NetApp AIPod Mini - Inferenza RAG aziendale con NetApp e Intel

NetApp artificial intelligence solutions

NetApp
February 12, 2026

Sommario

NetApp AIPod Mini - Inferenza RAG aziendale con NetApp e Intel

In questo documento viene presentato un progetto di riferimento convalidato di NetApp AIPod per Enterprise RAG con tecnologie e funzionalità combinate dei processori Intel Xeon 6 e delle soluzioni di gestione dati NetApp . La soluzione illustra un'applicazione ChatQnA downstream che sfrutta un ampio modello linguistico, fornendo risposte accurate e contestualmente rilevanti agli utenti simultanei. Le risposte vengono recuperate dal repository di conoscenze interno di un'organizzazione tramite una pipeline di inferenza RAG air-gapped.

Sathish Thyagarajan, Michael Oglesby, Arpita Mahajan, NetApp

Sintesi

Un numero crescente di organizzazioni sfrutta applicazioni di retrieval-augmented generation (RAG) e large language models (LLM) per interpretare i prompt degli utenti e generare risposte per aumentare la produttività e il valore di business. Questi prompt e risposte possono includere testo, codice, immagini o persino strutture proteiche terapeutiche recuperate dalla knowledge base interna di un'organizzazione, da data lakes, repository di codice e repository di documenti. Questo documento illustra il reference design della soluzione NetApp AIPod Mini, che comprende NetApp AFF storage e server con processori Intel Xeon 6. Include il software di gestione dei dati NetApp ONTAP combinato con Intel Advanced Matrix Extensions (Intel AMX) e il software Intel® AI for Enterprise RAG basato su Open Platform for Enterprise AI (OPEA). La soluzione NetApp AIPod Mini for enterprise RAG consente alle organizzazioni di aumentare un LLM pubblico in una soluzione privata di inferenza di intelligenza artificiale generativa (GenAI). La soluzione dimostra un'inferenza RAG efficiente e conveniente su scala enterprise, progettata per migliorare l'affidabilità e offrire un maggiore controllo sulle informazioni proprietarie.

Convalida del partner di storage Intel

I server basati su processori Intel Xeon 6 sono progettati per gestire carichi di lavoro di inferenza AI impegnativi, utilizzando Intel AMX per le massime prestazioni. Per consentire prestazioni di archiviazione e scalabilità ottimali, la soluzione è stata convalidata con successo utilizzando NetApp ONTAP, consentendo alle aziende di soddisfare le esigenze delle applicazioni RAG. Questa convalida è stata condotta su server dotati di processori Intel Xeon 6. Intel e NetApp hanno una solida partnership incentrata sulla fornitura di soluzioni di intelligenza artificiale ottimizzate, scalabili e allineate alle esigenze aziendali dei clienti.

Vantaggi dell'esecuzione di sistemi RAG con NetApp

Le applicazioni RAG prevedono il recupero di conoscenze dai repository di documenti aziendali in vari formati, come PDF, testo, CSV o Excel. Questi dati sono normalmente archiviati in soluzioni come uno storage a oggetti S3 o NFS on-premises come fonte per i dati. NetApp è stata leader nelle tecnologie di data management, mobilità dei dati, governance dei dati e sicurezza dei dati nell'ecosistema edge, data center e cloud. NetApp ONTAP data management fornisce storage enterprise-grade per supportare vari tipi di carichi di lavoro AI, inclusi batch e inferenza real-time, e offre alcuni dei seguenti vantaggi:

- Velocità e scalabilità. È possibile gestire grandi set di dati ad alta velocità per il controllo delle versioni, con la possibilità di scalare prestazioni e capacità in modo indipendente.
- Accesso ai dati. Il supporto multiprotocollo consente alle applicazioni client di leggere i dati utilizzando i protocolli di condivisione file S3, NFS e SMB. I bucket ONTAP S3 NAS possono facilitare l'accesso ai dati in scenari di inferenza LLM multimodali.
- Affidabilità e riservatezza. ONTAP offre protezione dei dati, protezione autonoma dai ransomware (ARP) NetApp integrata e provisioning dinamico dello storage, oltre a offrire crittografia basata sia su software che su hardware per migliorare la riservatezza e la sicurezza. ONTAP è conforme allo standard FIPS 140-2 per tutte le connessioni SSL.

Pubblico di destinazione

Questo documento è destinato ai decisori in materia di intelligenza artificiale, agli ingegneri dei dati, ai leader aziendali e ai dirigenti di reparto che desiderano sfruttare un'infrastruttura creata per fornire soluzioni RAG e GenAI aziendali. Una conoscenza pregressa dell'inferenza dell'IA, degli LLM, di Kubernetes e delle reti e dei relativi componenti sarà utile durante la fase di implementazione.

Requisiti tecnologici

Hardware

Tecnologie AI Intel®

Con Xeon 6 come CPU host, i sistemi accelerati traggono vantaggio da elevate prestazioni single-thread, maggiore larghezza di banda di memoria, maggiore affidabilità, disponibilità e manutenibilità (RAS) e più corsie I/O. Intel AMX accelera l'inferenza per INT8 e BF16 e offre supporto per modelli addestrati da FP16, con un massimo di 2.048 operazioni in virgola mobile per ciclo per core per INT8 e 1.024 operazioni in virgola mobile per ciclo per core per BF16/FP16. Per implementare una soluzione RAG utilizzando processori Xeon 6, in genere si consiglia una RAM minima di 250 GB e 500 GB di spazio su disco. Tuttavia, ciò dipende fortemente dalle dimensioni del modello LLM. Per ulteriori informazioni, fare riferimento a Intel "Processore Xeon 6" descrizione del prodotto.

Figura 1 - Server di elaborazione con processori Intel Xeon

6

Archiviazione NetApp AFF

I sistemi NetApp AFF A-Series di livello base e intermedio offrono prestazioni più potenti, densità e maggiore efficienza. I sistemi NetApp AFF A20, AFF A30 e AFF A50 forniscono un vero storage unificato che supporta blocchi, file e oggetti, basato su un singolo sistema operativo in grado di gestire, proteggere e mobilitare senza problemi i dati per le applicazioni RAG al costo più basso nel cloud ibrido.

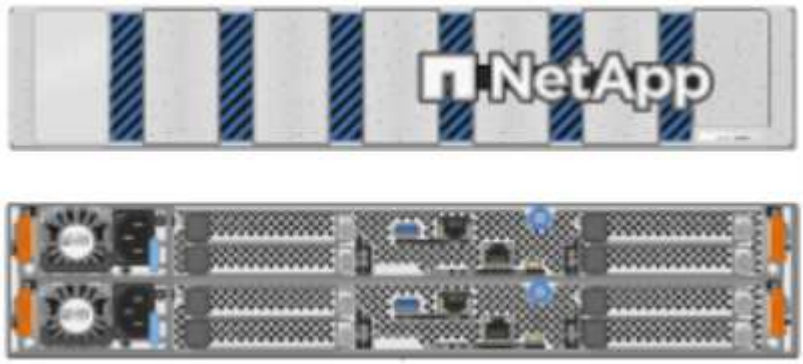


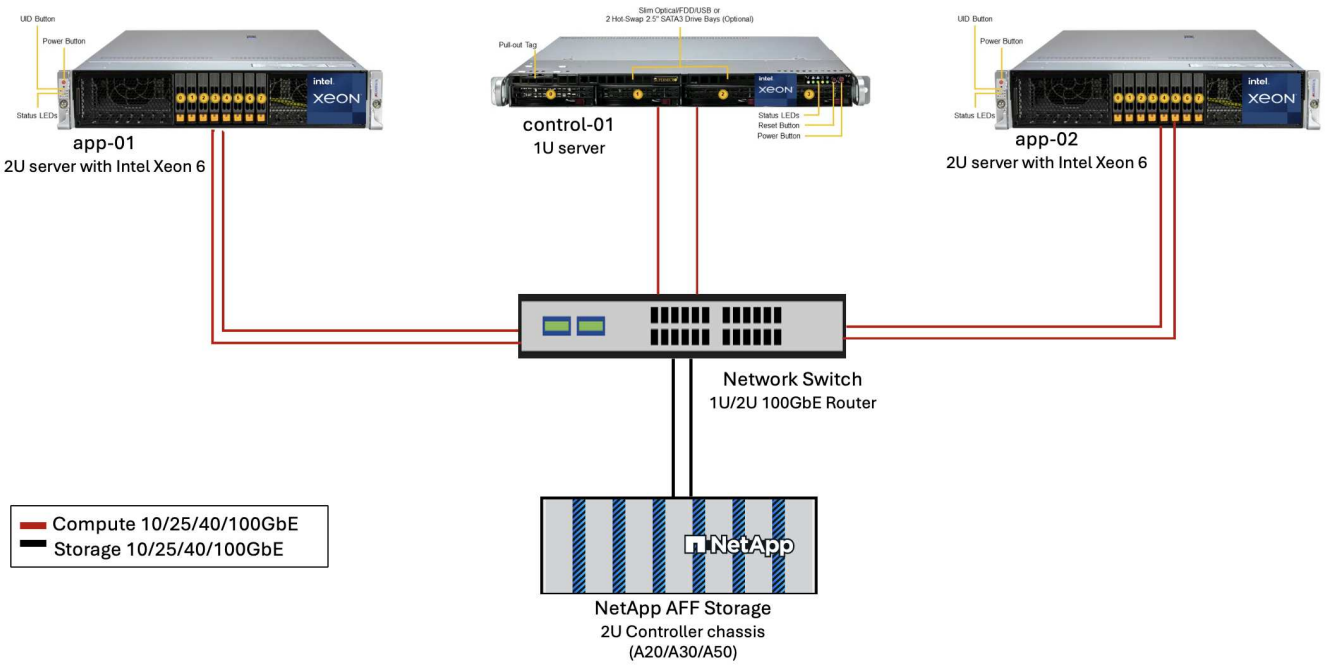
Figura 2 - Sistema NetApp AFF A-Series.

Hardware	Quantità	Commento
Intel Xeon 6th Gen (Granite Rapids)	2	Nodi di inferenza RAG—con processori Intel Xeon 6900-series (96 core) o Intel Xeon 6700-series (64 core) a doppio socket e RAM da 250GB a 3TB con DDR5 (6400MHz) o MRDIMM (8800MHz). Server 2U.
Server del piano di controllo con processore Intel	1	Piano di controllo Kubernetes/server 1U.

Hardware	Quantità	Commento
Scelta dello switch Ethernet da 100 Gb	1	Switch del centro dati.
NetApp AFF A20 (o AFF A30; AFF A50)	1	Capacità di archiviazione massima: 9,3 PB. Nota: Rete: porte 10/25/100 GbE.

Per la convalida di questo progetto di riferimento sono stati utilizzati server con processori Intel Xeon 6 di Supermicro (222HA-TN-OTO-37) e uno switch 100GbE di Arista (7280R3A).

Figura 3 - Architettura di distribuzione di AIPod Mini



Software

Piattaforma aperta per l'intelligenza artificiale aziendale

L'Open Platform for Enterprise AI (OPEA) è un'iniziativa open source guidata da Intel in collaborazione con i partner dell'ecosistema. Fornisce una piattaforma modulare di blocchi di costruzione componibili progettati per accelerare lo sviluppo di sistemi di intelligenza artificiale generativa all'avanguardia, con una forte attenzione al RAG. OPEA include un framework completo che comprende LLM, datastore, motori di prompt, progetti architettonici RAG e un metodo di valutazione in quattro fasi che valuta i sistemi di intelligenza artificiale generativa in base a prestazioni, funzionalità, affidabilità e prontezza aziendale.

L'OPEA è costituito essenzialmente da due componenti chiave:

- GenAIComps: un toolkit basato sui servizi composto da componenti di microservizi
- GenAIExamples: soluzioni pronte per l'implementazione come ChatQnA che dimostrano casi d'uso pratici

Per maggiori dettagli, vedere il "Documentazione del progetto OPEA"

Intel® AI for Enterprise RAG basato su OPEA

OPEA per Intel® AI for Enterprise RAG semplifica la trasformazione dei dati aziendali in informazioni fruibili. Basato su processori Intel Xeon, integra componenti di partner del settore per offrire un approccio semplificato all'implementazione di soluzioni enterprise. Si adatta perfettamente a framework di orchestrazione collaudati, offrendo la flessibilità e la scelta di cui la tua enterprise ha bisogno.

Basandosi sulle fondamenta di OPEA, Intel® AI for Enterprise RAG estende questa base con funzionalità chiave che migliorano scalabilità, sicurezza ed esperienza utente. Queste funzionalità includono service mesh per un'integrazione perfetta con le moderne architetture basate sui servizi, convalida pronta per la produzione per l'affidabilità della pipeline e un'interfaccia utente ricca di funzionalità per RAG as a service, che consente una facile gestione e monitoraggio dei workflow. Inoltre, il supporto di Intel e dei partner fornisce accesso a un ampio ecosistema di soluzioni, combinato con Identity and Access Management (IAM) integrato con UI e applicazioni per operazioni sicure e conformi. Guardrail programmabili forniscono un controllo granulare sul comportamento della pipeline, consentendo impostazioni di sicurezza e conformità personalizzate.

NetApp ONTAP

NetApp ONTAP è la tecnologia fondamentale su cui si fondano le soluzioni di archiviazione dati critici di NetApp. ONTAP include diverse funzionalità di gestione e protezione dei dati, come la protezione automatica contro i ransomware e gli attacchi informatici, funzionalità integrate di trasporto dei dati e capacità di efficienza di archiviazione. Questi vantaggi si applicano a una vasta gamma di architetture, da quelle on-premise a quelle multicloud ibride in NAS, SAN, storage object-defined e software-defined per distribuzioni LLM. È possibile utilizzare un server di archiviazione oggetti ONTAP S3 in un cluster ONTAP per distribuire applicazioni RAG, sfruttando l'efficienza di archiviazione e la sicurezza di ONTAP, fornite tramite utenti autorizzati e applicazioni client. Per maggiori informazioni, fare riferimento a "Scopri di più sulla configurazione ONTAP S3"

NetApp Trident

Il software NetApp Trident è un orchestratore di storage open source e completamente supportato per container e distribuzioni Kubernetes, tra cui Red Hat OpenShift. Trident funziona con l'intero portfolio di storage NetApp , incluso NetApp ONTAP , e supporta anche connessioni NFS e iSCSI. Per maggiori informazioni, fare riferimento a "NetApp Trident su Git"

Software	Versione	Commento
OPEA - Intel® AI for Enterprise RAG	2,0	Piattaforma RAG aziendale basata sui microservizi OPEA
Interfaccia di archiviazione del contenitore (driver CSI)	NetApp Trident 25.10	Abilita il provisioning dinamico, le copie Snapshot NetApp e i volumi.
Ubuntu	22.04.5	OS su cluster a due nodi.
Orchestrazione dei contenitori	Kubernetes 1.31.9 (installato dal playbook dell'infrastruttura Enterprise RAG)	Ambiente per eseguire il framework RAG
ONTAP	ONTAP 9.16.1P4 o superiore	Storage OS su AFF A20.

Distribuzione della soluzione

Stack software

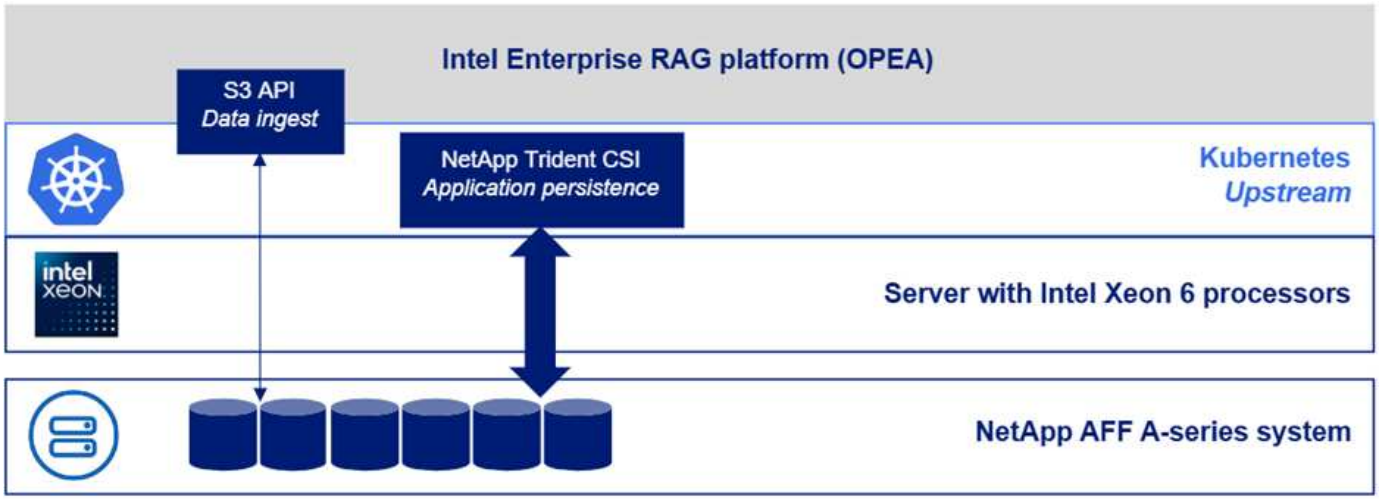
La soluzione è distribuita su un cluster Kubernetes costituito da nodi applicativi basati su Intel Xeon. Per

implementare l'alta disponibilità di base per il piano di controllo Kubernetes sono necessari almeno tre nodi. Abbiamo convalidato la soluzione utilizzando il seguente layout di cluster.

Tabella 3 - Layout del cluster Kubernetes

Nodo	Ruolo	Quantità
Server con processori Intel Xeon 6 e 1 TB di RAM	Nodo app, nodo piano di controllo	2
Server generico	Nodo del piano di controllo	1

La figura seguente illustra una "vista dello stack software" della soluzione.



Fasi di distribuzione

Distribuisci l'appliance di archiviazione ONTAP

Distribuisci e fornisci il tuo dispositivo di storage NetApp ONTAP . Fare riferimento al "Documentazione dei sistemi hardware ONTAP" per i dettagli.

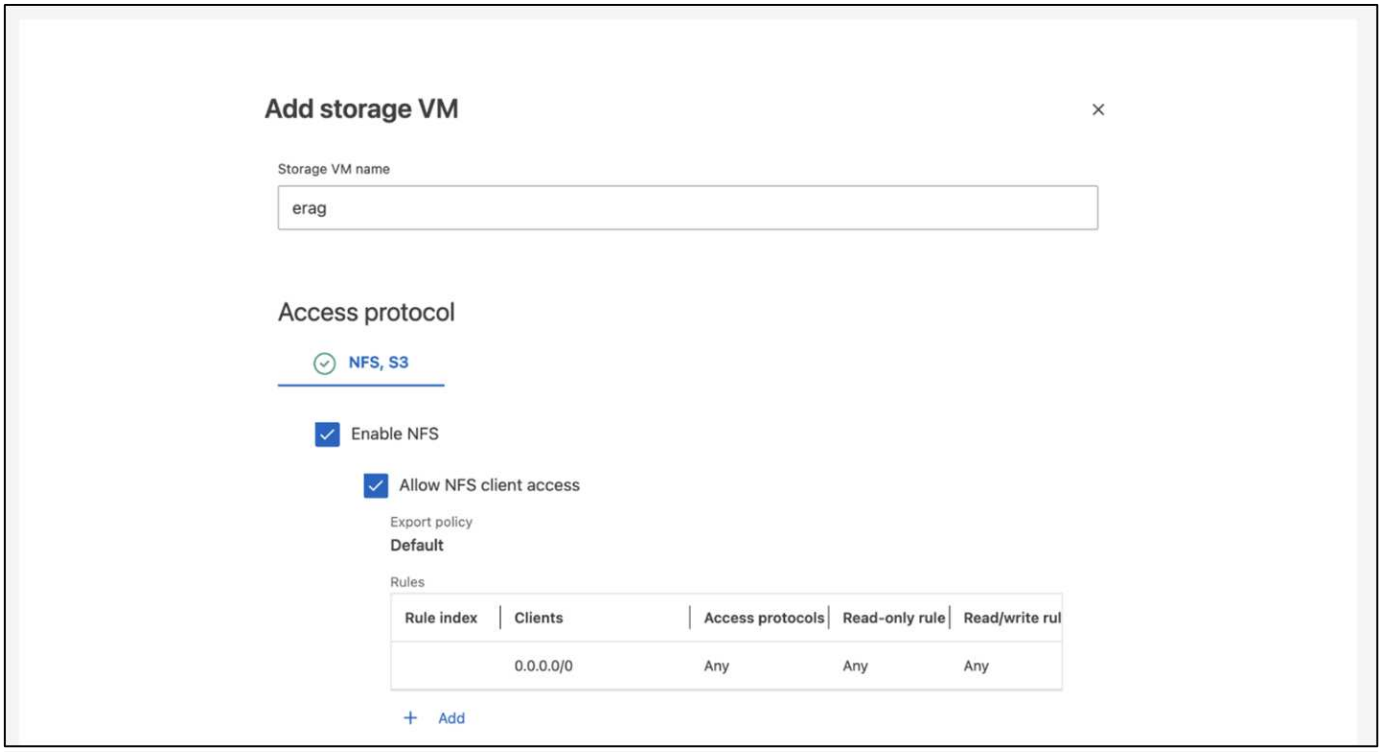
Configurare un ONTAP SVM per l'accesso NFS e S3

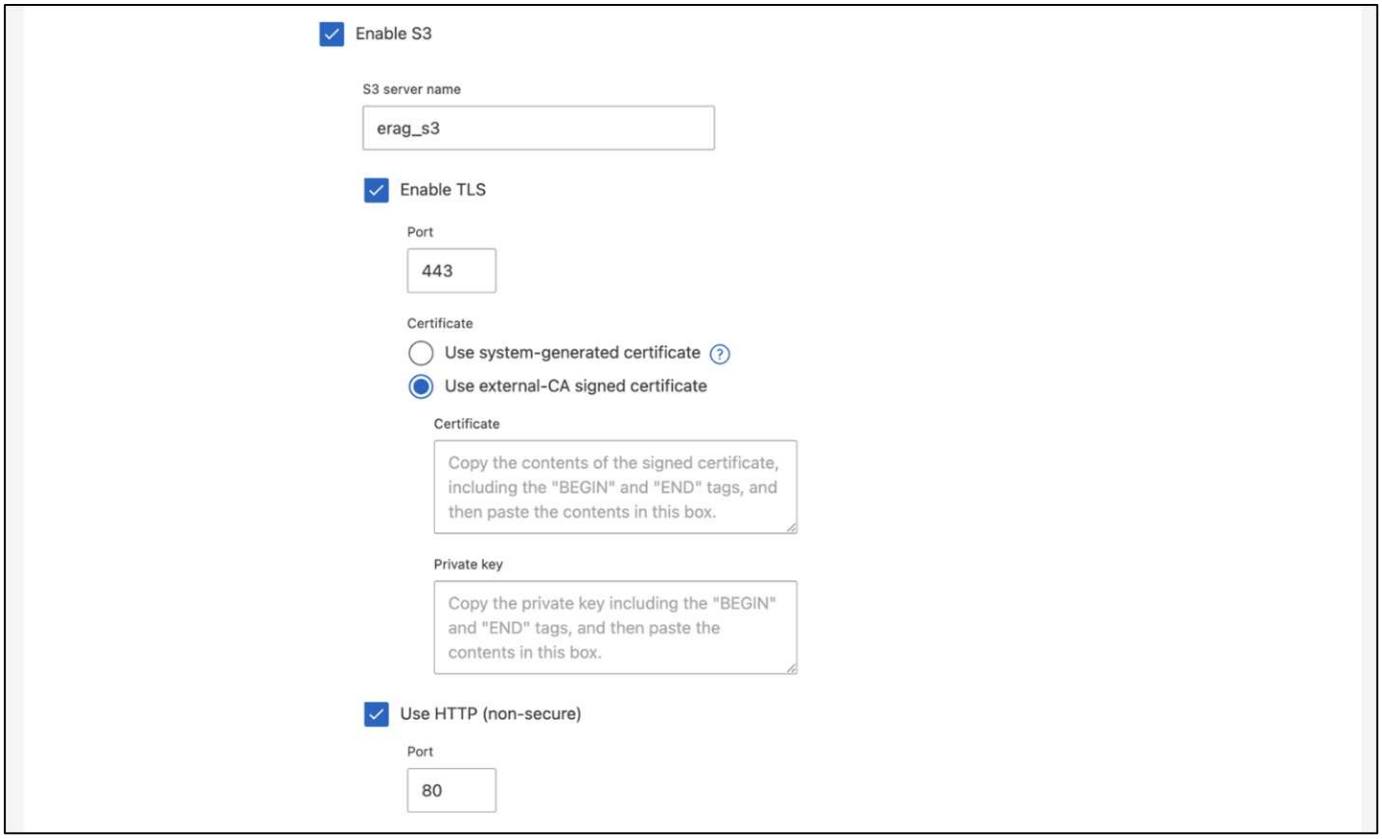
Configurare una macchina virtuale di archiviazione ONTAP (SVM) per l'accesso NFS e S3 su una rete accessibile dai nodi Kubernetes.

Per creare una SVM utilizzando ONTAP System Manager, accedere a Storage > Storage VM e fare clic sul pulsante + Aggiungi. Quando si abilita l'accesso S3 per la SVM, scegliere l'opzione per utilizzare un certificato firmato da una CA (autorità di certificazione) esterna, non un certificato generato dal sistema. È possibile utilizzare un certificato autofirmato oppure un certificato firmato da una CA pubblicamente attendibile. Per ulteriori dettagli, fare riferimento al "Documentazione ONTAP ."

La seguente schermata illustra la creazione di una SVM utilizzando ONTAP System Manager. Modifica i dettagli in base alle tue esigenze in base all'ambiente.

Figura 5 - Creazione di SVM tramite ONTAP System Manager.





Configurare le autorizzazioni S3

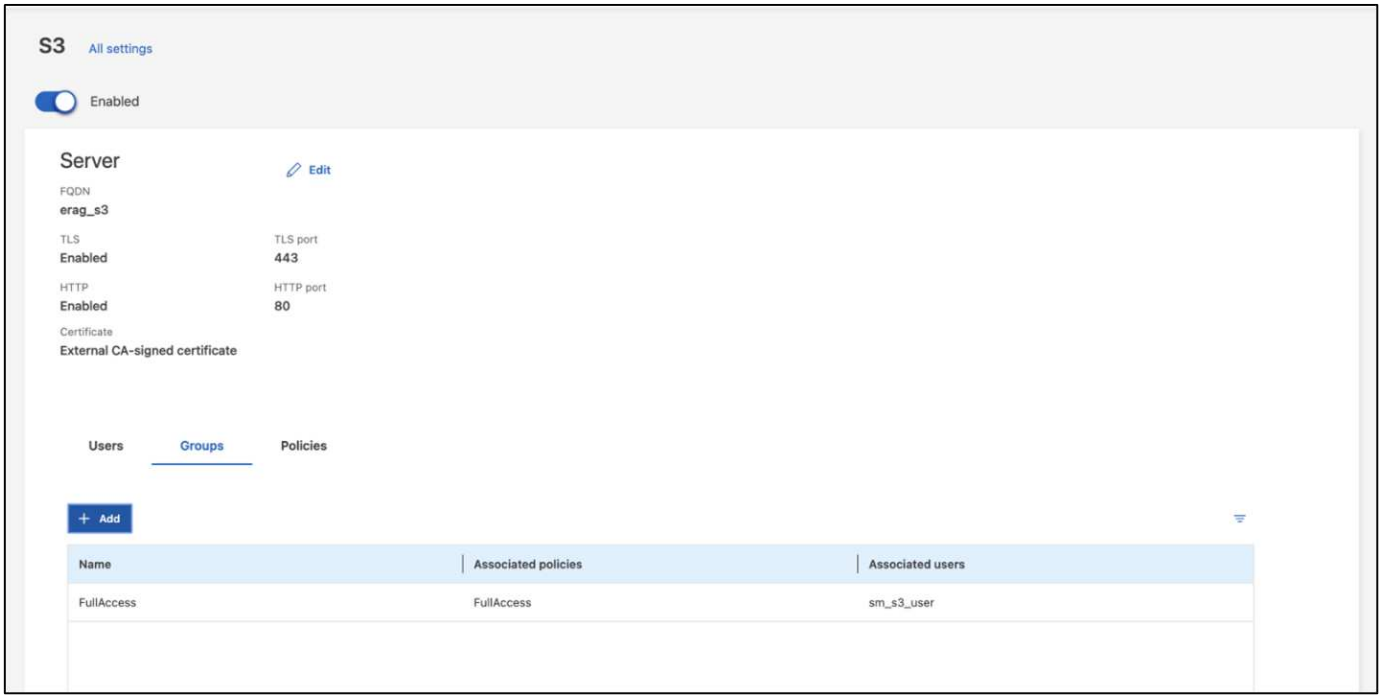
Configurare le impostazioni utente/gruppo S3 per l'SVM creato nel passaggio precedente. Assicurati di avere un utente con accesso completo a tutte le operazioni API S3 per quella SVM. Per maggiori dettagli, consultare la documentazione di ONTAP S3.

Nota: questo utente sarà necessario per il servizio di acquisizione dati dell'applicazione Intel® AI for Enterprise RAG. Se hai creato la tua SVM utilizzando ONTAP System Manager, System Manager avrà creato

automaticamente un utente `sm_s3_user` e una policy denominata `FullAccess` quando hai creato la tua SVM, ma non saranno state assegnate autorizzazioni a `sm_s3_user`.

Per modificare le autorizzazioni per questo utente, vai su Archiviazione > VM di archiviazione, fai clic sul nome della SVM creata nel passaggio precedente, fai clic su Impostazioni, quindi fai clic sull'icona della matita accanto a "S3". Per dare `sm_s3_user` accesso completo a tutte le operazioni API S3, crea un nuovo gruppo che associa `sm_s3_user` con il `FullAccess` politica come illustrato nello screenshot seguente.

Figura 6 - Autorizzazioni S3.

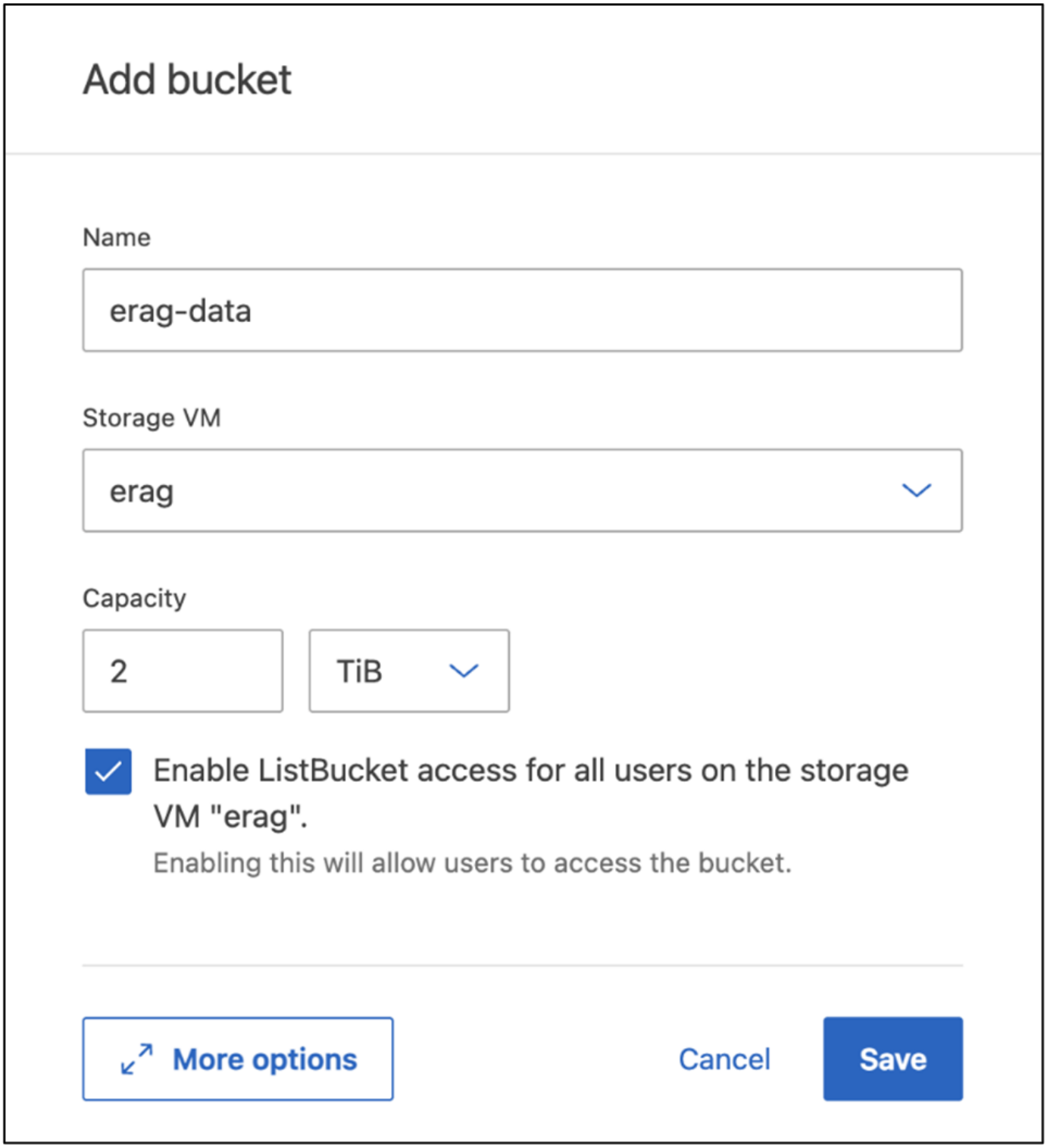


Crea un bucket S3

Crea un bucket S3 all'interno dell'SVM creato in precedenza. Per creare una SVM utilizzando ONTAP System Manager, vai su Storage > Bucket e fai clic sul pulsante + Aggiungi. Per ulteriori dettagli, fare riferimento alla documentazione ONTAP S3.

La seguente schermata illustra la creazione di un bucket S3 utilizzando ONTAP System Manager.

Figura 7 - Crea un bucket S3.



Configurare le autorizzazioni del bucket S3

Configurare le autorizzazioni per il bucket S3 creato nel passaggio precedente. Assicurati che l'utente configurato in un passaggio precedente disponga delle seguenti autorizzazioni: `GetObject, PutObject, DeleteObject, ListBucket, GetBucketAcl, GetObjectAcl, ListBucketMultipartUploads, ListMultipartUploadParts, GetObjectTagging, PutObjectTagging, DeleteObjectTagging, GetBucketLocation, GetBucketVersioning, PutBucketVersioning, ListBucketVersions, GetBucketPolicy, PutBucketPolicy, DeleteBucketPolicy, PutLifecycleConfiguration, GetLifecycleConfiguration, GetBucketCORS,`

`PutBucketCORS`.

Per modificare le autorizzazioni del bucket S3 tramite ONTAP System Manager, accedere a Storage > Bucket, fare clic sul nome del bucket, fare clic su Autorizzazioni e quindi su Modifica. Fare riferimento al "Documentazione ONTAP S3" per ulteriori dettagli.

La seguente schermata illustra le autorizzazioni bucket necessarie in ONTAP System Manager.

Figura 8 - Autorizzazioni del bucket S3.

Crea una regola di condivisione delle risorse multiorigine del bucket

Utilizzando l'interfaccia della riga di comando ONTAP , crea una regola CORS (cross-origin resource sharing) per il bucket creato in un passaggio precedente:

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

Questa regola consente a OPEA per Intel® AI for Enterprise RAG web application di interagire con il bucket dall'interno di un browser web.

Distribuisci server

Distribuisci i tuoi server e installa Ubuntu 22.04 LTS su ogni server. Dopo aver installato Ubuntu, installare le utility NFS su ogni server. Per installare le utilità NFS, eseguire il seguente comando:

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

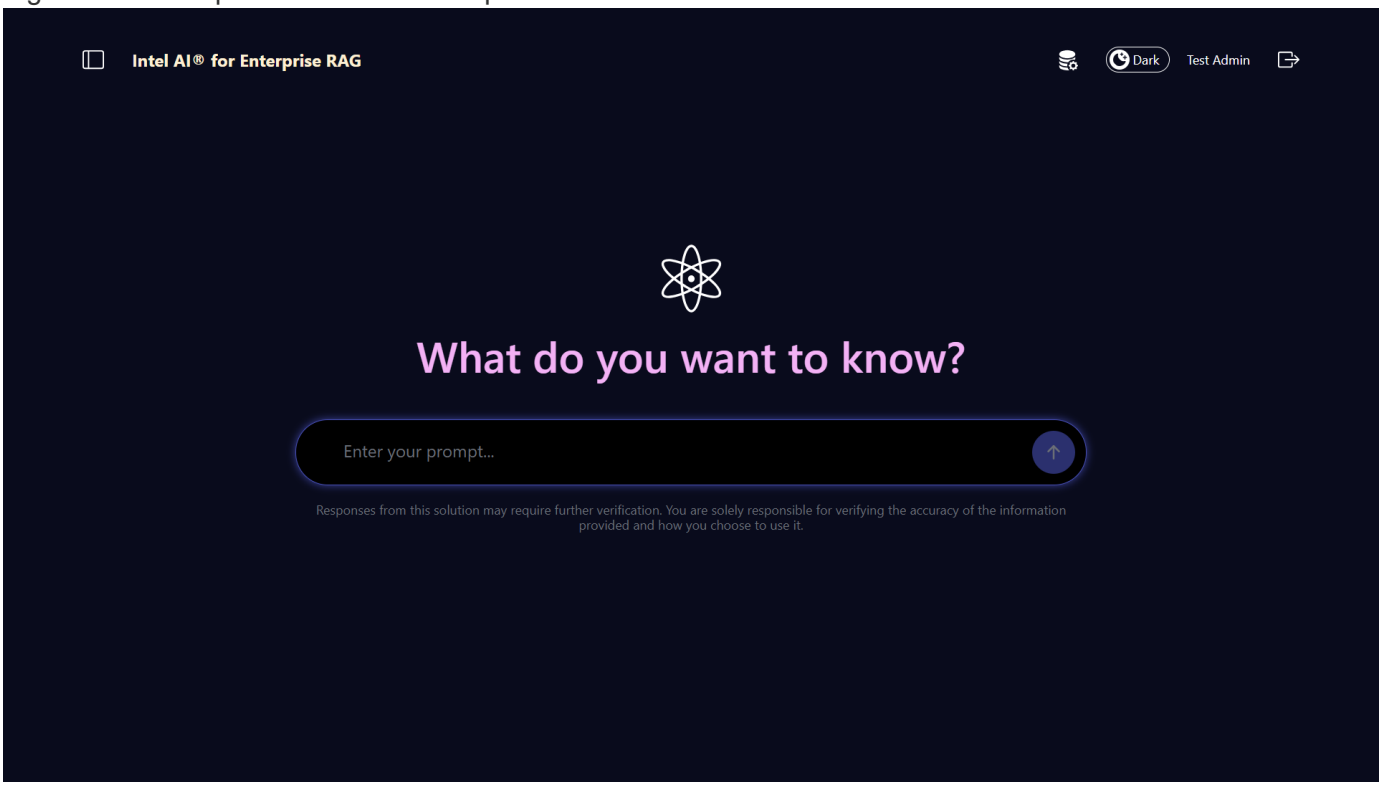
Distribuisci Enterprise RAG 2.0

Fare riferimento al seguente documento per un flusso di lavoro di distribuzione completo e dettagliato: NetApp AIPod Mini per ERAG - Fasi di deployment Tutti i prerequisiti, la preparazione dell'infrastruttura, i parametri di configurazione e le procedure di distribuzione sono documentati nella guida alla distribuzione sopra riportata.

Accedi a OPEA per Intel® AI for Enterprise RAG UI

Accedi all'OPEA per Intel® AI for Enterprise RAG UI. Consulta la "Documentazione di distribuzione di Intel® AI for Enterprise RAG" per i dettagli.

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



Acquisizione dati per RAG

Ora è possibile acquisire file da includere nell'aumento delle query basato su RAG. Esistono diverse opzioni per l'acquisizione dei file. Scegli l'opzione più adatta alle tue esigenze.

Nota: dopo che un file è stato acquisito, l'OPEA per Intel® AI for Enterprise RAG application controlla automaticamente la presenza di aggiornamenti al file e acquisisce gli aggiornamenti di conseguenza.

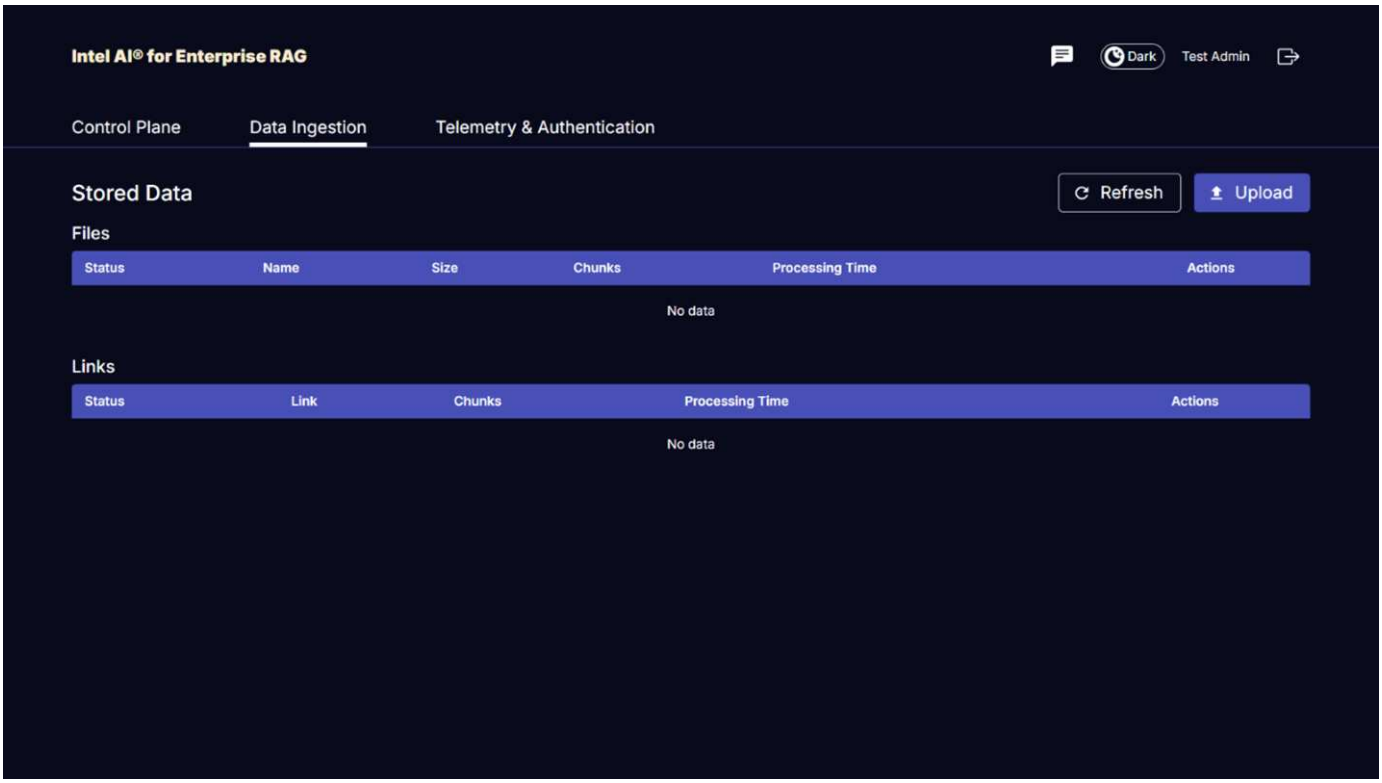
*Opzione 1: Carica direttamente nel tuo bucket S3 Per importare più file contemporaneamente, ti consigliamo di caricarli nel tuo bucket S3 (quello creato in precedenza) utilizzando il client S3 che preferisci. Tra i client S3 più diffusi figurano AWS CLI, Amazon SDK for Python (Boto3), s3cmd, S3 Browser, Cyberduck e Commander One. Se i file sono di un tipo supportato, tutti i file che carichi nel tuo bucket S3 verranno automaticamente importati dall'applicazione OPEA for Intel® AI for Enterprise RAG.

Nota: al momento della stesura di questo documento, sono supportati i seguenti tipi di file: PDF, HTML, TXT, DOC, DOCX, ADOC, PPT, PPTX, MD, XML, JSON, JSONL, YAML, XLS, XLSX, CSV, TIFF, JPG, JPEG, PNG, and SVG.

È possibile utilizzare l'OPEA for Intel® AI for Enterprise RAG UI per confermare che i file siano stati correttamente acquisiti. Consultare la documentazione di Intel® AI for Enterprise RAG UI per i dettagli. Si noti che l'applicazione potrebbe impiegare del tempo per acquisire un numero elevato di file.

*Opzione 2: Caricamento tramite l'UI Se devi importare solo un numero limitato di file, puoi importarli tramite l'OPEA for Intel® AI for Enterprise RAG UI. Consulta la documentazione di Intel® AI for Enterprise RAG UI per i dettagli.

Figura 10 - Interfaccia utente per l'acquisizione dei dati.



Eseguire query di chat

Ora puoi "chattare" con l'OPEA per Intel® AI for Enterprise RAG application utilizzando l'interfaccia utente di chat inclusa. Quando risponde alle tue query, l'applicazione esegue la RAG utilizzando i tuoi file acquisiti. Ciò significa che l'applicazione cerca automaticamente le informazioni rilevanti all'interno dei tuoi file acquisiti e incorpora queste informazioni quando risponde alle tue query.

Guida alle taglie

Nell'ambito del nostro impegno di convalida, abbiamo condotto test delle prestazioni in coordinamento con Intel. Da questi test sono emerse le indicazioni sulle dimensioni riportate nella tabella seguente.

Caratterizzazioni	Valore	Commento
Dimensioni del modello	20 miliardi di parametri	Llama-8B, Llama-13B, Mistral 7B, Qwen 14B, DeepSeek Distill 8B
Dimensione di input	~2k gettoni	~4 pagine
Dimensione di output	~2k gettoni	~4 pagine
Utenti simultanei	32	Per "utenti simultanei" si intendono le richieste rapide che inviano query contemporaneamente.

Nota: le indicazioni sul dimensionamento presentate sopra si basano sulla convalida delle prestazioni e sui risultati dei test raccolti utilizzando processori Intel Xeon 6 con 96 core. Per i clienti con token I/O e requisiti di dimensioni del modello simili, consigliamo di utilizzare server con processori Xeon 6 con 96 core. Per ulteriori dettagli sulla guida al dimensionamento, fare riferimento a "Guida al dimensionamento Intel® AI for Enterprise RAG"

Conclusione

I sistemi RAG enterprise e gli LLM sono tecnologie che lavorano insieme per aiutare le organizzazioni a fornire risposte accurate e consapevoli del contesto. Queste risposte implicano il recupero di informazioni basato su una vasta raccolta di dati aziendali privati e interni. Utilizzando RAG, API, incorporamenti vettoriali e sistemi di storage dalle performance elevate per interrogare repository di documenti che contengono dati aziendali, i dati vengono elaborati più rapidamente e in modo sicuro. Il NetApp AIPod Mini combina l'infrastruttura dati intelligente di NetApp con le funzionalità di gestione dei dati ONTAP, i processori Intel Xeon 6, Intel® AI for Enterprise RAG e lo stack software OPEA per aiutare a implementare applicazioni RAG dalle performance elevate e indirizzare le organizzazioni verso la leadership nell'AI.

Riconoscimento

Questo documento è stato redatto da Sathish Thyagarajan, Michael Oglesby e Arpita Mahajan, membri del team di NetApp Solutions Engineering. Gli autori desiderano inoltre ringraziare il team di prodotto Enterprise AI di Intel—Ajay Mungara, Mikolaj Zyczynski, Igor Konopko, Ramakrishna Karamsetty, Michal Prostko, Anna Alberska, Maciej Cichocki, Shreejan Mistry, Nicholas Rago e Ned Fiori—nonché gli altri membri del team di NetApp—Lawrence Bunka, Bobby Oommen e Jeff Liborio, per il loro continuo supporto e aiuto durante il processo di validazione della soluzione.

distinta base

Di seguito è riportato il BOM utilizzato per la convalida funzionale di questa soluzione e può essere utilizzato come riferimento. È possibile utilizzare qualsiasi server o componente di rete (o anche una rete esistente con larghezza di banda preferibilmente di 100 GbE) che si allinei alla seguente configurazione.

Per il server dell'app:

Codice parte	Descrizione del prodotto	Quantità
222HA-TN-OTO-37	Hyper SuperServer SYS-222HA-TN /2U	2
P4X-GNR6972P-SRPL2-UC	Processore Intel® Xeon® 6972P 96-Core 2.40GHz 480MB Cache (500W)	4
Memoria RAM	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 960 GB 1DWPD TLC D, 80 mm	2
	Alimentatore ridondante a singola uscita WS-1K63A-1R(x2)1U da 692W/1600W. Dissipazione del calore di 2361 BTU/ora con temperatura massima di 59 °C (circa)	4

Per il server di controllo:

Codice parte	Descrizione del prodotto	Quantità
511R-M-OTO-17	OTTIMIZZATO FINO A 1U X13SCH-SYS, CSE-813MF2TS-R0RCNBP, PWS-602A-1R	1
	RPL-E 6369P IP 8C/16T 3.3G 24MB 95W 1700 BO	1
Memoria RAM	MEM-DR516MB-EU48(x2)UDIMM ECC DDR5-4800 1Rx8 (16Gb) da 16 GB	1
	HDS-M2N4-960G0-E1-TXD-NON-080(x2) SSD M.2 NVMe PCIe4 960 GB 1DWPD TLC D, 80 mm	2

Per lo switch di rete:

Codice parte	Descrizione del prodotto	Quantità
DCS-7280CR3A	Arista 7280R3A 28x100 GbE	1

Archiviazione NetApp AFF :

Codice parte	Descrizione del prodotto	Quantità
AFF-A20A-100-C	Sistema AFF A20 HA, -C	1
X800-42U-R6-C	Jumper Crd, In-Cab, C13-C14, -C	2
X97602A-C	Alimentatore, 1600W, Titanio, -C	2
X66211B-2-N-C	Cavo, 100 GbE, QSFP28-QSFP28, Cu, 2 m, -C	4
X66240A-05-N-C	Cavo, 25 GbE, SFP28-SFP28, Cu, 0,5 m, -C	2
X5532A-N-C	Binario, 4 montanti, sottile, foro rotondo/quadrato, piccolo, regolabile, 24-32, -C	1
X4024A-2-A-C	Unità Pack 2X1,92 TB, NVMe4, SED, -C	6
X60130A-C	Modulo IO, 2PT, 100 GbE, -C	2
X60132A-C	Modulo IO, 4PT, 10/25 GbE, -C	2
SW-ONTAPB-FLASH-A20-C	SW, pacchetto base ONTAP , per TB, Flash, A20, -C	23

Lista di controllo per la preparazione dell’infrastruttura

Fare riferimento a NetApp AIPod Mini - Preparazione dell’infrastruttura per i dettagli.

Dove trovare ulteriori informazioni

Per saperne di più sulle informazioni descritte nel presente documento, consultare i seguenti documenti e/o siti web:

"Documentazione del prodotto NetApp"

"Progetto OPEA"

"Intel® AI ERAG Documentazione"

"Manuale di distribuzione OPEA Enterprise RAG" == Cronologia delle versioni

Versione	Data	Cronologia delle versioni del documento
Versione 1.0	Settembre 2025	Versione iniziale
Versione 2.0	Feb 2026	Aggiornato con OPEA-Intel® AI for Enterprise RAG 2.0

Informazioni sul copyright

Informazioni sul marchio commerciale