# NetApp AIPod Mini - Inferência RAG empresarial com NetApp e Intel

## NetApp artificial intelligence solutions

NetApp
February 12, 2026

# Índice

# 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, Arpita Mahajan, NetApp

## Sumário executivo

Um número crescente de organizações está utilizando aplicações de geração aumentada por recuperação (RAG) e grandes modelos de linguagem (LLMs) para interpretar solicitações de usuários e gerar respostas para aumentar a produtividade e o valor para os negócios. Essas solicitações e respostas podem incluir texto, código, imagens ou até mesmo estruturas de proteínas terapêuticas recuperadas da base de conhecimento interna da organização, data lakes, repositórios de código e repositórios de documentos. Este artigo descreve o projeto de referência da solução NetApp AIPod Mini, composta por NetApp AFF storage e servidores com processadores Intel Xeon 6. Inclui o software de gerenciamento de dados NetApp ONTAP combinado com Intel Advanced Matrix Extensions (Intel AMX) e o software Intel® AI for Enterprise RAG construído sobre Open Platform for Enterprise AI (OPEA). O NetApp AIPod Mini para enterprise RAG permite que as organizações aumentem um LLM público em uma solução privada de inferência de IA generativa (GenAI). A solução demonstra inferência RAG eficiente e econômica em escala empresarial, projetada para aumentar a confiabilidade e fornecer a você maior 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 ou Excel. Esses dados são normalmente armazenados em soluções como um storage de objetos S3 ou NFS on-premises como fonte para dados. NetApp tem sido líder em tecnologias de gerenciamento de dados, mobilidade de dados, governança de dados e segurança de dados em todo o ecossistema de edge, data center e nuvem. NetApp ONTAP data management fornece armazenamento de nível empresarial para suportar 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 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.

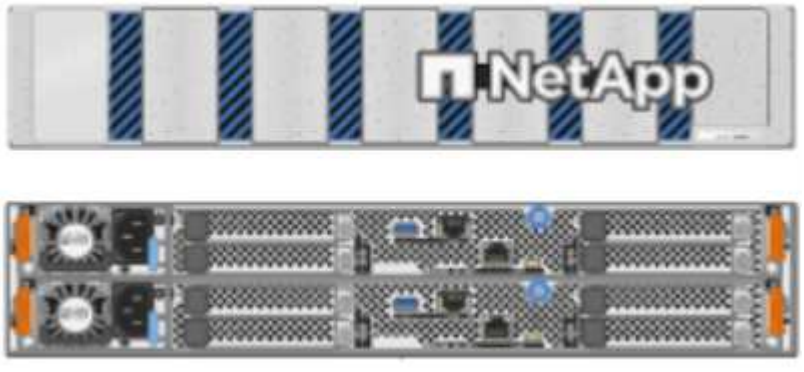


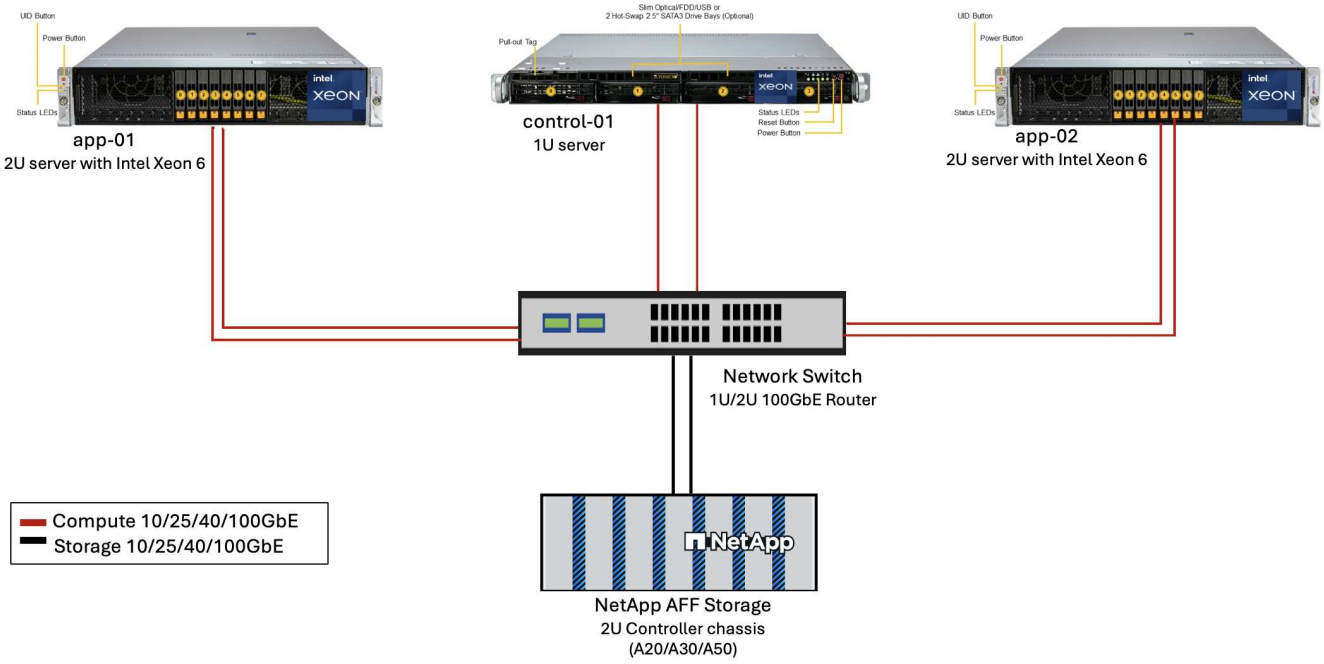

Figura 2 - Sistema NetApp AFF Série A.

| Hardware | Quantidade | Comentário |
| --- | --- | --- |
| Intel Xeon 6ª Geração (Granite Rapids) | 2 | Nós de inferência RAG — com processadores Intel Xeon 6900-series (96 núcleos) ou Intel Xeon 6700-series (64 núcleos) de dois soquetes e 250GB a 3TB de RAM com DDR5 (6400MHz) ou MRDIMM (8800MHz). Servidor 2U. |

| Hardware | Quantidade | Comentário |
| --- | --- | --- |
| Servidor de plano de controle com processador Intel | 1 | Plano de controle do Kubernetes/servidor 1U. |
| 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).

Figura 3 - AIPod Mini Arquitetura de Implantação




## 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 for Enterprise RAG desenvolvido por OPEA

OPEA para Intel® AI for Enterprise RAG simplifica a transformação dos dados da sua empresa em insights acionáveis. Impulsionado por processadores Intel Xeon, integra componentes de parceiros do setor para oferecer uma abordagem simplificada para a implementação de soluções empresariais. Ele escala perfeitamente com frameworks de orquestração comprovados, proporcionando a flexibilidade e a escolha que sua empresa precisa.

Com base no OPEA, Intel® AI for Enterprise RAG amplia essa base com principais recursos que aprimoram a escalabilidade, a segurança e a experiência do usuário. Esses recursos incluem funcionalidades de service mesh para integração perfeita com arquiteturas modernas baseadas em serviços, validação pronta para produção para garantir a confiabilidade do pipeline e uma interface de usuário rica em recursos para RAG como serviço, permitindo o gerenciamento e o monitoramento simplificados dos fluxos de trabalho. Além disso, o suporte da Intel e de parceiros oferece acesso a um amplo ecossistema de soluções, combinado com gerenciamento integrado de identidade e acesso (IAM) com interface de usuário e aplicativos para operações seguras e em conformidade. Guardrails programáveis proporcionam controle preciso 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 - Intel® AI for Enterprise RAG | 2,0 | Plataforma empresarial RAG baseada em microsserviços OPEA |
| Interface de armazenamento de contêiner (driver CSI) | NetApp Trident 25.10 | Permite provisionamento dinâmico, cópias do NetApp Snapshot e volumes. |
| Ubuntu | 22.04.5 | Sistema operacional em cluster de dois nós. |

| Software | Versão | Comentário |
|---|---|---|
| Orquestração de contêineres | Kubernetes 1.31.9 (Instalado pelo playbook de infraestrutura Enterprise RAG) | Ambiente para executar o framework RAG |
| ONTAP | ONTAP 9.16.1P4 ou superior | Sistema operacional de armazenamento no AFF A20. |

# Implantação da solução

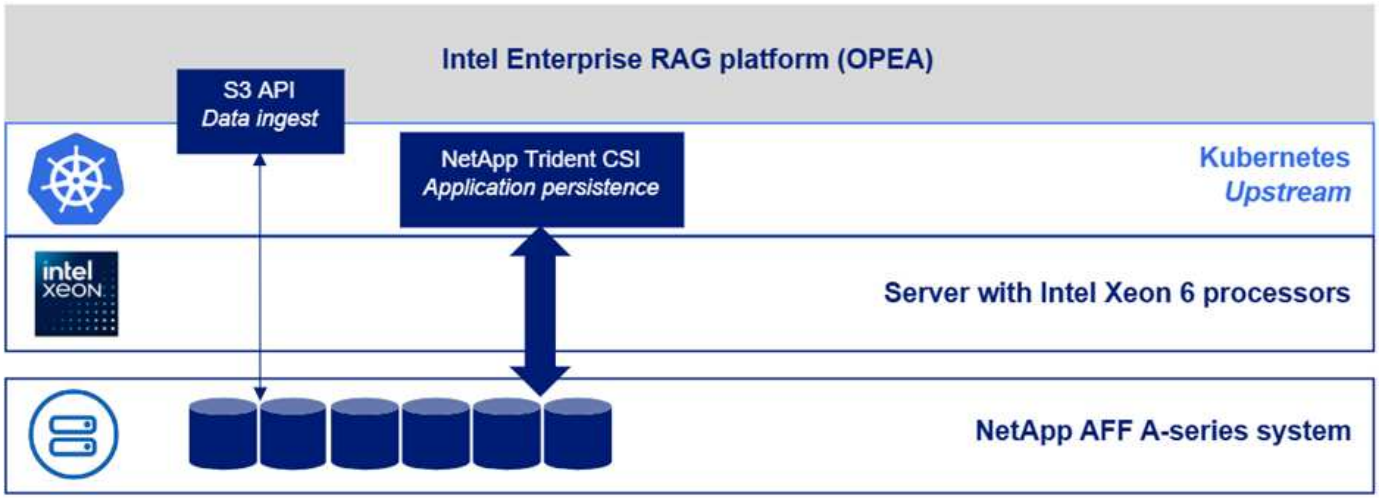
## Pilha de software

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

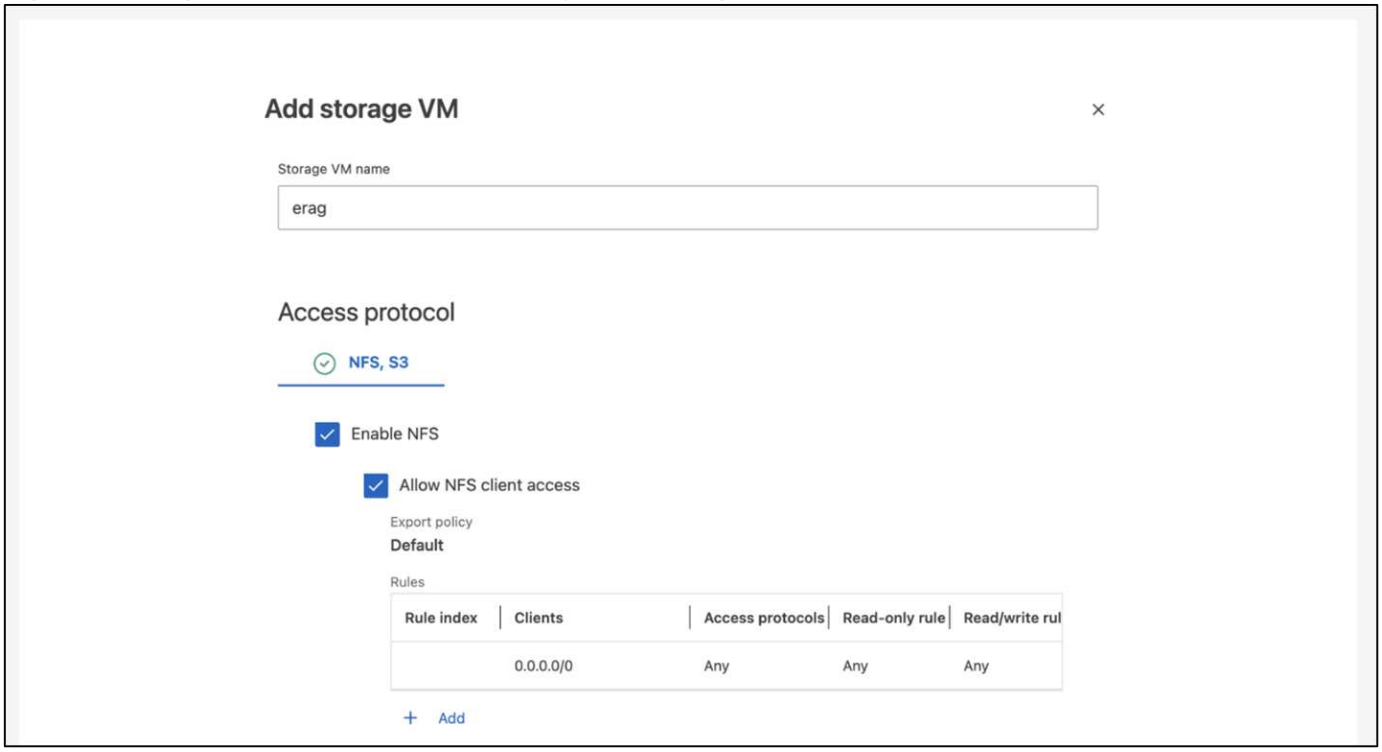
Configure uma máquina virtual de armazenamento ONTAP (SVM) para acesso NFS e S3 em uma rede que

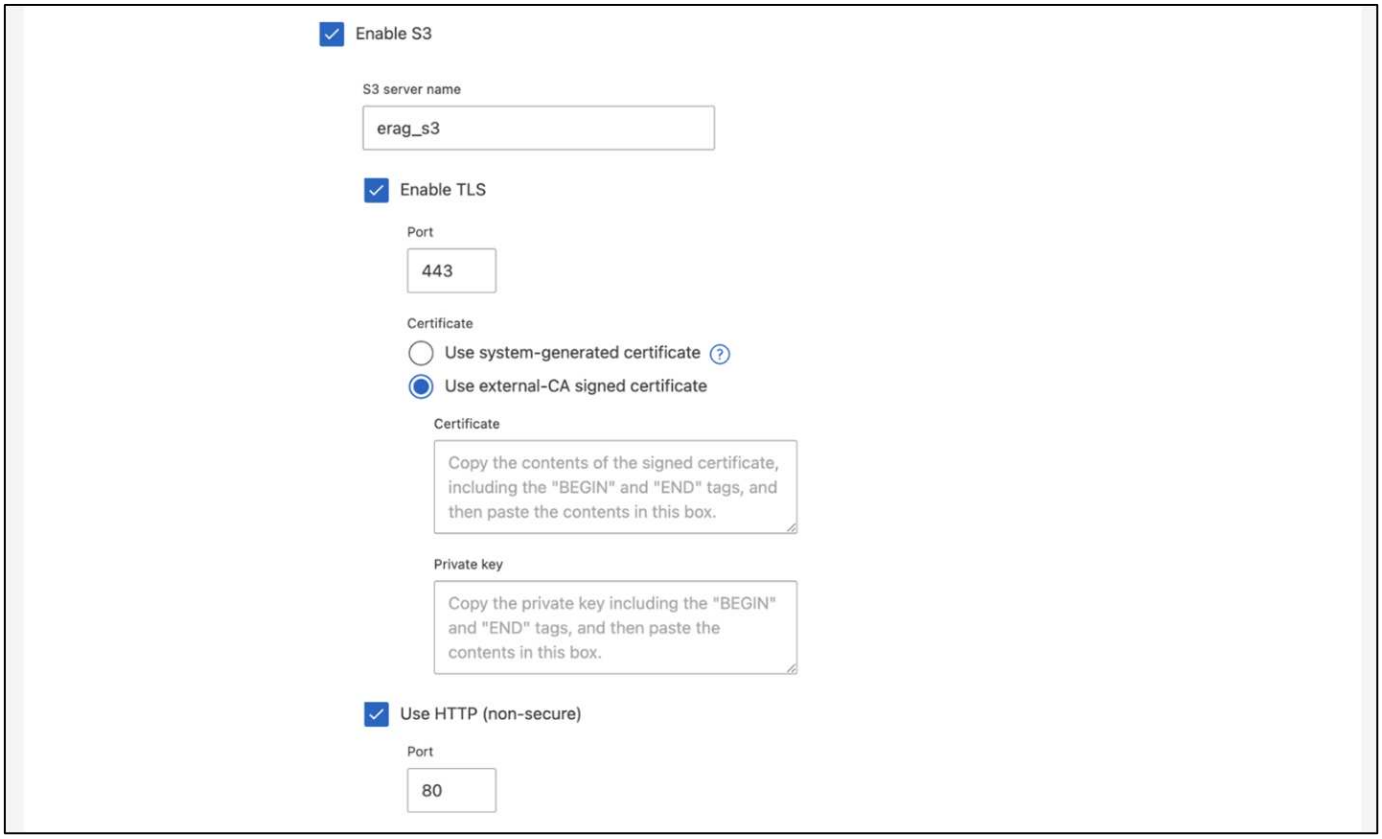
seja acessível pelos seus nós do Kubernetes.

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 5 - Criação de SVM usando ONTAP System Manager.





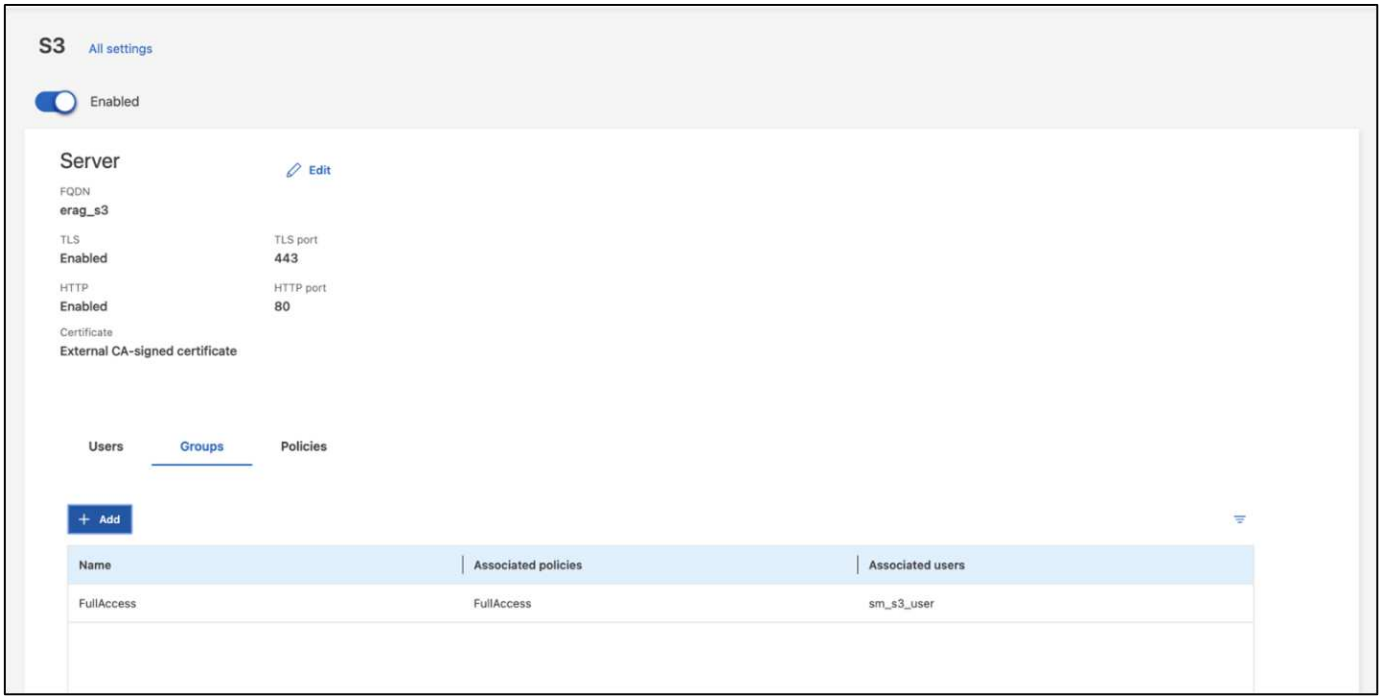

**Configurar permissões do S3**

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 sua SVM usando ONTAP System Manager, System Manager terá criado automaticamente um usuário chamado `sm_s3_user` e uma política chamada `FullAccess` quando você criou sua SVM, mas nenhuma permissão terá sido atribuída a `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 6 - Permissões S3.

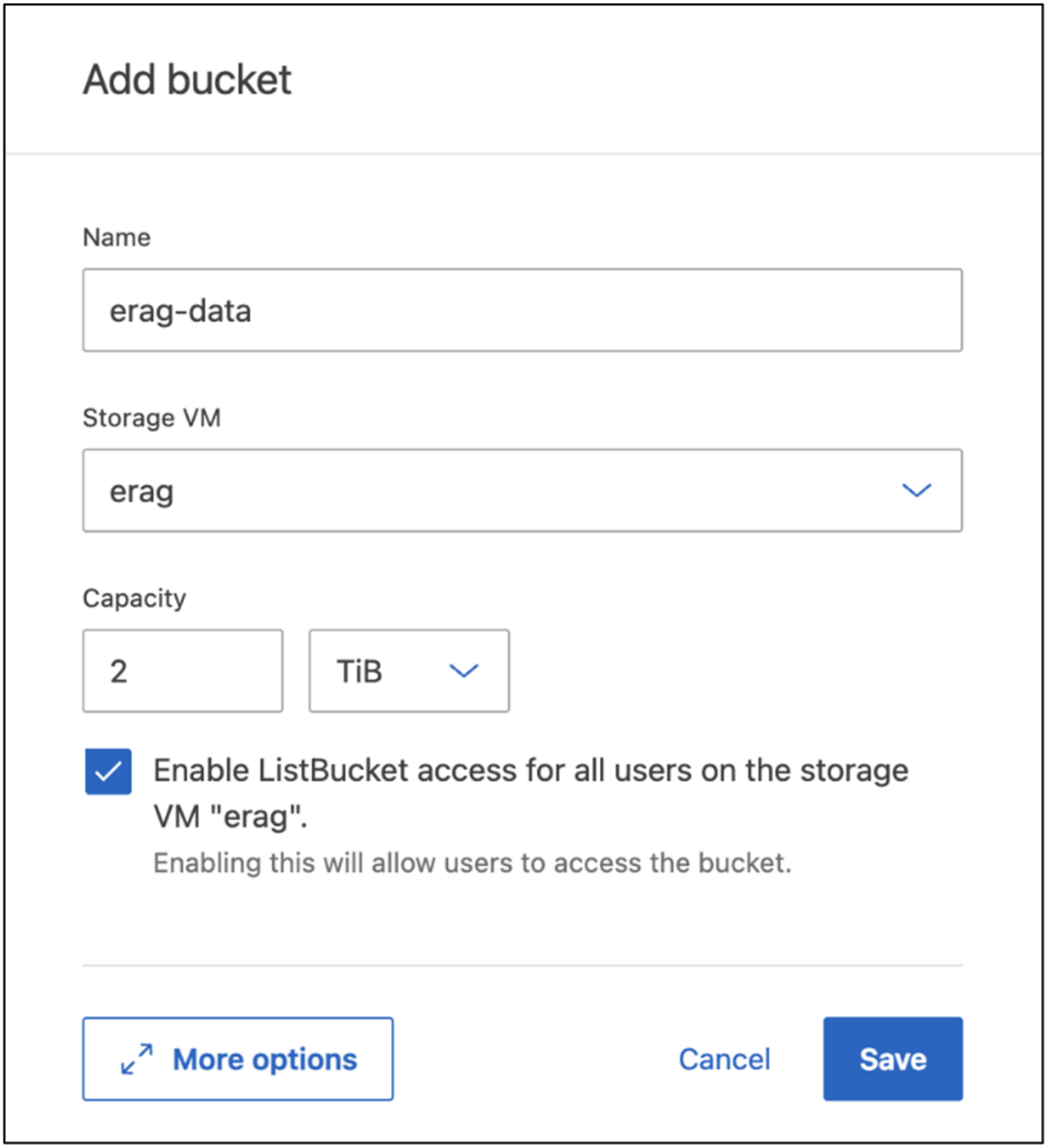



**Criar um bucket S3**

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 7 - Criar 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 8 - 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 OPEA para Intel® AI for Enterprise RAG web application interaja com o bucket a partir de um navegador 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
```

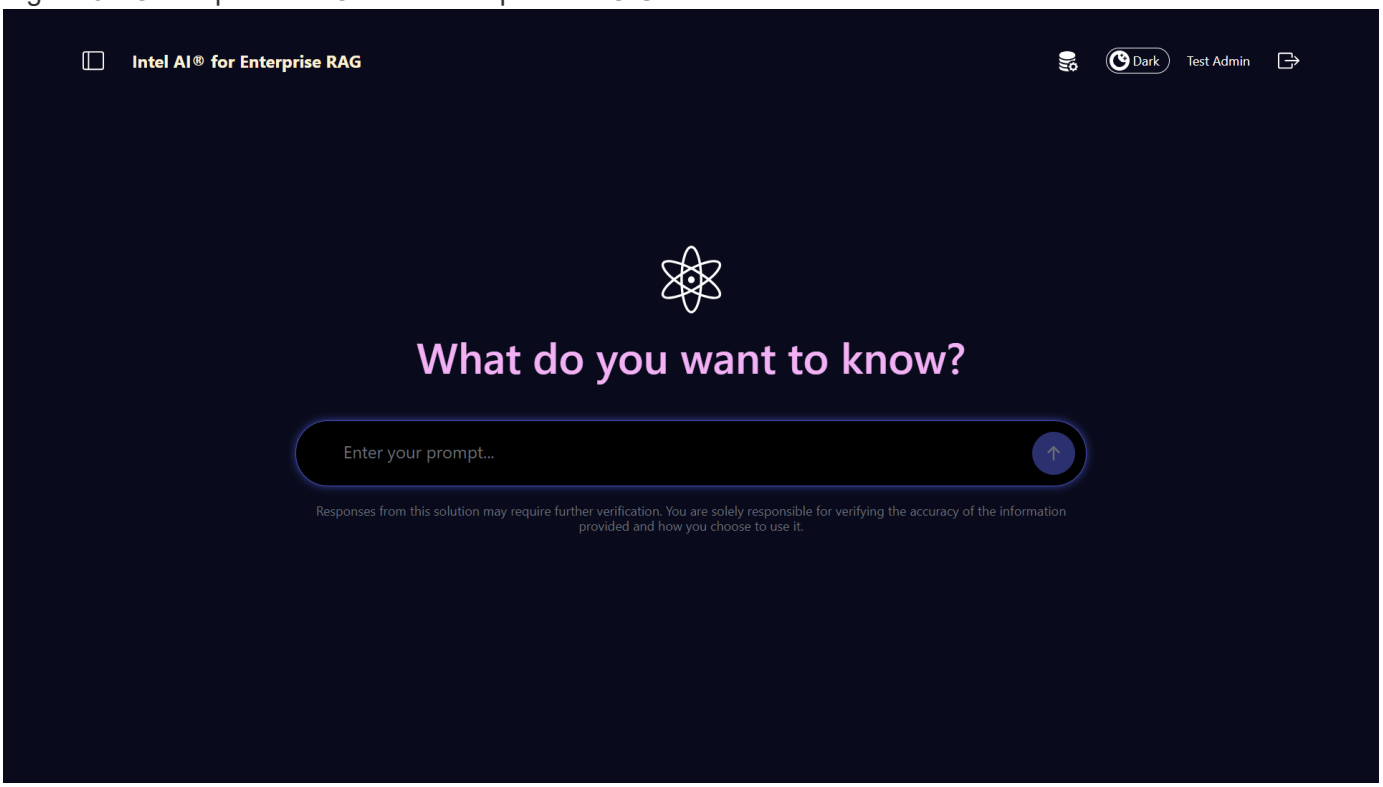
### Implantar Enterprise RAG 2.0

Consulte o seguinte documento para obter um fluxo de trabalho de implantação completo e passo a passo: NetApp AIPod Mini para ERAG - Etapas de implantaçãoTodos os pré-requisitos, preparação da infraestrutura, parâmetros de configuração e procedimentos de implantação estão documentados no guia de implantação acima.

### Acesse OPEA para Intel® AI for Enterprise RAG UI

Acesse o OPEA para Intel® AI for Enterprise RAG UI. Consulte o "Documentação de deployment do Intel® AI for Enterprise RAG" para obter detalhes.

Figura 9 - OPEA para Intel® AI for Enterprise RAG UI.




### 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.

Nota: Após a ingestão de um arquivo, o OPEA para Intel® AI for Enterprise RAG application verifica automaticamente se há atualizações no arquivo e as ingere conforme necessário.

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

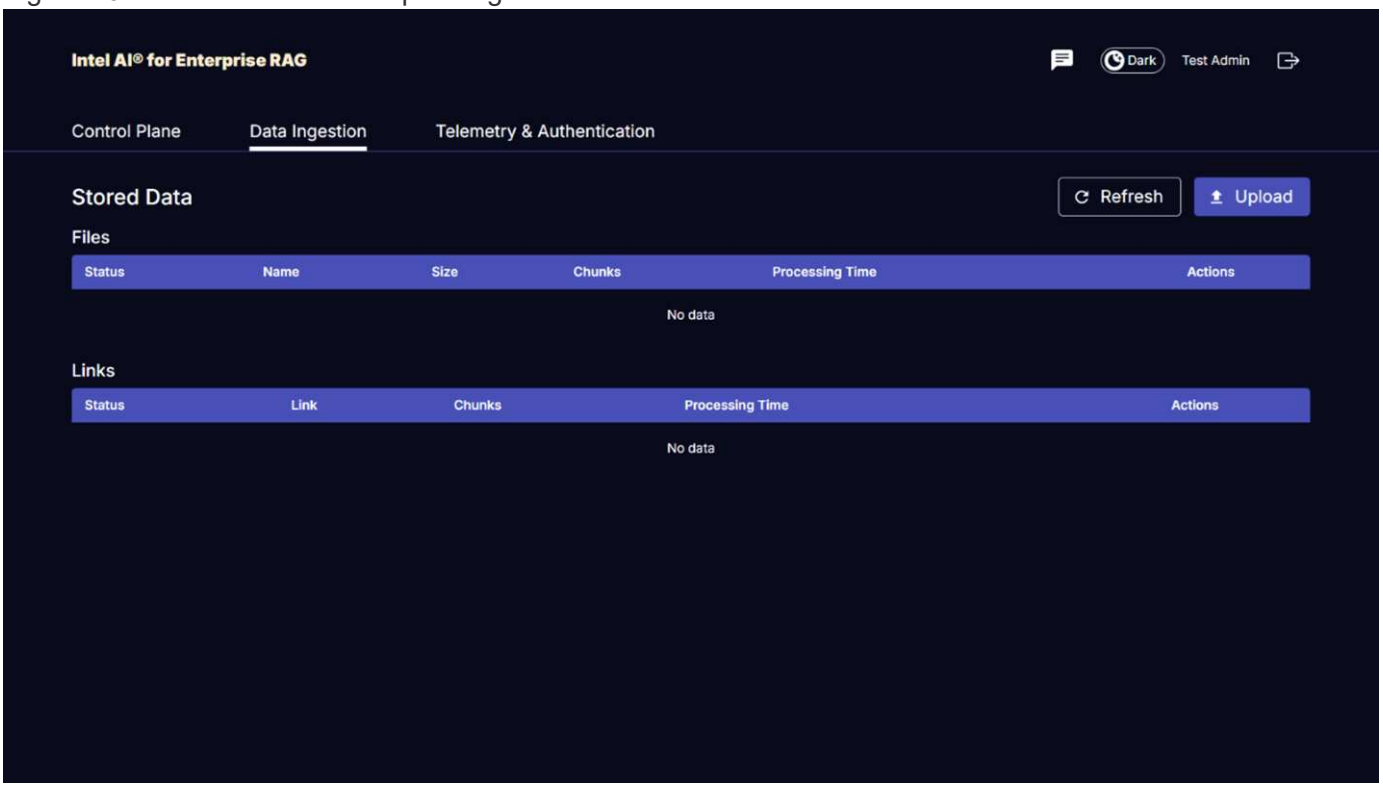
Observação: no momento da redação deste documento, os seguintes tipos de arquivo são suportados: PDF, HTML, TXT, DOC, DOCX, ADOC, PPT, PPTX, MD, XML, JSON, JSONL, YAML, XLS, XLSX, CSV, TIFF, JPG, JPEG, PNG e SVG.

Você pode usar o OPEA for Intel® AI for Enterprise RAG UI para confirmar se seus arquivos foram importados corretamente. Consulte a documentação do Intel® AI for Enterprise RAG UI para obter detalhes. Observe que pode levar algum tempo para o aplicativo importar um grande número de arquivos.

*Opção 2: Carregar usando a interface do usuário Se você precisar importar apenas um pequeno número de arquivos, poderá importá-los usando o OPEA para Intel® AI for Enterprise RAG UI. Consulte a documentação

do Intel® AI for Enterprise RAG UI para obter detalhes.

Figura 10 - Interface do usuário para ingestão de dados.




**Executar consultas de bate-papo**

Agora você pode "conversar" com o OPEA for Intel® AI for Enterprise RAG application usando a interface de chat incluída. Ao responder às suas perguntas, o aplicativo realiza 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 perguntas.

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

Nota: As orientações de dimensionamento apresentadas acima baseiam-se na validação de desempenho e nos resultados de testes realizados com processadores Intel Xeon 6 de 96 núcleos. Para clientes com requisitos semelhantes de tokens de E/S e tamanho de modelo, recomendamos o uso de servidores com processadores Xeon 6 de 96 núcleos. Para obter mais detalhes sobre o guia de dimensionamento, consulte

"Guia de dimensionamento RAG Intel® AI for Enterprise"

# Conclusão

Os sistemas RAG corporativos e os LLMs são tecnologias que trabalham juntas para ajudar as organizações a fornecer respostas precisas e com reconhecimento de contexto. Essas respostas envolvem a recuperação de informações com base em uma vasta coleção de dados internos e privados da empresa. Ao usar RAG, APIs, incorporações vetoriais e sistemas de armazenamento de alto desempenho para consultar repositórios de documentos que contêm dados da empresa, os dados são processados mais rapidamente e com segurança. O NetApp AIPod Mini combina a infraestrutura de dados inteligente da NetApp com as capacidades de gerenciamento de dados do ONTAP, processadores Intel Xeon 6, Intel® AI for Enterprise RAG e o conjunto 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 foi escrito por Sathish Thyagarajan, Michael Oglesby e Arpita Mahajan, membros da equipe de Engenharia de Soluções da NetApp. Os autores também gostariam de agradecer à equipe de produtos Enterprise AI da Intel—Ajay Mungara, Mikolaj Zyczynski, Igor Konopko, Ramakrishna Karamsetty, Michal Prostko, Anna Alberska, Maciej Cichocki, Shreejan Mistry, Nicholas Rago e Ned Fiori—bem como a outros membros da equipe da NetApp—Lawrence Bunka, Bobby Oommen e Jeff Liborio, pelo apoio e ajuda contínuos durante o processo de validação da 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-GNR6972P-SRPL2-UC | Processador Intel® Xeon® 6972P de 96 núcleos 2,40GHz 480MB de cache (500W) | 4 |
| BATER | MEM-DR564MC-ER64(x16)64GB DDR5-6400 2RX4 (16Gb) ECC RDIMM | 32 |
| | 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 |

# Lista de verificação de prontidão da infraestrutura

Consulte o NetApp AIPod Mini - Prontidão da Infraestrutura documento para obter mais detalhes.

# 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"

"Intel® AI ERAG Documentação"

"Manual de implantação do OPEA Enterprise RAG" == Histórico de versões

| Versão | Data | Histórico de versões do documento |
|---|---|---|
| Versão 1.0 | Setembro de 2025 | Lançamento inicial |
| Versão 2.0 | Fev 2026 | Atualizado com OPEA-Intel® AI for Enterprise RAG 2.0 |

## Informações sobre direitos autorais



## Informações sobre marcas comerciais