NetApp AIPod Mini - Inferenza RAG aziendale con NetApp e Intel

NetApp artificial intelligence solutions

NetApp
August 20, 2025

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

Sommario

NetApp AIPod Mini - Inferenza RAG aziendale con NetApp e Intel . . . 1
Sintesi . . . 1
Convalida del partner di storage Intel . . . 1
Vantaggi dell'esecuzione di sistemi RAG con NetApp . . . 1
Pubblico di destinazione . . . 2
Requisiti tecnologici . . . 2
Hardware . . . 2
Software . . . 4
Distribuzione della soluzione . . . 5
Stack software . . . 5
Fasi di distribuzione . . . 6
Abilita l'uso di ONTAP S3 . . . 11
Configurare le impostazioni di sincronizzazione pianificata . . . 11
Cambia le modalità di accesso al volume . . . 12
(Facoltativo) Disattiva la verifica del certificato SSL . . . 12
Guida alle taglie . . . 14
Conclusione . . . 15
Riconoscimento . . . 15
distinta base . . . 15
Dove trovare ulteriori informazioni . . . 17

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, NetApp

Sintesi

Un numero crescente di organizzazioni sfrutta le applicazioni di generazione aumentata del recupero (RAG) e i modelli linguistici di grandi dimensioni (LLM) per interpretare i prompt degli utenti e generare risposte per aumentare la produttività e il valore aziendale. Questi prompt e risposte possono includere testo, codice, immagini o persino strutture proteiche terapeutiche recuperate dalla knowledge base interna di un'organizzazione, dai data lake, dai repository di codice e dai repository di documenti. In questo documento viene illustrato il progetto di riferimento della soluzione NetApp AIPod Mini, comprendente storage NetApp AFF e server con processori Intel Xeon 6. Include il software di gestione dati NetApp ONTAP combinato con Intel Advanced Matrix Extensions (Intel AMX) e il software Intel AI for Enterprise Retrieval-augmented Generation (RAG) basato su Open Platform for Enterprise AI (OPEA). NetApp AIPod Mini per RAG aziendale consente alle organizzazioni di trasformare un LLM pubblico in una soluzione di inferenza di intelligenza artificiale generativa (GenAI) privata. La soluzione dimostra un'inferenza RAG efficiente e conveniente su scala aziendale, progettata per migliorare l'affidabilità e fornirti un controllo migliore sulle tue 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 comportano il recupero di conoscenze dai repository di documenti aziendali in vari formati, come PDF, testo, CSV, Excel o grafici di conoscenza. Questi dati vengono solitamente archiviati in soluzioni quali un archivio di oggetti S3 o NFS in locale come origine dei dati. NetApp è leader nelle tecnologie di gestione, mobilità, governance e sicurezza dei dati nell'ecosistema edge, data center e cloud. La gestione dei dati NetApp ONTAP fornisce storage di livello aziendale per supportare vari tipi di carichi di lavoro di

intelligenza artificiale, tra cui inferenza batch e in tempo reale, 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 di intelligenza artificiale 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.

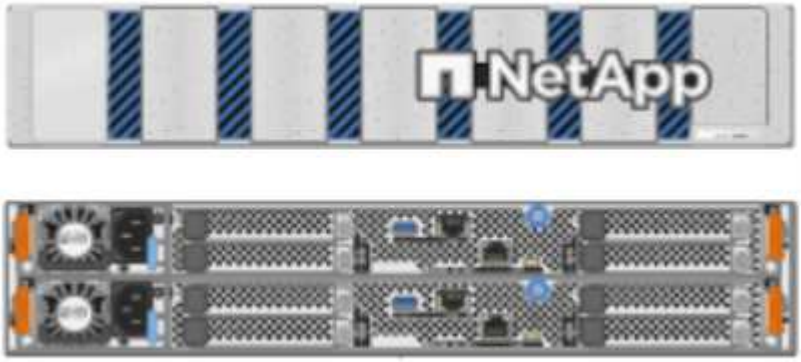


Figura 2 - Sistema NetApp AFF A-Series.

Hardware	Quantità	Commento
Server basato su Intel Xeon 6	2	Nodi di inferenza RAG: con processori Intel Xeon serie 6900 o Intel Xeon serie 6700 a doppio socket e RAM da 250 GB a 3 TB con DDR5 (6400 MHz) o MRDIMM (8800 MHz). 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).

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"

Inferenza Intel AI for Enterprise basata su OPEA

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

Basandosi sulle fondamenta di OPEA, Intel AI for Enterprise RAG amplia questa base con funzionalità chiave che migliorano la scalabilità, la sicurezza e l'esperienza utente. Queste funzionalità includono funzionalità di 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 come servizio, che consente una facile gestione e monitoraggio dei flussi di lavoro. Inoltre, il supporto di Intel e dei partner fornisce l'accesso a un ampio ecosistema di soluzioni, combinato con la gestione integrata dell'identità e degli accessi (IAM) con interfaccia utente e applicazioni per operazioni sicure e conformi. I guardrail programmabili forniscono un controllo dettagliato 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 per Intel AI per aziende RAG	1.1.2	Piattaforma RAG aziendale basata sui microservizi OPEA
Interfaccia di archiviazione del contenitore (driver CSI)	NetApp Trident 25.02	Abilita il provisioning dinamico, le copie Snapshot NetApp e i volumi.
Ubuntu	22.04.5	Sistema operativo su cluster a due nodi
Orchestrazione dei contenitori	Kubernetes 1.31.4	Ambiente per eseguire il framework RAG
ONTAP	ONTAP 9.16.1P4	Sistema operativo di archiviazione su AFF A20. È dotato di Vscan e ARP.

Distribuzione della soluzione

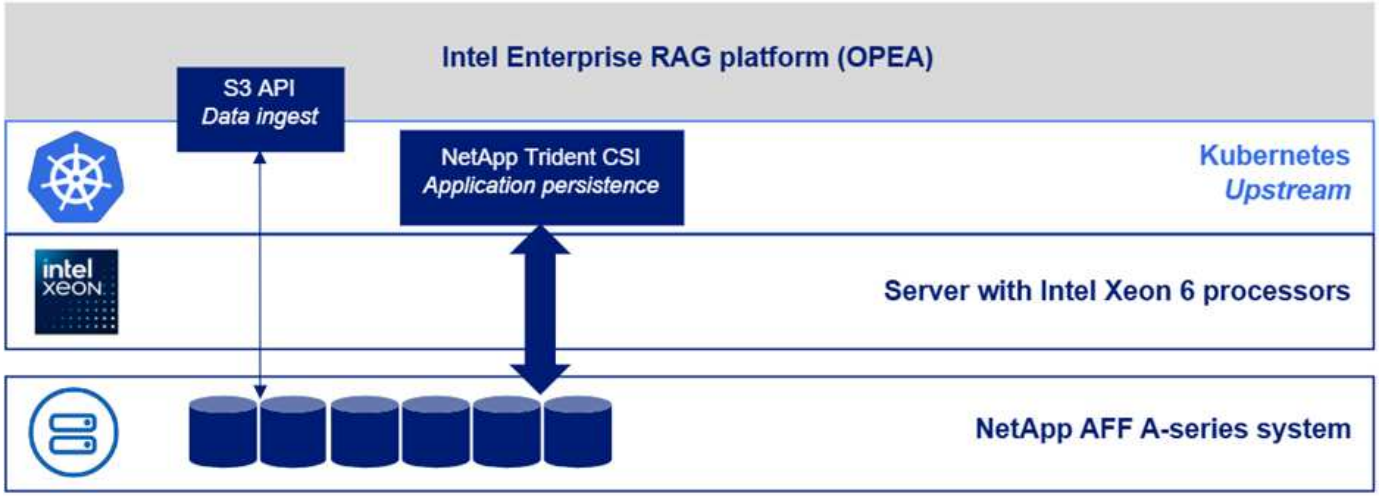
Stack software

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.

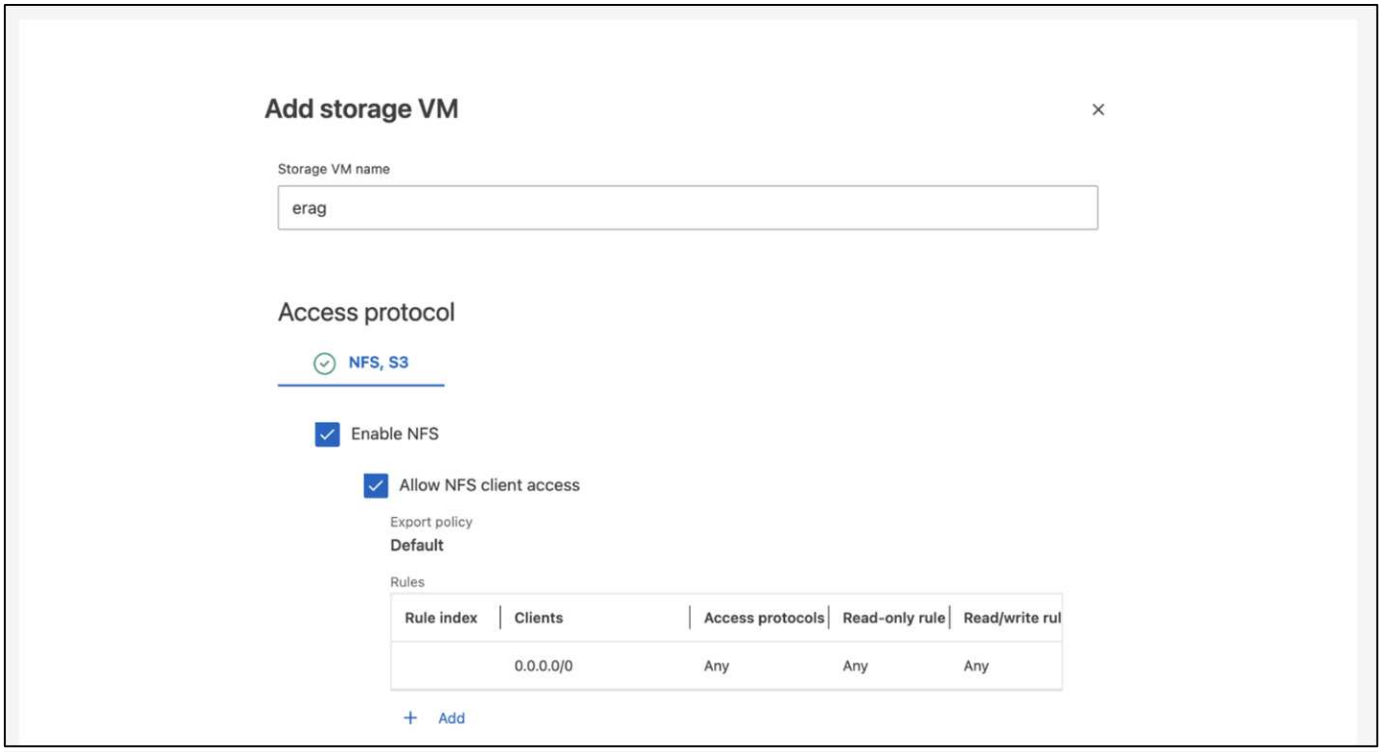
Configurare un ONTAP SVM per l'accesso NFS e S3

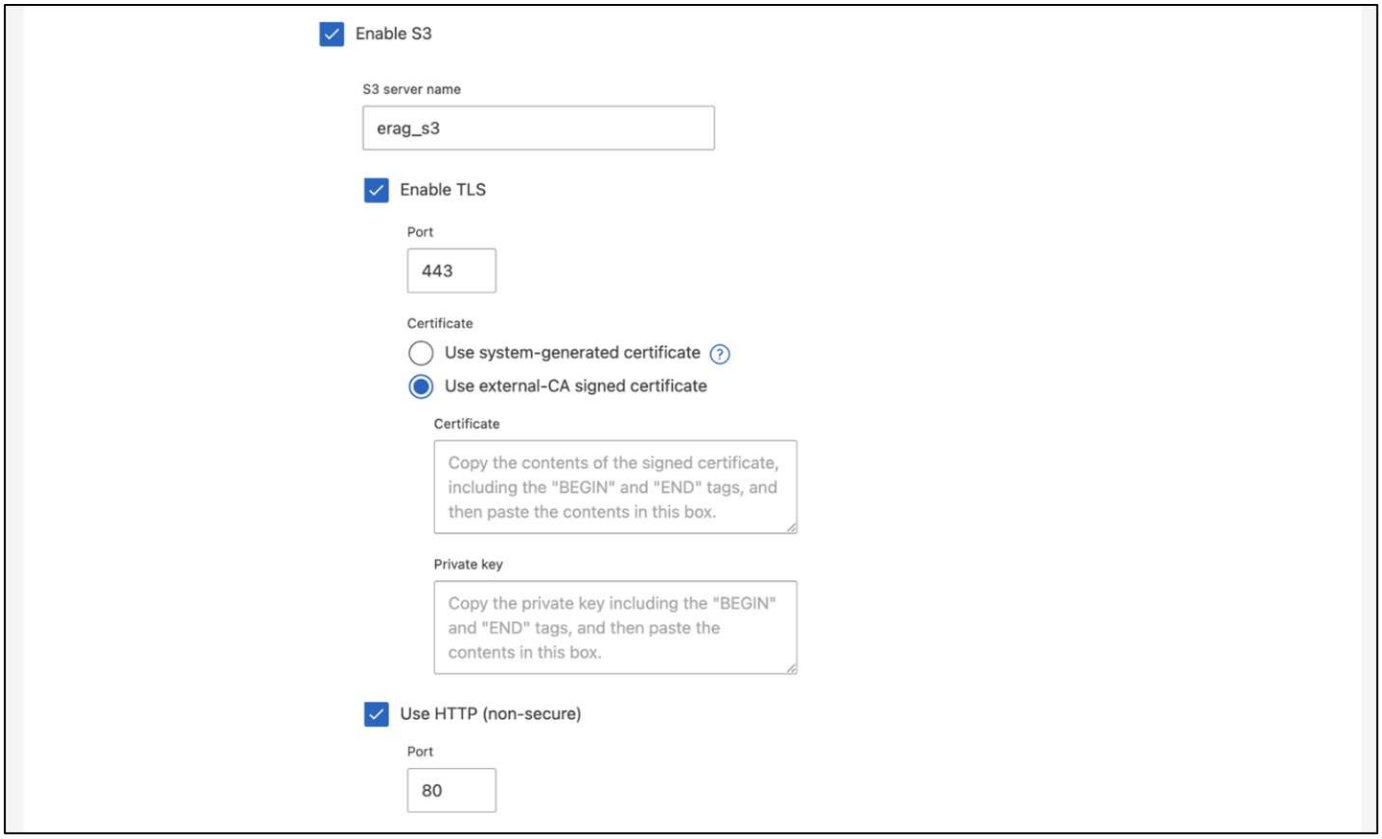
Configurare una macchina virtuale di archiviazione ONTAP (SVM) per l'accesso NFS e S3 su una rete accessibile dai nodi Kubernetes.

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 4 - Creazione di SVM tramite ONTAP System Manager.





Configurare le autorizzazioni S3

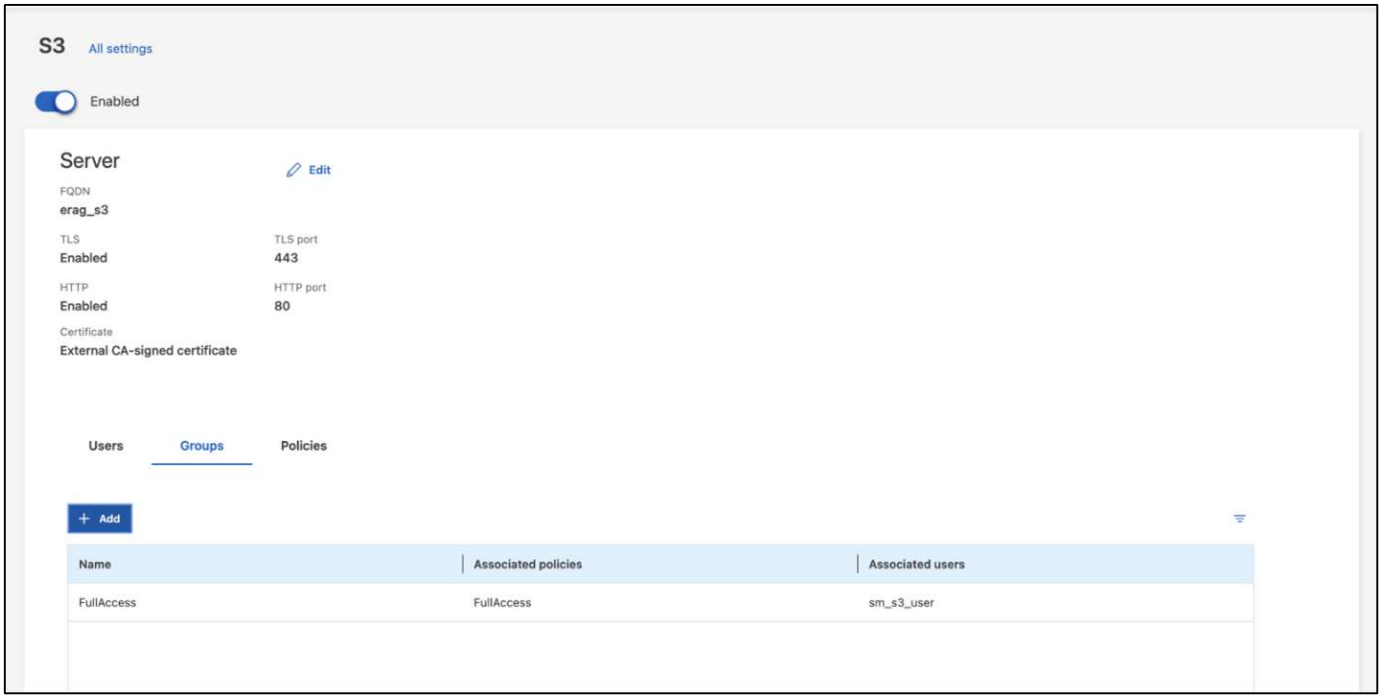
Configurare le impostazioni utente/gruppo S3 per l'SVM creato nel passaggio precedente. Assicurati di avere un utente con accesso completo a tutte le operazioni API S3 per quella SVM. Per maggiori dettagli, consultare la documentazione di ONTAP S3.

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

automaticamente un utente denominato `sm_s3_user` e una politica denominata `FullAccess` quando hai creato il tuo SVM, ma non ti saranno state assegnate autorizzazioni `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 5 - Autorizzazioni S3.

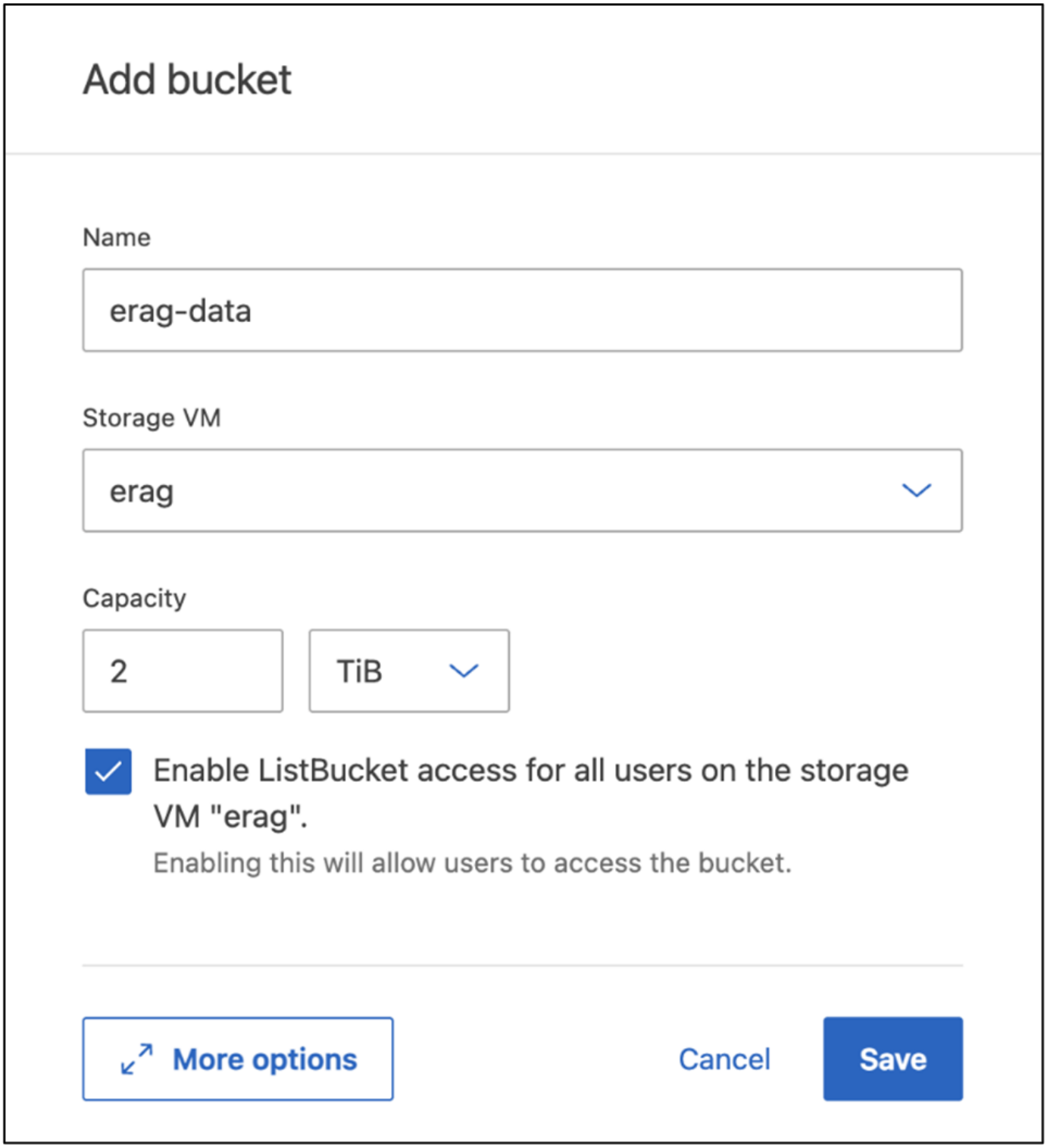


Crea un bucket S3

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 6 - Creazione di 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 7 - 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 all'applicazione web OPEA per Intel AI for Enterprise RAG 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
```

Installa Kubernetes

Installa Kubernetes sui tuoi server utilizzando Kubespray. Fare riferimento al "Documentazione di Kubespray" per i dettagli.

Installa il driver Trident CSI

Installa il driver NetApp Trident CSI nel tuo cluster Kubernetes. Fare riferimento al "Documentazione di installazione Trident" per i dettagli.

Crea un back-end Trident

Crea un back-end Trident per l'SVM creato in precedenza. Quando crei il tuo back-end, usa il `ontap-nas` autista. Fare riferimento al "Documentazione back-end Trident" per i dettagli.

Creare una classe di archiviazione

Crea una classe di archiviazione Kubernetes corrispondente al back-end Trident creato nel passaggio precedente. Per maggiori dettagli, fare riferimento alla documentazione sulla classe di archiviazione Trident .

OPEA per Intel AI per aziende RAG

Installa OPEA per Intel AI per Enterprise RAG nel tuo cluster Kubernetes. Fare riferimento al "Intel AI per la distribuzione RAG aziendale" documentazione per i dettagli. Assicuratevi di prendere nota delle modifiche necessarie al file di configurazione descritte più avanti in questo documento. È necessario apportare queste modifiche prima di eseguire il playbook di installazione affinché l'applicazione Intel AI for Enterprise RAG funzioni correttamente con il sistema di storage ONTAP .

Abilita l'uso di ONTAP S3

Quando si installa OPEA per Intel AI per Enterprise RAG, modificare il file di configurazione principale per abilitare l'uso di ONTAP S3 come repository dei dati di origine.

Per abilitare l'uso di ONTAP S3, impostare i seguenti valori all'interno di `edp` sezione.

Nota: per impostazione predefinita, l'applicazione Intel AI for Enterprise RAG acquisisce dati da tutti i bucket esistenti nella SVM. Se hai più bucket nel tuo SVM, puoi modificare `bucketNameRegexFilter` campo in modo che i dati vengano acquisiti solo da determinati bucket.

```
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: ".*"
```

Configurare le impostazioni di sincronizzazione pianificata

Quando si installa l'applicazione OPEA per Intel AI per Enterprise RAG, abilitare `scheduledSync` in modo che l'applicazione acquisisca automaticamente file nuovi o aggiornati dai bucket S3.

Quando `scheduledSync` è abilitato, l'applicazione controlla automaticamente i bucket S3 di origine per file nuovi o aggiornati. Tutti i file nuovi o aggiornati rilevati durante questo processo di sincronizzazione vengono

automaticamente acquisiti e aggiunti alla knowledge base RAG. L'applicazione controlla i bucket di origine in base a un intervallo di tempo preimpostato. L'intervallo di tempo predefinito è di 60 secondi, il che significa che l'applicazione verifica le modifiche ogni 60 secondi. Potresti voler modificare questo intervallo in base alle tue esigenze specifiche.

Per abilitare `scheduledSync` e impostare l'intervallo di sincronizzazione, impostare i seguenti valori in `deployment/components/edp/values.yaml:`

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

Cambia le modalità di accesso al volume

In `deployment/components/gmc/microservices-connector/helm/values.yaml` , per ogni volume nel `pvc` elenco, cambia il `accessMode` A `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
```

(Facoltativo) Disattiva la verifica del certificato SSL

Se hai utilizzato un certificato autofirmato quando hai abilitato l'accesso S3 per la tua SVM, devi disabilitare la verifica del certificato SSL. Se hai utilizzato un certificato firmato da una CA pubblicamente attendibile, puoi saltare questo passaggio.

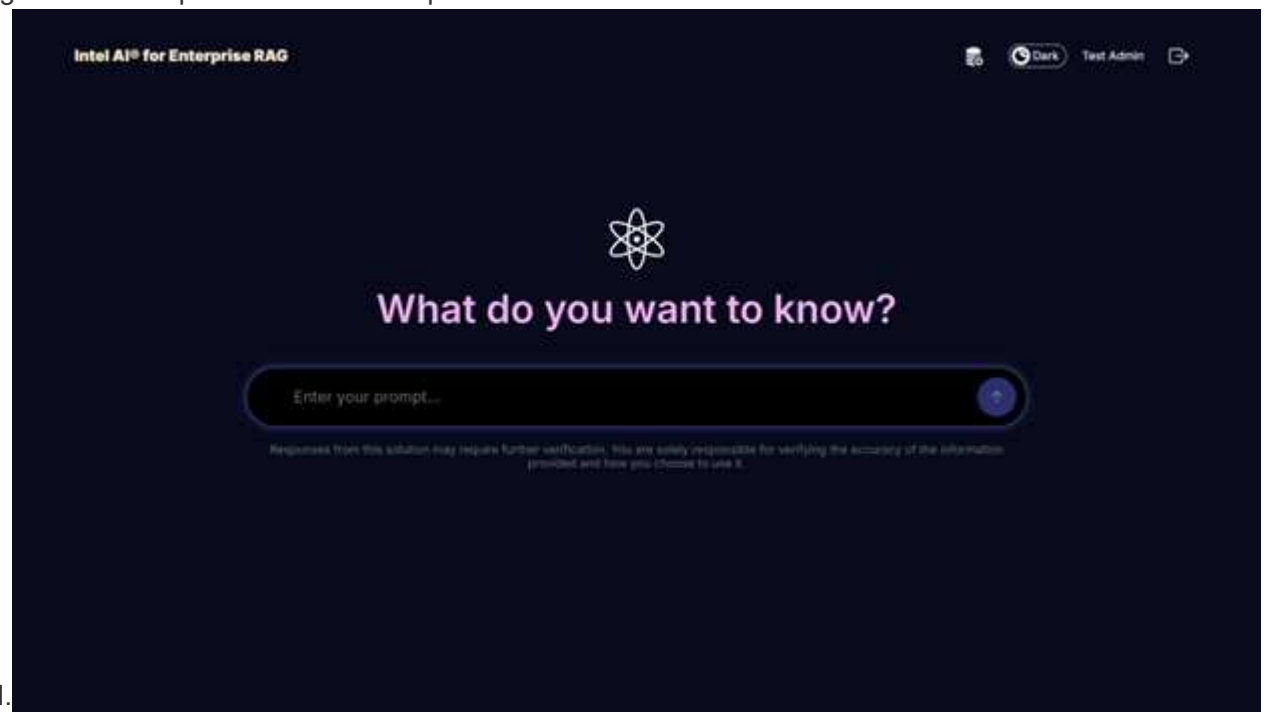
Per disabilitare la verifica del certificato SSL, impostare i seguenti valori in `deployment/components/edp/values.yaml:`

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

Accedi a OPEA per Intel AI for Enterprise RAG UI

Accedi all'interfaccia utente OPEA per Intel AI for Enterprise RAG. Fare riferimento al "Documentazione sulla distribuzione di Intel AI for Enterprise RAG" per i dettagli.

Figura 8: 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 l'acquisizione di un file, l'applicazione OPEA per Intel AI for Enterprise RAG verifica automaticamente la presenza di aggiornamenti al file e acquisisce gli aggiornamenti di conseguenza.

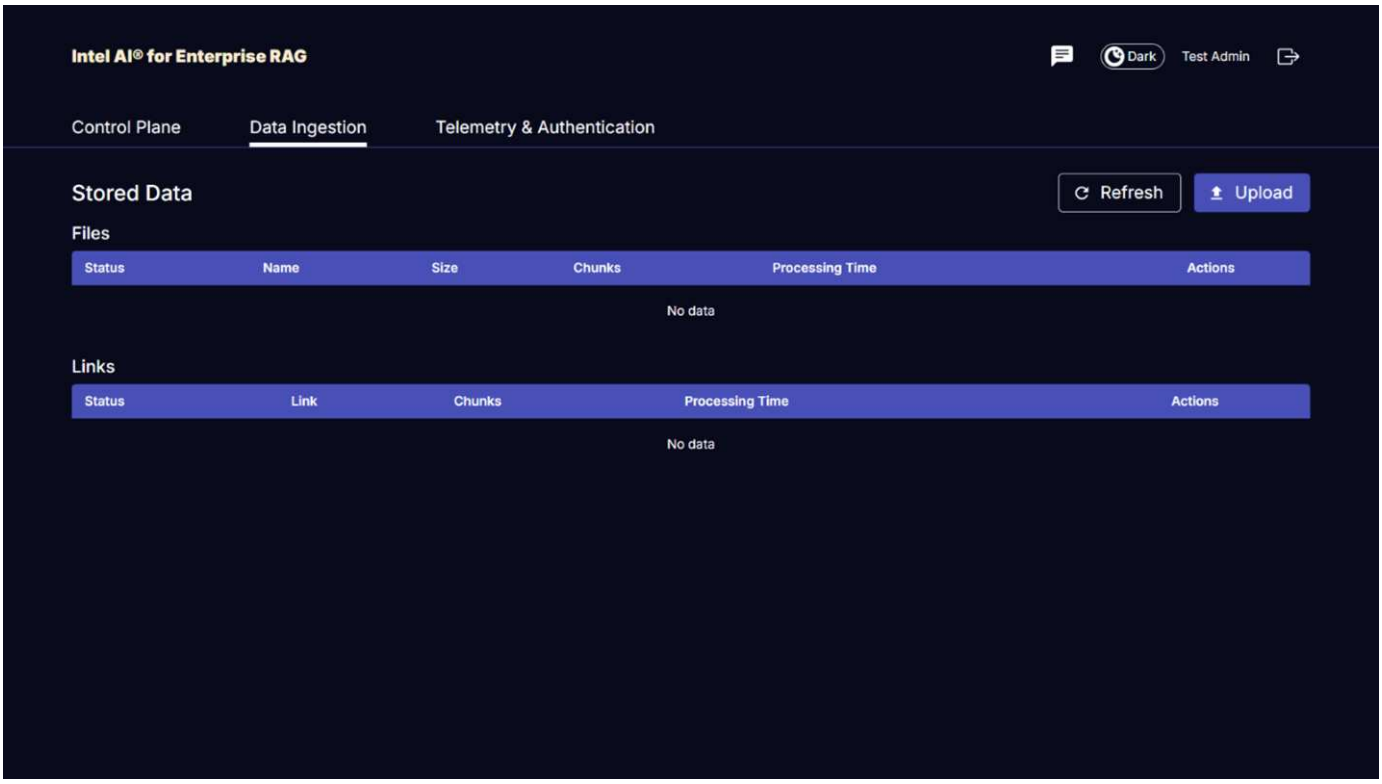
*Opzione 1: Carica direttamente sul tuo bucket S3 Per acquisire più file contemporaneamente, ti consigliamo di caricarli sul tuo bucket S3 (quello creato in precedenza) utilizzando il client S3 che preferisci. Tra i client S3 più diffusi figurano AWS CLI, Amazon SDK per Python (Boto3), s3cmd, S3 Browser, Cyberduck e Commander One. Se i file sono di un tipo supportato, tutti i file caricati nel bucket S3 verranno automaticamente acquisiti dall'applicazione OPEA per 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, PPT, PPTX, MD, XML, JSON, JSONL, YAML, XLS, XLSX, CSV, TIFF, JPG, JPEG, PNG e SVG.

È possibile utilizzare l’interfaccia utente OPEA per Intel AI for Enterprise RAG per confermare che i file siano stati acquisiti correttamente. Per maggiori dettagli, fare riferimento alla documentazione dell’interfaccia utente Intel AI for Enterprise RAG. Tieni presente che l’applicazione potrebbe impiegare del tempo per acquisire un numero elevato di file.

*Opzione 2: caricamento tramite l’interfaccia utente Se devi acquisire solo un numero limitato di file, puoi farlo utilizzando l’interfaccia utente OPEA per Intel AI for Enterprise RAG. Per maggiori dettagli, fare riferimento alla documentazione dell’interfaccia utente Intel AI for Enterprise RAG.

Figura 9 - Interfaccia utente per l’acquisizione dei dati.



Eseguire query di chat

Ora puoi "chattare" con l’applicazione OPEA per Intel AI for Enterprise RAG utilizzando l’interfaccia utente della chat inclusa. Quando risponde alle tue domande, l’applicazione esegue un RAG utilizzando i file ingeriti. Ciò significa che l’applicazione cerca automaticamente le informazioni rilevanti nei file ingeriti e le incorpora quando risponde alle tue domande.

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

Caratterizzazioni	Valore	Commento
Utenti simultanei	32	Per "utenti simultanei" si intendono le richieste rapide che inviano query contemporaneamente.

_Nota: le indicazioni sulle dimensioni 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 requisiti simili in termini di token I/O e dimensioni del modello, consigliamo di utilizzare server con processori Xeon 6 con 96 o 128 core.

Conclusione

I sistemi RAG aziendali e gli LLM sono tecnologie che lavorano insieme per aiutare le organizzazioni a fornire risposte accurate e contestualizzate. 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 archiviazione ad alte prestazioni per interrogare i repository di documenti contenenti dati aziendali, i dati vengono elaborati in modo più rapido e sicuro. NetApp AIPod Mini combina l'infrastruttura dati intelligente di NetApp con le funzionalità di gestione dati ONTAP e i processori Intel Xeon 6, Intel AI for Enterprise RAG e lo stack software OPEA per aiutare a implementare applicazioni RAG ad alte prestazioni e indirizzare le organizzazioni verso la leadership nell'intelligenza artificiale.

Riconoscimento

Questo documento è opera di Sathish Thyagarajan e Michael Ogelsby, membri del team 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, Shreejan Mistry e Ned Fiori) e gli altri membri del team di NetApp(Lawrence Bunka, Bobby Oommen e Jeff Liborio) per il loro continuo supporto e aiuto durante la convalida di questa 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-GNR6980P-SRPL2-UCC	Intel Xeon 6980P 2P 128C 2G 504M 500W SGX512	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

Codice parte	Descrizione del prodotto	Quantità
	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

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"

"Manuale di distribuzione OPEA Enterprise RAG"

Informazioni sul copyright

Copyright © 2025 NetApp, Inc. Tutti i diritti riservati. Stampato negli Stati Uniti d'America. Nessuna porzione di questo documento soggetta a copyright può essere riprodotta in qualsiasi formato o mezzo (grafico, elettronico o meccanico, inclusi fotocopie, registrazione, nastri o storage in un sistema elettronico) senza previo consenso scritto da parte del detentore del copyright.

Il software derivato dal materiale sottoposto a copyright di NetApp è soggetto alla seguente licenza e dichiarazione di non responsabilità:

IL PRESENTE SOFTWARE VIENE FORNITO DA NETAPP "COSÌ COM'È" E SENZA QUALSIVOGLIA TIPO DI GARANZIA IMPLICITA O ESPRESSA FRA CUI, A TITOLO ESEMPLIFICATIVO E NON ESAUSTIVO, GARANZIE IMPLICITE DI COMMERCIABILITÀ E IDONEITÀ PER UNO SCOPO SPECIFICO, CHE VENGONO DECLINATE DAL PRESENTE DOCUMENTO. NETAPP NON VERRÀ CONSIDERATA RESPONSABILE IN ALCUN CASO PER QUALSIVOGLIA DANNO DIRETTO, INDIRETTO, ACCIDENTALE, SPECIALE, ESEMPLARE E CONSEQUENZIALE (COMPRESI, A TITOLO ESEMPLIFICATIVO E NON ESAUSTIVO, PROCUREMENT O SOSTITUZIONE DI MERCI O SERVIZI, IMPOSSIBILITÀ DI UTILIZZO O PERDITA DI DATI O PROFITTI OPPURE INTERRUZIONE DELL'ATTIVITÀ AZIENDALE) CAUSATO IN QUALSIVOGLIA MODO O IN RELAZIONE A QUALUNQUE TEORIA DI RESPONSABILITÀ, SIA ESSA CONTRATTUALE, RIGOROSA O DOVUTA A INSOLVENZA (COMPRESA LA NEGLIGENZA O ALTRO) INSORTA IN QUALSIASI MODO ATTRAVERSO L'UTILIZZO DEL PRESENTE SOFTWARE ANCHE IN PRESENZA DI UN PREAVVISO CIRCA L'EVENTUALITÀ DI QUESTO TIPO DI DANNI.

NetApp si riserva il diritto di modificare in qualsiasi momento qualunque prodotto descritto nel presente documento senza fornire alcun preavviso. NetApp non si assume alcuna responsabilità circa l'utilizzo dei prodotti o materiali descritti nel presente documento, con l'eccezione di quanto concordato espressamente e per iscritto da NetApp. L'utilizzo o l'acquisto del presente prodotto non comporta il rilascio di una licenza nell'ambito di un qualche diritto di brevetto, marchio commerciale o altro diritto di proprietà intellettuale di NetApp.

Il prodotto descritto in questa guida può essere protetto da uno o più brevetti degli Stati Uniti, esteri o in attesa di approvazione.

LEGENDA PER I DIRITTI SOTTOPOSTI A LIMITAZIONE: l'utilizzo, la duplicazione o la divulgazione da parte degli enti governativi sono soggetti alle limitazioni indicate nel sottoparagrafo (b)(3) della clausola Rights in Technical Data and Computer Software del DFARS 252.227-7013 (FEB 2014) e FAR 52.227-19 (DIC 2007).

I dati contenuti nel presente documento riguardano un articolo commerciale (secondo la definizione data in FAR 2.101) e sono di proprietà di NetApp, Inc. Tutti i dati tecnici e il software NetApp forniti secondo i termini del presente Contratto sono articoli aventi natura commerciale, sviluppati con finanziamenti esclusivamente privati. Il governo statunitense ha una licenza irrevocabile limitata, non esclusiva, non trasferibile, non cedibile, mondiale, per l'utilizzo dei Dati esclusivamente in connessione con e a supporto di un contratto governativo statunitense in base al quale i Dati sono distribuiti. Con la sola esclusione di quanto indicato nel presente documento, i Dati non possono essere utilizzati, divulgati, riprodotti, modificati, visualizzati o mostrati senza la previa approvazione scritta di NetApp, Inc. I diritti di licenza del governo degli Stati Uniti per il Dipartimento della Difesa sono limitati ai diritti identificati nella clausola DFARS 252.227-7015(b) (FEB 2014).

Informazioni sul marchio commerciale

NETAPP, il logo NETAPP e i marchi elencati alla pagina http://www.netapp.com/TM sono marchi di NetApp, Inc. Gli altri nomi di aziende e prodotti potrebbero essere marchi dei rispettivi proprietari.