FSx for ONTAP 文件系统（使用现有 FSx for ONTAP 卷中的文件）
 - 添加文件系统（使用来自通用 SMB 或 NFS 共享的文件）

添加 FSx for ONTAP 文件系统

1. 选择文件系统：选择数据源文件所在的FSx for ONTAP文件系统，然后选择*下一步*。
2. 选择卷：选择数据源文件所在的卷，然后选择*下一步*。

选择使用SMB协议存储的文件时、您需要输入Active Directory信息、其中包括域、IP地址、用户名和密码。

3. 选择数据源：根据文件的保存位置选择数据源位置。该卷可以是整个卷，也可以是卷中的特定文件夹或子文件夹，然后选择*Next*。
4. 配置：配置数据源如何从文件中获取信息，以及它包含在扫描中的文件：

- 定义数据源：在*Chunking string*部分中，定义在数据源与知识库集成时，GenAI引擎如何将数据源内容拆分成块。您可以选择以下策略之一：
 - 多句拆分：将数据源中的信息组织为句子定义的区块。您可以选择每个区块包含多少个句子(最多100个)。
 - 基于重叠的区块：将数据源中的信息组织为字符定义的区块，这些区块可以重叠相邻区块。您可以选择每个区块的字符大小、以及每个区块与相邻区块重叠的程度。您可以将区块大小配置为50到3000个字符、并将重叠百分比配置为1到99%。



选择较高的重叠百分比会显著增加存储需求、但检索准确性只会稍有提高。

- 文件过滤：配置扫描中包含的文件：

- 在*文件类型支持*部分，选择包括所有类型的文件，或选择要包含在数据源扫描中的单个文件类型。

如果您包含图像或 PDF 文件，NetApp Workload Factory for GenAI 会解析图像中的文本（包括 PDF 文档中的图像），这会产生更高的成本。

如果包含图像中的文本数据、则在将扫描的文本数据从您的环境发送到AWS时、GenAI无法屏蔽图像中的个人身份信息(个人身份信息、(PII))。但是、一旦存储数据、所有的数据都会在GenAI数据库中屏蔽。



您选择将图像文件包括在扫描中与知识库聊天模式相关。如果扫描中包含图像文件、则聊天模式必须支持图像。如果在此处选择了图像文件类型、则无法将知识库切换到不支持图像文件的聊天模式。

- 在*文件修改时间过滤器*部分，选择根据修改时间启用或禁用包含文件。如果启用了修改时间筛选、请从列表选择一个日期范围。



如果您包含基于修改日期范围的文件、则只要不满足日期范围(这些文件在您指定的日期范围内未被修改)、这些文件就会从定期扫描中排除、并且数据源不会包含这些文件。

5. 在*权限感知*部分(仅当您选择的数据源位于使用SMB协议的卷上时才可用)中、您可以启用或禁用权限感知响应：
 - 已启用：访问此知识库的聊天机器人用户只会从他们有权访问的数据源获得查询响应。

- 已禁用：聊天机器人用户将使用所有集成数据源的内容接收响应。

6. 选择*Add*将此数据源添加到您的知识库中。

添加通用 **NFS** 文件系统

1. 选择文件系统：输入数据源文件所在的文件系统主机的 IP 地址或 FQDN，选择网络共享的 NFS 协议，然后选择*下一步*。
2. 选择数据源：根据文件的保存位置选择数据源位置。该卷可以是整个卷，也可以是卷中的特定文件夹或子文件夹，然后选择*Next*。



在某些情况下，您可能需要手动输入 NFS 导出名称，然后选择“检索目录”以显示可用目录。您可以选择整个导出，或仅选择导出中的特定文件夹。

3. 配置：配置数据源如何从文件中获取信息，以及它包含在扫描中的文件：

- 定义数据源：在*Chunking string*部分中，定义在数据源与知识库集成时，GenAI引擎如何将数据源内容拆分成块。您可以选择以下策略之一：
 - 多句拆分：将数据源中的信息组织为句子定义的区块。您可以选择每个区块包含多少个句子(最多100个)。
 - 基于重叠的区块：将数据源中的信息组织为字符定义的区块，这些区块可以重叠相邻区块。您可以选择每个区块的字符大小、以及每个区块与相邻区块重叠的程度。您可以将区块大小配置为50到3000个字符、并将重叠百分比配置为1到99%。



选择较高的重叠百分比会显著增加存储需求、但检索准确性只会稍有提高。

- 文件过滤：配置扫描中包含的文件：

- 在*文件类型支持*部分，选择包括所有类型的文件，或选择要包含在数据源扫描中的单个文件类型。

如果您包含图像或 PDF 文件，NetApp Workload Factory for GenAI 会解析图像中的文本（包括 PDF 文档中的图像），这会产生更高的成本。

如果包含图像中的文本数据、则在将扫描的文本数据从您的环境发送到AWS时、GenAI无法屏蔽图像中的个人身份信息(个人身份信息、(PI))。但是、一旦存储数据、所有的数据都会在GenAI数据库中屏蔽。



您选择将图像文件包括在扫描中与知识库聊天模式相关。如果扫描中包含图像文件、则聊天模式必须支持图像。如果在此处选择了图像文件类型、则无法将知识库切换到不支持图像文件的聊天模式。

- 在*文件修改时间过滤器*部分，选择根据修改时间启用或禁用包含文件。如果启用了修改时间筛选、请从列表中选择一个日期范围。



如果您包含基于修改日期范围的文件、则只要不满足日期范围(这些文件在您指定的日期范围内未被修改)、这些文件就会从定期扫描中排除、并且数据源不会包含这些文件。

4. 选择*添加数据源*将此数据源添加到您的知识库。

添加通用 SMB 文件系统

1. 选择文件系统：

- 输入数据源文件所在的文件系统主机的 IP 地址或 FQDN。
- 为网络共享选择 SMB 协议。
- 输入 Active Directory 信息，包括域、IP 地址、用户名和密码。
- 选择 * 下一步 *。

2. 选择数据源：根据文件的保存位置选择数据源位置。该卷可以是整个卷，也可以是卷中的特定文件夹或子文件夹，然后选择*Next*。



在某些情况下，您可能需要手动输入 SMB 共享名称，然后选择“检索目录”以显示可用目录。您可以选择整个共享，或仅选择共享中的特定文件夹。

3. 配置：配置数据源如何从文件中获取信息，以及它包含在扫描中的文件：

- 定义数据源：在*Chunking string*部分中，定义在数据源与知识库集成时，GenAI引擎如何将数据源内容拆分成块。您可以选择以下策略之一：
 - 多句拆分：将数据源中的信息组织为句子定义的区块。您可以选择每个区块包含多少个句子(最多100个)。
 - 基于重叠的区块：将数据源中的信息组织为字符定义的区块，这些区块可以重叠相邻区块。您可以选择每个区块的字符大小、以及每个区块与相邻区块重叠的程度。您可以将区块大小配置为50到3000个字符、并将重叠百分比配置为1到99%。



选择较高的重叠百分比会显著增加存储需求、但检索准确性只会稍有提高。

◦ 权限感知：启用或禁用权限感知响应：

- 已启用：访问此知识库的聊天机器人用户只会从他们有权访问的数据源获得查询响应。
- 已禁用：聊天机器人用户将使用所有集成数据源的内容接收响应。

◦ 文件过滤：配置扫描中包含的文件：

- 在*文件类型支持*部分，选择包括所有类型的文件，或选择要包含在数据源扫描中的单个文件类型。

如果您包含图像或 PDF 文件，NetApp Workload Factory for GenAI 会解析图像中的文本（包括 PDF 文档中的图像），这会产生更高的成本。

如果包含图像中的文本数据、则在将扫描的文本数据从您的环境发送到AWS时、GenAI无法屏蔽图像中的个人身份信息(个人身份信息、(PI))。但是、一旦存储数据、所有的数据都会在GenAI数据库中屏蔽。



您选择将图像文件包括在扫描中与知识库聊天模式相关。如果扫描中包含图像文件、则聊天模式必须支持图像。如果在此处选择了图像文件类型、则无法将知识库切换到不支持图像文件的聊天模式。

- 在*文件修改时间过滤器*部分，选择根据修改时间启用或禁用包含文件。如果启用了修改时间筛选、请从列表中选择一个日期范围。



如果您包含基于修改日期范围的文件、则只要不满足日期范围(这些文件在您指定的日期范围内未被修改)、这些文件就会从定期扫描中排除、并且数据源不会包含这些文件。

4. 选择*添加数据源*将此数据源添加到您的知识库。

结果

数据源开始嵌入到您的知识库中。当数据源完全嵌入时、状态将从"嵌入"更改为"嵌入"。

向知识库添加单个数据源后、您可以在本地的聊天机器人模拟器窗口中对其进行测试、并进行任何必要的更改、然后再将此聊天机器人提供给用户。您也可以按照相同的步骤向知识库添加其他数据源。

测试GenAI知识库

创建知识库后、您将能够使用聊天机器人模拟器在本地对其进行测试、并进行任何必要的更改、然后再通过聊天机器人应用程序向用户提供知识库。

关于此任务

您可以对知识库进行测试、以确保其按预期运行、并且可以自定义默认情况下希望此知识库的聊天机器人用户可以使用的对话开场白。聊天机器人模拟器对知识库中嵌入的所有数据源运行。

您可以通过在聊天机器人模拟器中与嵌入的数据源聊天来测试知识库。请注意、在本地测试知识库时、没有在GenAI矢量数据库中捕获任何交互或见解。

在为用户在应用程序中部署知识库之前、您将在 Workload Factory 中执行大部分测试。如果您需要更改数据源或聊天机器人操作，则需要在发布知识库之前进行更改。



您可以调整聊天机器人模拟器窗口的大小并重新命名、还可以将问题和回答复制到剪贴板。

测试聊天机器人时、您需要执行的一些任务包括：

- 输入大量与您的组织相关的问题、以确保答案符合预期。
- 在聊天机器人应用程序中自定义您希望默认为用户提供的对话开场白。
- 确保聊天机器人答案底部提供的归属内容包含正确的参考。

步骤

1. 从"Knowledge Bases"清单页面中、选择要测试的知识库。

聊天机器人模拟器显示在右窗格中。如果已定义、则还会显示现有的开场白。

2. 在聊天机器人输入字段中、输入提示或问题、然后选择 ► 以查看您的聊天机器人如何利用您的组织知识做出响应。



- 您可以通过展开响应下的 ***sources *** 列表来查看用于生成答案的源。此选项将提供用于生成答案的文件列表。通过将鼠标悬停在文件名上、您可以查看每个文件和卷路径中使用的数据块并将其复制到每个文件。
- 如果答案中包含表格，您可以对每列中的数据进行排序，然后将每个表格复制到剪贴板。
- 如果答案结果包含结构化数据，并且知识库启用了 **S3 Bucket** 功能，GenAI 会将结果存储在 S3 Bucket 中。您可以使用聊天会话中的 ***下载结果*** 链接从存储桶中下载结果。

3. 如果您需要更新任何数据源、以便知识库提供更有针对性的答案、请立即进行这些更改、然后重新测试知识库。

为GenAI知识库激活外部身份验证

激活知识库身份验证、以便在使用API端点将知识库与聊天机器人应用程序集成时需要进行令牌验证和ACL。激活身份验证时、您需要为JSON Web令牌配置设置、此令牌将用于向聊天机器人客户端发出的知识库API请求。

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. 从"Knowledge Bases"清单页面中、选择要激活身份验证的知识库。
4. 选择 **...** 并选择***管理知识库***。
5. 选择***操作***菜单，然后选择***管理身份验证设置***。
6. 设置身份验证：
 - a. 选择***激活身份验证设置***。
 - b. 提供所需信息。我们提供了一些示例、但您应从身份验证提供程序获取这些字段的值：
 - 算法：身份验证提供程序使用的签名算法。
 - **Audient**(可选)：包含令牌预期收件人的字符串(有时为URL)。
 - 颁发者：用于标识发出令牌的提供程序的字符串。

例如、Amazon Cognito使用以下格式的颁发者字符串：

```
https://cognito-idp-<region>.amazonaws.com/<UserPoolID>
```

其中 **<region>** 是包含用户池的AWS区域、**<UserPoolID>** 是您的用户池ID。您可以使用以下命令检索用户池ID：

```
aws cognito-idp list-user-pools --max-results=60 --output=table
```

- **JWKS URI**：提供验证此令牌签名所需的公共密钥的URI字符串。

例如、Amazon Cognito使用以下格式的JWKS URI字符串：

```
https://cognito-idp.<region>.amazonaws.com/<userPoolId>/.well-known/jwks.json
```

+ 其中 <region> 是包含用户池的AWS区域、 <UserPoolID> 是您的用户池ID。您可以使用以下命令检索用户池ID：

```
aws cognito-idp list-user-pools --max-results=60 --output=table
```

7. 选择 * 保存 *。

结果

知识库身份验证现在处于活动状态、您可以使用API端点与知识库进行交互、并将知识库与聊天机器人应用程序集成。

发布GenAI知识库并查看唯一端点

在本地构建并测试知识库后、您可以发布知识库、以便与聊天机器人应用程序集成、使用户能够查询知识库。

关于此任务

发布知识库使您能够在聊天应用程序中使用它。发布操作触发 Workload Factory API 生成并发布唯一端点。发布后，知识库可供聊天应用程序访问，并且 API 端点已准备好进行集成。

您发布的每个知识库都有唯一的端点。

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. 从"Knowledge Bases"清单页面中、选择要发布的知识库。
4. 选择  并选择*管理知识库*。

此页面显示已发布状态、数据源的嵌入状态、嵌入模式以及所有嵌入数据源的列表。

5. 选择*操作*菜单，然后选择*发布*。

Workload Factory 发布知识库。在知识库的详细信息页面上，状态从*未发布*更改为*已发布*。

现在、您可以获取有关知识库唯一端点的详细信息。

6. 在已发布状态旁边，选择*view*。

显示有关如何使用 Workload Factory API 访问知识库的详细信息。

7. 从*查看已发布信息*对话框中，复制可用于将知识库与应用程序集成的API端点。

要了解有关API端点的更多信息、请转到 ["API文档"](#)、然后选择*人工智能>外部*。

在使用这些端点之前、您需要从身份验证提供程序获取用户令牌。

结果

现在、您已发布知识库、并拥有可用于将知识库与聊天机器人应用程序集成的唯一端点。

使用GenAI外部示例聊天机器人应用程序

配置、激活和发布知识库后，外部应用程序开发人员可以配置和运行NetApp提供的开源示例聊天机器人应用程序，以便与您的知识库进行交互，并了解如何使用 Workload Factory API 创建自己的生成式 AI 应用程序。

步骤

1. ["创建知识库"](#)(英文)
2. ["激活身份验证"](#) 您创建的知识库。

这样、知识库便可对API请求进行身份验证、并在使用API端点时进行令牌验证和ACL。



与此知识库集成的外部聊天应用程序将需要使用您在知识库的身份验证设置中配置的相同身份验证提供程序(颁发者)。

3. ["发布知识库"](#) 为外部应用程序启用API访问。

知识库发布后、可以从外部访问API端点、您可以将知识库与外部聊天应用程序(例如、示例聊天机器人应用程序)集成在一起。

4. 从下载示例chatbot应用程序包 ["GitHub"](#)。
5. 按照软件包中README文件中的说明安装并运行chatbot应用程序。
6. 浏览以登录到 ["http://localhost:9091"](#) 应用程序。

此时将显示聊天机器人应用程序示例。

了解更多信息。

["工作负载工厂 API 文档"](#)

创建基于RAG的GenAI应用程序

构建知识库并测试聊天机器人后、您可以设置应用程序、使用户能够查询聊天机器人。

["了解如何在FSx for ONTAP上创建基于RAG的AI应用程序"](#)

您接下来可以使用**GenAI**做什么

现在、您已经使用企业数据创建了知识库并为用户部署了该知识库、您可以管理知识库、数据源和RAG基础架构、包括适用于ONTAP文件系统的FSx。

您可以执行以下任务来管理知识库组件：

- 更新数据源的内容或添加新数据源、并将这些更改与知识库和聊天机器人同步。
- 管理数据源设置、包括分块策略和权限感知(用于SMB文件访问)。
- 管理您的知识库设置、包括聊天模式和对话开场白。
- 取消发布知识库或在进行更改后重新发布知识库。
- 备份和保护FSx for ONTAP文件系统上的重要数据、以确保知识库数据和其他基础架构组件始终可用。

有关管理 FSx for ONTAP文件系统的信息，请访问 ["Amazon FSx for NetApp ONTAP的 Workload Factory 文档"](#) 查看您可以使用的备份和保护功能。

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。