NetApp AIPod Mini - Inferência RAG empresarial com NetApp e Intel

NetApp artificial intelligence solutions

NetApp
August 20, 2025

This PDF was generated from https://docs.netapp.com/pt-br/netapp-solutions-ai/infra/ai-minipod.html on August 20, 2025. Always check docs.netapp.com for the latest.

Índice

NetApp AIPod Mini - Inferência RAG empresarial com NetApp e Intel . . . 1
Sumário executivo . . . 1
Validação de parceiros de armazenamento Intel . . . 1
Vantagens de executar sistemas RAG com NetApp . . . 1
Público-alvo . . . 2
Requisitos de tecnologia . . . 2
Hardware . . . 2
Software . . . 4
Implantação da solução . . . 5
Pilha de software . . . 5
Etapas de implantação . . . 6
Habilitar o uso do ONTAP S3 . . . 11
Configurar as configurações de sincronização agendada . . . 11
Alterar modos de acesso de volume . . . 12
(Opcional) Desabilitar verificação de certificado SSL . . . 12
Orientação de dimensionamento . . . 14
Conclusão . . . 15
Reconhecimento . . . 15
Lista de materiais . . . 15
Onde encontrar informações adicionais . . . 17

NetApp AIPod Mini - Inferência RAG empresarial com NetApp e Intel

Este artigo apresenta um projeto de referência validado do NetApp AIPod para Enterprise RAG com tecnologias e recursos combinados de processadores Intel Xeon 6 e soluções de gerenciamento de dados NetApp . A solução demonstra um aplicativo ChatQnA downstream que aproveita um grande modelo de linguagem, fornecendo respostas precisas e contextualmente relevantes para usuários simultâneos. As respostas são recuperadas do repositório de conhecimento interno de uma organização por meio de um pipeline de inferência RAG isolado.

Sathish Thyagarajan, Michael Oglesby, NetApp

Sumário executivo

Um número crescente de organizações está aproveitando aplicativos de geração aumentada de recuperação (RAG) e modelos de grande linguagem (LLMs) para interpretar prompts do usuário e gerar respostas para aumentar a produtividade e o valor comercial. Esses prompts e respostas podem incluir texto, código, imagens ou até mesmo estruturas de proteínas terapêuticas recuperadas da base de conhecimento interna de uma organização, data lakes, repositórios de código e repositórios de documentos. Este artigo aborda o design de referência da solução NetApp AIPod Mini, que inclui armazenamento NetApp AFF e servidores com processadores Intel Xeon 6. Inclui o software de gerenciamento de dados NetApp ONTAP combinado com o Intel Advanced Matrix Extensions (Intel AMX) e o software Intel AI for Enterprise Retrieval-Augmented Generation (RAG) desenvolvido na Open Platform for Enterprise AI (OPEA). O NetApp AIPod Mini para RAG empresarial permite que as organizações ampliem um LLM público em uma solução de inferência de IA generativa privada (GenAI). A solução demonstra inferência RAG eficiente e econômica em escala empresarial, projetada para aumentar a confiabilidade e fornecer melhor controle sobre suas informações proprietárias.

Validação de parceiros de armazenamento Intel

Servidores equipados com processadores Intel Xeon 6 são desenvolvidos para lidar com cargas de trabalho exigentes de inferência de IA, usando Intel AMX para desempenho máximo. Para permitir desempenho e escalabilidade ideais de armazenamento, a solução foi validada com sucesso usando o NetApp ONTAP, permitindo que as empresas atendam às necessidades dos aplicativos RAG. Esta validação foi realizada em servidores com processadores Intel Xeon 6. A Intel e a NetApp têm uma forte parceria focada em fornecer soluções de IA otimizadas, escaláveis e alinhadas aos requisitos de negócios do cliente.

Vantagens de executar sistemas RAG com NetApp

As aplicações RAG envolvem a recuperação de conhecimento dos repositórios de documentos das empresas em vários tipos, como PDF, texto, CSV, Excel ou gráficos de conhecimento. Esses dados normalmente são armazenados em soluções como armazenamento de objetos S3 ou NFS local como fonte de dados. A NetApp é líder em tecnologias de gerenciamento de dados, mobilidade de dados, governança de dados e segurança de dados em todo o ecossistema de borda, data center e nuvem. O gerenciamento de dados do NetApp

ONTAP fornece armazenamento de nível empresarial para dar suporte a vários tipos de cargas de trabalho de IA, incluindo inferência em lote e em tempo real, e oferece alguns dos seguintes benefícios:

- Velocidade e escalabilidade. Você pode manipular grandes conjuntos de dados em alta velocidade para controle de versão com a capacidade de dimensionar o desempenho e a capacidade de forma independente.
- Acesso a dados. O suporte multiprotocolo permite que aplicativos cliente leiam dados usando os protocolos de compartilhamento de arquivos S3, NFS e SMB. Os buckets NAS ONTAP S3 podem facilitar o acesso a dados em cenários de inferência LLM multimodal.
- Confiabilidade e confidencialidade. O ONTAP fornece proteção de dados, NetApp Autonomous Ransomware Protection (ARP) integrada e provisionamento dinâmico de armazenamento, além de oferecer criptografia baseada em software e hardware para aumentar a confidencialidade e a segurança. O ONTAP é compatível com FIPS 140-2 para todas as conexões SSL.

Público-alvo

Este documento é destinado a tomadores de decisão de IA, engenheiros de dados, líderes empresariais e executivos departamentais que desejam aproveitar uma infraestrutura criada para fornecer soluções empresariais de RAG e GenAI. Conhecimento prévio de inferência de IA, LLMs, Kubernetes e redes e seus componentes ajudará durante a fase de implementação.

Requisitos de tecnologia

Hardware

Tecnologias de IA da Intel

Com o Xeon 6 como CPU host, os sistemas acelerados se beneficiam de alto desempenho de thread único; maior largura de banda de memória; maior confiabilidade, disponibilidade e capacidade de manutenção (RAS); e mais faixas de E/S. O Intel AMX acelera a inferência para INT8 e BF16 e oferece suporte para modelos treinados em FP16, com até 2.048 operações de ponto flutuante por ciclo por núcleo para INT8 e 1.024 operações de ponto flutuante por ciclo por núcleo para BF16/FP16. Para implantar uma solução RAG usando processadores Xeon 6, geralmente é recomendado um mínimo de 250 GB de RAM e 500 GB de espaço em disco. No entanto, isso depende muito do tamanho do modelo LLM. Para obter mais informações, consulte o site da Intel "Processador Xeon 6" resumo do produto.

Figura 1 - Servidor de computação com processadores Intel Xeon

6

Armazenamento NetApp AFF

Os sistemas NetApp AFF A-Series de nível básico e médio oferecem desempenho mais potente, densidade e maior eficiência. Os sistemas NetApp AFF A20, AFF A30 e AFF A50 fornecem armazenamento verdadeiramente unificado que suporta blocos, arquivos e objetos, com base em um único sistema operacional que pode gerenciar, proteger e mobilizar dados para aplicativos RAG com o menor custo em nuvem híbrida.

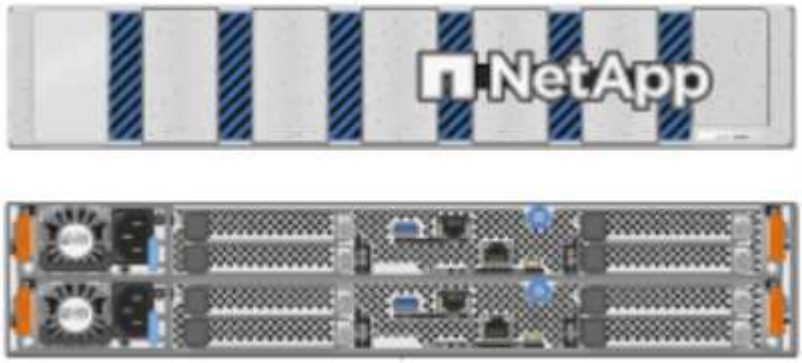


Figura 2 - Sistema NetApp AFF Série A.

Hardware	Quantidade	Comentário
Servidor baseado em Intel Xeon 6	2	Nós de inferência RAG — com processadores Intel Xeon série 6900 ou Intel Xeon série 6700 de soquete duplo e 250 GB a 3 TB de RAM com DDR5 (6400 MHz) ou MRDIMM (8800 MHz). Servidor 2U.
Servidor de plano de controle com processador Intel	1	Plano de controle do Kubernetes/servidor 1U.

Hardware	Quantidade	Comentário
Escolha de switch Ethernet de 100 Gb	1	Switch de data center.
NetApp AFF A20 (ou AFF A30; AFF A50)	1	Capacidade máxima de armazenamento: 9,3 PB. Observação: Rede: portas 10/25/100 GbE.

Para a validação deste projeto de referência, foram utilizados servidores com processadores Intel Xeon 6 da Supermicro (222HA-TN-OTO-37) e um switch 100GbE da Arista (7280R3A).

Software

Plataforma aberta para IA empresarial

A Open Platform for Enterprise AI (OPEA) é uma iniciativa de código aberto liderada pela Intel em colaboração com parceiros do ecossistema. Ele fornece uma plataforma modular de blocos de construção componíveis, projetados para acelerar o desenvolvimento de sistemas de IA generativa de ponta, com forte foco em RAG. O OPEA inclui uma estrutura abrangente com LLMs, armazenamentos de dados, mecanismos de prompt, projetos arquitetônicos RAG e um método de avaliação de quatro etapas que avalia sistemas de IA generativa com base em desempenho, recursos, confiabilidade e prontidão empresarial.

Em sua essência, a OPEA compreende dois componentes principais:

- GenAIComps: um kit de ferramentas baseado em serviços composto por componentes de microsserviços
- GenAIExamples: soluções prontas para implantação, como ChatQnA, que demonstram casos de uso práticos

Para mais detalhes, consulte o "Documentação do Projeto OPEA"

Intel AI para inferência empresarial com tecnologia OPEA

OPEA para Intel AI for Enterprise RAG simplifica a transformação de dados corporativos em insights práticos. Equipado com processadores Intel Xeon, ele integra componentes de parceiros do setor para oferecer uma abordagem simplificada para implantação de soluções empresariais. Ele se adapta perfeitamente a estruturas de orquestração comprovadas, proporcionando a flexibilidade e as opções que sua empresa precisa.

Com base na OPEA, o Intel AI for Enterprise RAG amplia essa base com recursos importantes que melhoram a escalabilidade, a segurança e a experiência do usuário. Esses recursos incluem capacidades de malha de serviço para integração perfeita com arquiteturas modernas baseadas em serviços, validação pronta para produção para confiabilidade de pipeline e uma interface de usuário rica em recursos para RAG como serviço, permitindo fácil gerenciamento e monitoramento de fluxos de trabalho. Além disso, o suporte da Intel e de parceiros fornece acesso a um amplo ecossistema de soluções, combinado com Gerenciamento de Identidade e Acesso (IAM) integrado com interface de usuário e aplicativos para operações seguras e compatíveis. Os guardrails programáveis fornecem controle detalhado sobre o comportamento do pipeline, permitindo configurações personalizadas de segurança e conformidade.

NetApp ONTAP

O NetApp ONTAP é a tecnologia fundamental que sustenta as soluções críticas de armazenamento de dados da NetApp. O ONTAP inclui vários recursos de gerenciamento e proteção de dados, como proteção automática contra ransomware contra ataques cibernéticos, recursos integrados de transporte de dados e

recursos de eficiência de armazenamento. Esses benefícios se aplicam a uma variedade de arquiteturas, desde locais até multicloud híbrida em NAS, SAN, objeto e armazenamento definido por software para implantações de LLM. Você pode usar um servidor de armazenamento de objetos ONTAP S3 em um cluster ONTAP para implantar aplicativos RAG, aproveitando as eficiências de armazenamento e a segurança do ONTAP, fornecidas por usuários autorizados e aplicativos clientes. Para mais informações, consulte "Saiba mais sobre a configuração do ONTAP S3"

NetApp Trident

O software NetApp Trident é um orquestrador de armazenamento de código aberto e totalmente compatível para contêineres e distribuições Kubernetes, incluindo o Red Hat OpenShift. O Trident funciona com todo o portfólio de armazenamento da NetApp , incluindo o NetApp ONTAP , e também oferece suporte a conexões NFS e iSCSI. Para mais informações, consulte "NetApp Trident no Git"

Software	Versão	Comentário
OPEA para Intel AI para Enterprise RAG	1.1.2	Plataforma empresarial RAG baseada em microsserviços OPEA
Interface de armazenamento de contêiner (driver CSI)	NetApp Trident 25.02	Permite provisionamento dinâmico, cópias do NetApp Snapshot e volumes.
Ubuntu	22.04.5	SO em cluster de dois nós
Orquestração de contêineres	Kubernetes 1.31.4	Ambiente para executar o framework RAG
ONTAP	ONTAP 9.16.1P4	Sistema operacional de armazenamento no AFF A20. Possui Vscan e ARP.

Implantação da solução

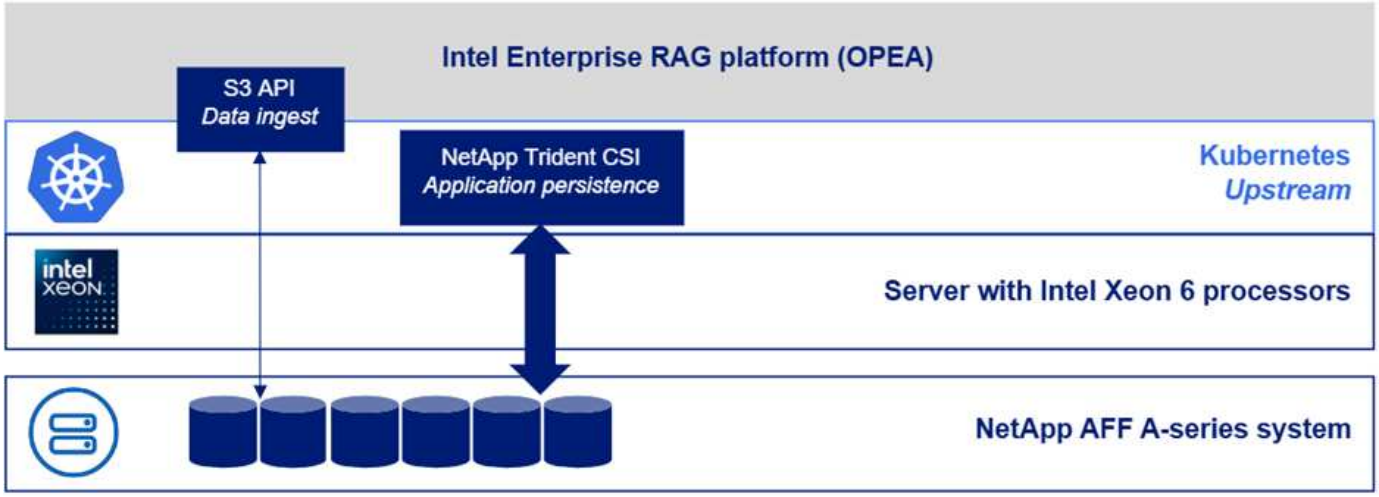
Pilha de software

A solução é implantada em um cluster Kubernetes que consiste em nós de aplicativos baseados em Intel Xeon. Pelo menos três nós são necessários para implementar alta disponibilidade básica para o plano de controle do Kubernetes. Validamos a solução usando o seguinte layout de cluster.

Tabela 3 - Layout do cluster Kubernetes

Nó	Papel	Quantidade
Servidores com processadores Intel Xeon 6 e 1 TB de RAM	Nó de aplicativo, nó do plano de controle	2
Servidor genérico	Nó do plano de controle	1

A figura a seguir descreve uma "visão da pilha de software" da solução.



Etapas de implantação

Implantar dispositivo de armazenamento ONTAP

Implante e provisione seu dispositivo de armazenamento NetApp ONTAP . Consulte o "Documentação dos sistemas de hardware ONTAP" para mais detalhes.

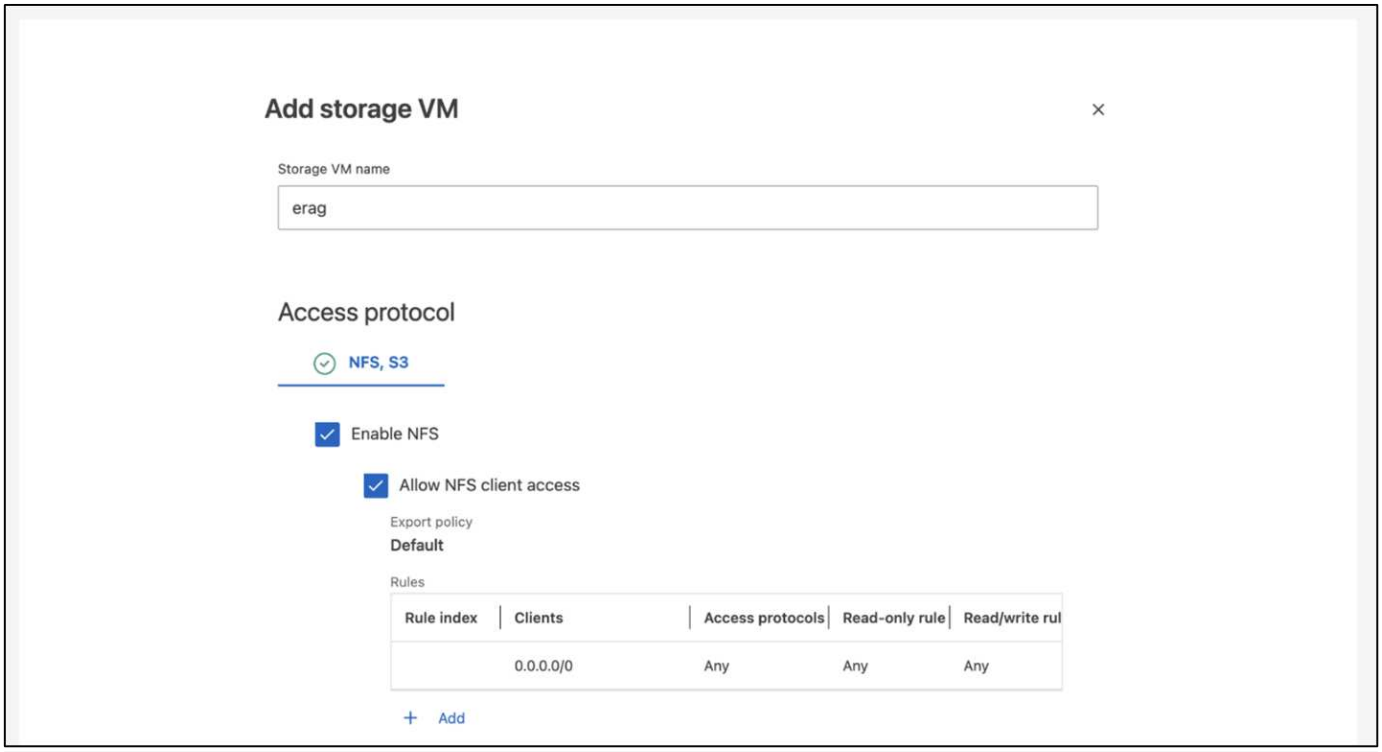
Configurar um ONTAP SVM para acesso NFS e S3

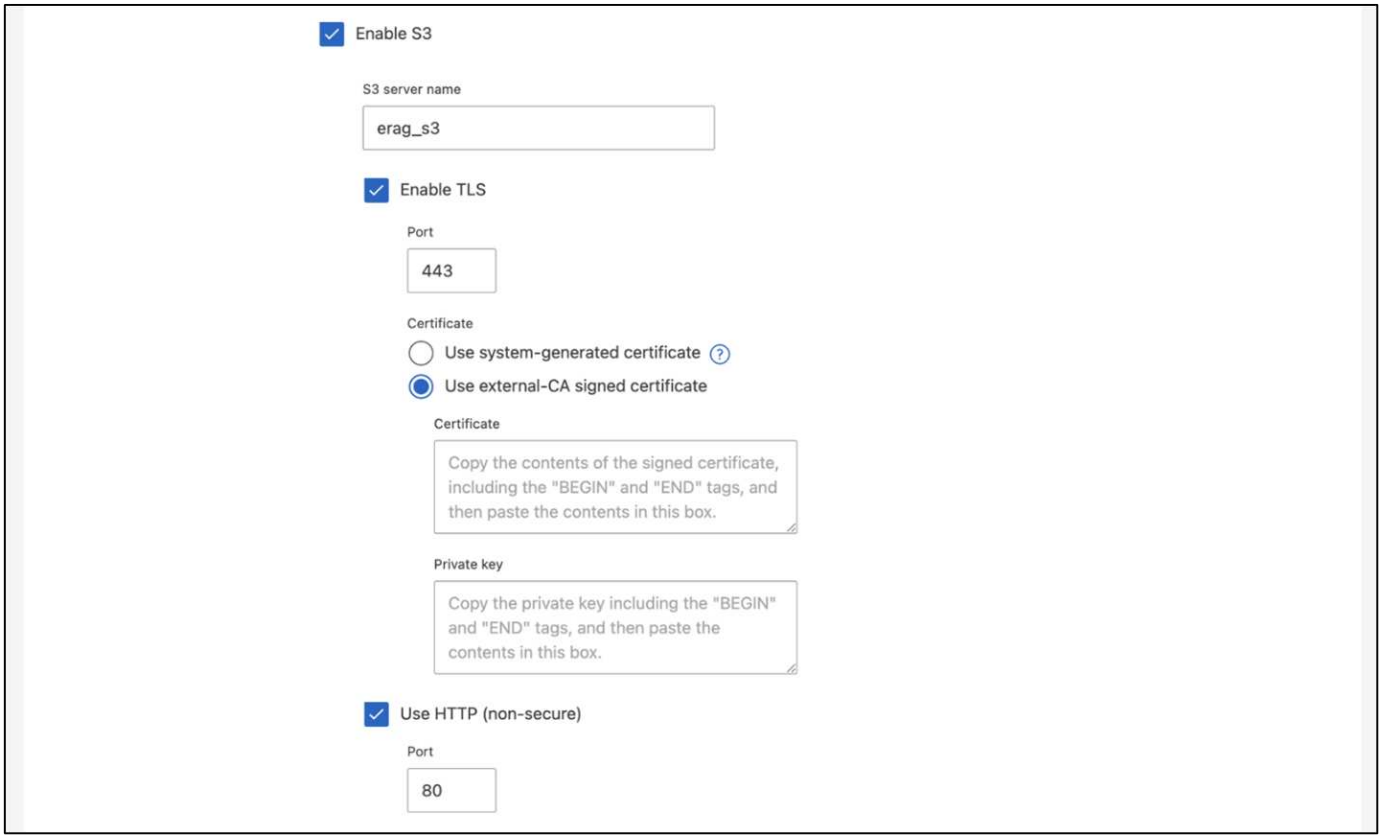
Configure uma máquina virtual de armazenamento ONTAP (SVM) para acesso NFS e S3 em uma rede que seja acessível pelos seus nós do Kubernetes.

Para criar uma SVM usando o ONTAP System Manager, navegue até Armazenamento > VMs de armazenamento e clique no botão + Adicionar. Ao habilitar o acesso S3 para sua SVM, escolha a opção de usar um certificado assinado por uma CA (autoridade de certificação) externa, não um certificado gerado pelo sistema. Você pode usar um certificado autoassinado ou um certificado assinado por uma CA publicamente confiável. Para obter detalhes adicionais, consulte o "Documentação do ONTAP ."

A captura de tela a seguir descreve a criação de um SVM usando o ONTAP System Manager. Modifique os detalhes conforme necessário com base no seu ambiente.

Figura 4 - Criação de SVM usando o ONTAP System Manager.





Configurar permissões do S3

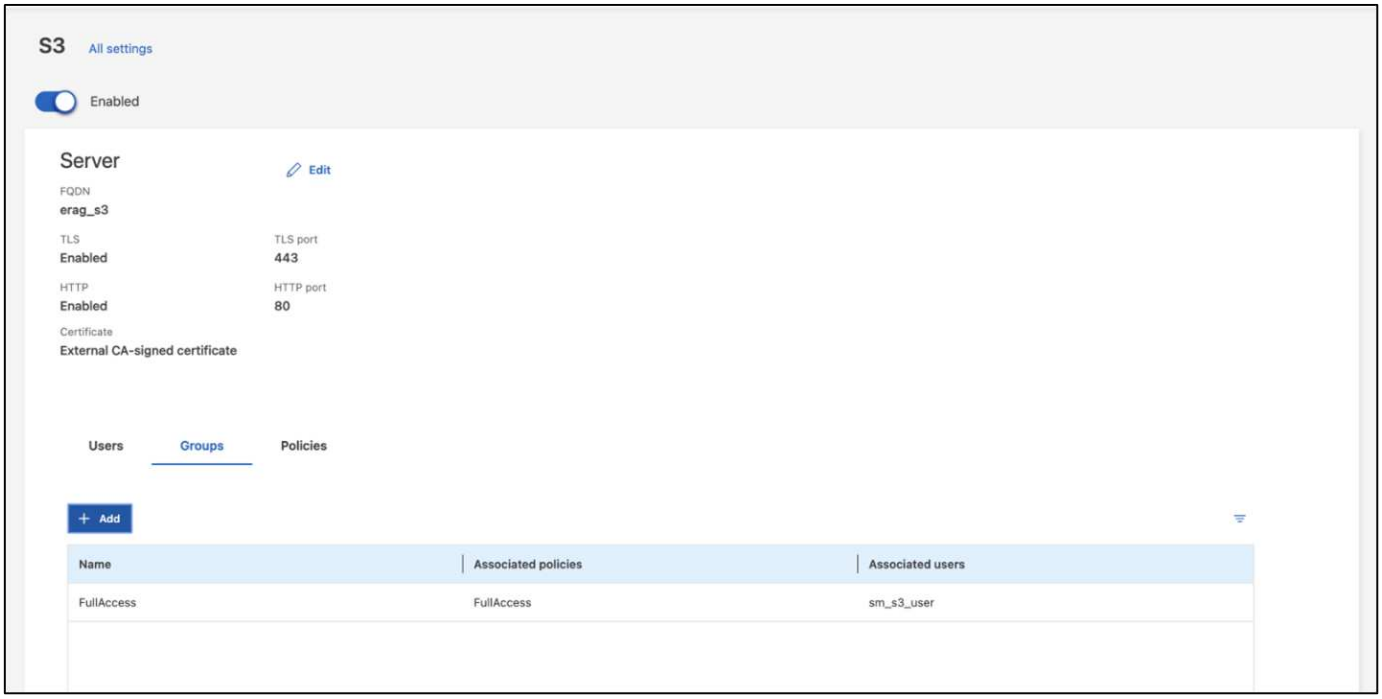
Configure as definições de usuário/grupo do S3 para o SVM que você criou na etapa anterior. Certifique-se de ter um usuário com acesso total a todas as operações da API do S3 para esse SVM. Consulte a documentação do ONTAP S3 para obter detalhes.

Observação: este usuário será necessário para o serviço de ingestão de dados do aplicativo Intel AI for Enterprise RAG. Se você criou seu SVM usando o ONTAP System Manager, o System Manager terá criado

automaticamente um usuário chamado `sm_s3_user` e uma política chamada `FullAccess` quando você criou seu SVM, mas nenhuma permissão foi atribuída a ele `sm_s3_user` .

Para editar as permissões deste usuário, navegue até Armazenamento > VMs de armazenamento, clique no nome da SVM que você criou na etapa anterior, clique em Configurações e, em seguida, clique no ícone de lápis ao lado de "S3". Para dar `sm_s3_user` acesso total a todas as operações da API S3, crie um novo grupo que associe `sm_s3_user` com o `FullAccess` política conforme ilustrado na captura de tela a seguir.

Figura 5 - Permissões do S3.

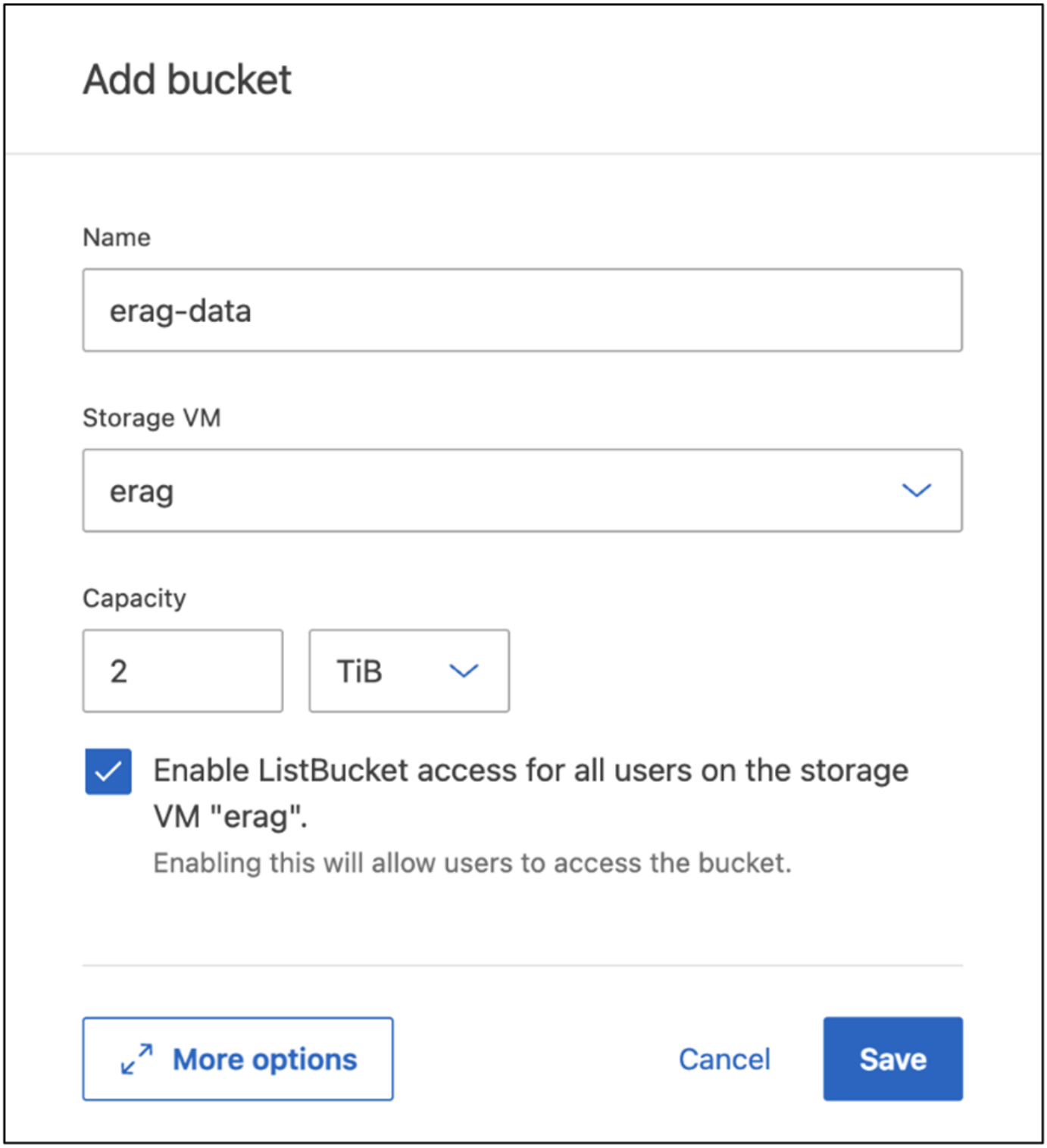


Criar um bucket S3

Crie um bucket S3 dentro do SVM que você criou anteriormente. Para criar um SVM usando o ONTAP System Manager, navegue até Armazenamento > Buckets e clique no botão + Adicionar. Para obter detalhes adicionais, consulte a documentação do ONTAP S3.

A captura de tela a seguir descreve a criação de um bucket S3 usando o ONTAP System Manager.

Figura 6 - Crie um bucket
S3.



Configurar permissões do bucket S3

Configure permissões para o bucket S3 que você criou na etapa anterior. Certifique-se de que o usuário configurado na etapa anterior tenha as seguintes permissões: `GetObject`, `PutObject`, `DeleteObject`, `ListBucket`, `GetBucketAcl`, `GetObjectAcl`, `ListBucketMultipartUploads`, `ListMultipartUploadParts`, `GetObjectTagging`, `PutObjectTagging`, `DeleteObjectTagging`, `GetBucketLocation`, `GetBucketVersioning`, `PutBucketVersioning`, `ListBucketVersions`, `GetBucketPolicy`, `PutBucketPolicy`, `DeleteBucketPolicy`, `PutLifecycleConfiguration`, `GetLifecycleConfiguration`, `GetBucketCORS`,

`PutBucketCORS.`

Para editar as permissões do bucket S3 usando o ONTAP System Manager, navegue até Armazenamento > Buckets, clique no nome do seu bucket, clique em Permissões e, em seguida, clique em Editar. Consulte o "Documentação do ONTAP S3" para obter detalhes adicionais.

A captura de tela a seguir descreve as permissões de bucket necessárias no ONTAP System Manager.

Figura 7 - Permissões do bucket

S3.

Criar regra de compartilhamento de recursos de origem cruzada de bucket

Usando a CLI do ONTAP , crie uma regra de compartilhamento de recursos de origem cruzada (CORS) para o bucket que você criou em uma etapa anterior:

```
ontap::> bucket cors-rule create -vserver erag -bucket erag-data -allowed
-origins *erag.com -allowed-methods GET,HEAD,PUT,DELETE,POST -allowed
-headers *
```

Esta regra permite que o aplicativo web OPEA para Intel AI for Enterprise RAG interaja com o bucket de dentro de um navegador da web.

Implantar servidores

Implante seus servidores e instale o Ubuntu 22.04 LTS em cada servidor. Após a instalação do Ubuntu, instale os utilitários NFS em todos os servidores. Para instalar os utilitários NFS, execute o seguinte comando:

```
apt-get update && apt-get install nfs-common
```

Instalar o Kubernetes

Instale o Kubernetes em seus servidores usando o Kubespray. Consulte o "Documentação do Kubespray" para mais detalhes.

Instalar o driver Trident CSI

Instale o driver NetApp Trident CSI no seu cluster Kubernetes. Consulte o "Documentação de instalação do Trident" para mais detalhes.

Crie um back-end Trident

Crie um back-end Trident para o SVM que você criou anteriormente. Ao criar seu back-end, use o `ontap-nas` motorista. Consulte o "Documentação de back-end do Trident" para mais detalhes.

Criar uma classe de armazenamento

Crie uma classe de armazenamento do Kubernetes correspondente ao back-end do Trident que você criou na etapa anterior. Consulte a documentação da classe de armazenamento Trident para obter detalhes.

OPEA para Intel AI para Enterprise RAG

Instale o OPEA para Intel AI for Enterprise RAG no seu cluster Kubernetes. Consulte o "Implantação do Intel AI para Enterprise RAG" documentação para mais detalhes. Não deixe de anotar as modificações necessárias no arquivo de configuração, descritas mais adiante neste artigo. Você deve fazer essas modificações antes de executar o manual de instalação para que o aplicativo Intel AI for Enterprise RAG funcione corretamente com seu sistema de armazenamento ONTAP .

Habilitar o uso do ONTAP S3

Ao instalar o OPEA para Intel AI for Enterprise RAG, edite seu arquivo de configuração principal para habilitar o uso do ONTAP S3 como seu repositório de dados de origem.

Para habilitar o uso do ONTAP S3, defina os seguintes valores em `edp` seção.

Observação: por padrão, o aplicativo Intel AI for Enterprise RAG ingere dados de todos os buckets existentes no seu SVM. Se você tiver vários buckets em seu SVM, poderá modificá-los `bucketNameRegexFilter` campo para que os dados sejam ingeridos apenas de determinados buckets.

```
edp:
  enabled: true
  namespace: edp
  dpGuard:
    enabled: false
  storageType: s3compatible
  s3compatible:
    region: "us-east-1"
    accessKeyId: "<your_access_key>"
    secretAccessKey: "<your_secret_key>"
    internalUrl: "https://<your_ONTAP_S3_interface>"
    externalUrl: "https://<your_ONTAP_S3_interface>"
    bucketNameRegexFilter: ".*"
```

Configurar as configurações de sincronização agendada

Ao instalar o aplicativo OPEA para Intel AI for Enterprise RAG, habilite `scheduledSync` para que o aplicativo ingira automaticamente arquivos novos ou atualizados dos seus buckets do S3.

Quando `scheduledSync` estiver habilitado, o aplicativo verificará automaticamente seus buckets S3 de origem em busca de arquivos novos ou atualizados. Todos os arquivos novos ou atualizados encontrados

como parte desse processo de sincronização são automaticamente ingeridos e adicionados à base de conhecimento do RAG. O aplicativo verifica seus buckets de origem com base em um intervalo de tempo predefinido. O intervalo de tempo padrão é 60 segundos, o que significa que o aplicativo verifica alterações a cada 60 segundos. Talvez você queira alterar esse intervalo para atender às suas necessidades específicas.

Para habilitar `scheduledSync` e defina o intervalo de sincronização, defina os seguintes valores em `deployment/components/edp/values.yaml:`

```
celery:
  config:
    scheduledSync:
      enabled: true
      syncPeriodSeconds: "60"
```

Alterar modos de acesso de volume

Em `deployment/components/gmc/microservices-connector/helm/values.yaml` , para cada volume no `pvc` lista, altere o `accessMode` para `ReadWriteMany` .

```
pvc:
  modelLlm:
    name: model-volume-llm
    accessMode: ReadWriteMany
    storage: 100Gi
  modelEmbedding:
    name: model-volume-embedding
    accessMode: ReadWriteMany
    storage: 20Gi
  modelReranker:
    name: model-volume-reranker
    accessMode: ReadWriteMany
    storage: 10Gi
  vectorStore:
    name: vector-store-data
    accessMode: ReadWriteMany
    storage: 20Gi
```

(Opcional) Desabilitar verificação de certificado SSL

Se você usou um certificado autoassinado ao habilitar o acesso S3 para seu SVM, deverá desabilitar a verificação do certificado SSL. Se você usou um certificado assinado por uma CA publicamente confiável, pode pular esta etapa.

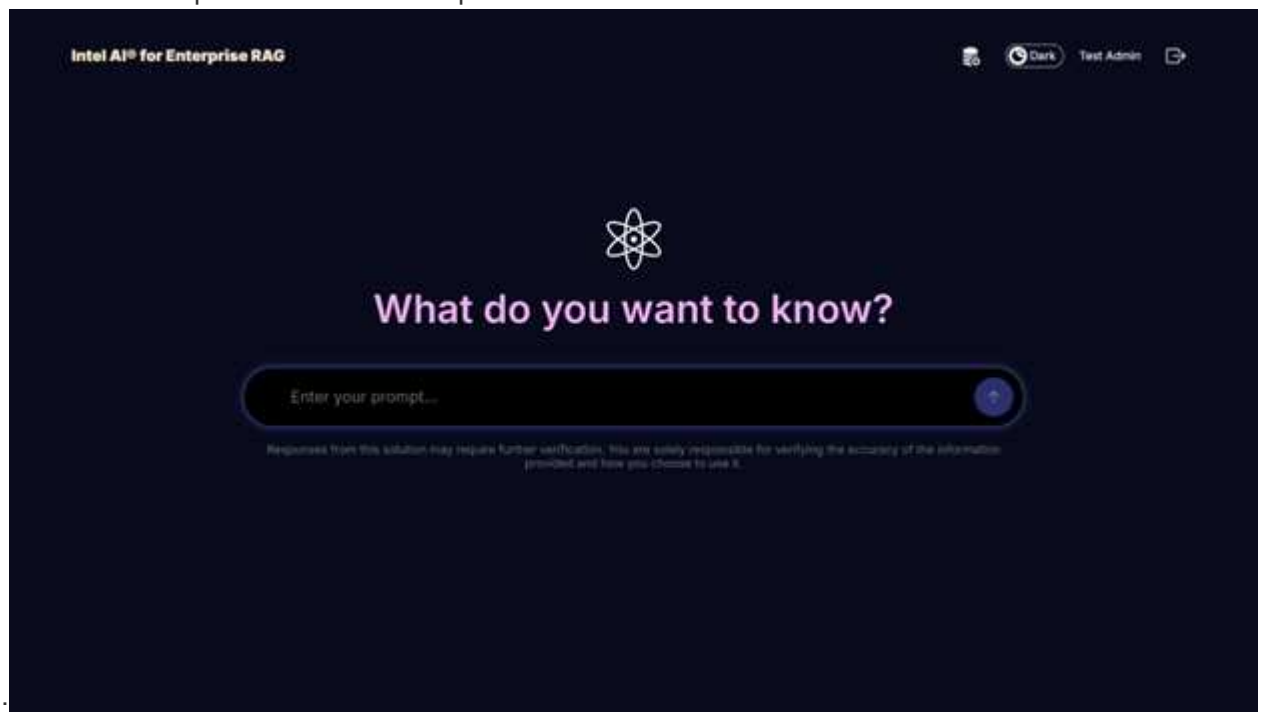
Para desabilitar a verificação do certificado SSL, defina os seguintes valores em `deployment/components/edp/values.yaml:`

```
edpExternalUrl: "https://s3.erag.com"
edpExternalSecure: "true"
edpExternalCertVerify: "false"
edpInternalUrl: "edp-minio:9000"
edpInternalSecure: "true"
edpInternalCertVerify: "false"
```

Acesse OPEA para Intel AI for Enterprise RAG UI

Acesse o OPEA para Intel AI for Enterprise RAG UI. Consulte o "Documentação de implantação do Intel AI for Enterprise RAG" para mais detalhes.

Figura 8 – UI OPEA para Intel AI for Enterprise RAG.



Ingerir dados para RAG

Agora você pode ingerir arquivos para inclusão no aumento de consulta baseado em RAG. Há várias opções para ingestão de arquivos. Escolha a opção apropriada para suas necessidades.

Observação: depois que um arquivo é ingerido, o aplicativo OPEA para Intel AI for Enterprise RAG verifica automaticamente se há atualizações no arquivo e as ingere adequadamente.

*Opção 1: Carregar diretamente para o seu bucket S3 Para ingerir muitos arquivos de uma vez, recomendamos carregar os arquivos para o seu bucket S3 (o bucket que você criou anteriormente) usando o cliente S3 de sua escolha. Os clientes S3 populares incluem o AWS CLI, o Amazon SDK para Python (Boto3), s3cmd, S3 Browser, Cyberduck e Commander One. Se os arquivos forem de um tipo compatível, todos os arquivos que você enviar para seu bucket S3 serão ingeridos automaticamente pelo aplicativo OPEA for Intel AI for Enterprise RAG.

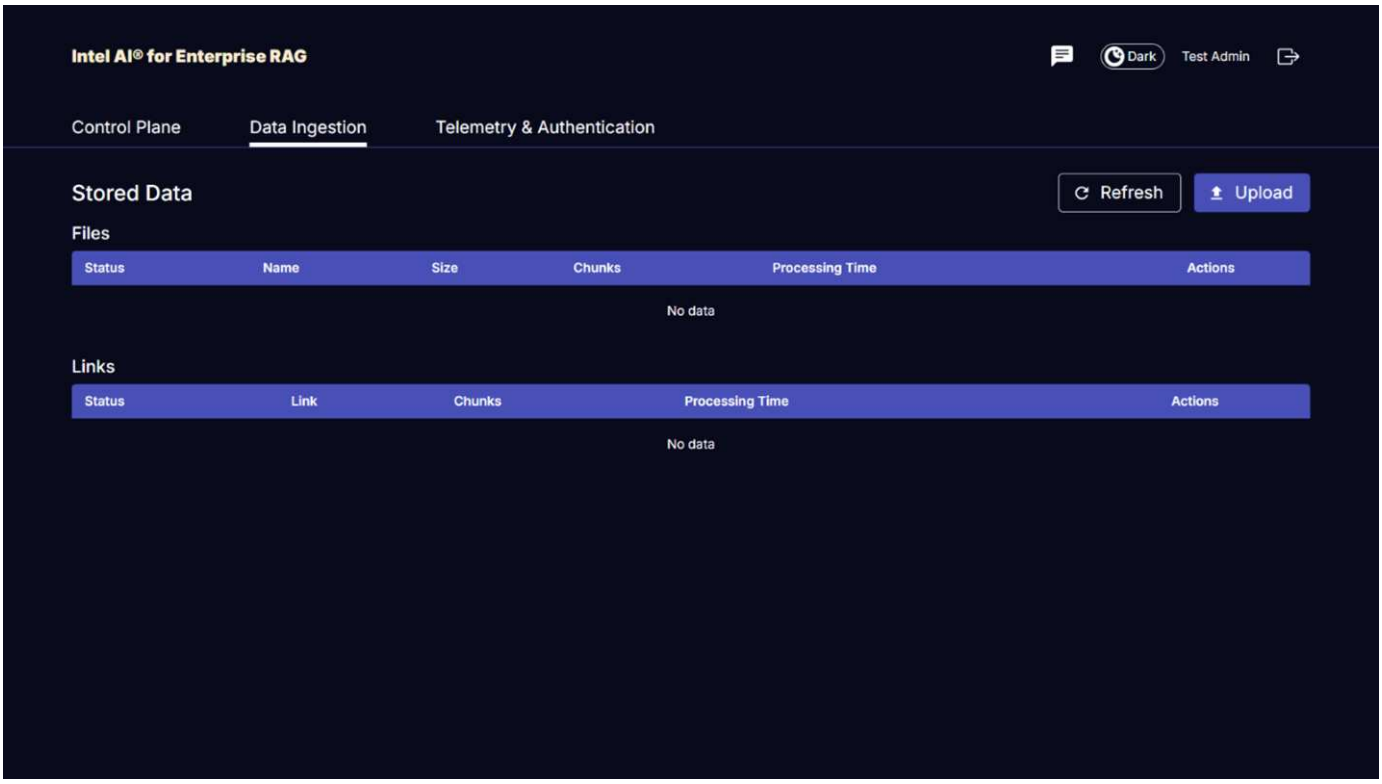
Observação: no momento em que este artigo foi escrito, os seguintes tipos de arquivo eram suportados: PDF, HTML, TXT, DOC, DOCX, PPT, PPTX, MD, XML, JSON, JSONL, YAML, XLS, XLSX, CSV, TIFF, JPG, JPEG,

PNG e SVG.

Você pode usar o OPEA para Intel AI for Enterprise RAG UI para confirmar se seus arquivos foram ingeridos corretamente. Consulte a documentação da interface de usuário Intel AI for Enterprise RAG para obter detalhes. Observe que pode levar algum tempo para o aplicativo ingerir um grande número de arquivos.

*Opção 2: Carregar usando a interface do usuário. Se você precisar ingerir apenas um pequeno número de arquivos, poderá ingeri-los usando a interface do usuário do OPEA para Intel AI for Enterprise RAG. Consulte a documentação da interface de usuário Intel AI for Enterprise RAG para obter detalhes.

Figura 9 - Interface de usuário de ingestão de dados.



Executar consultas de bate-papo

Agora você pode "conversar" com o aplicativo OPEA for Intel AI for Enterprise RAG usando a interface de bate-papo incluída. Ao responder às suas perguntas, o aplicativo executa o RAG usando seus arquivos ingeridos. Isso significa que o aplicativo busca automaticamente informações relevantes dentro dos seus arquivos ingeridos e incorpora essas informações ao responder às suas consultas.

Orientação de dimensionamento

Como parte do nosso esforço de validação, conduzimos testes de desempenho em coordenação com a Intel. Esse teste resultou nas orientações de dimensionamento descritas na tabela a seguir.

Caracterizações	Valor	Comentário
Tamanho do modelo	20 bilhões de parâmetros	Lhama-8B, Lhama-13B, Mistral 7B, Qwen 14B, DeepSeek Distill 8B
Tamanho da entrada	~2 mil tokens	~4 páginas

Caracterizações	Valor	Comentário
Tamanho da saída	~2 mil tokens	~4 páginas
Usuários simultâneos	32	"Usuários simultâneos" referem-se a solicitações rápidas que enviam consultas ao mesmo tempo.

_Observação: as orientações de dimensionamento apresentadas acima são baseadas na validação de desempenho e nos resultados de testes coletados usando processadores Intel Xeon 6 com 96 núcleos. Para clientes com tokens de E/S e requisitos de tamanho de modelo semelhantes, recomendamos usar servidores com processadores Xeon 6 com 96 ou 128 núcleos.

Conclusão

Os sistemas RAG empresariais e LLMs são tecnologias que trabalham juntos para ajudar as organizações a fornecer respostas precisas e contextualizadas. Essas respostas envolvem recuperação de informações com base em uma vasta coleção de dados empresariais privados e internos. Ao usar RAG, APIs, embeddings de vetores e sistemas de armazenamento de alto desempenho para consultar repositórios de documentos que contêm dados da empresa, os dados são processados de forma mais rápida e segura. O NetApp AIPod Mini combina a infraestrutura de dados inteligente da NetApp com recursos de gerenciamento de dados ONTAP e processadores Intel Xeon 6, Intel AI for Enterprise RAG e a pilha de software OPEA para ajudar a implantar aplicativos RAG de alto desempenho e colocar as organizações no caminho da liderança em IA.

Reconhecimento

Este documento é trabalho de Sathish Thyagarajan e Michael Ogelsby, membros da equipe de engenharia de soluções da NetApp . Os autores também gostariam de agradecer à equipe de produtos de IA empresarial da Intel — Ajay Mungara, Mikolaj Zyczynski, Igor Konopko, Ramakrishna Karamsetty, Michal Prostko, Shreejan Mistry e Ned Fiori — e outros membros da equipe da NetApp— Lawrence Bunka, Bobby Oommen e Jeff Liborio — por seu suporte e ajuda contínuos durante a validação desta solução.

Lista de materiais

A seguir está a lista de materiais usada para a validação funcional desta solução e pode ser usada como referência. Qualquer servidor ou componente de rede (ou mesmo uma rede existente com largura de banda preferencialmente de 100 GbE) que esteja alinhado com a seguinte configuração pode ser usado.

Para o servidor de aplicativos:

Número da peça	Descrição do produto	Quantidade
222HA-TN-OTO-37	Hyper SuperServer SYS-222HA-TN /2U	2
P4X-GNR6980P-SRPL2-UCC	Intel Xeon 6980P 2P 128C 2G 504M 500W SGX512	4
BATER	MEM-DR564MC-ER64(x16)64GB DDR5-6400 2RX4 (16Gb) ECC RDIMM	32

Número da peça	Descrição do produto	Quantidade
	HDS-M2N4-960G0-E1-TXD-NON-080(x2) SSD M.2 NVMe PCIe4 960GB 1DWPD TLC D, 80mm	2
	Fonte de alimentação redundante de saída única WS-1K63A-1R(x2)1U 692W/1600W. Dissipação de calor de 2361 BTU/h com temperatura máxima de 59 °C (aprox.)	4

Para o servidor de controle:

Número da peça	Descrição do produto	Quantidade
511R-M-OTO-17	OTIMIZADO UP 1U X13SCH-SYS, CSE-813MF2TS-R0RCNBP, PWS-602A-1R	1
	RPL-E 6369P IP 8C/16T 3.3G 24MB 95W 1700 BO	1
BATER	MEM-DR516MB-EU48(x2)16GB DDR5-4800 1Rx8 (16Gb) ECC UDIMM	1
	HDS-M2N4-960G0-E1-TXD-NON-080(x2) SSD M.2 NVMe PCIe4 960GB 1DWPD TLC D, 80mm	2

Para o switch de rede:

Número da peça	Descrição do produto	Quantidade
DCS-7280CR3A	Arista 7280R3A 28x100 GbE	1

Armazenamento NetApp AFF :

Número da peça	Descrição do produto	Quantidade
AFF-A20A-100-C	Sistema AFF A20 HA, -C	1
X800-42U-R6-C	Jumper Crd, na cabine, C13-C14, -C	2
X97602A-C	Fonte de alimentação, 1600 W, titânio, -C	2
X66211B-2-N-C	Cabo, 100GbE, QSFP28-QSFP28, Cu, 2m, -C	4
X66240A-05-N-C	Cabo, 25GbE, SFP28-SFP28, Cu, 0,5m, -C	2

X5532A-N-C	Trilho, 4 postes, fino, furo quadrado/redondo, pequeno, ajuste, 24-32, -C	1
X4024A-2-A-C	Pacote de unidade 2X1,92 TB, NVMe4, SED, -C	6
X60130A-C	Módulo IO, 2PT, 100GbE, -C	2
X60132A-C	Módulo IO, 4PT, 10/25GbE, -C	2
SW-ONTAPB-FLASH-A20-C	SW, pacote base ONTAP , por TB, Flash, A20, -C	23

Onde encontrar informações adicionais

Para saber mais sobre as informações descritas neste documento, revise os seguintes documentos e/ou sites:

"Documentação do produto NetApp"

"Projeto OPEA"

"Manual de implantação do OPEA Enterprise RAG"

Informações sobre direitos autorais

Copyright © 2025 NetApp, Inc. Todos os direitos reservados. Impresso nos EUA. Nenhuma parte deste documento protegida por direitos autorais pode ser reproduzida de qualquer forma ou por qualquer meio — gráfico, eletrônico ou mecânico, incluindo fotocópia, gravação, gravação em fita ou storage em um sistema de recuperação eletrônica — sem permissão prévia, por escrito, do proprietário dos direitos autorais.

O software derivado do material da NetApp protegido por direitos autorais está sujeito à seguinte licença e isenção de responsabilidade:

ESTE SOFTWARE É FORNECIDO PELA NETAPP "NO PRESENTE ESTADO" E SEM QUAISQUER GARANTIAS EXPRESSAS OU IMPLÍCITAS, INCLUINDO, SEM LIMITAÇÕES, GARANTIAS IMPLÍCITAS DE COMERCIALIZAÇÃO E ADEQUAÇÃO A UM DETERMINADO PROPÓSITO, CONFORME A ISENÇÃO DE RESPONSABILIDADE DESTE DOCUMENTO. EM HIPÓTESE ALGUMA A NETAPP SERÁ RESPONSÁVEL POR QUALQUER DANO DIRETO, INDIRETO, INCIDENTAL, ESPECIAL, EXEMPLAR OU CONSEQUENCIAL (INCLUINDO, SEM LIMITAÇÕES, AQUISIÇÃO DE PRODUTOS OU SERVIÇOS SOBRESSALENTES; PERDA DE USO, DADOS OU LUCROS; OU INTERRUPÇÃO DOS NEGÓCIOS), INDEPENDENTEMENTE DA CAUSA E DO PRINCÍPIO DE RESPONSABILIDADE, SEJA EM CONTRATO, POR RESPONSABILIDADE OBJETIVA OU PREJUÍZO (INCLUINDO NEGLIGÊNCIA OU DE OUTRO MODO), RESULTANTE DO USO DESTE SOFTWARE, MESMO SE ADVERTIDA DA RESPONSABILIDADE DE TAL DANO.

A NetApp reserva-se o direito de alterar quaisquer produtos descritos neste documento, a qualquer momento e sem aviso. A NetApp não assume nenhuma responsabilidade nem obrigação decorrentes do uso dos produtos descritos neste documento, exceto conforme expressamente acordado por escrito pela NetApp. O uso ou a compra deste produto não representam uma licença sob quaisquer direitos de patente, direitos de marca comercial ou quaisquer outros direitos de propriedade intelectual da NetApp.

O produto descrito neste manual pode estar protegido por uma ou mais patentes dos EUA, patentes estrangeiras ou pedidos pendentes.

LEGENDA DE DIREITOS LIMITADOS: o uso, a duplicação ou a divulgação pelo governo estão sujeitos a restrições conforme estabelecido no subparágrafo (b)(3) dos Direitos em Dados Técnicos - Itens Não Comerciais no DFARS 252.227-7013 (fevereiro de 2014) e no FAR 52.227- 19 (dezembro de 2007).

Os dados aqui contidos pertencem a um produto comercial e/ou serviço comercial (conforme definido no FAR 2.101) e são de propriedade da NetApp, Inc. Todos os dados técnicos e software de computador da NetApp fornecidos sob este Contrato são de natureza comercial e desenvolvidos exclusivamente com despesas privadas. O Governo dos EUA tem uma licença mundial limitada, irrevogável, não exclusiva, intransferível e não sublicenciável para usar os Dados que estão relacionados apenas com o suporte e para cumprir os contratos governamentais desse país que determinam o fornecimento de tais Dados. Salvo disposição em contrário no presente documento, não é permitido usar, divulgar, reproduzir, modificar, executar ou exibir os dados sem a aprovação prévia por escrito da NetApp, Inc. Os direitos de licença pertencentes ao governo dos Estados Unidos para o Departamento de Defesa estão limitados aos direitos identificados na cláusula 252.227-7015(b) (fevereiro de 2014) do DFARS.

Informações sobre marcas comerciais

NETAPP, o logotipo NETAPP e as marcas listadas em http://www.netapp.com/TM são marcas comerciais da NetApp, Inc. Outros nomes de produtos e empresas podem ser marcas comerciais de seus respectivos proprietários.