NetApp AIPod Mini - NetApp 및 Intel을 사용한 엔터프라이즈 RAG 추론

NetApp artificial intelligence solutions

NetApp
February 12, 2026

목차

NetApp AIPod Mini - NetApp 및 Intel을 사용한 엔터프라이즈 RAG 추론

본 논문에서는 Intel Xeon 6 프로세서와 NetApp 데이터 관리 솔루션의 기술과 기능을 결합한 Enterprise RAG용 NetApp AIPod 의 검증된 참조 설계를 제시합니다. 이 솔루션은 대규모 언어 모델을 활용하여 동시 사용자에게 정확하고 상황에 맞는 응답을 제공하는 다운스트림 ChatQnA 애플리케이션을 보여줍니다. 응답은 공기 간격이 있는 RAG 추론 파이프라인을 통해 조직의 내부 지식 저장소에서 검색됩니다.

Sathish Thyagarajan, Michael Oglesby, Arpita Mahajan, NetApp

요약

점점 더 많은 조직들이 생산성과 비즈니스 가치를 높이기 위해 검색 증강 생성(RAG) 애플리케이션과 대규모 언어 모델(LLM)을 활용하여 사용자 프롬프트를 해석하고 응답을 생성하고 있습니다. 이러한 프롬프트와 응답에는 텍스트, 코드, 이미지, 심지어 조직의 내부 지식 기반, 데이터 레이크, 코드 저장소, 문서 저장소에서 검색된 치료용 단백질 구조까지 포함될 수 있습니다. 본 문서에서는 NetApp AFF 스토리지와 Intel Xeon 6 프로세서가 탑재된 서버로 구성된 NetApp AIPod Mini 솔루션의 참조 설계를 다룹니다. 이 솔루션에는 Intel Advanced Matrix Extensions(Intel AMX)와 결합된 NetApp ONTAP 데이터 관리 소프트웨어, 그리고 Open Platform for Enterprise AI(OPEA) 기반의 Intel® AI for Enterprise RAG 소프트웨어가 포함됩니다. 엔터프라이즈 RAG용 NetApp AIPod Mini를 통해 조직은 퍼블릭 LLM을 프라이빗 생성형 AI(GenAI) 추론 솔루션으로 확장할 수 있습니다. 이 솔루션은 기업 규모에서 효율적이고 비용 효율적인 RAG 추론을 제공하며, 신뢰성을 향상시키고 독점 정보에 대한 더 나은 제어 권한을 제공하도록 설계되었습니다.

인텔 스토리지 파트너 검증

Intel Xeon 6 프로세서로 구동되는 서버는 최대 성능을 위해 Intel AMX를 사용하여 까다로운 AI 추론 워크로드를 처리하도록 제작되었습니다. 최적의 스토리지 성능과 확장성을 구현하기 위해 NetApp ONTAP 사용하여 솔루션의 검증이 성공적으로 완료되었으며, 이를 통해 기업은 RAG 애플리케이션의 요구 사항을 충족할 수 있습니다. 이 검증은 Intel Xeon 6 프로세서가 장착된 서버에서 수행되었습니다. Intel과 NetApp 최적화되고 확장 가능하며 고객 비즈니스 요구 사항에 맞는 AI 솔루션을 제공하는 데 중점을 둔 강력한 파트너십을 맺고 있습니다.

NetApp 으로 RAG 시스템을 실행하는 이점

RAG 애플리케이션은 PDF, 텍스트, CSV 또는 Excel과 같은 다양한 형식의 기업 문서 저장소에서 지식을 검색하는 작업을 포함합니다. 이러한 데이터는 일반적으로 S3 객체 스토리지 또는 온프레미스 NFS와 같은 솔루션에 데이터 소스로 저장됩니다. NetApp은 엣지, 데이터 센터 및 클라우드 생태계 전반에 걸쳐 데이터 관리, 데이터 이동성, 데이터 거버넌스 및 데이터 보안 기술 분야의 선두 주자입니다. NetApp ONTAP 데이터 관리는 배치 및 실시간 추론을 포함한 다양한 유형의 AI 워크로드를 지원하는 엔터프라이즈급 스토리지를 제공하며 다음과 같은 이점을 제공합니다:

- 속도와 확장성. 독립적으로 성능과 용량을 확장할 수 있는 기능을 통해 버전 관리를 위해 대용량 데이터 세트를 고속으로 처리할 수 있습니다.
- 데이터 접근. 다중 프로토콜 지원을 통해 클라이언트 애플리케이션은 S3, NFS, SMB 파일 공유 프로토콜을 사용하여 데이터를 읽을 수 있습니다. ONTAP S3 NAS 버킷은 다중 모드 LLM 추론 시나리오에서 데이터 액세스를

용이하게 할 수 있습니다.

- 신뢰성과 기밀성. ONTAP 데이터 보호, 내장형 NetApp 자율형 랜섬웨어 보호(ARP), 동적 스토리지 프로비저닝을 제공하며, 소프트웨어 및 하드웨어 기반 암호화를 모두 제공하여 기밀성과 보안을 강화합니다. ONTAP 은 모든 SSL 연결에 대해 FIPS 140-2를 준수합니다.

타겟 고객층

이 문서는 엔터프라이즈 RAG 및 GenAI 솔루션을 제공하기 위해 구축된 인프라를 활용하고자 하는 AI 의사결정권자, 데이터 엔지니어, 비즈니스 리더 및 부서 임원을 대상으로 합니다. AI 추론, LLM, Kubernetes, 네트워킹 및 구성 요소에 대한 사전 지식이 있으면 구현 단계에 도움이 됩니다.

기술 요구 사항

하드웨어

Intel® AI 기술

호스트 CPU로 Xeon 6을 사용하면 가속 시스템은 높은 단일 스레드 성능, 더 높은 메모리 대역폭, 향상된 안정성, 가용성, 서비스 용이성(RAS) 및 더 많은 I/O 레인의 이점을 누릴 수 있습니다. Intel AMX는 INT8 및 BF16에 대한 추론을 가속화하고 FP16으로 훈련된 모델을 지원하며, INT8의 경우 코어당 사이클당 최대 2,048개의 부동 소수점 연산, BF16/FP16의 경우 코어당 사이클당 최대 1,024개의 부동 소수점 연산을 지원합니다. Xeon 6 프로세서를 사용하여 RAG 솔루션을 배포하려면 일반적으로 최소 250GB의 RAM과 500GB의 디스크 공간이 권장됩니다. 하지만 이는 LLM 모델 크기에 크게 좌우됩니다. 자세한 내용은 Intel을 참조하세요. "제온 6 프로세서" 제품 개요.

그림 1 - Intel Xeon 6 프로세서가 탑재된 컴퓨팅

서버

NetApp AFF 스토리지

보급형 및 중급형 NetApp AFF A-Series 시스템은 더욱 강력한 성능, 밀도, 그리고 더 높은 효율성을 제공합니다. NetApp AFF A20, AFF A30 및 AFF A50 시스템은 단일 OS를 기반으로 블록, 파일 및 객체를 지원하는 진정한 통합 스토리지를 제공하여 하이브리드 클라우드 전반에서 가장 낮은 비용으로 RAG 애플리케이션의 데이터를 원활하게 관리, 보호 및 모바일화할 수 있습니다.

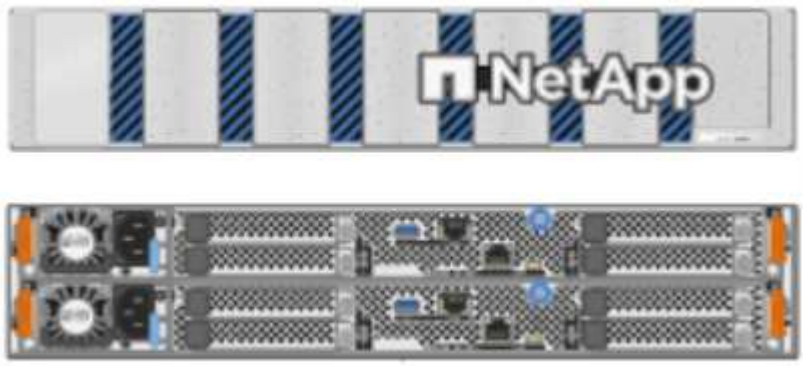
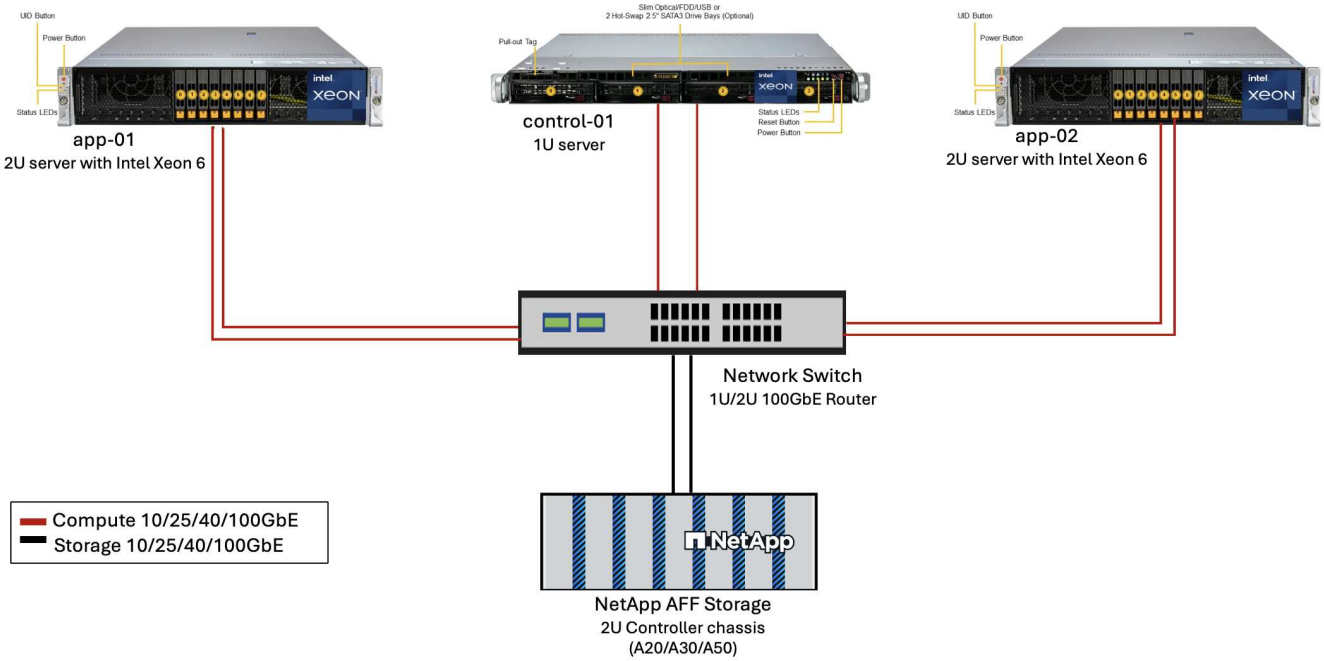


그림 2 - NetApp AFF A-시리즈 시스템.

하드웨어	수량	논평
Intel Xeon 6세대(Granite Rapids)	2	RAG 추론 노드—듀얼 소켓 Intel Xeon 6900-series(96코어) 또는 Intel Xeon 6700-series(64코어) 프로세서와 DDR5(6400MHz) 또는 MRDIMM(8800MHz)를 사용하는 250GB~3TB RAM. 2U 서버.
Intel 프로세서가 탑재된 제어 평면 서버	1	쿠버네티스 제어 평면/1U 서버.
100Gb 이더넷 스위치 선택	1	데이터 센터 스위치.
NetApp AFF A20(또는 AFF A30; AFF A50)	1	최대 저장 용량: 9.3PB. 참고: 네트워킹: 10/25/100GbE 포트.

이 참조 설계의 검증을 위해 Supermicro(222HA-TN-OTO-37)의 Intel Xeon 6 프로세서가 장착된 서버와 Arista(7280R3A)의 100GbE 스위치가 사용되었습니다.

그림 3 - AIPod Mini 배포 아키텍처



소프트웨어

엔터프라이즈 **AI**를 위한 오픈 플랫폼

OPEA(Open Platform for Enterprise AI)는 인텔이 생태계 파트너와 협력하여 주도하는 오픈 소스 이니셔티브입니다. RAG에 중점을 두고 최첨단 생성적 AI 시스템 개발을 가속화하도록 설계된 구성 가능한 구성 요소의 모듈식 플랫폼을 제공합니다. OPEA에는 LLM, 데이터 저장소, 프롬프트 엔진, RAG 아키텍처 청사진, 성능, 기능, 신뢰성, 기업 준비 상태를 기준으로 생성 AI 시스템을 평가하는 4단계 평가 방법 등을 갖춘 포괄적인 프레임워크가 포함되어 있습니다.

OPEA의 핵심은 두 가지 핵심 구성 요소로 구성됩니다.

- GenAIComps: 마이크로서비스 구성 요소로 구성된 서비스 기반 툴킷
- GenAIExamples: 실제 사용 사례를 보여주는 ChatQnA와 같은 즉시 배포 가능한 솔루션

자세한 내용은 다음을 참조하세요. "OPEA 프로젝트 문서"

OPEA 기반 Intel® AI for Enterprise RAG

OPEA for Intel® AI for Enterprise RAG는 기업 데이터를 실행 가능한 인사이트로 변환하는 과정을 간소화합니다. Intel Xeon 프로세서 기반의 OPEA는 업계 파트너사의 구성 요소를 통합하여 기업 솔루션 배포를 위한 효율적인 접근 방식을 제공합니다. 검증된 오케스트레이션 프레임워크를 통해 원활하게 확장되며, 기업에 필요한 유연성과 선택권을 제공합니다.

OPEA를 기반으로 구축된 Intel® AI for Enterprise RAG는 확장성, 보안 및 사용자 경험을 향상시키는 주요 기능을 추가하여 이 기반을 확장합니다. 이러한 기능에는 최신 서비스 기반 아키텍처와의 원활한 통합을 위한 서비스 메시 기능, 파이프라인 안정성을 위한 프로덕션 환경 수준의 검증 기능, 워크플로우의 손쉬운 관리 및 모니터링을 지원하는 서비스형 RAG를 위한 풍부한 기능을 갖춘 UI가 포함됩니다. 또한 Intel 및 파트너 지원을 통해 광범위한 솔루션 에코시스템을 이용할 수 있으며, 통합된 ID 및 액세스 관리(IAM) 기능과 UI 및 애플리케이션을 통해 안전하고 규정을 준수하는 운영을 지원합니다. 프로그래밍 가능한 가드레일을 통해 파이프라인 동작을 세밀하게 제어하고 맞춤형 보안 및 규정 준수 설정을 구현할 수 있습니다.

NetApp ONTAP

NetApp ONTAP 은 NetApp의 중요 데이터 스토리지 솔루션을 뒷받침하는 기반 기술입니다. ONTAP 에는 사이버 공격에 대한 자동 랜섬웨어 보호, 내장형 데이터 전송 기능, 스토리지 효율성 기능 등 다양한 데이터 관리 및 데이터 보호 기능이 포함되어 있습니다. 이러한 이점은 LLM 배포를 위한 NAS, SAN, 객체 및 소프트웨어 정의 스토리지의 온프레미스부터 하이브리드 멀티클라우드까지 다양한 아키텍처에 적용됩니다. ONTAP 클러스터에서 ONTAP S3 개체 스토리지 서버를 사용하면 RAG 애플리케이션을 배포하고, 권한이 있는 사용자와 클라이언트 애플리케이션을 통해 제공되는 ONTAP 의 스토리지 효율성과 보안을 활용할 수 있습니다. 자세한 내용은 다음을 참조하세요. "ONTAP S3 구성에 대해 알아보세요"

NetApp Trident

NetApp Trident 소프트웨어는 Red Hat OpenShift를 포함하여 컨테이너와 Kubernetes 배포판을 위한 오픈 소스이자 완벽하게 지원되는 스토리지 오케스트레이터입니다. Trident NetApp ONTAP 포함한 전체 NetApp 스토리지 포트폴리오와 호환되며 NFS 및 iSCSI 연결도 지원합니다. 자세한 내용은 다음을 참조하세요. "Git에서 NetApp Trident"

소프트웨어	버전	논평
OPEA - 엔터프라이즈 RAG용 Intel® AI	2.0	OPEA 마이크로서비스 기반 엔터프라이즈 RAG 플랫폼
컨테이너 스토리지 인터페이스(CSI 드라이버)	NetApp Trident 25.10	동적 프로비저닝, NetApp 스냅샷 복사본 및 볼륨을 활성화합니다.
우분투	22.04.5	2노드 클러스터의 OS.
컨테이너 오케스트레이션	Kubernetes 1.31.9(Enterprise RAG 인프라 플레이북으로 설치됨)	RAG 프레임워크를 실행하기 위한 환경
ONTAP	ONTAP 9.16.1P4 이상	AFF A20의 스토리지 OS.

솔루션 구축

소프트웨어 스택

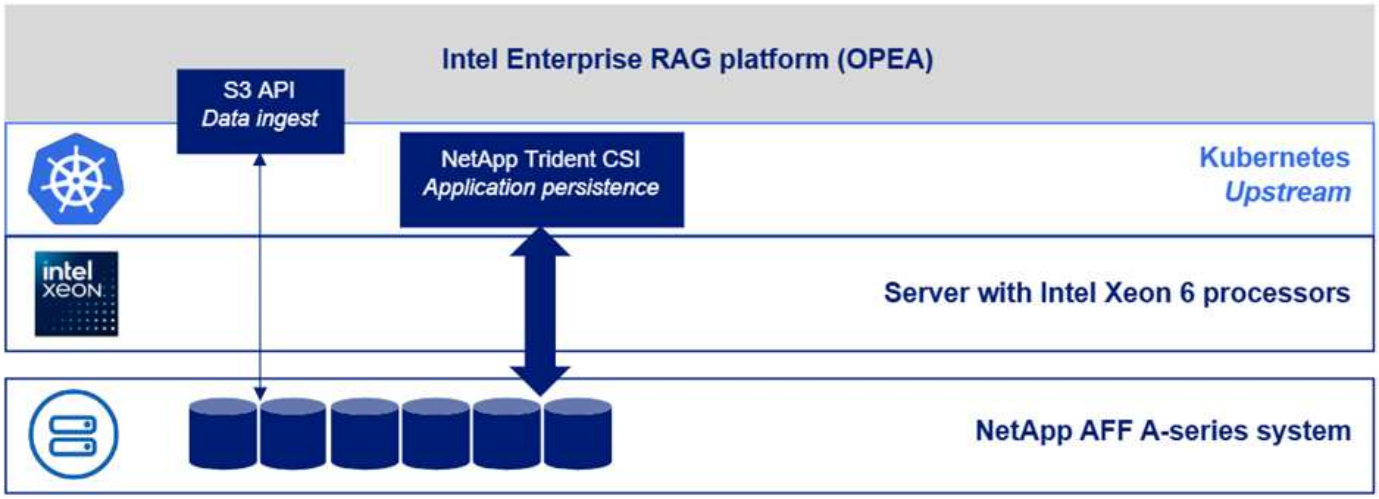
이 솔루션은 Intel Xeon 기반 앱 노드로 구성된 Kubernetes 클러스터에 배포됩니다. Kubernetes 제어 평면의 기본적인 고가용성을 구현하려면 최소 3개의 노드가 필요합니다. 다음 클러스터 레이아웃을 사용하여 솔루션을 검증했습니다.

표 3 - 쿠버네티스 클러스터 레이아웃

마디	역할	수량
Intel Xeon 6 프로세서와 1TB RAM을 탑재한 서버	앱 노드, 제어 평면 노드	2
일반 서버	제어 평면 노드	1

다음 그림은 솔루션의 "소프트웨어 스택 뷰"를 보여줍니다

.



배포 단계

ONTAP 스토리지 어플라이언스 배포

NetApp ONTAP 스토리지 어플라이언스를 배포하고 프로비저닝합니다. 를 참조하세요 "ONTAP 하드웨어 시스템 문서" 자세한 내용은.

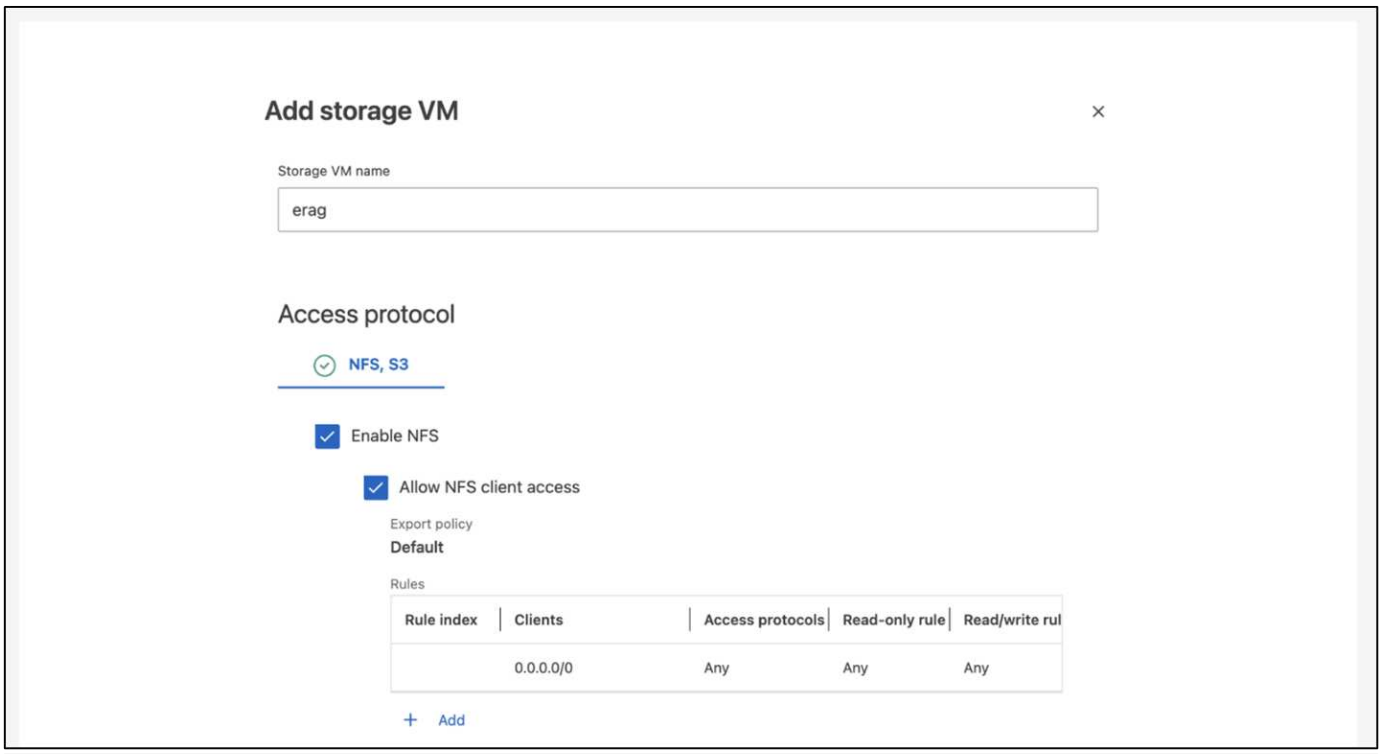
NFS 및 **S3** 액세스를 위한 **ONTAP SVM** 구성

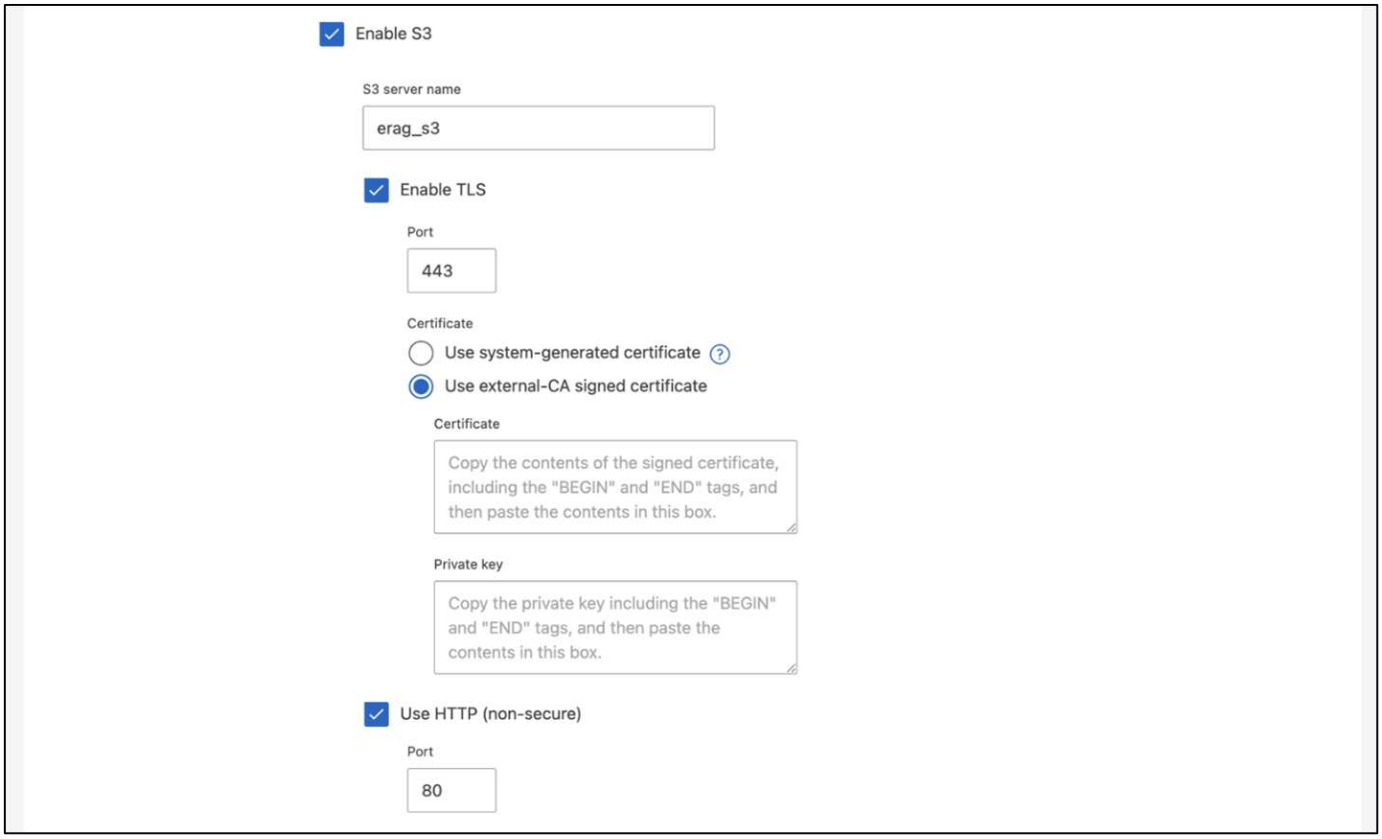
Kubernetes 노드에서 액세스할 수 있는 네트워크에서 NFS 및 S3 액세스를 위한 ONTAP 스토리지 가상 머신(SVM)을 구성합니다.

ONTAP System Manager를 사용하여 SVM을 생성하려면 스토리지 > 스토리지 VM으로 이동한 다음 + 추가 버튼을 클릭합니다. SVM에 대한 S3 액세스를 활성화할 때 시스템에서 생성된 인증서가 아닌 외부 CA(인증 기관) 서명 인증서를 사용하는 옵션을 선택하세요. 자체 서명된 인증서나 공개적으로 신뢰할 수 있는 CA에서 서명한 인증서를 사용할 수 있습니다. 추가 세부 사항은 다음을 참조하세요. "ONTAP 문서."

다음 스크린샷은 ONTAP 시스템 관리자를 사용하여 SVM을 만드는 방법을 보여줍니다. 사용자의 환경에 맞게 세부 정보를 수정하세요.

그림 5 - ONTAP System Manager를 사용한 SVM 생성
.





S3 권한 구성

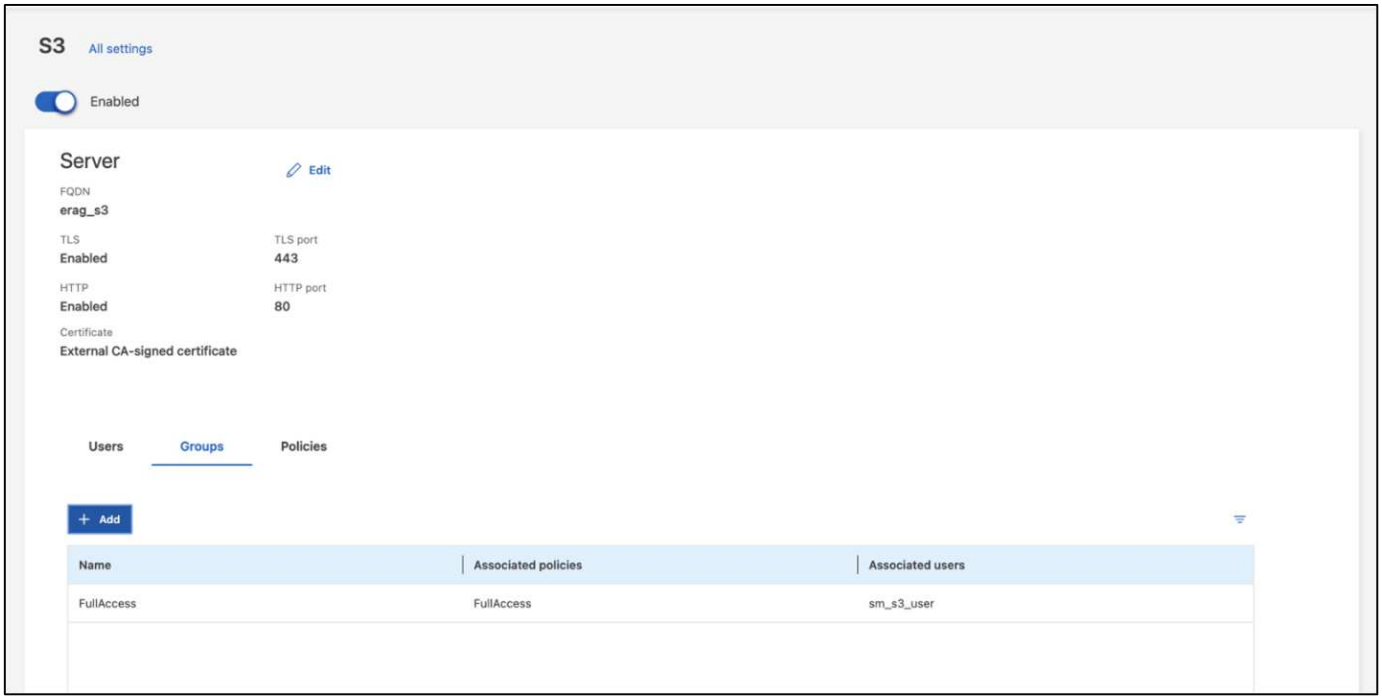
이전 단계에서 생성한 SVM에 대한 S3 사용자/그룹 설정을 구성합니다. 해당 SVM에 대한 모든 S3 API 작업에 대한 전체 액세스 권한이 있는 사용자가 있는지 확인하세요. 자세한 내용은 ONTAP S3 문서를 참조하세요.

참고: 이 사용자는 Intel® AI for Enterprise RAG 애플리케이션의 데이터 수집 서비스에 필요합니다. ONTAP System Manager를 사용하여 SVM을 생성한 경우 System Manager는 SVM 생성 시 `sm_s3_user`라는 사용자와

`FullAccess`라는 정책을 자동으로 생성하지만 `sm_s3_user`에는 권한이 할당되지 않습니다.

이 사용자의 권한을 편집하려면 저장소 > 저장소 VM으로 이동한 후 이전 단계에서 만든 SVM의 이름을 클릭하고 설정을 클릭한 다음 "S3" 옆에 있는 연필 아이콘을 클릭합니다. 주다 `sm_s3_user` 모든 S3 API 작업에 대한 전체 액세스, 연결하는 새 그룹 생성 `sm_s3_user` 와 함께 `FullAccess` 다음 스크린샷에 표시된 대로 정책입니다.

그림 6 - S3 권한

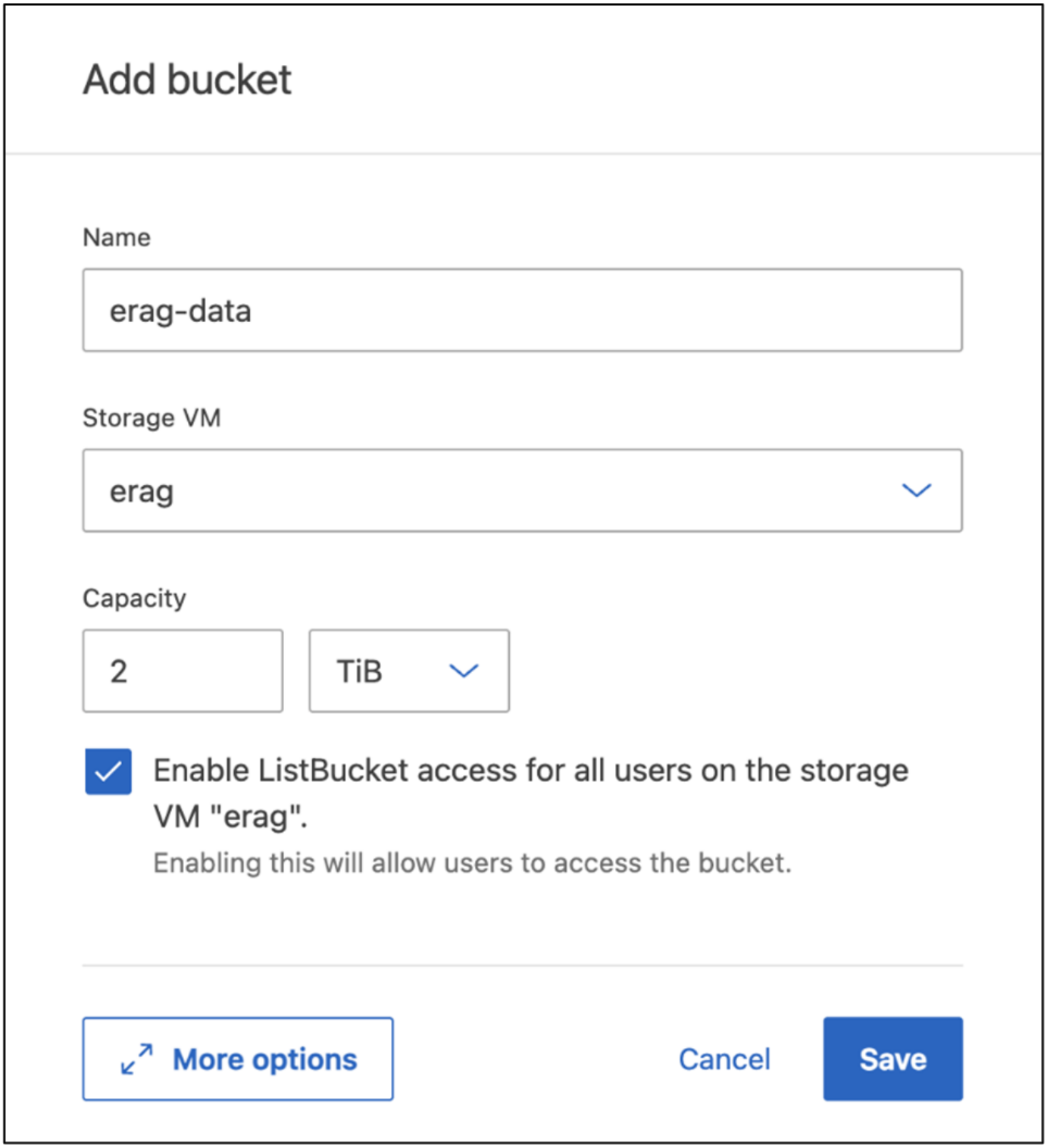


S3 버킷 생성

이전에 만든 SVM 내에 S3 버킷을 만듭니다. ONTAP 시스템 관리자를 사용하여 SVM을 생성하려면 스토리지 > 버킷으로 이동한 다음 + 추가 버튼을 클릭합니다. 자세한 내용은 ONTAP S3 설명서를 참조하세요.

다음 스크린샷은 ONTAP System Manager를 사용하여 S3 버킷을 생성하는 방법을 보여줍니다.

그림 7 - S3 버킷
생성



S3 버킷 권한 구성

이전 단계에서 생성한 S3 버킷에 대한 권한을 구성합니다. 이전 단계에서 구성한 사용자에게 다음 권한이 있는지 확인하세요. `GetObject`, `PutObject`, `DeleteObject`, `ListBucket`, `GetBucketAcl`, `GetObjectAcl`, `ListBucketMultipartUploads`, `ListMultipartUploadParts`, `GetObjectTagging`, `PutObjectTagging`, `DeleteObjectTagging`, `GetBucketLocation`, `GetBucketVersioning`, `PutBucketVersioning`, `ListBucketVersions`, `GetBucketPolicy`, `PutBucketPolicy`, `DeleteBucketPolicy`, `PutLifecycleConfiguration`, `GetLifecycleConfiguration`, `GetBucketCORS`, `PutBucketCORS`.

ONTAP System Manager를 사용하여 S3 버킷 권한을 편집하려면 저장소 > 버킷으로 이동한 후 버킷 이름을 클릭하고 권한을 클릭한 다음 편집을 클릭합니다. 를 참조하세요 "ONTAP S3 문서" 추가 세부 사항은 다음을 참조하세요.

다음 스크린샷은 ONTAP 시스템 관리자에서 필요한 버킷 권한을 보여줍니다.

그림 8 - S3 버킷 권한

버킷 교차 출처 리소스 공유 규칙 생성

ONTAP CLI를 사용하여 이전 단계에서 만든 버킷에 대한 버킷 CORS(교차 출처 리소스 공유) 규칙을 만듭니다.

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

이 규칙을 사용하면 Intel® AI for Enterprise RAG 웹 애플리케이션용 OPEA가 웹 브라우저 내에서 버킷과 상호 작용할 수 있습니다.

서버 배포

서버를 배포하고 모든 서버에 Ubuntu 22.04 LTS를 설치합니다. Ubuntu를 설치한 후 모든 서버에 NFS 유틸리티를 설치합니다. NFS 유틸리티를 설치하려면 다음 명령을 실행하세요.

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

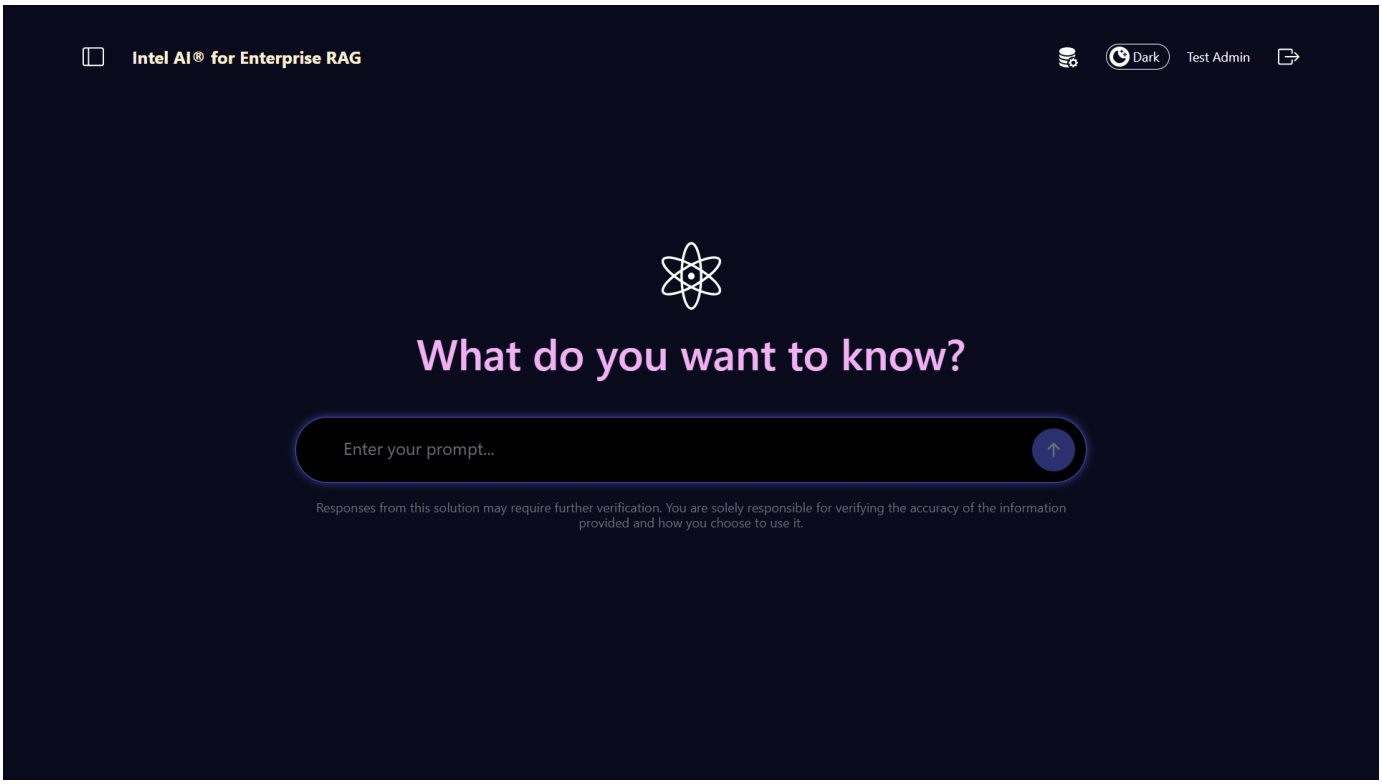
Enterprise RAG 2.0 배포

전체 단계별 배포 워크플로는 다음 문서를 참조하십시오. NetApp AIPod Mini for ERAG - 배포 단계 모든 사전 요구 사항, 인프라 준비, 구성 매개변수 및 배포 절차는 위의 배포 가이드에 설명되어 있습니다.

Enterprise RAG UI용 **Intel® AI**용 **OPEA** 액세스

Intel® AI for Enterprise RAG UI용 OPEA에 액세스하세요. 자세한 내용은 "Intel® AI for Enterprise RAG 배포 설명서"을(를) 참조하십시오.

그림 9 - Intel® AI for Enterprise RAG UI용 OPEA.



RAG에 대한 데이터 수집

이제 RAG 기반 쿼리 증강에 포함할 파일을 수집할 수 있습니다. 파일을 수집하는 데에는 여러 가지 옵션이 있습니다. 귀하의 필요에 맞는 적절한 옵션을 선택하세요.

참고: 파일이 수집된 후 Intel® AI for Enterprise RAG 애플리케이션용 OPEA는 파일 업데이트를 자동으로 확인하고 그에 따라 업데이트를 수집합니다.

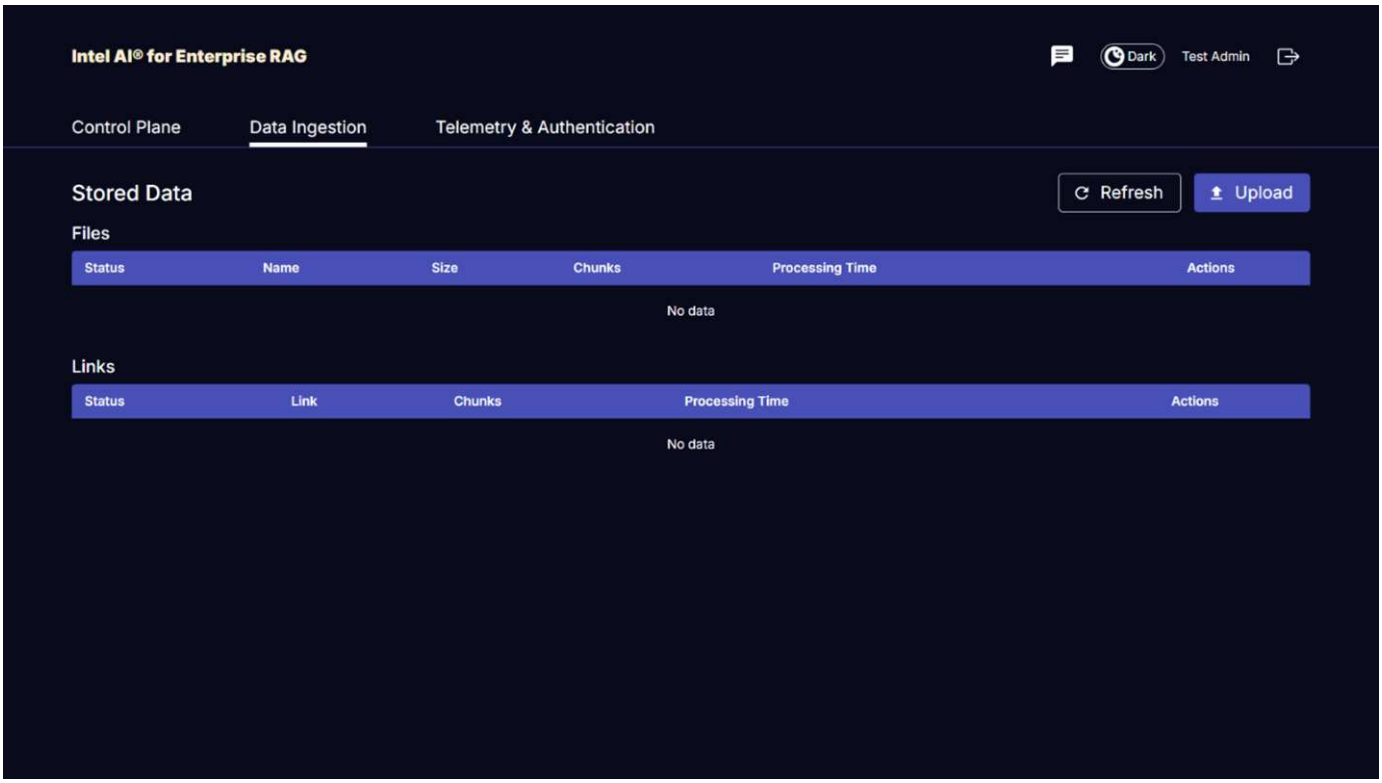
*옵션 1: S3 버킷에 직접 업로드하기 여러 파일을 한 번에 수집하려면 원하는 S3 클라이언트를 사용하여 이전에 생성한 S3 버킷에 파일을 업로드하는 것이 좋습니다. 널리 사용되는 S3 클라이언트로는 AWS CLI, Amazon SDK for Python(Boto3), s3cmd, S3 Browser, Cyberduck, Commander One 등이 있습니다. 파일 형식이 지원되는 경우, S3 버킷에 업로드된 모든 파일은 OPEA for Intel® AI for Enterprise RAG 애플리케이션에 자동으로 수집됩니다.

참고: 이 글을 작성하는 시점에서 지원되는 파일 형식은 PDF, HTML, TXT, DOC, DOCX, ADOC, PPT, PPTX, MD, XML, JSON, JSONL, YAML, XLS, XLSX, CSV, TIFF, JPG, JPEG, PNG 및 SVG입니다.

OPEA for Intel® AI for Enterprise RAG UI를 사용하여 파일이 제대로 수집되었는지 확인할 수 있습니다. 자세한 내용은 Intel® AI for Enterprise RAG UI 설명서를 참조하십시오. 대량의 파일을 수집하는 데 다소 시간이 걸릴 수 있습니다.

*옵션 2: UI를 사용한 업로드 소수의 파일만 수집해야 하는 경우 OPEA for Intel® AI for Enterprise RAG UI를 사용하여 파일을 수집할 수 있습니다. 자세한 내용은 Intel® AI for Enterprise RAG UI 설명서를 참조하십시오.

그림 10 - 데이터 수집 UI.



채팅 쿼리 실행

이제 포함된 채팅 UI를 사용하여 Intel® AI for Enterprise RAG 애플리케이션용 OPEA와 "채팅"할 수 있습니다. 애플리케이션은 사용자의 질문에 응답할 때 수집된 파일을 사용하여 RAG를 수행합니다. 즉, 애플리케이션이 수집된 파일에서 관련 정보를 자동으로 검색하고 해당 정보를 질문에 대한 응답에 통합합니다.

사이즈 가이드

검증 노력의 일환으로 우리는 인텔과 협력하여 성능 테스트를 실시했습니다. 이 테스트를 통해 다음 표에 설명된 크기 지침이 도출되었습니다.

특성화	가치	논평
모델 사이즈	200억 개의 매개변수	라마-8B, 라마-13B, 미스트랄 7B, 퀀 14B, 딥시크 디스틸 8B
입력 크기	~2k 토큰	~4페이지
출력 크기	~2k 토큰	~4페이지
동시 사용자	32	"동시 사용자"란 동시에 쿼리를 제출하는 프롬프트 요청을 말합니다.

_참고: 위에 제시된 사이징 지침은 96개의 코어가 있는 Intel Xeon 6 프로세서를 사용하여 수집한 성능 검증 및 테스트 결과를 기반으로 합니다. 유사한 I/O 토큰 및 모델 크기 요구 사항이 있는 고객의 경우 96개의 코어가 있는 Xeon 6 프로세서가 장착된 서버를 사용하는 것이 좋습니다. 사이징 가이드에 대한 자세한 내용은 "Intel® AI for Enterprise RAG 사이징 가이드"을(를) 참조하십시오.

결론

엔터프라이즈 RAG 시스템과 LLM은 조직이 정확하고 상황에 맞는 응답을 제공할 수 있도록 함께 작동하는 기술입니다. 이러한 응답에는 방대한 양의 비공개 및 내부 기업 데이터를 기반으로 한 정보 검색이 포함됩니다. RAG, API, 벡터 임베딩 및 고성능 스토리지 시스템을 사용하여 회사 데이터가 포함된 문서 저장소를 쿼리함으로써 데이터를 더 빠르고 안전하게 처리할 수 있습니다. NetApp AIPod Mini는 NetApp의 지능형 데이터 인프라와 ONTAP 데이터 관리 기능, Intel Xeon 6 프로세서, Intel® AI for Enterprise RAG 및 OPEA 소프트웨어 스택을 결합하여 고성능 RAG 애플리케이션을 배포하고 조직이 AI 리더십을 확보할 수 있도록 지원합니다.

승인

이 문서는 NetApp 솔루션 엔지니어링 팀의 Sathish Thyagarajan, Michael Oglesby, Arpita Mahajan이 작성했습니다. 또한, 솔루션 검증 과정에서 지속적인 지원과 도움을 주신 Intel의 엔터프라이즈 AI 제품 팀(Ajay Mungara, Mikolaj Zyczynski, Igor Konopko, Ramakrishna Karamsetty, Michal Prostko, Anna Alberska, Maciej Cichocki, Shreejan Mistry, Nicholas Rago, Ned Fiori)과 NetApp(Lawrence Bunka, Bobby Oommen, Jeff Liborio)에게 감사를 표합니다.

재료 목록

다음은 이 솔루션의 기능 검증에 사용된 BOM이며 참조로 사용할 수 있습니다. 다음 구성에 맞는 모든 서버나 네트워킹 구성 요소(또는 기존 네트워크(대역폭이 100GbE인 경우))를 사용할 수 있습니다.

앱 서버의 경우:

부품번호	제품 설명	수량
222HA-TN-OTO-37	하이퍼 슈퍼서버 SYS-222HA-TN /2U	2
P4X-GNR6972P-SRPL2-UC	Intel® Xeon® 6972P 프로세서 96코어 2.40GHz 480MB 캐시(500W)	4
숫양	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 중복 단일 출력 전원 공급 장치. 최대 온도 59°C(대략)에서 열 방출은 2361 BTU/Hr입니다.	4

제어 서버의 경우:

부품번호	제품 설명	수량
511R-M-OTO-17	최적화된 1U X13SCH-SYS, CSE-813MF2TS-R0RCNBP, PWS-602A-1R	1

	RPL-E 6369P IP 8C/16T 3.3G 24MB 95W 1700 BO	1
숫양	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

네트워크 스위치의 경우:

부품번호	제품 설명	수량
DCS-7280CR3A	아리스타 7280R3A 28x100GbE	1

NetApp AFF 스토리지:

부품번호	제품 설명	수량
AFF-A20A-100-C	AFF A20 HA 시스템, -C	1
X800-42U-R6-C	점퍼 케이블, 인캡, C13-C14, -C	2
X97602A-C	전원 공급 장치, 1600W, 티타늄, -C	2
X66211B-2-N-C	케이블, 100GbE, QSFP28-QSFP28, Cu, 2m, -C	4
X66240A-05-N-C	케이블, 25GbE, SFP28-SFP28, Cu, 0.5m, -C	2
X5532A-N-C	레일, 4-포스트, 얇은, 라운드/사각형 구멍, 소형, 조정식, 24-32, -C	1
X4024A-2-A-C	드라이브 팩 2X1.92TB, NVMe4, SED, -C	6
X60130A-C	IO 모듈, 2PT, 100GbE, -C	2
X60132A-C	IO 모듈, 4PT, 10/25GbE, -C	2
SW-ONTAPB-FLASH-A20-C	SW, ONTAP 기본 패키지, TB당, 플래시, A20, -C	23

인프라 준비 상태 점검표

자세한 내용은 NetApp AIPod Mini - 인프라 준비 상태를 참조하십시오.

추가 정보를 찾을 수 있는 곳

이 문서에 설명된 정보에 대해 자세히 알아보려면 다음 문서 및/또는 웹사이트를 검토하세요.

"NetApp 제품 설명서"

"OPEA 프로젝트"

"Intel® AI ERAG 문서"

"OPEA Enterprise RAG 배포 플레이북" == 버전 기록

버전	날짜	문서 버전 기록
버전 1.0	2025년 9월	초기 릴리스
버전 2.0	2026년 2월	OPEA-Intel® AI for Enterprise RAG 2.0으로 업데이트되었습니다.