Documentação do NetApp Workload Factory para GenAI

GenAI

NetApp
October 06, 2025

Índice

Documentação do NetApp Workload Factory para GenAI

Notas de lançamento

Novidades do NetApp Workload Factory para GenAI

Saiba o que há de novo com o recurso de cargas de trabalho de IA generativa do Workload Factory.

05 de outubro de 2025

BlueXP workload factory agora NetApp Workload Factory

O BlueXP foi renomeado e redesenhado para refletir melhor o papel que ele tem no gerenciamento da sua infraestrutura de dados. Como resultado, a BlueXP workload factory foi renomeada para NetApp Workload Factory.

Suporte para adicionar fontes de dados NFS/SMB genéricas em conectores NetApp para Amazon Q Business

Usando a API do Workload Factory, agora você pode adicionar uma fonte de dados de um compartilhamento genérico NFSv3, NFSv4 ou SMB a um NetApp Connector para Amazon Q Business. Isso permite que você inclua arquivos armazenados em volumes hospedados por sistemas de arquivos diferentes do Amazon FSx for NetApp ONTAP.

"Crie um conector NetApp para o Amazon Q Business"

"Adicionar fontes de dados a um conector"

Configuração avançada de chat para bases de conhecimento

Agora você pode configurar definições avançadas de chat aplicáveis ao modelo de chat da base de conhecimento, como duração da resposta, temperatura, configurações de raciocínio e muito mais. Algumas dessas configurações, como configurações de tempo de modificação e atualidade, configurações de recuperação avançada e prompt do sistema, estão disponíveis somente usando a API do Workload Factory.

"Crie uma base de conhecimento do GenAI"

A seleção do tipo de inferência agora é compatível com modelos de incorporação, bate-papo e reclassificação

Se o modelo de incorporação, bate-papo ou reclassificação escolhido tiver configurações de inferência, agora você pode selecionar um tipo de inferência. Isso permite que você ajuste melhor o desempenho do chatbot e os requisitos de recursos às suas necessidades.

"Crie uma base de conhecimento do GenAI"

03 de agosto de 2025

Armazenamento seguro para resultados de dados estruturados

Se os resultados da consulta do chatbot contiverem dados estruturados, o GenAI poderá armazenar os resultados em um bucket do Amazon S3. Quando esses resultados são armazenados em um bucket S3, você pode baixá-los usando o link de download na sessão de bate-papo.

"Crie uma base de conhecimento do GenAI"

Disponibilidade do servidor MCP

A NetApp agora fornece um servidor Model Context Protocol (MCP) com o NetApp Workload Factory para GenAI. Você pode instalar o servidor localmente para permitir que clientes MCP externos descubram e recuperem resultados de consultas de uma base de conhecimento do GenAI.

"Servidor NetApp Workload Factory GenAI MCP"

29 de junho de 2025

Suporte para fontes de dados hospedadas em sistemas de arquivos NFS/SMB genéricos

Agora você pode adicionar uma fonte de dados de um compartilhamento SMB ou NFS genérico. Isso permite incluir arquivos armazenados em volumes hospedados por sistemas de arquivos diferentes do Amazon FSx para NetApp ONTAP.

"Adicionar fontes de dados a uma base de conhecimento"

"Adicionar fontes de dados a um conector"

03 de junho de 2025

Tracker disponível para operações de monitoramento e rastreamento

O recurso de monitoramento do Tracker agora está disponível no GenAI. Você pode usar o Tracker para monitorar e acompanhar o progresso e o status de operações pendentes, em andamento e concluídas, revisar detalhes de tarefas e subtarefas operacionais, diagnosticar problemas ou falhas, editar parâmetros de operações com falha e repetir operações com falha.

"Monitore as operações de carga de trabalho com o Tracker no NetApp Workload Factory"

Escolha um modelo de reranker para uma base de conhecimento

Agora você pode aumentar a relevância dos resultados de consultas reclassificadas selecionando um modelo de reclassificação específico para usar com uma base de conhecimento. O GenAI suporta os modelos Cohere Rerank e Amazon Rerank.

"Crie uma base de conhecimento do GenAI"

04 de maio de 2025

Suporte para NetApp Connector para Amazon Q Business

Esta versão do GenAI apresenta suporte ao NetApp Connector para Amazon Q Business, permitindo que você crie conectores para o Amazon Q Business. Aproveite de forma rápida e fácil o assistente de IA do Amazon Q Business com menos configuração inicial do que criar uma base de conhecimento do GenAI para a Amazon bedrock.

"Crie um conector NetApp para o Amazon Q Business"

Suporte aprimorado ao modelo de chat

O GenAI agora suporta os seguintes modelos de bate-papo adicionais para bases de conhecimento:

- "Modelos Mistral AI"
- "Modelos de texto Amazon Titan"
- "Modelos Meta Llama"
- "Jamba 1,5 modelos"
- "Modelos Cohere Command"
- "Modelos Deepseek"

O GenAI suporta os modelos de cada provedor que a Amazon bedrock suporta: "Modelos de base suportados na Amazon bedrock"

"Crie uma base de conhecimento do GenAI"

Terminologia de permissões atualizada

A interface do usuário e a documentação do Workload Factory agora usam "somente leitura" para se referir às permissões de leitura e "leitura/gravação" para se referir às permissões de automação.

02 de março de 2025

Aprimoramentos de chatbot incorporados

Agora você pode copiar perguntas e respostas diretamente para a área de transferência, ajustar o tamanho da janela de bate-papo e alterar seu título. Além disso, as respostas de bate-papo agora podem incluir tabelas, que também são copiáveis.

"Teste uma base de conhecimento do GenAI"

Suporte a citações de resposta de chat

As respostas de bate-papo agora incluem citações que listam os arquivos e blocos de dados que foram usados para gerar a resposta.

"Teste uma base de conhecimento do GenAI"

Suporte aprimorado ao tipo de arquivo

Esta versão do GenAI oferece suporte aprimorado a arquivos:

- Os modelos de chat apresentam suporte CSV melhorado. Isso permite respostas mais úteis ao consultar dados de arquivos CSV.
- O GenAI agora pode ingerir arquivos Apache Parquet a partir de fontes de dados.
- O GenAI agora suporta a introdução de arquivos DOCX do Microsoft Word que incluem imagens. As imagens incorporadas nos documentos DOCX são digitalizadas e os insights de texto das imagens incorporadas são incluídos nas respostas às consultas da base de conhecimento.

"Formatos de arquivo de origem de dados suportados"

02 de fevereiro de 2025

Suporte para modelos de fundação Amazon Nova

O GenAI agora suporta os modelos de fundação Amazon Nova. São suportados Amazon Nova Micro, Amazon Nova Lite e Amazon Nova Pro.

"Requisitos do GenAI"

Filtragem do tipo de arquivo para fontes de dados

O GenAI agora suporta a seleção de tipos de arquivo específicos para incluir na digitalização da fonte de dados quando você adiciona uma fonte de dados.

"Adicione fontes de dados à base de conhecimento"

Filtragem de data de modificação de arquivo para fontes de dados

O GenAI agora suporta arquivos de filtragem para incluir na verificação da fonte de dados por data de modificação quando você adiciona uma fonte de dados. Você pode escolher um intervalo de datas de modificação para os arquivos incluídos.

"Adicione fontes de dados à base de conhecimento"

Suporte para arquivos de imagem e suporte aprimorado para arquivos PDF

O GenAI agora oferece suporte ao aprimoramento de respostas a consultas de base de conhecimento com insights de imagens e descrições de gráficos, bem como texto de documentos, levando a respostas mais ricas e de maior qualidade. O GenAI agora pode digitalizar arquivos de imagem e imagens dentro de arquivos PDF (também conhecido como suporte a arquivos multimodais). Se você optar por digitalizar imagens ou arquivos PDF, o texto das imagens (incluindo imagens incorporadas em documentos PDF) é digitalizado na fonte de dados e os insights das digitalizações são incluídos nas respostas às consultas da base de conhecimento.

"Adicione fontes de dados à base de conhecimento"

Pesquisa híbrida e suporte de reclassificação

O GenAI agora pode aumentar significativamente a relevância e a precisão dos resultados de pesquisa usando a pesquisa híbrida e reclassificando os resultados. A pesquisa híbrida combina os pontos fortes da pesquisa tradicional baseada em palavras-chave com técnicas avançadas de pesquisa semântica baseadas em vetores densos. Os resultados padrão de pesquisa de palavras-chave são aumentados com correspondências próximas e nuance linguística, aumentando a relevância. Em seguida, o GenAI refina esses resultados ainda mais usando modelos avançados de reclassificação, como cohere Rerank e Amazon Rerank, e retorna os resultados mais relevantes. Esta capacidade está disponível para bases de conhecimento recém-criadas.

"Saiba mais sobre o NetApp Workload Factory para GenAI"

05 de janeiro de 2025

Nome do instantâneo personalizado

Agora você pode fornecer um nome de instantâneo para um instantâneo ad-hoc.

"Proteja uma base de conhecimento com snapshots"

Nome de instância personalizado do mecanismo de AI

Agora você pode dar um nome personalizado à instância do mecanismo de AI durante a implantação.

"Implantar a infraestrutura do GenAI"

Reconstruir a infraestrutura do GenAI corrompida ou ausente

Se a instância do seu mecanismo de IA for corrompida ou excluída de alguma forma, você pode deixar o Workload Factory reconstruí-la para você. O Workload Factory reconecta automaticamente suas bases de conhecimento à infraestrutura após a reconstrução ser concluída, para que elas estejam prontas para uso.

"Solução de problemas"

01 de dezembro de 2024

Clonar uma base de conhecimento a partir de um instantâneo

O NetApp Workload Factory para GenAI agora oferece suporte à clonagem de uma base de conhecimento a partir de um snapshot. Isso permite a recuperação rápida de bases de conhecimento e a criação de novas bases de conhecimento com fontes de dados existentes, além de ajudar na recuperação e no desenvolvimento de dados.

"Clonar uma base de conhecimento"

Detecção e replicação de clusters ONTAP no local

Descubra e replique dados de cluster ONTAP locais em um sistema de arquivos FSx for ONTAP para que eles possam ser usados para enriquecer bases de conhecimento de IA. Todos os fluxos de trabalho de descoberta e replicação no local são possíveis no novo menu **On-Premises ONTAP** no inventário de armazenamento.

"Descubra um cluster ONTAP no local"

3 de novembro de 2024

Mascarar informações pessoais identificáveis com guardrails de dados

A carga de trabalho de IA generativa apresenta o recurso de proteção de dados, alimentado pela classificação do NetApp Console. O recurso de proteção de dados identifica e mascara Informações Pessoais Identificáveis (PII), ajudando você a manter a conformidade e fortalecer a segurança dos seus dados organizacionais confidenciais.

"Crie uma base de conhecimento do GenAI"

"Saiba mais sobre a classificação do NetApp Console"

29 de setembro de 2024

Suporte a snapshot e restauração para volumes da base de conhecimento

Agora, você pode proteger seus dados generativos de workloads de AI usando uma cópia pontual de uma base de conhecimento. Isso permite que você proteja seus dados contra perda acidental ou alterações de

teste nas configurações da base de conhecimento. Você pode restaurar a versão anterior do volume da base de conhecimento a qualquer momento.

"Tire um instantâneo de um volume da base de conhecimento"

"Restaurar um snapshot de um volume da base de conhecimento"

Pausar digitalizações programadas

Agora você pode pausar digitalizações de origem de dados agendadas. Por padrão, workloads de AI generativo varre cada fonte de dados diariamente para ingerir novos dados em cada base de conhecimento. Se você não quiser que as alterações mais recentes sejam ingeridas (durante o teste ou durante a restauração de um instantâneo, por exemplo), você pode pausar as verificações agendadas e retomá-las a qualquer momento.

"Gerenciar bases de conhecimento"

Volumes de proteção de dados agora compatíveis com bases de conhecimento

Ao selecionar um volume da base de conhecimento, agora você pode escolher um volume de proteção de dados que faça parte de uma relação de replicação do NetApp SnapMirror. Isso permite armazenar bases de conhecimento em volumes que já estão protegidos pela replicação do SnapMirror.

"Identifique as fontes de dados a serem integradas em sua base de conhecimento"

1 de setembro de 2024

Estratégias adicionais de divisão

Agora, as cargas de trabalho generativas de IA oferecem suporte a volumes de várias frases e conjuntos baseados em sobreposição para fontes de dados.

Volume dedicado para cada base de conhecimento

Agora, cria um volume dedicado do Amazon FSX for NetApp ONTAP para cada nova base de conhecimento, permitindo políticas de snapshot individuais para cada base de conhecimento e proteção aprimorada contra falhas e envenenamento de dados.

4 de agosto de 2024

Integração do Amazon CloudWatch Logs

As cargas de trabalho generativas de IA agora estão integradas ao Amazon CloudWatch Logs, permitindo que você monitore arquivos de log de cargas de trabalho generativas de IA.

Exemplo de aplicativo chatbot

O aplicativo de exemplo NetApp Workload Factory GenAI permite que você teste a autenticação e a recuperação da sua base de conhecimento publicada do NetApp Workload Factory interagindo diretamente com ela em um aplicativo de chatbot baseado na Web.

7 de julho de 2024

Lançamento inicial do Workload Factory para GenAI

A versão inicial inclui a capacidade de desenvolver uma base de conhecimento que é personalizada incorporando os dados da sua organização. A base de conhecimento pode ser acessada por um aplicativo de chatbot para seus usuários. Essa capacidade garante respostas precisas e relevantes a perguntas específicas da organização, aumentando a satisfação e a produtividade de todos os seus usuários.

Saiba mais sobre o NetApp Workload Factory para GenAI

Saiba mais sobre o NetApp Workload Factory para GenAI

O NetApp Workload Factory para GenAI permite que você integre os sistemas de arquivos Amazon FSx for NetApp ONTAP com os modelos de base do GenAI. Isso fornece armazenamento de alto desempenho com um rico conjunto de recursos de proteção, segurança e otimização de custos para seus conjuntos de dados de IA.

O que é o NetApp Workload Factory para GenAI?

O NetApp Workload Factory para GenAI permite que você use suas fontes de dados empresariais no Amazon FSx for NetApp ONTAP com aplicativos Generative AI. Utilizando a geração aumentada de recuperação (RAG), você pode conectar rapidamente fontes de dados a modelos básicos disponíveis via Amazon Bedrock ou Amazon Q Business para desenvolver aplicativos com tecnologia de IA generativa, como assistentes virtuais, chatbots de perguntas e respostas, sumarização de documentos, criação de conteúdo, etc.

O uso de IA generativa com seus dados organizacionais permite que você aproveite seu próprio conhecimento e experiência, e não confie apenas na inteligência do modelo com base em dados públicos nos quais os modelos foram treinados. O uso do RAG para personalizar os modelos garante respostas precisas e relevantes a perguntas específicas da organização, aumentando a produtividade e a eficiência para os usuários de seus aplicativos usando IA generativa.

Desenvolver um aplicativo GenAI que é adaptado aos dados da sua organização permite que você aproveite seu próprio conhecimento e experiência. Essa capacidade de personalização garante respostas precisas e relevantes a perguntas específicas da organização, aumentando a satisfação e a produtividade de todos os seus usuários.

Se "crie uma base de conhecimento"você , o GenAI ingere dados de suas fontes de dados, armazena os resultados vetorizados em um banco de dados e oferece controle total sobre como usar os dados ingeridos para responder a consultas. Essa abordagem requer uma configuração mais inicial, mas permite que você escolha diferentes modelos de bate-papo para resultados diferentes. Se "definir um conector NetApp para Amazon Q Business"você , os dados de suas fontes de dados serão ingeridos pelo Amazon Q Business e armazenados em um índice. Essa abordagem requer menos configuração inicial, mas oferece menos controle sobre os resultados.

Para obter mais informações sobre o Workload Factory, consulte o "Visão geral da Workload Factory" .

Benefícios de usar o GenAI para criar aplicativos de IA generativos

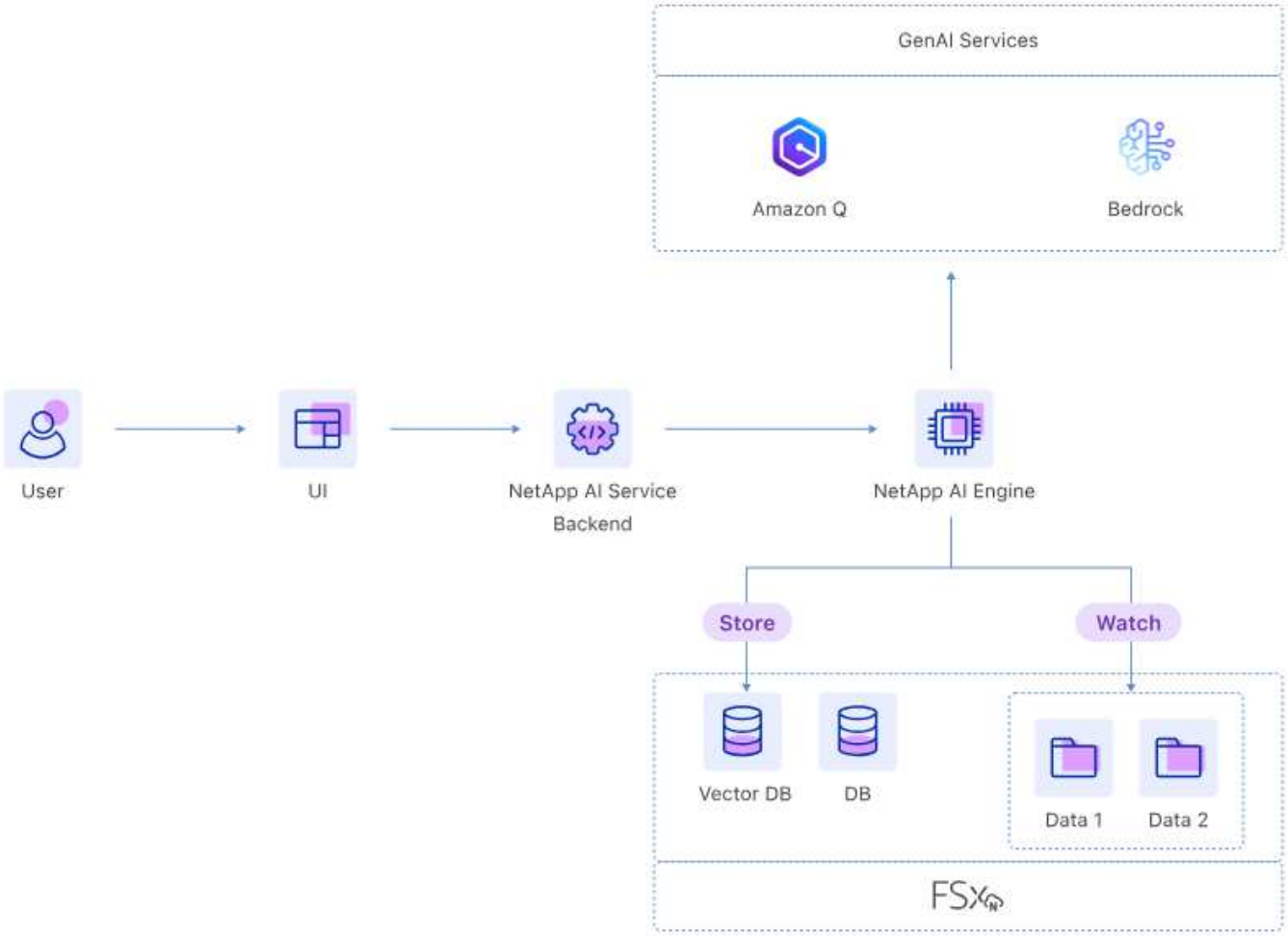
O NetApp Workload Factory para GenAI simplifica o processo de implantação da infraestrutura necessária para criar aplicativos de IA generativa usando geração aumentada de recuperação (RAG). Especificamente, o GenAI oferece os seguintes benefícios:

- Sem a necessidade de um conhecimento profundo da infraestrutura de dados, base e modelos de linguagem, os administradores e desenvolvedores DE TI podem acelerar o desenvolvimento de aplicações utilizando a automação fornecida pelo GenAI. Os administradores e desenvolvedores de dados podem criar bases de conhecimento empresariais de forma fácil e rápida que incorporam dados não estruturados da sua organização para serem usados por aplicativos de IA generativos.

- Melhore a segurança preservando as permissões do usuário em arquivos incorporados nas bases de conhecimento para garantir que a segurança e a privacidade dos dados sejam mantidas. Um aplicativo, como um chatbot, pode ser desenvolvido para fornecer apenas aos usuários autenticados respostas com base nos dados aos quais os usuários têm acesso.
- Mantenha seus dados empresariais privados e seguros na conta de cliente da AWS, onde seus dados organizacionais nunca serão expostos externamente.
- Acelere o desenvolvimento de aplicativos GenAI, como um chatbot de Q&A, usando frameworks de código aberto, como o LangChain, utilizando a API GenAI para provisionar e gerenciar bases de conhecimento e conetores, conversar com uma base de conhecimento e armazenar e recuperar histórico de bate-papo.
- Melhore a proteção e a disponibilidade de dados com a implantação da infraestrutura generativa de dados de AI nos sistemas de arquivos FSX para NetApp ONTAP e aproveite os recursos do ONTAP, como alta disponibilidade, snapshots para proteção e recuperação de dados locais, SnapMirror para recuperação de desastres e SnapVault para fazer backup da sua infraestrutura de dados.
- Reduza os custos gerais de storage para uma infraestrutura de dados generativa de AI aproveitando os recursos de eficiência de dados da ONTAP, como deduplicação, compressão e compactação de dados, disposição em camadas e thin Provisioning.
- Obtenha resultados de alta qualidade a partir dos seus dados com os recursos de pesquisa híbrida e classificação fornecidos pelo GenAI. A busca híbrida combinada com a reclassificação melhoram significativamente a relevância dos resultados da pesquisa. Esses recursos estão disponíveis por meio da Amazon AWS e dependem da região.

Como o GenAI funciona

O GenAI usa os dados privados da sua organização para complementar a inteligência do modelo (com base nos dados em que foi treinado) para fornecer respostas personalizadas às perguntas feitas pelos seus usuários em sua organização. Primeiro, você implanta a infraestrutura necessária para uma estrutura RAG, cria uma base de conhecimento ou define um conetor usando fontes de dados e modelos de base da sua organização disponíveis via Amazon bedrock ou Amazon Q Business e coneta um aplicativo (como um chatbot de perguntas e respostas) à base de conhecimento ou ao conetor.



Como o NetApp Workload Factory para GenAI ajuda a criar aplicativos de IA generativa

O GenAI ajuda a criar aplicativos de IA generativos usando o RAG das seguintes maneiras:

- Implanta a infraestrutura necessária para a estrutura de geração aumentada de recuperação (RAG) para trabalhar com fontes de dados no FSX para sistemas de arquivos ONTAP e Amazon bedrock ou Amazon Q Business. A infraestrutura inclui a instância do NetApp GenAI Engine para gerenciamento de dados, um banco de dados vetorial incorporado (LanceDB) e armazenamento no sistema de arquivos FSX for ONTAP para o banco de dados vetorial.
- Ajuda a conetar as fontes de dados a incorporações e modelos de linguagem disponíveis via Amazon bedrock ou Amazon Q Business para incorporar fontes de dados e recuperar as respostas para consultas de usuários. As fontes de dados, juntamente com os modelos e sua configuração, são apresentadas como bases de conhecimento do FSX for ONTAP.
- Ingere dados de origem na base de conhecimento ou conetor para incorporar arquivos de origem em compartilhamentos SMB e exportações de NFS nos sistemas de arquivos FSX for ONTAP, juntamente com o armazenamento de permissões de arquivos para arquivos em compartilhamentos SMB.
- Constrói automaticamente perguntas iniciais de conversação com base no conteúdo em bases de conhecimento.
- Fornece um simulador de bate-papo para os administradores de dados testar conversas com bases de conhecimento.
- Fornece uma interface de conetor simples para que você possa conetar o GenAI com o Amazon Q

Business, de forma rápida e fácil, utilizando os recursos deste assistente de IA.

Ferramentas para usar o NetApp Workload Factory

Você pode usar o NetApp Workload Factory com as seguintes ferramentas:

- **Console do Workload Factory**: O console do Workload Factory fornece uma visão visual e holística de seus aplicativos e projetos.
- * NetApp Console*: O NetApp Console oferece uma experiência de interface híbrida para que você possa usar o Workload Factory junto com outros serviços de dados do NetApp .
- **Pergunte-me**: use o assistente de IA Ask me para fazer perguntas e saber mais sobre o Workload Factory sem sair do console do Workload Factory. Acesse Pergunte-me no menu de ajuda do Workload Factory.
- **CloudShell CLI**: O Workload Factory inclui um CloudShell CLI para gerenciar e operar ambientes AWS e NetApp em todas as contas a partir de um único CLI baseado em navegador. Acesse o CloudShell na barra superior do console do Workload Factory.
- **API REST**: Use as APIs REST do Workload Factory para implantar e gerenciar seu FSx para sistemas de arquivos ONTAP e outros recursos da AWS.
- **CloudFormation**: use o código do AWS CloudFormation para executar as ações definidas no console do Workload Factory para modelar, provisionar e gerenciar recursos da AWS e de terceiros da pilha do CloudFormation na sua conta da AWS.
- **Provedor do Terraform NetApp Workload Factory**: use o Terraform para criar e gerenciar fluxos de trabalho de infraestrutura gerados no console do Workload Factory.

Custo

Não há custo para usar o recurso GenAI do Workload Factory.

No entanto, você precisará pagar pelos recursos da AWS que implantar para oferecer suporte à infraestrutura de IA generativa. Por exemplo, você pagará a AWS pela Amazon bedrock ou Amazon Q Business, o FSX for ONTAP file system e a capacidade de armazenamento e a instância do GenAI Engine EC2.

Algumas operações multimodais, como a digitalização de imagens para informações de texto, podem usar mais recursos e, portanto, incorrer em um custo mais alto. Algumas operações de configuração, como alterar as configurações de uma base de conhecimento, podem fazer com que as fontes de dados sejam digitalizadas novamente, e as verificações de origem de dados também podem incorrer em um custo mais alto.

Licenciamento

Não são necessárias licenças especiais da NetApp para usar os recursos de IA do Workload Factory.

Regiões

O Workload Factory é suportado em todas as regiões comerciais onde o FSx for ONTAP é suportado."Veja as regiões da Amazon suportadas."

As seguintes regiões da AWS não são suportadas:

- Regiões da China
- Regiões GovCloud (EUA)

- Nuvem Secreta
- Nuvem Top Secret

Componentes do motor NetApp GenAI

Ao implantar a infraestrutura GenAI, o Workload Factory cria uma instância EC2 para o mecanismo GenAI. Ele também cria uma função do IAM, um grupo de segurança e endpoints privados para esta instância. Talvez você queira entender mais detalhes sobre esses componentes que o Workload Factory cria no seu ambiente AWS.

Tipo de instância EC2

m5.large

Função do IAM

A instância do mecanismo GenAI precisa de permissões para enviar partes de dados para o modelo de incorporação na Amazon bedrock e para se comunicar com o back-end do Serviço de IA da NetApp. A função do IAM inclui as seguintes permissões:

Permissões de função do IAM

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Action": [
        "iam:CreateRole",
        "iam:CreatePolicy",
        "iam:AttachRolePolicy",
        "iam:PassRole"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "ssm:DescribeDocument",
        "ssm:DescribeAssociation",
        "ssm:GetDeployablePatchSnapshotForInstance",
        "ssm:GetManifest",
        "ssm:ListInstanceAssociations",
        "ssm:ListAssociations",
        "ssm:PutInventory",
        "ssm:PutComplianceItems",
        "ssm:PutConfigurePackageResult",
        "ssm:UpdateAssociationStatus",
        "ssm:UpdateInstanceAssociationStatus",
        "ssm:UpdateInstanceInformation",
        "ssmmessages:CreateControlChannel",
        "ssmmessages:CreateDataChannel",
        "ssmmessages:OpenControlChannel",
        "ssmmessages:OpenDataChannel"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "ssm:GetParameter"
      ],
      "Resource": "arn:aws:ssm:*:*:parameter/netapp/wlmai/*",
      "Effect": "Allow"
    },
    {
      "Action": [
```

```
            "fsx:DescribeVolumes",
            "fsx:DescribeStorageVirtualMachines",
            "fsx:DescribeFileSystems"
        ],
        "Resource": "*",
        "Effect": "Allow"
    },
    {
        "Action": [
            "fsx:TagResource",
            "fsx:ListTagsForResource"
        ],
        "Resource": [
            "arn:aws:fsx:*:*:storage-virtual-machine/*/*",
            "arn:aws:fsx:*:*:volume/*/*"
        ],
        "Effect": "Allow"
    },
    {
        "Action": [
            "fsx:CreateVolume"
        ],
        "Resource": [
            "arn:aws:fsx:*:*:volume/*/*",
            "arn:aws:fsx:*:*:storage-virtual-machine/*/*"
        ],
        "Effect": "Allow"
    },
    {
        "Condition": {
            "StringLike": {
                "aws:ResourceTag/netapp:wlmai:<id>:kbId": "*"
            }
        },
        "Action": "fsx:DeleteVolume",
        "Resource": [
            "arn:aws:fsx:*:*:volume/*/*",
            "arn:aws:fsx:*:*:backup/*"
        ],
        "Effect": "Allow"
    },
    {
        "Condition": {
            "StringLike": {
                "aws:ResourceTag/netapp:wlmai:<id>:qConnectorId": "*"
            }
```

```
    },
    "Action": "fsx:DeleteVolume",
    "Resource": [
      "arn:aws:fsx:*:*:volume/*/*",
      "arn:aws:fsx:*:*:backup/*"
    ],
    "Effect": "Allow"
  },
  {
    "Condition": {
      "StringLike": {
        "aws:ResourceTag/netapp:wlmai:<id>": "*"
      }
    },
    "Action": "fsx:UntagResource",
    "Resource": "arn:aws:fsx:*:*:storage-virtual-machine/*/*",
    "Effect": "Allow"
  },
  {
    "Condition": {
      "StringLike": {
        "aws:ResourceTag/netapp:wlmai:<id>:kbId": "*"
      }
    },
    "Action": "fsx:UntagResource",
    "Resource": "arn:aws:fsx:*:*:volume/*/*",
    "Effect": "Allow"
  },
  {
    "Condition": {
      "StringLike": {
        "aws:ResourceTag/netapp:wlmai:<id>:qConnectorId": "*"
      }
    },
    "Action": "fsx:UntagResource",
    "Resource": "arn:aws:fsx:*:*:volume/*/*",
    "Effect": "Allow"
  },
  {
    "Action": [
      "bedrock:InvokeModel",
      "bedrock:Rerank",
      "bedrock:GetFoundationModel",
      "bedrock:GetInferenceProfile",
      "bedrock:GetModelInvocationLoggingConfiguration",
      "bedrock:PutModelInvocationLoggingConfiguration"
```

```
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "ec2messages:GetMessages",
        "ec2messages:GetEndpoint",
        "ec2messages:AcknowledgeMessage",
        "ec2messages:DeleteMessage",
        "ec2messages:FailMessage",
        "ec2messages:SendReply"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "qbusiness:ListWebExperiences",
        "qbusiness:ListApplications",
        "qbusiness:GetApplication",
        "qbusiness:CreateDataSource",
        "qbusiness:DeleteDataSource",
        "qbusiness:ListIndices",
        "qbusiness:StartDataSourceSyncJob",
        "qbusiness:StopDataSourceSyncJob",
        "qbusiness:ListDataSourceSyncJobs",
        "qbusiness:BatchPutDocument",
        "qbusiness:BatchDeleteDocument"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "logs:DescribeLogGroups"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "logs:DescribeLogStreams",
        "logs:PutLogEvents",
        "logs:CreateLogStream",
        "logs:CreateLogGroup"
```

```
        ],
        "Resource": [
          "arn:aws:logs:*:*:log-group:/aws/bedrock*",
          "arn:aws:logs:*:*:log-group:/netapp/wlmai/*:log-stream:*",
          "arn:aws:logs:*:*:log-group:/netapp/wlmai/*"
        ],
        "Effect": "Allow"
      },
      {
        "Action": [
          "s3:GetObject",
          "s3:PutObject"
        ],
        "Resource": "*",
        "Effect": "Allow"
      },
      {
        "Action": [
          "kms:Decrypt",
          "kms:GenerateDataKey"
        ],
        "Resource": "*",
        "Effect": "Allow"
      }
    ]
}
```

Grupo de segurança

As regras de saída estão abertas a todo o tráfego, enquanto as regras de entrada são completamente fechadas.

Endpoints privados

Se a VPC de destino ainda não os tiver, o Workload Factory criará endpoints privados para a instância EC2 do mecanismo GenAI para que ele possa se comunicar com os seguintes serviços da AWS:

- Amazon bedrock
  - bedrock
  - bedrock-runtime
  - bedrock-agent-runtime
- Amazon Elastic Container Registry (ECR)
  - API
  - docker
- AWS Systems Manager (SSM)
  - ssm

    - ec2messages
    - mensagens de smsm
- Amazon FSX para NetApp ONTAP
- Amazon CloudWatch

Use o GenAI para criar bases de conhecimento para a Amazon bedrock

Comece agora

Início rápido para bases de conhecimento GenAI

Comece a criar uma base de conhecimento ou o Amazon Q Business Connector usando os dados da sua organização que existem nos sistemas de arquivos do Amazon FSX para NetApp ONTAP. Um aplicativo como um chatbot acessará essa base de conhecimento ou conetor para fornecer respostas focadas na organização aos usuários finais.

Efetue login no Workload Factory

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

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 em sua base de conhecimento ou conetor, acesso ao serviço de AI da Amazon bedrock ou ao aplicativo Amazon Q Business e muito mais.

"Saiba mais sobre os requisitos do GenAI".

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

As fontes de dados que você integrará em sua base de conhecimento 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".

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 usado para armazenar o banco de dados de vetor usado pela base de conhecimento.

"Saiba como implantar a infraestrutura da base de conhecimento".

O que vem a seguir

Agora você pode criar uma base de conhecimento para fornecer respostas focadas na organização aos usuários finais.

Requisitos da base de conhecimento GenAI

Certifique-se de que o Workload Factory e o AWS estejam configurados corretamente antes de criar sua base de conhecimento. Isso inclui ter suas credenciais de login da AWS, um sistema de arquivos FSx for ONTAP implantado que contém as fontes de dados que você deseja integrar à sua base de conhecimento, acesso ao serviço Amazon Bedrock AI e muito mais.

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 do modo *Basic* e do modo *Read-Only* não são suportadas neste 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 da base de conhecimento GenAI

Se você planeja trabalhar com bases de conhecimento, certifique-se de que seu ambiente atenda aos seguintes requisitos.

Amazon bedrock

O Amazon bedrock permite que você use modelos de base e fornece os recursos para criar aplicativos de IA generativos.

Antes de começar a usar o NetApp Workload Factory para GenAI, você precisa configurar o Amazon Bedrock. Sua implantação do GenAI deve estar em uma região da AWS que tenha o Amazon Bedrock habilitado.

- "Documentação da AWS: Configurar o Amazon bedrock"
- "Documentação da AWS: Regiões e modelos compatíveis para bases de conhecimento da Amazon bedrock"

O GenAI reclassifica os resultados da pesquisa por padrão para melhorar a relevância do resultado. Para obter os melhores resultados, certifique-se de que a configuração do modelo de base do Amazon bedrock inclua acesso a um modelo de reclassificação, como cohere Rerank ou Amazon Rerank, se disponível em sua região.

Modelo de incorporação

Você deve habilitar o modelo de incorporação que você planeja usar antes de criar sua base de conhecimento. Os seguintes modelos de incorporação são suportados:

- Titãs incorporações G1 - texto
- Texto de incorporação Titan v2
- Incorporações multimodais Titan G1
- Incorpore o inglês
- Incorpore multilíngue

  "Saiba mais sobre o Amazon Titan"

Modelo de chat

Você deve habilitar o modelo básico de bate-papo que planeja usar antes de criar sua base de conhecimento. Como o suporte ao modelo varia de acordo com a região da AWS, "A documentação da AWS" consulte para verificar quais modelos você pode usar nas regiões em que planeja implantar sua base de conhecimento.

O GenAI suporta vários modelos de Antropometria, Amazon, Mistral AI, Meta, Jamba e cohere.

Saiba mais sobre como usar esses modelos no Amazon bedrock:

- "Claude de antrópico em Amazon bedrock"
- "Introdução ao Amazon Nova no console Amazon bedrock"
- "Modelos Mistral AI"
- "Modelos de texto Amazon Titan"
- "Modelos Meta Llama"
- "Jamba modelos"
- "Modelos Cohere Command"

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 o banco de dados de vetor usado pela base de dados de conhecimento.

  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 integrando em sua base de conhecimento.

  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. O sistema de arquivos deve estar em uma região da AWS que tenha o Amazon bedrock habilitado.
- 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 uma base de conhecimento ou conetor

Identifique ou crie os documentos (fontes de dados) que residem no seu sistema de arquivos do FSX for ONTAP que você integrará em sua base de conhecimento. Essas fontes de dados permitem que a base de conhecimento 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 base de conhecimento, 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 bancos de conhecimento 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 a base de conhecimento 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 sua base de conhecimento, e 2 dos arquivos são arquivos de recursos humanos que contêm informações restritas, apenas usuários de chatbot autenticados para acessar esses arquivos 2 receberão respostas do chatbot que incluem dados desses arquivos.

Formatos de arquivo de origem de dados suportados

Os seguintes formatos de arquivo de fonte de dados são atualmente suportados pelas bases de conhecimento do Workload Factory GenAI.

Formato do ficheiro	Extensão
Apache Parquet [1]	.parquet
Valores separados por vírgulas file[1]	.csv
Formato de intercâmbio de gráficos	.gif
JPEG	.jpg or.jpeg
JSON e JSONPnote:aviso de isenção de responsabilidade[]	.json
Markdown	.md
Microsoft Word	.doc ou .docx
Texto simples	.txt
Formato de documento portátil	.pdf
Gráficos de rede portáteis	.png
Imagem WebP	.webp

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

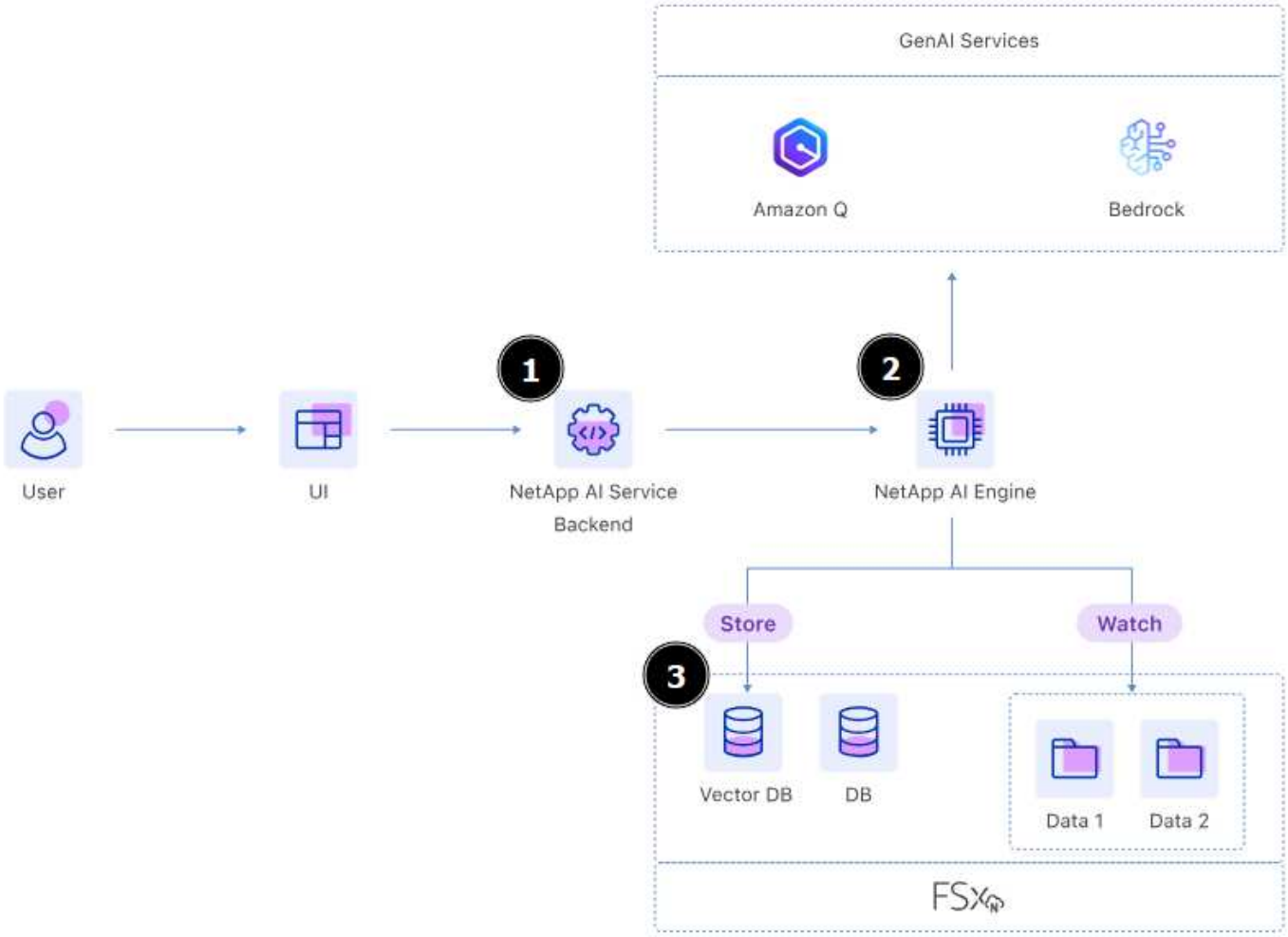
O sistema de arquivos FSX for ONTAP fornece o armazenamento para o seu sistema GenAI.

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.

Crie uma base de conhecimento do GenAI

Depois de implantar a infraestrutura de IA e identificar as fontes de dados que você integrará em sua base de conhecimento a partir dos seus armazenamentos de dados do FSx for ONTAP , você estará pronto para criar a base de conhecimento usando o Workload Factory. Como parte desta etapa, você também definirá as características da IA e criará iniciadores de conversas.

Certifique-se de que seu ambiente atenda "requisitos"às bases de conhecimento para antes de prosseguir.

Sobre esta tarefa

As bases de conhecimento têm duas modalidades de integração de dados - *public mode* e *Enterprise mode*.

Modo público

Uma base de conhecimento pode ser usada sem integrar fontes de dados da sua organização. Neste caso, um aplicativo integrado à base de conhecimento só fornecerá resultados de informações publicamente disponíveis na internet. Isso é conhecido como integração *public mode*.

Modo empresarial

Na maioria dos casos, você vai querer integrar fontes de dados da sua organização na base de conhecimento. Isso é conhecido como integração *Enterprise mode* porque fornece conhecimento da sua empresa.

Fontes de dados da sua organização podem conter Informações de Identificação Pessoal (PII). Para proteger essas informações confidenciais, você pode habilitar *guardrails de dados* ao criar e configurar bases de conhecimento. Os guardrails de dados, alimentados pela NetApp Data Classification, identificam e mascaram PII, tornando-os inacessíveis e irrecuperáveis.

"Saiba mais sobre a classificação de dados da NetApp" .

O NetApp Workload Factory para GenAI não mascara informações pessoais confidenciais (SPii). Consulte"tipos de dados pessoais sensíveis" para mais informações sobre esse tipo de dados.

Os guardrails de dados podem ser ativados ou desativados a qualquer momento. Se você alternar a ativação dos guardrails de dados, o Workload Factory verificará toda a base de conhecimento do zero, o que gerará um custo.

Crie e configure a base de conhecimento

A base de conhecimento define caraterísticas como os modelos de IA bedrock e o formato de incorporação que você deseja usar para criar sua base de conhecimento.

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. No menu Bases de conhecimento e conectores, selecione o menu suspenso **Criar novo** e escolha **Base de conhecimento NetApp GenAI para Bedrock**.
4. Na página Criar base de conhecimento do NetApp GenAI, configure as definições da base de conhecimento:

Detalhes da base de conhecimento

1. **Nome**: Insira o nome que deseja usar para a base de conhecimento.
2. **Descrição**: Insira uma descrição detalhada para a base de conhecimento.
3. **Bedrock**: Escolha a região onde o Amazon Bedrock está disponível para sua conta da AWS.

Ingestão

1. **Modelo de incorporação**:
    - Escolha um modelo de incorporação para usar na base de conhecimento. O modelo de incorporação define como seus dados serão convertidos em incorporações vetoriais para a base de conhecimento. O Workload Factory oferece suporte aos seguintes modelos:
    - Titãs incorporações G1 - texto
    - Texto de incorporação Titan v2
    - Incorporações multimodais Titan G1
    - Incorpore o inglês
    - Incorpore multilíngue

      Observe que você já deve ter habilitado o modelo de incorporação da Amazon bedrock.

    "Saiba mais sobre o Amazon Titan"

    - Se aplicável, selecione o tipo de inferência que corresponde à configuração do modelo de incorporação selecionado.
2. **Guardrails de dados**: escolha se deseja habilitar ou desabilitar os guardrails de dados. "Saiba mais sobre guardrails de dados, com tecnologia da NetApp Data Classification" .

   Os pré-requisitos a seguir devem ser atendidos para habilitar os corrimões de dados.

    - Uma conta de serviço é necessária para se comunicar com o NetApp Data Classification. Você deve ter a função *Administrador da organização* na sua conta de locação do NetApp Console para criar uma conta de serviço. Um membro que tem a função de administrador da organização pode concluir todas as ações no ."Aprenda como adicionar uma função a um membro no NetApp Console"
    - O mecanismo de IA deve ter acesso ao"Ponto de extremidade da API do console NetApp" .
    - Você precisará fazer o seguinte conforme descrito em"Documentação de classificação de dados da NetApp" :
        i. Criar um agente de console
        ii. Certifique-se de que seu ambiente atenda aos pré-requisitos

iii. Implantar classificação de dados do NetApp

> O recurso de guardrails de dados não é suportado ao inserir arquivos de dados estruturados, como CSV, JSON, JSONP ou Parquet.

Configurações de bate-papo e recuperação

1. **Modelo de bate-papo**:
   - Escolha entre vários modelos de bate-papo integrados ao Amazon Bedrock. Observe que você já deve ter habilitado o modelo de bate-papo do Amazon Bedrock.
   - Se aplicável, selecione o tipo de inferência que corresponde à configuração do modelo selecionado.
2. **Configurações de bate-papo**:
   - Escolha uma temperatura para o chatbot para configurar a aleatoriedade e a criatividade das respostas. Uma temperatura mais baixa resulta em respostas mais previsíveis, e uma temperatura mais alta resulta em respostas mais variadas.
   - Escolha um comprimento máximo de resposta para configurar o quão detalhadas as respostas serão. Respostas mais longas usam mais tokens de resposta e podem gerar custos mais altos.
3. **Modo de pensamento**: Quando o modo de pensamento está ativado, o chatbot levará mais tempo para processar as consultas e os resultados geralmente serão mais precisos. Ao ativar o modo de raciocínio, você pode controlar quantos tokens de raciocínio são usados ao gerar resultados. Usar mais fichas de raciocínio pode levar a respostas mais precisas, mas pode incorrer em um custo mais alto.
4. **Reclassificação**: habilite ou desabilite a reclassificação, o que pode melhorar a relevância e a qualidade dos resultados da consulta. Escolha um modelo de bate-papo padrão ou um modelo de reclassificação especializado para usar na reclassificação. As opções do modelo Reranker só serão exibidas se estiverem disponíveis na sua região. Selecione o tipo de inferência que corresponde à configuração do modelo selecionado.
5. **Entradas de conversação**: Escolha se você deseja fornecer até quatro prompts iniciais de conversação que são exibidos para usuários que interagem com um chatbot que usa essa base de conhecimento. Recomendamos que ative esta definição.

   Se você ativar os iniciantes de conversação, "modo automático" é selecionado por padrão. O "modo manual" só pode ser ativado depois de adicionar fontes de dados à sua base de conhecimento. "Saiba como modificar as configurações da base de conhecimento".

Definições de armazenamento

1. *Sistema de arquivos FSx para ONTAP *: quando você define uma nova base de conhecimento, o Workload Factory cria um novo volume Amazon FSx for NetApp ONTAP para armazená-la. Escolha um nome de sistema de arquivos existente e SVM (também chamado de VM de armazenamento) onde o novo volume será criado.
2. **Política de snapshot**: escolha uma política de snapshot na lista de políticas existentes definidas no inventário de armazenamento do Workload Factory. Snapshots recorrentes da base de conhecimento serão criados automaticamente com uma frequência baseada na política de snapshot selecionada.
3. **S3 Bucket**: Se os resultados da consulta do chatbot contiverem dados estruturados, o GenAI pode armazenar os resultados em um bucket S3. Para usar esse recurso, ative a configuração **Ativar bucket S3** e escolha um bucket S3 associado à sua conta na lista. Quando esses resultados são armazenados em um bucket S3, você pode baixá-los usando o link de download na sessão de bate-papo.

   Se a política de snapshot que você precisa não existir, você pode "criar uma política de snapshot" na VM

de armazenamento que contém o volume.

4. Selecione **criar base de conhecimento** para adicionar a base de conhecimento ao GenAI.

   Um indicador de progresso é exibido enquanto a base de conhecimento é criada.

   Depois que a base de conhecimento é criada, você tem a opção de adicionar uma fonte de dados à sua nova base de conhecimento ou terminar o processo sem adicionar uma fonte de dados. Recomendamos que você selecione **Adicionar fonte de dados** e adicione uma ou mais fontes de dados agora.

Adicione fontes de dados à base de conhecimento

Você pode adicionar uma ou mais fontes de dados para preencher a base de conhecimento com os dados da sua organização.

Sobre esta tarefa

O número máximo de fontes de dados suportadas é 10.

Passos

1. Depois de selecionar **Adicionar fonte de dados**, selecione o tipo de fonte de dados que deseja adicionar:
   - Adicionar FSx para sistema de arquivos ONTAP (usar arquivos de um volume FSx para ONTAP existente)
   - Adicionar sistema de arquivos (usar arquivos de um compartilhamento SMB ou NFS genérico)

Adicionar um FSx para sistema de arquivos ONTAP

1. * Selecione um sistema de arquivos*: Selecione o sistema de arquivos FSX for ONTAP onde seus arquivos de origem de dados residem e selecione **Next**.
2. **Selecione um volume**: Selecione o volume no qual os arquivos de origem de dados residem e selecione **Next**.

   Ao selecionar arquivos armazenados usando o protocolo SMB, você precisará inserir as informações do ative Directory, que incluem o domínio, o endereço IP, o nome de usuário e a senha.

3. **Selecione uma fonte de dados**: Selecione a localização da fonte de dados com base no local onde você salvou os arquivos. Este pode ser um volume inteiro, ou apenas uma pasta específica ou subpasta no volume, e selecione **Next**.
4. * Configurações*: Configure como a fonte de dados ingere informações de seus arquivos e quais arquivos ela inclui em varreduras:

   - **Definir fonte de dados**: Na seção **Estratégia de Chunking**, defina como o mecanismo GenAI divide o conteúdo da fonte de dados em blocos quando a fonte de dados é integrada a uma base de conhecimento. Você pode escolher uma das seguintes estratégias:
     - * Agrupamento de frases múltiplas*: Organiza informações de sua fonte de dados em blocos definidos por sentença. Você pode escolher quantas frases compõem cada pedaço (até 100).
     - * Agrupamento baseado em sobreposição*: Organiza informações de sua fonte de dados em blocos definidos por carateres que podem sobrepor blocos vizinhos. Você pode escolher o tamanho de cada pedaço em carateres, e quanto cada pedaço se sobrepõe com pedaços adjacentes. Você pode configurar um tamanho de bloco entre 50 e 3000 carateres e uma porcentagem de sobreposição entre 1 e 99%.

       Escolher uma alta porcentagem de sobreposição pode aumentar significativamente os requisitos de armazenamento com apenas pequenas melhorias na precisão de recuperação.

   - * Filtragem de arquivos*: Configure quais arquivos estão incluídos nas digitalizações:
     - Na seção **suporte a tipos de arquivo**, escolha incluir todos os tipos de arquivos ou selecionar tipos de arquivo individuais para inclusão nas verificações de origem de dados.

       Se você incluir imagens ou arquivos PDF, o NetApp Workload Factory for GenAI analisará o texto nas imagens (incluindo imagens em documentos PDF), e isso incorrerá em um custo mais alto.

     Ao incluir dados de texto de imagens, o GenAI não consegue mascarar informações de identificação pessoal (PII) da imagem à medida que os dados de texto digitalizados são enviados do seu ambiente para a AWS. No entanto, uma vez que os dados são armazenados, todas as PII são mascaradas no banco de dados do GenAI.

     Sua escolha de incluir arquivos de imagem em digitalizações está relacionada ao modelo de bate-papo da base de conhecimento. Se você incluir arquivos de imagem em digitalizações, o modelo de bate-papo deve suportar imagens. Se os tipos de arquivo de imagem estiverem selecionados aqui, você não poderá alternar a base de conhecimento para um modelo de chat que não suporte arquivos de imagem.

   - Na seção **filtro de tempo de modificação de arquivo**, escolha ativar ou desativar a inclusão de

arquivos com base em seu tempo de modificação. Se ativar a filtragem de hora de modificação, selecione um intervalo de datas na lista.

> Se você incluir arquivos com base em um intervalo de datas de modificação, assim que o intervalo de datas não for satisfeito (os arquivos não foram modificados dentro do intervalo de datas especificado), os arquivos serão excluídos da verificação periódica e a fonte de dados não incluirá esses arquivos.

5. Na seção **reconhecimento de permissão**, que está disponível somente quando a fonte de dados selecionada estiver em um volume que usa o protocolo SMB, você pode ativar ou desativar respostas com reconhecimento de permissão:
   - **Habilitado**: Os usuários do chatbot que acessam essa base de conhecimento só receberão respostas a consultas de fontes de dados às quais têm acesso.
   - **Disabled**: Os usuários do chatbot receberão respostas usando conteúdo de todas as fontes de dados integradas.
6. Selecione **Add** para adicionar esta fonte de dados à sua base de conhecimento.

Adicionar um sistema de arquivos NFS genérico

1. **Selecione um sistema de arquivos**: insira o endereço IP ou FQDN do host do sistema de arquivos onde seus arquivos de fonte de dados residem, escolha o protocolo NFS para o compartilhamento de rede e selecione **Avançar**.
2. **Selecione uma fonte de dados**: Selecione a localização da fonte de dados com base no local onde você salvou os arquivos. Este pode ser um volume inteiro, ou apenas uma pasta específica ou subpasta no volume, e selecione **Next**.

> Em alguns casos, pode ser necessário inserir o nome da exportação NFS manualmente e selecionar **Recuperar diretórios** para exibir os diretórios disponíveis. Você pode optar por selecionar a exportação inteira ou apenas pastas específicas da exportação.

3. * Configurações*: Configure como a fonte de dados ingere informações de seus arquivos e quais arquivos ela inclui em varreduras:
   - **Definir fonte de dados**: Na seção **Estratégia de Chunking**, defina como o mecanismo GenAI divide o conteúdo da fonte de dados em blocos quando a fonte de dados é integrada a uma base de conhecimento. Você pode escolher uma das seguintes estratégias:
     - * Agrupamento de frases múltiplas*: Organiza informações de sua fonte de dados em blocos definidos por sentença. Você pode escolher quantas frases compõem cada pedaço (até 100).
     - * Agrupamento baseado em sobreposição*: Organiza informações de sua fonte de dados em blocos definidos por carateres que podem sobrepor blocos vizinhos. Você pode escolher o tamanho de cada pedaço em carateres, e quanto cada pedaço se sobrepõe com pedaços adjacentes. Você pode configurar um tamanho de bloco entre 50 e 3000 carateres e uma porcentagem de sobreposição entre 1 e 99%.

> Escolher uma alta porcentagem de sobreposição pode aumentar significativamente os requisitos de armazenamento com apenas pequenas melhorias na precisão de recuperação.

   - * Filtragem de arquivos*: Configure quais arquivos estão incluídos nas digitalizações:

- Na seção **suporte a tipos de arquivo**, escolha incluir todos os tipos de arquivos ou selecionar tipos de arquivo individuais para inclusão nas verificações de origem de dados.

  Se você incluir imagens ou arquivos PDF, o NetApp Workload Factory for GenAI analisará o texto nas imagens (incluindo imagens em documentos PDF), e isso incorrerá em um custo mais alto.

Ao incluir dados de texto de imagens, o GenAI não consegue mascarar informações de identificação pessoal (PII) da imagem à medida que os dados de texto digitalizados são enviados do seu ambiente para a AWS. No entanto, uma vez que os dados são armazenados, todas as PII são mascaradas no banco de dados do GenAI.

Sua escolha de incluir arquivos de imagem em digitalizações está relacionada ao modelo de bate-papo da base de conhecimento. Se você incluir arquivos de imagem em digitalizações, o modelo de bate-papo deve suportar imagens. Se os tipos de arquivo de imagem estiverem selecionados aqui, você não poderá alternar a base de conhecimento para um modelo de chat que não suporte arquivos de imagem.

- Na seção **filtro de tempo de modificação de arquivo**, escolha ativar ou desativar a inclusão de arquivos com base em seu tempo de modificação. Se ativar a filtragem de hora de modificação, selecione um intervalo de datas na lista.

Se você incluir arquivos com base em um intervalo de datas de modificação, assim que o intervalo de datas não for satisfeito (os arquivos não foram modificados dentro do intervalo de datas especificado), os arquivos serão excluídos da verificação periódica e a fonte de dados não incluirá esses arquivos.

4. Selecione **Adicionar fonte de dados** para adicionar esta fonte de dados à sua base de conhecimento.

Adicionar um sistema de arquivos SMB genérico

1. **Selecione o sistema de arquivos**:
   a. Digite o endereço IP ou FQDN do host do sistema de arquivos onde seus arquivos de fonte de dados residem.
   b. Escolha o protocolo SMB para o compartilhamento de rede.
   c. Insira as informações do Active Directory, que incluem o domínio, endereço IP, nome de usuário e senha.
   d. Selecione **seguinte**.
2. **Selecione uma fonte de dados**: Selecione a localização da fonte de dados com base no local onde você salvou os arquivos. Este pode ser um volume inteiro, ou apenas uma pasta específica ou subpasta no volume, e selecione **Next**.

Em alguns casos, pode ser necessário inserir o nome do compartilhamento SMB manualmente e selecionar **Recuperar diretórios** para exibir os diretórios disponíveis. Você pode optar por selecionar o compartilhamento inteiro ou apenas pastas específicas do compartilhamento.

3. * Configurações*: Configure como a fonte de dados ingere informações de seus arquivos e quais arquivos ela inclui em varreduras:

- **Definir fonte de dados**: Na seção **Estratégia de Chunking**, defina como o mecanismo GenAI divide o conteúdo da fonte de dados em blocos quando a fonte de dados é integrada a uma base de conhecimento. Você pode escolher uma das seguintes estratégias:
  - * Agrupamento de frases múltiplas*: Organiza informações de sua fonte de dados em blocos definidos por sentença. Você pode escolher quantas frases compõem cada pedaço (até 100).
  - * Agrupamento baseado em sobreposição*: Organiza informações de sua fonte de dados em blocos definidos por carateres que podem sobrepor blocos vizinhos. Você pode escolher o tamanho de cada pedaço em carateres, e quanto cada pedaço se sobrepõe com pedaços adjacentes. Você pode configurar um tamanho de bloco entre 50 e 3000 carateres e uma porcentagem de sobreposição entre 1 e 99%.

    Escolher uma alta porcentagem de sobreposição pode aumentar significativamente os requisitos de armazenamento com apenas pequenas melhorias na precisão de recuperação.

- **Consciente de permissão**: Habilita ou desabilita respostas cientes de permissão:
  - **Habilitado**: Os usuários do chatbot que acessam essa base de conhecimento só receberão respostas a consultas de fontes de dados às quais têm acesso.
  - **Disabled**: Os usuários do chatbot receberão respostas usando conteúdo de todas as fontes de dados integradas.
- * Filtragem de arquivos*: Configure quais arquivos estão incluídos nas digitalizações:
  - Na seção **suporte a tipos de arquivo**, escolha incluir todos os tipos de arquivos ou selecionar tipos de arquivo individuais para inclusão nas verificações de origem de dados.

    Se você incluir imagens ou arquivos PDF, o NetApp Workload Factory for GenAI analisará o texto nas imagens (incluindo imagens em documentos PDF), e isso incorrerá em um custo mais alto.

  Ao incluir dados de texto de imagens, o GenAI não consegue mascarar informações de identificação pessoal (PII) da imagem à medida que os dados de texto digitalizados são enviados do seu ambiente para a AWS. No entanto, uma vez que os dados são armazenados, todas as PII são mascaradas no banco de dados do GenAI.

  Sua escolha de incluir arquivos de imagem em digitalizações está relacionada ao modelo de bate-papo da base de conhecimento. Se você incluir arquivos de imagem em digitalizações, o modelo de bate-papo deve suportar imagens. Se os tipos de arquivo de imagem estiverem selecionados aqui, você não poderá alternar a base de conhecimento para um modelo de chat que não suporte arquivos de imagem.

- Na seção **filtro de tempo de modificação de arquivo**, escolha ativar ou desativar a inclusão de arquivos com base em seu tempo de modificação. Se ativar a filtragem de hora de modificação, selecione um intervalo de datas na lista.

  Se você incluir arquivos com base em um intervalo de datas de modificação, assim que o intervalo de datas não for satisfeito (os arquivos não foram modificados dentro do intervalo de datas especificado), os arquivos serão excluídos da verificação periódica e a fonte de dados não incluirá esses arquivos.

4. Selecione **Adicionar fonte de dados** para adicionar esta fonte de dados à sua base de conhecimento.

Resultado

A fonte de dados começa a ser incorporada na sua base de conhecimento. O status muda de "incorporação" para "incorporada" quando a fonte de dados está completamente incorporada.

Depois de adicionar uma única fonte de dados à base de conhecimento, você pode testá-la localmente na janela do simulador do chatbot e fazer as alterações necessárias antes de tornar o chatbot disponível para seus usuários. Você também pode seguir os mesmos passos para adicionar fontes de dados adicionais à base de conhecimento.

Teste uma base de conhecimento do GenAI

Depois de criar a base de conhecimento, você poderá testá-la localmente usando o simulador de chatbot e fazer as alterações necessárias antes de tornar a base de conhecimento disponível para seus usuários através de um aplicativo de chatbot.

Sobre esta tarefa

Você testa sua base de conhecimento para garantir que ela funcione como você espera e pode personalizar os iniciantes de conversação que você deseja estar disponível por padrão para os usuários de chatbot dessa base de conhecimento. O simulador de chatbot é executado contra todas as fontes de dados que foram incorporadas na base de conhecimento.

Você pode testar uma base de conhecimento conversando com suas fontes de dados incorporadas no simulador de chatbot. Observe que nenhuma interação ou insights são capturados no banco de dados de vetores GenAI ao testar a base de conhecimento localmente.

Você executará a maior parte dos seus testes no Workload Factory antes de implantar a base de conhecimento em um aplicativo para seus usuários. Se você precisar fazer alterações na sua fonte de dados ou na operação do chatbot, faça isso agora, antes de publicar sua base de conhecimento.

Você pode redimensionar e renomear a janela do simulador do chatbot e copiar perguntas e respostas para a área de transferência.

Algumas das tarefas que você vai querer executar para testar seu chatbot são:

- Insira um grande número de perguntas relevantes para sua organização para garantir que as respostas sejam as esperadas.
- Personalize os iniciantes de conversação que você deseja estar disponível por padrão para seus usuários no aplicativo chatbot.
- Certifique-se de que o conteúdo atribuído fornecido na parte inferior das respostas do chatbot contenha as referências corretas.

Passos

1. Na página de inventário das bases de conhecimento, selecione a base de conhecimento que pretende testar.

   O simulador de chatbot aparece no painel direito. Se definido, os iniciantes de conversação existentes também são exibidos.

2. No campo de entrada do chatbot, insira um prompt ou pergunta e selecione para ver como seu chatbot responde com seu conhecimento organizacional.

- Você pode ver as fontes usadas para produzir a resposta expandindo a lista **fontes** sob a resposta. Isso fornece uma lista de arquivos usados para gerar a resposta. Você pode exibir e copiar os blocos de dados usados de cada arquivo e caminho de volume para cada arquivo passando o Mouse sobre o nome do arquivo.
- Se tabelas estiverem incluídas na resposta, você poderá classificar os dados em cada coluna e copiar cada tabela para a área de transferência.
- Se os resultados das respostas contiverem dados estruturados e o recurso **S3 Bucket** estiver habilitado para a base de conhecimento, o GenAI armazenará os resultados em um bucket S3. Você pode baixar os resultados do bucket usando o link **Baixar resultados** na sessão de bate-papo.

3. Se você precisar atualizar qualquer uma de suas fontes de dados para que sua base de conhecimento forneça respostas mais focadas, faça essas alterações agora e teste novamente a base de conhecimento.

Ative a autenticação externa para uma base de conhecimento do GenAI

Ative a autenticação para uma base de conhecimento para que a validação de token e ACLs sejam necessárias ao usar os endpoints da API para integrar uma base de conhecimento com um aplicativo de chatbot. Ao ativar a autenticação, você configura as configurações de um Token Web JSON que será usado para solicitações de API para uma base de conhecimento de clientes de chatbot.

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. Na página de inventário das bases de conhecimento, selecione a base de conhecimento para a qual pretende ativar a autenticação.
4. Selecione ••• e selecione **Gerenciar base de conhecimento**.
5. Selecione o menu **ações** e selecione **Gerenciar configurações de autenticação**.
6. Configurar autenticação:

   a. Selecione **Ativar definições de autenticação**.

   b. Forneça as informações necessárias. Exemplos são fornecidos, mas você deve obter os valores para esses campos do seu provedor de autenticação:

      - **Algoritmos**: O algoritmo de assinatura que o provedor de autenticação usa.
      - **Audience** (Opcional): Uma string contendo o destinatário pretendido do token (às vezes uma URL).
      - **Emissor**: Uma string que identifica o provedor que emitiu o token.

        Por exemplo, o Amazon Cognito usa strings de emissor com o seguinte formato:

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

`<region>`Onde está a região da AWS que contém o pool de usuários `<UserPoolID>` e é o ID do pool de usuários. Você pode recuperar seu ID do pool de usuários usando o seguinte comando:

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

- * JWKS URI*: A cadeia de carateres URI que fornece chaves públicas necessárias para verificar assinaturas deste token.

  Por exemplo, o Amazon Cognito usa strings URI JWKS com o seguinte formato:

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

+
`<region>`Onde está a região da AWS que contém o pool de usuários `<UserPoolID>` e é o ID do pool de usuários. Você pode recuperar seu ID do pool de usuários usando o seguinte comando:

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

7. Selecione **Guardar**.

Resultado

A autenticação para a base de conhecimento está ativa e você pode usar endpoints de API para interagir com a base de conhecimento e integrar a base de conhecimento com um aplicativo de chatbot.

Publique uma base de conhecimento do GenAI e visualize o endpoint exclusivo

Depois de criar e testar sua base de conhecimento localmente, você pode publicar a base de conhecimento para que ela possa ser integrada a um aplicativo de chatbot que permitirá que seus usuários consultem a base de conhecimento.

Sobre esta tarefa

Publicar a base de conhecimento permite que você a utilize em aplicativos de bate-papo. A ação de publicação aciona a API do Workload Factory para gerar e publicar endpoints exclusivos. Após a publicação, a base de conhecimento se torna acessível para aplicativos de bate-papo, e os pontos de extremidade da API estão prontos para integração.

Cada base de conhecimento que você publica tem endpoints exclusivos.

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. Na página de inventário bases de conhecimento, selecione a base de conhecimento que deseja publicar.
4. Selecione ••• e selecione **Gerenciar base de conhecimento**.

   Esta página exibe o status publicado, o status de incorporação das fontes de dados, o modo de incorporação e a lista de todas as fontes de dados incorporadas.

5. Selecione o menu **ações** e selecione **publicar**.

   O Workload Factory publica a base de conhecimento. Na página de detalhes da base de conhecimento, o status muda de **Não publicado** para **Publicado**.

   Agora você pode obter detalhes sobre o endpoint exclusivo para a base de conhecimento.

6. Ao lado do status publicado, selecione **Exibir**.

   Detalhes sobre como acessar a base de conhecimento usando a API do Workload Factory são exibidos.

7. Na caixa de diálogo **Exibir informações publicadas**, copie os endpoints da API que você pode usar para integrar a base de conhecimento a um aplicativo.

   Para saber mais sobre os endpoints da API, vá para o "Documentação do API" e selecione **AI > Externo**.

   Antes de poder usar esses endpoints, você precisa obter um token de usuário do seu provedor de autenticação.

Resultado

Agora você tem uma base de conhecimento publicada e o endpoint exclusivo que você pode usar para integrar a base de conhecimento com um aplicativo de chatbot.

Use o aplicativo de chatbot de exemplo externo GenAI

Depois de configurar, ativar e publicar uma base de conhecimento, os desenvolvedores de aplicativos externos podem configurar e executar o aplicativo de chatbot de exemplo de código aberto fornecido pela NetApp para interagir com sua base de conhecimento e aprender como usar a API do Workload Factory para criar seus próprios aplicativos de IA generativos.

Passos

1. "Crie uma base de conhecimento".
2. "Ative a autenticação" para a base de conhecimento que você criou.

   Isso permite que a base de conhecimento autentique solicitações de API e torna necessária a validação de token e ACLs ao usar os endpoints da API.

   > Os aplicativos de bate-papo externos que se integram a essa base de conhecimento precisarão usar o mesmo provedor de autenticação (emissor) que você configura nas configurações de autenticação da base de conhecimento.

3. "Publique a base de conhecimento" Para habilitar o acesso à API para aplicativos externos.

Depois que uma base de conhecimento é publicada, os endpoints da API são acessíveis externamente e você pode integrar a base de conhecimento com um aplicativo de chat externo (como o exemplo de aplicativo chatbot).

4. Faça o download do pacote de aplicativos chatbot de exemplo "GitHub" do .
5. Instale e execute o aplicativo chatbot seguindo as instruções no arquivo README incluído no pacote.
6. Navegue até para "http://localhost:9091" para iniciar sessão na aplicação.

   O exemplo de aplicativo chatbot é exibido.

Saiba mais

"Documentação da API do Workload Factory"

Crie um aplicativo GenAI baseado em RAG

Depois de criar sua base de conhecimento e testar seu chatbot, você estará pronto para configurar o aplicativo que permitirá que seus usuários consultem o chatbot.

"Saiba como criar um aplicativo de IA baseado em RAG no FSX for ONTAP"

O que você pode fazer a seguir com o GenAI

Agora que você criou uma base de conhecimento usando seus dados empresariais e implantou-os para seus usuários, você pode gerenciar a base de conhecimento, as fontes de dados e a infraestrutura RAG, incluindo os sistemas de arquivos FSX for ONTAP.

Algumas das tarefas que você pode executar para gerenciar seus componentes da base de conhecimento são:

- Atualize o conteúdo de suas fontes de dados ou adicione novas fontes de dados e sincronize essas alterações com sua base de conhecimento e chatbot.
- Gerencie suas configurações de origem de dados, incluindo a estratégia de divisão e o reconhecimento de permissões (para acesso a arquivos SMB).
- Gerencie suas configurações da base de conhecimento, incluindo o modelo de bate-papo e os iniciantes de conversa.
- Despublique uma base de conhecimento ou republique-a depois de fazer alterações.
- Faça backup e proteja os dados importantes no seu sistema de arquivos FSX for ONTAP para garantir que seus dados da base de conhecimento e outros componentes de infraestrutura estejam sempre disponíveis.

Para obter informações sobre como gerenciar seu sistema de arquivos FSx for ONTAP , acesse o "Documentação do Workload Factory para Amazon FSx for NetApp ONTAP" para visualizar os recursos de backup e proteção que você pode usar.

[1] o recurso de guardrails de dados não é suportado ao inserir arquivos de dados estruturados em bases de conhecimento.

Use o GenAI para criar conetores para o Amazon Q Business

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

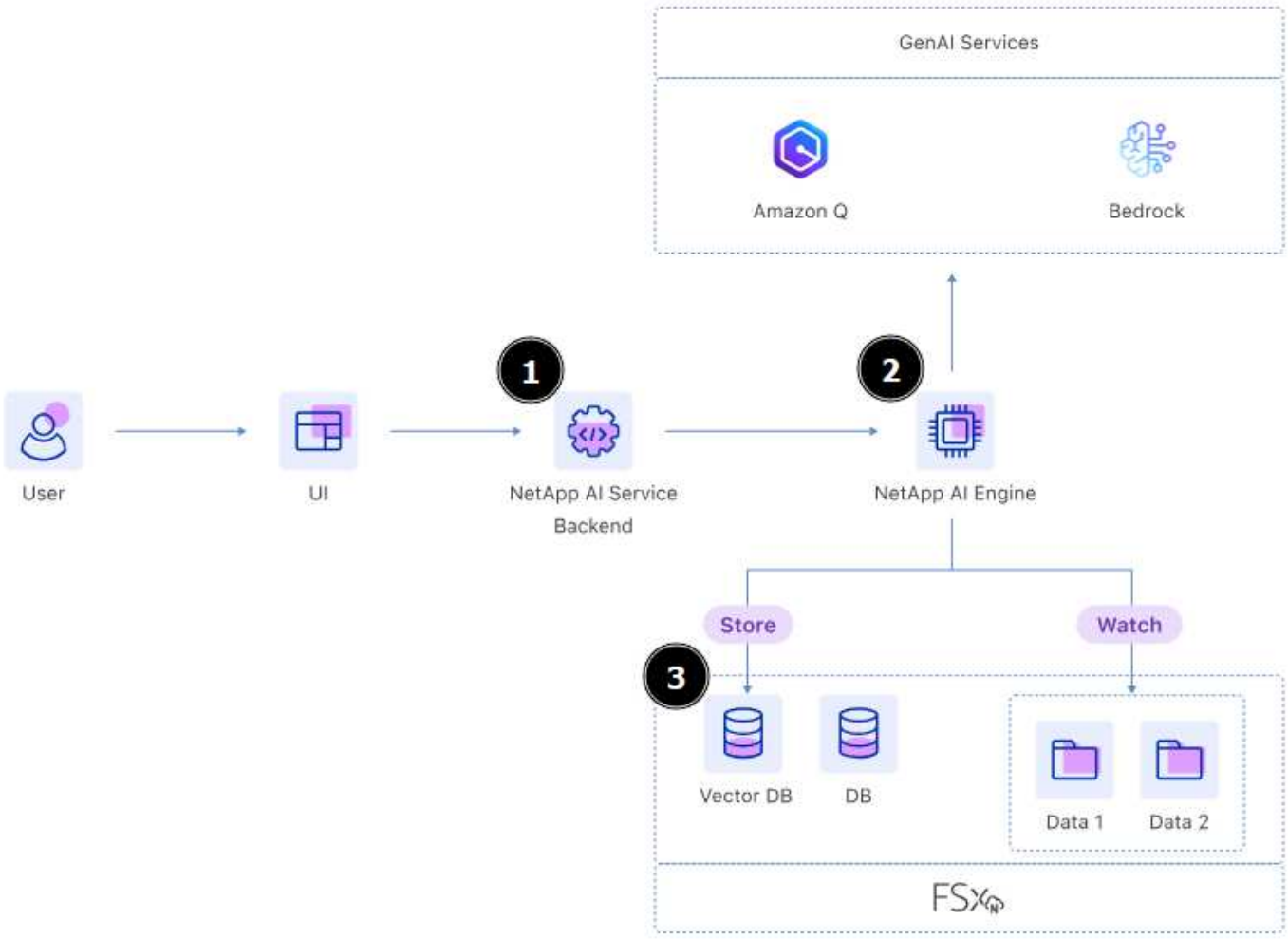
O sistema de arquivos FSX for ONTAP fornece o armazenamento para o seu sistema GenAI.

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.

Crie um conector NetApp para o Amazon Q Business

Depois de implantar a infraestrutura de IA e identificar as fontes de dados que você usará dos seus datastores do FSx para ONTAP, você estará pronto para definir um NetApp Connector para o Amazon Q Business.

Certifique-se de que seu ambiente atenda ao "requisitos" para Amazon Q Business antes de prosseguir.

Sobre esta tarefa

Fontes de dados da sua organização podem conter Informações de Identificação Pessoal (PII). Para proteger

essas informações confidenciais, você pode habilitar *guardrails de dados* ao definir um conector. Os guardrails de dados, alimentados pela NetApp Data Classification, identificam e mascaram PII, tornando-os inacessíveis e irrecuperáveis.

"Saiba mais sobre a classificação de dados da NetApp" .

O NetApp Workload Factory para GenAI não mascara informações pessoais confidenciais (SPII). Consulte"tipos de dados pessoais sensíveis" para mais informações sobre esse tipo de dados.

Os guardrails de dados podem ser ativados ou desativados a qualquer momento. Se você alternar a ativação dos guardrails de dados, o Workload Factory verificará toda a fonte de dados do zero, o que pode gerar um custo.

Defina um conetor

Crie um conector NetApp para o Amazon Q Business. O conector permite a comunicação de API e fonte de dados entre a GenAI e o Amazon Q Business.

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. No menu Bases de conhecimento e conectores, selecione o menu suspenso **Criar novo** e escolha **Conector Amazon Q Business**.
4. Na página Definir conetor, configure as definições do conetor:
   a. **Nome**: Insira o nome que deseja usar para o conetor.
   b. **Descrição**: Introduza uma descrição detalhada do conetor.
   c. **Amazon Q**: A região e o nome do aplicativo para a instância do Amazon Q Business que você deseja integrar.
   d. **Guardrails de dados**: escolha se deseja habilitar ou desabilitar os guardrails de dados. "Saiba mais sobre guardrails de dados, com tecnologia da NetApp Data Classification" .

      Os pré-requisitos a seguir devem ser atendidos para habilitar os corrimões de dados.

      - Uma conta de serviço é necessária para se comunicar com o NetApp Data Classification. Você deve ter a função *Administrador da organização* na sua conta de locação do NetApp Console para criar uma conta de serviço. Um membro que tem a função de administrador da organização pode concluir todas as ações no NetApp Console."Aprenda como adicionar uma função a um membro no NetApp Console"
      - O mecanismo de IA deve ter acesso ao"Ponto de extremidade da API do console NetApp" .
      - Você precisará fazer o seguinte conforme descrito em"Documentação de classificação de dados da NetApp" :
        A. Criar um agente de console
        B. Certifique-se de que seu ambiente atenda aos pré-requisitos
        C. Implantar classificação de dados do NetApp

Quando você ativa o recurso de guardrails de dados, o GenAI processa arquivos .txt, .md, .csv, .docx e .pdf inserindo apenas texto simples (excluindo imagem incorporada ou texto de Mídia) e mascarando quaisquer dados privados ou confidenciais. Todos os outros tipos de arquivo são processados normalmente sem mascarar dados privados ou confidenciais.

e. *FSx para sistema de arquivos ONTAP *: quando você define um novo NetApp Connector para o Amazon Q Business, o Workload Factory cria um novo volume Amazon FSx for NetApp ONTAP para armazenar as informações do conector. Escolha um sistema de arquivos existente e uma SVM (também chamada de VM de armazenamento) onde o novo volume será criado.

f. **Política de snapshot**: escolha uma política de snapshot na lista de políticas existentes definidas no inventário de armazenamento do Workload Factory. O GenAI cria automaticamente snapshots recorrentes do volume que armazena as informações do conector em uma frequência baseada na política de snapshot selecionada.

   Se a política de snapshot que você precisa não existir, você pode "criar uma política de snapshot" na VM de armazenamento que contém o volume.

5. Selecione **Create Connector** para integrar o Amazon Q Business com o GenAI.

   Um indicador de progresso aparece enquanto o conetor é criado.

   Depois que o conetor é criado, você tem a opção de adicionar uma fonte de dados ao conetor para que o Amazon Q Business ingere seus dados e os adicione ao seu índice. Recomendamos que você selecione **Adicionar fonte de dados** e adicione uma ou mais fontes de dados agora.

Adicione fontes de dados ao conetor

Você pode adicionar uma ou mais fontes de dados para preencher o índice do Amazon Q Business com os dados da sua organização.

Sobre esta tarefa

- O número máximo de fontes de dados suportadas é 10.
- Consulte o "Documentação do Amazon Q Business" para obter restrições de serviço específicas do índice Amazon Q Business.

Passos

1. Depois de selecionar **Adicionar fonte de dados**, a página **Selecionar um sistema de arquivos** é exibida.
2. * Selecione um sistema de arquivos*: Selecione o sistema de arquivos FSX for ONTAP onde seus arquivos de origem de dados residem e selecione **Next**.
3. **Selecione um volume**: Selecione o volume no qual os arquivos de origem de dados residem e selecione **Next**.

   Ao selecionar arquivos armazenados usando o protocolo SMB, você precisará inserir as informações do ative Directory, que incluem o domínio, o endereço IP, o nome de usuário e a senha.

4. **Selecione uma fonte de dados**: Selecione a localização da fonte de dados com base no local onde você salvou os arquivos. Este pode ser um volume inteiro, ou apenas uma pasta específica ou subpasta no volume, e selecione **Next**.
5. * Configurações*: Configure como a fonte de dados ingere informações de seus arquivos e quais arquivos

ela inclui em varreduras:

- * Filtragem de arquivos*: Configure quais arquivos estão incluídos nas digitalizações:
  - Na seção **suporte a tipos de arquivo**, escolha incluir todos os tipos de arquivos ou selecionar tipos de arquivo individuais para inclusão nas verificações de origem de dados.
  - Na seção **filtro de tempo de modificação de arquivo**, escolha ativar ou desativar a inclusão de arquivos com base em seu tempo de modificação. Se ativar a filtragem de hora de modificação, selecione um intervalo de datas na lista.

> Se você incluir arquivos com base em um intervalo de datas de modificação, assim que o intervalo de datas não for satisfeito (os arquivos não foram modificados dentro do intervalo de datas especificado), os arquivos serão excluídos da verificação periódica e a fonte de dados não incluirá esses arquivos.

6. Na seção **reconhecimento de permissão**, que está disponível somente quando a fonte de dados selecionada estiver em um volume que usa o protocolo SMB, você pode ativar ou desativar respostas com reconhecimento de permissão:
   - **Enabled**: Os usuários do chatbot que acessarem este conetor receberão apenas respostas a consultas de fontes de dados às quais tenham acesso.
   - **Disabled**: Os usuários do chatbot receberão respostas usando conteúdo de todas as fontes de dados integradas.

> As permissões de grupo do ative Directory não são suportadas para fontes de dados do Amazon Q Business Connector.

7. Selecione **Adicionar** para adicionar essa fonte de dados ao Amazon Q Business Connector.

Resultado

A fonte de dados está incorporada ao índice do Amazon Q Business. O status muda de "incorporação" para "incorporada" quando a fonte de dados está completamente incorporada.

Depois de adicionar uma única fonte de dados ao conetor, você pode testá-la no ambiente de chatbot do Amazon Q Business e fazer as alterações necessárias antes de disponibilizar o serviço aos usuários. Você também pode seguir os mesmos passos para adicionar fontes de dados adicionais ao conetor.

Administrar e monitorar

Gerenciar a infraestrutura do GenAI

Você pode ver detalhes sobre a infraestrutura do GenAI RAG implantada ou remover a infraestrutura do chatbot se não precisar mais dela.

Veja informações sobre a infraestrutura

Você pode ver informações sobre a infraestrutura do chatbot.

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. Selecione o menu **Infraestrutura**.
4. Veja informações sobre a infraestrutura, que inclui detalhes sobre os seguintes componentes:
    - Definições AWS
    - Configurações de infraestrutura
    - O mecanismo AI
    - O banco de dados vetorial

Remova a infraestrutura

Se você não precisar mais da infraestrutura de chatbot implantada para um ou mais chatbots, poderá removê-la do Workload Factory.

Todos os chatbots que foram implantados nessa infraestrutura serão desativados e todo o histórico de chat será excluído.

Esta operação remove apenas os links para a infraestrutura de IA do Workload Factory; ela não remove todos os componentes da AWS. Você precisará excluir manualmente os seguintes componentes de infraestrutura da AWS:

- A instância da VM
- Endpoints privados
- O volume no sistema de arquivos FSX for ONTAP que contém os bancos de dados AI
- A função do IAM
- A política
- O grupo de segurança

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. Selecione o menu **Infraestrutura**.

4. Selecione ••• e selecione **Remover a infraestrutura do chatbot**.
5. Confirme se deseja excluir a infraestrutura e selecione **Remover**.

Resultado

Os componentes de infraestrutura do chatbot são removidos do Workload Factory.

Gerenciar bases de conhecimento do GenAI

Depois de criar uma base de conhecimento, você pode visualizar os detalhes da base de conhecimento, modificar a base de conhecimento, integrar fontes de dados adicionais ou excluir a base de conhecimento.

Exibir informações sobre uma base de conhecimento

Você pode exibir informações sobre as configurações de uma base de conhecimento e a fonte de dados integrada.

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. Selecione a base de conhecimento que pretende visualizar.

   Se definido, as entradas de conversa que estão sendo usadas atualmente são exibidas no painel direito.

4. Para visualizar os detalhes da base de conhecimento, ••• selecione e selecione **Gerenciar base de conhecimento**.

   Esta página exibe o status publicado, o status de incorporação das fontes de dados, o modo de incorporação, a lista de todas as fontes de dados incorporadas e muito mais.

   O menu **ações** permite gerenciar a base de conhecimento se você quiser fazer alterações.

Edite uma base de conhecimento

Você pode atualizar uma base de conhecimento alterando algumas configurações ou pode adicionar ou remover fontes de dados.

Cada vez que você adicionar, modificar ou remover fontes de dados da base de conhecimento, você deve sincronizar a fonte de dados para que ela seja reindexada à base de conhecimento. A sincronização é incremental, portanto, o Amazon bedrock só processa os objetos no volume do FSX for ONTAP que foram adicionados, modificados ou excluídos desde a última sincronização.

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. Na página de inventário das bases de conhecimento, selecione a base de conhecimento que pretende atualizar.
4. Selecione ••• e selecione **Gerenciar base de conhecimento**.

Esta página exibe o status publicado, o status de incorporação das fontes de dados, o modo de incorporação, a lista de todas as fontes de dados incorporadas e muito mais.

5. Selecione o menu **ações** e selecione **Editar base de conhecimento**.
6. Na página Editar base de conhecimento, você pode alterar o nome da base de conhecimento, a descrição, o modelo de incorporação, o modelo de bate-papo, a ativação de recursos, escolher se os iniciadores de conversas serão criados automaticamente ou manualmente e a política de instantâneo usada para o volume que contém a base de conhecimento.

   Se você usar o modo Manual para iniciantes de conversação, você também pode alterar os iniciantes de conversação aqui.

   > Cada varredura da base de conhecimento, que inclui incorporação, custos. Se os guardrails de dados estiverem ativados após a criação de uma base de conhecimento, a base de conhecimento será digitalizada novamente e incorrerá em custos. Da mesma forma, se você alterar os modelos de chat, o GenAI fará uma nova varredura das fontes de dados associadas (incorrendo em um custo).

7. Selecione **Salvar** depois de fazer suas alterações.

Proteja uma base de conhecimento com snapshots

Você pode proteger os dados da sua base de conhecimento tirando e restaurando snapshots dos volumes da sua base de conhecimento. Você pode restaurar a partir de um instantâneo para reverter para a versão anterior da base de conhecimento a qualquer momento.

Os snapshots podem ser mais rápidos e eficientes em storage do que os backups, além de permitir que você proteja cada base de conhecimento usando uma política de proteção diferente. Alguns dos cenários em que os instantâneos podem ser úteis são:

- Perda ou corrupção acidental de dados
- Recuperação de dados incorretos que estão sendo ingeridos na base de conhecimento
- Testando diferentes fontes de dados ou estratégias de divisão, e revertendo rapidamente quando o teste estiver concluído

Tire um instantâneo de um volume da base de conhecimento

Você pode salvar o estado de uma base de conhecimento tirando um instantâneo manual do volume da base de conhecimento.

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. Na página de inventário bases de conhecimento, selecione a base de conhecimento que deseja proteger.
4. Selecione ••• e selecione **Gerenciar base de conhecimento**.

   Esta página exibe o status publicado, o status de incorporação das fontes de dados, o modo de incorporação, a lista de todas as fontes de dados incorporadas e muito mais.

5. Selecione o menu **ações** e selecione **Snapshot > Create new snapshot**.
6. Opcionalmente, selecione **Definir nome do instantâneo** e insira um nome personalizado para o

instantâneo.

Definir um nome personalizado pode ajudá-lo a determinar melhor o conteúdo de um snapshot se você precisar restaurá-lo no futuro.

7. Selecione **criar**.

   Um instantâneo da base de conhecimento é criado.

Restaurar um snapshot de um volume da base de conhecimento

Você pode restaurar um instantâneo manual ou programado de um volume da base de conhecimento a qualquer momento.

> Não é possível restaurar um instantâneo usando a IU de cargas de trabalho de IA generativa se o banco de dados armazenado no volume estiver corrompido ou tiver sido excluído. Como solução alternativa, você pode restaurar o instantâneo usando o "CLI do ONTAP" no cluster do ONTAP onde o volume está hospedado.

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. Na página de inventário bases de conhecimento, selecione a base de conhecimento que deseja restaurar.
4. Selecione ••• e selecione **Gerenciar base de conhecimento**.

   Esta página exibe o status publicado, o status de incorporação das fontes de dados, o modo de incorporação, a lista de todas as fontes de dados incorporadas e muito mais.

5. Selecione o menu **ações** e selecione **Snapshot > Restore snapshot**.

   É apresentada a caixa de diálogo de seleção de instantâneos, onde pode ver uma lista dos instantâneos criados para esta base de dados de conhecimento.

6. (Opcional) Deselecione a opção **Pausa em execução e verificações agendadas após restaurar o instantâneo** se quiser que as verificações de origem de dados agendadas e atualmente em execução continuem após a restauração do instantâneo.

   Esta opção está ativada por predefinição para garantir que uma verificação não aconteça enquanto a base de conhecimento estiver num estado parcialmente restaurado ou que uma verificação não atualize uma base de conhecimento recentemente restaurada com dados mais antigos.

7. Selecione o instantâneo que pretende restaurar a partir da lista.
8. Selecione **Restaurar**.

Clonar uma base de conhecimento

Você pode criar uma nova base de conhecimento a partir de um snapshot da base de conhecimento. Isso é útil se a base de conhecimento original estiver corrompida ou perdida.

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. Na página de inventário bases de conhecimento, selecione a base de conhecimento que deseja restaurar.
4. Selecione ••• e selecione **Gerenciar base de conhecimento**.

   Esta página exibe o status publicado, o status de incorporação das fontes de dados, o modo de incorporação, a lista de todas as fontes de dados incorporadas e muito mais.

5. Selecione o menu **ações** e selecione **Snapshot > Clone base de conhecimento**.

   A caixa de diálogo clone é exibida.

6. Opcionalmente, desmarque a opção **Pausa em execução e verificações agendadas após clonar o instantâneo** se quiser que as verificações de origem de dados agendadas e atualmente em execução continuem após o instantâneo ser clonado.

   Esta opção está ativada por predefinição para garantir que uma verificação não aconteça enquanto a base de conhecimento estiver num estado parcialmente restaurado ou que uma verificação não atualize uma base de conhecimento recentemente restaurada com dados mais antigos.

7. Selecione o instantâneo que deseja clonar na lista.
8. Selecione **continuar**.
9. Insira um nome para a nova base de conhecimento.
10. Escolha um SVM de sistema de arquivos e nome de volume para a nova base de conhecimento.
11. Selecione **Clone**.

Adicione fontes de dados adicionais a uma base de conhecimento

Você pode incorporar fontes de dados adicionais em sua base de conhecimento para preenchê-la com dados adicionais da organização.

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. Na página de inventário das bases de dados de conhecimento, selecione a base de conhecimento onde deseja adicionar a fonte de dados.
4. Selecione ••• e selecione **Adicionar fonte de dados**.
5. Selecione o tipo de fonte de dados que você deseja adicionar:
   - Adicionar FSx para sistema de arquivos ONTAP (usar arquivos de um volume FSx para ONTAP existente)
   - Adicionar sistema de arquivos (usar arquivos de um compartilhamento SMB ou NFS genérico)

Adicionar um FSx para sistema de arquivos ONTAP

1. * Selecione um sistema de arquivos*: Selecione o sistema de arquivos FSX for ONTAP onde seus arquivos de origem de dados residem e selecione **Next**.
2. **Selecione um volume**: Selecione o volume no qual os arquivos de origem de dados residem e selecione **Next**.

   Ao selecionar arquivos armazenados usando o protocolo SMB, você precisará inserir as informações do ative Directory, que incluem o domínio, o endereço IP, o nome de usuário e a senha.

3. **Selecione uma fonte de dados**: Selecione a localização da fonte de dados com base no local onde você salvou os arquivos. Este pode ser um volume inteiro, ou apenas uma pasta específica ou subpasta no volume, e selecione **Next**.
4. * Configurações*: Configure como a fonte de dados ingere informações de seus arquivos e quais arquivos ela inclui em varreduras:

   - **Definir fonte de dados**: Na seção **Estratégia de Chunking**, defina como o mecanismo GenAI divide o conteúdo da fonte de dados em blocos quando a fonte de dados é integrada a uma base de conhecimento. Você pode escolher uma das seguintes estratégias:
     - * Agrupamento de frases múltiplas*: Organiza informações de sua fonte de dados em blocos definidos por sentença. Você pode escolher quantas frases compõem cada pedaço (até 100).
     - * Agrupamento baseado em sobreposição*: Organiza informações de sua fonte de dados em blocos definidos por carateres que podem sobrepor blocos vizinhos. Você pode escolher o tamanho de cada pedaço em carateres, e quanto cada pedaço se sobrepõe com pedaços adjacentes. Você pode configurar um tamanho de bloco entre 50 e 3000 carateres e uma porcentagem de sobreposição entre 1 e 99%.

       Escolher uma alta porcentagem de sobreposição pode aumentar significativamente os requisitos de armazenamento com apenas pequenas melhorias na precisão de recuperação.

   - * Filtragem de arquivos*: Configure quais arquivos estão incluídos nas digitalizações:
     - Na seção **suporte a tipos de arquivo**, escolha incluir todos os tipos de arquivos ou selecionar tipos de arquivo individuais para inclusão nas verificações de origem de dados.

       Se você incluir imagens ou arquivos PDF, o NetApp Workload Factory for GenAI analisará o texto nas imagens (incluindo imagens em documentos PDF), e isso incorrerá em um custo mais alto.

     Ao incluir dados de texto de imagens, o GenAI não consegue mascarar informações de identificação pessoal (PII) da imagem à medida que os dados de texto digitalizados são enviados do seu ambiente para a AWS. No entanto, uma vez que os dados são armazenados, todas as PII são mascaradas no banco de dados do GenAI.

     Sua escolha de incluir arquivos de imagem em digitalizações está relacionada ao modelo de bate-papo da base de conhecimento. Se você incluir arquivos de imagem em digitalizações, o modelo de bate-papo deve suportar imagens. Se os tipos de arquivo de imagem estiverem selecionados aqui, você não poderá alternar a base de conhecimento para um modelo de chat que não suporte arquivos de imagem.

   - Na seção **filtro de tempo de modificação de arquivo**, escolha ativar ou desativar a inclusão de

arquivos com base em seu tempo de modificação. Se ativar a filtragem de hora de modificação, selecione um intervalo de datas na lista.

> Se você incluir arquivos com base em um intervalo de datas de modificação, assim que o intervalo de datas não for satisfeito (os arquivos não foram modificados dentro do intervalo de datas especificado), os arquivos serão excluídos da verificação periódica e a fonte de dados não incluirá esses arquivos.

5. Na seção **reconhecimento de permissão**, que está disponível somente quando a fonte de dados selecionada estiver em um volume que usa o protocolo SMB, você pode ativar ou desativar respostas com reconhecimento de permissão:
   - **Habilitado**: Os usuários do chatbot que acessam essa base de conhecimento só receberão respostas a consultas de fontes de dados às quais têm acesso.
   - **Disabled**: Os usuários do chatbot receberão respostas usando conteúdo de todas as fontes de dados integradas.
6. Selecione **Add** para adicionar esta fonte de dados à sua base de conhecimento.

Adicionar um sistema de arquivos NFS genérico

1. **Selecione um sistema de arquivos**: insira o endereço IP ou FQDN do host do sistema de arquivos onde seus arquivos de fonte de dados residem, escolha o protocolo NFS para o compartilhamento de rede e selecione **Avançar**.
2. **Selecione uma fonte de dados**: Selecione a localização da fonte de dados com base no local onde você salvou os arquivos. Este pode ser um volume inteiro, ou apenas uma pasta específica ou subpasta no volume, e selecione **Next**.

> Em alguns casos, pode ser necessário inserir o nome da exportação NFS manualmente e selecionar **Recuperar diretórios** para exibir os diretórios disponíveis. Você pode optar por selecionar a exportação inteira ou apenas pastas específicas da exportação.

3. * Configurações*: Configure como a fonte de dados ingere informações de seus arquivos e quais arquivos ela inclui em varreduras:
   - **Definir fonte de dados**: Na seção **Estratégia de Chunking**, defina como o mecanismo GenAI divide o conteúdo da fonte de dados em blocos quando a fonte de dados é integrada a uma base de conhecimento. Você pode escolher uma das seguintes estratégias:
     - * Agrupamento de frases múltiplas*: Organiza informações de sua fonte de dados em blocos definidos por sentença. Você pode escolher quantas frases compõem cada pedaço (até 100).
     - * Agrupamento baseado em sobreposição*: Organiza informações de sua fonte de dados em blocos definidos por carateres que podem sobrepor blocos vizinhos. Você pode escolher o tamanho de cada pedaço em carateres, e quanto cada pedaço se sobrepõe com pedaços adjacentes. Você pode configurar um tamanho de bloco entre 50 e 3000 carateres e uma porcentagem de sobreposição entre 1 e 99%.

> Escolher uma alta porcentagem de sobreposição pode aumentar significativamente os requisitos de armazenamento com apenas pequenas melhorias na precisão de recuperação.

   - * Filtragem de arquivos*: Configure quais arquivos estão incluídos nas digitalizações:

- Na seção **suporte a tipos de arquivo**, escolha incluir todos os tipos de arquivos ou selecionar tipos de arquivo individuais para inclusão nas verificações de origem de dados.

  Se você incluir imagens ou arquivos PDF, o NetApp Workload Factory for GenAI analisará o texto nas imagens (incluindo imagens em documentos PDF), e isso incorrerá em um custo mais alto.

Ao incluir dados de texto de imagens, o GenAI não consegue mascarar informações de identificação pessoal (PII) da imagem à medida que os dados de texto digitalizados são enviados do seu ambiente para a AWS. No entanto, uma vez que os dados são armazenados, todas as PII são mascaradas no banco de dados do GenAI.

Sua escolha de incluir arquivos de imagem em digitalizações está relacionada ao modelo de bate-papo da base de conhecimento. Se você incluir arquivos de imagem em digitalizações, o modelo de bate-papo deve suportar imagens. Se os tipos de arquivo de imagem estiverem selecionados aqui, você não poderá alternar a base de conhecimento para um modelo de chat que não suporte arquivos de imagem.

- Na seção **filtro de tempo de modificação de arquivo**, escolha ativar ou desativar a inclusão de arquivos com base em seu tempo de modificação. Se ativar a filtragem de hora de modificação, selecione um intervalo de datas na lista.

Se você incluir arquivos com base em um intervalo de datas de modificação, assim que o intervalo de datas não for satisfeito (os arquivos não foram modificados dentro do intervalo de datas especificado), os arquivos serão excluídos da verificação periódica e a fonte de dados não incluirá esses arquivos.

4. Selecione **Adicionar fonte de dados** para adicionar esta fonte de dados à sua base de conhecimento.

Adicionar um sistema de arquivos SMB genérico

1. **Selecione o sistema de arquivos**:
   a. Digite o endereço IP ou FQDN do host do sistema de arquivos onde seus arquivos de fonte de dados residem.
   b. Escolha o protocolo SMB para o compartilhamento de rede.
   c. Insira as informações do Active Directory, que incluem o domínio, endereço IP, nome de usuário e senha.
   d. Selecione **seguinte**.
2. **Selecione uma fonte de dados**: Selecione a localização da fonte de dados com base no local onde você salvou os arquivos. Este pode ser um volume inteiro, ou apenas uma pasta específica ou subpasta no volume, e selecione **Next**.

Em alguns casos, pode ser necessário inserir o nome do compartilhamento SMB manualmente e selecionar **Recuperar diretórios** para exibir os diretórios disponíveis. Você pode optar por selecionar o compartilhamento inteiro ou apenas pastas específicas do compartilhamento.

3. * Configurações*: Configure como a fonte de dados ingere informações de seus arquivos e quais arquivos ela inclui em varreduras:

- **Definir fonte de dados**: Na seção **Estratégia de Chunking**, defina como o mecanismo GenAI divide o conteúdo da fonte de dados em blocos quando a fonte de dados é integrada a uma base de conhecimento. Você pode escolher uma das seguintes estratégias:
  - * Agrupamento de frases múltiplas*: Organiza informações de sua fonte de dados em blocos definidos por sentença. Você pode escolher quantas frases compõem cada pedaço (até 100).
  - * Agrupamento baseado em sobreposição*: Organiza informações de sua fonte de dados em blocos definidos por carateres que podem sobrepor blocos vizinhos. Você pode escolher o tamanho de cada pedaço em carateres, e quanto cada pedaço se sobrepõe com pedaços adjacentes. Você pode configurar um tamanho de bloco entre 50 e 3000 carateres e uma porcentagem de sobreposição entre 1 e 99%.

    Escolher uma alta porcentagem de sobreposição pode aumentar significativamente os requisitos de armazenamento com apenas pequenas melhorias na precisão de recuperação.

- **Consciente de permissão**: Habilita ou desabilita respostas cientes de permissão:
  - **Habilitado**: Os usuários do chatbot que acessam essa base de conhecimento só receberão respostas a consultas de fontes de dados às quais têm acesso.
  - **Disabled**: Os usuários do chatbot receberão respostas usando conteúdo de todas as fontes de dados integradas.
- * Filtragem de arquivos*: Configure quais arquivos estão incluídos nas digitalizações:
  - Na seção **suporte a tipos de arquivo**, escolha incluir todos os tipos de arquivos ou selecionar tipos de arquivo individuais para inclusão nas verificações de origem de dados.

    Se você incluir imagens ou arquivos PDF, o NetApp Workload Factory for GenAI analisará o texto nas imagens (incluindo imagens em documentos PDF), e isso incorrerá em um custo mais alto.

  Ao incluir dados de texto de imagens, o GenAI não consegue mascarar informações de identificação pessoal (PII) da imagem à medida que os dados de texto digitalizados são enviados do seu ambiente para a AWS. No entanto, uma vez que os dados são armazenados, todas as PII são mascaradas no banco de dados do GenAI.

  Sua escolha de incluir arquivos de imagem em digitalizações está relacionada ao modelo de bate-papo da base de conhecimento. Se você incluir arquivos de imagem em digitalizações, o modelo de bate-papo deve suportar imagens. Se os tipos de arquivo de imagem estiverem selecionados aqui, você não poderá alternar a base de conhecimento para um modelo de chat que não suporte arquivos de imagem.

- Na seção **filtro de tempo de modificação de arquivo**, escolha ativar ou desativar a inclusão de arquivos com base em seu tempo de modificação. Se ativar a filtragem de hora de modificação, selecione um intervalo de datas na lista.

  Se você incluir arquivos com base em um intervalo de datas de modificação, assim que o intervalo de datas não for satisfeito (os arquivos não foram modificados dentro do intervalo de datas especificado), os arquivos serão excluídos da verificação periódica e a fonte de dados não incluirá esses arquivos.

4. Selecione **Adicionar fonte de dados** para adicionar esta fonte de dados à sua base de conhecimento.

Resultado

A fonte de dados está integrada à sua base de conhecimento.

Sincronize suas fontes de dados com uma base de conhecimento

As fontes de dados são sincronizadas com a base de conhecimento associada automaticamente uma vez por dia, para que quaisquer alterações na fonte de dados sejam refletidas no chatbot. Se você fizer alterações em qualquer uma de suas fontes de dados e quiser sincronizar os dados imediatamente, poderá executar uma sincronização sob demanda.

A sincronização é incremental, portanto, o Amazon bedrock só processa os objetos em suas fontes de dados que foram adicionados, modificados ou excluídos desde a última sincronização.

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. Na página de inventário bases de conhecimento, selecione a base de conhecimento que deseja sincronizar.
4. Selecione ••• e selecione **Gerenciar base de conhecimento**.
5. Selecione o menu **ações** e selecione **Digitalizar agora**.

   Você verá uma mensagem informando que suas fontes de dados estão sendo digitalizadas e uma mensagem final quando a digitalização estiver concluída.

Resultado

A base de conhecimento é sincronizada com as fontes de dados anexadas e qualquer chatbot ativo começará a usar as informações mais recentes de suas fontes de dados.

Pausar ou retomar uma sincronização agendada

Se pretender pausar ou retomar a próxima sincronização (digitalização) das fontes de dados, pode fazê-lo a qualquer momento. Talvez seja necessário pausar a próxima sincronização agendada se você fizer alterações em uma fonte de dados e não quiser que a sincronização aconteça durante a janela de mudança.

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. No menu Bases de conhecimento e conectores, selecione a base de conhecimento cujas verificações você deseja pausar ou retomar.
4. Selecione ••• e selecione **Gerenciar base de conhecimento**.
5. Selecione o menu **ações** e selecione **Digitalizar > Pausar digitalização agendada** ou **Digitalizar > Retomar digitalização agendada**.

   Você verá uma mensagem informando que a próxima digitalização agendada foi pausada ou retomada.

Avalie modelos de bate-papo antes de criar uma base de conhecimento

Você pode avaliar os modelos básicos de bate-papo disponíveis antes de criar uma base de conhecimento para que você possa ver qual modelo funciona melhor para sua implementação. Como o suporte ao modelo

varia de acordo com a região da AWS, "Esta página de documentação da AWS" consulte para verificar quais modelos você pode usar nas regiões em que planeja implantar sua base de conhecimento.

> Esta funcionalidade só está disponível quando não foram criadas bases de conhecimento — quando não existem bases de conhecimento na página de inventário bases de conhecimento.

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. Na página de inventário do Knowledge base, você verá a opção de selecionar o modelo de bate-papo no lado direito da página do Chatbot.
4. Selecione o modelo de chat na lista e insira um conjunto de perguntas na área de prompt para ver como o chatbot responde.
5. Experimente vários modelos para ver qual modelo é melhor para sua implementação.

Resultado

Use esse modelo de chat ao criar sua base de conhecimento.

Despublique sua base de conhecimento

Depois de publicar sua base de conhecimento para que ela possa ser integrada a um aplicativo de chatbot, você pode despublicá-la se quiser desativar o aplicativo de chatbot de acessar a base de conhecimento.

A despublicação da base de conhecimento impede que qualquer aplicativo de bate-papo funcione. O endpoint de API exclusivo no qual a base de conhecimento estava acessível está desativado.

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. Na página de inventário bases de conhecimento, selecione a base de conhecimento que pretende anular a publicação.
4. Selecione ••• e selecione **Gerenciar base de conhecimento**.

   Esta página exibe o status publicado, o status de incorporação das fontes de dados, o modo de incorporação e a lista de todas as fontes de dados incorporadas.

5. Selecione o menu **ações** e selecione **Unpublish**.

Resultado

A base de conhecimento está desativada e não é mais acessível por um aplicativo de chatbot.

Excluir uma base de conhecimento

Se você não precisar mais de uma base de conhecimento, poderá excluí-la. Quando você exclui uma base de conhecimento, ela é removida do Workload Factory e o volume que contém a base de conhecimento é excluído. Todos os aplicativos ou chatbots que estiverem usando a base de conhecimento deixarão de funcionar. A exclusão de uma base de conhecimento não é reversível.

Ao excluir uma base de conhecimento, você também deve desassociar a base de conhecimento de quaisquer

agentes a que está associada para excluir totalmente todos os recursos associados à base de conhecimento.

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. Na página de inventário das bases de conhecimento, selecione a base de conhecimento que pretende eliminar.
4. Selecione ••• e selecione **Gerenciar base de conhecimento**.
5. Selecione o menu **ações** e selecione **Excluir base de conhecimento**.
6. Na caixa de diálogo Excluir base de conhecimento, confirme se deseja excluí-la e selecione **Excluir**.

Resultado

A base de conhecimento é removida do Workload Factory e seu volume associado é excluído.

Gerencie os conetores do Amazon Q Business

Depois de criar um conetor para o Amazon Q Business, você pode exibir os detalhes do conetor, modificar o conetor, integrar fontes de dados adicionais ou excluir o conetor.

Exibir informações sobre um conetor

Pode visualizar informações sobre as definições de um conetor e as fontes de dados integradas.

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. Na página de inventário Bases de conhecimento e conectores, selecione o conector que você deseja visualizar.
4. Para ver os detalhes do conetor, •••selecione e selecione **Manage Connector**.

   Esta página exibe o status publicado, o status de incorporação das fontes de dados, o modo de incorporação, a lista de todas as fontes de dados incorporadas e muito mais.

   O menu **ações** permite gerenciar o conetor se você quiser fazer alterações.

Edite um conetor

Pode atualizar um conetor alterando algumas definições ou pode adicionar ou remover fontes de dados.

Sempre que você adicionar, modificar ou remover fontes de dados do conetor, o GenAI precisa enviar as informações de origem de dados para o Amazon Q Business para que elas sejam reindexadas. A sincronização é incremental, então o Amazon Q Business processa apenas os objetos no volume do FSX for ONTAP que foram adicionados, modificados ou excluídos desde a última sincronização.

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. Na página de inventário bases de conhecimento e conetores, selecione o conetor que pretende atualizar.
4. Selecione ••• e selecione **Manage Connector**.

   Esta página exibe o status publicado, o status de incorporação das fontes de dados, o modo de incorporação, a lista de todas as fontes de dados incorporadas e muito mais.

5. Selecione o menu **ações** e selecione **Editar conetor**.
6. Na página Editar conetor, você pode alterar o nome do conetor, a descrição, o modelo de incorporação, a habilitação dos corrimões de dados e a política de snapshot usada para o volume que contém o conetor.

   Toda varredura de fonte de dados, que inclui incorporação, incorre em um custo. Se você ativar os corrimões de dados depois que um conetor foi criado, a fonte de dados será digitalizada novamente e incorrerá em custos.

7. Selecione **Salvar** depois de fazer alterações.

Adicione fontes de dados adicionais a um conetor

Você pode incorporar fontes de dados adicionais no seu conetor para preenchê-lo com dados adicionais da organização.

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. A partir da página de inventário bases de dados e conetores de conhecimento, selecione o conetor onde pretende adicionar a fonte de dados.
4. Selecione ••• e selecione **Adicionar fonte de dados**.
5. Selecione o tipo de fonte de dados que você deseja adicionar:
   - Adicionar FSx para sistema de arquivos ONTAP (usar arquivos de um volume FSx para ONTAP existente)
   - Adicionar sistema de arquivos (usar arquivos de um compartilhamento SMB ou NFS genérico)

Adicionar um FSx para sistema de arquivos ONTAP

1. * Selecione um sistema de arquivos*: Selecione o sistema de arquivos FSX for ONTAP onde seus arquivos de origem de dados residem e selecione **Next**.
2. **Selecione um volume**: Selecione o volume no qual os arquivos de origem de dados residem e selecione **Next**.

   Ao selecionar arquivos armazenados usando o protocolo SMB, você precisará inserir as informações do ative Directory, que incluem o domínio, o endereço IP, o nome de usuário e a senha.

3. **Selecione uma fonte de dados**: Selecione a localização da fonte de dados com base no local onde você salvou os arquivos. Este pode ser um volume inteiro, ou apenas uma pasta específica ou subpasta no volume, e selecione **Next**.
4. * Configurações*: Configure como a fonte de dados ingere informações de seus arquivos e quais arquivos ela inclui em varreduras:

   - **Definir fonte de dados**: Na seção **Estratégia de Chunking**, defina como o mecanismo GenAI divide o conteúdo da fonte de dados em blocos quando a fonte de dados é integrada a uma base de conhecimento. Você pode escolher uma das seguintes estratégias:
     - * Agrupamento de frases múltiplas*: Organiza informações de sua fonte de dados em blocos definidos por sentença. Você pode escolher quantas frases compõem cada pedaço (até 100).
     - * Agrupamento baseado em sobreposição*: Organiza informações de sua fonte de dados em blocos definidos por carateres que podem sobrepor blocos vizinhos. Você pode escolher o tamanho de cada pedaço em carateres, e quanto cada pedaço se sobrepõe com pedaços adjacentes. Você pode configurar um tamanho de bloco entre 50 e 3000 carateres e uma porcentagem de sobreposição entre 1 e 99%.

       Escolher uma alta porcentagem de sobreposição pode aumentar significativamente os requisitos de armazenamento com apenas pequenas melhorias na precisão de recuperação.

   - * Filtragem de arquivos*: Configure quais arquivos estão incluídos nas digitalizações:
     - Na seção **suporte a tipos de arquivo**, escolha incluir todos os tipos de arquivos ou selecionar tipos de arquivo individuais para inclusão nas verificações de origem de dados.

       Se você incluir imagens ou arquivos PDF, o NetApp Workload Factory for GenAI analisará o texto nas imagens (incluindo imagens em documentos PDF), e isso incorrerá em um custo mais alto.

     Ao incluir dados de texto de imagens, o GenAI não consegue mascarar informações de identificação pessoal (PII) da imagem à medida que os dados de texto digitalizados são enviados do seu ambiente para a AWS. No entanto, uma vez que os dados são armazenados, todas as PII são mascaradas no banco de dados do GenAI.

     Sua escolha de incluir arquivos de imagem em digitalizações está relacionada ao modelo de bate-papo da base de conhecimento. Se você incluir arquivos de imagem em digitalizações, o modelo de bate-papo deve suportar imagens. Se os tipos de arquivo de imagem estiverem selecionados aqui, você não poderá alternar a base de conhecimento para um modelo de chat que não suporte arquivos de imagem.

   - Na seção **filtro de tempo de modificação de arquivo**, escolha ativar ou desativar a inclusão de

arquivos com base em seu tempo de modificação. Se ativar a filtragem de hora de modificação, selecione um intervalo de datas na lista.

> Se você incluir arquivos com base em um intervalo de datas de modificação, assim que o intervalo de datas não for satisfeito (os arquivos não foram modificados dentro do intervalo de datas especificado), os arquivos serão excluídos da verificação periódica e a fonte de dados não incluirá esses arquivos.

5. Na seção **reconhecimento de permissão**, que está disponível somente quando a fonte de dados selecionada estiver em um volume que usa o protocolo SMB, você pode ativar ou desativar respostas com reconhecimento de permissão:
   - **Habilitado**: Os usuários do chatbot que acessam essa base de conhecimento só receberão respostas a consultas de fontes de dados às quais têm acesso.
   - **Disabled**: Os usuários do chatbot receberão respostas usando conteúdo de todas as fontes de dados integradas.
6. Selecione **Add** para adicionar esta fonte de dados à sua base de conhecimento.

Adicionar um sistema de arquivos NFS genérico

1. **Selecione um sistema de arquivos**: insira o endereço IP ou FQDN do host do sistema de arquivos onde seus arquivos de fonte de dados residem, escolha o protocolo NFS para o compartilhamento de rede e selecione **Avançar**.
2. **Selecione uma fonte de dados**: Selecione a localização da fonte de dados com base no local onde você salvou os arquivos. Este pode ser um volume inteiro, ou apenas uma pasta específica ou subpasta no volume, e selecione **Next**.

> Em alguns casos, pode ser necessário inserir o nome da exportação NFS manualmente e selecionar **Recuperar diretórios** para exibir os diretórios disponíveis. Você pode optar por selecionar a exportação inteira ou apenas pastas específicas da exportação.

3. * Configurações*: Configure como a fonte de dados ingere informações de seus arquivos e quais arquivos ela inclui em varreduras:
   - **Definir fonte de dados**: Na seção **Estratégia de Chunking**, defina como o mecanismo GenAI divide o conteúdo da fonte de dados em blocos quando a fonte de dados é integrada a uma base de conhecimento. Você pode escolher uma das seguintes estratégias:
     - * Agrupamento de frases múltiplas*: Organiza informações de sua fonte de dados em blocos definidos por sentença. Você pode escolher quantas frases compõem cada pedaço (até 100).
     - * Agrupamento baseado em sobreposição*: Organiza informações de sua fonte de dados em blocos definidos por carateres que podem sobrepor blocos vizinhos. Você pode escolher o tamanho de cada pedaço em carateres, e quanto cada pedaço se sobrepõe com pedaços adjacentes. Você pode configurar um tamanho de bloco entre 50 e 3000 carateres e uma porcentagem de sobreposição entre 1 e 99%.

> Escolher uma alta porcentagem de sobreposição pode aumentar significativamente os requisitos de armazenamento com apenas pequenas melhorias na precisão de recuperação.

   - * Filtragem de arquivos*: Configure quais arquivos estão incluídos nas digitalizações:

- Na seção **suporte a tipos de arquivo**, escolha incluir todos os tipos de arquivos ou selecionar tipos de arquivo individuais para inclusão nas verificações de origem de dados.

  Se você incluir imagens ou arquivos PDF, o NetApp Workload Factory for GenAI analisará o texto nas imagens (incluindo imagens em documentos PDF), e isso incorrerá em um custo mais alto.

Ao incluir dados de texto de imagens, o GenAI não consegue mascarar informações de identificação pessoal (PII) da imagem à medida que os dados de texto digitalizados são enviados do seu ambiente para a AWS. No entanto, uma vez que os dados são armazenados, todas as PII são mascaradas no banco de dados do GenAI.

Sua escolha de incluir arquivos de imagem em digitalizações está relacionada ao modelo de bate-papo da base de conhecimento. Se você incluir arquivos de imagem em digitalizações, o modelo de bate-papo deve suportar imagens. Se os tipos de arquivo de imagem estiverem selecionados aqui, você não poderá alternar a base de conhecimento para um modelo de chat que não suporte arquivos de imagem.

- Na seção **filtro de tempo de modificação de arquivo**, escolha ativar ou desativar a inclusão de arquivos com base em seu tempo de modificação. Se ativar a filtragem de hora de modificação, selecione um intervalo de datas na lista.

Se você incluir arquivos com base em um intervalo de datas de modificação, assim que o intervalo de datas não for satisfeito (os arquivos não foram modificados dentro do intervalo de datas especificado), os arquivos serão excluídos da verificação periódica e a fonte de dados não incluirá esses arquivos.

4. Selecione **Adicionar fonte de dados** para adicionar esta fonte de dados à sua base de conhecimento.

Adicionar um sistema de arquivos SMB genérico

1. **Selecione o sistema de arquivos**:
   a. Digite o endereço IP ou FQDN do host do sistema de arquivos onde seus arquivos de fonte de dados residem.
   b. Escolha o protocolo SMB para o compartilhamento de rede.
   c. Insira as informações do Active Directory, que incluem o domínio, endereço IP, nome de usuário e senha.
   d. Selecione **seguinte**.
2. **Selecione uma fonte de dados**: Selecione a localização da fonte de dados com base no local onde você salvou os arquivos. Este pode ser um volume inteiro, ou apenas uma pasta específica ou subpasta no volume, e selecione **Next**.

Em alguns casos, pode ser necessário inserir o nome do compartilhamento SMB manualmente e selecionar **Recuperar diretórios** para exibir os diretórios disponíveis. Você pode optar por selecionar o compartilhamento inteiro ou apenas pastas específicas do compartilhamento.

3. * Configurações*: Configure como a fonte de dados ingere informações de seus arquivos e quais arquivos ela inclui em varreduras:

- **Definir fonte de dados**: Na seção **Estratégia de Chunking**, defina como o mecanismo GenAI divide o conteúdo da fonte de dados em blocos quando a fonte de dados é integrada a uma base de conhecimento. Você pode escolher uma das seguintes estratégias:
  - * Agrupamento de frases múltiplas*: Organiza informações de sua fonte de dados em blocos definidos por sentença. Você pode escolher quantas frases compõem cada pedaço (até 100).
  - * Agrupamento baseado em sobreposição*: Organiza informações de sua fonte de dados em blocos definidos por carateres que podem sobrepor blocos vizinhos. Você pode escolher o tamanho de cada pedaço em carateres, e quanto cada pedaço se sobrepõe com pedaços adjacentes. Você pode configurar um tamanho de bloco entre 50 e 3000 carateres e uma porcentagem de sobreposição entre 1 e 99%.

    Escolher uma alta porcentagem de sobreposição pode aumentar significativamente os requisitos de armazenamento com apenas pequenas melhorias na precisão de recuperação.

- **Consciente de permissão**: Habilita ou desabilita respostas cientes de permissão:
  - **Habilitado**: Os usuários do chatbot que acessam essa base de conhecimento só receberão respostas a consultas de fontes de dados às quais têm acesso.
  - **Disabled**: Os usuários do chatbot receberão respostas usando conteúdo de todas as fontes de dados integradas.
- * Filtragem de arquivos*: Configure quais arquivos estão incluídos nas digitalizações:
  - Na seção **suporte a tipos de arquivo**, escolha incluir todos os tipos de arquivos ou selecionar tipos de arquivo individuais para inclusão nas verificações de origem de dados.

    Se você incluir imagens ou arquivos PDF, o NetApp Workload Factory for GenAI analisará o texto nas imagens (incluindo imagens em documentos PDF), e isso incorrerá em um custo mais alto.

  Ao incluir dados de texto de imagens, o GenAI não consegue mascarar informações de identificação pessoal (PII) da imagem à medida que os dados de texto digitalizados são enviados do seu ambiente para a AWS. No entanto, uma vez que os dados são armazenados, todas as PII são mascaradas no banco de dados do GenAI.

  Sua escolha de incluir arquivos de imagem em digitalizações está relacionada ao modelo de bate-papo da base de conhecimento. Se você incluir arquivos de imagem em digitalizações, o modelo de bate-papo deve suportar imagens. Se os tipos de arquivo de imagem estiverem selecionados aqui, você não poderá alternar a base de conhecimento para um modelo de chat que não suporte arquivos de imagem.

- Na seção **filtro de tempo de modificação de arquivo**, escolha ativar ou desativar a inclusão de arquivos com base em seu tempo de modificação. Se ativar a filtragem de hora de modificação, selecione um intervalo de datas na lista.

  Se você incluir arquivos com base em um intervalo de datas de modificação, assim que o intervalo de datas não for satisfeito (os arquivos não foram modificados dentro do intervalo de datas especificado), os arquivos serão excluídos da verificação periódica e a fonte de dados não incluirá esses arquivos.

4. Selecione **Adicionar fonte de dados** para adicionar esta fonte de dados à sua base de conhecimento.

Resultado

A fonte de dados está integrada ao seu conetor.

Sincronize as fontes de dados com um conetor

As fontes de dados são sincronizadas automaticamente com o conetor associado uma vez por dia, de modo que quaisquer alterações na fonte de dados sejam refletidas no Amazon Q Business. Se você fizer alterações em qualquer uma de suas fontes de dados e quiser sincronizar (digitalizar) os dados imediatamente, poderá executar uma sincronização sob demanda.

A sincronização é incremental, portanto, o Amazon Q Business só processa os objetos em suas fontes de dados que foram adicionados, modificados ou excluídos desde a última sincronização.

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. No menu Bases de conhecimento e conectores, selecione o conector que você deseja sincronizar.
4. Selecione ••• e selecione **Manage Connector**.
5. Selecione o menu **ações** e selecione **Digitalizar agora**.

   Você verá uma mensagem informando que suas fontes de dados estão sendo digitalizadas e uma mensagem final quando a digitalização estiver concluída.

Resultado

O conetor é sincronizado com as fontes de dados anexadas e o Amazon Q Business começará a usar as informações mais recentes de suas fontes de dados.

Pausar ou retomar uma sincronização agendada

Se pretender pausar ou retomar a próxima sincronização (digitalização) das fontes de dados, pode fazê-lo a qualquer momento. Talvez seja necessário pausar a próxima sincronização agendada se você fizer alterações em uma fonte de dados e não quiser que a sincronização aconteça durante a janela de mudança.

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. Na página de inventário do conetor, selecione o conetor para o qual deseja pausar ou retomar exames.
4. Selecione ••• e selecione **Manage Connector**.
5. Selecione o menu **ações** e selecione **Digitalizar > Pausar digitalização agendada** ou **Digitalizar > Retomar digitalização agendada**.

   Você verá uma mensagem informando que a próxima digitalização agendada foi pausada ou retomada.

Eliminar um conetor

Se você não precisar mais de um conector, poderá excluí-lo. Quando você exclui um conector, ele é removido do Workload Factory e o volume que contém o conector é excluído. A exclusão de um conector não é reversível.

Ao excluir um conetor, você também deve desassociar o conetor de qualquer agente ao qual está associado para excluir totalmente todos os recursos associados ao 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. Na página de inventário bases de conhecimento e conetores, selecione o conetor que pretende eliminar.
4. Selecione ••• e selecione **Manage Connector**.
5. Selecione o menu **ações** e selecione **Excluir conetor**.
6. Na caixa de diálogo Excluir conetor, confirme se deseja excluí-lo e selecione **Excluir**.

Resultado

O conector é removido do Workload Factory e seu volume associado é excluído.

Gerenciar fontes de dados do GenAI

Depois de criar uma base de conhecimento ou um conetor usando fontes de dados no seu sistema de arquivos FSX for ONTAP, você pode visualizar os detalhes da fonte de dados, atualizar ou alterar o conteúdo da fonte de dados, editar configurações da fonte de dados ou excluir a fonte de dados.

Exibir informações sobre uma fonte de dados

Você pode exibir informações sobre o conteúdo de uma fonte de dados e pode visualizar seu status de incorporação com a base de conhecimento ou o conetor. Como as fontes de dados estão associadas a uma base de conhecimento ou conetor, você precisará escolher primeiro a base de conhecimento ou o conetor antes de poder visualizar os detalhes da fonte de dados.

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. Selecione a base de conhecimento ou o conetor onde reside a fonte de dados e, em seguida, •••selecione e selecione **Gerenciar base de conhecimento** ou **Gerenciar conetor**.

   A parte inferior da página lista as fontes de dados associadas.

4. Expanda cada linha selecionando o ⌄ para exibir informações detalhadas sobre cada fonte de dados, como o sistema de arquivos FSX for ONTAP, o volume e o caminho onde reside a fonte de dados.

   Ele também lista as informações de incorporação e se essa fonte de dados está atualmente incorporada na base de conhecimento ou no conetor.

Editar as definições da fonte de dados

Você pode editar informações sobre uma fonte de dados integrada a uma base de conhecimento ou conetor. A maioria das informações é corrigida depois que você adicionou uma fonte de dados, mas você pode fazer alterações em algumas das configurações (como definição de divisão ou reconhecimento de permissão).

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. Na página de inventário das bases de conhecimento, selecione a base de conhecimento onde reside a fonte de dados e, em seguida, ••• selecione e selecione **Gerenciar base de conhecimento**.

   A parte inferior da página lista as fontes de dados que fazem parte desta base de conhecimento.
4. Na linha da fonte de dados que você deseja editar, ••• selecione e selecione **Editar fonte de dados**.
5. Na página Editar fonte de dados, ⌄ selecione para expandir a linha para definição de bloco.
6. Atualize as configurações para a estratégia e configuração de agrupamento, e reconhecimento de permissões (para volumes SMB) e selecione **Salvar**.

Resultado

As configurações da fonte de dados são atualizadas e o sistema de AI sincroniza a fonte de dados de modo que ela seja reindexada à base de conhecimento.

Atualize o conteúdo de uma fonte de dados existente

Você pode alterar o conteúdo de uma fonte de dados a qualquer momento para adicionar ou atualizar seus dados organizacionais. Se essa fonte de dados estiver sendo usada ativamente em uma base de conhecimento, você deve sincronizar a fonte de dados para que ela seja reindexada à base de conhecimento. A sincronização é incremental, portanto, o Amazon bedrock só processa os objetos no volume do FSX for ONTAP que foram adicionados, modificados ou excluídos desde a última sincronização.

As fontes de dados são sincronizadas automaticamente com a base de conhecimento uma vez por dia, para que quaisquer alterações na fonte de dados sejam refletidas no chatbot. Se você fizer alterações em uma fonte de dados e quiser sincronizar os dados imediatamente, poderá "execute uma sincronização sob demanda".

Eliminar uma fonte de dados

Se você não precisar mais de uma fonte de dados para fazer parte da sua base de conhecimento, você pode excluí-la.

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. Na página de inventário da base de conhecimento, selecione a base de conhecimento onde reside a fonte de dados e, em seguida, ••• selecione e selecione **Gerenciar base de conhecimento**.

   A parte inferior da página lista as fontes de dados que fazem parte desta base de conhecimento.
4. Na linha da fonte de dados que você deseja excluir, ••• selecione e selecione **Excluir fonte de dados**.
5. Na caixa de diálogo Excluir fonte de dados, confirme se deseja excluí-la e selecione **Confirm**.

Resultado

A fonte de dados é removida da base de conhecimento e o sistema de IA remove as informações indexadas sobre essa fonte de dados da base de conhecimento. Qualquer informação dessa fonte de dados não estará mais disponível para chatbots que estejam usando a base de conhecimento.

Monitore as operações de carga de trabalho com o Tracker no NetApp Workload Factory

Monitore e acompanhe a execução de operações de carga de trabalho e monitore o progresso das tarefas com o Tracker no NetApp Workload Factory.

Sobre esta tarefa

O NetApp Workload Factory fornece o Tracker, um recurso de monitoramento, para que você possa monitorar e rastrear o progresso e o status das operações de carga de trabalho, revisar detalhes de tarefas e subtarefas de operação e diagnosticar quaisquer problemas ou falhas.

Várias ações estão disponíveis no Tracker. Você pode filtrar tarefas por período de tempo (últimas 24 horas, 7 dias, 14 dias ou 30 dias), carga de trabalho, status e usuário; encontrar trabalhos usando a função de pesquisa; e fazer download da tabela de tarefas como um arquivo CSV. Você pode atualizar o Rastreador a qualquer momento e tentar novamente rapidamente uma operação com falha ou editar parâmetros para uma operação com falha e tentar a operação novamente.

O Tracker suporta dois níveis de monitoramento, dependendo da operação. Cada tarefa, como a implantação do sistema de arquivos, exibe a descrição da tarefa, status, hora de início, duração da tarefa, usuário, região, recurso proxy, ID da tarefa e todas as subtarefas relacionadas. Você pode visualizar as respostas da API para entender o que aconteceu durante a operação.

Monitore níveis de tarefa com exemplos

- Nível 1 (tarefa): Controla a implantação do sistema de arquivos.
- Nível 2 (subtarefa): Controla as subtarefas relacionadas com a implementação do sistema de ficheiros.

Estado da operação

O status da operação no Rastreador é o seguinte *em andamento*, *sucesso* e *falha*.

Frequência de funcionamento

A frequência de funcionamento baseia-se no tipo de trabalho e na programação de trabalhos.

Retenção de eventos

Os eventos são mantidos na interface do usuário por 30 dias.

Monitorizar e monitorizar as operações

Rastreie e monitore operações no console do Workload Factory com o Tracker.

Passos

1. Inicie sessão utilizando uma das "experiências de console".
2. No menu de carga de trabalho, selecione **Administração** e depois selecione **Rastreador**.
3. No Tracker, use os filtros ou pesquise para restringir os resultados dos trabalhos. Você também pode baixar um relatório de empregos.

Exibir solicitação de API

Visualize a solicitação de API na caixa de código para uma tarefa no Tracker.

Passos

1. No Rastreador, selecione uma tarefa.
2. Selecione o menu de ações e depois selecione **Exibir solicitação de API**.

Tente novamente uma operação com falha

Tente novamente uma operação com falha no Tracker. Você também pode copiar a mensagem de erro de uma operação com falha.

Você pode tentar novamente uma operação com falha até 10 vezes.

Passos

1. No Rastreador, selecione uma operação com falha.
2. Selecione o menu de ações e depois selecione **Repetir**.

Resultado

A operação é reiniciada.

Edite e tente novamente uma operação com falha

Edite os parâmetros da operação com falha e tente novamente a operação fora do Rastreador.

Passos

1. No Rastreador, selecione uma operação com falha.
2. Selecione o menu de ações e depois selecione **Editar e tentar novamente**.

   Você é redirecionado para a página de operação, onde você pode editar os parâmetros e tentar novamente a operação.

Resultado

A operação é reiniciada. Aceda a Rastreador para ver o estado da operação.

Conhecimento e apoio

Registre-se para obter suporte para o NetApp Workload Factory para GenAI

O registro de suporte é necessário para receber suporte técnico específico para o NetApp Workload Factory e suas soluções e serviços de armazenamento. Você deve se registrar para obter suporte no NetApp Console, que é um console baseado na Web separado do Workload Factory.

O registro para suporte não habilita o suporte da NetApp para um serviço de arquivo do provedor de nuvem. Para obter suporte técnico relacionado a um serviço de arquivo do provedor de nuvem, sua infraestrutura ou qualquer solução que use o serviço, consulte "Obter ajuda" na documentação do Workload Factory para esse produto.

"Amazon FSX para ONTAP"

Visão geral do Registro de suporte

O registro da assinatura de suporte do ID da sua conta (seu número de série 960xxxxxxxx de 20 dígitos localizado na página Recursos de suporte no NetApp Console) serve como seu único ID de assinatura de suporte. Cada assinatura de suporte em nível de conta da NetApp deve ser registrada.

O registro habilita recursos como abertura de tickets de suporte e geração automática de casos. O registro é concluído adicionando contas do NetApp Support Site (NSS) ao NetApp Console, conforme descrito abaixo.

Registe a sua conta para obter assistência NetApp

Para se registrar para obter suporte e ativar o direito ao suporte, um usuário em sua conta deve associar uma conta do Site de Suporte NetApp ao login do NetApp Console. A maneira como você se registra para o suporte da NetApp depende se você já tem uma conta no NetApp Support Site (NSS).

Cliente existente com uma conta NSS

Se você for um cliente NetApp com uma conta NSS, basta se registrar para receber suporte pelo NetApp Console.

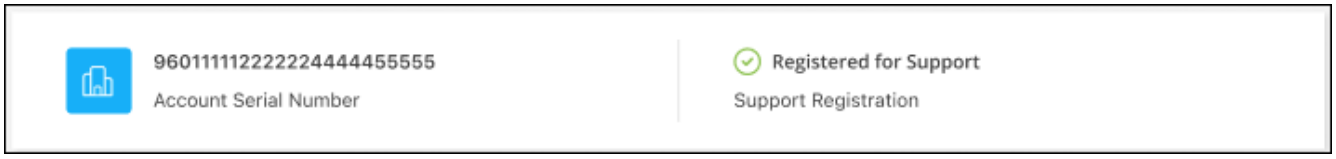
Passos

1. No canto superior direito do console do Workload Factory, selecione **Ajuda > Suporte**.

   Selecionar esta opção abre o NetApp Console em uma nova guia do navegador e carrega o painel de Suporte.

2. No menu do NetApp Console, selecione **Administração** e, em seguida, selecione **Credenciais**.
3. Selecione **credenciais do usuário**.
4. Selecione **Adicionar credenciais NSS** e siga o prompt de autenticação do site de suporte da NetApp (NSS).
5. Para confirmar que o processo de Registro foi bem-sucedido, selecione o ícone Ajuda e selecione **suporte**.

A página **recursos** deve mostrar que sua conta está registrada para suporte.



Observe que outros usuários do NetApp Console não verão o mesmo status de registro de suporte se não tiverem associado uma conta do NetApp Support Site ao login do NetApp Console. No entanto, isso não significa que sua conta NetApp não esteja registrada para suporte. Desde que um usuário na conta tenha seguido essas etapas, sua conta foi registrada.

Cliente existente, mas sem conta NSS

Se você já for um cliente da NetApp com licenças e números de série, mas *nenhuma* conta NSS, será necessário criar uma conta NSS e associá-la ao seu login do NetApp Console.

Passos

1. Crie uma conta do site de suporte da NetApp preenchendo o. "Formulário de Registro do usuário do site de suporte da NetApp"
   a. Certifique-se de selecionar o nível de usuário apropriado, que normalmente é **Cliente NetApp/Usuário final**.
   b. Certifique-se de copiar o número de série da conta NetApp (960xxxx) usado acima para o campo de número de série. Isso acelerará o processamento da conta.
2. Associe sua nova conta NSS ao login do NetApp Console concluindo as etapas emCliente existente com uma conta NSS .

Novo na NetApp

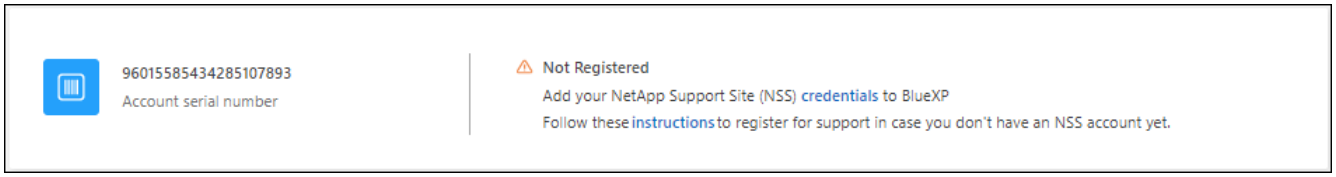
Se você é novo no NetApp e não tem uma conta NSS, siga cada passo abaixo.

Passos

1. No canto superior direito do console do Workload Factory, selecione **Ajuda > Suporte**.

   Selecionar esta opção abre o NetApp Console em uma nova guia do navegador e carrega o painel de Suporte.

2. Localize o número de série da ID da conta na página recursos de suporte.



3. Navegue "Site de Registro de suporte da NetApp" e selecione **não sou um Cliente NetApp registado**.
4. Preencha os campos obrigatórios (aqueles com asteriscos vermelhos).
5. No campo **linha de produtos**, selecione **Cloud Manager** e, em seguida, selecione seu provedor de cobrança aplicável.
6. Copie o número de série da sua conta a partir da etapa 2 acima, complete a verificação de segurança e confirme se leu a Política de Privacidade de dados globais da NetApp.

Um e-mail é enviado imediatamente para a caixa de correio fornecida para finalizar esta transação segura. Certifique-se de verificar suas pastas de spam se o e-mail de validação não chegar em poucos minutos.

7. Confirme a ação a partir do e-mail.

   A confirmação envia sua solicitação à NetApp e recomenda que você crie uma conta do site de suporte da NetApp.

8. Crie uma conta do site de suporte da NetApp preenchendo o. "Formulário de Registro do usuário do site de suporte da NetApp"
   a. Certifique-se de selecionar o nível de usuário apropriado, que normalmente é **Cliente NetApp/Usuário final**.
   b. Certifique-se de copiar o número de série da conta (960xxxx) usado acima para o campo de número de série. Isto irá acelerar o processamento da conta.

Depois de terminar

O NetApp deve entrar em Contato com você durante esse processo. Este é um exercício de integração única para novos usuários.

Depois de ter sua conta do Site de Suporte NetApp , associe a conta ao seu login do Console NetApp concluindo as etapas emCliente existente com uma conta NSS .

Solução de problemas do GenAI

Aprenda a contornar alguns problemas comuns que você pode encontrar.

Problemas e soluções comuns

Se você tiver um desses problemas, use as etapas na coluna solução alternativa para tentar resolvê-lo.

Área	Problema	Causa	Solução alternativa
Implantação	A implantação falha porque o volume já existe.	O NetApp Workload Factory for GenAI precisa criar um novo volume durante o processo de implantação, mas já existe um volume usando o nome que você especificou.	Especifique um nome exclusivo a ser usado para o novo volume e tente implantar novamente.
Implantação	A implantação falha porque o NetApp Workload Factory para GenAI não consegue montar o volume.	Uma ou mais portas de entrada necessárias para o FSX for NetApp ONTAP são fechadas ou filtradas.	Abra as seguintes portas de entrada:

| Protocolo | Porta | Finalidade

| Todo o ICMP | Tudo | Fazer ping na instância

| HTTPS | 443 | Acesso do conetor ao LIF de gerenciamento fsxadmin para enviar chamadas de API para o

FSX

| SSH | 22 | Acesso SSH ao endereço IP do LIF de gerenciamento de cluster ou um LIF de gerenciamento de nó

| TCP | 111 | Chamada de procedimento remoto para NFS

| TCP | 139 | Sessão de serviço NetBIOS para CIFS

| TCP | 161-162 | Protocolo de gerenciamento de rede simples

| TCP | 445 | Microsoft SMB/CIFS sobre TCP com enquadramento NetBIOS

| TCP | 635 | Montagem em NFS

| TCP | 749 | Kerberos

| TCP | 2049 | Daemon do servidor NFS

| TCP | 3260 | Acesso iSCSI através do iSCSI data LIF

| TCP | 4045 | Daemon de bloqueio NFS

| TCP | 4046 | Monitor de status da rede para NFS

| TCP | 10000 | Backup usando NDMP

| TCP | 11104 | Gestão de sessões de comunicação entre clusters para SnapMirror

| TCP | 11105 | Transferência de dados SnapMirror usando LIFs entre clusters

| UDP | 111 | Chamada de procedimento remoto para NFS

| UDP | 161-162 | Protocolo de gerenciamento de rede simples

| UDP | 635 | Montagem em NFS

| UDP | 2049 | Daemon do servidor NFS

| UDP | 4045 | Daemon de bloqueio NFS

| UDP | 4046 | Monitor de status da rede para NFS

| UDP | 4049 | Protocolo rquotad NFS

Manutenção	O mecanismo de IA falha ao iniciar e você vê o erro "erro de instância do mecanismo de IA" na página **bases de conhecimento**.	A instância do mecanismo de IA foi corrompida ou não existe.	Selecione o botão **Reconstruir**. O NetApp Workload Factory para GenAI reconstrói a infraestrutura e exibe o progresso da reconstrução. Quando concluído, suas bases de conhecimento serão reconectadas à infraestrutura reconstruída e a lista de bases de conhecimento será exibida.
Manutenção	O mecanismo de IA falha ao iniciar, e você vê o erro "a instância do mecanismo GenAI está parada" na página **bases de conhecimento**.	A instância do mecanismo de IA não está em execução.	Use o Console de Gerenciamento da AWS ou a CLI da AWS para iniciar a instância do mecanismo de IA.
Manutenção	O mecanismo de IA não arranca e vê o erro "o servidor do motor GenAI não está a responder" na página **bases de dados de conhecimento**.	A instância do mecanismo de IA não está respondendo.	Use as seguintes etapas de recuperação: **Passos** 1. Modifique o grupo de segurança da instância do mecanismo GenAI para habilitar o acesso SSH à instância do mecanismo GenAI. 2. Faça login na instância usando SSH. 3. Execute o seguinte comando: `docker-compose up`

Manutenção	A instância do Docker de backend usada pelo NetApp Workload Factory para GenAI falhou ao iniciar.	O volume foi excluído e a instância EC2 foi reiniciada.	Use as seguintes etapas de recuperação:
			
			Passos
			1. Crie um novo volume no FSX for NetApp ONTAP. Por exemplo, o nome do volume pode ser `netapp_ai` e o caminho do volume pode ser `/netapp_ai`.
			2. SSH para a instância do Amazon EC2.
			3. Listar os volumes:
			`docker volume list`
			4. Retire o volume antigo:
			`docker volume rm ec2-user_persistent_folder`
			5. Abra o `docker-compose.yml` arquivo usando um editor de texto.
			6. `volumes`Na secção , altere o caminho do dispositivo para o novo caminho do volume. Por exemplo:

Manutenção	A instância do Docker de backend usada pelo NetApp Workload Factory para GenAI falhou ao iniciar.	O volume raiz foi excluído.	Crie um volume com um nome e um caminho e reinicie a instância do Docker de back-end do Amazon EC2.
Manutenção	A instância do Docker de backend usada pelo NetApp Workload Factory para GenAI falhou ao iniciar.	O volume raiz foi excluído.	Crie um volume com um nome e um caminho e reinicie a instância do Docker de back-end do Amazon EC2.

```
"addr=svm-
[illegible]
d8a72.\
fs-
[illegible]
[illegible].\
fsx.us-east-
1.amazonaws.
com,nolock,s
oft,rw"
device:
':/netapp_ai
' # Path to
new volume
```

Obtenha ajuda com o NetApp Workload Factory para GenAI

A NetApp fornece suporte para o Workload Factory e seus serviços de nuvem de diversas maneiras. Há diversas opções gratuitas de autoatendimento disponíveis 24 horas por dia, 7 dias por semana, como artigos da base de conhecimento (KB) e um fórum da comunidade. Seu cadastro no suporte inclui suporte técnico remoto por meio de tickets online.

Obtenha suporte para o FSX for ONTAP

Para obter suporte técnico relacionado ao FSx for ONTAP, sua infraestrutura ou qualquer solução que use o serviço, consulte "Obter ajuda" na documentação do Workload Factory para esse produto.

"Amazon FSX para ONTAP"

Para receber suporte técnico específico para o Workload Factory e suas soluções e serviços de armazenamento, use as opções de suporte descritas abaixo.

Use opções de suporte autônomo

Estas opções estão disponíveis gratuitamente, 24 horas por dia, 7 dias por semana:

- Documentação

  A documentação do Workload Factory que você está visualizando no momento.

- "Base de conhecimento"

  Pesquise na base de conhecimento do Workload Factory para encontrar artigos úteis para solucionar problemas.

- "Comunidades"

  Junte-se à comunidade do Workload Factory para acompanhar discussões em andamento ou criar novas.

Crie um caso com o suporte do NetApp

Além das opções de suporte autônomo acima, você pode trabalhar com um especialista de suporte da NetApp para resolver quaisquer problemas depois de ativar o suporte.

Antes de começar

Para usar o recurso **Criar um caso**, você deve primeiro se registrar para obter suporte. Associe suas credenciais do site de suporte da NetApp ao seu login do Workload Factory. "Saiba como se inscrever para obter suporte" .

Passos

1. No canto superior direito do console do Workload Factory, selecione **Ajuda > Suporte**.

   Selecionar esta opção abre o NetApp Console em uma nova guia do navegador e carrega o painel de Suporte.

2. Na página **recursos**, escolha uma das opções disponíveis em suporte técnico:

   a. Selecione **Ligue para nós** se quiser falar com alguém no telefone. Você será direcionado para uma página no NetApp.com que lista os números de telefone que você pode ligar.

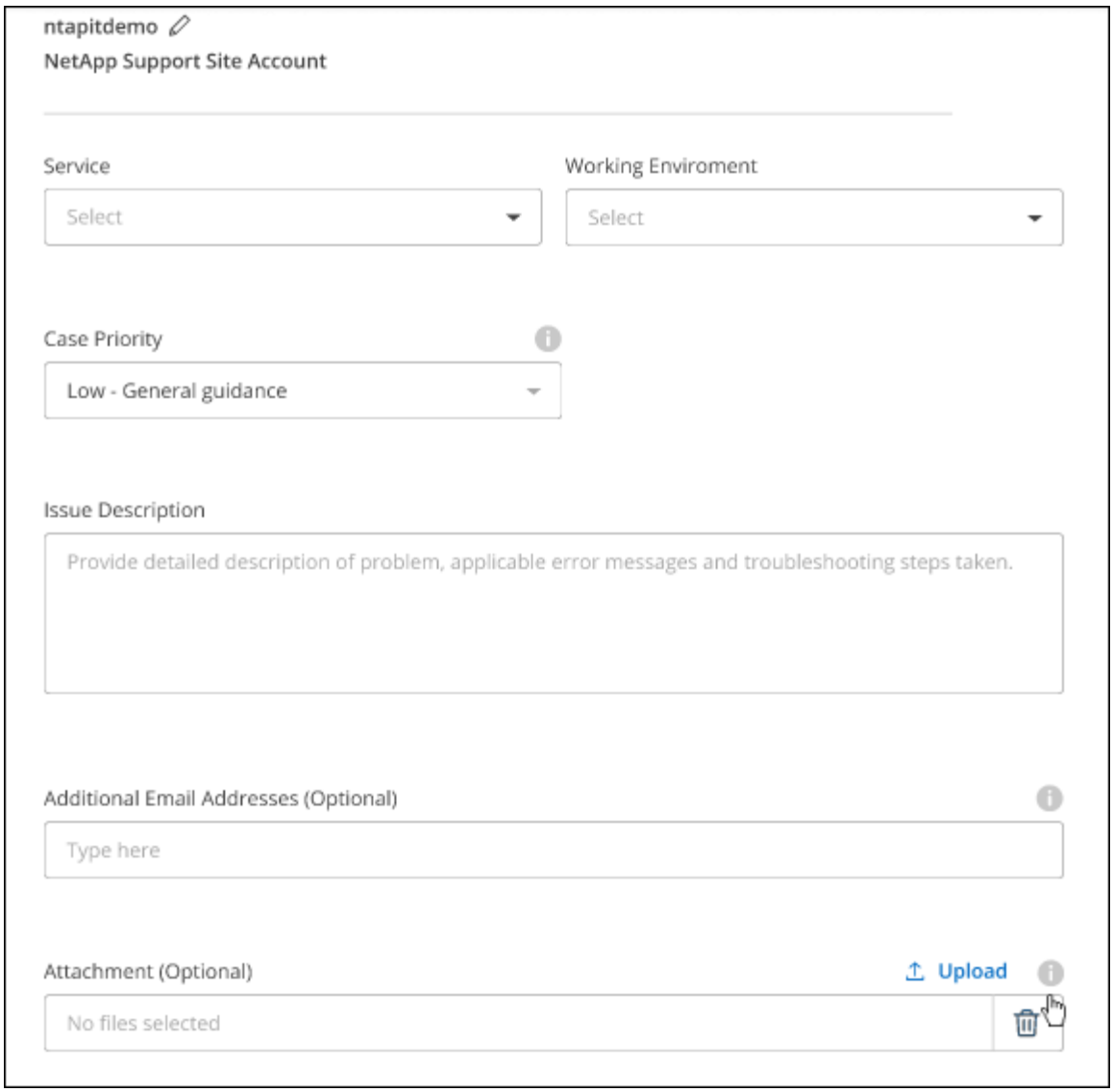
   b. Selecione **criar um caso** para abrir um ticket com um especialista em suporte da NetApp:

      - **Serviço**: Selecione **fábrica de carga de trabalho**.
      - **Prioridade do caso**: Escolha a prioridade para o caso, que pode ser baixa, média, alta ou Crítica.

        Para saber mais detalhes sobre essas prioridades, passe o Mouse sobre o ícone de informações ao lado do nome do campo.

      - **Descrição do problema**: Forneça uma descrição detalhada do seu problema, incluindo quaisquer mensagens de erro aplicáveis ou etapas de solução de problemas que você executou.
      - **Endereços de e-mail adicionais**: Insira endereços de e-mail adicionais se você quiser que outra pessoa saiba sobre esse problema.
      - **Anexo (Opcional)**: Carregue até cinco anexos, um de cada vez.

        Os anexos estão limitados a 25 MB por ficheiro. As seguintes extensões de arquivo são suportadas: txt, log, pdf, jpg/jpeg, rtf, doc/docx, xls/xlsx e csv.



Depois de terminar

Um pop-up aparecerá com o número do seu caso de suporte. Um especialista em suporte da NetApp irá rever o seu caso e voltar para você em breve.

Para obter um histórico de seus casos de suporte, você pode selecionar **Configurações > linha do tempo** e procurar ações chamadas "criar caso de suporte". Um botão à direita permite expandir a ação para ver detalhes.

É possível que você encontre a seguinte mensagem de erro ao tentar criar um caso:

"Você não está autorizado a criar um caso contra o serviço selecionado"

Esse erro pode significar que a conta NSS e a empresa registrada à qual ela está associada não são a mesma empresa registrada para o número de série da conta do NetApp Console (por exemplo, 960xxxx) ou o número de série do sistema. Você pode buscar assistência usando uma das seguintes opções:

- Use o chat no produto
- Envie um caso não técnico em https://mysupport.netapp.com/site/help

Gerenciar seus casos de suporte (prévia)

Você pode visualizar e gerenciar casos de suporte ativos e resolvidos diretamente do NetApp Console. Você pode gerenciar os casos associados à sua conta NSS e à sua empresa.

O gerenciamento de casos está disponível como uma prévia. Planejamos refinar essa experiência e adicionar melhorias nos próximos lançamentos. Por favor, envie-nos feedback usando o chat no produto.

Observe o seguinte:

- O painel de gerenciamento de casos na parte superior da página oferece duas visualizações:
  - A vista à esquerda mostra o total de casos abertos nos últimos 3 meses pela conta do usuário NSS que você forneceu.
  - A visualização à direita mostra o total de casos abertos nos últimos 3 meses ao nível da sua empresa com base na sua conta NSS de utilizador.

  Os resultados na tabela refletem os casos relacionados à exibição selecionada.
- Você pode adicionar ou remover colunas de interesse e pode filtrar o conteúdo de colunas como prioridade e Status. Outras colunas fornecem apenas capacidades de ordenação.

  Veja os passos abaixo para obter mais detalhes.
- Em um nível por caso, oferecemos a capacidade de atualizar notas de caso ou fechar um caso que ainda não esteja no status fechado ou pendente fechado.

Passos

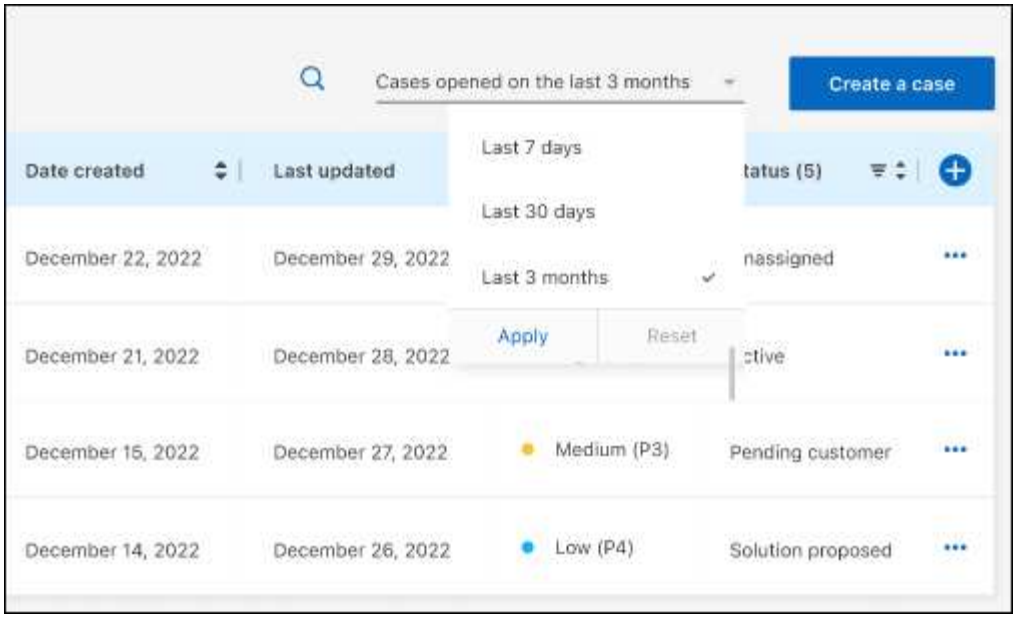
1. No canto superior direito do console do Workload Factory, selecione **Ajuda > Suporte**.

   Selecionar esta opção abre o NetApp Console em uma nova guia do navegador e carrega o painel de Suporte.

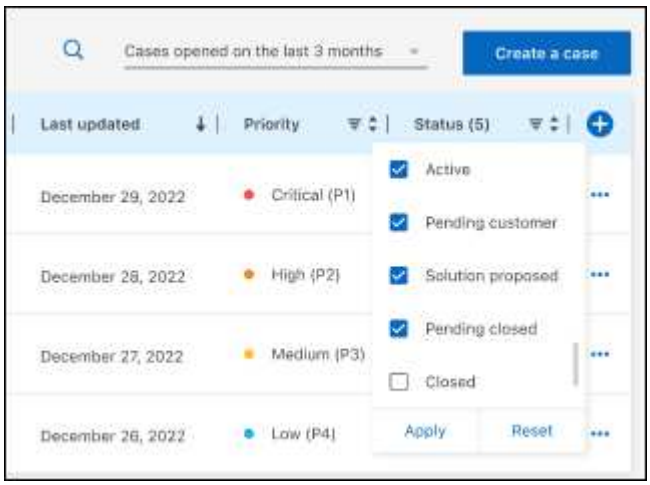
2. Selecione **Gerenciamento de casos** e, se solicitado, adicione sua conta NSS ao NetApp Console.

   A página **Gerenciamento de casos** mostra casos abertos relacionados à conta NSS associada à sua conta de usuário do NetApp Console. Esta é a mesma conta NSS que aparece no topo da página **Gerenciamento NSS**.

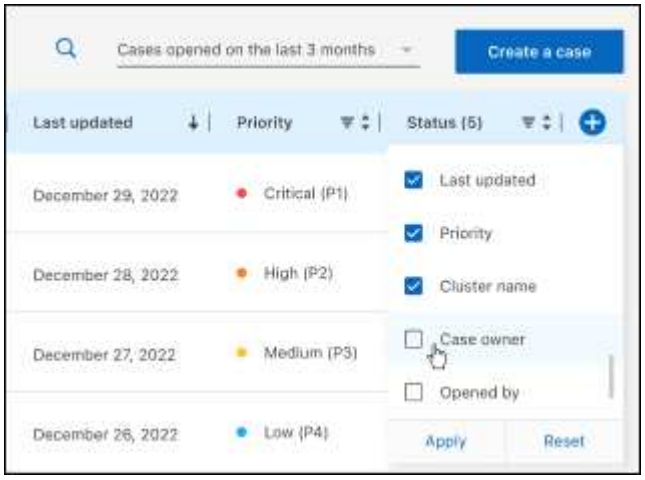
3. Opcionalmente, modifique as informações exibidas na tabela:
   - Em **casos da organização**, selecione **Exibir** para ver todos os casos associados à sua empresa.
   - Modifique o intervalo de datas escolhendo um intervalo de datas exato ou escolhendo um intervalo de tempo diferente.



◦ Filtre o conteúdo das colunas.



◦ Altere as colunas que aparecem na tabela selecionando [O ícone de mais que aparece na tabela] e escolhendo as colunas que você deseja exibir.



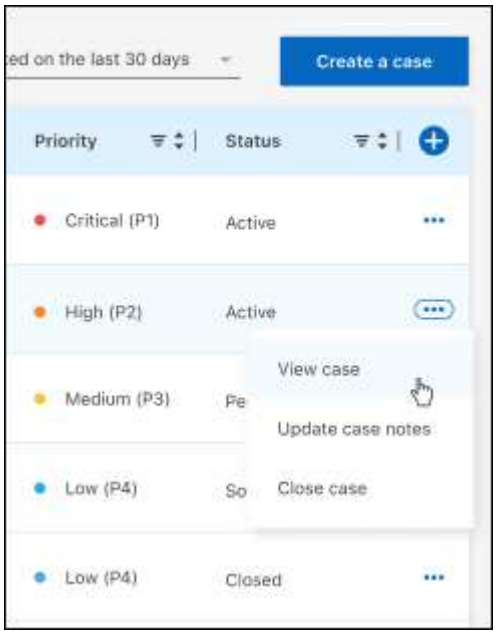
4. Gerencie um caso existente [Um ícone com três pontos que aparece na última coluna da

tabela]selecionando e selecionando uma das opções disponíveis:

- **Ver caso**: Veja detalhes completos sobre um caso específico.
- * Atualizar notas de caso*: Forneça detalhes adicionais sobre o seu problema ou selecione **carregar arquivos** para anexar até um máximo de cinco arquivos.

  Os anexos estão limitados a 25 MB por ficheiro. As seguintes extensões de arquivo são suportadas: txt, log, pdf, jpg/jpeg, rtf, doc/docx, xls/xlsx e csv.

- * Fechar caso*: Forneça detalhes sobre por que você está fechando o caso e selecione **Fechar caso**.



Avisos legais do NetApp Workload Factory para GenAI

Avisos legais fornecem acesso a declarações de direitos autorais, marcas registradas, patentes e muito mais.

Direitos de autor

"https://www.netapp.com/company/legal/copyright/"

Marcas comerciais

NetApp, o logotipo DA NetApp e as marcas listadas na página de marcas comerciais da NetApp são marcas comerciais da NetApp, Inc. Outros nomes de produtos e empresas podem ser marcas comerciais de seus respetivos proprietários.

"https://www.netapp.com/company/legal/trademarks/"

Patentes

Uma lista atual de patentes de propriedade da NetApp pode ser encontrada em:

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

Política de privacidade

"https://www.netapp.com/company/legal/privacy-policy/"

Código aberto

Os arquivos de aviso fornecem informações sobre direitos autorais de terceiros e licenças usadas no software NetApp.

"Fábrica de carga de trabalho da NetApp"

Informações sobre direitos autorais

Informações sobre marcas comerciais