NetApp AIPod Mini - NetAppとIntelによるエンタープライズRAG推論

NetApp artificial intelligence solutions

NetApp
August 20, 2025

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

目次

NetApp AIPod Mini - NetAppとIntelによるエンタープライズRAG推論 . . . 1
エグゼクティブサマリー . . . 1
Intel ストレージ パートナーの検証 . . . 1
NetAppでRAGシステムを実行する利点 . . . 1
対象 . . . 2
テクノロジ要件 . . . 2
ハードウェア . . . 2
ソフトウェア . . . 4
ソリューションの展開 . . . 5
ソフトウェアスタック . . . 5
導入手順 . . . 6
ONTAP S3の使用を有効にする . . . 11
スケジュールされた同期設定を構成する . . . 11
ボリュームアクセスモードを変更する . . . 12
（オプション）SSL証明書の検証を無効にする . . . 12
サイズガイド . . . 14
まとめ . . . 15
了承 . . . 15
部品表 . . . 15
詳細情報の入手方法 . . . 17

NetApp AIPod Mini - NetAppとIntelによるエンタープライズRAG推論

このホワイト ペーパーでは、Intel Xeon 6 プロセッサとNetAppデータ管理ソリューションのテクノロジと組み合わせた機能を備えた、エンタープライズ RAG 向けNetApp AIPodの検証済みリファレンス デザインについて説明します。このソリューションは、大規模な言語モデルを活用したダウンストリーム ChatQnA アプリケーションを示し、同時ユーザーに正確でコンテキストに適した応答を提供します。応答は、エアギャップ RAG 推論パイプラインを通じて組織の内部知識リポジトリから取得されます。

Sathish Thyagarajan、Michael Oglesby、 NetApp

エグゼクティブサマリー

生産性とビジネス価値を高めるために、検索拡張生成 (RAG) アプリケーションと大規模言語モデル (LLM) を活用して、ユーザーのプロンプトを解釈し、応答を生成する組織が増えています。これらのプロンプトと応答には、組織の内部ナレッジベース、データ レイク、コード リポジトリ、ドキュメント リポジトリから取得されたテキスト、コード、画像、さらには治療用タンパク質構造などが含まれます。このホワイト ペーパーでは、 NetApp AFFストレージと Intel Xeon 6 プロセッサを搭載したサーバーで構成されるNetApp AIPod Mini ソリューションのリファレンス デザインについて説明します。これには、Intel Advanced Matrix Extensions (Intel AMX) と組み合わせたNetApp ONTAPデータ管理ソフトウェアと、Open Platform for Enterprise AI (OPEA) 上に構築された Intel AI for Enterprise Retrieval-augmented Generation (RAG) ソフトウェアが含まれます。 NetApp AIPod Mini for enterprise RAG を使用すると、組織はパブリック LLM をプライベート生成 AI (GenAI) 推論ソリューションに拡張できます。このソリューションは、信頼性を高め、独自の情報をより適切に制御できるように設計された、エンタープライズ規模での効率的でコスト効率の高い RAG 推論を実現します。

Intel ストレージ パートナーの検証

Intel Xeon 6 プロセッサーを搭載したサーバーは、Intel AMX を使用して最高のパフォーマンスを実現し、要求の厳しい AI 推論ワークロードを処理できるように構築されています。最適なストレージ パフォーマンスとスケーラビリティを実現するために、このソリューションはNetApp ONTAPを使用して検証されており、企業は RAG アプリケーションのニーズを満たすことができます。この検証は、Intel Xeon 6 プロセッサを搭載したサーバーで実施されました。 Intel とNetApp は、最適化され、拡張可能で、顧客のビジネス要件に適合した AI ソリューションの提供に重点を置いた強力なパートナーシップを築いています。

NetAppでRAGシステムを実行する利点

RAG アプリケーションでは、PDF、テキスト、CSV、Excel、ナレッジ グラフなど、さまざまな形式で企業のドキュメント リポジトリからナレッジを取得します。このデータは通常、データのソースとして S3 オブジェクト ストレージやオンプレミスの NFS などのソリューションに保存されます。 NetApp は、エッジ、データセンター、クラウドのエコシステム全体にわたって、データ管理、データ モビリティ、データ ガバナンス、データ セキュリティ テクノロジーのリーダーです。 NetApp ONTAPデータ管理は、バッチやリアルタイム推論などのさまざまなタイプの AI ワークロードをサポートするエンタープライズ グレードのストレージを

提供し、次のような利点をもたらします。

- 速度とスケーラビリティ。パフォーマンスと容量を個別に拡張できるため、大規模なデータセットを高速に処理してバージョン管理を行うことができます。
- データアクセス。マルチプロトコル サポートにより、クライアント アプリケーションは S3、NFS、SMB ファイル共有プロトコルを使用してデータを読み取ることができます。 ONTAP S3 NAS バケットは、マルチモーダル LLM 推論シナリオでのデータ アクセスを容易にします。
- 信頼性と機密性。 ONTAP は、データ保護、組み込みのNetApp Autonomous Ransomware Protection (ARP)、ストレージの動的プロビジョニングを提供し、機密性とセキュリティを強化するためにソフトウェアベースとハードウェアベースの両方の暗号化を提供します。 ONTAP は、すべての SSL 接続に関して FIPS 140-2 に準拠しています。

対象

このドキュメントは、エンタープライズ RAG および GenAI ソリューションを提供するために構築されたインフラストラクチャを活用したい AI 意思決定者、データ エンジニア、ビジネス リーダー、部門幹部を対象としています。 AI 推論、LLM、Kubernetes、ネットワークとそのコンポーネントに関する事前の知識は、実装フェーズで役立ちます。

テクノロジ要件

ハードウェア

インテル**AI**テクノロジー

Xeon 6 をホスト CPU として使用することで、高速化されたシステムは、高いシングルスレッド パフォーマンス、メモリ帯域幅の拡大、信頼性、可用性、保守性 (RAS) の向上、および I/O レーンの増加といったメリットを享受できます。 Intel AMX は、INT8 および BF16 の推論を高速化し、FP16 トレーニング済みモデルのサポートを提供します。INT8 の場合はコアあたりサイクルあたり最大 2,048 回の浮動小数点演算、BF16/FP16 の場合はコアあたりサイクルあたり最大 1,024 回の浮動小数点演算が可能です。 Xeon 6 プロセッサを使用して RAG ソリューションを展開するには、通常、最低 250 GB の RAM と 500 GB のディスク容量が推奨されます。ただし、これは LLM モデルのサイズに大きく依存します。詳細については、Intel "Xeon 6プロセッサ"製品概要。

図1 - Intel Xeon 6プロセッサを搭載したコンピューティングサーバ

—

NetApp AFFストレージ

エントリーレベルおよびミッドレベルのNetApp AFF A シリーズ システムは、より強力なパフォーマンス、密度、および優れた効率性を提供します。 NetApp AFF A20、 AFF A30、 AFF A50 システムは、ハイブリッドクラウド全体で最低コストで RAG アプリケーションのデータをシームレスに管理、保護、および移動できる単一の OS に基づいて、ブロック、ファイル、およびオブジェクトをサポートする真の統合ストレージを提供します。

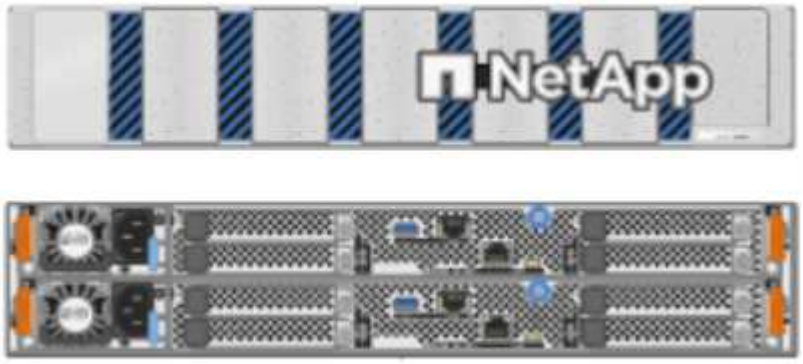


図 2 - NetApp AFF A シリーズ システム。

ハードウェア	量	コメント
Intel Xeon 6ベースのサーバー	2	RAG 推論ノード - デュアル ソケット Intel Xeon 6900 シリーズまたは Intel Xeon 6700 シリーズ プロセッサと、DDR5 (6400 MHz) または MRDIMM (8800 MHz) の 250 GB ～ 3 TB の RAM を搭載。 2Uサーバー。

ハードウェア	量	コメント
Intel プロセッサを搭載したコントロール プレーン サーバー	1	Kubernetes コントロール プレーン/1U サーバー。
100Gbイーサネットスイッチの選択	1	データセンタースイッチ。
NetApp AFF A20（またはAFF A30、 AFF A50）	1	最大ストレージ容量: 9.3PB。注: ネットワーク: 10/25/100 GbE ポート。

このリファレンス デザインの検証には、Supermicro の Intel Xeon 6 プロセッサ (222HA-TN-OTO-37) を搭載したサーバーと Arista の 100GbE スイッチ (7280R3A) が使用されました。

ソフトウェア

エンタープライズ**AI**向けオープンプラットフォーム

Open Platform for Enterprise AI (OPEA) は、Intel がエコシステム パートナーと協力して主導するオープン ソース イニシアチブです。 RAG に重点を置いた最先端の生成 AI システムの開発を加速するように設計された、構成可能なビルディング ブロックのモジュール式プラットフォームを提供します。 OPEA には、LLM、データストア、プロンプト エンジン、RAG アーキテクチャ ブループリント、パフォーマンス、機能、信頼性、エンタープライズ準備に基づいて生成 AI システムを評価する 4 段階の評価方法を備えた包括的なフレームワークが含まれています。

OPEA は、主に次の 2 つの主要コンポーネントで構成されています。

- GenAIComps: マイクロサービスコンポーネントで構成されたサービスベースのツールキット
- GenAIExamples: ChatQnAのような、実用的なユースケースを示すすぐに導入できるソリューション

詳細については、 "OPEAプロジェクトドキュメント"

OPEA を搭載した **Intel AI for Enterprise** 推論

Intel AI for Enterprise RAG の OPEA は、エンタープライズ データを実用的な洞察へと変換することを簡素化します。 Intel Xeon プロセッサーを搭載し、業界パートナーのコンポーネントを統合して、エンタープライズ ソリューションを導入するための合理的なアプローチを提供します。実績のあるオーケストレーション フレームワークを使用してシームレスに拡張し、企業が必要とする柔軟性と選択肢を提供します。

OPEA の基盤を基に、Intel AI for Enterprise RAG は、スケーラビリティ、セキュリティ、ユーザー エクスペリエンスを強化する主要な機能によってこの基盤を拡張します。これらの機能には、最新のサービスベースのアーキテクチャとのシームレスな統合を実現するサービス メッシュ機能、パイプラインの信頼性を本番環境で検証できる機能、ワークフローの管理と監視を容易にする RAG as a Service 用の機能豊富な UI などが含まれます。さらに、インテルとパートナーのサポートにより、安全でコンプライアンスに準拠した運用を実現する UI とアプリケーションを備えた統合 ID およびアクセス管理 (IAM) と組み合わせた、幅広いソリューション エコシステムへのアクセスが提供されます。プログラム可能なガードレールにより、パイプラインの動作をきめ細かく制御でき、セキュリティとコンプライアンスの設定をカスタマイズできます。

NetApp ONTAP

NetApp ONTAP は、NetApp の重要なデータ ストレージ ソリューションを支える基盤テクノロジーです。 ONTAP には、サイバー攻撃に対する自動ランサムウェア保護、組み込みのデータ転送機能、ストレージ効率

機能など、さまざまなデータ管理およびデータ保護機能が含まれています。これらの利点は、オンプレミスから、NAS、SAN、オブジェクト、LLM 展開のソフトウェア定義ストレージのハイブリッド マルチクラウドまで、さまざまなアーキテクチャに適用されます。 ONTAPクラスタ内のONTAP S3 オブジェクト ストレージ サーバを使用して RAG アプリケーションを導入し、承認されたユーザーとクライアント アプリケーションを通じて提供されるONTAPのストレージ効率とセキュリティを活用できます。詳細については、 "ONTAP S3 の構成について学ぶ"

NetAppTrident

NetApp Tridentソフトウェアは、Red Hat OpenShift を含むコンテナおよび Kubernetes ディストリビューション向けのオープンソースで完全にサポートされているストレージ オーケストレーターです。 Trident は、NetApp ONTAPを含むNetAppストレージ ポートフォリオ全体と連携し、NFS および iSCSI 接続もサポートします。詳細については、 "Git 上のNetApp Trident"

ソフトウェア	バージョン	コメント
エンタープライズ RAG 向けインテル AI の OPEA	1.1.2	OPEAマイクロサービスに基づくエンタープライズRAGプラットフォーム
コンテナ ストレージ インターフェース (CSI ドライバー)	NetAppTrident25.02	動的プロビジョニング、 NetAppスナップショット コピー、ボリュームを有効にします。
Ubuntu	22.04.5	2ノードクラスタ上のOS
コンテナオーケストレーション	Kubernetes 1.31.4	RAGフレームワークを実行する環境
ONTAP	ONTAP 9.16.1P4	AFF A20 上のストレージ OS。Vscan と ARP を備えています。

ソリューションの展開

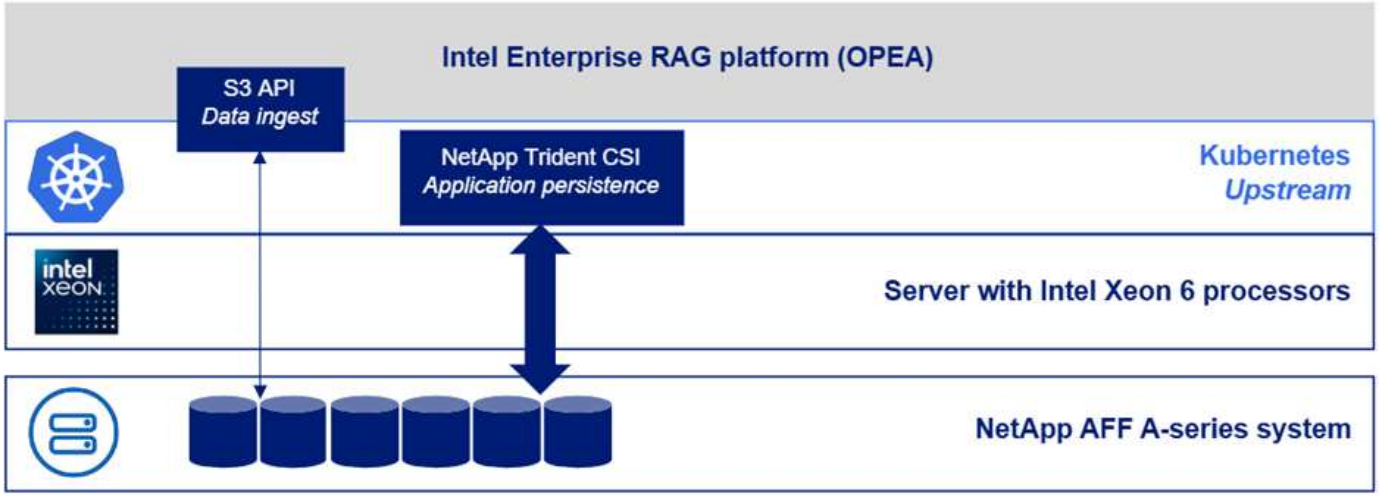
ソフトウェアスタック

このソリューションは、Intel Xeon ベースのアプリケーション ノードで構成される Kubernetes クラスターに展開されます。 Kubernetes コントロール プレーンの基本的な高可用性を実装するには、少なくとも 3 つのノードが必要です。次のクラスター レイアウトを使用してソリューションを検証しました。

表3 - Kubernetesクラスタレイアウト

ノード	ロール	数量
Intel Xeon 6プロセッサと1TB RAMを搭載したサーバー	アプリノード、コントロールプレーンノード	2
汎用サーバー	コントロールプレーンノード	1

次の図は、ソリューションの「ソフトウェア スタック ビュー」を示しています。



導入手順

ONTAPストレージアプライアンスを導入する

NetApp ONTAPストレージ アプライアンスを導入およびプロビジョニングします。参照 "ONTAPハードウェア システムのドキュメント"詳細については。

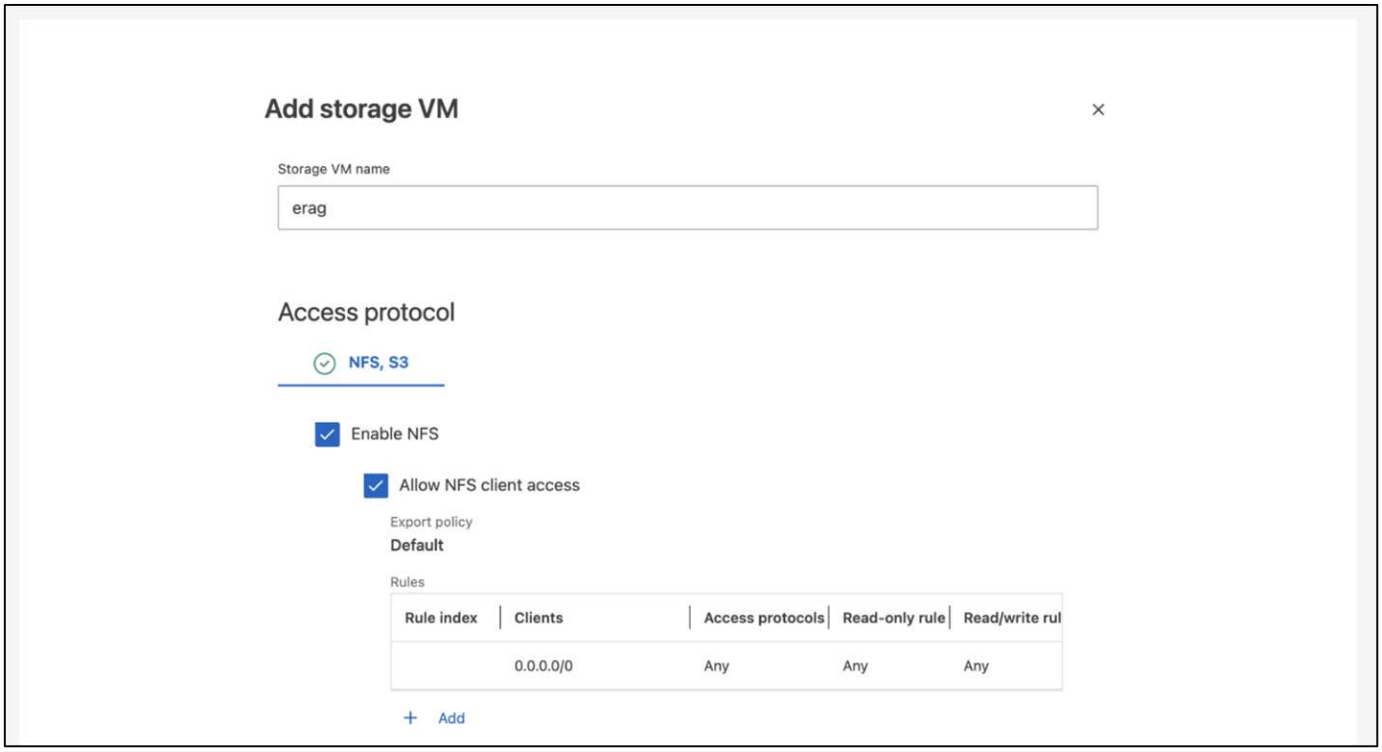
NFSおよび**S3**アクセス用に**ONTAP SVM**を構成する

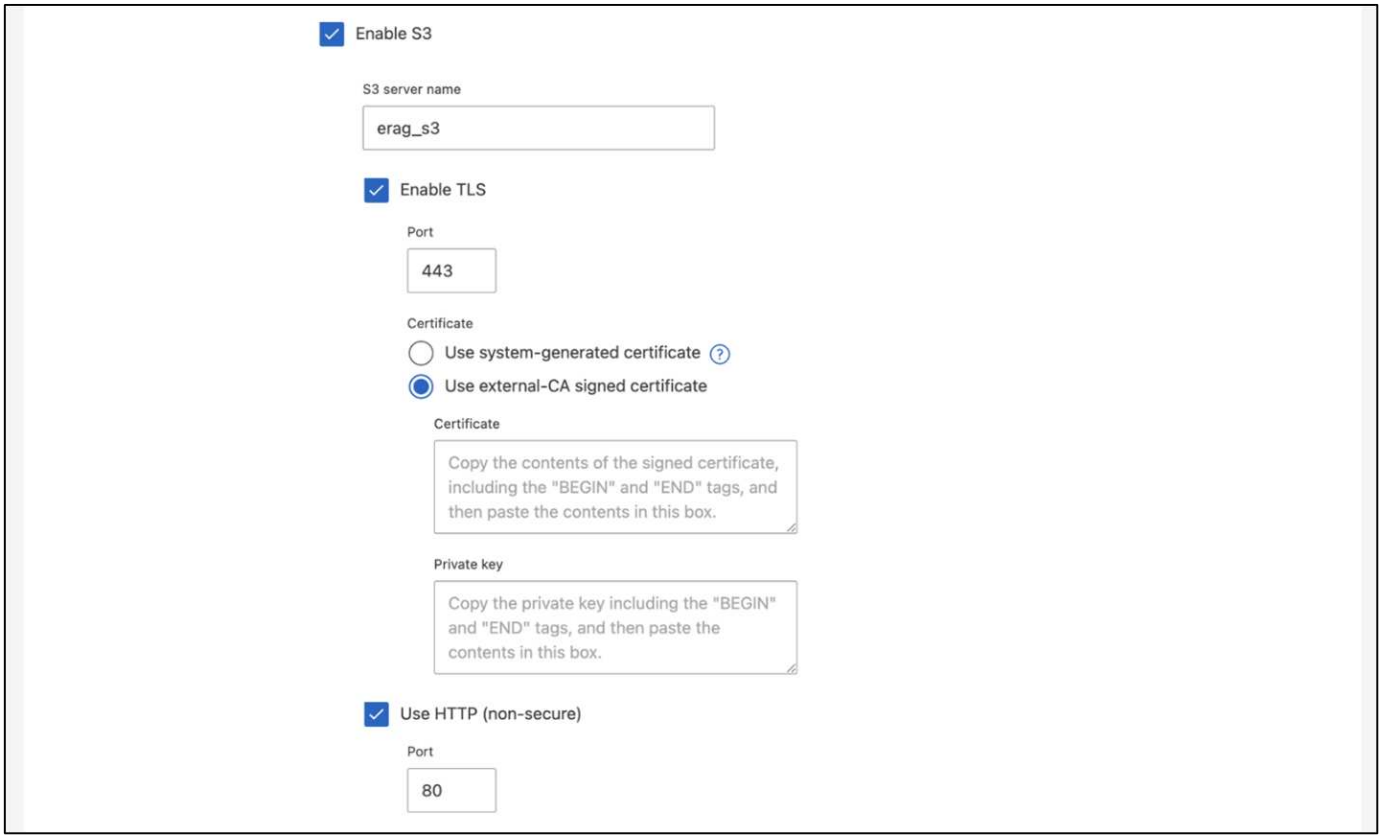
Kubernetes ノードからアクセス可能なネットワーク上で、NFS および S3 アクセス用のONTAPストレージ仮想マシン (SVM) を構成します。

ONTAP System Manager を使用して SVM を作成するには、[ストレージ] > [ストレージ VM] に移動し、[+ 追加] ボタンをクリックします。 SVM の S3 アクセスを有効にするときは、システム生成の証明書ではなく、外部 CA (証明機関) 署名付き証明書を使用するオプションを選択します。自己署名証明書または公的に信頼された CA によって署名された証明書のいずれかを使用できます。詳細については、 "ONTAP のドキュメント。"

次のスクリーンショットは、 ONTAP System Manager を使用して SVM を作成する様子を示しています。環境に応じて必要に応じて詳細を変更します。

図 4 - ONTAP System Manager を使用した SVM の作成。





S3の権限を設定する

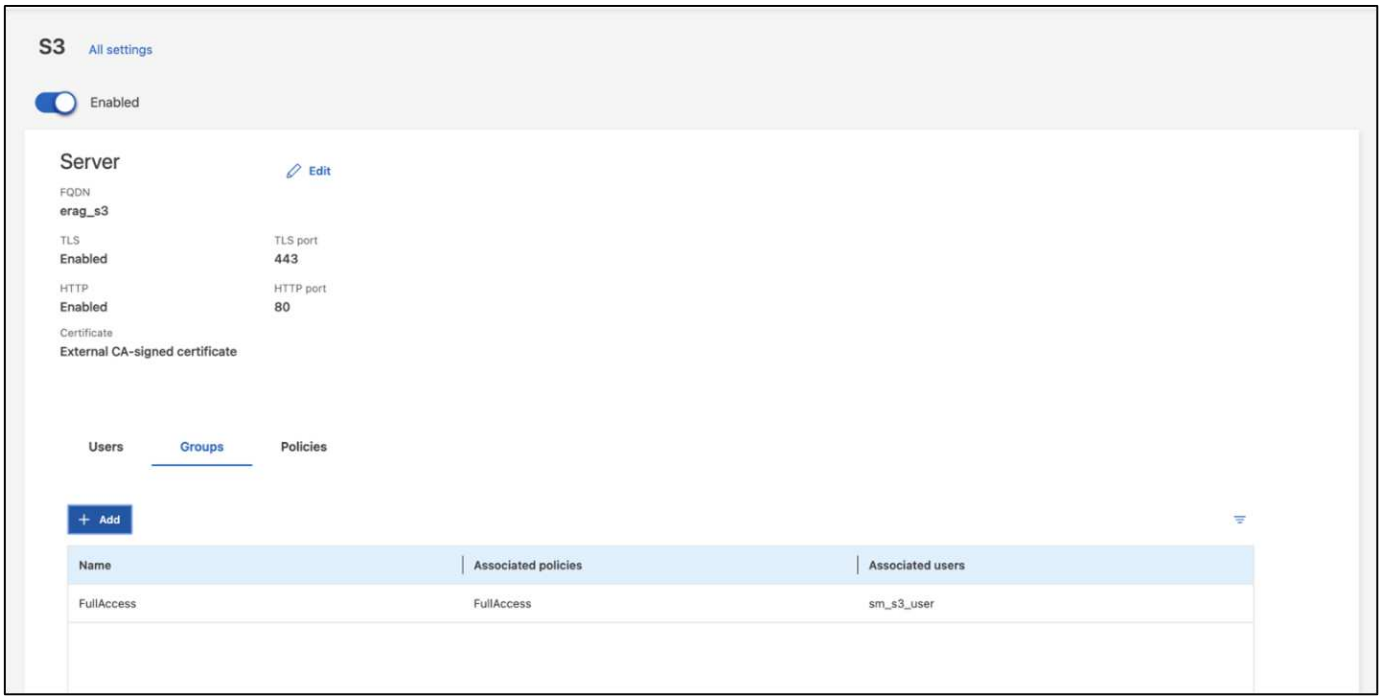
前の手順で作成した SVM の S3 ユーザー/グループ設定を構成します。その SVM のすべての S3 API 操作へのフルアクセス権を持つユーザーがいることを確認します。詳細については、 ONTAP S3 のドキュメントを参照してください。

注: このユーザーは、Intel AI for Enterprise RAG アプリケーションのデータ取り込みサービスに必要になりま

す。 ONTAP System Managerを使用してSVMを作成した場合、System Managerは自動的に次の名前のユーザーを作成します。 `sm_s3_user`そして、 `FullAccess` SVMを作成したときに権限が割り当てられていません `sm_s3_user`。

このユーザーの権限を編集するには、[ストレージ] > [ストレージ VM] に移動し、前の手順で作成した SVM の名前をクリックして、[設定] をクリックし、[S3] の横にある鉛筆アイコンをクリックします。与える `sm_s3_user`すべてのS3 API操作へのフルアクセスを付与するには、関連付ける新しいグループを作成します。 `sm_s3_user`と `FullAccess`次のスクリーンショットに示すようなポリシーです。

図 5 - S3 のアクセス許可。

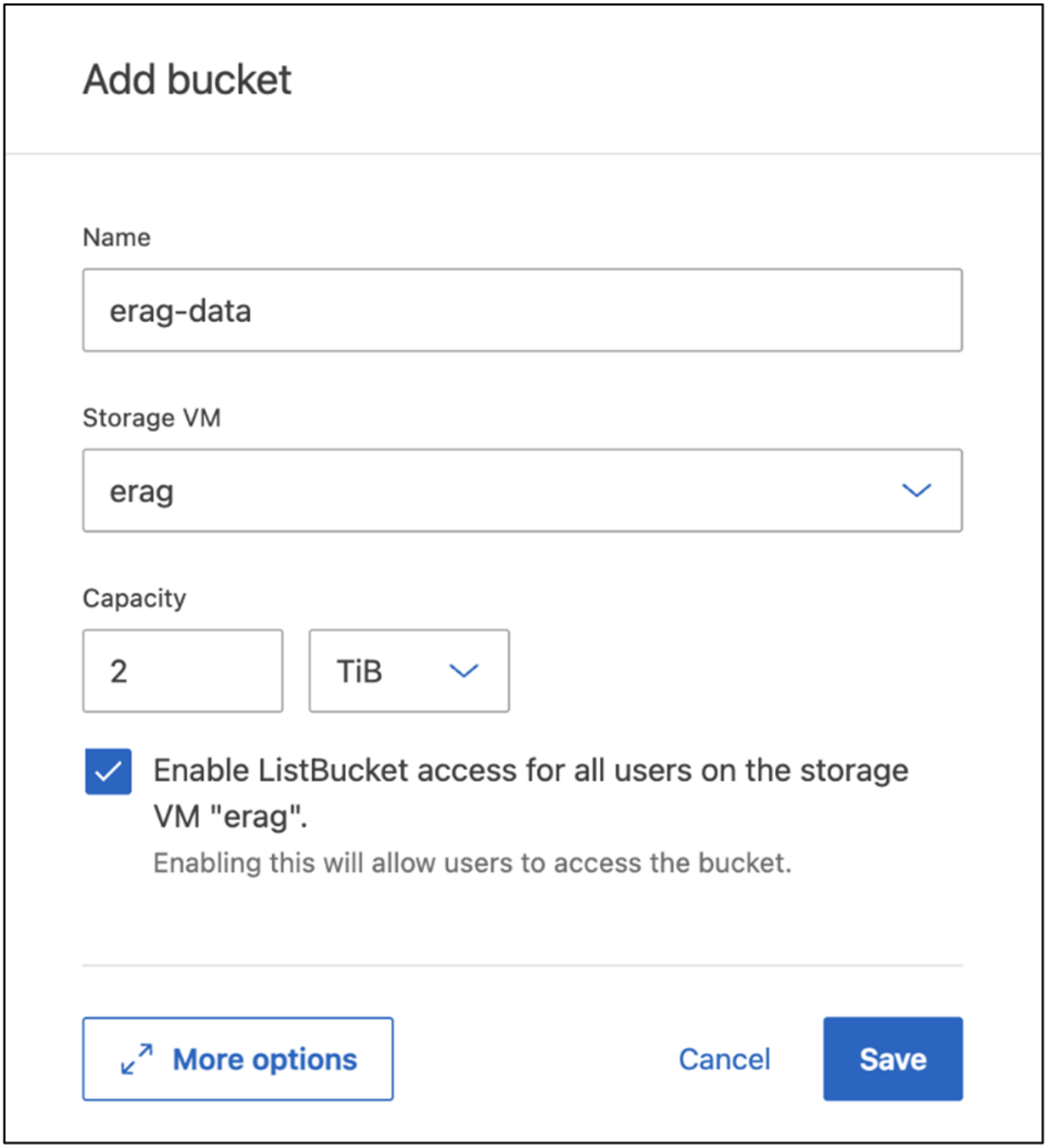


S3バケットを作成する

先ほど作成した SVM 内に S3 バケットを作成します。 ONTAP System Manager を使用して SVM を作成するには、[ストレージ] > [バケット] に移動し、[+ 追加] ボタンをクリックします。詳細については、 ONTAP S3 のドキュメントを参照してください。

次のスクリーンショットは、 ONTAP System Manager を使用して S3 バケットを作成する様子を示しています。

図 6 - S3 バケットを作成する。



S3バケットの権限を設定する

前の手順で作成した S3 バケットのアクセス許可を設定します。前の手順で設定したユーザーに次の権限があることを確認します。 `GetObject`, `PutObject`, `DeleteObject`, `ListBucket`, `GetBucketAcl`, `GetObjectAcl`, `ListBucketMultipartUploads`, `ListMultipartUploadParts`, `GetObjectTagging`, `PutObjectTagging`, `DeleteObjectTagging`, `GetBucketLocation`, `GetBucketVersioning`, `PutBucketVersioning`, `ListBucketVersions`, `GetBucketPolicy`, `PutBucketPolicy`, `DeleteBucketPolicy`, `PutLifecycleConfiguration`, `GetLifecycleConfiguration`, `GetBucketCORS`, `PutBucketCORS`.

ONTAP System Manager を使用して S3 バケットの権限を編集するには、[ストレージ] > [バケット] に移動し、バケットの名前をクリックして、[権限] をクリックし、[編集] をクリックします。参照 "ONTAP S3 ドキュメント"詳細については、こちらをご覧ください。

次のスクリーンショットは、 ONTAP System Manager で必要なバケット権限を示しています。

図 7 - S3 バケットのアクセス許可。

バケットのクロスオリジンリソース共有ルールを作成する

ONTAP CLI を使用して、前の手順で作成したバケットのバケット クロスオリジン リソース共有 (CORS) ルールを作成します。

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

このルールにより、OPEA for Intel AI for Enterprise RAG Web アプリケーションは、Web ブラウザー内からバケットと対話できるようになります。

サーバーの展開

サーバーを展開し、各サーバーに Ubuntu 22.04 LTS をインストールします。 Ubuntu をインストールしたら、すべてのサーバーに NFS ユーティリティをインストールします。 NFS ユーティリティをインストールするには、次のコマンドを実行します。

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

Kubernetesをインストールする

Kubespray を使用してサーバーに Kubernetes をインストールします。参照 "Kubespray ドキュメント"詳細については。

Trident CSIドライバをインストールする

Kubernetes クラスターにNetApp Trident CSI ドライバーをインストールします。参照 "Tridentインストールドキュメント"詳細については。

Tridentバックエンドを作成する

以前に作成した SVM のTridentバックエンドを作成します。バックエンドを作成するときは、 `ontap-nas`ドライバ。参照 "Tridentバックエンドドキュメント"詳細については。

ストレージクラスを作成する

前の手順で作成したTridentバックエンドに対応する Kubernetes ストレージ クラスを作成します。詳細については、 Tridentストレージ クラスのドキュメントを参照してください。

エンタープライズ **RAG** 向けインテル **AI** の **OPEA**

Kubernetes クラスターに Intel AI for Enterprise RAG 用の OPEA をインストールします。参照 "エンタープライズ向けインテル AI RAG 展開"詳細についてはドキュメントを参照してください。このホワイト ペーパーの後半で説明する、必要な構成ファイルの変更に注意してください。 Intel AI for Enterprise RAG アプリケーションがONTAPストレージ システムで正しく動作するためには、インストール プレイブックを実行する前にこれらの変更を行う必要があります。

ONTAP S3の使用を有効にする

Intel AI for Enterprise RAG 用の OPEA をインストールするときは、メイン構成ファイルを編集して、 ONTAP S3 をソース データ リポジトリとして使用できるようにします。

ONTAP S3の使用を有効にするには、 `edp`セクション。

注: デフォルトでは、Intel AI for Enterprise RAG アプリケーションは、SVM 内の既存のすべてのバケットからデータを取り込みます。 SVMに複数のバケットがある場合は、 `bucketNameRegexFilter`特定のバケットからのみデータが取り込まれるようにフィールドを設定します。

```
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: ".*"
```

スケジュールされた同期設定を構成する

OPEA for Intel AI for Enterprise RAGアプリケーションをインストールするときは、 `scheduledSync`アプリケーションが S3 バケットから新しいファイルや更新されたファイルを自動的に取り込むようになります。

いつ `scheduledSync`有効になっている場合、アプリケーションはソース S3 バケットで新しいファイルや更

新されたファイルを自動的にチェックします。この同期プロセスの一環として見つかった新しいファイルまたは更新されたファイルは自動的に取り込まれ、RAG ナレッジ ベースに追加されます。アプリケーションは、事前に設定された時間間隔に基づいてソース バケットをチェックします。デフォルトの時間間隔は 60 秒です。つまり、アプリケーションは 60 秒ごとに変更をチェックします。特定のニーズに合わせてこの間隔を変更する必要があるかもしれません。

有効にするには `scheduledSync`同期間隔を設定するには、次の値を設定します。`deployment/components/edp/values.yaml:`

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

ボリュームアクセスモードを変更する

で `deployment/components/gmc/microservices-connector/helm/values.yaml`、各巻ごとに `pvc`リストを変更する `accessMode`に `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
```

（オプション）SSL証明書の検証を無効にする

SVM の S3 アクセスを有効にするときに自己署名証明書を使用した場合は、SSL 証明書の検証を無効にする必要があります。公的に信頼された CA によって署名された証明書を使用した場合は、この手順をスキップできます。

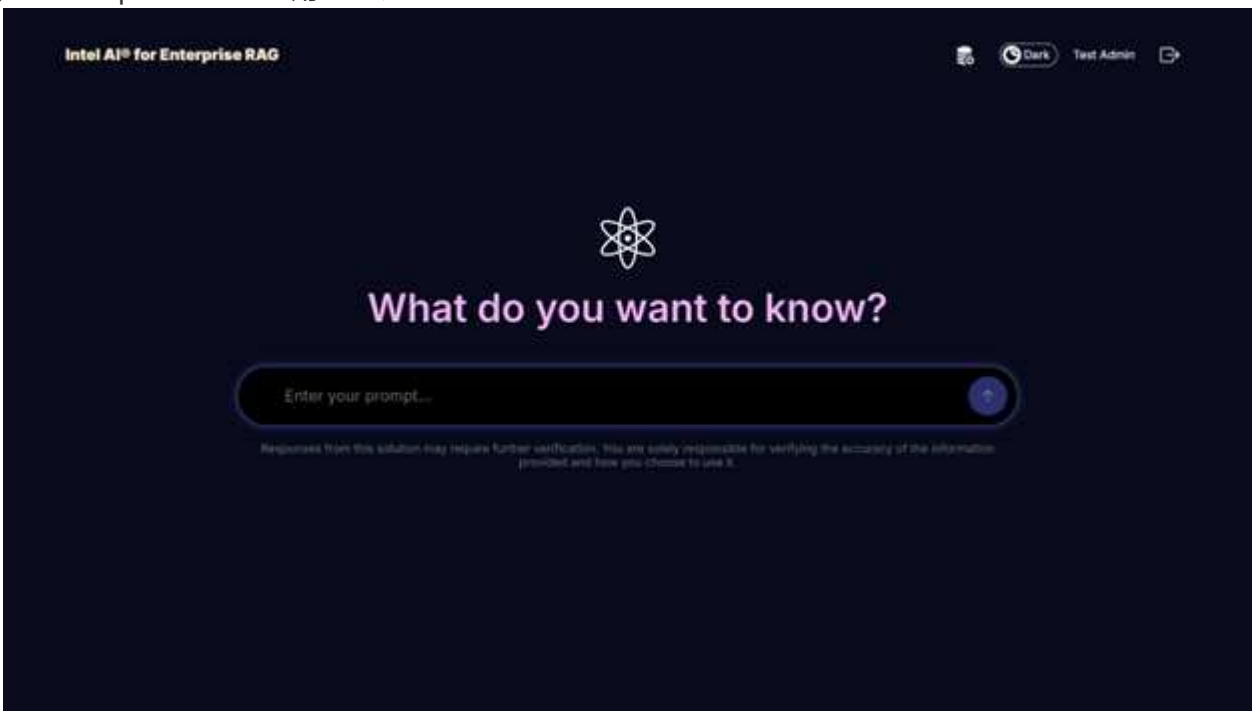
SSL証明書の検証を無効にするには、次の値を設定します。`deployment/components/edp/values.yaml:`

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

Enterprise RAG UI 向けインテル **AI** の **OPEA** にアクセスする

Intel AI for Enterprise RAG UI の OPEA にアクセスします。参照 "Intel AI for Enterprise RAG 導入ドキュメント"詳細については。

図 8 - Enterprise RAG UI 用インテル AI の OPEA



。

RAGのデータを取り込む

RAG ベースのクエリ拡張に含めるファイルを取り込むことができるようになりました。ファイルの取り込みには複数のオプションがあります。ニーズに応じて適切なオプションを選択してください。

注: ファイルが取り込まれた後、OPEA for Intel AI for Enterprise RAG アプリケーションはファイルの更新を自動的にチェックし、それに応じて更新を取り込みます。

*オプション 1: S3 バケットに直接アップロードする 一度に多くのファイルを取り込むには、任意の S3 クライアントを使用して、ファイルを S3 バケット (以前に作成したバケット) にアップロードすることをお勧めします。一般的な S3 クライアントには、AWS CLI、Amazon SDK for Python (Boto3)、s3cmd、S3 Browser、Cyberduck、Commander One などがあります。ファイルがサポートされているタイプである場合、S3 バケットにアップロードしたファイルは、OPEA for Intel AI for Enterprise RAG アプリケーションによって自動的に取り込まれます。

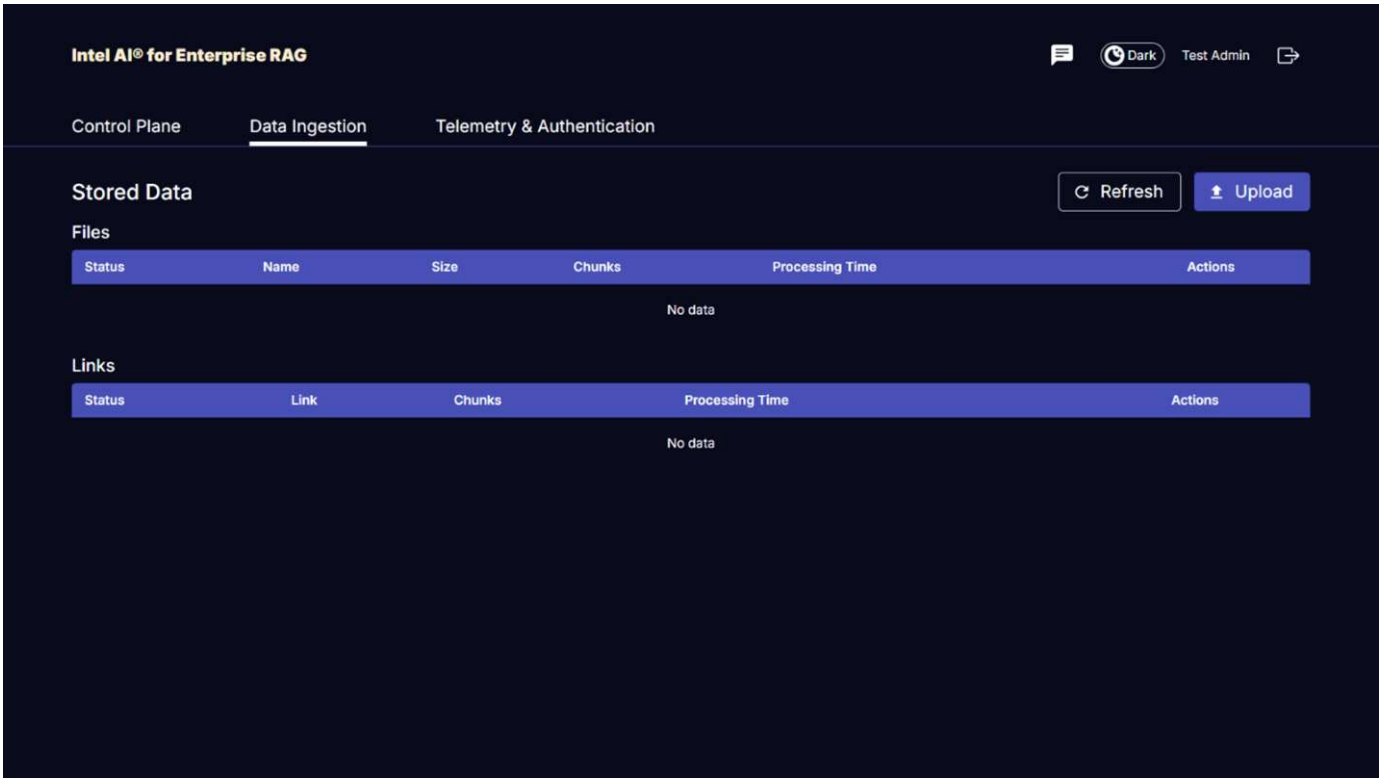
注: この文書の執筆時点では、PDF、HTML、TXT、DOC、DOCX、PPT、PPTX、MD、XML、JSON

、JSONL、YAML、XLS、XLSX、CSV、TIFF、JPG、JPEG、PNG、SVG のファイル タイプがサポートされています。

OPEA for Intel AI for Enterprise RAG UI を使用して、ファイルが適切に取り込まれたことを確認できます。詳細については、Intel AI for Enterprise RAG UI のドキュメントを参照してください。アプリケーションが大量のファイルを取り込むには時間がかかる場合があることに注意してください。

*オプション 2: UI を使用してアップロードする 少数のファイルのみを取り込む必要がある場合は、OPEA for Intel AI for Enterprise RAG UI を使用して取り込むことができます。詳細については、Intel AI for Enterprise RAG UI のドキュメントを参照してください。

図 9 - データ取り込み UI
。



チャットクエリを実行する

付属のチャット UI を使用して、OPEA for Intel AI for Enterprise RAG アプリケーションと「チャット」できるようになりました。クエリに応答する際、アプリケーションは取り込んだファイルを使用して RAG を実行します。つまり、アプリケーションは取り込んだファイル内の関連情報を自動的に検索し、クエリに応答するときにこの情報を組み込みます。

サイズガイド

検証作業の一環として、Intel と連携してパフォーマンス テストを実施しました。このテストの結果、次の表に示すサイズ設定のガイドラインが得られました。

特徴づけ	**Value**	コメント
モデルサイズ	200億のパラメータ	ラマ-8B、ラマ-13B、ミストラル7B、クウェン14B、ディープシーク・ディスティル8B
入力サイズ	約2,000トークン	約4ページ
出力サイズ	約2,000トークン	約4ページ
同時ユーザー数	32	「同時ユーザー」とは、同時にクエリを送信しているプロンプト要求を指します。

_注: 上記のサイズ設定ガイダンスは、96 コアの Intel Xeon 6 プロセッサを使用して収集されたパフォーマンス検証およびテスト結果に基づいています。同様の I/O トークンとモデル サイズ要件を持つお客様には、96 または 128 コアの Xeon 6 プロセッサを搭載したサーバーの使用をお勧めします。

まとめ

エンタープライズ RAG システムと LLM は連携して動作し、組織が正確でコンテキストに応じた応答を提供できるようにするテクノロジーです。これらの対応には、膨大な個人データと社内企業データに基づく情報検索が含まれます。 RAG、API、ベクトル埋め込み、高性能ストレージ システムを使用して、企業データを含むドキュメント リポジトリをクエリすることで、データはより高速かつ安全に処理されます。 NetApp AIPod Mini は、NetApp のインテリジェント データ インフラストラクチャとONTAPデータ管理機能、Intel Xeon 6 プロセッサ、Intel AI for Enterprise RAG、OPEA ソフトウェア スタックを組み合わせ、高性能 RAG アプリケーションの導入を支援し、組織を AI リーダーシップへと導きます。

了承

このドキュメントは、 NetAppソリューション エンジニアリング チームのメンバーである Sathish Thyagarajan と Michael Ogelsby が作成したものです。著者らはまた、NetAppの他のチーム メンバー (Lawrence Bunka、Bobby Oommen、Jeff Liborio) にも感謝の意を表します。

部品表

以下は、このソリューションの機能検証に使用された BOM であり、参照として使用できます。次の構成に適合する任意のサーバーまたはネットワーク コンポーネント (または、100GbE 帯域幅が望ましい既存のネットワーク) を使用できます。

アプリサーバーの場合:

部品番号	製品説明	量
222HA-TN-OTO-37	ハイパースーパーサーバー SYS-222HA-TN /2U	2
P4X-GNR6980P-SRPL2-UCC	インテル Xeon 6980P 2P 128C 2G 504M 500W SGX512	4
RAM	MEM-DR564MC-ER64(x16)64GB DDR5-6400 2RX4 (16Gb) ECC RDIMM	32

部品番号	製品説明	量
	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 冗長シングル出力電源。熱放散は 2361 BTU/時、最高温度は 59 °C（約）	4

制御サーバーの場合:

部品番号	製品説明	量
511R-M-OTO-17	1U X13SCH-SYS、CSE-813MF2TS-R0RCNBP、PWS-602A-1Rまで最適化	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

ネットワーク スイッチの場合:

部品番号	製品説明	量
DCS-7280CR3A	アリスタ 7280R3A 28x100 GbE	1

NetApp AFFストレージ:

部品番号	製品説明	量
AFF-A20A-100-C	AFF A20 HAシステム、-C	1
X800-42U-R6-C	ジャンパーカード、インキャブ、C13-C14、-C	2
X97602A-C	電源、1600W、チタン、-C	2
X66211B-2-NC	ケーブル、100GbE、QSFP28-QSFP28、Cu、2m、-C	4
X66240A-05-NC	ケーブル、25GbE、SFP28-SFP28、Cu、0.5m、-C	2
X5532A-NC	レール、4ポスト、薄型、Rnd/Sq-Hole、Sm、Adj、24-32、-C	1
X4024A-2-AC	ドライブ パック 2X1.92TB、NVMe4、SED、-C	6

X60130A-C	IO モジュール、2PT、100GbE、-C	2
X60132A-C	IOモジュール、4PT、10/25GbE、-C	2
SW-ONTAPB-FLASH-A20-C	SW、 ONTAPベース パッケージ、TB 単位、フラッシュ、A20、-C	23

詳細情報の入手方法

このドキュメントに記載されている情報の詳細については、次のドキュメントや Web サイトを参照してください。

"NetApp製品ドキュメント"

"OPEAプロジェクト"

"OPEA エンタープライズ RAG 導入プレイブック"

著作権に関する情報

Copyright © 2025 NetApp, Inc. All Rights Reserved. Printed in the U.S.このドキュメントは著作権によって保護されています。著作権所有者の書面による事前承諾がある場合を除き、画像媒体、電子媒体、および写真複写、記録媒体、テープ媒体、電子検索システムへの組み込みを含む機械媒体など、いかなる形式および方法による複製も禁止します。

ネットアップの著作物から派生したソフトウェアは、次に示す使用許諾条項および免責条項の対象となります。

このソフトウェアは、ネットアップによって「現状のまま」提供されています。ネットアップは明示的な保証、または商品性および特定目的に対する適合性の暗示的保証を含み、かつこれに限定されないいかなる暗示的な保証も行いません。ネットアップは、代替品または代替サービスの調達、使用不能、データ損失、利益損失、業務中断を含み、かつこれに限定されない、このソフトウェアの使用により生じたすべての直接的損害、間接的損害、偶発的損害、特別損害、懲罰的損害、必然的損害の発生に対して、損失の発生の可能性が通知されていたとしても、その発生理由、根拠とする責任論、契約の有無、厳格責任、不法行為（過失またはそうでない場合を含む）にかかわらず、一切の責任を負いません。

ネットアップは、ここに記載されているすべての製品に対する変更を随時、予告なく行う権利を保有します。ネットアップによる明示的な書面による合意がある場合を除き、ここに記載されている製品の使用により生じる責任および義務に対して、ネットアップは責任を負いません。この製品の使用または購入は、ネットアップの特許権、商標権、または他の知的所有権に基づくライセンスの供与とはみなされません。

このマニュアルに記載されている製品は、1つ以上の米国特許、その他の国の特許、および出願中の特許によって保護されている場合があります。

権利の制限について：政府による使用、複製、開示は、DFARS 252.227-7013（2014年2月）およびFAR 5252.227-19（2007年12月）のRights in Technical Data -Noncommercial Items（技術データ - 非商用品目に関する諸権利）条項の(b)(3)項、に規定された制限が適用されます。

本書に含まれるデータは商用製品および / または商用サービス（FAR 2.101の定義に基づく）に関係し、データの所有権はNetApp, Inc.にあります。本契約に基づき提供されるすべてのネットアップの技術データおよびコンピュータ ソフトウェアは、商用目的であり、私費のみで開発されたものです。米国政府は本データに対し、非独占的かつ移転およびサブライセンス不可で、全世界を対象とする取り消し不能の制限付き使用権を有し、本データの提供の根拠となった米国政府契約に関連し、当該契約の裏付けとする場合にのみ本データを使用できます。前述の場合を除き、NetApp, Inc.の書面による許可を事前に得ることなく、本データを使用、開示、転載、改変するほか、上演または展示することはできません。国防総省にかかる米国政府のデータ使用権については、DFARS 252.227-7015(b)項（2014年2月）で定められた権利のみが認められます。

商標に関する情報

NetApp、NetAppのロゴ、http://www.netapp.com/TMに記載されているマークは、NetApp, Inc.の商標です。その他の会社名と製品名は、それを所有する各社の商標である場合があります。