Comece agora

GenAI

NetApp
October 06, 2025

Índice

Comece agora

Início rápido para conetores GenAI

Comece a criar um NetApp Connector para o Amazon Q Business usando os dados da sua organização que existem no Amazon FSx para sistemas de arquivos NetApp ONTAP. Depois de criar um conetor, os usuários finais podem acessar o assistente do Amazon Q Business para respostas focadas na organização para suas perguntas.

1 Efetue login no Workload Factory

Você precisará "crie uma conta no Workload Factory" e faça login usando um dos "experiências de console" .

2 Configure seu ambiente para atender aos requisitos do GenAI

Você precisará de credenciais da AWS para implantar a infraestrutura da AWS, um sistema de arquivos FSX for ONTAP implantado e descoberto, a lista de fontes de dados que você deseja integrar no seu conetor, acesso ao aplicativo Amazon Q Business e muito mais.

"Saiba mais sobre os requisitos do GenAI".

3 Identifique o sistema de arquivos FSX for ONTAP que contém as fontes de dados

As fontes de dados que você integrará em seu conetor podem estar localizadas em um único sistema de arquivos FSX for ONTAP ou em vários sistemas de arquivos FSX for ONTAP. Se esses sistemas estiverem em VPCs diferentes, eles devem estar acessíveis dentro da mesma rede ou os VPCs devem ser direcionados e usando a mesma região e conta da AWS que o mecanismo de IA.

"Saiba como identificar fontes de dados".

4 Implantar a infraestrutura do GenAI

Inicie o assistente de implantação de infraestrutura para implantar a infraestrutura do GenAI em seu ambiente AWS. Esse processo implanta uma instância do EC2 para o mecanismo do NetApp GenAI e um volume em um sistema de arquivos FSX for ONTAP para conter os bancos de dados do NetApp AI Engine. O volume é utilizado para armazenar informações sobre o conetor.

"Saiba como implantar a infraestrutura do GenAI".

O que vem a seguir

Agora você pode criar um conetor para o Amazon Q Business para fornecer respostas focadas na organização aos usuários finais.

Requisitos do conetor GenAI

Certifique-se de que o Workload Factory e o AWS estejam configurados corretamente antes de criar um NetApp Connector para o Amazon Q Business.

Requisitos básicos do GenAI

A GenAI tem requisitos gerais que seu ambiente precisa atender antes de começar.

Login e conta do Workload Factory

Você precisará "crie uma conta no Workload Factory" e faça login usando um dos "experiências de console" .

Credenciais e permissões da AWS

Você precisa adicionar credenciais da AWS ao Workload Factory com permissões de leitura/gravação, o que significa que você usará o Workload Factory no modo *leitura/gravação* para o GenAI.

As permissões dos modos *básico* e *somente leitura* não são suportadas no momento.

Ao configurar suas credenciais, selecionar permissões como mostrado abaixo fornece acesso total para gerenciar os sistemas de arquivos FSX for ONTAP e implantar e gerenciar a instância do GenAI EC2 e outros recursos da AWS necessários para sua base de conhecimento e chatbot.

"Aprenda como adicionar credenciais da AWS ao Workload Factory"

Requisitos para o NetApp Connector para Amazon Q Business

Certifique-se de que seu ambiente atenda aos seguintes requisitos específicos para o Amazon Q Business Connectors.

Aplicação Amazon Q Business

Você precisa criar um aplicativo do Amazon Q Business ou usar um existente.

- Verifique se o aplicativo existe em uma das regiões da AWS.
- Certifique-se de que "criou um índice" tem para a aplicação.
- Certifique-se de que a aplicação não está num estado com falha.

FSX para sistema de arquivos ONTAP

Você precisa de um mínimo de um sistema de arquivos FSX for ONTAP:

- Um sistema de arquivos será usado (ou criado, se não existir) pelo mecanismo NetApp GenAI para armazenar informações sobre o conetor.

  Este sistema de arquivos FSX for ONTAP deve usar o FlexVol volumes. Os volumes FlexGroup não são compatíveis.

- Um ou mais sistemas de arquivos conterá as fontes de dados que você estará adicionando ao seu conetor.

  Um sistema de arquivos FSX for ONTAP pode ser usado para ambos esses fins, ou você pode usar vários sistemas de arquivos FSX for ONTAP.

- Você precisará conhecer a região, a VPC e a sub-rede da AWS onde reside o sistema de arquivos do AWS FSX for ONTAP.
- Você precisará considerar os pares de chave/valor de tag que deseja aplicar aos recursos da AWS que fazem parte dessa implantação (opcional).
- Você precisará saber as informações do par de chaves que permitem que você se conete com

segurança à instância do mecanismo de IA do NetApp.

"Saiba como implantar e gerenciar os sistemas de arquivos do FSX for ONTAP"

Identificar fontes de dados para adicionar a um conetor

Identifique ou crie os documentos (fontes de dados) que residem no seu sistema de arquivos FSX for ONTAP que você integrará no seu conetor. Essas fontes de dados permitem que o Amazon Q Business forneça respostas precisas e personalizadas para consultas de usuários com base em dados relevantes para sua organização.

Número máximo de fontes de dados

O número máximo de fontes de dados suportadas é 10.

Localização das fontes de dados

As fontes de dados podem ser armazenadas em um único volume ou em uma pasta dentro de um volume, em um compartilhamento SMB ou exportação NFS em um sistema de arquivos do Amazon FSX for NetApp ONTAP. As fontes de dados também podem ser armazenadas no Amazon FSX for NetApp ONTAP volumes que estão em uma relação de proteção de dados da NetApp SnapMirror.

Não é possível selecionar documentos individuais dentro de um volume ou pasta, portanto, você deve garantir que cada volume ou pasta que contém fontes de dados não contenha documentos estranhos que não devem ser integrados à sua base de conhecimento.

Você pode adicionar várias fontes de dados a cada conetor, mas todas elas precisam residir nos sistemas de arquivos do FSX for ONTAP que estão acessíveis a partir da sua conta da AWS.

O tamanho máximo de arquivo para cada fonte de dados é de 50 MB.

Protocolos compatíveis

Os conetores dão suporte a dados de volumes que usam protocolos NFS ou SMB/CIFS. Ao selecionar arquivos armazenados usando o protocolo SMB, você precisará inserir as informações do ative Directory para que o conetor possa acessar os arquivos nesses volumes. Isso inclui o domínio do ative Directory, o endereço IP, o nome de usuário e a senha.

Ao armazenar sua fonte de dados em um compartilhamento (arquivo ou diretório) acessado pelo SMB, os dados só podem ser acessados por usuários ou grupos de chatbot que têm permissões para acessar esse compartilhamento. Quando esta "capacidade de reconhecimento de permissões" está ativada, o sistema de IA compara o e-mail do utilizador em auth0 com os utilizadores autorizados a visualizar ou utilizar os ficheiros na partilha SMB. O chatbot fornecerá respostas com base nas permissões do usuário para os arquivos incorporados.

Por exemplo, se você integrou arquivos 10 (fontes de dados) em seu conetor, e 2 dos arquivos são arquivos de recursos humanos que contêm informações restritas, apenas os usuários do chatbot que são autenticados para acessar esses arquivos 2 receberão respostas do chatbot que incluem dados desses arquivos.

Quando você adiciona fontes de dados a um Amazon Q Business Connector, apenas as permissões de usuário se aplicam a arquivos de origem de dados. As permissões de grupo não são aplicadas.

Se um arquivo em sua fonte de dados não tiver texto (por exemplo, uma imagem livre de texto), o Amazon Q Business não o indexa, mas Registra uma entrada no Amazon CloudWatch Logs observando a ausência de texto.

Formatos de arquivo de origem de dados suportados

Os seguintes formatos de arquivo de fonte de dados são atualmente suportados pelo NetApp Connector para Amazon Q Business.

Formato do ficheiro	Extensão
Arquivo de valores separados por vírgula	.csv
JSON e JSONP	.json
Markdown	.md
Microsoft Word	.docx
Texto simples	.txt
Formato de documento portátil	.pdf
Microsoft PowerPoint	.ppt ou .pptx
Hypertext Markup Language	.html
Extensible Markup Language (linguagem de marcação extensível)	.xml
XSLT	.xslt
Microsoft Excel	.xls
Formato Rich Text	.rtf

Implantar a infraestrutura do GenAI

Você precisa implantar a infraestrutura do GenAI para a estrutura RAG em seu ambiente antes de criar bases de conhecimento, conetores e aplicativos do FSX for ONTAP para sua organização. Os principais componentes da infraestrutura são o serviço Amazon bedrock, uma instância de máquina virtual para o mecanismo NetApp GenAI e um sistema de arquivos FSX for ONTAP.

A infraestrutura implantada pode oferecer suporte a várias bases de conhecimento, chatbots e conetores, portanto, você normalmente só precisará executar essa tarefa uma vez.

Detalhes da infraestrutura

Sua implantação do GenAI deve estar em uma região da AWS que tenha o Amazon bedrock habilitado. "Veja a lista de regiões suportadas"

A infraestrutura consiste nos seguintes componentes.

Serviço Amazon bedrock

O Amazon bedrock é um serviço totalmente gerenciado que permite que você use os modelos de base (FMS) das principais empresas de IA por meio de uma única API. Ele também fornece os recursos de que você precisa para criar aplicativos de IA generativos seguros.

"Saiba mais sobre a Amazon bedrock"

Amazon Q Business

O Amazon Q baseia-se no Amazon bedrock para fornecer um assistente de IA generativa totalmente gerenciado que você pode usar para responder perguntas e gerar conteúdo com base em informações de suas fontes de dados.

"Saiba mais sobre o Amazon Q Business"

Máquina virtual para o motor NetApp GenAI

O mecanismo NetApp GenAI é implantado durante esse processo. Ele fornece o poder de processamento para obter os dados de suas fontes de dados e, em seguida, gravar esses dados no banco de dados vetorial.

FSX para sistema de arquivos ONTAP

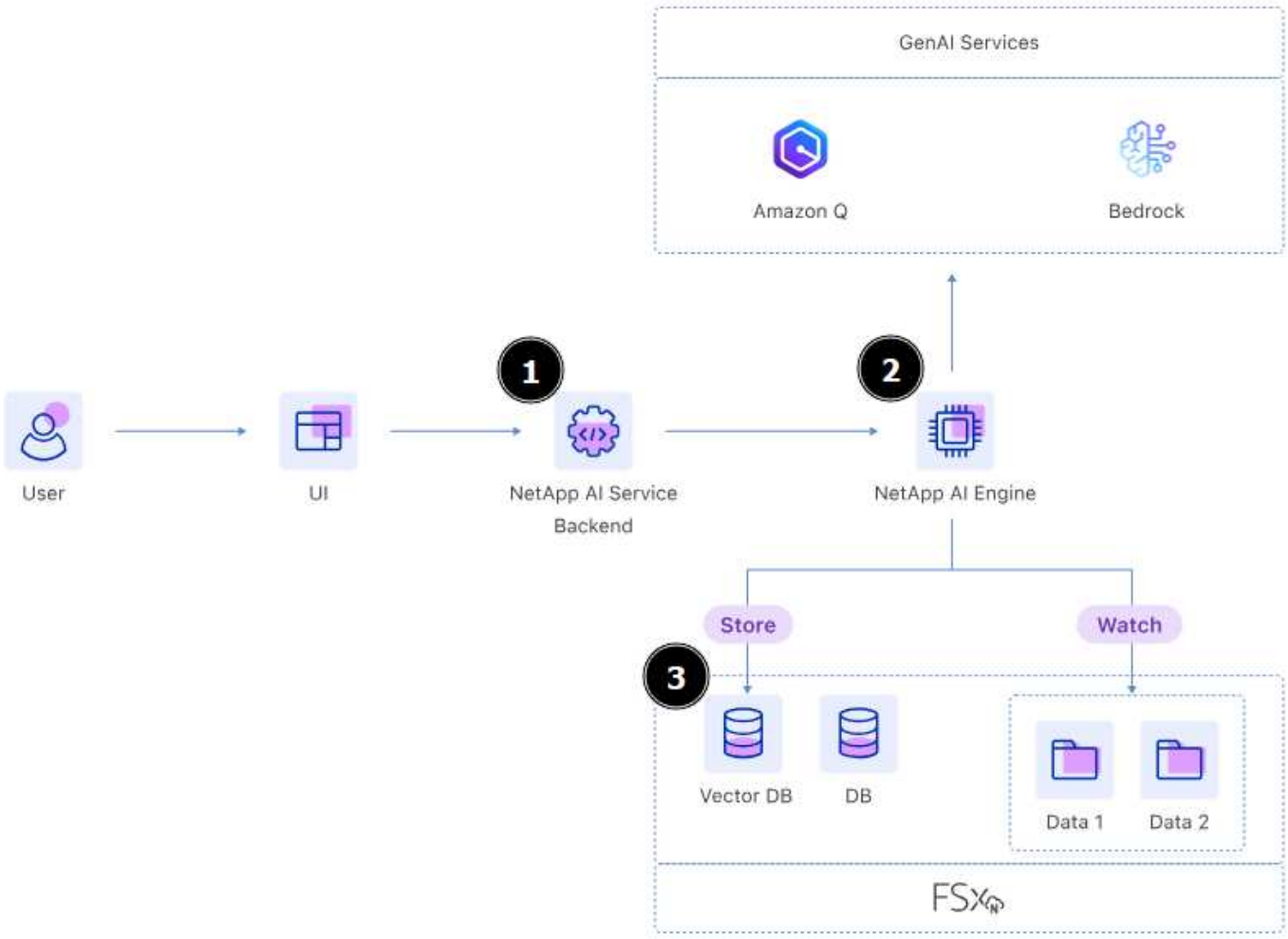
O sistema de arquivos FSX for ONTAP fornece o armazenamento para o seu sistema GenAI.

Um único volume é implantado que conterá o banco de dados vetorial que armazena os dados gerados pelo modelo básico com base em suas fontes de dados.

As fontes de dados que você integrará em sua base de conhecimento podem residir no mesmo sistema de arquivos FSX for ONTAP ou em um sistema diferente.

O mecanismo NetApp GenAI monitora e interage com ambos esses volumes.

A imagem a seguir mostra a infraestrutura do GenAI. Os componentes numerados 1, 2 e 3 são acionados durante este procedimento. Os outros elementos devem estar no lugar antes de iniciar a implantação.



Implantar a infraestrutura do GenAI

Você precisará inserir suas credenciais da AWS e selecionar o sistema de arquivos FSX for ONTAP para implantar a infraestrutura de geração de recuperação aumentada (RAG).

Antes de começar

Certifique-se de que seu ambiente atenda aos requisitos de bases de conhecimento ou conetores, dependendo do que você escolher, antes de iniciar este procedimento.

- "Requisitos da base de conhecimento"
- "Requisitos do conetor"

Passos

1. Faça login no Workload Factory usando um dos"experiências de console" .
2. No bloco cargas de trabalho de IA, selecione **Deploy & Manage**.
3. Revise o diagrama de infraestrutura e selecione **Next**.
4. Preencha os itens na seção **AWS settings**:
   a. **Credenciais da AWS**: Selecione ou adicione as credenciais da AWS que fornecem permissões para implantar os recursos da AWS.
   b. **Localização**: Selecione uma região, VPC e sub-rede da AWS.

A implantação do GenAI deve estar em uma região da AWS que tenha o Amazon bedrock habilitado. "Veja a lista de regiões suportadas"

5. Preencha os itens na seção **Configurações de infra-estrutura**:
   a. **Tags**: insira quaisquer pares de chave/valor de tag que você deseja aplicar a todos os recursos da AWS que fazem parte desta implantação. Essas tags são visíveis no AWS Management Console e na área de informações de infraestrutura do Workload Factory e podem ajudar você a controlar os recursos do Workload Factory.
6. Preencha a seção **conetividade**:
   a. **Par de chaves**: Selecione um par de chaves que permita que você se conete com segurança à instância do mecanismo NetApp GenAI.
7. Complete a seção **AI Engine**:
   a. **Nome da instância**: Opcionalmente, selecione **Definir nome da instância** e insira um nome personalizado para a instância do mecanismo de IA. O nome da instância aparece no AWS Management Console e na área de informações de infraestrutura do Workload Factory e pode ajudar você a controlar os recursos do Workload Factory.
8. Selecione **Deploy** para iniciar a implantação.

Se a implantação falhar com um erro de credenciais, você poderá obter mais detalhes de erro selecionando os hiperlinks dentro da mensagem de erro. Você pode ver uma lista de permissões ausentes ou bloqueadas, bem como uma lista de permissões que a carga de trabalho do GenAI precisa para que ela possa implantar a infraestrutura do GenAI.

Resultado

A Workload Factory começa a implantar a infraestrutura do chatbot. Esse processo pode levar até 10 minutos.

Durante o processo de implantação, os seguintes itens são configurados:

- A rede é configurada juntamente com os endpoints privados.
- A função do IAM, o perfil da instância e o grupo de segurança são criados.
- A instância de máquina virtual para o mecanismo GenAI é implantada.
- O Amazon bedrock está configurado para enviar logs para o Amazon CloudWatch Logs, usando um grupo de log com o prefixo `/aws/bedrock/`.
- O mecanismo GenAI está configurado para enviar logs para o Amazon CloudWatch Logs, usando um grupo de logs com o nome `/netapp/wlmai/<tenancyAccountId>/randomId` , onde `<tenancyAccountID>` é o "ID da conta do console NetApp" para o usuário atual.

Informações sobre direitos autorais

Informações sobre marcas comerciais