



NetApp Workload Factory for GenAI 文档 GenAI

NetApp
October 06, 2025

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

目录

NetApp Workload Factory for GenAI 文档	1
发行说明	2
NetApp Workload Factory for GenAI 的新功能	2
2025年10月5日	2
2025年8月3日	2
2025年6月29日	3
2025年6月3日	3
2025年5月4日	3
2025年3月2日	4
2025年2月2日	4
2025年1月5日	5
2024年12月1日	5
2024年11月3日	6
2024年9月29日	6
2024年9月1日	7
2024 年 8 月 4 日	7
2024年7月7日	7
了解NetApp Workload Factory for GenAI	8
了解NetApp Workload Factory for GenAI	8
NetApp Workload Factory for GenAI 是什么？	8
使用GenAI创建生成性AI应用程序的优势	8
GenAI的工作原理	9
NetApp Workload Factory for GenAI 如何帮助构建生成式 AI 应用程序	9
使用NetApp Workload Factory 的工具	10
成本	10
许可	10
区域	10
NetApp GenAI引擎的组件	11
使用GenAI为Amazon Brock构建知识库	18
开始使用	18
GenAI知识库快速入门	18
GenAI知识库要求	18
确定要添加到知识库或连接器的数据源	20
部署GenAI基础架构	21
创建GenAI知识库	24
创建并配置知识库	25
向知识库添加数据源	27
测试GenAI知识库	31
为GenAI知识库激活外部身份验证	32

发布GenAI知识库并查看唯一端点	33
使用GenAI外部示例聊天机器人应用程序	34
了解更多信息。	34
创建基于RAG的GenAI应用程序	34
您接下来可以使用GenAI做什么	35
使用GenAI为Amazon Q Business创建连接器	36
开始使用	36
GenAI连接器快速入门	36
GenAI连接器要求	36
确定要添加到连接器的数据源	38
部署GenAI基础架构	39
为 Amazon Q Business 创建 NetApp 连接器	41
定义接头	42
将数据源添加到连接器	43
管理和监控	45
管理GenAI基础架构	45
查看有关基础架构的信息	45
删除基础架构	45
管理GenAI知识库	46
查看有关知识库的信息	46
编辑知识库	46
使用快照保护知识库	47
向知识库添加其他数据源	49
将数据源与知识库同步	53
在创建知识库之前评估聊天模式	54
取消发布知识库	54
删除知识库	54
管理Amazon Q Business连接器	55
查看有关连接器的信息	55
编辑连接器	55
向连接器添加其他数据源	56
将数据源与连接器同步	60
删除连接器	61
管理GenAI数据源	61
查看有关数据源的信息	61
编辑数据源设置	61
更新现有数据源的内容	62
删除数据源	62
使用NetApp Workload Factory 中的 Tracker 监控工作负载操作	62
跟踪和监控操作	63
查看API请求	63

重试失败的操作	63
编辑失败的操作并重试	64
知识和支持	65
注册以获取NetApp Workload Factory for GenAI 的支持	65
支持注册概述	65
注册您的帐户以获得NetApp支持	65
GenAI故障排除	67
常见问题和解决方案	67
获取有关NetApp Workload Factory for GenAI 的帮助	71
获取FSx for ONTAP支持	71
使用自助支持选项	71
向NetApp支持部门创建案例	71
管理支持案例(预览)	74
NetApp Workload Factory for GenAI 法律声明	77
版权	77
商标	77
专利	77
隐私政策	77
开放源代码	77

NetApp Workload Factory for GenAI 文档

发行说明

NetApp Workload Factory for GenAI 的新功能

了解 Workload Factory 的生成式 AI 工作负载功能的新功能。

2025年10月5日

BlueXP workload factory现为**NetApp**工作负载工厂

BlueXP已重命名并重新设计，以更好地反映其在管理数据基础设施中的作用。因此，BlueXP workload factory 已重命名为NetApp工作负载工厂。

支持在适用于 **Amazon Q Business** 的**NetApp**连接器中添加通用 **NFS/SMB** 数据源

使用 Workload Factory API，您现在可以将数据源从通用 NFSv3、NFSv4 或 SMB 共享添加到NetApp Connector for Amazon Q Business。这使您能够包含存储在由Amazon FSx for NetApp ONTAP以外的文件系统托管的卷上的文件。

["为 Amazon Q Business 创建 NetApp 连接器"](#)

["向连接器添加数据源"](#)

知识库的高级聊天配置

您现在可以配置适用于知识库聊天模型的高级聊天设置，例如响应长度、温度、推理设置等。其中一些设置（例如最近时间和修改时间设置、高级检索设置和系统提示）仅可使用 Workload Factory API。

["创建GenAI知识库"](#)

现支持嵌入、聊天和重新排名模型的推理类型选择

如果您选择的嵌入、聊天或重新排名模型具有推理设置，您现在可以选择推理类型。这使您能够更好地根据您的需要调整聊天机器人的性能和资源需求。

["创建GenAI知识库"](#)

2025年8月3日

结构化数据结果的安全存储

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 Amazon S3 存储桶中。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。

["创建GenAI知识库"](#)

MCP 服务器可用性

NetApp现在为 GenAI 提供带有NetApp Workload Factory 的模型上下文协议 (MCP) 服务器。您可以在本地安装

服务器，以允许外部 MCP 客户端从 GenAI 知识库中发现和检索查询结果。

["NetApp Workload Factory GenAI MCP 服务器"](#)

2025年6月29日

支持托管在通用 **NFS/SMB** 文件系统上的数据源

您现在可以从通用 SMB 或 NFS 共享添加数据源。这使您能够包含存储在 Amazon FSx for NetApp ONTAP 以外的文件系统托管的卷上的文件。

["向知识库添加数据源"](#)

["向连接器添加数据源"](#)

2025年6月3日

跟踪器可用于监控和跟踪操作

GenAI 现已提供 Tracker 监控功能。您可以使用 Tracker 监控和跟踪待处理、正在进行和已完成操作的进度和状态，查看操作任务和子任务的详细信息，诊断任何问题或故障，编辑失败操作的参数，以及重试失败的操作。

["使用NetApp Workload Factory 中的 Tracker 监控工作负载操作"](#)

为知识库选择重排序模型

现在，您可以通过选择与知识库配合使用的特定重排序模型来提高重排序查询结果的相关性。GenAI支持 Cohere Rerank 和 Amazon Rerank 模型。

["创建GenAI知识库"](#)

2025年5月4日

支持 **NetApp Connector for Amazon Q Business**

此版本的 GenAI 引入了对 NetApp Connector for Amazon Q Business 的支持，使您能够为 Amazon Q Business 创建连接器。快速轻松地利用Amazon Q Business AI助理、与构建Amazon Brock的GenAI知识库相比、初始配置更少。

["为 Amazon Q Business 创建 NetApp 连接器"](#)

增强了聊天模式支持

现在、GenAI还支持在知识库中使用以下聊天模式：

- ["Mistral AI型号"](#)
- ["Amazon T人 文本模型"](#)
- ["Meta Llama模型"](#)
- ["JAMBA 1.5型号"](#)

- ["Cohere命令模型"](#)
- ["DeepSeek型号"](#)

GenAI支持Amazon Brock支持的每个提供商的型号：["Amazon Brock中支持的基础模型"](#)

["创建GenAI知识库"](#)

更新了权限术语

Workload Factory 用户界面和文档现在使用“只读”来指代读取权限，使用“读/写”来指代自动化权限。

2025年3月2日

嵌入式聊天机器人增强功能

现在、您可以将问题和回答直接复制到剪贴板、调整聊天窗口的大小并更改其标题。此外、聊天响应现在可以包括表、这些表也是可复制的。

["测试GenAI知识库"](#)

聊天响应引用支持

聊天回复现在包含引用、列出用于生成回复的文件和数据块。

["测试GenAI知识库"](#)

增强了文件类型支持

此版本的GenAI提供了增强的文件支持：

- 聊天模式改进了CSV支持。这样、在从CSV文件查询数据时、可以做出更有用的响应。
- 现在、GenAI可以从数据源中安装Apache Parquet文件。
- GenAI现在支持载入包含图像的Microsoft Word DOCX文件。对DOCX文档中嵌入的图像进行扫描、对知识库查询的响应中包含嵌入图像中的文本洞察信息。

["支持的数据源文件格式"](#)

2025年2月2日

支持**Amazon Nova**基础模型

现在、GenAI支持Amazon Nova基础模型。支持Amazon Nova Micro、Amazon Nova Lite和Amazon Nova Pro。

["GenAI要求"](#)

数据源的文件类型筛选

现在、GenAI支持在添加数据源时选择要包括在数据源扫描中的特定文件类型。

"向知识库添加数据源"

数据源的文件修改日期筛选

现在、GenAI支持在添加数据源时按修改日期筛选要包含在数据源扫描中的文件。您可以为包含的文件选择修改日期范围。

"向知识库添加数据源"

支持图像文件和增强的PDF文件支持

现在、GenAI支持通过从图像和图形说明以及文档文本中获得洞察力来增强对知识库查询的响应、从而获得更丰富、质量更高的答案。现在、GenAI可以扫描PDF文件中的图像文件和图像(也称为多模式文件支持)。如果您选择扫描图像或PDF文件、则图像中的文本(包括PDF文档中嵌入的图像)将扫描到数据源中、扫描中的见解将包括在知识库查询的响应中。

"向知识库添加数据源"

混合搜索和重新搜索支持

现在、GenAI可以使用混合搜索并重新排列结果、从而显著提高搜索结果的相关性和准确性。混合搜索将基于关键字的传统搜索的优势与基于密集矢量的高级语法搜索技术相结合。标准关键字搜索结果通过近似匹配和语言细微差别得到增强、从而增强相关性。然后、GenAI使用Cohere Rerank和Amazon Rerank等高级重新排名模型进一步细化这些结果、并返回最相关的结果。此功能可用于新创建的知识库。

"了解NetApp Workload Factory for GenAI"

2025年1月5日

自定义快照名称

现在、您可以为临时快照提供快照名称。

"使用快照保护知识库"

自定义AI引擎实例名称

现在、您可以在部署期间为AI引擎实例提供一个自定义名称。

"部署GenAI基础架构"

重建损坏或缺失的GenAI基础架构

如果您的AI引擎实例损坏或以某种方式被删除,您可以让Workload Factory为您重建它。重建完成后,工作负载工厂会自动将您的知识库重新连接到基础设施,以便随时可以使用。

"故障排除"

2024年12月1日

从快照克隆一个信息库克隆

NetApp Workload Factory for GenAI 现在支持从快照克隆知识库。从而可以快速恢复知识库，并利用现有数据源创建新的知识库，有助于数据恢复和开发。

["克隆知识库"](#)

内部ONTAP集群发现和复制

发现并将本地ONTAP集群数据复制到 FSx for ONTAP文件系统，以便用于丰富 AI 知识库。所有本地发现和复制工作流程均可通过存储清单中新的 **On-Premises ONTAP** 菜单实现。

["发现内部 ONTAP 集群"](#)

2024年11月3日

使用数据防护屏蔽个人身份信息

生成式人工智能工作负载引入了由NetApp控制台分类提供支持的数据护栏功能。数据护栏功能可识别和屏蔽个人身份信息 (PII)，帮助您保持合规性并加强敏感组织数据的安全性。

["创建GenAI知识库"](#)

["了解NetApp控制台分类"](#)

2024年9月29日

为知识库卷提供快照和还原支持

现在、您可以通过创建知识库的时间点副本来保护生成性AI工作负载数据。这样、您就可以保护数据、防止意外丢失或测试对知识库设置所做的更改。您可以随时还原知识库卷的先前版本。

["为知识库卷创建快照"](#)

["还原知识库卷的快照"](#)

暂停计划的扫描

现在、您可以暂停计划的数据源扫描。默认情况下、生成性AI工作负载每天都会扫描每个数据源、以便将新数据导入到每个知识库中。如果您不希望(例如在测试或还原快照期间)导入最新更改、您可以随时暂停计划的扫描并恢复它们。

["管理知识库"](#)

现在、知识库可支持数据保护卷

现在、在选择知识库卷时、您可以选择属于NetApp SnapMirror复制关系的数据保护卷。这样、您就可以将知识库存储在已受SnapMirror复制保护的卷上。

["确定要集成到知识库中的数据源"](#)

2024年9月1日

其他分块策略

生成型AI工作负载现在支持对数据源执行多句分块和基于重叠的分块。

为每个知识库提供专用卷

现在、生成性AI工作负载会为每个新知识库创建一个专用的Amazon FSx for NetApp ONTAP卷、从而为每个知识库启用单独的快照策略、并提高对故障和数据中毒的防护能力。

2024 年 8 月 4 日

Amazon CloudWatch Logs集成

生成型AI工作负载现已与Amazon CloudWatch Logs集成、使您能够监控生成型AI工作负载日志文件。

聊天机器人应用程序示例

NetApp Workload Factory GenAI 示例应用程序使您能够通过基于 Web 的聊天机器人应用程序中直接与已发布的NetApp Workload Factory 知识库进行交互来测试其身份验证和检索。

2024年7月7日

GenAI 工作负载工厂的初始版本

初始版本支持开发一个知识库、该知识库可通过嵌入组织的数据进行自定义。您的用户可以通过聊天机器人应用程序访问知识库。此功能可确保准确、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。

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、从而提高使用生成式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创建生成式AI应用程序的优势

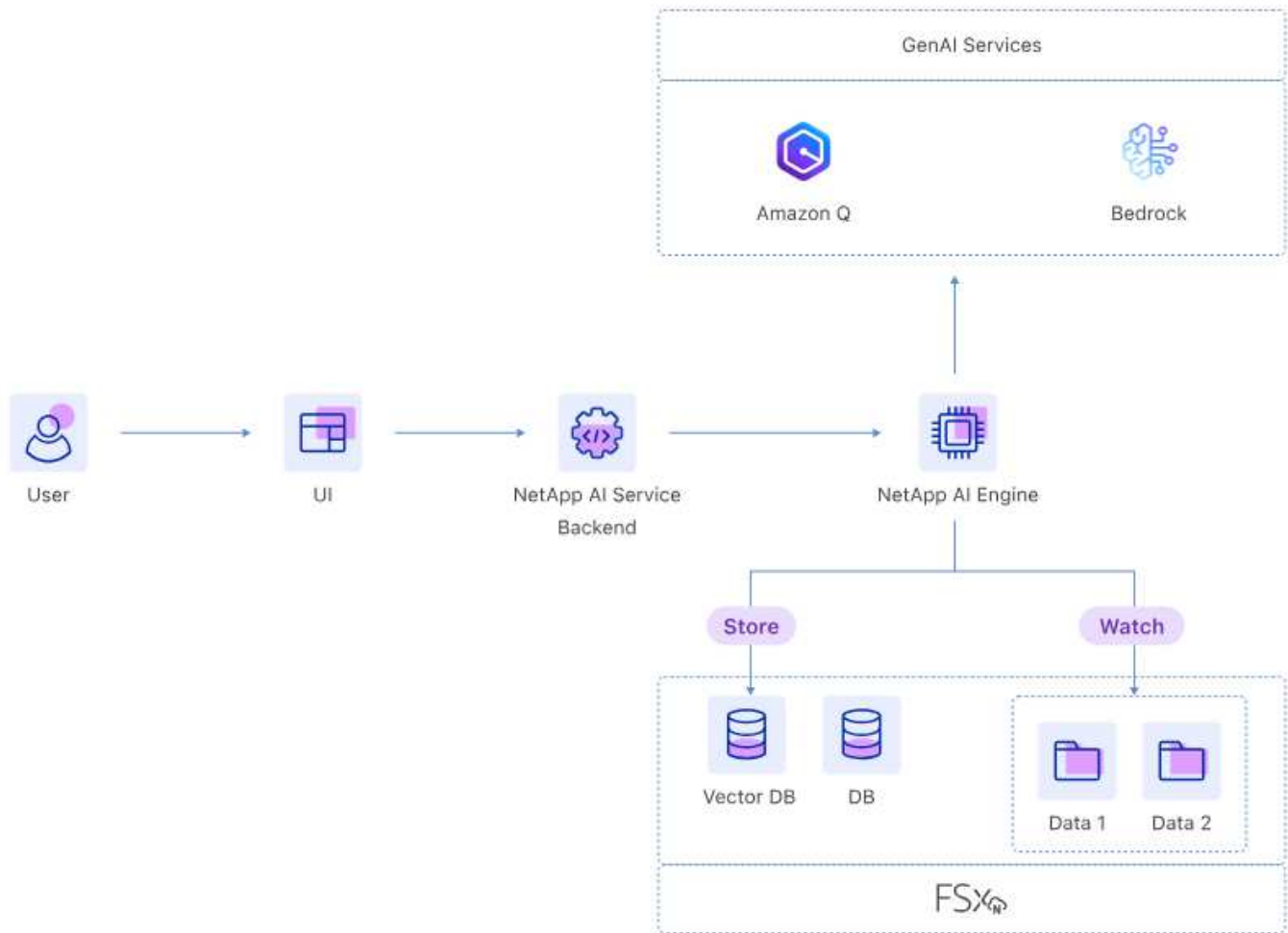
NetApp Workload Factory for GenAI 简化了使用检索增强生成 (RAG) 构建生成式 AI 应用程序所需的基础架构的部署过程。具体来说，GenAI 具有以下优势：

- 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"
}

```

```

    }
  },
  "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

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

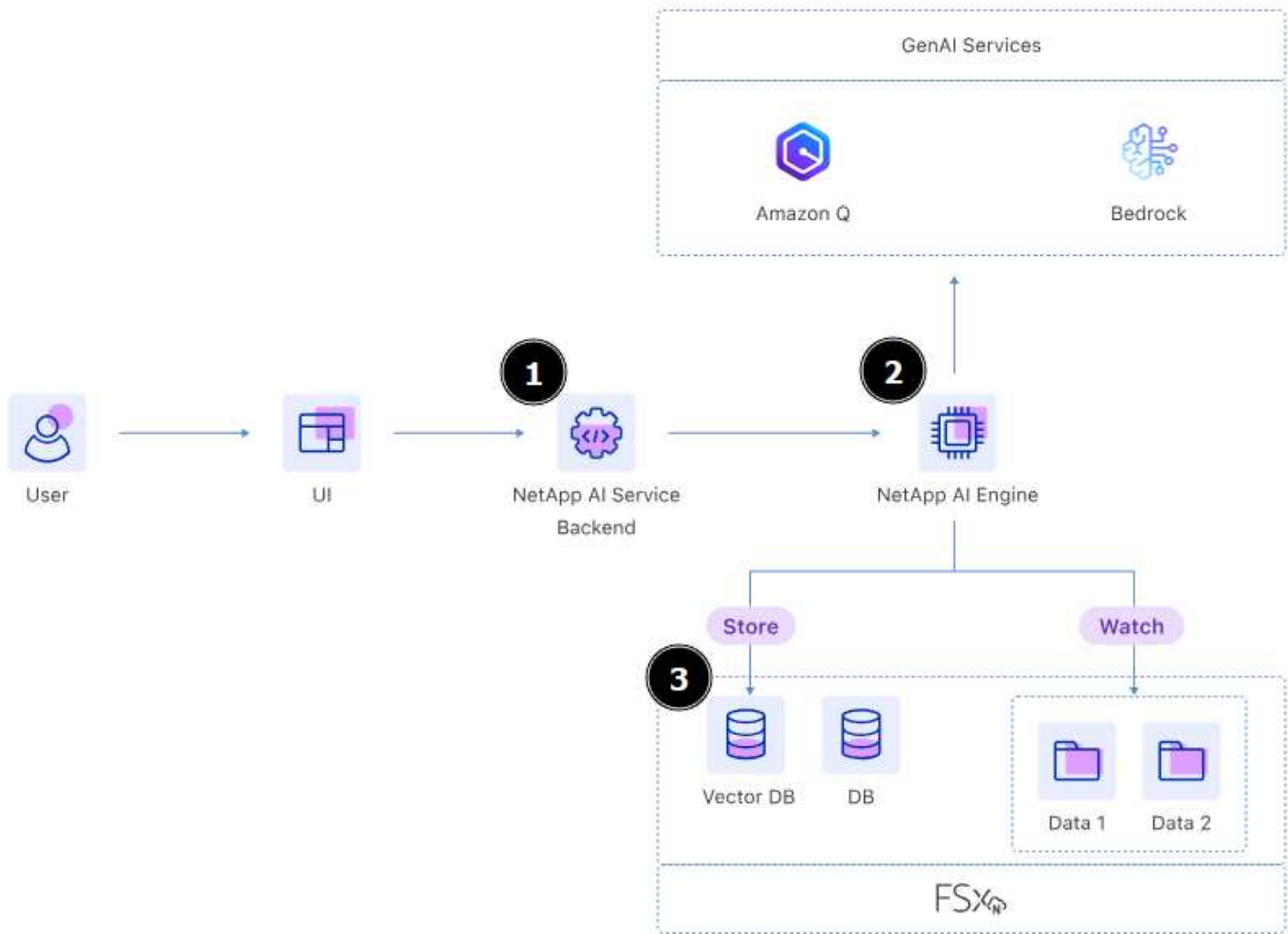
FSx for ONTAP文件系统为您的GenAI系统提供存储。

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、(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。

使用GenAI为Amazon Q Business创建连接器

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

文件格式	扩展
降价	.MD
Microsoft Word	.docx
纯文本	.txt
可移植文档格式	.pdf
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

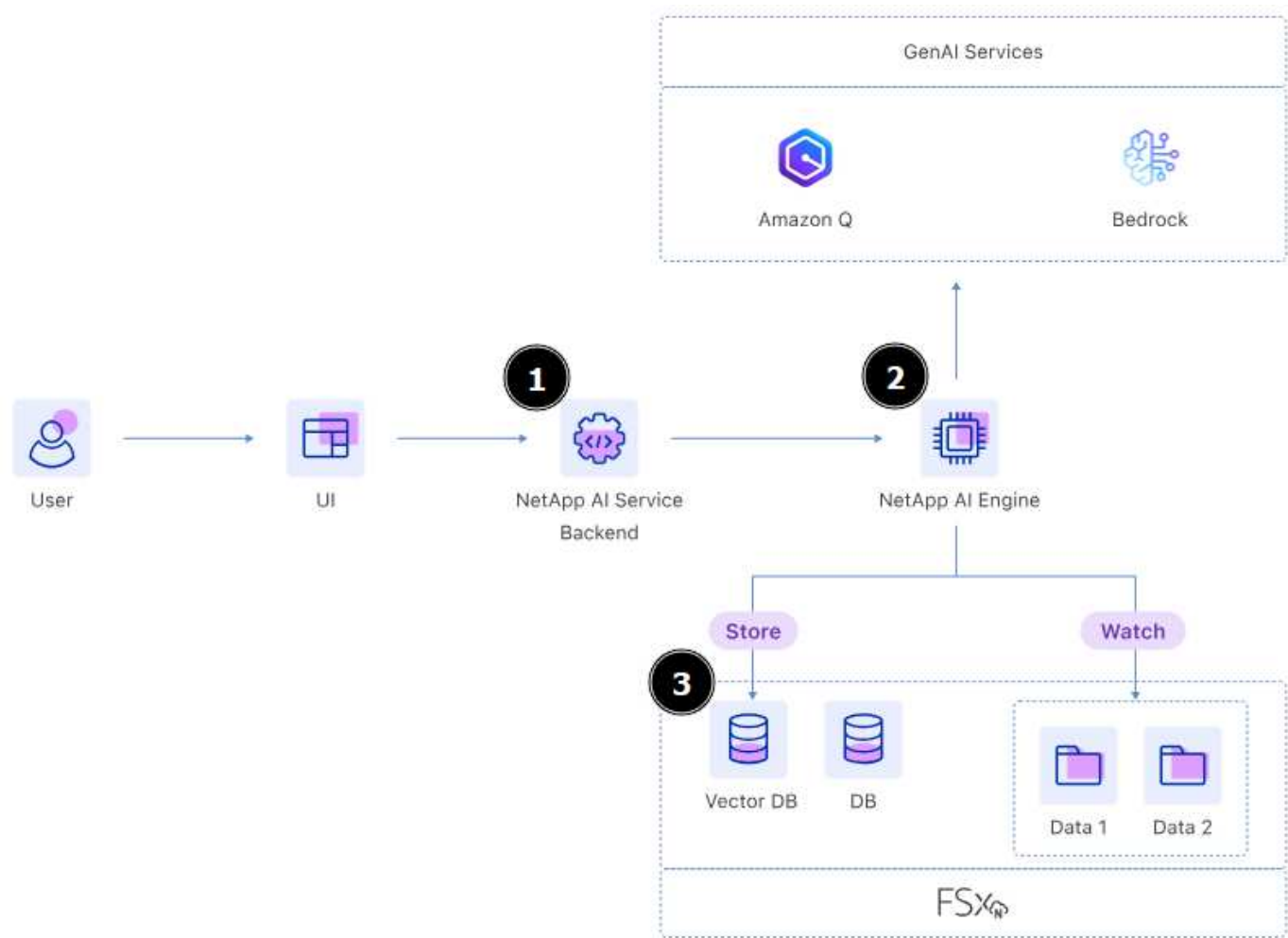
FSx for ONTAP文件系统为您的GenAI系统提供存储。

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。

为 Amazon Q Business 创建 NetApp 连接器

部署 AI 基础架构并确定将从 FSx for ONTAP 数据存储区使用的数据源后，您就可以为 Amazon Q Business 定义 NetApp 连接器了。

在继续操作之前、请确保您的环境符合["要求"](#)Amazon Q Business的要求。

关于此任务

您组织的数据源可能包含个人身份信息 (PII)。为了保护这些敏感信息，您可以在定义连接器时启用 `_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数据源，这可能会产生成本。

定义接头

为 Amazon Q Business 创建 NetApp 连接器。该连接器支持 GenAI 和 Amazon Q Business 之间的 API 和数据源通信。

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*Amazon Q 商业连接器*。
4. 在"定义连接器"页面上、配置连接器设置：
 - a. **Name**：输入要用于连接器的名称。
 - b. **Description**：输入连接器的详细说明。
 - c. **Amazon Q**：要集成的Amazon Q Business实例的区域和应用程序名称。
 - d. 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在NetApp控制台中完成所有操作。["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"](#)：
 - A. 创建控制台代理
 - B. 确保您的环境满足前提条件
 - C. 部署NetApp数据分类



启用数据保障功能后，GenAI将通过仅输入纯文本(不包括嵌入的图像或媒体文本)并屏蔽任何私有或敏感数据来处理.txt、.MD、.csv、.DOCX和.pdf文件。所有其他文件类型均正常处理、不会屏蔽私有或敏感数据。

- e. **FSx for ONTAP**文件系统：当您为 Amazon Q Business 定义新的NetApp Connector 时，Workload Factory 会创建一个新的Amazon FSx for NetApp ONTAP卷来存储连接器信息。选择将在其中创建新卷的现有文件系统和 SVM（也称为存储 VM）。
- f. 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。 GenAI 会根据您选择的快照策略以一定频率自动创建存储连接器信息的卷的定期快照。

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。

5. 选择*创建连接器*以将Amazon Q Business与GenAI集成。

创建连接器时、将显示进度指示器。

创建连接器后、您可以选择向连接器添加数据源、以便Amazon Q Business可以先插入数据、然后将其添加到索引中。建议您选择*添加数据源*，然后立即添加一个或多个数据源。

将数据源添加到连接器

您可以添加一个或多个数据源、以便使用您组织的数据填充Amazon Q Business索引。

关于此任务

- 支持的最大数据源数为10。
- 有关亚马逊Q业务索引的特定服务限制、请参阅 ["Amazon Q Business文档"](#)。

步骤

1. 选择*添加数据源*后，将出现*选择文件系统*页面。
2. 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*。
3. 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。

4. 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*。
5. 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：
 - 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。
 - 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。

6. 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：访问此连接器的聊天机器人用户将仅获得其有权访问的数据源的查询响应。
 - 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。



Amazon Q Business Connector数据源不支持Active Directory组权限。

7. 选择*添加*以将此数据源添加到Amazon Q Business连接器。

结果

数据源嵌入到Amazon Q Business索引中。当数据源完全嵌入时、状态将从"嵌入"更改为"嵌入"。

向连接器添加单个数据源后、您可以在Amazon Q Business聊天机器人环境中对其进行测试、并进行任何必要的更改、然后再向用户提供服务。您也可以按照相同的步骤向连接器添加其他数据源。

管理和监控

管理GenAI基础架构

您可以查看有关已部署的GenAI RAG基础架构的详细信息、或者删除不再需要的聊天机器人基础架构。

查看有关基础架构的信息

您可以查看有关聊天机器人基础架构的信息。

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. 选择“基础设施”菜单。
4. 查看有关基础架构的信息、其中包括有关以下组件的详细信息：
 - AWS 设置
 - 基础架构设置
 - AI引擎
 - 引导程序数据库

删除基础架构

如果您不再需要为一个或多个聊天机器人部署的聊天机器人基础架构，则可以将其从 Workload Factory 中删除。



此基础架构上部署的所有聊天机器人都将被禁用、并且所有聊天历史记录都将被删除。

此操作仅从 Workload Factory 中删除到 AI 基础设施的链接；它不会从 AWS 中删除所有组件。您需要从 AWS 中手动删除以下基础设施组件：

- VM实例
- 专用端点
- FSx for ONTAP文件系统上包含AI数据库的卷
- IAM角色
- 策略
- 安全组

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. 选择“基础设施”菜单。
4. 选择  并选择*删除聊天机器人基础架构*。
5. 确认要删除基础架构，然后选择*Remove*。

结果

聊天机器人基础设施组件已从 Workload Factory 中删除。

管理GenAI知识库

创建知识库后、您可以查看知识库详细信息、修改知识库、集成其他数据源或删除知识库。

查看有关知识库的信息

您可以查看有关知识库和集成的数据源设置的信息。

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. 选择要查看的知识库。

如果已定义、则当前正在使用的对话启动器将显示在右窗格中。

4. 要查看知识库详细信息，请选择，然后选择  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、所有嵌入数据源的列表等。

如果要进行任何更改，可使用*Actions*菜单管理知识库。

编辑知识库

您可以通过更改某些设置来更新知识库、也可以添加或删除数据源。

每次在知识库中添加、修改或删除数据源时、都必须同步数据源、以便将其重新编制到知识库索引。同步是增量的、因此Amazon Bonrock仅处理FSx for ONTAP卷中自上次同步以来添加、修改或删除的对象。

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、选择要更新的知识库。
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、所有嵌入数据源的列表等。

5. 选择*操作*菜单，然后选择*编辑知识库*。

6. 在编辑知识库页面中，您可以更改知识库名称、描述、嵌入模型、聊天模型、功能启用，选择是自动还是手动创建对话启动器，以及包含知识库的卷使用的快照策略。

如果您使用"手动"模式启动对话、您也可以在此处更改启动对话的内容。



每个知识库扫描(包括嵌入)都将产生成本。如果在创建知识库后启用了数据防护、则会再次扫描知识库并产生成本。同样、如果您更改聊天模式、GenAI将重新扫描关联的数据源(产生成本)。

7. 完成更改后选择*保存*。

使用快照保护知识库

您可以通过创建和还原知识库卷的快照来保护知识库数据。您可以随时从快照还原以还原到先前版本的知识库。

与备份相比、快照速度更快、存储效率更高、并支持您使用不同的保护策略来保护每个知识库。快照可能会很有用的一些情形包括：

- 意外数据丢失或损坏
- 从知识文库中输入的错误数据中恢复
- 测试不同的数据源或分块策略、并在测试完成后快速还原

为知识库卷创建快照

您可以通过手动创建知识库卷的快照来保存知识库的状态。

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、选择要保护的知识库。
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、所有嵌入数据源的列表等。

5. 选择*操作*菜单，然后选择*快照>创建新快照*。
6. (可选)选择*Define Snapshot name*并为快照输入自定义名称。

定义自定义名称可以帮助您更好地确定快照的内容、以便将来需要还原快照。

7. 选择 * 创建 *。

创建知识库快照。

还原知识库卷的快照

您可以随时还原知识库卷的手动快照或计划快照。



如果存储在卷上的数据库已损坏或被删除、则无法使用生成型AI工作负载UI还原快照。解决方法是、您可以在托管此卷的ONTAP集群上使用还原此快照 ["ONTAP 命令行界面"](#)。

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、选择要还原的知识库。
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、所有嵌入数据源的列表等。

5. 选择*Actions*菜单，然后选择*Snapshot > Restore Snap照*。

此时将显示快照选择对话框、您可以在此对话框中查看为此知识库创建的快照列表。

6. (可选)如果希望在恢复快照后继续执行计划的和当前正在运行的数据源扫描，请取消选择*恢复快照后暂停运行和计划扫描*选项。

默认情况下、此选项处于启用状态、以确保在知识库处于部分还原状态时不会进行扫描、或者扫描不会使用旧数据更新新还原的知识库。

7. 从列表中选择要还原的快照。
8. 选择 * 还原 *。

克隆知识库

您可以从知识库快照创建新的知识库。如果原始知识库已损坏或丢失、则此功能非常有用。

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、选择要还原的知识库。
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、所有嵌入数据源的列表等。

5. 选择*操作*菜单，然后选择*快照>克隆知识库*。

此时将显示克隆对话框。

6. 或者，如果希望在克隆快照后继续执行计划的和当前正在运行的数据源扫描，请取消选择*Pause running and Scheduled scans after克隆快照*选项。

默认情况下、此选项处于启用状态、以确保在知识库处于部分还原状态时不会进行扫描、或者扫描不会使用旧数据更新新还原的知识库。

7. 从列表中选择要克隆的快照。

8. 选择 * 继续 *。
9. 输入新知识库的名称。
10. 为新知识库选择文件系统SVM和卷名称。
11. 选择 * 克隆 *。

向知识库添加其他数据源

您可以在知识库中嵌入其他数据源、以使用其他组织数据填充该知识库。

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、选择要添加数据源的知识库。
4. 选择，然后选择  添加数据源。
5. 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。

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

数据源已集成到您的知识库中。

将数据源与知识库同步

数据源每天自动与相关知识库同步一次、以便任何数据源更改都反映在聊天机器人中。如果您对任何数据源进行了更改、并且希望立即同步数据、则可以执行按需同步。

同步是增量的、因此Amazon Bonck仅处理数据源中自上次同步以来添加、修改或删除的对象。

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、选择要同步的知识库。
4. 选择 ... 并选择*管理知识库*。
5. 选择*Actions*菜单，然后选择*Scan Now (立即扫描)*。

您将看到一条消息、指示正在扫描数据源、扫描完成后、您将看到最后一条消息。

结果

知识库与所连接的数据源同步、任何活动的聊天机器人都将开始使用您的数据源中的最新信息。

暂停或恢复计划的同步

如果要暂停或恢复数据源的下次同步(扫描)、您可以随时执行此操作。如果要更改数据源、并且不希望在更改窗口期间进行同步、则可能需要暂停下一次计划同步。

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、选择要暂停或恢复扫描的知识库。
4. 选择 ... 并选择*管理知识库*。
5. 选择*Actions*菜单，然后选择*Scan > Pause Scheduled scan-或*Scan > Resume Scheduled scan-。

您将看到一条消息、指出下一个计划扫描已暂停或恢复。

在创建知识库之前评估聊天模式

您可以在创建知识库之前评估可用的基础聊天模式、以便了解哪种模式最适合您的实施。由于型号支持因AWS地区而异、请参见 ["此AWS文档页面"](#) 以验证您可以在计划部署知识库的地区使用哪些型号。



只有在未创建任何知识库时(即"知识库"清单页面中不存在任何知识库时)、此功能才可用。

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. 在知识库清单页面中、您将在页面右侧看到选择聊天机器人聊天模式的选项。
4. 从列表中选择聊天模式、然后在提示区域输入一组问题、以查看聊天机器人的响应方式。
5. 尝试多种模式、了解哪种模式最适合您的实施。

结果

创建知识库时、请使用聊天模式。

取消发布知识库

在发布知识库以便与聊天机器人应用程序集成后、如果要禁止聊天机器人应用程序访问知识库、可以取消发布。

取消发布知识库会停止任何聊天应用程序的工作。已禁用可访问知识库的唯一API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、选择要取消发布的知识库。
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*Actions*菜单，然后选择*Unpub*。

结果

此知识库已禁用、无法再由聊天机器人应用程序访问。

删除知识库

如果您不再需要知识库，您可以将其删除。当您删除知识库时，它将从 Workload Factory 中删除，并且包含该知识库的卷也将被删除。任何使用知识库的应用程序或聊天机器人都将停止工作。删除知识库是不可逆的。

删除知识库时、您还应解除知识库与关联的任何代理的关联、以完全删除与知识库关联的所有资源。

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、选择要删除的知识库。
4. 选择  并选择*管理知识库*。
5. 选择*操作*菜单，然后选择*删除知识库*。
6. 在“删除知识库”对话框中，确认要将其删除，然后选择*Delete*。

结果

知识库已从 Workload Factory 中删除，并且其关联卷也已被删除。

管理Amazon Q Business连接器

为Amazon Q Business创建连接器之后、您可以查看连接器详细信息、修改连接器、集成其他数据源或删除连接器。

查看有关连接器的信息

您可以查看有关连接器设置和集成的数据源的信息。

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. 从知识库和连接器清单页面中，选择要查看的连接器。
4. 要查看连接器详细信息，请选择，然后选择  **Manage connector**。

此页面显示已发布状态、数据源的嵌入状态、嵌入模式、所有嵌入数据源的列表等。

如果要进行任何更改，可使用*Actions*菜单来管理连接器。

编辑连接器

您可以通过更改某些设置来更新连接器、也可以添加或删除数据源。

每次在连接器中添加、修改或删除数据源时、GenAI都需要将数据源信息发送到Amazon Q Business、以便对其重新编制索引。同步是增量操作、因此Amazon Q Business仅处理FSx for ONTAP卷中自上次同步以来添加、修改或删除的对象。

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 & Connectors (知识库和连接器)"清单页面中、选择要更新的连接器。
4. 选择  并选择*Manage connector*。

此页面显示已发布状态、数据源的嵌入状态、嵌入模式、所有嵌入数据源的列表等。

5. 选择*Actions*菜单，然后选择*Edit connector*。
6. 在"Edit connector"页面中、您可以更改连接器名称、说明、嵌入模型、数据防护栏启用以及用于包含该连接器的卷的快照策略。



每次数据源扫描(包括嵌入)都会带来成本。如果在创建连接器后启用数据防护、则会再次扫描数据源并产生成本。

7. 进行更改后选择*保存*。

向连接器添加其他数据源

您可以在连接器中嵌入其他数据源、以使用其他组织数据填充该数据源。

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 & Connectors (知识库和连接器)"清单页面中、选择要添加数据源的连接器。
4. 选择，然后选择  添加数据源。
5. 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、(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

数据源集成到连接器中。

将数据源与连接器同步

数据源每天自动与关联连接器同步一次、以便任何数据源更改都反映在Amazon Q Business中。如果您对任何数据源进行了更改、并且希望立即同步(扫描)数据、则可以执行按需同步。

同步是增量操作、因此Amazon Q Business仅处理数据源中自上次同步以来添加、修改或删除的对象。

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，选择要同步的连接器。
4. 选择...并选择*Manage connector*。
5. 选择*Actions*菜单，然后选择*Scan Now (立即扫描)*。

您将看到一条消息、指示正在扫描数据源、扫描完成后、您将看到最后一条消息。

结果

此连接器将与所连接的数据源同步、Amazon Q Business将开始使用您的数据源中的最新信息。

暂停或恢复计划的同步

如果要暂停或恢复数据源的下次同步(扫描)、您可以随时执行此操作。如果要更改数据源、并且不希望在更改窗口期间进行同步、则可能需要暂停下一次计划同步。

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. 从连接器清单页面中、选择要暂停或恢复扫描的连接器。
4. 选择...并选择*Manage connector*。
5. 选择*Actions*菜单，然后选择*Scan > Pause Scheduled scan-或*Scan > Resume Scheduled scan-。

您将看到一条消息、指出下一个计划扫描已暂停或恢复。

删除连接器

如果您不再需要连接器，您可以将其删除。当您删除连接器时，它将从 Workload Factory 中删除，并且包含该连接器的卷也将被删除。删除连接器是不可逆的。

删除连接器时，还应解除连接器与与其关联的任何代理的关联，以完全删除与连接器关联的所有资源。

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 & Connectors (知识库和连接器)" 清单页面中、选择要删除的连接器。
4. 选择  并选择 ***Manage connector***。
5. 选择 ***Actions*** 菜单，然后选择 ***Delete connector***。
6. 在删除连接器对话框中，确认要将其删除，然后选择 ***Delete***。

结果

连接器已从 Workload Factory 中移除，并且其关联的卷已被删除。

管理 GenAI 数据源

在 FSx for ONTAP 文件系统上使用数据源创建知识库或连接器之后、您可以查看数据源详细信息、更新或更改数据源内容、编辑数据源设置或删除数据源。

查看有关数据源的信息

您可以查看有关数据源内容的信息、也可以通过知识库或连接器查看其嵌入状态。由于数据源与知识库或连接器相关联、因此您需要先选择知识库或连接器、然后才能查看数据源详细信息。

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. 选择数据源所在的知识库或连接器，然后选择并选择 ***管理知识库  或 *管理连接器。**

页面底部列出了关联的数据源。

4. 通过选择展开每行  以查看有关每个数据源的详细信息、例如 FSx for ONTAP 文件系统、卷以及数据源所在的路径。

此外、还会列出嵌入信息以及该数据源当前是嵌入到知识库还是连接器中。

编辑数据源设置

您可以编辑有关已与知识库或连接器集成的数据源的信息。在添加数据源后、大多数信息都会得到修复、但您可以对部分配置进行更改(例如、chunking 定义或权限感知)。

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. 在知识库清单页面中，选择数据源所在的知识库，然后选择  并选择*Manage Knowledge bas*。

页面底部列出了此知识库中的数据源。

4. 在要编辑的数据源所在行中，选择  ，然后选择*Edit data source*。
5. 在编辑数据源页面中、选择以展开区块定义行。
6. 更新分块策略和配置以及权限感知(对于SMB卷)的设置、然后选择*保存*。

结果

此时将更新数据源设置、AI系统将同步数据源、以便将其重新编制到知识库中。

更新现有数据源的内容

您可以随时更改数据源的内容、以添加或更新组织数据。如果此数据源正在知识库中使用、则必须同步此数据源、以便将其重新编制到知识库索引。同步是增量的、因此Amazon Bonrock仅处理FSx for ONTAP卷中自上次同步以来添加、修改或删除的对象。

数据源每天自动与知识库同步一次、以便任何数据源更改都反映在聊天机器人中。如果您对数据源进行了更改，并且希望立即同步数据，则可以 ["执行按需同步"](#)。

删除数据源

如果您不再需要某个数据源加入知识库、可以将其删除。

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. 从知识库清单页面中，选择数据源所在的知识库，然后选择并选择*管理知识库  *。

页面底部列出了此知识库中的数据源。

4. 在要删除的数据源所在的行中，选择，然后选择  **Delete data source**。
5. 在删除数据源对话框中，确认要将其删除，然后选择*Confirm *

结果

数据源将从知识库中删除、而AI系统将从知识库中删除有关此数据源的索引信息。使用知识库的聊天机器人将无法再使用该数据源中的任何信息。

使用NetApp Workload Factory 中的 Tracker 监控工作负载操作

使用NetApp Workload Factory 中的 Tracker 监控和跟踪工作负载操作的执行情况并监控任务进度。

关于此任务

NetApp Workload Factory 提供监控功能 Tracker，因此您可以监控和跟踪工作负载操作的进度和状态、查看操作任务和子任务的详细信息以及诊断任何问题或故障。

Tracker提供了多种操作。您可以按时间范围(过去24小时、7天、14天或30天)、工作负载、状态和用户筛选作业；使用搜索功能查找作业；以及将作业表下载为CSV文件。您可以随时刷新Tracker、并快速重试失败的操作、或者编辑失败操作的参数并重试此操作。

Tracker支持两个级别的监控、具体取决于操作。每个任务(例如文件系统部署)都会显示任务说明、状态、开始时间、任务工期、用户、区域、代理资源、任务ID以及所有相关子任务。您可以查看API响应以了解操作期间发生的情况。

带有示例的跟踪器任务级别

- 级别1 (任务): 跟踪文件系统部署。
- 级别2 (子任务): 跟踪与文件系统部署相关的子任务。

操作状态

跟踪器中的操作状态如下: `_in Progress _`、`_Success _`和`_Failed _`。

操作频率

操作频率取决于作业类型和作业计划。

事件保留

事件将在用户界面中保留30天。

跟踪和监控操作

使用 Tracker 跟踪和监控 Workload Factory 控制台中的操作。

步骤

1. 使用其中一个登录["控制台体验"](#)。
2. 从工作量菜单中，选择*管理*，然后选择*跟踪器*。
3. 在 Tracker 中，使用过滤器或搜索来缩小工作结果范围。您还可以下载工作报告。

查看API请求

在跟踪程序的"Code"框中查看某个任务的API请求。

步骤

1. 在Tracker中、选择一个任务。
2. 选择操作菜单，然后选择*查看 API 请求*。

重试失败的操作

在Tracker中重试失败的操作。您还可以复制失败操作的错误消息。



对于失败的操作、您最多可以重试10次。

步骤

1. 在Tracker中、选择失败的操作。
2. 选择操作菜单，然后选择*重试*。

结果

此时将重新启动此操作。

编辑失败的操作并重试

编辑失败操作的参数、然后在Tracker外部重试此操作。

步骤

1. 在Tracker中、选择失败的操作。
2. 选择操作菜单，然后选择*编辑并重试*。

系统会将您重定向到操作页面、您可以在其中编辑参数并重试此操作。

结果

此时将重新启动此操作。进入跟踪器以查看操作状态。

知识和支持

注册以获取NetApp Workload Factory for GenAI 的支持

需要注册支持才能获得针对NetApp Workload Factory 及其存储解决方案和服务的技术支持。您必须注册NetApp Console 的支持，它是独立于 Workload Factory 的基于 Web 的控制台。

注册支持并不能使NetApp获得云提供商文件服务的支持。有关云提供商文件服务、其基础设施或使用该服务的任何解决方案的技术支持，请参阅该产品的 Workload Factory 文档中的“获取帮助”。

["适用于 ONTAP 的 Amazon FSX"](#)

支持注册概述

注册您的帐户 ID 支持订阅（位于NetApp控制台中的“支持资源”页面上的 20 位 960xxxxxxxx 序列号）作为您的单一支持订阅 ID。每个NetApp帐户级支持订阅都必须注册。

注册后即可实现开立支持票和自动生成案例等功能。通过将NetApp支持站点 (NSS) 帐户添加到NetApp控制台即可完成注册，如下所述。

注册您的帐户以获得NetApp支持

要注册支持并激活支持权利，您帐户中的一名用户必须将NetApp支持站点帐户与其NetApp控制台登录名关联。如何注册NetApp支持取决于您是否已经拥有NetApp支持站点 (NSS) 帐户。

具有NSS帐户的现有客户

如果您是拥有 NSS 帐户的NetApp客户，则只需通过NetApp控制台注册即可获得支持。

步骤

1. 在 Workload Factory 控制台的右上角，选择 帮助 > 支持。

选择此选项将在新浏览器选项卡中打开NetApp控制台并加载支持仪表板。

2. 从NetApp控制台菜单中，选择 管理，然后选择 凭据。
3. 选择*用户凭据*。
4. 选择*添加NSS凭证*，然后按照NetApp 支持站点(NSS)鉴定提示进行操作。
5. 要确认注册过程是否成功，请选择帮助图标，然后选择*Support*。

“资源”页面应显示您的帐户已注册支持。



9601111122222444455555
Account Serial Number



Registered for Support
Support Registration

请注意，如果其他NetApp控制台用户尚未将NetApp支持站点帐户与其NetApp控制台登录名关联，则他们将看不到相同的支持注册状态。但是，这并不意味着您的NetApp帐户未注册支持。只要账户中有一位用户遵循了这些步骤，那么您的账户就注册成功了。

现有客户、但无NSS帐户

如果您是现有NetApp客户，拥有现有许可证和序列号但没有 NSS 帐户，则需要创建一个 NSS 帐户并将其与您的NetApp控制台登录关联。

步骤

1. 完成以创建NetApp支持站点帐户 "[NetApp 支持站点 用户注册表](#)"
 - a. 请务必选择适当的用户级别、通常为* NetApp客户/最终用户*。
 - b. 请务必复制上面用于序列号字段的NetApp帐户序列号 (960xxxx)。这将加快账户处理速度。
2. 完成以下步骤，将您的新 NSS 帐户与您的NetApp控制台登录关联[具有NSS帐户的现有客户](#)。

NetApp的新品牌

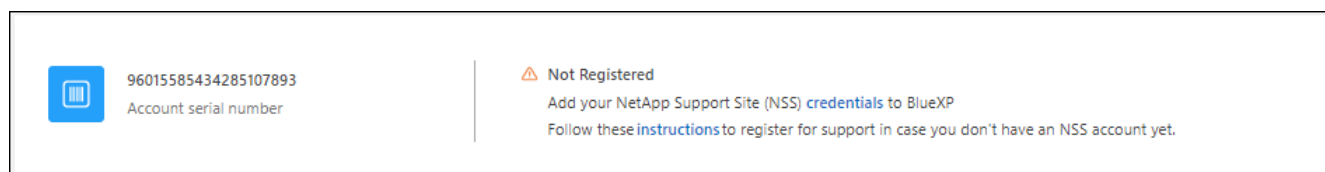
如果您是NetApp的新客户、并且没有NSS帐户、请按照以下每个步骤进行操作。

步骤

1. 在 Workload Factory 控制台的右上角，选择 帮助 > 支持。

选择此选项将在新浏览器选项卡中打开NetApp控制台并加载支持仪表板。

2. 从支持资源页面中找到您的帐户ID序列号。



3. 导航到 "[NetApp的支持注册站点](#)" 并选择*我不是NetApp的注册客户*。
4. 填写必填字段(带有红色星号的字段)。
5. 在*产品线*字段中、选择*云管理器*、然后选择适用的计费提供商。
6. 复制上述第2步中的帐户序列号、完成安全检查、然后确认您已阅读NetApp的全球数据隐私政策。

系统会立即向提供的邮箱发送一封电子邮件、以完成此安全事务。如果验证电子邮件未在几分钟内收到、请务必检查您的垃圾邮件文件夹。

7. 在电子邮件中确认操作。

确认将向NetApp提交您的请求、并建议您创建NetApp 支持站点 帐户。

8. 完成以创建NetApp支持站点帐户 "[NetApp 支持站点 用户注册表](#)"
 - a. 请务必选择适当的用户级别、通常为* NetApp客户/最终用户*。
 - b. 请务必复制上面用于序列号字段的帐户序列号(960xxxx)。这样可以加快帐户处理速度。

完成后

在此过程中、NetApp应与您联系。这是针对新用户的一次性入职练习。

拥有NetApp支持站点帐户后，请按照以下步骤将该帐户与您的NetApp控制台登录关联[具有NSS帐户的现有客户](#)。

GenAI故障排除

了解如何解决您可能遇到的一些常见问题。

常见问题和解决方案

如果您遇到上述问题之一、可以使用"解决方法"列中的步骤尝试解决该问题。

区域	问题描述	发生原因	临时解决策
部署	由于卷已存在、部署失败。	NetApp Workload Factory for GenAI 需要在部署过程中创建一个新卷，但使用您指定的名称的卷已经存在。	指定要用于新卷的唯一名称、然后重新尝试部署。
部署	部署失败，因为NetApp Workload Factory for GenAI 无法挂载该卷。	FSx for NetApp ONTAP所需的一个或多个入站端口已关闭或筛选。	打开以下入站端口：

| 协议 | 端口 | 目的

| 所有 ICMP | 全部 | Ping 实例

| HTTPS | 443 | 从Connector访问fsxadmin管理LIF、以便向FSX发送API调用

| SSH | 22 | SSH 访问集群管理 LIF 或节点管理 LIF 的 IP 地址

| TCP | 111 | 远程过程调用 NFS

| TCP | 139 | 用于 CIFS 的 NetBIOS 服务会话

| TCP | 161-162 | 简单网络管理协议

| TCP | 445 | Microsoft SMB/CIFS over TCP （通过 TCP ）和 NetBIOS 成帧

| TCP | 635 | NFS 挂载

| TCP | 749 | Kerberos

| TCP | 2049 | NFS 服务器守护进程

| TCP | 3260 | 通过 iSCSI 数据 LIF 进行 iSCSI 访问

| TCP | 4045 | NFS 锁定守护进程

| TCP | 4046 | NFS 的网络状态监视器

| TCP | 10000 | 使用 NDMP 备份

| TCP | 11104 | 管理 SnapMirror 的集群间通信会话

| TCP | 11105 | 使用集群间 LIF 进行 SnapMirror 数据传输

| UDP | 111 | 远程过程调用 NFS

| UDP | 161-162 | 简单网络管理协议

| UDP | 635 | NFS 挂载

| UDP | 2049 | NFS 服务器守护进程

| UDP | 4045 | NFS 锁定守护进程

| UDP | 4046 | NFS 的网络状态监视器

| UDP | 4049 | NFS Rquotad 协议

维护	AI引擎启动失败，您在*Knowledge Bases*页面上看到错误"AI引擎实例错误"。	AI引擎实例已损坏或不存在。	选择*重建*按钮。NetApp Workload Factory for GenAI 重建基础架构并显示重建进度。完成后，您的知识库将重新连接到重建的基础设施，并显示知识库列表。
维护	AI引擎启动失败，您在*Knowledge Bases*页面上看到错误"The GenAI Engine Instance is Stop"(GenAI引擎实例已停止)。	AI引擎实例未运行。	使用AWS管理控制台或AWS命令行界面启动AI引擎实例。

维护	AI引擎启动失败，您在*Knowledge Bases*页面上看到错误"The GenAI引擎server is not starting"(GenAI引擎服务器未响应)。	AI引擎实例未响应。	请使用以下恢复步骤： 步骤 1. 修改GenAI引擎实例安全组、以启用对GenAI引擎实例的SSH访问。 2. 使用SSH登录到实例。 3. 运行以下命令： <pre>docker- compose up</pre>
----	--	------------	---

维护	NetApp Workload Factory for GenAI 使用的后端 Docker 实例无法启动。	已删除此卷并重新启动EC2实例。	请使用以下恢复步骤： 步骤 1. 在FSx for NetApp ONTAP上创建新卷。 例如，卷名可以是 netapp_ai，卷路径可以是 /netapp_ai。 2. 通过SSH连接到Amazon EC2实例。 3. 列出卷： <pre>docker volume list</pre> 4. 删除旧卷： <pre>docker volume rm ec2- user_persist ent_folder</pre> 5. `docker-compose.yml` 使用文本编辑器打开文件。 6. 在 `volumes` 部分中、 将设备路径更改为新卷路径。例如：
----	--	------------------	---

维护	NetApp Workload Factory for GenAI 使用的后端 Docker 实例无法启动。	已删除此根卷。	使用名称和路径创建一个卷、然后从Amazon EC2 重新启动后端Docker实例。
维护	NetApp Workload Factory for GenAI 使用的后端 Docker 实例无法启动。	已删除此根卷。	使用名称和路径创建一个卷、然后从Amazon EC2 重新启动后端Docker实例。

获取有关NetApp Workload Factory for GenAI 的帮助

NetApp通过多种方式为 Workload Factory 及其云服务提供支持。全天候提供广泛的免费自助支持选项，例如知识库 (KB) 文章和社区论坛。您的支持注册包含通过网络工单获取的远程技术支持。

获取FSx for ONTAP支持

有关 FSx for ONTAP、其基础架构或任何使用该服务的解决方案的技术支持，请参阅该产品的Workload Factory 文档中的“获取帮助”。

"适用于 ONTAP 的 Amazon FSX"

要获得特定于Workload Factory及其存储解决方案和服务的技术支持、请使用下面所述的支持选项。

使用自助支持选项

这些选项每周 7 天，每天 24 小时免费提供：

- 文档
您当前正在查看的 Workload Factory 文档。
- "知识库"
搜索 Workload Factory 知识库以查找有助于解决问题的文章。
- "社区"
加入 Workload Factory 社区以关注正在进行的讨论或创建新的讨论。

向NetApp支持部门创建案例

除了上述自助支持选项之外、您还可以在激活支持后与NetApp支持专家合作解决任何问题。

开始之前

要使用*创建案例*功能，您必须首先注册支持。将您的NetApp支持站点凭据与您的 Workload Factory 登录名关联。"了解如何注册获取支持"。

步骤

```
'nfs'
o:
"addr=svm-
0d673008aaca
12bc3.\
fs-
fsx-workload-
1.amazonaws.
com,nolock,s
oft,rw"
device:
':/netapp_ai
' # Path to
new volume
```

1. 在 Workload Factory 控制台的右上角，选择 帮助 > 支持。

选择此选项将在新浏览器选项卡中打开NetApp控制台并加载支持仪表板。

2. 在*资源*页面上、在技术支持下选择一个可用选项：

- a. 如果您想通过电话与某人通话，请选择*呼叫我们*。系统会将您定向到netapp.com上的一个页面、其中列出了您可以拨打的电话号码。

- b. 选择*创建案例*向NetApp支持专家开立TT：

- 服务：选择*工作负载工厂*。
- 案例优先级：选择案例的优先级、可以是"低"、"中"、"高"或"严重"。

要了解有关这些优先级的更多详细信息、请将鼠标悬停在字段名称旁边的信息图标上。

- *问题描述*：提供问题的详细问题描述、包括任何适用的错误消息或您执行的故障排除步骤。
- 其他电子邮件地址：如果您希望其他人了解此问题描述、请输入其他电子邮件地址。
- 附件(可选)：一次最多上传五个附件。

每个文件的附件数限制为25 MB。支持以下文件扩展名：txt、log、pdf、jp6/jpeu、rtf、doc/docx、xls/xlsx和csv。

ntapitdemo 

NetApp Support Site Account

Service

Select ▼

Working Enviroment

Select ▼

Case Priority 

Low - General guidance ▼

Issue Description

Provide detailed description of problem, applicable error messages and troubleshooting steps taken.

Additional Email Addresses (Optional) 

Type here

Attachment (Optional)

No files selected

 Upload 

完成后

此时将显示一个弹出窗口、其中包含您的支持案例编号。NetApp支持专家将审核您的案例、并尽快与您联系。

要查看支持案例的历史记录，您可以选择*设置>时间线*并查找名为“创建支持案例”的操作。最右侧的按钮可用于展开操作以查看详细信息。

尝试创建案例时、您可能会遇到以下错误消息：

"您无权针对选定服务创建案例"

此错误可能意味着 NSS 帐户及其关联的记录公司与NetApp控制台帐户序列号的记录公司不同（即。960xxxx）或系统序列号。您可以使用以下选项之一寻求帮助：

- 使用产品内聊天功能
- 请通过提交非技术案例 <https://mysupport.netapp.com/site/help>

管理支持案例(预览)

您可以直接从NetApp控制台查看和管理活动和已解决的支持案例。您可以管理与您的 NSS 帐户和公司相关的案例。

案例管理以预览形式提供。我们计划改进此体验、并在即将发布的版本中添加增强功能。请通过产品内聊天向我们发送反馈。

请注意以下事项：

- 页面顶部的案例管理信息板提供了两个视图：
 - 左侧视图显示了您提供的用户NSS帐户在过去3个月内打开的案例总数。
 - 右侧视图显示了过去3个月内根据用户NSS帐户在公司级别开立的案例总数。

此表中的结果反映了与选定视图相关的案例。

- 您可以添加或删除感兴趣的列、也可以筛选优先级和状态等列的内容。其他列仅提供排序功能。

有关更多详细信息、请查看以下步骤。

- 在每个案例级别、我们可以更新案例备注或关闭尚未处于"已关闭"或"待关闭"状态的案例。

步骤

1. 在 Workload Factory 控制台的右上角，选择 帮助 > 支持。

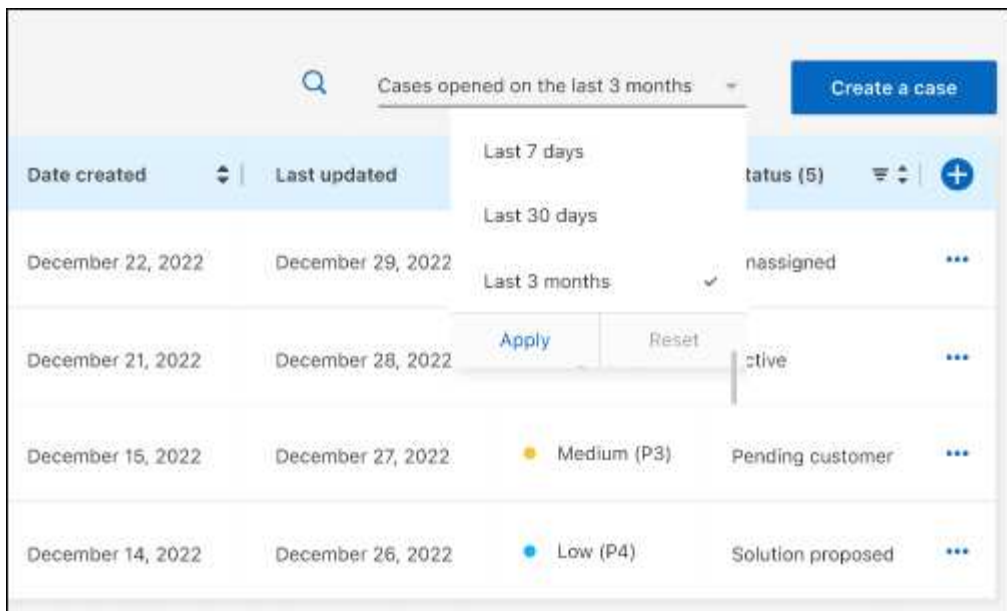
选择此选项将打开NetApp控制台的新浏览器选项卡并加载支持仪表板。

2. 选择*案例管理*，如果出现提示，请将您的 NSS 帐户添加到NetApp控制台。

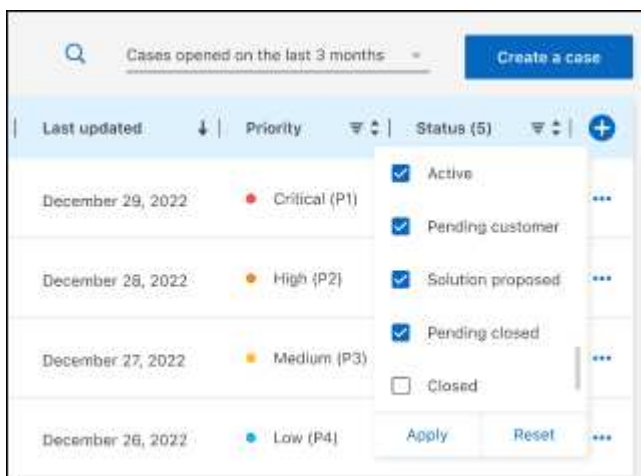
案例管理*页面显示与您的**NetApp**控制台用户帐户关联的 **NSS** 帐户相关的未结案例。这与出现在 ***NSS 管理** 页面顶部的 NSS 帐户相同。

3. 也可以修改表中显示的信息：

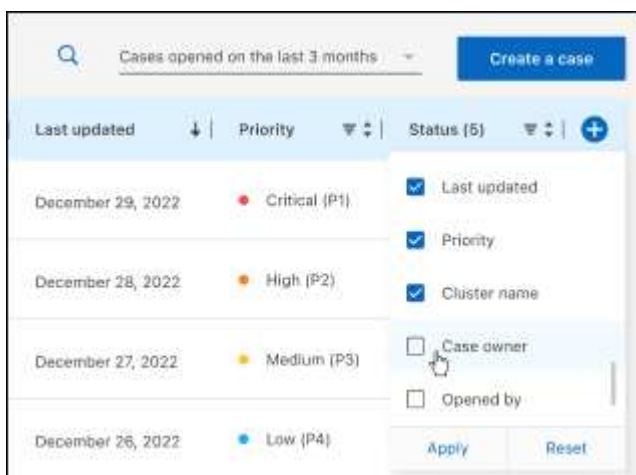
- 在“组织案例”下，选择“查看”以查看与贵公司关联的所有案例。
- 通过选择确切的日期范围或选择其他时间范围来修改日期范围。



- 筛选列的内容。



- 通过选择并选择要显示的列来更改表中 [表中显示的加号图标] 显示的列。

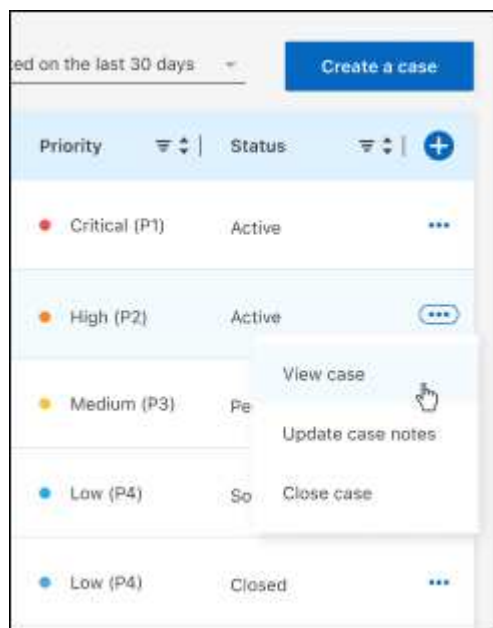


4. 通过选择并选择一个可用选项来管理现有案例 [包含三个点的图标、显示在表的最后一列中]：

- 查看案例：查看有关特定案例的完整详细信息。
- 更新案例注释：提供有关您的问题的更多详细信息、或者选择*上传文件*最多附加五个文件。

每个文件的附件数限制为25 MB。支持以下文件扩展名：txt、log、pdf、jp6/jpeu、rtf、doc/docx、xls/xlsx和csv。

- 关闭案例：提供关闭案例的详细原因，然后选择*关闭案例*。



NetApp Workload Factory for GenAI 法律声明

法律声明提供对版权声明、商标、专利等的访问。

版权

["https://www.netapp.com/company/legal/copyright/"](https://www.netapp.com/company/legal/copyright/)

商标

NetApp、NetApp 徽标和 NetApp 商标页面上列出的标记是 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。

["https://www.netapp.com/company/legal/trademarks/"](https://www.netapp.com/company/legal/trademarks/)

专利

有关 NetApp 拥有的专利的最新列表，请访问：

<https://www.netapp.com/pdf.html?item=/media/11887-patentspage.pdf>

隐私政策

["https://www.netapp.com/company/legal/privacy-policy/"](https://www.netapp.com/company/legal/privacy-policy/)

开放源代码

通知文件提供有关 NetApp 软件中使用的第三方版权和许可证的信息。

["NetApp工作负载工厂"](#)

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。