NetApp AIPod Mini – Enterprise RAG Inferencing mit NetApp und Intel

NetApp artificial intelligence solutions

NetApp
August 20, 2025

Inhalt

NetApp AIPod Mini – Enterprise RAG Inferencing mit NetApp und Intel

Dieses Dokument stellt ein validiertes Referenzdesign von NetApp AIPod für Enterprise RAG mit Technologien und kombinierten Funktionen von Intel Xeon 6-Prozessoren und NetApp -Datenverwaltungslösungen vor. Die Lösung demonstriert eine nachgelagerte ChatQnA-Anwendung, die ein großes Sprachmodell nutzt und gleichzeitigen Benutzern genaue, kontextrelevante Antworten liefert. Die Antworten werden über eine Air-Gap-RAG-Inferenzpipeline aus dem internen Wissensspeicher einer Organisation abgerufen.

Sathish Thyagarajan, Michael Oglesby, NetApp

Zusammenfassung

Immer mehr Unternehmen nutzen Retrieval-Augmented Generation (RAG)-Anwendungen und große Sprachmodelle (LLMs), um Benutzereingaben zu interpretieren und Antworten zu generieren und so die Produktivität und den Geschäftswert zu steigern. Diese Eingabeaufforderungen und Antworten können Text, Code, Bilder oder sogar therapeutische Proteinstrukturen umfassen, die aus der internen Wissensdatenbank, Datenseen, Code-Repositories und Dokument-Repositories einer Organisation abgerufen werden. Dieses Dokument behandelt das Referenzdesign der NetApp AIPod Mini-Lösung, die aus NetApp AFF -Speicher und Servern mit Intel Xeon 6-Prozessoren besteht. Es umfasst die Datenverwaltungssoftware NetApp ONTAP in Kombination mit Intel Advanced Matrix Extensions (Intel AMX) und die auf Open Platform for Enterprise AI (OPEA) basierende Software Intel AI for Enterprise Retrieval-augmented Generation (RAG). Mit dem NetApp AIPod Mini für Enterprise RAG können Unternehmen ein öffentliches LLM zu einer privaten generativen KI-Inferenzlösung (GenAI) erweitern. Die Lösung demonstriert effizientes und kostengünstiges RAG-Inferencing im Unternehmensmaßstab, das die Zuverlässigkeit verbessern und Ihnen eine bessere Kontrolle über Ihre geschützten Informationen bieten soll.

Validierung durch Intel-Speicherpartner

Server mit Intel Xeon 6-Prozessoren sind für die Verarbeitung anspruchsvoller KI-Inferenz-Workloads ausgelegt und nutzen Intel AMX für maximale Leistung. Um optimale Speicherleistung und Skalierbarkeit zu ermöglichen, wurde die Lösung erfolgreich mit NetApp ONTAP validiert, sodass Unternehmen die Anforderungen von RAG-Anwendungen erfüllen können. Diese Validierung wurde auf Servern mit Intel Xeon 6-Prozessoren durchgeführt. Intel und NetApp pflegen eine starke Partnerschaft, deren Schwerpunkt auf der Bereitstellung optimierter, skalierbarer und auf die Geschäftsanforderungen der Kunden abgestimmter KI-Lösungen liegt.

Vorteile des Betriebs von RAG-Systemen mit NetApp

Bei RAG-Anwendungen geht es um den Abruf von Wissen aus den Dokumentenspeichern von Unternehmen in verschiedenen Formaten wie PDF, Text, CSV, Excel oder Wissensgraphen. Diese Daten werden normalerweise in Lösungen wie einem S3-Objektspeicher oder NFS vor Ort als Datenquelle gespeichert. NetApp ist ein führender Anbieter von Technologien für Datenmanagement, Datenmobilität, Datenverwaltung und Datensicherheit im gesamten Ökosystem von Edge, Rechenzentrum und Cloud. Das NetApp ONTAP Datenmanagement bietet Speicher der Enterprise-Klasse zur Unterstützung verschiedener Arten von KI-

Workloads, einschließlich Batch- und Echtzeit-Inferenz, und bietet einige der folgenden Vorteile:

- Geschwindigkeit und Skalierbarkeit. Sie können große Datensätze mit hoher Geschwindigkeit für die Versionierung verarbeiten und dabei Leistung und Kapazität unabhängig voneinander skalieren.
- Datenzugriff. Durch die Multiprotokollunterstützung können Clientanwendungen Daten mithilfe der Dateifreigabeprotokolle S3, NFS und SMB lesen. ONTAP S3 NAS-Buckets können den Datenzugriff in multimodalen LLM-Inferenzszenarien erleichtern.
- Zuverlässigkeit und Vertraulichkeit. ONTAP bietet Datenschutz, integrierten NetApp Autonomous Ransomware Protection (ARP) und dynamische Speicherbereitstellung und bietet sowohl software- als auch hardwarebasierte Verschlüsselung zur Verbesserung der Vertraulichkeit und Sicherheit. ONTAP ist für alle SSL-Verbindungen mit FIPS 140-2 kompatibel.

Zielgruppe

Dieses Dokument richtet sich an KI-Entscheidungsträger, Dateningenieure, Unternehmensleiter und Abteilungsleiter, die die Vorteile einer Infrastruktur nutzen möchten, die für die Bereitstellung von RAG- und GenAI-Unternehmenslösungen entwickelt wurde. Vorkenntnisse in KI-Inferenz, LLMs, Kubernetes sowie Netzwerken und deren Komponenten sind in der Implementierungsphase hilfreich.

Technologieanforderungen

Hardware

Intel KI-Technologien

Mit Xeon 6 als Host-CPU profitieren beschleunigte Systeme von hoher Single-Thread-Leistung, höherer Speicherbandbreite, verbesserter Zuverlässigkeit, Verfügbarkeit und Wartungsfreundlichkeit (RAS) und mehr E/A-Lanes. Intel AMX beschleunigt die Inferenz für INT8 und BF16 und bietet Unterstützung für FP16-trainierte Modelle mit bis zu 2.048 Gleitkommaoperationen pro Zyklus pro Kern für INT8 und 1.024 Gleitkommaoperationen pro Zyklus pro Kern für BF16/FP16. Für die Bereitstellung einer RAG-Lösung mit Xeon 6-Prozessoren werden im Allgemeinen mindestens 250 GB RAM und 500 GB Festplattenspeicher empfohlen. Dies hängt jedoch stark von der Größe des LLM-Modells ab. Weitere Informationen finden Sie im Intel "Xeon 6 Prozessor" Produktbeschreibung.

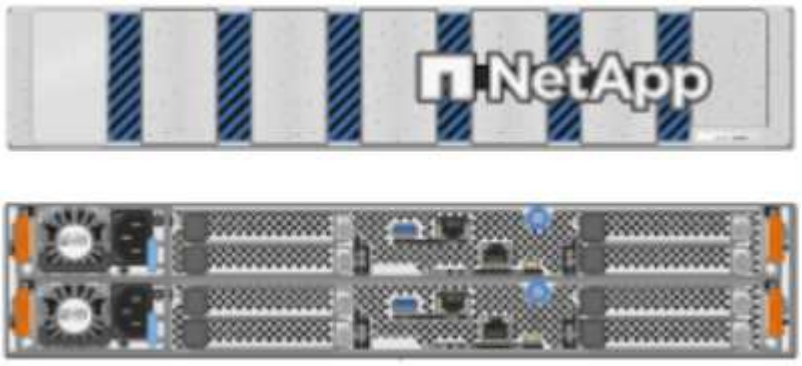
Abbildung 1 – Compute-Server mit Intel Xeon 6-

Prozessoren

NetApp AFF Speicher

Die NetApp AFF A-Series-Systeme der Einstiegs- und Mittelklasse bieten mehr Leistung, Dichte und höhere Effizienz. Die Systeme NetApp AFF A20, AFF A30 und AFF A50 bieten echten Unified Storage, der Block-, Datei- und Objektspeicher unterstützt und auf einem einzigen Betriebssystem basiert, das Daten für RAG-Anwendungen nahtlos verwalten, schützen und mobilisieren kann – und das zu den niedrigsten Kosten in der gesamten Hybrid Cloud.

Abbildung 2 – NetApp AFF A-Series-



System.

Hardware	Menge	Kommentar
Intel Xeon 6-basierter Server	2	RAG-Inferenzknoten – mit Dual-Socket-Prozessoren der Intel Xeon 6900-Serie oder Intel Xeon 6700-Serie und 250 GB bis 3 TB RAM mit DDR5 (6400 MHz) oder MRDIMM (8800 MHz). 2HE-Server.

Hardware	Menge	Kommentar
Control Plane Server mit Intel-Prozessor	1	Kubernetes-Steuerebene/1U-Server.
Auswahl eines 100-Gb-Ethernet-Switches	1	Rechenzentrums-Switch.
NetApp AFF A20 (oder AFF A30; AFF A50)	1	Maximale Speicherkapazität: 9,3 PB. Hinweis: Netzwerk: 10/25/100 GbE-Ports.

Zur Validierung dieses Referenzdesigns wurden Server mit Intel Xeon 6 Prozessoren von Supermicro (222HA-TN-OTO-37) und ein 100GbE Switch von Arista (7280R3A) verwendet.

Software

Offene Plattform für Enterprise-KI

Die Open Platform for Enterprise AI (OPEA) ist eine Open-Source-Initiative unter der Leitung von Intel in Zusammenarbeit mit Ökosystempartnern. Es bietet eine modulare Plattform aus zusammensetzbaren Bausteinen, die die Entwicklung hochmoderner generativer KI-Systeme beschleunigen soll, mit einem starken Fokus auf RAG. OPEA umfasst ein umfassendes Framework mit LLMs, Datenspeichern, Prompt-Engines, RAG-Architekturentwürfen und einer vierstufigen Bewertungsmethode, die generative KI-Systeme anhand von Leistung, Funktionen, Vertrauenswürdigkeit und Unternehmensbereitschaft bewertet.

Im Kern besteht OPEA aus zwei Schlüsselkomponenten:

- GenAIComps: ein servicebasiertes Toolkit bestehend aus Microservice-Komponenten
- GenAIExamples: einsatzbereite Lösungen wie ChatQnA, die praktische Anwendungsfälle demonstrieren

Weitere Einzelheiten finden Sie im "OPEA-Projektdokumentation"

Intel AI für Enterprise-Inferenz mit OPEA

OPEA für Intel AI for Enterprise RAG vereinfacht die Umwandlung Ihrer Unternehmensdaten in umsetzbare Erkenntnisse. Es basiert auf Intel Xeon-Prozessoren und integriert Komponenten von Branchenpartnern, um einen optimierten Ansatz für die Bereitstellung von Unternehmenslösungen zu bieten. Es lässt sich nahtlos mit bewährten Orchestrierungsframeworks skalieren und bietet die Flexibilität und Auswahl, die Ihr Unternehmen benötigt.

Aufbauend auf der Grundlage von OPEA erweitert Intel AI for Enterprise RAG diese Basis um wichtige Funktionen, die Skalierbarkeit, Sicherheit und Benutzererfahrung verbessern. Zu diesen Funktionen gehören Service-Mesh-Funktionen für die nahtlose Integration in moderne servicebasierte Architekturen, eine produktionsreife Validierung der Pipeline-Zuverlässigkeit und eine funktionsreiche Benutzeroberfläche für RAG als Service, die eine einfache Verwaltung und Überwachung von Arbeitsabläufen ermöglicht. Darüber hinaus bieten Intel und der Partner-Support Zugriff auf ein breites Ökosystem von Lösungen, kombiniert mit integriertem Identity and Access Management (IAM) mit UI und Anwendungen für sichere und konforme Vorgänge. Programmierbare Leitplanken bieten eine fein abgestufte Kontrolle über das Pipeline-Verhalten und ermöglichen benutzerdefinierte Sicherheits- und Compliance-Einstellungen.

NetApp ONTAP

NetApp ONTAP ist die grundlegende Technologie, die den kritischen Datenspeicherlösungen von NetApp zugrunde liegt. ONTAP umfasst verschiedene Datenverwaltungs- und Datenschutzfunktionen, wie z. B.

automatischen Ransomware-Schutz vor Cyberangriffen, integrierte Datentransportfunktionen und Speichereffizienzfunktionen. Diese Vorteile gelten für eine Reihe von Architekturen, von lokalen bis hin zu hybriden Multiclouds in NAS, SAN, Objekt- und softwaredefiniertem Speicher für LLM-Bereitstellungen. Sie können einen ONTAP S3-Objektspeicherserver in einem ONTAP Cluster zum Bereitstellen von RAG-Anwendungen verwenden und dabei die Speichereffizienz und Sicherheit von ONTAP nutzen, die durch autorisierte Benutzer und Clientanwendungen bereitgestellt wird. Weitere Informationen finden Sie unter "Erfahren Sie mehr über die ONTAP S3-Konfiguration"

NetApp Trident

Die NetApp Trident -Software ist ein Open-Source- und vollständig unterstützter Speicherorchestrator für Container und Kubernetes-Distributionen, einschließlich Red Hat OpenShift. Trident funktioniert mit dem gesamten NetApp -Speicherportfolio, einschließlich NetApp ONTAP , und unterstützt auch NFS- und iSCSI-Verbindungen. Weitere Informationen finden Sie unter "NetApp Trident auf Git"

Software	Version	Kommentar
OPEA für Intel AI für Enterprise RAG	1.1.2	Enterprise-RAG-Plattform basierend auf OPEA-Microservices
Container Storage Interface (CSI-Treiber)	NetApp Trident 25.02	Ermöglicht dynamische Bereitstellung, NetApp Snapshot-Kopien und Volumes.
Ubuntu	22.04.5	Betriebssystem auf einem Cluster mit zwei Knoten
Container-Orchestrierung	Kubernetes 1.31.4	Umgebung zum Ausführen des RAG-Frameworks
ONTAP	ONTAP 9.16.1P4	Speicherbetriebssystem auf AFF A20. Es verfügt über Vscan und ARP.

Lösungsbereitstellung

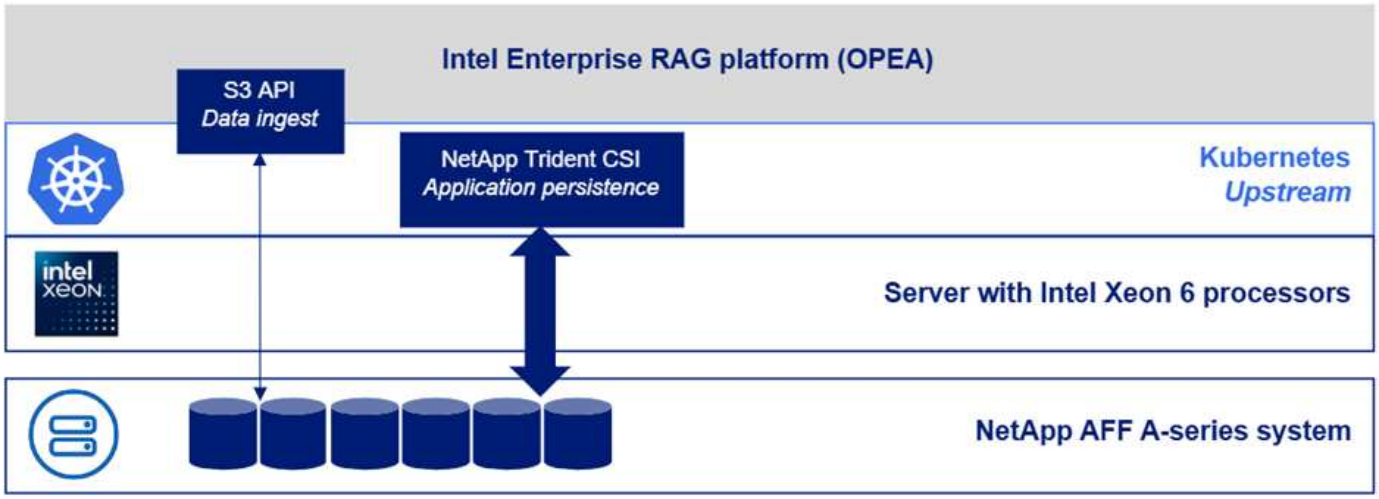
Software-Stack

Die Lösung wird auf einem Kubernetes-Cluster bereitgestellt, der aus Intel Xeon-basierten App-Knoten besteht. Um eine grundlegende Hochverfügbarkeit für die Kubernetes-Steuerebene zu implementieren, sind mindestens drei Knoten erforderlich. Wir haben die Lösung mithilfe des folgenden Cluster-Layouts validiert.

Tabelle 3 – Kubernetes-Cluster-Layout

Node	Rolle	Menge
Server mit Intel Xeon 6 Prozessoren und 1TB RAM	App-Knoten, Steuerebenenknoten	2
Generischer Server	Steuerebenenknoten	1

Die folgende Abbildung zeigt eine „Software-Stack-Ansicht“ der Lösung.



Bereitstellungsschritte

Bereitstellen des ONTAP Speichergeräts

Implementieren und Bereitstellen Ihres NetApp ONTAP Speichergeräts. Weitere Informationen finden Sie im "Dokumentation zu ONTAP -Hardwaresystemen" für Details.

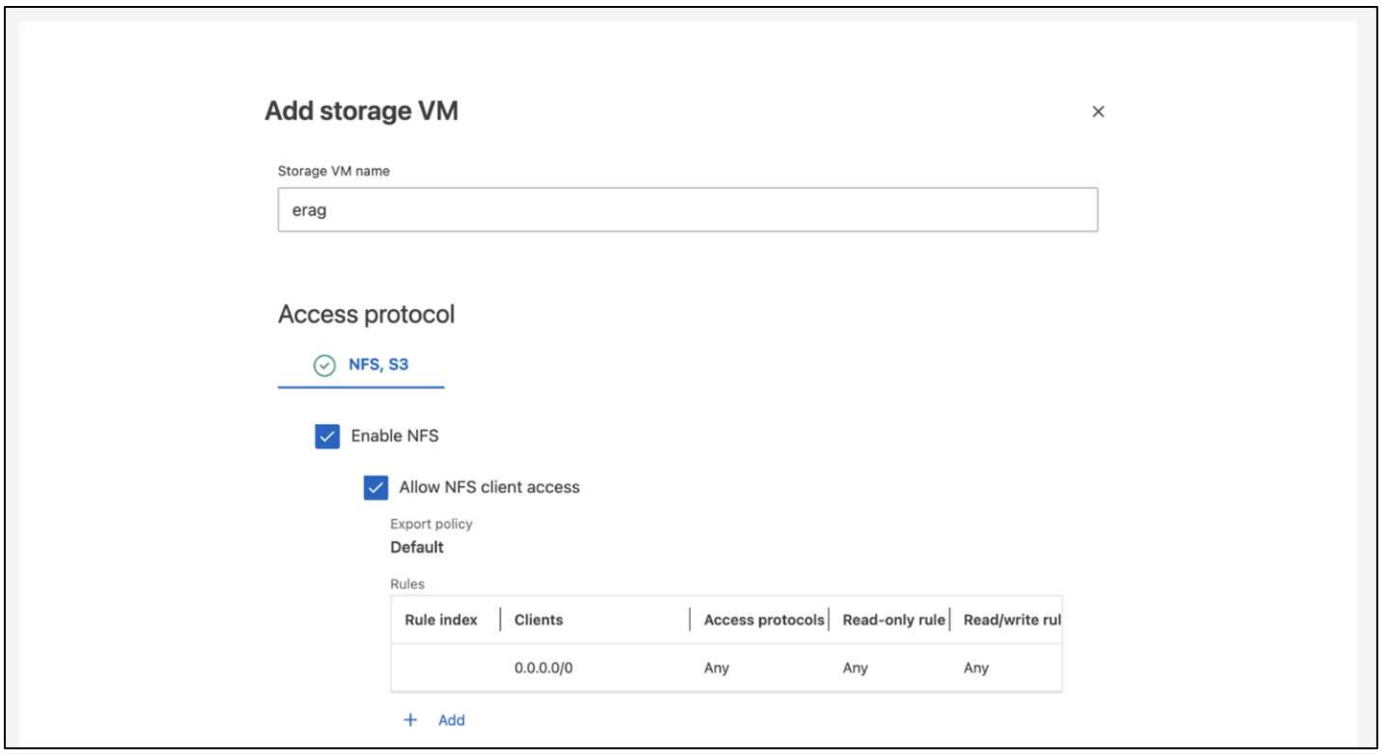
Konfigurieren Sie ein ONTAP SVM für den NFS- und S3-Zugriff

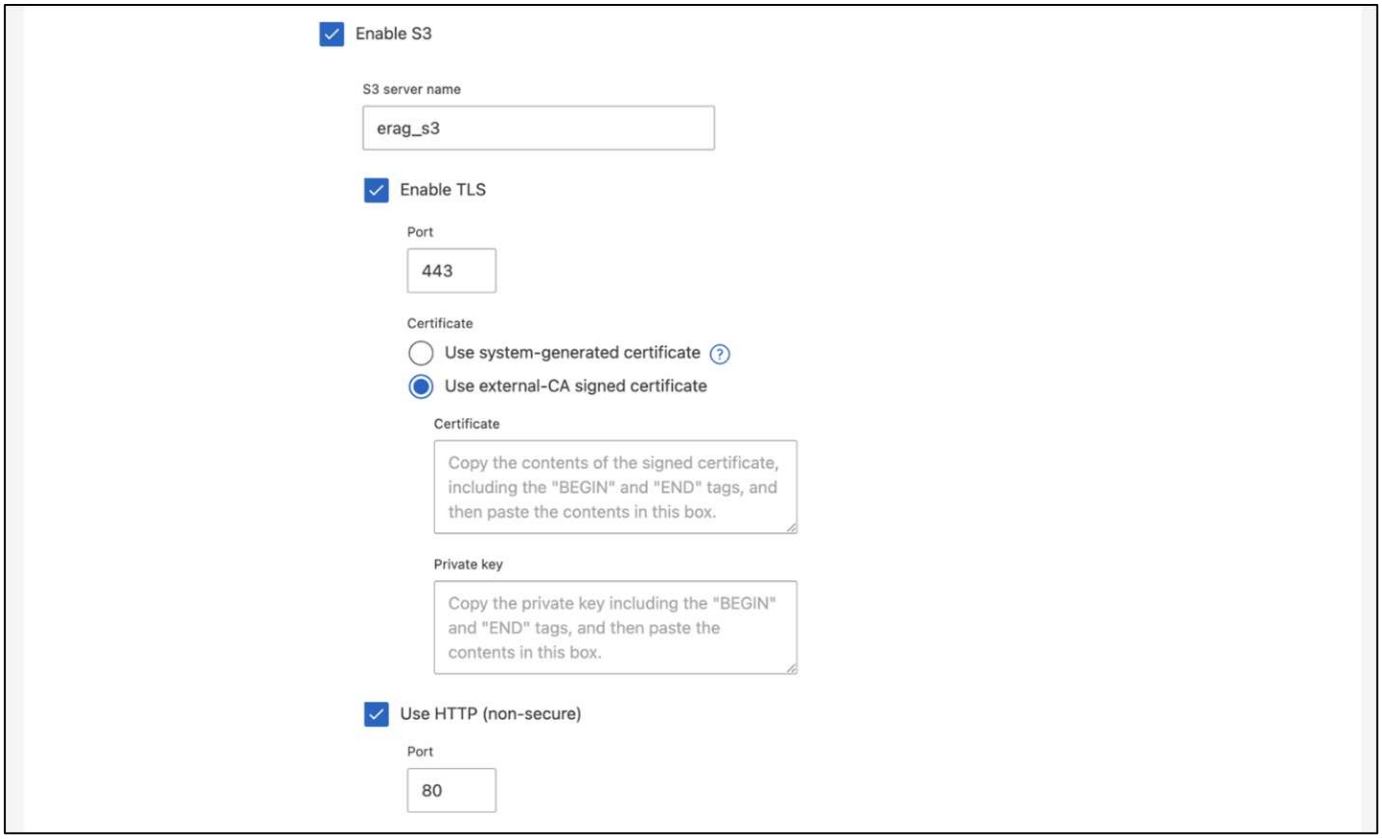
Konfigurieren Sie eine ONTAP Storage Virtual Machine (SVM) für den NFS- und S3-Zugriff in einem Netzwerk, auf das Ihre Kubernetes-Knoten zugreifen können.

Um eine SVM mit ONTAP System Manager zu erstellen, navigieren Sie zu Speicher > Speicher-VMs und klicken Sie auf die Schaltfläche + Hinzufügen. Wenn Sie den S3-Zugriff für Ihre SVM aktivieren, wählen Sie die Option zur Verwendung eines von einer externen Zertifizierungsstelle (CA) signierten Zertifikats und nicht eines systemgenerierten Zertifikats. Sie können entweder ein selbstsigniertes Zertifikat oder ein Zertifikat verwenden, das von einer öffentlich vertrauenswürdigen Zertifizierungsstelle signiert wurde. Weitere Einzelheiten finden Sie im "ONTAP -Dokumentation."

Der folgende Screenshot zeigt die Erstellung einer SVM mit ONTAP System Manager. Ändern Sie die Details je nach Bedarf entsprechend Ihrer Umgebung.

Abbildung 4 – SVM-Erstellung mit ONTAP System Manager.





Konfigurieren von S3-Berechtigungen

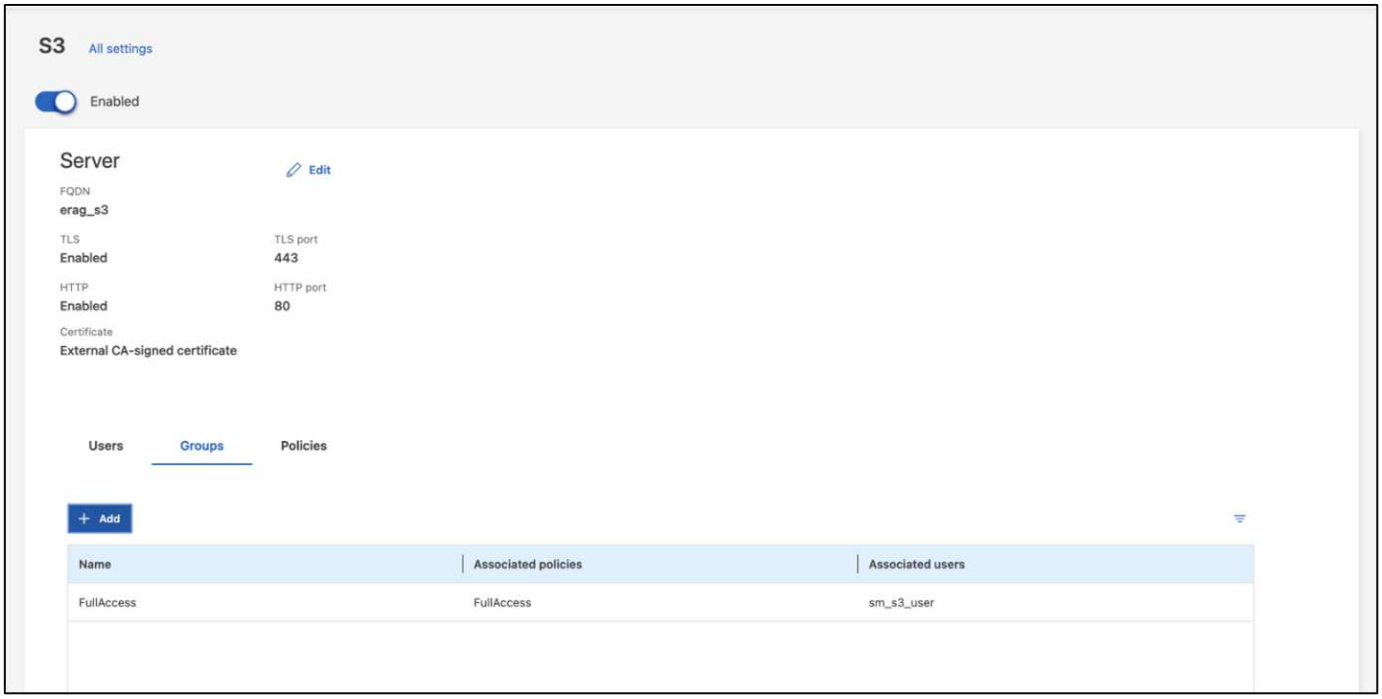
Konfigurieren Sie die S3-Benutzer-/Gruppeneinstellungen für die SVM, die Sie im vorherigen Schritt erstellt haben. Stellen Sie sicher, dass Sie einen Benutzer mit vollem Zugriff auf alle S3-API-Operationen für diese SVM haben. Weitere Informationen finden Sie in der ONTAP S3-Dokumentation.

Hinweis: Dieser Benutzer wird für den Datenaufnahmedienst der Intel AI for Enterprise RAG-Anwendung benötigt. Wenn Sie Ihre SVM mit ONTAP System Manager erstellt haben, hat System Manager automatisch

einen Benutzer mit dem Namen erstellt. `sm_s3_user` und eine Richtlinie namens `FullAccess` beim Erstellen Ihrer SVM, aber es wurden keine Berechtigungen zugewiesen `sm_s3_user` .

Um die Berechtigungen für diesen Benutzer zu bearbeiten, navigieren Sie zu „Speicher“ > „Speicher-VMs“, klicken Sie auf den Namen der SVM, die Sie im vorherigen Schritt erstellt haben, klicken Sie auf „Einstellungen“ und dann auf das Stiftsymbol neben „S3“. Geben `sm_s3_user` Vollzugriff auf alle S3-API-Operationen, erstellen Sie eine neue Gruppe, die verknüpft `sm_s3_user` mit dem `FullAccess` Richtlinie, wie im folgenden Screenshot dargestellt.

Abbildung 5 – S3-Berechtigungen.

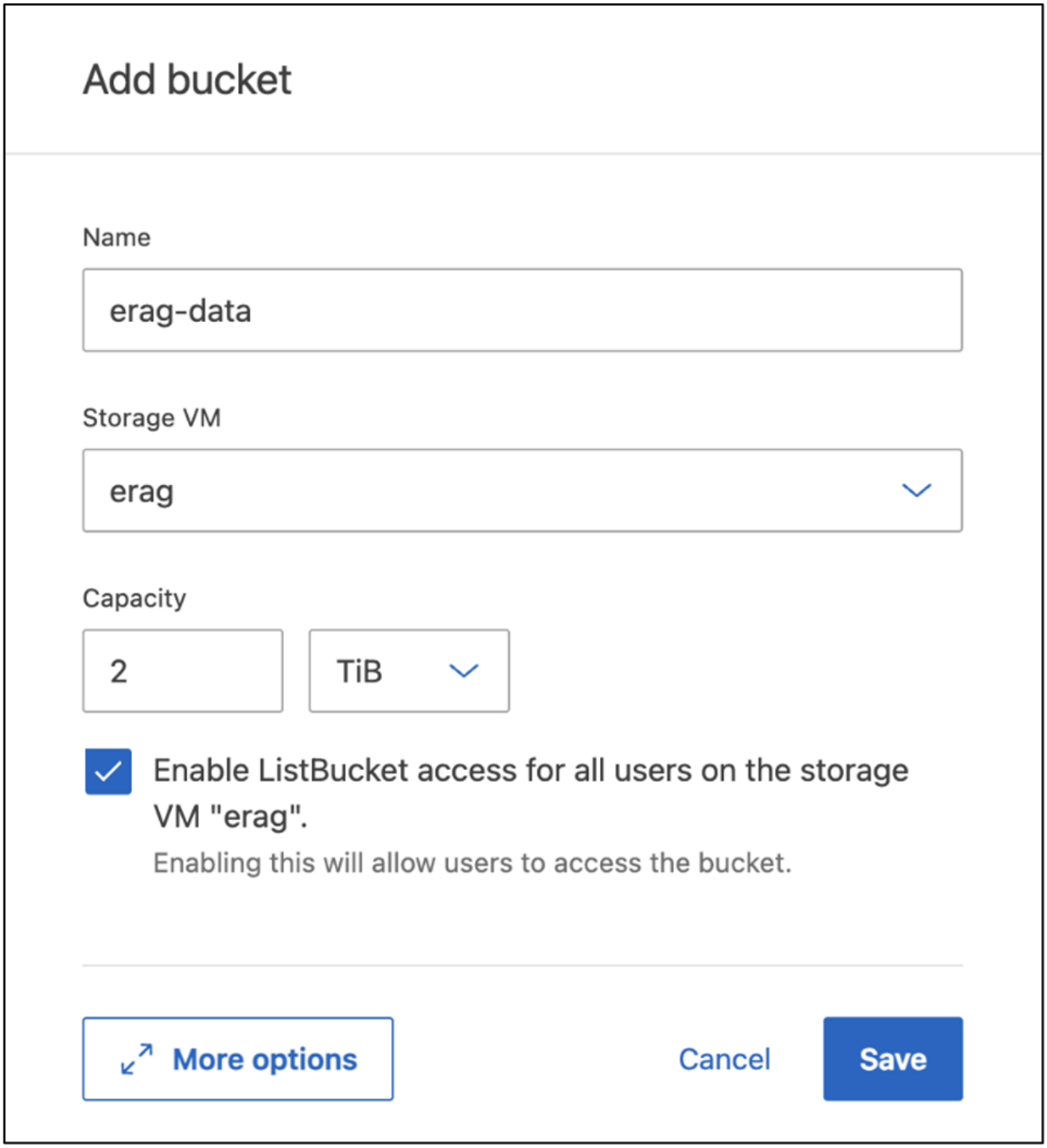


Erstellen eines S3-Buckets

Erstellen Sie einen S3-Bucket innerhalb der SVM, die Sie zuvor erstellt haben. Um eine SVM mit ONTAP System Manager zu erstellen, navigieren Sie zu Speicher > Buckets und klicken Sie auf die Schaltfläche + Hinzufügen. Weitere Einzelheiten finden Sie in der ONTAP S3-Dokumentation.

Der folgende Screenshot zeigt die Erstellung eines S3-Buckets mit ONTAP System Manager.

Abbildung 6 – Erstellen Sie einen S3-
Bucket.



Konfigurieren von S3-Bucket-Berechtigungen

Konfigurieren Sie die Berechtigungen für den S3-Bucket, den Sie im vorherigen Schritt erstellt haben. Stellen Sie sicher, dass der Benutzer, den Sie in einem vorherigen Schritt konfiguriert haben, über die folgenden Berechtigungen verfügt: `GetObject, PutObject, DeleteObject, ListBucket, GetBucketAcl, GetObjectAcl, ListBucketMultipartUploads, ListMultipartUploadParts, GetObjectTagging, PutObjectTagging, DeleteObjectTagging, GetBucketLocation, GetBucketVersioning, PutBucketVersioning, ListBucketVersions, GetBucketPolicy, PutBucketPolicy, DeleteBucketPolicy, PutLifecycleConfiguration,`

`GetLifecycleConfiguration, GetBucketCORS, PutBucketCORS.`

Um S3-Bucket-Berechtigungen mit ONTAP System Manager zu bearbeiten, navigieren Sie zu Speicher > Buckets, klicken Sie auf den Namen Ihres Buckets, klicken Sie auf Berechtigungen und dann auf Bearbeiten. Weitere Informationen finden Sie im "ONTAP S3-Dokumentation" für weitere Einzelheiten.

Der folgende Screenshot zeigt die erforderlichen Bucket-Berechtigungen im ONTAP System Manager.

Abbildung 7 – S3-Bucket-Berechtigungen.

Erstellen einer Bucket-Cross-Origin-Ressourcenfreigaberegel

Erstellen Sie mithilfe der ONTAP CLI eine Bucket-Cross-Origin-Resource-Sharing-Regel (CORS) für den Bucket, den Sie im vorherigen Schritt erstellt haben:

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

Diese Regel ermöglicht es der OPEA für die Intel AI for Enterprise RAG-Webanwendung, über einen Webbrowser mit dem Bucket zu interagieren.

Bereitstellen von Servern

Stellen Sie Ihre Server bereit und installieren Sie Ubuntu 22.04 LTS auf jedem Server. Installieren Sie nach der Installation von Ubuntu die NFS-Dienstprogramme auf jedem Server. Führen Sie zum Installieren der NFS-Dienstprogramme den folgenden Befehl aus:

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

Installieren Sie Kubernetes

Installieren Sie Kubernetes mit Kubespray auf Ihren Servern. Weitere Informationen finden Sie im "Kubespray-Dokumentation" für Details.

Installieren Sie den Trident CSI-Treiber

Installieren Sie den NetApp Trident CSI-Treiber in Ihrem Kubernetes-Cluster. Weitere Informationen finden Sie im "Trident -Installationsdokumentation" für Details.

Erstellen Sie ein Trident -Backend

Erstellen Sie ein Trident Backend für die SVM, die Sie zuvor erstellt haben. Verwenden Sie beim Erstellen Ihres Backends die `ontap-nas` Treiber. Weitere Informationen finden Sie im "Trident -Backend -Dokumentation" für Details.

Erstellen einer Speicherklasse

Erstellen Sie eine Kubernetes-Speicherklasse, die dem Trident Back-End entspricht, das Sie im vorherigen Schritt erstellt haben. Weitere Informationen finden Sie in der Dokumentation zur Trident -Speicherklasse.

OPEA für Intel AI für Enterprise RAG

Installieren Sie OPEA für Intel AI for Enterprise RAG in Ihrem Kubernetes-Cluster. Weitere Informationen finden Sie im "Intel KI für Enterprise RAG-Bereitstellung" Einzelheiten finden Sie in der Dokumentation. Beachten Sie unbedingt die erforderlichen Änderungen an der Konfigurationsdatei, die später in diesem Dokument beschrieben werden. Sie müssen diese Änderungen vornehmen, bevor Sie das Installations-Playbook ausführen, damit die Intel AI for Enterprise RAG-Anwendung ordnungsgemäß mit Ihrem ONTAP Speichersystem funktioniert.

Aktivieren Sie die Verwendung von ONTAP S3

Bearbeiten Sie beim Installieren von OPEA für Intel AI for Enterprise RAG Ihre Hauptkonfigurationsdatei, um die Verwendung von ONTAP S3 als Quelldaten-Repository zu ermöglichen.

Um die Verwendung von ONTAP S3 zu ermöglichen, legen Sie die folgenden Werte innerhalb der `edp` Abschnitt.

Hinweis: Standardmäßig nimmt die Intel AI for Enterprise RAG-Anwendung Daten aus allen vorhandenen Buckets in Ihrem SVM auf. Wenn Sie mehrere Buckets in Ihrem SVM haben, können Sie die `bucketNameRegexFilter` Feld, sodass Daten nur aus bestimmten Buckets aufgenommen werden.

```
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: ".*"
```

Konfigurieren der Einstellungen für die geplante Synchronisierung

Aktivieren Sie bei der Installation der OPEA für Intel AI for Enterprise RAG-Anwendung `scheduledSync` damit die Anwendung automatisch neue oder aktualisierte Dateien aus Ihren S3-Buckets aufnimmt.

Wann `scheduledSync` aktiviert ist, überprüft die Anwendung Ihre Quell-S3-Buckets automatisch auf neue oder aktualisierte Dateien. Alle neuen oder aktualisierten Dateien, die im Rahmen dieses Synchronisierungsprozesses gefunden werden, werden automatisch aufgenommen und der RAG-Wissensdatenbank hinzugefügt. Die Anwendung überprüft Ihre Quell-Buckets basierend auf einem voreingestellten Zeitintervall. Das Standardzeitintervall beträgt 60 Sekunden, was bedeutet, dass die Anwendung alle 60 Sekunden nach Änderungen sucht. Möglicherweise möchten Sie dieses Intervall Ihren speziellen Anforderungen entsprechend ändern.

So aktivieren Sie `scheduledSync` und legen Sie das Synchronisierungsintervall fest, legen Sie die folgenden Werte fest `deployment/components/edp/values.yaml`:

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

Ändern der Volume-Zugriffsmodi

In `deployment/components/gmc/microservices-connector/helm/values.yaml` , für jedes Volumen in der `pvc` Liste, ändern Sie die `accessMode` Zu `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
```

(Optional) Deaktivieren Sie die SSL-Zertifikatüberprüfung

Wenn Sie beim Aktivieren des S3-Zugriffs für Ihre SVM ein selbstsigniertes Zertifikat verwendet haben, müssen Sie die SSL-Zertifikatsüberprüfung deaktivieren. Wenn Sie ein Zertifikat verwendet haben, das von einer öffentlich vertrauenswürdigen Zertifizierungsstelle signiert ist, können Sie diesen Schritt überspringen.

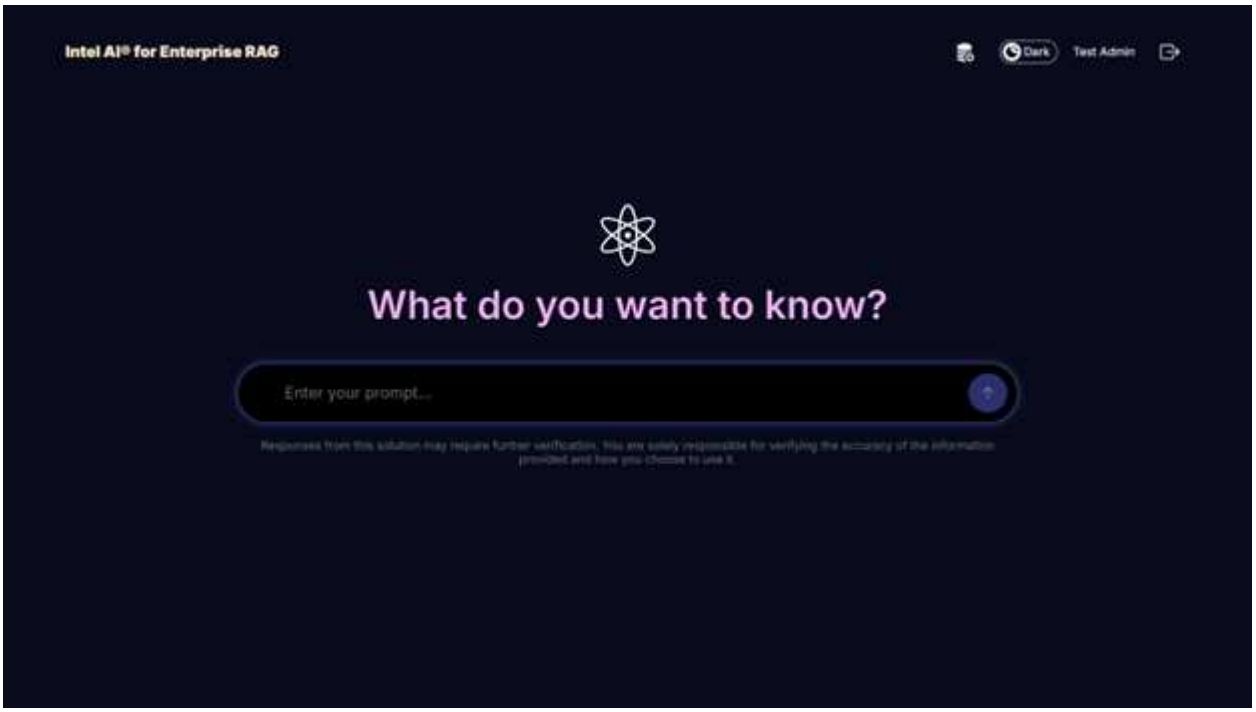
Um die SSL-Zertifikatsüberprüfung zu deaktivieren, legen Sie die folgenden Werte fest in `deployment/components/edp/values.yaml`:

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

Greifen Sie auf OPEA für Intel AI für Enterprise RAG UI zu

Greifen Sie auf die OPEA für die Intel AI for Enterprise RAG-Benutzeroberfläche zu. Weitere Informationen finden Sie im "Intel AI for Enterprise RAG-Bereitstellungsdokumentation" für Details.

Abbildung 8 – OPEA für Intel AI für Enterprise RAG-Benutzeroberfläche.



Daten für RAG aufnehmen

Sie können jetzt Dateien zur Einbeziehung in die RAG-basierte Abfrageerweiterung aufnehmen. Es gibt mehrere Optionen zum Einlesen von Dateien. Wählen Sie die passende Option für Ihre Anforderungen.

Hinweis: Nachdem eine Datei aufgenommen wurde, sucht die OPEA für Intel AI for Enterprise RAG-Anwendung automatisch nach Aktualisierungen der Datei und nimmt die Aktualisierungen entsprechend auf.

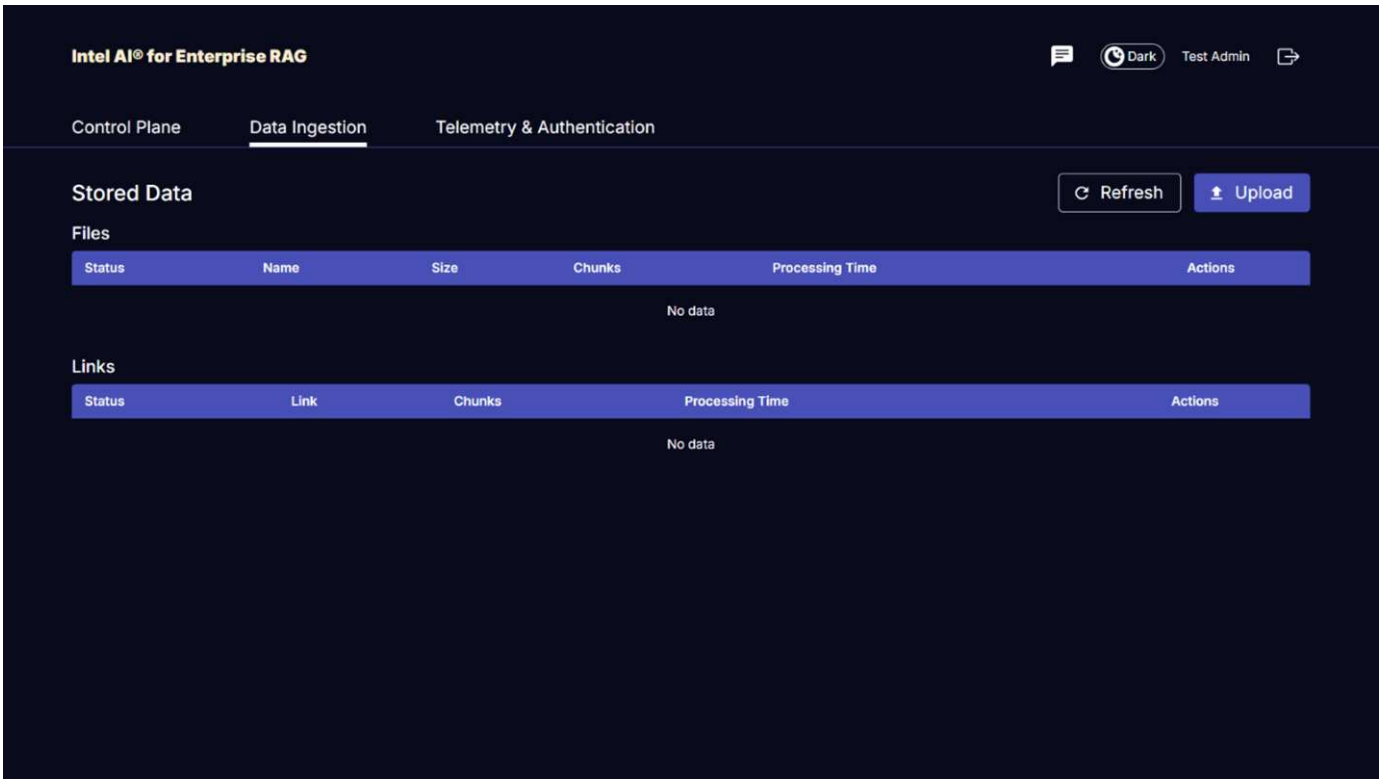
*Option 1: Direkt in Ihren S3-Bucket hochladen. Um viele Dateien auf einmal aufzunehmen, empfehlen wir, die Dateien mit dem S3-Client Ihrer Wahl in Ihren S3-Bucket (den Bucket, den Sie zuvor erstellt haben) hochzuladen. Zu den beliebten S3-Clients gehören die AWS CLI, das Amazon SDK für Python (Boto3), s3cmd, S3 Browser, Cyberduck und Commander One. Wenn es sich bei den Dateien um einen unterstützten Typ handelt, werden alle Dateien, die Sie in Ihren S3-Bucket hochladen, automatisch von der OPEA für die Intel AI for Enterprise RAG-Anwendung aufgenommen.

Hinweis: Zum Zeitpunkt der Erstellung dieses Dokuments werden die folgenden Dateitypen unterstützt: PDF, HTML, TXT, DOC, DOCX, PPT, PPTX, MD, XML, JSON, JSONL, YAML, XLS, XLSX, CSV, TIFF, JPG, JPEG, PNG und SVG.

Sie können die OPEA für die Intel AI for Enterprise RAG-Benutzeroberfläche verwenden, um zu bestätigen, dass Ihre Dateien ordnungsgemäß aufgenommen wurden. Weitere Informationen finden Sie in der Intel AI for Enterprise RAG UI-Dokumentation. Beachten Sie, dass es einige Zeit dauern kann, bis die Anwendung eine große Anzahl von Dateien aufgenommen hat.

*Option 2: Hochladen über die Benutzeroberfläche. Wenn Sie nur eine kleine Anzahl von Dateien aufnehmen müssen, können Sie diese über die OPEA für Intel AI for Enterprise RAG-Benutzeroberfläche aufnehmen. Weitere Informationen finden Sie in der Intel AI for Enterprise RAG UI-Dokumentation.

Abbildung 9 – Benutzeroberfläche zur Datenaufnahme.



Chat-Abfragen ausführen

Sie können jetzt mit der OPEA für die Intel AI for Enterprise RAG-Anwendung „chatten“, indem Sie die enthaltene Chat-Benutzeroberfläche verwenden. Bei der Beantwortung Ihrer Anfragen führt die Anwendung RAG mithilfe Ihrer aufgenommenen Dateien durch. Dies bedeutet, dass die Anwendung automatisch nach relevanten Informationen in Ihren aufgenommenen Dateien sucht und diese Informationen bei der Beantwortung Ihrer Anfragen berücksichtigt.

Größenberatung

Im Rahmen unserer Validierungsbemühungen haben wir in Abstimmung mit Intel Leistungstests durchgeführt. Das Ergebnis dieser Tests sind die in der folgenden Tabelle aufgeführten Größenrichtlinien.

Charakterisierungen	Wert	Kommentar
Modellgröße	20 Milliarden Parameter	Llama-8B, Llama-13B, Mistral 7B, Qwen 14B, DeepSeek Distill 8B
Eingabegröße	~2.000 Token	~4 Seiten
Ausgabegröße	~2.000 Token	~4 Seiten
Gleichzeitige Benutzer	32	„Gleichzeitige Benutzer“ bezieht sich auf Eingabeaufforderungen, die gleichzeitig Abfragen übermitteln.

Hinweis: Die oben aufgeführten Größenrichtlinien basieren auf Leistungsvalidierungen und Testergebnissen, die mit Intel Xeon 6-Prozessoren mit 96 Kernen gesammelt wurden. Für Kunden mit ähnlichen Anforderungen an E/A-Token und Modellgröße empfehlen wir die Verwendung von Servern mit Xeon 6-Prozessoren mit 96 oder 128 Kernen.

Abschluss

Enterprise-RAG-Systeme und LLMs sind Technologien, die zusammenarbeiten, um Unternehmen dabei zu helfen, genaue und kontextbezogene Antworten zu geben. Diese Antworten beinhalten die Informationsbeschaffung auf der Grundlage einer umfangreichen Sammlung privater und interner Unternehmensdaten. Durch die Verwendung von RAG, APIs, Vektoreinbettungen und Hochleistungsspeichersystemen zum Abfragen von Dokumentenspeichern, die Unternehmensdaten enthalten, werden die Daten schneller und sicherer verarbeitet. Der NetApp AIPod Mini kombiniert die intelligente Dateninfrastruktur von NetApp mit ONTAP Datenverwaltungsfunktionen und Intel Xeon 6-Prozessoren, Intel AI für Enterprise RAG und dem OPEA-Software-Stack, um die Bereitstellung leistungsstarker RAG-Anwendungen zu unterstützen und Unternehmen auf den Weg zur KI-Führung zu bringen.

Anerkennung

Dieses Dokument ist das Werk von Sathish Thyagarajan und Michael Ogelsby, Mitgliedern des NetApp Solutions Engineering-Teams. Die Autoren möchten sich außerdem beim Enterprise AI-Produktteam bei Intel – Ajay Mungara, Mikolaj Zyczynski, Igor Konopko, Ramakrishna Karamsetty, Michal Prostko, Shreejan Mistry und Ned Fiori – und den anderen Teammitgliedern bei NetApp– Lawrence Bunka, Bobby Oommen und Jeff Liborio – für ihre kontinuierliche Unterstützung und Hilfe während der Validierung dieser Lösung bedanken.

Stückliste

Die folgende Stückliste wurde für die Funktionsvalidierung dieser Lösung verwendet und kann als Referenz verwendet werden. Es kann jeder Server oder jede Netzwerkkomponente (oder sogar ein vorhandenes Netzwerk mit vorzugsweise 100 GbE Bandbreite) verwendet werden, die mit der folgenden Konfiguration übereinstimmt.

Für den App-Server:

Teilenummer	Produktbeschreibung	Menge
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
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 960GB 1DWPD TLC D, 80mm	2
	WS-1K63A-1R(x2)1U 692W/1600W redundantes Netzteil mit Einzelausgang. Wärmeableitung von 2361 BTU/h bei einer maximalen Temperatur von 59 °C (ca.)	4

Für den Kontrollserver:

Teilenummer	Produktbeschreibung	Menge
511R-M-OTO-17	OPTIMIERT UP 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, 80mm	2

Für den Netzwerk-Switch:

Teilenummer	Produktbeschreibung	Menge
DCS-7280CR3A	Arista 7280R3A 28x100 GbE	1

NetApp AFF -Speicher:

Teilenummer	Produktbeschreibung	Menge
AFF-A20A-100-C	AFF A20 HA System, -C	1
X800-42U-R6-C	Überbrückungsbatterie, In-Cab, C13-C14, -C	2
X97602A-C	Netzteil, 1600 W, Titan, -C	2

X66211B-2-N-C	Kabel, 100GbE, QSFP28-QSFP28, Cu, 2m, -C	4
X66240A-05-N-C	Kabel, 25GbE, SFP28-SFP28, Cu, 0,5m, -C	2
X5532A-N-C	Schiene, 4-Pfosten, dünn, rund/quadratisch, klein, verstellbar, 24–32, -C	1
X4024A-2-A-C	Laufwerkspaket 2 x 1,92 TB, NVMe4, SED, -C	6
X60130A-C	IO-Modul, 2PT, 100GbE, -C	2
X60132A-C	IO-Modul, 4PT, 10/25GbE, -C	2
SW-ONTAPB-FLASH-A20-C	SW, ONTAP -Basispaket, pro TB, Flash, A20, -C	23

Wo Sie weitere Informationen finden

Weitere Informationen zu den in diesem Dokument beschriebenen Informationen finden Sie in den folgenden Dokumenten und/oder auf den folgenden Websites:

"NetApp Produktdokumentation"

"OPEA-Projekt"

"Playbook zur OPEA Enterprise RAG-Bereitstellung"