NetApp AIPod Mini: Inferencia RAG empresarial con NetApp e Intel

NetApp artificial intelligence solutions

NetApp
February 12, 2026

Tabla de contenidos

NetApp AIPod Mini: Inferencia RAG empresarial con NetApp e Intel . . . 1
- Resumen ejecutivo . . . 1
- Validación de socios de almacenamiento de Intel . . . 1
- Ventajas de ejecutar sistemas RAG con NetApp . . . 1
- Público objetivo . . . 2
- Requisitos tecnológicos . . . 2
  - Hardware . . . 2
  - Software . . . 4
- Implementación de la solución . . . 6
  - Pila de software . . . 6
  - Pasos de implementación . . . 6
- Guía de tallas . . . 12
- Conclusión . . . 13
- Reconocimiento . . . 13
- Lista de materiales . . . 13
- Lista de comprobación Infra Readiness . . . 15
- Dónde encontrar información adicional . . . 15

NetApp AIPod Mini: Inferencia RAG empresarial con NetApp e Intel

Este documento presenta un diseño de referencia validado de NetApp AIPod para Enterprise RAG con tecnologías y capacidades combinadas de procesadores Intel Xeon 6 y soluciones de gestión de datos de NetApp . La solución demuestra una aplicación ChatQnA que aprovecha un modelo de lenguaje amplio y brinda respuestas precisas y contextualmente relevantes a usuarios simultáneos. Las respuestas se recuperan del repositorio de conocimiento interno de una organización a través de una tubería de inferencia RAG con espacio de aire.

Sathish Thyagarajan, Michael Oglesby, Arpita Mahajan, NetApp

Resumen ejecutivo

Cada vez son más las organizaciones que aprovechan las aplicaciones de generación aumentada por recuperación (RAG) y los grandes modelos lingüísticos (LLMs) para interpretar las indicaciones de los usuarios y generar respuestas para aumentar la productividad y el valor empresarial. Estas indicaciones y respuestas pueden incluir texto, código, imágenes o incluso estructuras de proteínas terapéuticas recuperadas de la base de conocimientos interna de una organización, lagos de datos, repositorios de código y repositorios de documentos. Este documento cubre el diseño de referencia de la solución NetApp AIPod Mini, que incluye almacenamiento NetApp AFF y servidores con procesadores Intel Xeon 6. Incluye el software de gestión de datos NetApp ONTAP combinado con Intel Advanced Matrix Extensions (Intel AMX) y el software Intel® AI for Enterprise RAG construido sobre Open Platform for Enterprise AI (OPEA). El NetApp AIPod Mini para enterprise RAG permite a las organizaciones aumentar un LLM público en una solución privada de inferencia de IA generativa (GenAI). La solución demuestra una inferencia RAG eficiente y rentable a escala empresarial, diseñada para mejorar la fiabilidad y darte un mejor control sobre tu información propietaria.

Validación de socios de almacenamiento de Intel

Los servidores equipados con procesadores Intel Xeon 6 están diseñados para manejar cargas de trabajo de inferencia de IA exigentes, utilizando Intel AMX para lograr el máximo rendimiento. Para permitir un rendimiento y una escalabilidad de almacenamiento óptimos, la solución se ha validado con éxito utilizando NetApp ONTAP, lo que permite a las empresas satisfacer las necesidades de las aplicaciones RAG. Esta validación se realizó en servidores con procesadores Intel Xeon 6. Intel y NetApp tienen una sólida asociación centrada en ofrecer soluciones de IA optimizadas, escalables y alineadas con los requisitos comerciales del cliente.

Ventajas de ejecutar sistemas RAG con NetApp

Las aplicaciones RAG implican la recuperación de conocimientos de los repositorios de documentos de las empresas en varios tipos como PDF, texto, CSV o Excel. Estos datos normalmente se almacenan en soluciones como un almacenamiento de objetos S3 o NFS en las instalaciones como fuente de datos. NetApp ha sido líder en gestión de datos, movilidad de datos, gobernanza de datos y tecnologías de seguridad de datos en todo el ecosistema de edge, centro de datos y cloud. NetApp ONTAP data management proporciona

almacenamiento de clase empresarial para soportar varios tipos de cargas de trabajo de IA, incluyendo inferencias por lotes y en tiempo real, y ofrece algunos de los siguientes beneficios:

- Velocidad y escalabilidad. Puede manejar grandes conjuntos de datos a alta velocidad para el control de versiones con la capacidad de escalar el rendimiento y la capacidad de forma independiente.
- Acceso a datos. La compatibilidad con múltiples protocolos permite que las aplicaciones cliente lean datos mediante los protocolos de intercambio de archivos S3, NFS y SMB. Los buckets NAS de ONTAP S3 pueden facilitar el acceso a los datos en escenarios de inferencia LLM multimodal.
- Confiabilidad y confidencialidad. ONTAP proporciona protección de datos, protección autónoma contra ransomware (ARP) de NetApp integrada y aprovisionamiento dinámico de almacenamiento, y ofrece cifrado basado en software y hardware para mejorar la confidencialidad y la seguridad. ONTAP cumple con FIPS 140-2 para todas las conexiones SSL.

Público objetivo

Este documento está dirigido a tomadores de decisiones de IA, ingenieros de datos, líderes empresariales y ejecutivos departamentales que desean aprovechar una infraestructura diseñada para brindar soluciones empresariales RAG y GenAI. El conocimiento previo de inferencia de IA, LLM, Kubernetes, redes y sus componentes ayudará durante la fase de implementación.

Requisitos tecnológicos

Hardware

Tecnologías de Intel® AI

Con Xeon 6 como CPU host, los sistemas acelerados se benefician de un alto rendimiento de un solo subproceso, mayor ancho de banda de memoria, confiabilidad, disponibilidad y capacidad de servicio (RAS) mejoradas y más líneas de E/S. Intel AMX acelera la inferencia para INT8 y BF16 y ofrece soporte para modelos entrenados con FP16, con hasta 2048 operaciones de punto flotante por ciclo por núcleo para INT8 y 1024 operaciones de punto flotante por ciclo por núcleo para BF16/FP16. Para implementar una solución RAG utilizando procesadores Xeon 6, generalmente se recomienda una RAM mínima de 250 GB y 500 GB de espacio en disco. Sin embargo, esto depende en gran medida del tamaño del modelo LLM. Para obtener más información, consulte Intel "Procesador Xeon 6" Descripción del producto.

Figura 1 - Servidor de cómputo con procesadores Intel Xeon

6

Almacenamiento AFF de NetApp

Los sistemas NetApp AFF Serie A de nivel básico y medio ofrecen mayor rendimiento, densidad y eficiencia. Los sistemas NetApp AFF A20, AFF A30 y AFF A50 proporcionan un verdadero almacenamiento unificado que admite bloques, archivos y objetos, basado en un único sistema operativo que puede administrar, proteger y movilizar datos sin problemas para aplicaciones RAG al menor costo en la nube híbrida.

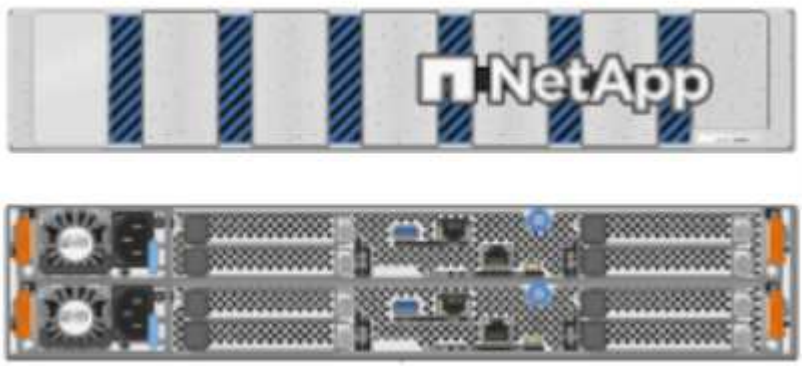


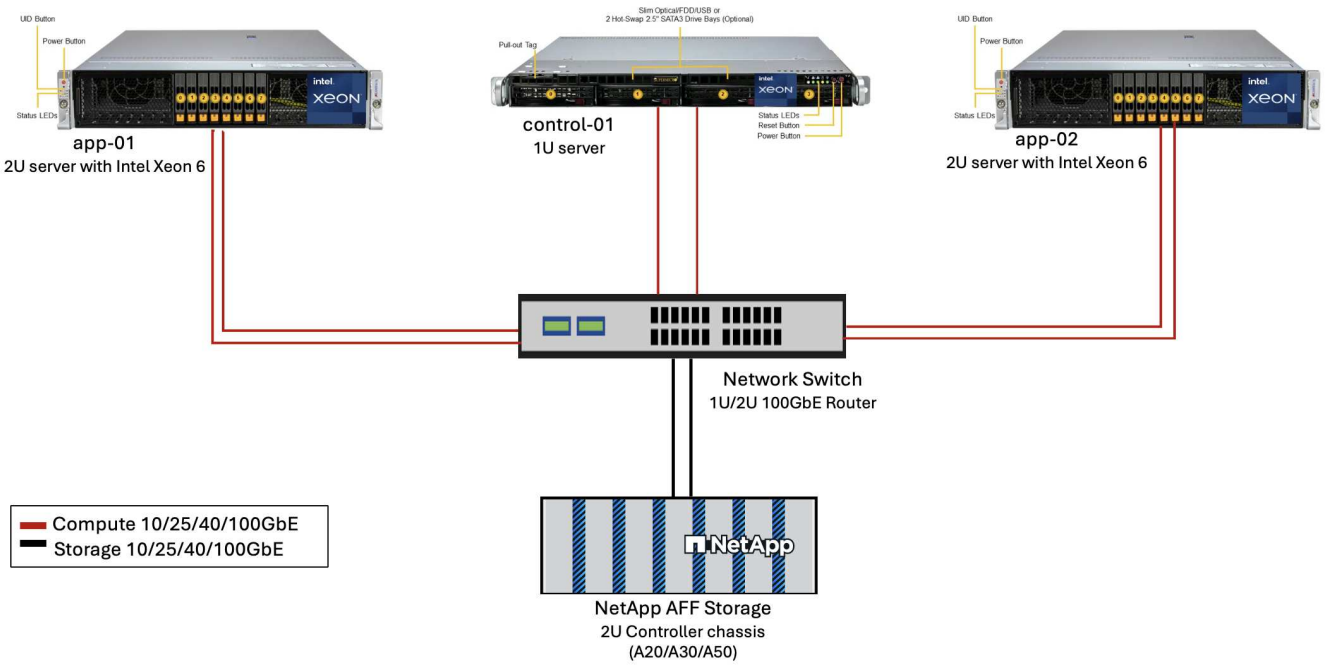
Figura 2 - Sistema NetApp AFF Serie A.

Hardware	Cantidad	Comentario
Intel Xeon 6th Gen (Granite Rapids)	2	Nodos de inferencia RAG, con procesadores Intel Xeon 6900-series de doble zócalo (96 núcleos) o Intel Xeon 6700-series (64 núcleos) y de 250GB a 3TB de RAM con DDR5 (6400MHz) o MRDIMM (8800MHz). Servidor 2U.
Servidor de plano de control con procesador Intel	1	Plano de control de Kubernetes/servidor 1U.

Hardware	Cantidad	Comentario
Elección de conmutador Ethernet de 100 Gb	1	Conmutador de centro de datos.
NetApp AFF A20 (o AFF A30; AFF A50)	1	Capacidad máxima de almacenamiento: 9.3PB. Nota: Redes: puertos 10/25/100 GbE.

Para la validación de este diseño de referencia se utilizaron servidores con procesadores Intel Xeon 6 de Supermicro (222HA-TN-OTO-37) y un switch 100GbE de Arista (7280R3A).

Figura 3 - Arquitectura de despliegue de AIPod Mini



Software

Plataforma abierta para IA empresarial

La Plataforma Abierta para IA Empresarial (OPEA) es una iniciativa de código abierto liderada por Intel en colaboración con socios del ecosistema. Proporciona una plataforma modular de bloques de construcción componibles diseñados para acelerar el desarrollo de sistemas de IA generativa de vanguardia, con un fuerte enfoque en RAG. OPEA incluye un marco integral que incluye LLM, almacenes de datos, motores de solicitud, planos arquitectónicos RAG y un método de evaluación de cuatro pasos que evalúa los sistemas de IA generativa en función del rendimiento, las características, la confiabilidad y la preparación empresarial.

En esencia, la OPEA consta de dos componentes clave:

- GenAIComps: un conjunto de herramientas basado en servicios compuesto por componentes de microservicios
- Ejemplos de GenAI: soluciones listas para implementar como ChatQnA que demuestran casos de uso prácticos

Para más detalles, consulte la "Documentación del proyecto OPEA"

Intel® AI for Enterprise RAG con tecnología OPEA

OPEA para Intel® AI for Enterprise RAG simplifica transformar los datos de tu empresa en información práctica. Impulsado por procesadores Intel Xeon, integra componentes de partners del sector para ofrecer un enfoque optimizado para poner en marcha soluciones empresariales. Se adapta perfectamente con marcos de orquestación probados, proporcionando la flexibilidad y la capacidad de elección que tu empresa necesita.

Partiendo de los cimientos de OPEA, Intel® AI for Enterprise RAG amplía esta base con funciones clave que mejoran la escalabilidad, la seguridad y la experiencia del usuario. Estas funciones incluyen capacidades de malla de servicios para una integración perfecta con arquitecturas modernas basadas en servicios, validación lista para producción para la fiabilidad de la canalización y una interfaz de usuario rica en funciones para RAG como servicio, permitiendo una gestión y supervisión sencillas de los flujos de trabajo. Además, Intel y el soporte de sus partners proporcionan acceso a un amplio ecosistema de soluciones, combinado con la gestión de identidades y accesos (IAM) integrada con la interfaz de usuario y las aplicaciones para operaciones seguras y conformes. Las barandillas programables proporcionan un control detallado sobre el comportamiento de la canalización, permitiendo personalizar los ajustes de seguridad y conformidad.

ONTAP de NetApp

NetApp ONTAP es la tecnología fundamental que sustenta las soluciones de almacenamiento de datos críticos de NetApp. ONTAP incluye varias funciones de gestión y protección de datos, como protección automática contra ransomware contra ciberataques, funciones de transporte de datos integradas y capacidades de eficiencia de almacenamiento. Estos beneficios se aplican a una variedad de arquitecturas, desde locales hasta multicloud híbrido en NAS, SAN, objetos y almacenamiento definido por software para implementaciones LLM. Puede utilizar un servidor de almacenamiento de objetos ONTAP S3 en un clúster ONTAP para implementar aplicaciones RAG, aprovechando la eficiencia de almacenamiento y la seguridad de ONTAP, proporcionada a través de usuarios autorizados y aplicaciones cliente. Para obtener más información, consulte "Obtenga más información sobre la configuración de ONTAP S3"

Trident de NetApp

El software NetApp Trident es un orquestador de almacenamiento de código abierto y totalmente compatible con contenedores y distribuciones de Kubernetes, incluido Red Hat OpenShift. Trident funciona con todo el portafolio de almacenamiento de NetApp , incluido NetApp ONTAP y también admite conexiones NFS e iSCSI. Para obtener más información, consulte "NetApp Trident en Git"

Software	Versión	Comentario
OPEA - Intel® AI for Enterprise RAG	2,0	Plataforma RAG empresarial basada en microservicios OPEA
Interfaz de almacenamiento de contenedores (controlador CSI)	NetApp Trident 25.10	Permite el aprovisionamiento dinámico, copias Snapshot de NetApp y volúmenes.
Ubuntu	22.04.5	SO en clúster de dos nodos.
Orquestación de contenedores	Kubernetes 1.31.9 (instalado por el playbook de infraestructura Enterprise RAG)	Entorno para ejecutar el marco RAG
ONTAP	ONTAP 9.16.1P4 o superior	Storage OS en AFF A20.

Implementación de la solución

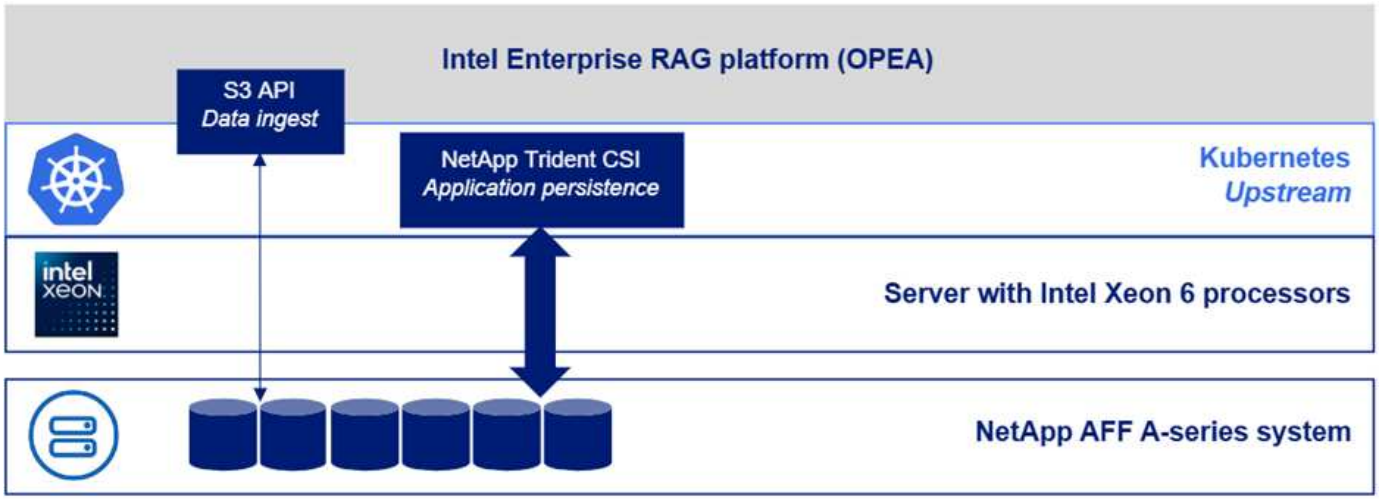
Pila de software

La solución se implementa en un clúster de Kubernetes que consta de nodos de aplicaciones basados en Intel Xeon. Se requieren al menos tres nodos para implementar alta disponibilidad básica para el plano de control de Kubernetes. Validamos la solución utilizando el siguiente diseño de clúster.

Tabla 3: Disposición del clúster de Kubernetes

Node	Role	Cantidad
Servidores con procesadores Intel Xeon 6 y 1 TB de RAM	Nodo de aplicación, nodo de plano de control	2
Servidor genérico	Nodo del plano de control	1

La siguiente figura muestra una "vista de la pila de software" de la solución.



Pasos de implementación

Implementar el dispositivo de almacenamiento ONTAP

Implemente y aprovisione su dispositivo de almacenamiento NetApp ONTAP . Consulte la "Documentación de los sistemas de hardware de ONTAP" Para más detalles.

Configurar una SVM de ONTAP para acceso NFS y S3

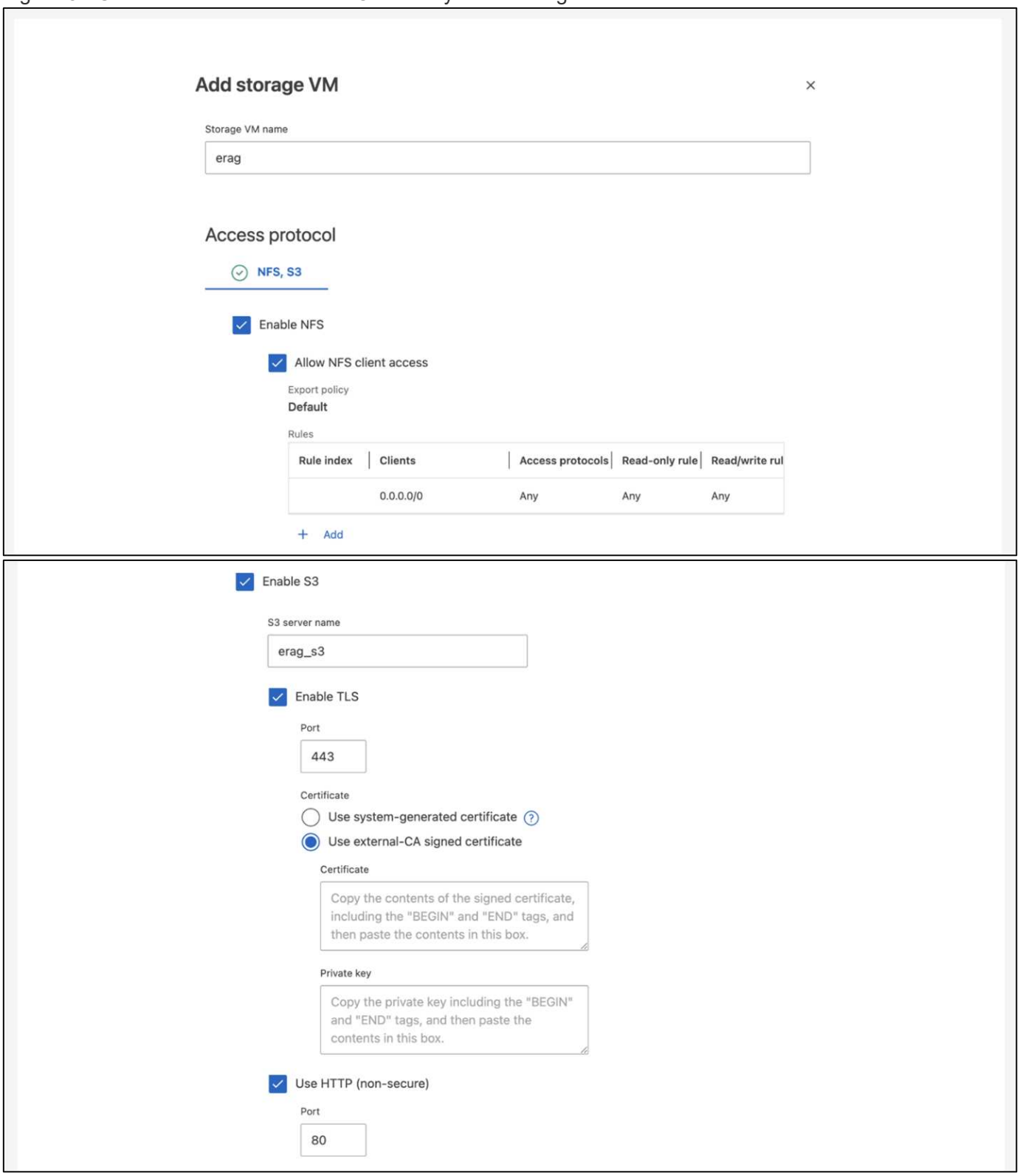
Configure una máquina virtual de almacenamiento ONTAP (SVM) para acceso NFS y S3 en una red a la que puedan acceder sus nodos de Kubernetes.

Para crear una SVM usando ONTAP System Manager, navegue a Almacenamiento > Máquinas virtuales de almacenamiento y haga clic en el botón + Agregar. Al habilitar el acceso S3 para su SVM, elija la opción para utilizar un certificado firmado por una CA externa (autoridad de certificación), no un certificado generado por el sistema. Puede utilizar un certificado autofirmado o un certificado firmado por una CA de confianza pública. Para obtener más detalles, consulte la "Documentación de ONTAP ."

La siguiente captura de pantalla muestra la creación de una SVM utilizando ONTAP System Manager.

Modifique los detalles según sea necesario en función de su entorno.

Figura 5 - Creación de SVM mediante ONTAP System Manager.



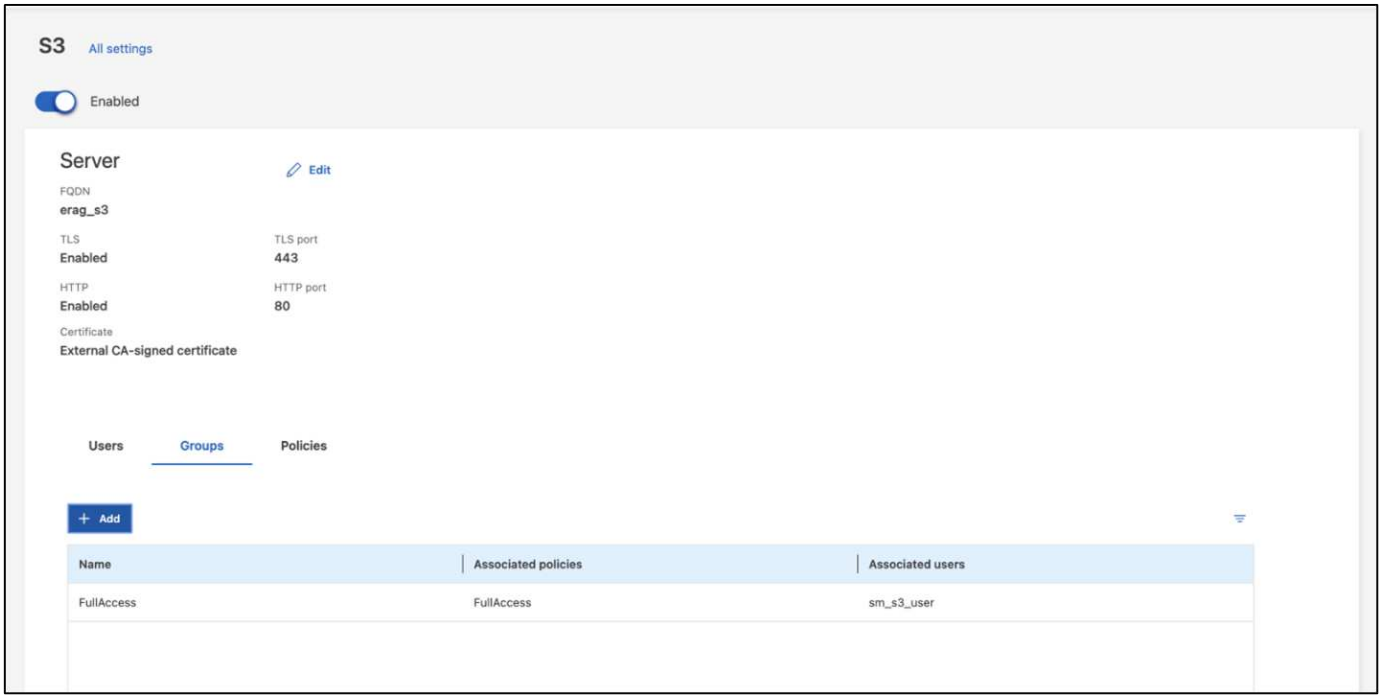
Configurar permisos de S3

Configure los ajustes de usuario/grupo S3 para la SVM que creó en el paso anterior. Asegúrese de tener un usuario con acceso completo a todas las operaciones de API S3 para esa SVM. Consulte la documentación de ONTAP S3 para obtener más detalles.

Nota: este usuario será necesario para el servicio de ingesta de datos de la aplicación Intel® AI for Enterprise RAG. Si creaste tu SVM usando ONTAP System Manager, System Manager habrá creado automáticamente un usuario llamado `sm_s3_user` y una política llamada `FullAccess` cuando creaste tu SVM, pero no se habrán asignado permisos a `sm_s3_user`.

Para editar los permisos de este usuario, vaya a Almacenamiento > Máquinas virtuales de almacenamiento, haga clic en el nombre de la SVM que creó en el paso anterior, haga clic en Configuración y, luego, haga clic en el ícono de lápiz junto a "S3". Dar `sm_s3_user` acceso completo a todas las operaciones de la API de S3, crear un nuevo grupo que asocie `sm_s3_user` con el `FullAccess` política como se muestra en la siguiente captura de pantalla.

Figura 6 - Permisos S3.

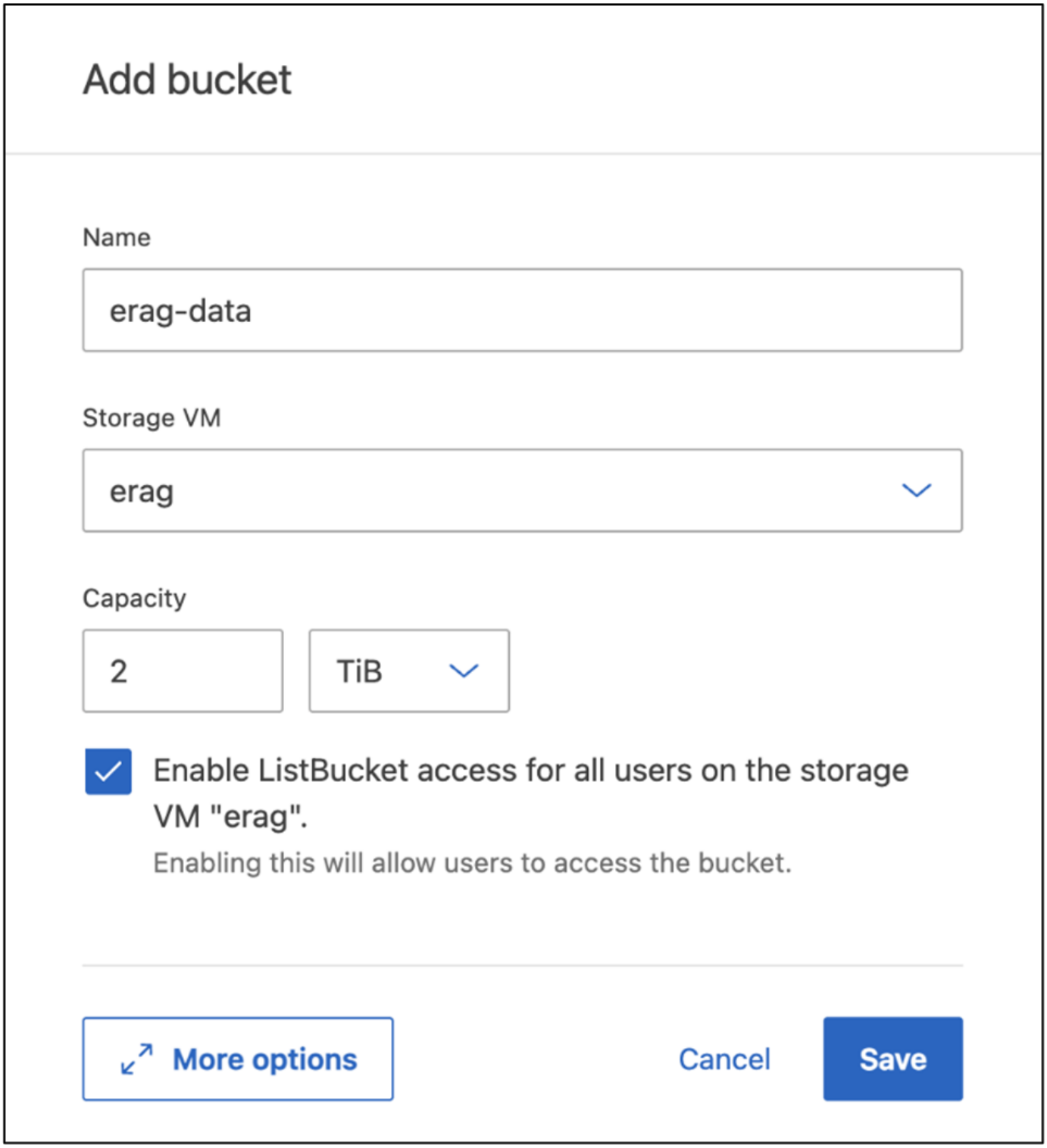


Crear un bucket S3

Crea un bucket S3 dentro del SVM que creaste anteriormente. Para crear una SVM usando ONTAP System Manager, navegue a Almacenamiento > Cubos y haga clic en el botón + Agregar. Para obtener detalles adicionales, consulte la documentación de ONTAP S3.

La siguiente captura de pantalla muestra la creación de un depósito S3 mediante ONTAP System Manager.

Figura 7 - Crear un bucket S3.



Configurar los permisos del bucket S3

Configure los permisos para el depósito S3 que creó en el paso anterior. Asegúrese de que el usuario que configuró en un paso anterior tenga los siguientes permisos: `GetObject, PutObject, DeleteObject, ListBucket, GetBucketAcl, GetObjectAcl, ListBucketMultipartUploads, ListMultipartUploadParts, GetObjectTagging, PutObjectTagging, DeleteObjectTagging, GetBucketLocation, GetBucketVersioning, PutBucketVersioning, ListBucketVersions, GetBucketPolicy, PutBucketPolicy, DeleteBucketPolicy, PutLifecycleConfiguration, GetLifecycleConfiguration, GetBucketCORS,`

`PutBucketCORS`.

Para editar los permisos de un depósito S3 mediante ONTAP System Manager, navegue a Almacenamiento > Depósitos, haga clic en el nombre de su depósito, haga clic en Permisos y, luego, haga clic en Editar. Consulte la "Documentación de ONTAP S3" Para más detalles.

La siguiente captura de pantalla muestra los permisos de depósito necesarios en ONTAP System Manager.

Figura 8 - Permisos del bucket S3.

Crear una regla de uso compartido de recursos de origen cruzado de bucket

Con la CLI de ONTAP , cree una regla de uso compartido de recursos de origen cruzado (CORS) para el depósito que creó en un paso 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 regla permite que OPEA for Intel® AI for Enterprise RAG web application interactúe con el bucket desde un navegador web.

Implementar servidores

Implemente sus servidores e instale Ubuntu 22.04 LTS en cada servidor. Después de instalar Ubuntu, instale las utilidades NFS en cada servidor. Para instalar las utilidades NFS, ejecute el siguiente comando:

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

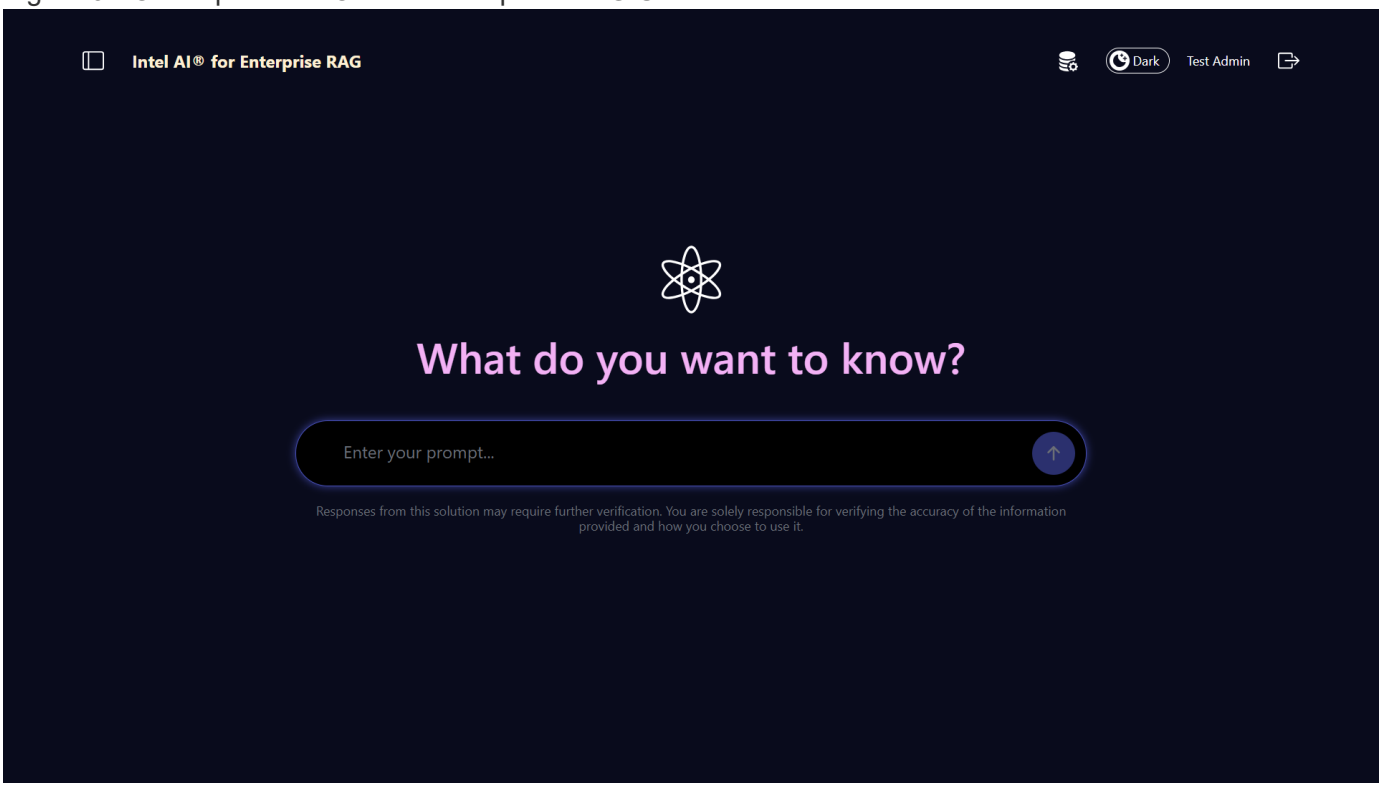
Implementa Enterprise RAG 2.0

Consulta el siguiente documento para ver el flujo de trabajo de puesta en marcha paso a paso: NetApp AIPod Mini para ERAG - Pasos de la implementación todos los requisitos previos, la preparación de la infraestructura, los parámetros de configuración y los procedimientos de puesta en marcha están documentados en la guía de puesta en marcha anterior.

Accede a la interfaz de usuario RAG de OPEA para Intel® AI for Enterprise

Accede a la RAG UI de OPEA para Intel® AI for Enterprise. Consulta el "Documentación de puesta en marcha de Intel® AI for Enterprise RAG" para más detalles.

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



Ingerir datos para RAG

Ahora puede ingerir archivos para incluirlos en la ampliación de consultas basada en RAG. Existen múltiples opciones para ingerir archivos. Elija la opción adecuada a sus necesidades.

Nota: después de que se ha ingestado un archivo, la OPEA para Intel® AI for Enterprise RAG application comprueba automáticamente si hay actualizaciones en el archivo y las ingiere en consecuencia.

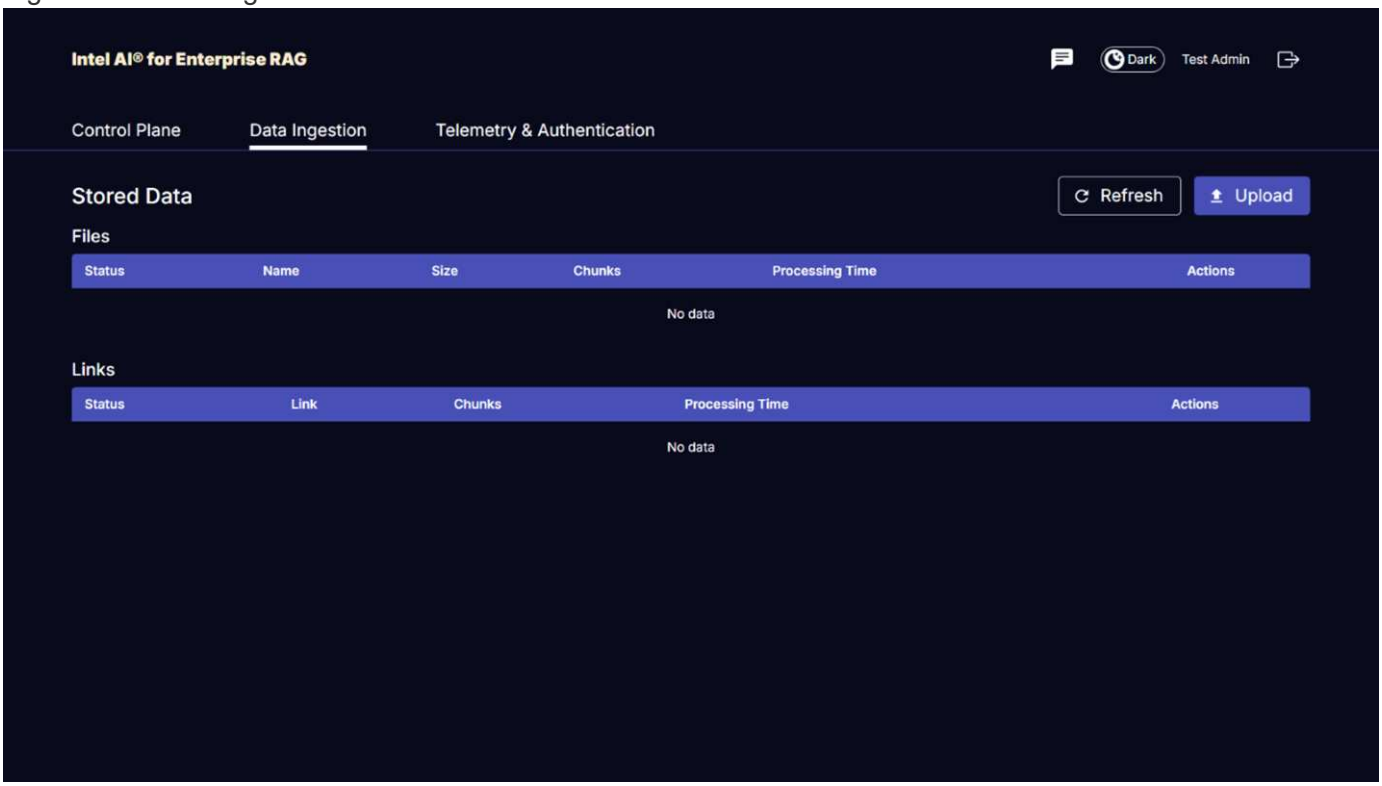
*Opción 1: subir directamente a tu bucket de S3 Para ingestar muchos archivos a la vez, te recomendamos subir los archivos a tu bucket de S3 (el bucket que creaste anteriormente) usando el cliente de S3 que prefieras. Los clientes de S3 más populares incluyen AWS CLI, Amazon SDK for Python (Boto3), s3cmd, S3 Browser, Cyberduck y Commander One. Si los archivos son de un tipo admitido, cualquier archivo que subas a tu bucket de S3 será ingerido automáticamente por la aplicación OPEA for Intel® AI for Enterprise RAG.

Nota: en el momento de redactar este documento, se admiten los siguientes tipos de archivo: PDF, HTML, TXT, DOC, DOCX, ADOC, PPT, PPTX, MD, XML, JSON, JSONL, YAML, XLS, XLSX, CSV, TIFF, JPG, JPEG, PNG y SVG.

Puedes usar la OPEA para Intel® AI for Enterprise RAG UI para confirmar que tus archivos se ingirieron correctamente. Consulta la documentación de Intel® AI for Enterprise RAG UI para más detalles. Ten en cuenta que la aplicación puede tardar un poco en ingerir una gran cantidad de archivos.

*Opción 2: carga mediante la interfaz de usuario Si necesitas cargar solo un pequeño número de archivos, puedes cargarlos usando la OPEA para Intel® AI for Enterprise RAG UI. Consulta la documentación de Intel® AI for Enterprise RAG UI para más detalles.

Figura 10 - UI de ingesta de datos.



Ejecutar consultas de chat

Ahora puedes "chatear" con la OPEA for Intel® AI for Enterprise RAG application usando la interfaz de chat incluida. Al responder a tus consultas, la aplicación realiza RAG usando tus archivos ingestados. Esto significa que la aplicación busca automáticamente información relevante dentro de tus archivos ingestados e incorpora esta información al responder a tus consultas.

Guía de tallas

Como parte de nuestro esfuerzo de validación, realizamos pruebas de rendimiento en coordinación con Intel. Esta prueba dio como resultado la guía de tamaño que se describe en la siguiente tabla.

Caracterizaciones	Valor	Comentario
Tamaño del modelo	20 mil millones de parámetros	Llama-8B, Llama-13B, Mistral 7B, Qwen 14B, DeepSeek Distill 8B
Tamaño de entrada	~2k tokens	~4 páginas
Tamaño de salida	~2k tokens	~4 páginas
Usuarios concurrentes	32	"Usuarios concurrentes" se refiere a solicitudes que envían consultas al mismo tiempo.

Nota: la guía de dimensionamiento presentada anteriormente se basa en la validación del rendimiento y en los resultados de pruebas realizadas con procesadores Intel Xeon 6 de 96 núcleos. Para clientes con requisitos de tokens de E/S y tamaño de modelo similares, recomendamos usar servidores con procesadores Xeon 6 de 96 núcleos. Para más detalles sobre la guía de dimensionamiento, consulta "Guía de dimensionamiento de Intel® AI for Enterprise RAG"

Conclusión

Los sistemas RAG empresariales y los LLM son tecnologías que trabajan conjuntamente para ayudar a las organizaciones a proporcionar respuestas precisas y conscientes del contexto. Estas respuestas implican la recuperación de información basada en una amplia colección de datos privados e internos de la empresa. Al usar RAG, APIs, incrustaciones vectoriales y sistemas de almacenamiento de alto rendimiento para consultar repositorios de documentos que contienen datos de la empresa, los datos se procesan más rápido y de forma segura. El NetApp AIPod Mini combina la infraestructura de datos inteligente de NetApp con las capacidades de gestión de datos de ONTAP y los procesadores Intel Xeon 6, Intel® AI for Enterprise RAG y la pila de software OPEA para ayudar a desplegar aplicaciones RAG de alto rendimiento y poner a las organizaciones en el camino hacia el liderazgo en IA.

Reconocimiento

Este documento es obra de Sathish Thyagarajan, Michael Oglesby y Arpita Mahajan, miembros del equipo de ingeniería de soluciones de NetApp. Los autores también quieren agradecer al equipo de producto de Enterprise AI en Intel—Ajay Mungara, Mikolaj Zyczynski, Igor Konopko, Ramakrishna Karamsetty, Michal Prostko, Anna Alberska, Maciej Cichocki, Shreejan Mistry, Nicholas Rago y Ned Fiori—así como a otros miembros del equipo en NetApp—Lawrence Bunka, Bobby Oommen y Jeff Liborio, por su apoyo y ayuda continuos durante el proceso de validación de la solución.

Lista de materiales

La siguiente fue la lista de materiales utilizada para la validación funcional de esta solución y puede usarse como referencia. Se podría utilizar cualquier servidor o componente de red (o incluso una red existente con un ancho de banda preferiblemente de 100 GbE) que se ajuste a la siguiente configuración.

Para el servidor de aplicaciones:

Número de pieza	Descripción del Producto	Cantidad
222HA-TN-OTO-37	Hyper SuperServer SYS-222HA-TN /2U	2
P4X-GNR6972P-SRPL2-UC	Procesador Intel® Xeon® 6972P de 96 núcleos 2,40GHz 480MB caché (500W)	4
RAM	MEM-DR564MC-ER64(x16) 64 GB DDR5-6400 2RX4 (16 GB) ECC RDIMM	32
	HDS-M2N4-960G0-E1-TXD-NON-080(x2) SSD M.2 NVMe PCIe4 960GB 1DWPD TLC D, 80 mm	2
	Fuente de alimentación de salida única redundante WS-1K63A-1R(x2)1U 692W/1600W. Disipación de calor de 2361 BTU/hora con temperatura máxima de 59 C (aprox.)	4

Para el servidor de control:

Número de pieza	Descripción del Producto	Cantidad
511R-M-OTO-17	OPTIMIZADO HASTA 1U X13SCH-SYS, CSE-813MF2TS-R0RCNBP, PWS-602A-1R	1
	RPL-E 6369P IP 8C/16T 3.3G 24MB 95W 1700 BO	1
RAM	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, 80 mm	2

Para el conmutador de red:

Número de pieza	Descripción del Producto	Cantidad
DCS-7280CR3A	Arista 7280R3A 28x100 GbE	1

Almacenamiento AFF de NetApp :

Número de pieza	Descripción del Producto	Cantidad
AFF-A20A-100-C	Sistema AFF A20 HA, -C	1
X800-42U-R6-C	Cable de puente, en cabina, C13-C14, -C	2
X97602A-C	Fuente de alimentación, 1600 W, titanio, -C	2
X66211B-2-N-C	Cable, 100 GbE, QSFP28-QSFP28, Cu, 2 m, -C	4
X66240A-05-N-C	Cable, 25 GbE, SFP28-SFP28, Cu, 0,5 m, -C	2
X5532A-N-C	Riel, 4 postes, delgado, agujero redondo/cuadrado, pequeño, ajustable, 24-32, -C	1
X4024A-2-A-C	Paquete de unidades 2X1,92 TB, NVMe4, SED, -C	6
X60130A-C	Módulo de E/S, 2PT, 100 GbE, -C	2
X60132A-C	Módulo de E/S, 4 PT, 10/25 GbE, -C	2
SW-ONTAPB-FLASH-A20-C	SW, paquete básico de ONTAP , por TB, Flash, A20, -C	23

Lista de comprobación Infra Readiness

Consulta el NetApp AIPod Mini - Preparación de infraestructura para más detalles.

Dónde encontrar información adicional

Para obtener más información sobre la información que se describe en este documento, revise los siguientes documentos y/o sitios web:

"Documentación de productos de NetApp"

"Proyecto OPEA"

"Intel® AI ERAG Documentación"

"Manual de implementación de OPEA Enterprise RAG" == Historial de versiones

Versión	Fecha	Historial de versiones del documento
Versión 1.0	Septiembre 2025	Lanzamiento inicial
Versión 2.0	Feb 2026	Actualizado con OPEA-Intel® AI for Enterprise RAG 2.0