NetApp AIPod Mini - Inférence RAG d’entreprise avec NetApp et Intel

NetApp artificial intelligence solutions

NetApp
August 20, 2025

Sommaire

NetApp AIPod Mini - Inférence RAG d’entreprise avec NetApp et Intel

Cet article présente une conception de référence validée de NetApp AIPod pour Enterprise RAG avec les technologies et les capacités combinées des processeurs Intel Xeon 6 et des solutions de gestion de données NetApp . La solution démontre une application ChatQnA en aval exploitant un modèle linguistique de grande taille, fournissant des réponses précises et contextuellement pertinentes aux utilisateurs simultanés. Les réponses sont récupérées à partir du référentiel de connaissances interne d’une organisation via un pipeline d’inférence RAG isolé.

Sathish Thyagarajan, Michael Oglesby, NetApp

Résumé exécutif

Un nombre croissant d’organisations exploitent des applications de génération augmentée de récupération (RAG) et des modèles de langage volumineux (LLM) pour interpréter les invites des utilisateurs et générer des réponses afin d’augmenter la productivité et la valeur commerciale. Ces invites et réponses peuvent inclure du texte, du code, des images ou même des structures de protéines thérapeutiques récupérées à partir de la base de connaissances interne d’une organisation, des lacs de données, des référentiels de codes et des référentiels de documents. Cet article couvre la conception de référence de la solution NetApp AIPod Mini, comprenant le stockage NetApp AFF et les serveurs équipés de processeurs Intel Xeon 6. Il comprend le logiciel de gestion de données NetApp ONTAP combiné à Intel Advanced Matrix Extensions (Intel AMX) et au logiciel Intel AI for Enterprise Retrieval-augmented Generation (RAG) basé sur Open Platform for Enterprise AI (OPEA). Le NetApp AIPod Mini pour RAG d’entreprise permet aux organisations d’augmenter un LLM public en une solution d’inférence d’IA générative privée (GenAI). La solution démontre une inférence RAG efficace et rentable à l’échelle de l’entreprise, conçue pour améliorer la fiabilité et vous offrir un meilleur contrôle sur vos informations propriétaires.

Validation des partenaires de stockage Intel

Les serveurs équipés de processeurs Intel Xeon 6 sont conçus pour gérer des charges de travail d’inférence d’IA exigeantes, en utilisant Intel AMX pour des performances maximales. Pour permettre des performances de stockage et une évolutivité optimales, la solution a été validée avec succès à l’aide de NetApp ONTAP, permettant aux entreprises de répondre aux besoins des applications RAG. Cette validation a été réalisée sur des serveurs équipés de processeurs Intel Xeon 6. Intel et NetApp ont un partenariat solide axé sur la fourniture de solutions d’IA optimisées, évolutives et alignées sur les exigences commerciales des clients.

Avantages de l’exécution de systèmes RAG avec NetApp

Les applications RAG impliquent la récupération de connaissances à partir des référentiels de documents des entreprises sous différents types tels que PDF, texte, CSV, Excel ou graphiques de connaissances. Ces données sont normalement stockées dans des solutions telles qu’un stockage d’objets S3 ou NFS sur site comme source de données. NetApp est un leader dans les technologies de gestion des données, de mobilité des données, de gouvernance des données et de sécurité des données dans l’écosystème de la périphérie, du

centre de données et du cloud. La gestion des données NetApp ONTAP fournit un stockage de niveau entreprise pour prendre en charge différents types de charges de travail d'IA, y compris l'inférence par lots et en temps réel, et offre certains des avantages suivants :

- Vitesse et évolutivité. Vous pouvez gérer de grands ensembles de données à grande vitesse pour le contrôle de version avec la possibilité de faire évoluer les performances et la capacité de manière indépendante.
- Accès aux données. La prise en charge multiprotocole permet aux applications clientes de lire des données à l'aide des protocoles de partage de fichiers S3, NFS et SMB. Les buckets NAS ONTAP S3 peuvent faciliter l'accès aux données dans les scénarios d'inférence LLM multimodaux.
- Fiabilité et confidentialité. ONTAP offre une protection des données, une protection autonome contre les ransomwares NetApp intégrée (ARP) et un provisionnement dynamique du stockage. Il offre également un cryptage logiciel et matériel pour améliorer la confidentialité et la sécurité. ONTAP est conforme à la norme FIPS 140-2 pour toutes les connexions SSL.

Public cible

Ce document est destiné aux décideurs en matière d'IA, aux ingénieurs de données, aux chefs d'entreprise et aux cadres de service qui souhaitent profiter d'une infrastructure conçue pour fournir des solutions RAG et GenAI d'entreprise. Une connaissance préalable de l'inférence de l'IA, des LLM, de Kubernetes, de la mise en réseau et de ses composants sera utile pendant la phase de mise en œuvre.

Exigences technologiques

Matériel

Technologies d'IA d'Intel

Avec Xeon 6 comme processeur hôte, les systèmes accélérés bénéficient de performances monothread élevées, d'une bande passante mémoire plus élevée, d'une fiabilité, d'une disponibilité et d'une facilité d'entretien (RAS) améliorées et de davantage de voies d'E/S. Intel AMX accélère l'inférence pour INT8 et BF16 et offre une prise en charge des modèles formés FP16, avec jusqu'à 2 048 opérations à virgule flottante par cycle et par cœur pour INT8 et 1 024 opérations à virgule flottante par cycle et par cœur pour BF16/FP16. Pour déployer une solution RAG utilisant des processeurs Xeon 6, une RAM minimale de 250 Go et un espace disque de 500 Go sont généralement recommandés. Cependant, cela dépend fortement de la taille du modèle LLM. Pour plus d'informations, reportez-vous à la documentation Intel "Processeur Xeon 6" fiche produit.

Figure 1 - Serveur de calcul avec processeurs Intel Xeon

6

Stockage NetApp AFF

Les systèmes NetApp AFF A-Series d'entrée de gamme et de milieu de gamme offrent des performances plus puissantes, une densité plus élevée et une plus grande efficacité. Les systèmes NetApp AFF A20, AFF A30 et AFF A50 offrent un véritable stockage unifié qui prend en charge les blocs, les fichiers et les objets, basé sur un système d'exploitation unique capable de gérer, de protéger et de mobiliser de manière transparente les données pour les applications RAG au coût le plus bas dans le cloud hybride.

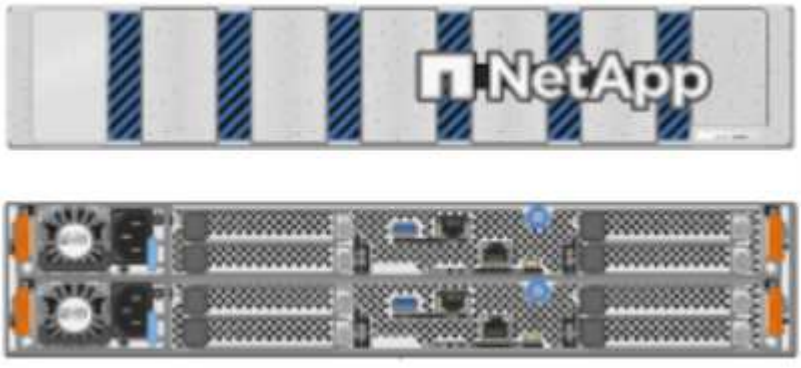


Figure 2 - Système NetApp AFF série A.

Matériel	Quantité	Commentaire
Serveur basé sur Intel Xeon 6	2	Nœuds d'inférence RAG : avec processeurs Intel Xeon série 6900 ou Intel Xeon série 6700 à double socket et 250 Go à 3 To de RAM avec DDR5 (6 400 MHz) ou MRDIMM (8 800 MHz). Serveur 2U.
Serveur de plan de contrôle avec processeur Intel	1	Plan de contrôle Kubernetes/serveur 1U.

Matériel	Quantité	Commentaire
Choix de commutateur Ethernet 100 Gb	1	Commutateur de centre de données.
NetApp AFF A20 (ou AFF A30 ; AFF A50)	1	Capacité de stockage maximale : 9,3 Po. Remarque : mise en réseau : ports 10/25/100 GbE.

Pour la validation de cette conception de référence, des serveurs équipés de processeurs Intel Xeon 6 de Supermicro (222HA-TN-OTO-37) et d'un commutateur 100GbE d'Arista (7280R3A) ont été utilisés.

Logiciels

Plateforme ouverte pour l'IA d'entreprise

L'Open Platform for Enterprise AI (OPEA) est une initiative open source dirigée par Intel en collaboration avec des partenaires de l'écosystème. Il fournit une plate-forme modulaire de blocs de construction composables conçus pour accélérer le développement de systèmes d'IA génératifs de pointe, avec un fort accent sur RAG. OPEA comprend un cadre complet comprenant des LLM, des magasins de données, des moteurs d'invite, des plans architecturaux RAG et une méthode d'évaluation en quatre étapes qui évalue les systèmes d'IA génératifs en fonction des performances, des fonctionnalités, de la fiabilité et de la préparation de l'entreprise.

À la base, l'OPEA comprend deux éléments clés :

- GenAIComps : une boîte à outils basée sur les services composée de composants de microservices
- Exemples GenAI : des solutions prêtes à être déployées comme ChatQnA qui illustrent des cas d'utilisation pratiques

Pour plus de détails, voir le "Documentation du projet OPEA"

Intel AI for Enterprise inference optimisé par OPEA

OPEA pour Intel AI for Enterprise RAG simplifie la transformation de vos données d'entreprise en informations exploitables. Alimenté par des processeurs Intel Xeon, il intègre des composants de partenaires industriels pour offrir une approche simplifiée du déploiement de solutions d'entreprise. Il s'adapte de manière transparente aux cadres d'orchestration éprouvés, offrant la flexibilité et le choix dont votre entreprise a besoin.

S'appuyant sur les fondations d'OPEA, Intel AI for Enterprise RAG étend cette base avec des fonctionnalités clés qui améliorent l'évolutivité, la sécurité et l'expérience utilisateur. Ces fonctionnalités incluent des capacités de maillage de services pour une intégration transparente avec les architectures modernes basées sur les services, une validation prête pour la production pour la fiabilité du pipeline et une interface utilisateur riche en fonctionnalités pour RAG en tant que service, permettant une gestion et une surveillance faciles des flux de travail. De plus, Intel et le support de ses partenaires offrent l'accès à un vaste écosystème de solutions, associé à une gestion intégrée des identités et des accès (IAM) avec interface utilisateur et applications pour des opérations sécurisées et conformes. Les garde-fous programmables offrent un contrôle précis du comportement du pipeline, permettant des paramètres de sécurité et de conformité personnalisés.

NetApp ONTAP

NetApp ONTAP est la technologie fondamentale qui sous-tend les solutions de stockage de données critiques de NetApp. ONTAP comprend diverses fonctionnalités de gestion et de protection des données, telles que la protection automatique contre les ransomwares contre les cyberattaques, des fonctionnalités de transport de données intégrées et des capacités d'efficacité de stockage. Ces avantages s'appliquent à une gamme

d’architectures, du sur site au multicloud hybride en NAS, SAN, objet et stockage défini par logiciel pour les déploiements LLM. Vous pouvez utiliser un serveur de stockage d’objets ONTAP S3 dans un cluster ONTAP pour déployer des applications RAG, en tirant parti de l’efficacité du stockage et de la sécurité d' ONTAP, fournies par les utilisateurs autorisés et les applications clientes. Pour plus d’informations, reportez-vous à "En savoir plus sur la configuration ONTAP S3"

NetApp Trident

Le logiciel NetApp Trident est un orchestrateur de stockage open source et entièrement pris en charge pour les conteneurs et les distributions Kubernetes, y compris Red Hat OpenShift. Trident fonctionne avec l’ensemble du portefeuille de stockage NetApp , y compris NetApp ONTAP , et prend également en charge les connexions NFS et iSCSI. Pour plus d’informations, reportez-vous à "NetApp Trident sur Git"

Logiciel	Version	Commentaire
OPEA pour Intel AI pour Enterprise RAG	1.1.2	Plateforme RAG d’entreprise basée sur les microservices OPEA
Interface de stockage de conteneurs (pilote CSI)	NetApp Trident 25.02	Permet le provisionnement dynamique, les copies NetApp Snapshot et les volumes.
Ubuntu	22.04.5	Système d’exploitation sur un cluster à deux nœuds
Orchestration des conteneurs	Kubernetes 1.31.4	Environnement pour exécuter le framework RAG
ONTAP	ONTAP 9.16.1P4	Stockage OS sur AFF A20. Il comprend Vscan et ARP.

Déploiement de la solution

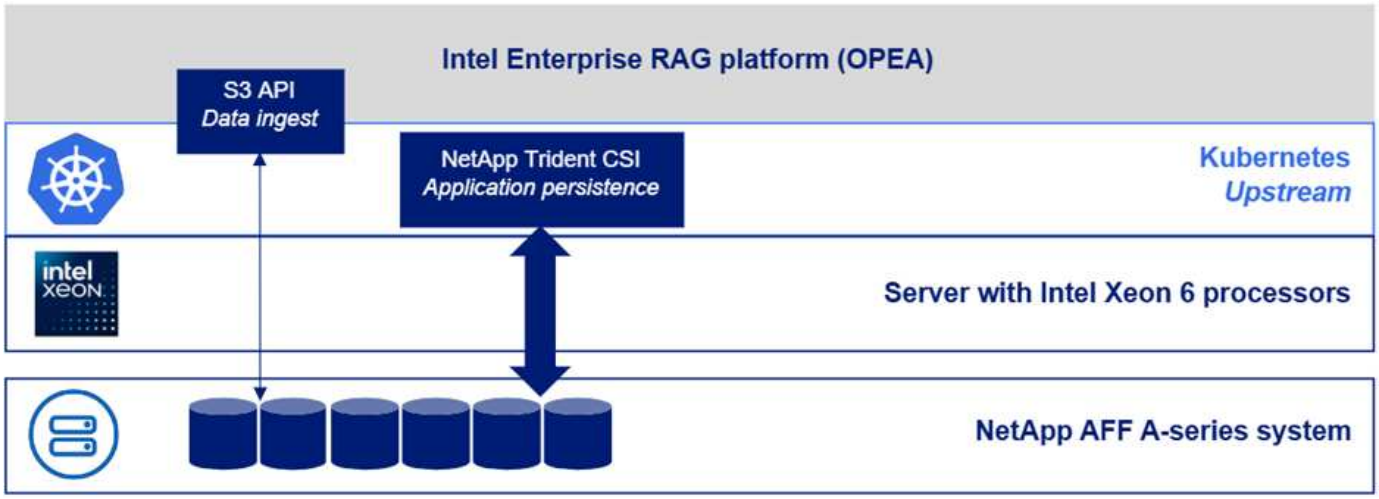
Pile logicielle

La solution est déployée sur un cluster Kubernetes composé de nœuds d’application basés sur Intel Xeon. Au moins trois nœuds sont nécessaires pour implémenter la haute disponibilité de base pour le plan de contrôle Kubernetes. Nous avons validé la solution en utilisant la disposition de cluster suivante.

Tableau 3 - Disposition du cluster Kubernetes

Nœud	Rôle	Quantité
Serveurs avec processeurs Intel Xeon 6 et 1 To de RAM	Nœud d’application, nœud de plan de contrôle	2
Serveur générique	Nœud de plan de contrôle	1

La figure suivante illustre une « vue de la pile logicielle » de la solution.



Étapes de déploiement

Déployer un dispositif de stockage ONTAP

Déployez et provisionnez votre appliance de stockage NetApp ONTAP . Se référer à la "Documentation des systèmes matériels ONTAP" pour plus de détails.

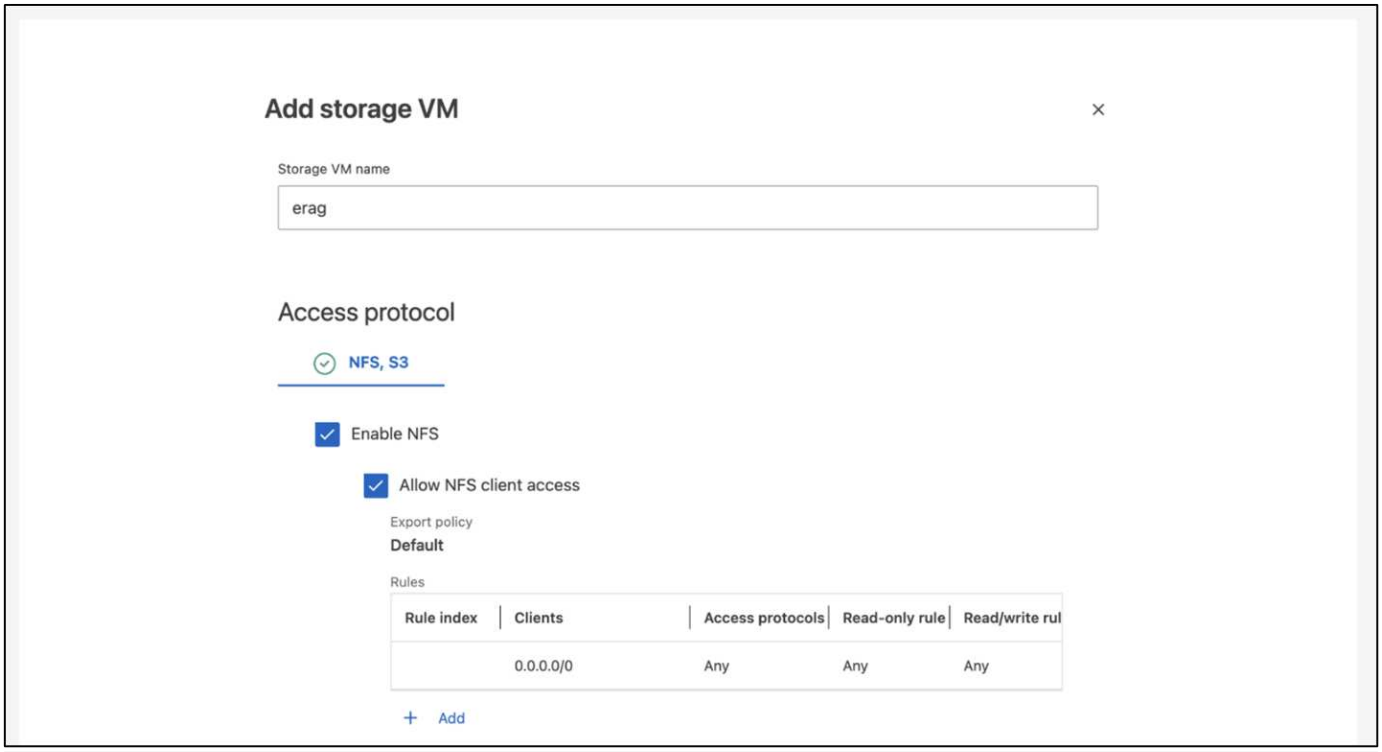
Configurer une SVM ONTAP pour l'accès NFS et S3

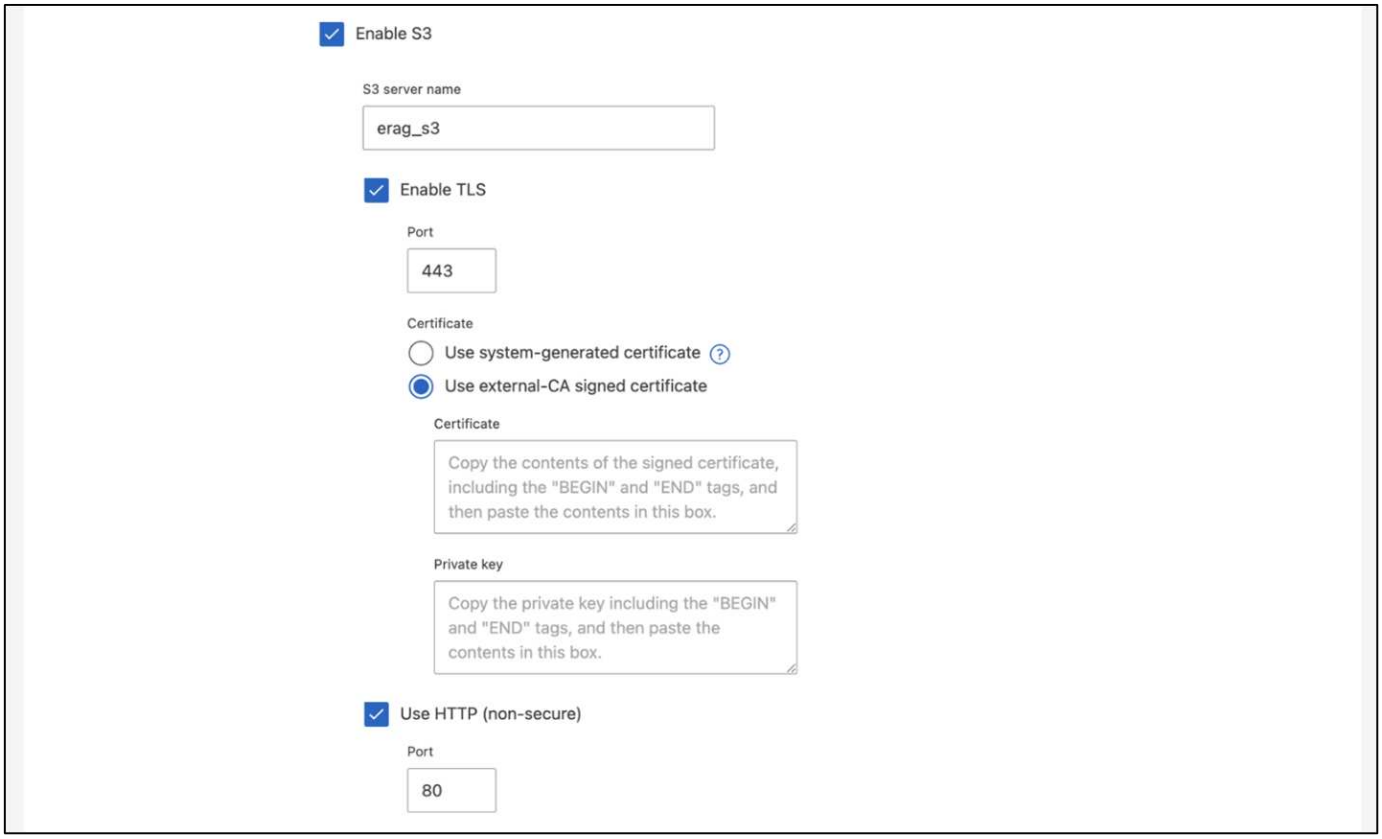
Configurez une machine virtuelle de stockage ONTAP (SVM) pour l'accès NFS et S3 sur un réseau accessible par vos nœuds Kubernetes.

Pour créer une SVM à l'aide d' ONTAP System Manager, accédez à Stockage > Machines virtuelles de stockage, puis cliquez sur le bouton + Ajouter. Lorsque vous activez l'accès S3 pour votre SVM, choisissez l'option permettant d'utiliser un certificat signé par une autorité de certification externe, et non un certificat généré par le système. Vous pouvez utiliser un certificat auto-signé ou un certificat signé par une autorité de certification publiquement approuvée. Pour plus de détails, reportez-vous à la "Documentation ONTAP ."

La capture d'écran suivante illustre la création d'un SVM à l'aide d' ONTAP System Manager. Modifiez les détails selon vos besoins en fonction de votre environnement.

Figure 4 - Création de SVM à l'aide d' ONTAP System Manager.





Configurer les autorisations S3

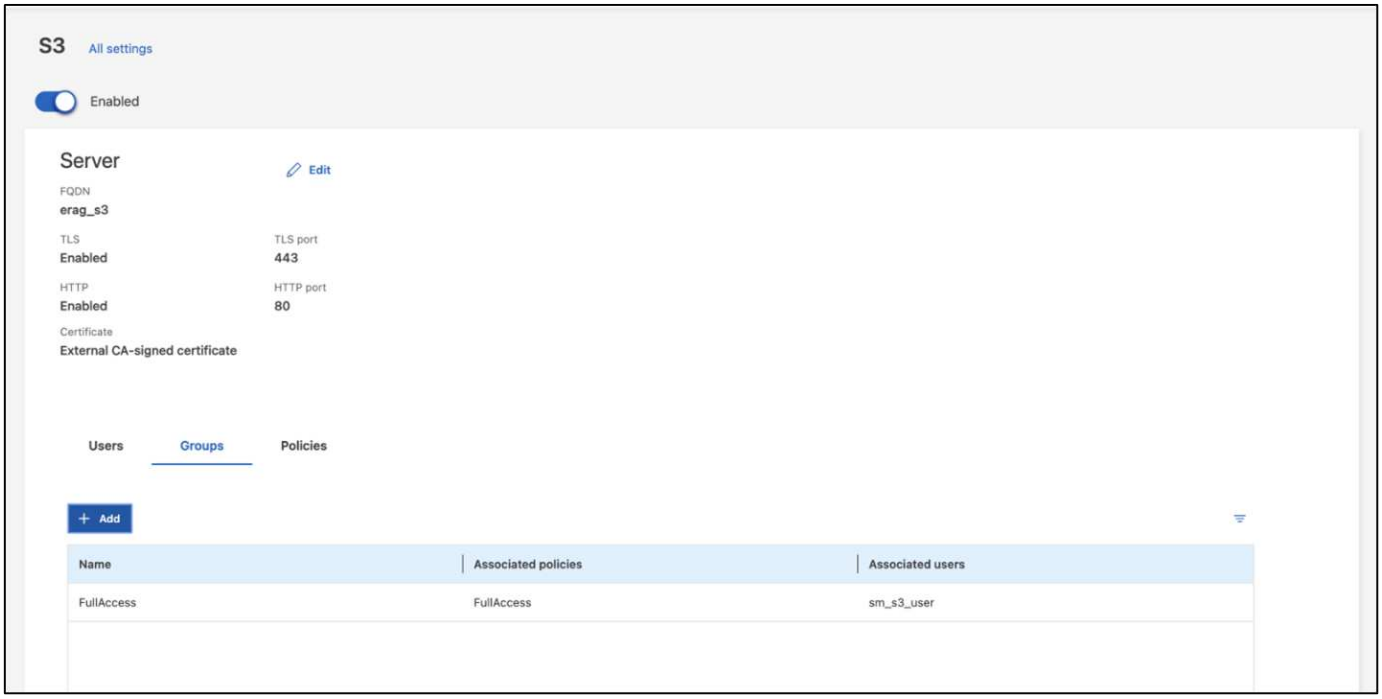
Configurez les paramètres utilisateur/groupe S3 pour le SVM que vous avez créé à l'étape précédente. Assurez-vous d'avoir un utilisateur avec un accès complet à toutes les opérations de l'API S3 pour ce SVM. Reportez-vous à la documentation ONTAP S3 pour plus de détails.

Remarque : cet utilisateur sera nécessaire pour le service d'ingestion de données de l'application Intel AI for Enterprise RAG. Si vous avez créé votre SVM à l'aide d' ONTAP System Manager, System Manager aura

automatiquement créé un utilisateur nommé `sm_s3_user` et une politique nommée `FullAccess` lorsque vous avez créé votre SVM, mais aucune autorisation ne vous aura été attribuée `sm_s3_user` .

Pour modifier les autorisations de cet utilisateur, accédez à Stockage > Machines virtuelles de stockage, cliquez sur le nom de la SVM que vous avez créée à l'étape précédente, cliquez sur Paramètres, puis sur l'icône en forme de crayon à côté de « S3 ». Donner `sm_s3_user` accès complet à toutes les opérations de l'API S3, créez un nouveau groupe qui associe `sm_s3_user` avec le `FullAccess` politique telle que décrite dans la capture d'écran suivante.

Figure 5 - Autorisations S3.

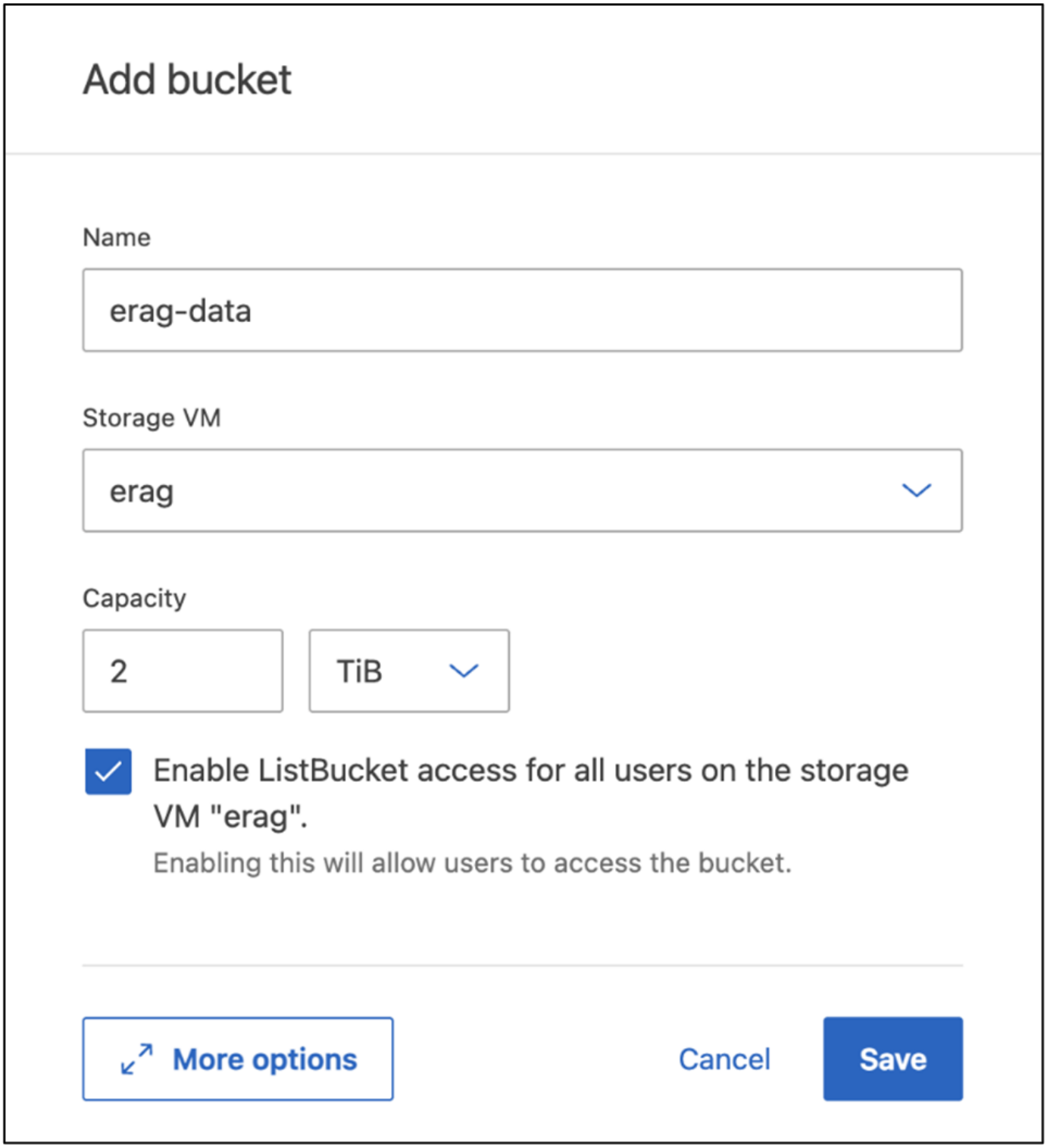


Créer un bucket S3

Créez un bucket S3 dans la SVM que vous avez créée précédemment. Pour créer un SVM à l'aide d' ONTAP System Manager, accédez à Stockage > Buckets et cliquez sur le bouton + Ajouter. Pour plus de détails, reportez-vous à la documentation ONTAP S3.

La capture d'écran suivante illustre la création d'un bucket S3 à l'aide d' ONTAP System Manager.

Figure 6 – Créer un bucket
S3.



Configurer les autorisations du compartiment S3

Configurez les autorisations pour le compartiment S3 que vous avez créé à l'étape précédente. Assurez-vous que l'utilisateur que vous avez configuré à une étape précédente dispose des autorisations suivantes : `GetObject, PutObject, DeleteObject, ListBucket, GetBucketAcl, GetObjectAcl, ListBucketMultipartUploads, ListMultipartUploadParts, GetObjectTagging, PutObjectTagging, DeleteObjectTagging, GetBucketLocation, GetBucketVersioning, PutBucketVersioning, ListBucketVersions, GetBucketPolicy, PutBucketPolicy, DeleteBucketPolicy, PutLifecycleConfiguration, GetLifecycleConfiguration,`

`GetBucketCORS, PutBucketCORS.`

Pour modifier les autorisations du compartiment S3 à l'aide d' ONTAP System Manager, accédez à Stockage > Compartiments, cliquez sur le nom de votre compartiment, cliquez sur Autorisations, puis sur Modifier. Se référer à la "Documentation ONTAP S3" pour plus de détails.

La capture d'écran suivante illustre les autorisations de compartiment nécessaires dans ONTAP System Manager.

Figure 7 - Autorisations du compartiment S3.

Créer une règle de partage de ressources inter-origines de bucket

À l'aide de l'interface de ligne de commande ONTAP , créez une règle de partage de ressources inter-origines (CORS) pour le bucket que vous avez créé à l'étape précédente :

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

Cette règle permet à l'application Web OPEA pour Intel AI for Enterprise RAG d'interagir avec le bucket à partir d'un navigateur Web.

Déployer des serveurs

Déployez vos serveurs et installez Ubuntu 22.04 LTS sur chaque serveur. Une fois Ubuntu installé, installez les utilitaires NFS sur chaque serveur. Pour installer les utilitaires NFS, exécutez la commande suivante :

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

Installer Kubernetes

Installez Kubernetes sur vos serveurs à l'aide de Kubespray. Se référer à la "Documentation de Kubespray" pour plus de détails.

Installer le pilote Trident CSI

Installez le pilote NetApp Trident CSI dans votre cluster Kubernetes. Se référer à la "Documentation d'installation de Trident" pour plus de détails.

Créer un back-end Trident

Créez un back-end Trident pour le SVM que vous avez créé précédemment. Lors de la création de votre back-end, utilisez le `ontap-nas` conducteur. Se référer à la "Documentation du back-end de Trident" pour plus de détails.

Créer une classe de stockage

Créez une classe de stockage Kubernetes correspondant au back-end Trident que vous avez créé à l'étape précédente. Reportez-vous à la documentation de la classe de stockage Trident pour plus de détails.

OPEA pour Intel AI pour Enterprise RAG

Installez OPEA pour Intel AI for Enterprise RAG dans votre cluster Kubernetes. Se référer à la "Déploiement de l'IA Intel pour le RAG d'entreprise" documentation pour plus de détails. Assurez-vous de prendre note des modifications requises du fichier de configuration qui sont décrites plus loin dans ce document. Vous devez effectuer ces modifications avant d'exécuter le playbook d'installation afin que l'application Intel AI for Enterprise RAG fonctionne correctement avec votre système de stockage ONTAP .

Activer l'utilisation d' ONTAP S3

Lors de l'installation d'OPEA pour Intel AI for Enterprise RAG, modifiez votre fichier de configuration principal pour permettre l'utilisation d' ONTAP S3 comme référentiel de données source.

Pour activer l'utilisation d' ONTAP S3, définissez les valeurs suivantes dans le `edp` section.

Remarque : par défaut, l'application Intel AI for Enterprise RAG ingère les données de tous les compartiments existants dans votre SVM. Si vous avez plusieurs buckets dans votre SVM, vous pouvez modifier le `bucketNameRegexFilter` champ afin que les données soient ingérées uniquement à partir de certains compartiments.

```
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: ".*"
```

Configurer les paramètres de synchronisation planifiée

Lors de l'installation de l'application OPEA pour Intel AI for Enterprise RAG, activez `scheduledSync` afin que l'application ingère automatiquement les fichiers nouveaux ou mis à jour à partir de vos buckets S3.

Quand `scheduledSync` est activé, l'application vérifie automatiquement vos buckets S3 sources pour les fichiers nouveaux ou mis à jour. Tous les fichiers nouveaux ou mis à jour trouvés dans le cadre de ce processus de synchronisation sont automatiquement ingérés et ajoutés à la base de connaissances RAG. L'application vérifie vos buckets sources en fonction d'un intervalle de temps prédéfini. L'intervalle de temps par défaut est de 60 secondes, ce qui signifie que l'application vérifie les modifications toutes les 60 secondes. Vous souhaiterez peut-être modifier cet intervalle en fonction de vos besoins spécifiques.

Pour activer `scheduledSync` et définissez l'intervalle de synchronisation, définissez les valeurs suivantes dans `deployment/components/edp/values.yaml`:

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

Modifier les modes d'accès au volume

Dans `deployment/components/gmc/microservices-connector/helm/values.yaml` , pour chaque volume du `pvc` liste, changer le `accessMode` à `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
```

(Facultatif) Désactiver la vérification du certificat SSL

Si vous avez utilisé un certificat auto-signé lors de l'activation de l'accès S3 pour votre SVM, vous devez désactiver la vérification du certificat SSL. Si vous avez utilisé un certificat signé par une autorité de certification publiquement approuvée, vous pouvez ignorer cette étape.

Pour désactiver la vérification du certificat SSL, définissez les valeurs suivantes dans

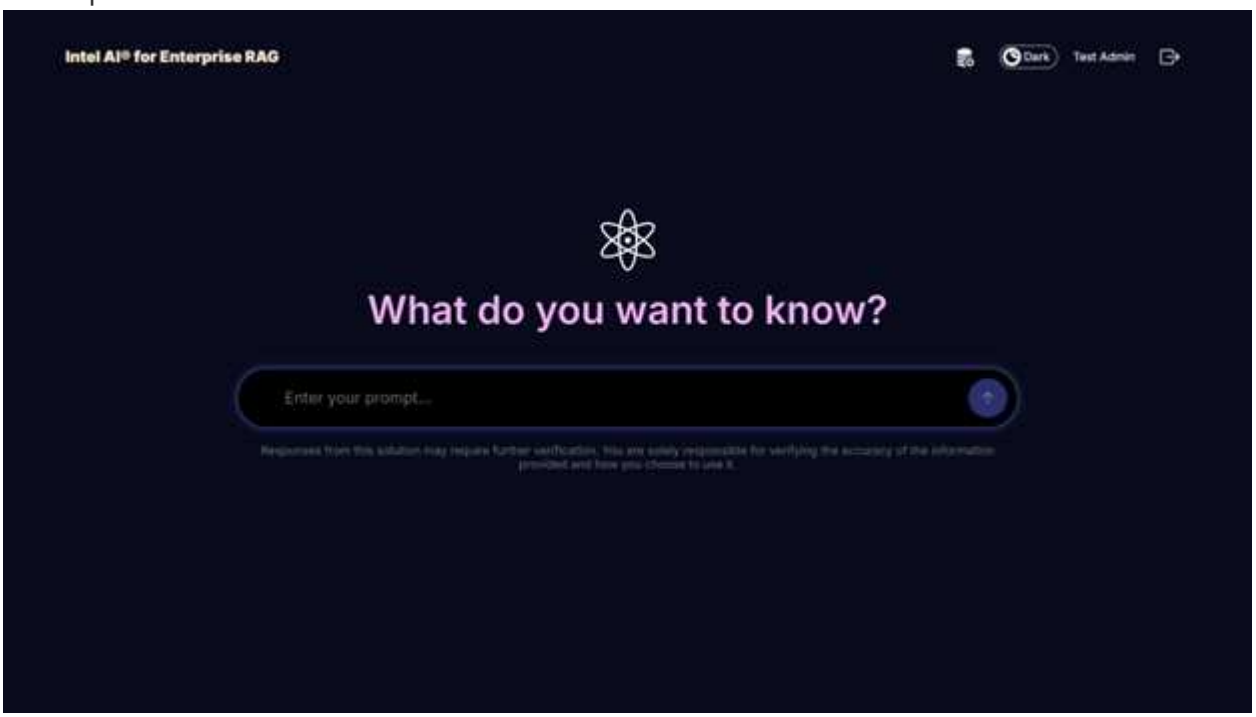
deployment/components/edp/values.yaml:

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

Accédez à OPEA pour Intel AI pour l'interface utilisateur RAG d'entreprise

Accédez à l'interface utilisateur RAG d'OPEA pour Intel AI for Enterprise. Se référer à la "Documentation sur le déploiement d'Intel AI for Enterprise RAG" pour plus de détails.

Figure 8 - OPEA pour Intel AI pour l'interface utilisateur RAG d'entreprise.



Ingérer des données pour RAG

Vous pouvez désormais ingérer des fichiers à inclure dans l'augmentation des requêtes basée sur RAG. Il existe plusieurs options pour ingérer des fichiers. Choisissez l'option appropriée à vos besoins.

Remarque : une fois qu'un fichier a été ingéré, l'application OPEA pour Intel AI for Enterprise RAG recherche automatiquement les mises à jour du fichier et ingère les mises à jour en conséquence.

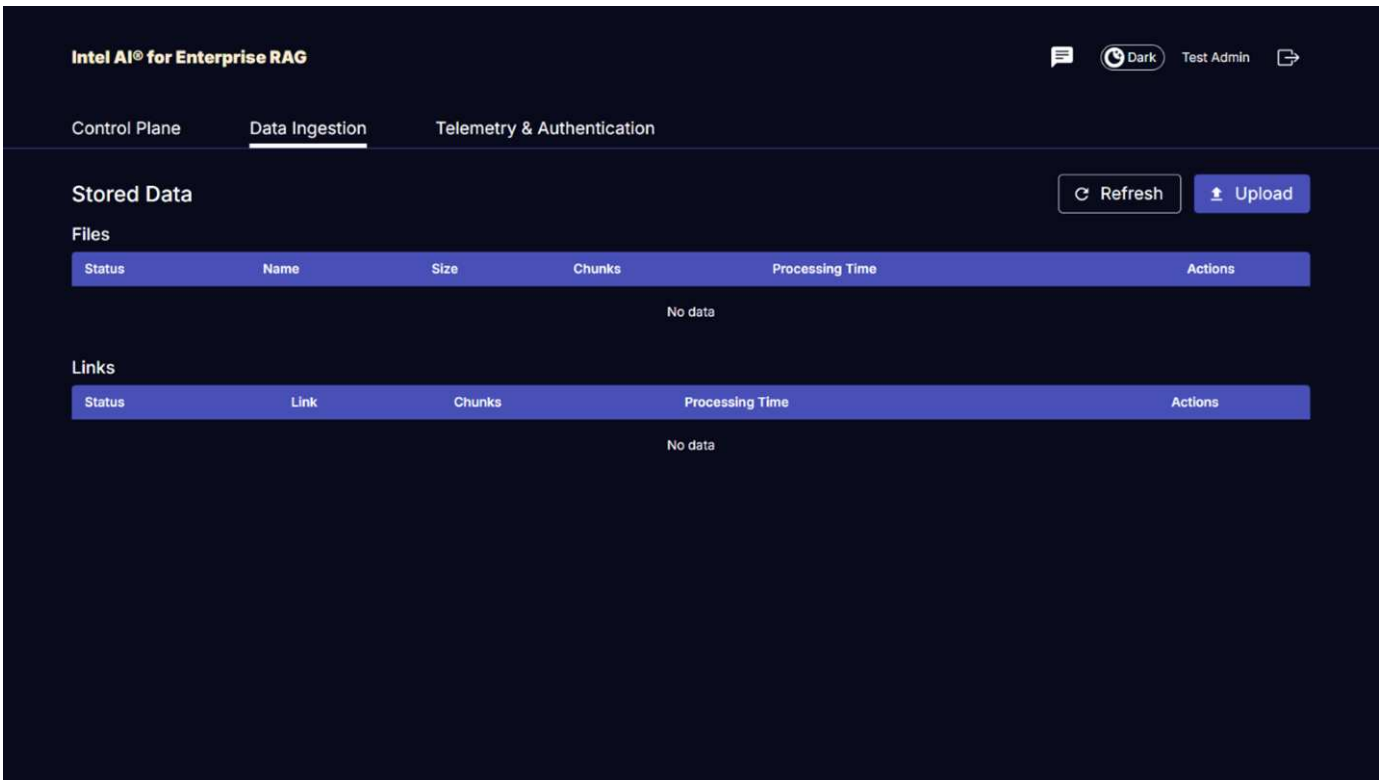
*Option 1 : télécharger directement dans votre bucket S3 Pour ingérer plusieurs fichiers à la fois, nous vous recommandons de télécharger les fichiers dans votre bucket S3 (le bucket que vous avez créé précédemment) en utilisant le client S3 de votre choix. Les clients S3 populaires incluent AWS CLI, Amazon SDK pour Python (Boto3), s3cmd, S3 Browser, Cyberduck et Commander One. Si les fichiers sont d'un type pris en charge, tous les fichiers que vous téléchargez dans votre compartiment S3 seront automatiquement ingérés par l'application OPEA pour Intel AI for Enterprise RAG.

Remarque : au moment de la rédaction de cet article, les types de fichiers suivants sont pris en charge : PDF, HTML, TXT, DOC, DOCX, PPT, PPTX, MD, XML, JSON, JSONL, YAML, XLS, XLSX, CSV, TIFF, JPG, JPEG, PNG et SVG.

Vous pouvez utiliser l'interface utilisateur OPEA pour Intel AI for Enterprise RAG pour confirmer que vos fichiers ont été correctement ingérés. Reportez-vous à la documentation de l'interface utilisateur Intel AI for Enterprise RAG pour plus de détails. Notez que l'application peut prendre un certain temps pour ingérer un grand nombre de fichiers.

*Option 2 : Télécharger à l'aide de l'interface utilisateur Si vous devez ingérer uniquement un petit nombre de fichiers, vous pouvez les ingérer à l'aide de l'interface utilisateur RAG d'OPEA pour Intel AI for Enterprise. Reportez-vous à la documentation de l'interface utilisateur Intel AI for Enterprise RAG pour plus de détails.

Figure 9 - Interface utilisateur d'ingestion de données.



Exécuter des requêtes de chat

Vous pouvez désormais « discuter » avec l'application OPEA pour Intel AI for Enterprise RAG en utilisant l'interface utilisateur de chat incluse. Lorsqu'elle répond à vos requêtes, l'application exécute RAG en utilisant vos fichiers ingérés. Cela signifie que l'application recherche automatiquement les informations pertinentes dans vos fichiers ingérés et intègre ces informations lorsqu'elle répond à vos requêtes.

Guide de dimensionnement

Dans le cadre de nos efforts de validation, nous avons effectué des tests de performances en coordination avec Intel. Ces tests ont permis d'établir les recommandations de dimensionnement décrites dans le tableau suivant.

Caractérisations	Valeur	Commentaire
Taille du modèle	20 milliards de paramètres	Llama-8B, Llama-13B, Mistral 7B, Qwen 14B, DeepSeek Distill 8B
Taille d'entrée	~2 000 jetons	~4 pages
Taille de sortie	~2 000 jetons	~4 pages
Utilisateurs simultanés	32	« Utilisateurs simultanés » fait référence aux demandes d'invite qui soumettent des requêtes en même temps.

_Remarque : les conseils de dimensionnement présentés ci-dessus sont basés sur la validation des performances et les résultats des tests recueillis à l'aide de processeurs Intel Xeon 6 avec 96 cœurs. Pour les clients ayant des exigences similaires en matière de jetons d'E/S et de taille de modèle, nous recommandons d'utiliser des serveurs équipés de processeurs Xeon 6 à 96 ou 128 cœurs.

Conclusion

Les systèmes RAG d'entreprise et les LLM sont des technologies qui fonctionnent ensemble pour aider les organisations à fournir des réponses précises et contextuelles. Ces réponses impliquent une recherche d'informations basée sur une vaste collection de données privées et internes à l'entreprise. En utilisant RAG, les API, les intégrations vectorielles et les systèmes de stockage hautes performances pour interroger les référentiels de documents contenant des données d'entreprise, les données sont traitées plus rapidement et en toute sécurité. Le NetApp AIPod Mini combine l'infrastructure de données intelligente de NetApp avec les capacités de gestion des données ONTAP et les processeurs Intel Xeon 6, Intel AI for Enterprise RAG et la pile logicielle OPEA pour aider à déployer des applications RAG hautes performances et mettre les organisations sur la voie du leadership en matière d'IA.

Reconnaissance

Ce document est l'œuvre de Sathish Thyagarajan et Michael Ogelsby, membres de l'équipe d'ingénierie des solutions NetApp . Les auteurs souhaitent également remercier l'équipe produit Enterprise AI d'Intel (Ajay Mungara, Mikolaj Zyczynski, Igor Konopko, Ramakrishna Karamsetty, Michal Prostko, Shreejan Mistry et Ned Fiori) ainsi que les autres membres de l'équipe de NetApp(Lawrence Bunka, Bobby Oommen et Jeff Liborio) pour leur soutien et leur aide continus lors de la validation de cette solution.

Nomenclature

Ce qui suit est la nomenclature utilisée pour la validation fonctionnelle de cette solution et peut être utilisée comme référence. Tout serveur ou composant réseau (ou même un réseau existant avec de préférence une bande passante de 100 GbE) qui correspond à la configuration suivante peut être utilisé.

Pour le serveur d'applications :

Référence	Description du produit	Quantité
222HA-TN-OTO-37	Hyper SuperServer SYS-222HA-TN /2U	2

Référence	Description du produit	Quantité
P4X-GNR6980P-SRPL2-UCC	Processeur Intel Xeon 6980P 2 cœurs 128C 2 Go 504 Mo 500 W SGX512	4
BÉLIER	Mémoire RDIMM ECC MEM-DR564MC-ER64(x16) 64 Go DDR5-6400 2RX4 (16 Go)	32
	HDS-M2N4-960G0-E1-TXD-NON-080(x2) SSD M.2 NVMe PCIe4 960 Go 1DWPD TLC D, 80 mm	2
	Alimentation à sortie unique redondante WS-1K63A-1R(x2)1U 692W/1600W. Dissipation thermique de 2361 BTU/h avec une température maximale de 59 °C (environ)	4

Pour le serveur de contrôle :

Référence	Description du produit	Quantité
511R-M-OTO-17	OPTIMISÉ 1U X13SCH-SYS, CSE-813MF2TS-R0RCNBP, PWS-602A-1R	1
	RPL-E 6369P IP 8C/16T 3,3G 24 Mo 95 W 1700 mAh	1
BÉLIER	MEM-DR516MB-EU48(x2)16 Go DDR5-4800 1Rx8 (16 Go) ECC UDIMM	1
	HDS-M2N4-960G0-E1-TXD-NON-080(x2) SSD M.2 NVMe PCIe4 960 Go 1DWPD TLC D, 80 mm	2

Pour le commutateur réseau :

Référence	Description du produit	Quantité
DCS-7280CR3A	Arista 7280R3A 28x100 GbE	1

Stockage NetApp AFF :

Référence	Description du produit	Quantité
AFF-A20A-100-C	Système AFF A20 HA, -C	1
X800-42U-R6-C	Jumper Crd, en cabine, C13-C14, -C	2
X97602A-C	Alimentation, 1600 W, titane, -C	2

X66211B-2-N-C	Câble, 100 GbE, QSFP28-QSFP28, Cu, 2 m, -C	4
X66240A-05-N-C	Câble, 25 GbE, SFP28-SFP28, Cu, 0,5 m, -C	2
X5532A-N-C	Rail, 4 poteaux, mince, trou rond/carré, petit, réglable, 24-32, -C	1
X4024A-2-A-C	Pack de disques 2 x 1,92 To, NVMe4, SED, -C	6
X60130A-C	Module d'E/S, 2PT, 100 GbE, -C	2
X60132A-C	Module d'E/S, 4 PT, 10/25 GbE, -C	2
SW-ONTAPB-FLASH-A20-C	SW, package de base ONTAP , par To, Flash, A20, -C	23

Où trouver des informations supplémentaires

Pour en savoir plus sur les informations décrites dans ce document, consultez les documents et/ou sites Web suivants :

"Documentation produit NetApp"

"Projet OPEA"

"Manuel de déploiement d'OPEA Enterprise RAG"

Informations sur le copyright

Informations sur les marques commerciales