ORACLE

오라클의 생성형 AI 지원 전략 및 주요 기술

미래 비즈니스를 꿈꾸는 오라클의 생성형 AI, 무엇이 다를까?

장성우
Cloud Engineering, Oracle Korea
Jan 25, 2024

Safe harbor statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

Overall Message

오라클의 생성형 AI 지원 전략과 기업 데이터 플랫폼과의 통합 방안

- 오라클의 생성형 AI 지원 전략
  - ✓ AI Infrastructure : NVIDIA와의 협력을 통한 GPU SuperCluster 지원
  - ✓ App embedding with Gen AI : Fusion Application에 Gen AI 내재화
  - ✓ Gen AI Service : SOTA LLM(Llama2,Cohere)을 활용한 fine-tuning 및 inference의 end-to-end process 지원
  - ✓ Vector type in Oracle DB 23c : Oracle DB의 Vector DB로의 활용 지원 → 기존 투자 활용, 쉬운 구축, 성능/안정성
  - ✓ RAG Agent : 기업 데이터 플랫폼과 결합된 RAG 구축 지원
- 생성형 AI와 기업 데이터 플랫폼과의 통합 방안
  - ✓ Oracle Modern Data Platform with in-DB ML and Gen AI
  - ✓ Data Platform과 생성형 AI의 유기적인 연계를 통한 기업 수준의 통합된 AI / Data platform 지원

Agenda

1. 오라클의 생성형 AI 지원 전략
2. OCI Generative AI Service
3. Oracle DB 23c에서의 Vector Type과 RAG 지원
4. Oracle Modern AI / Data Platform
5. Summary

오라클의 생성형 AI 지원 전략

모든 계층에서 AI를 구현하는 오라클

Oracle Cloud Applications

- Oracle의 클라우드 애플리케이션 및 데이터베이스 포트폴리오 전반에 생성형 AI를 통합하여 제공
- Oracle의 제품과 서비스가 AI 기술을 활용하여 사용자의 업무 효율성을 향상시키고, 비즈니스 프로세스를 자동화하며, 데이터 분석 및 인사이트를 개선

OCI Generative AI service

- 미세 조정(fine-tuned)이 가능한 LLM이 탑재된 새로운 AI 서비스
- 오라클은 Meta 및 Cohere와 협력하여 생성형 AI 서비스를 제공
- GA(Generally Available)

OCI AI Infrastructure

- 모델 학습 및 추론을 위한 고속, 저비용의 AI 인프라
- OCI Supercluster를 통한 업계 최고의 확장성과 초저지연 고대역폭의 AI에 최적화된 네트워크로 설계

오라클의 AI 지원 전략 및 서비스

Applications

- Fusion Applications
- NetSuite
- Fusion Analytics
- Industry Applications
- 3rd Party Applications

AI Services

- NEW OCI Generative AI
- NEW OCI GenAI Agents
- Digital Assistant
- Speech
- Language
- Vision
- Document Understanding

Data Platforms

- OCI Data Science
- AI Vector Search in Oracle Database
- MySQL HeatWave Vector Store
- OCI Data Labeling

Data

AI Infrastructure

- Compute bare metal instances and VMs with NVIDIA GPUs
- OCI Supercluster with RDMA networking
- Block, object, and file storage; HPC filesystems

Oracle AI Partners

OCI Generative AI Service : LLM 기반 학습/추론 지원 서비스

GA(Generally Available) : 1/24

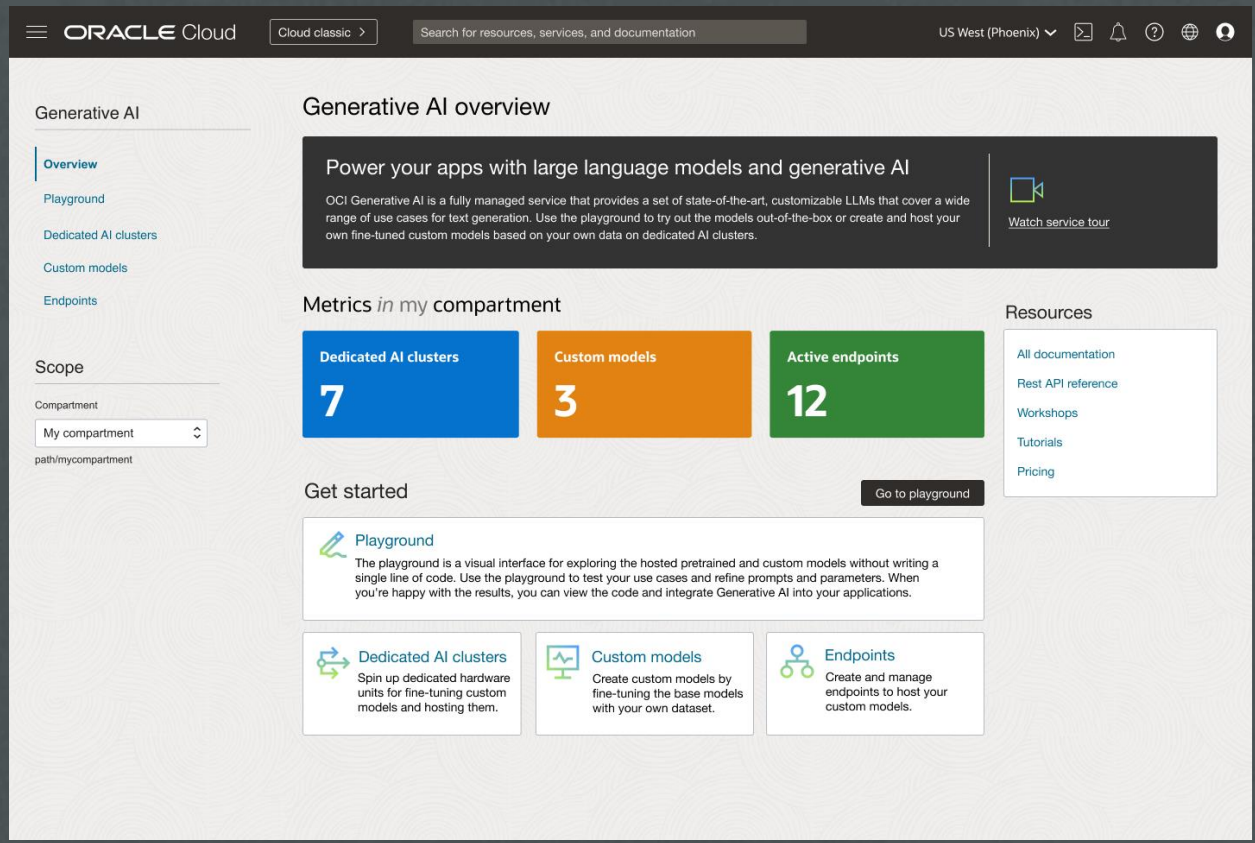
고성능 사전 학습 모델
Llama2 및 Cohere LLM을 FM로 제공

Fine-tuning 지원
자체 데이터로 사용자 전용 모델로 Fine-Tuning하여 특정 문제를 해결하거나 특정 도메인을 최적화

OCI 기반 호스팅
모든 LLM 데이터 스토리지 및 LLM 프로세싱, 모델 버전, 업그레이드 관리는 OCI의 완전 관리형 서비스 형태로 제공

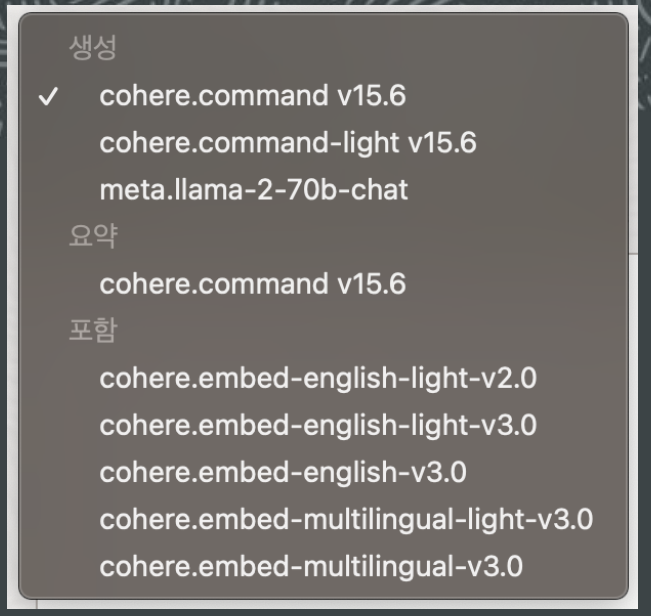
고객 개인정보 보호
고객이 제공한 학습 및 추론 데이터는 안전하게 보호되며 다른 고객이 접근 불가. Cohere와 데이터가 공유되지 않음

PAYG 및 단일 테넌트, 전용 하드웨어
안정적인 성능을 위해 토큰 단위로 지불하거나 전용 하드웨어에서 모델을 호스팅



OCI Generative AI Service가 지원하는 LLM

엔터프라이즈 기업을 위한 SOTA LLM 모델 제공



Llama-2 70B

Meta에서 개발한 llama-2 70B 파라미터 텍스트 생성 모델
연구 및 상업적 용도로 무료로 사용 가능한 개방형 LLM

cohere

Command

Command는 Cohere의 고성능 생성형 LLM 모델
모델 크기: 6B (Medium), 52B 파라미터 (XL)
Command XL 모델이 Command M 모델 보다 더 높은 정확성, Command M은 비용 및 처리 속도에 강점

Summarize

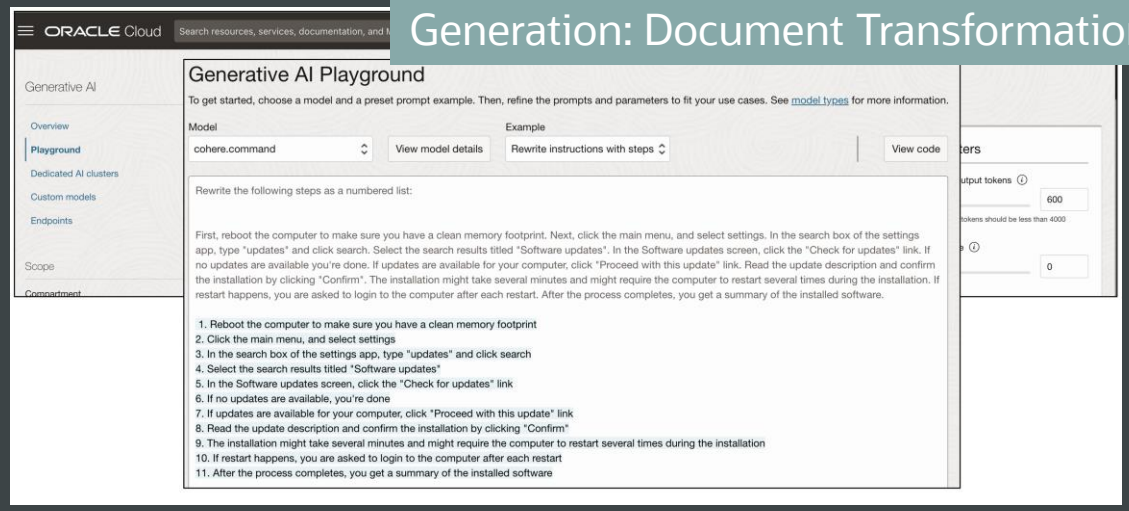
Summarize 모델은 문서에서 가장 중요한 정보를 정확하게 포착하고 고품질 요약을 생성하는 특화 모델

Embed

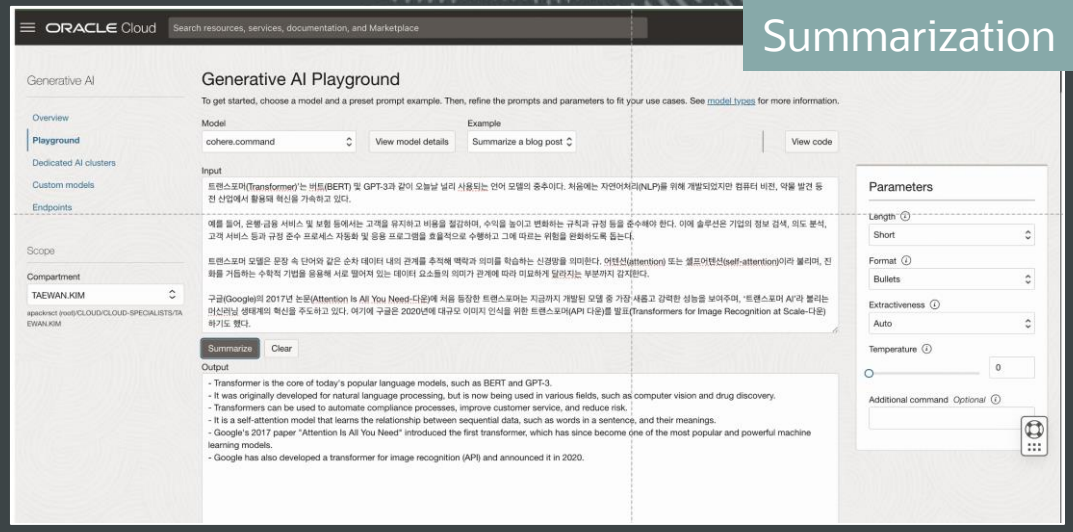
Embed는 텍스트를 벡터로 임베딩하는 모델로 영어 및 다국어 언어 임베딩 모델(V3) 지원.
더 작고 빠르지만 성능이 약간 떨어지는 '라이트' 버전의 모델도 제공 (영어전용)

OCI Generative AI 서비스: Playground

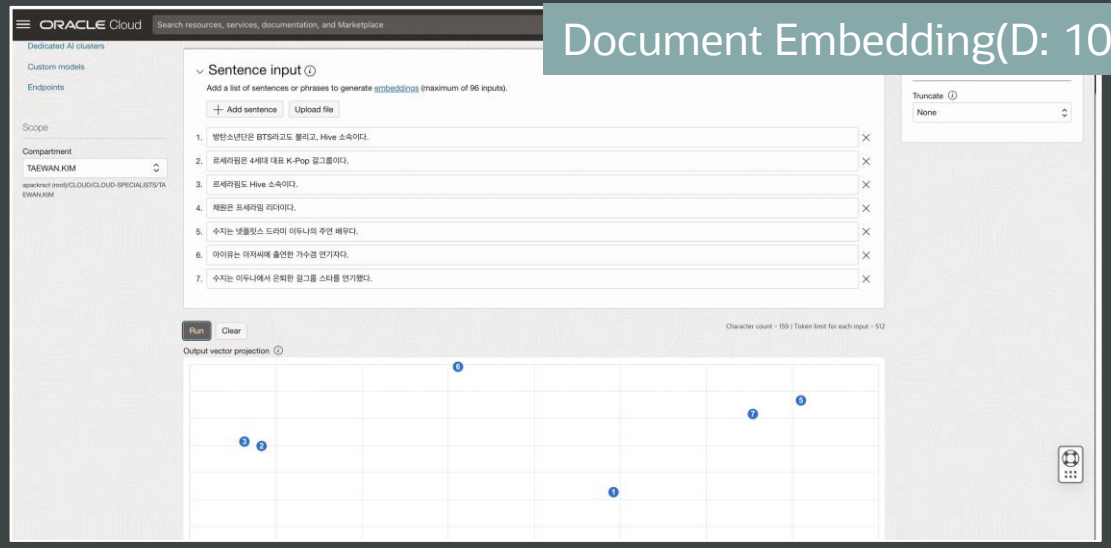
Generation: Document Transformation



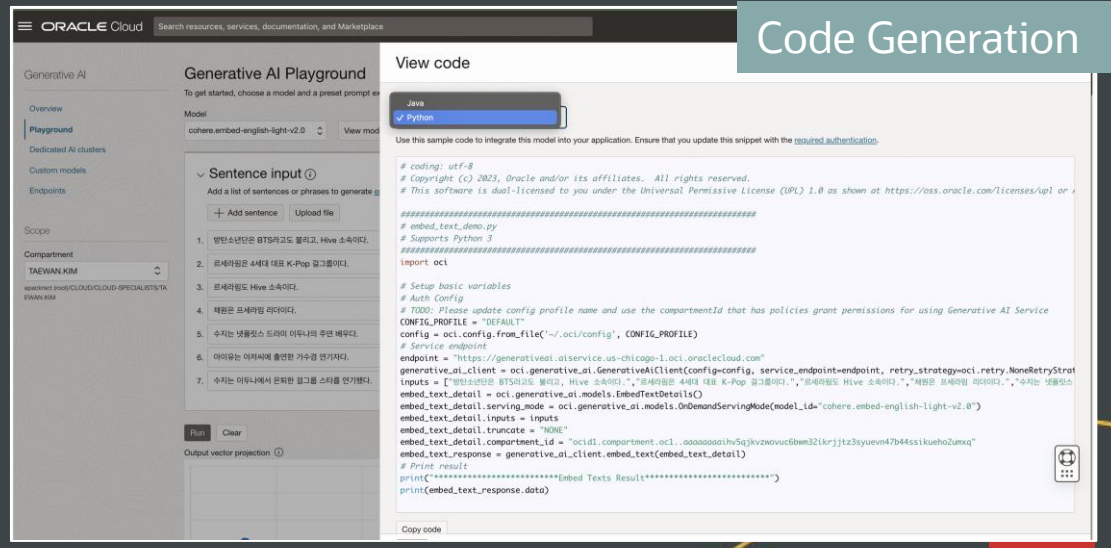
Summarization



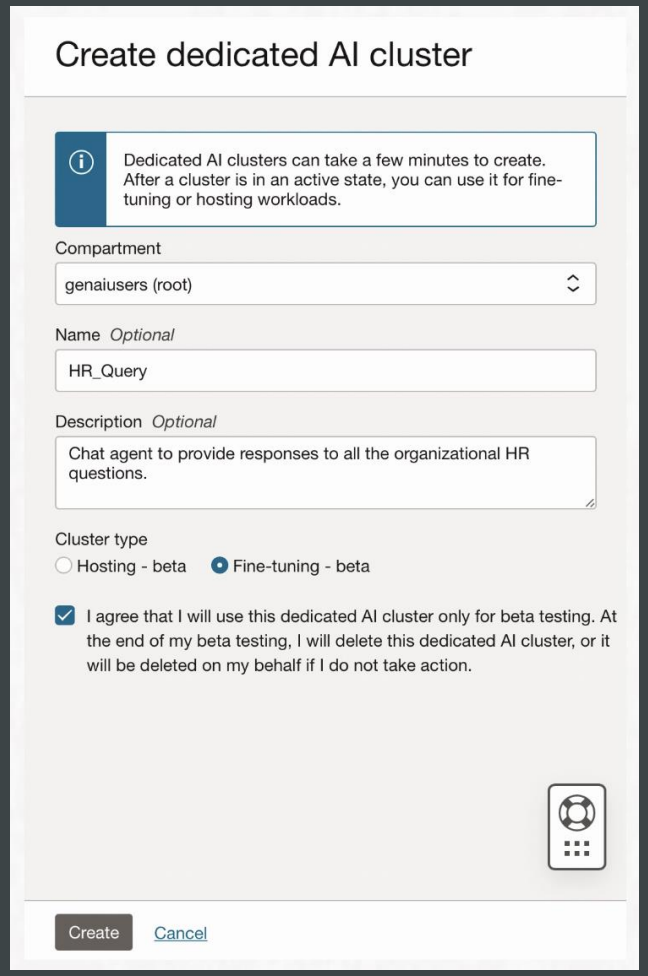
Document Embedding(D: 1024)

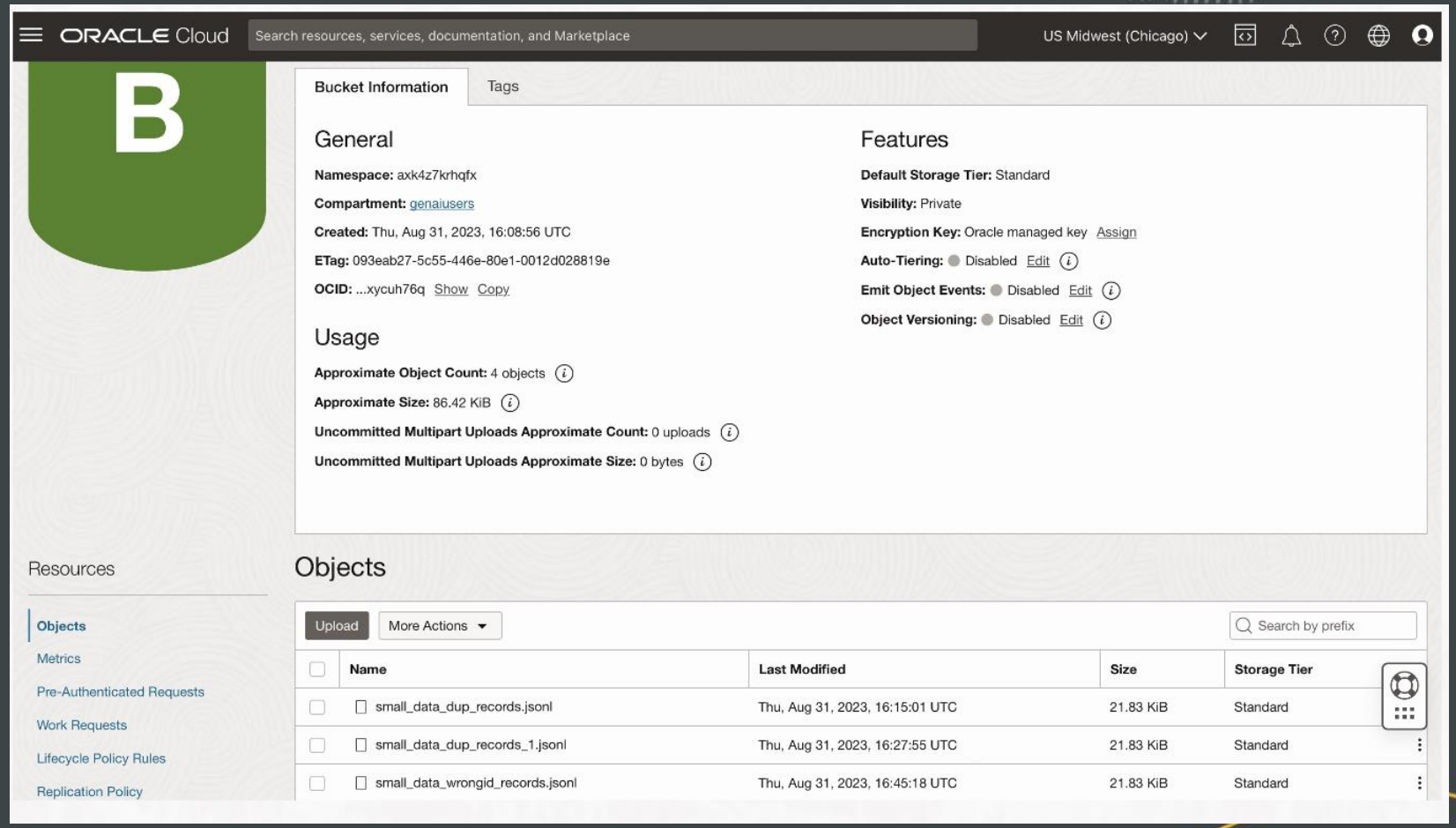


Code Generation

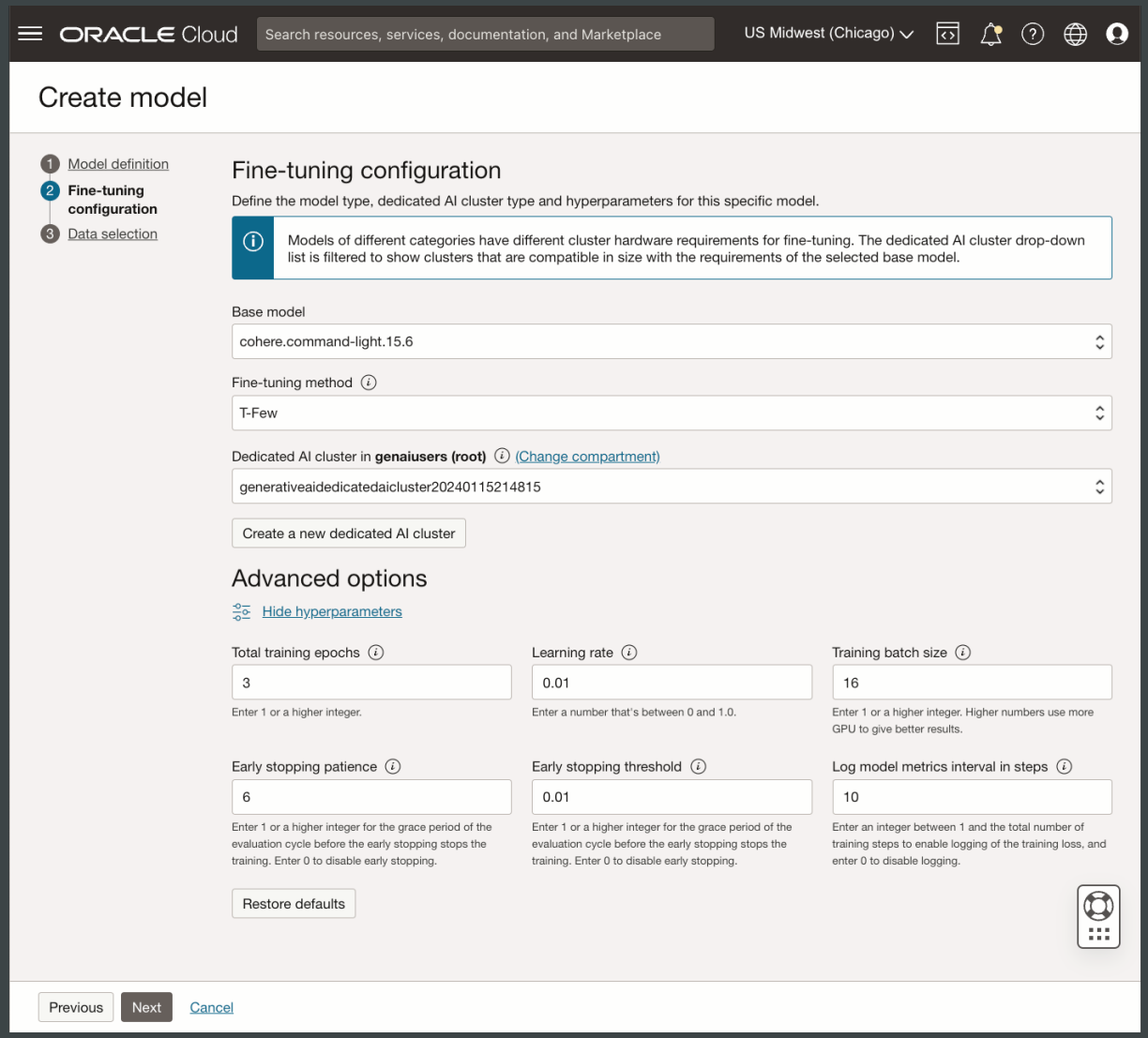


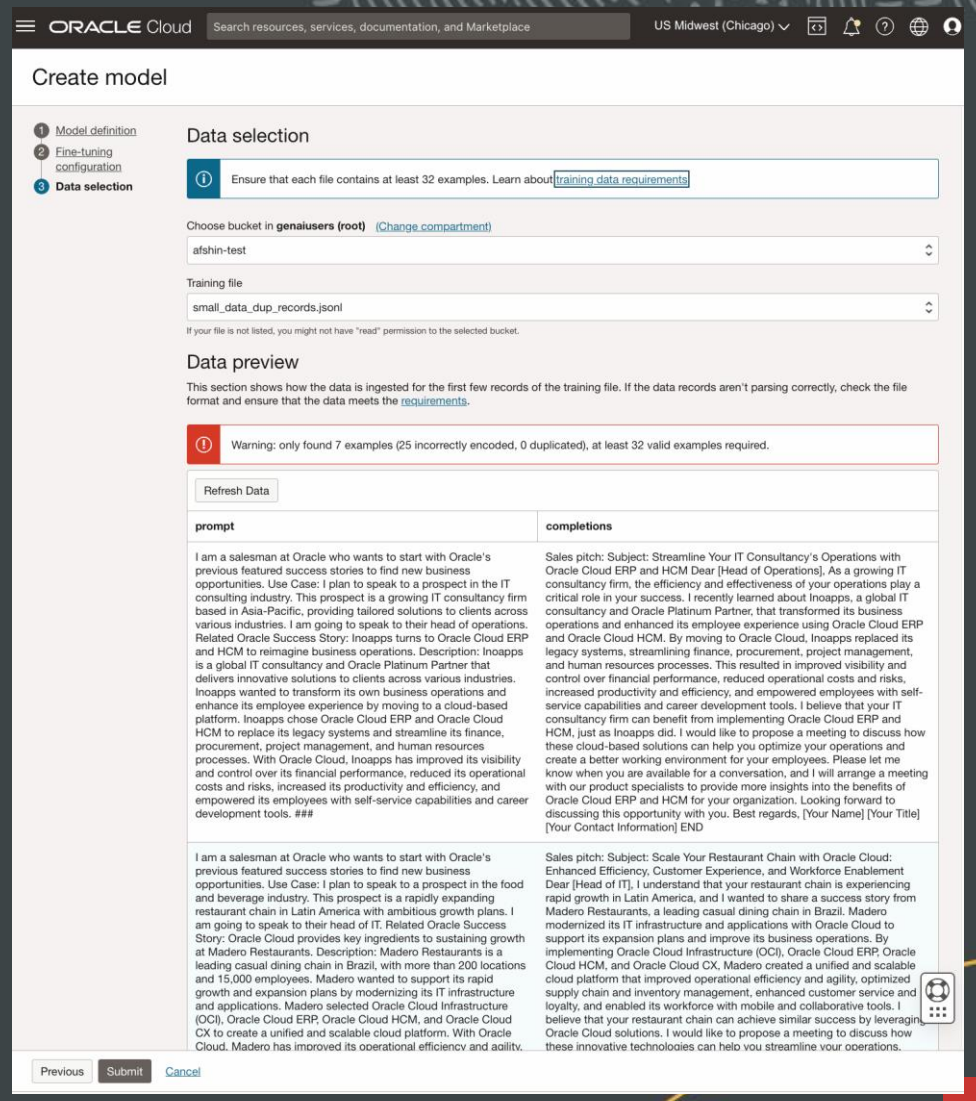
OCI Generative AI 서비스: 전용 모델



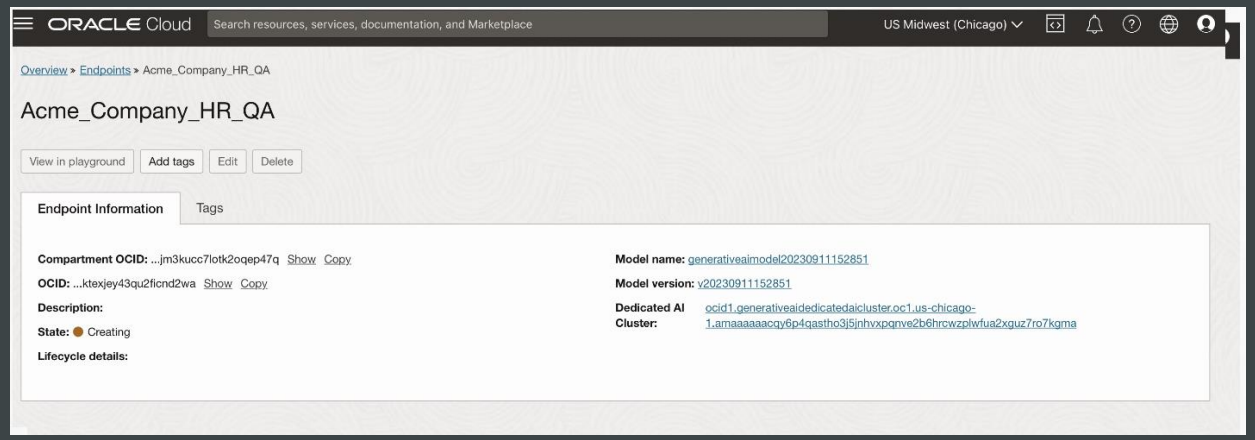


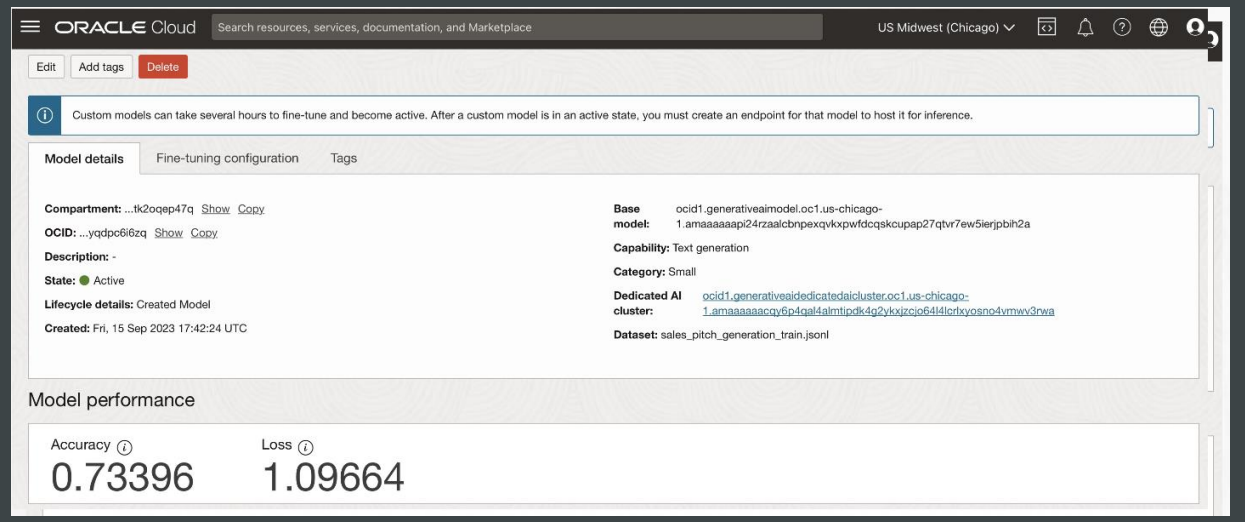
OCI Generative AI 서비스: LLM 모델 Fine-Tuning

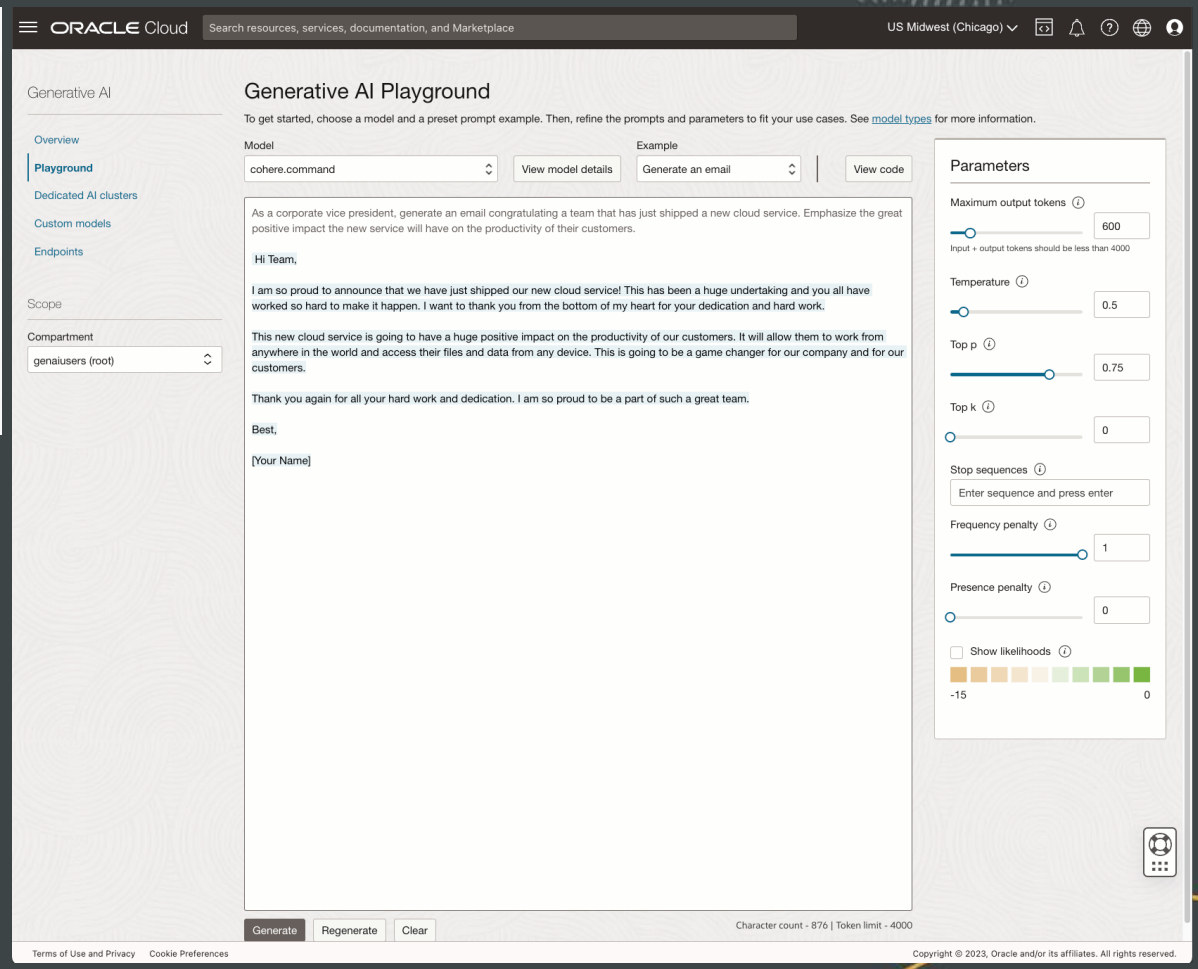




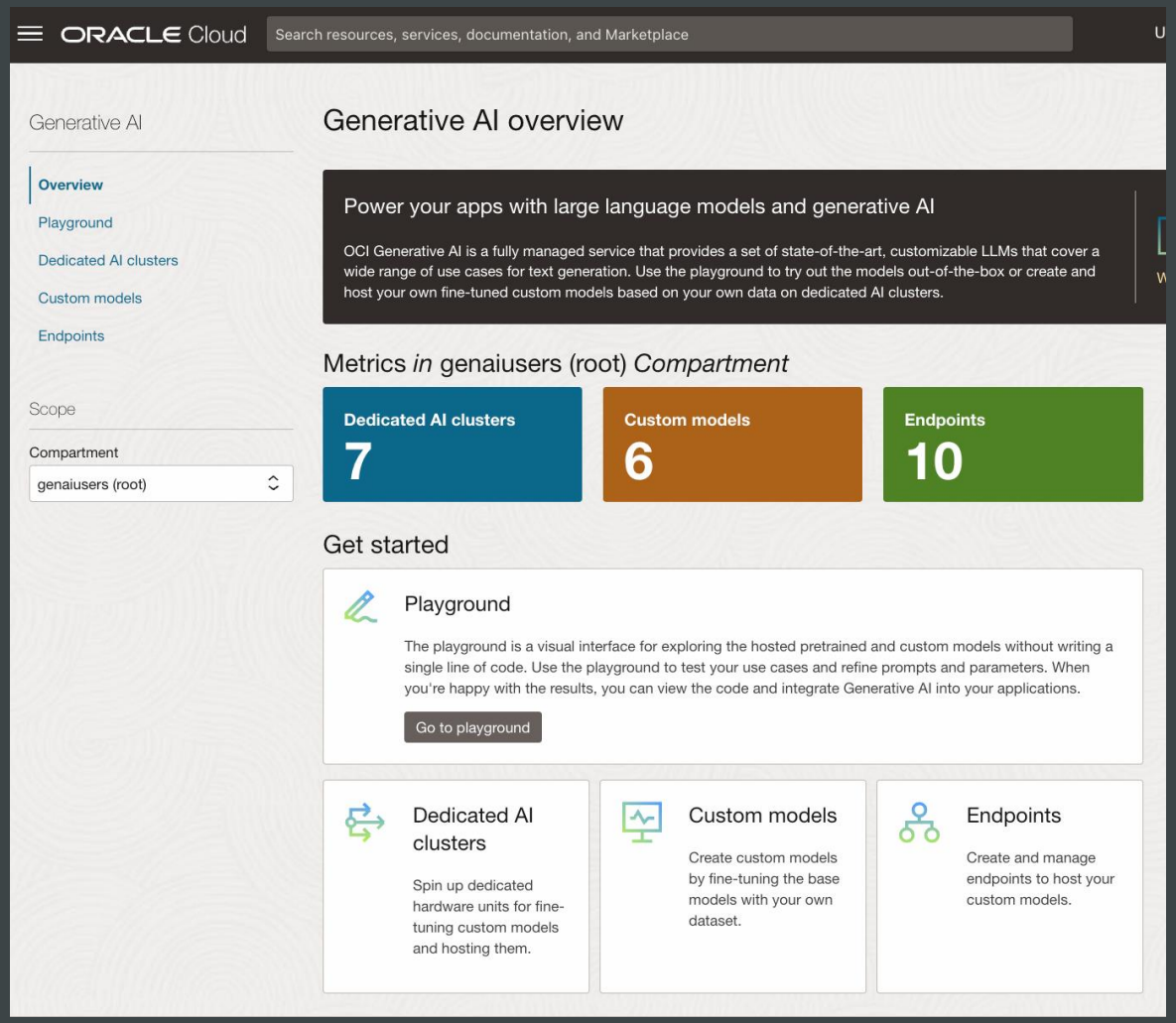
OCI Generative AI 서비스: LLM 모델 Fine-Tuning

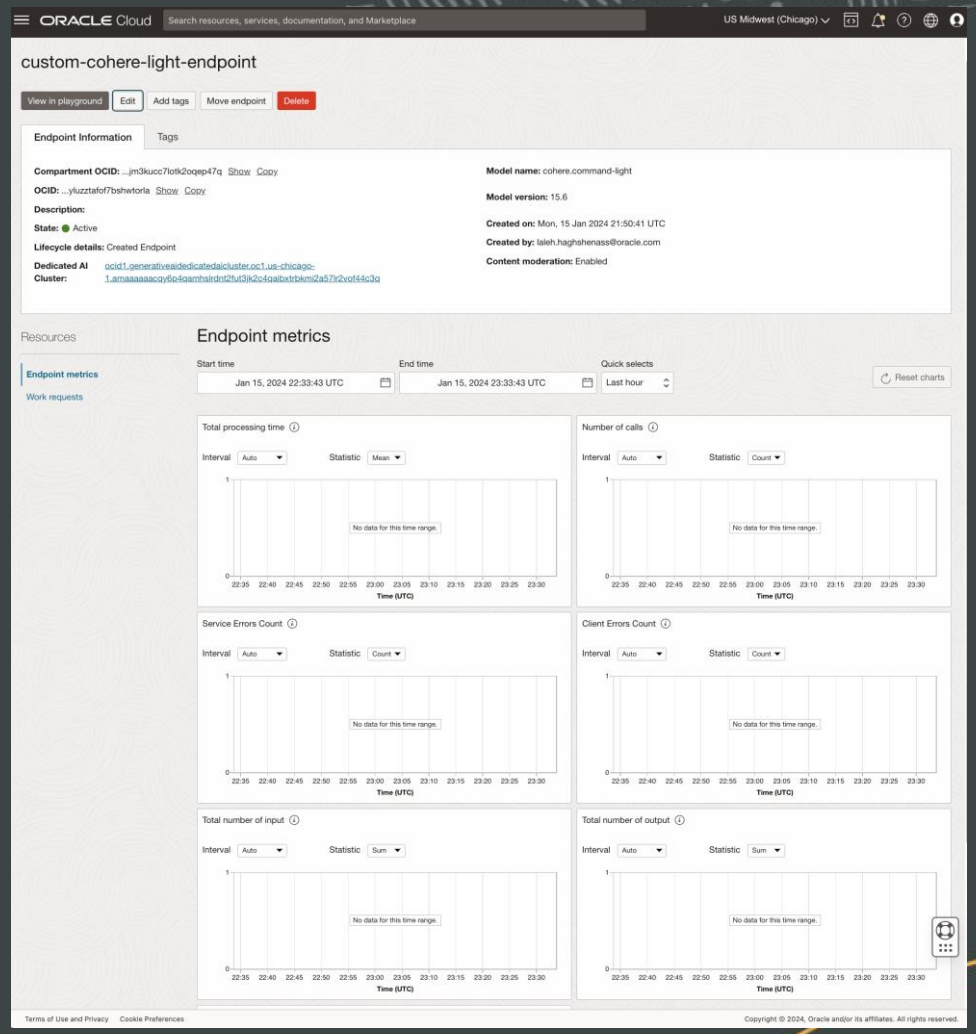






OCI Generative AI 서비스: LLM 모델 관리

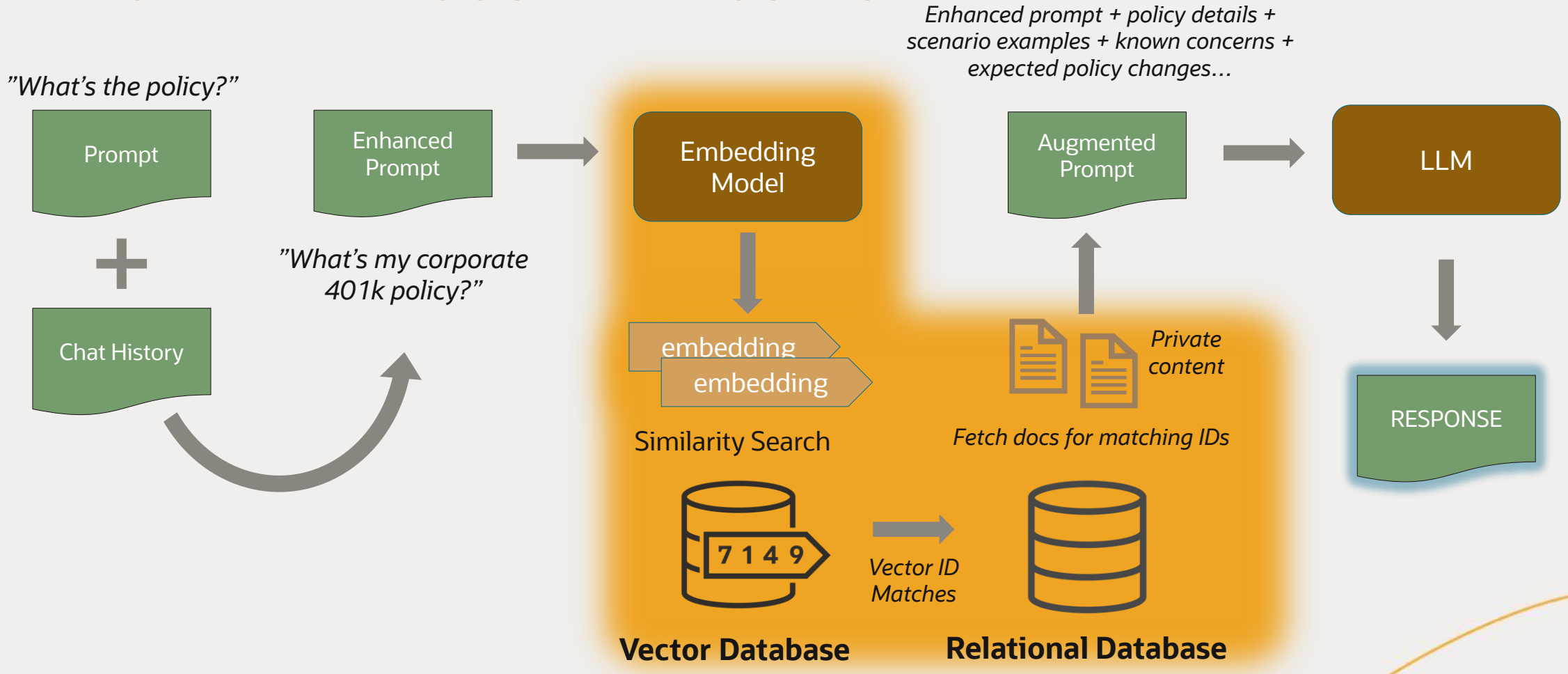




RAG(Retrieval Augmented Generation)

기업 지식/데이터를 활용하여 hallucination 없는 Q&A 지원

- ✓ **Enterprise-level Gen AI Application 개발을 위해서는 RAG 지원은 필수**
- ✓ **오라클은 Oracle DB 내에서 Vector DB의 구성 지원**



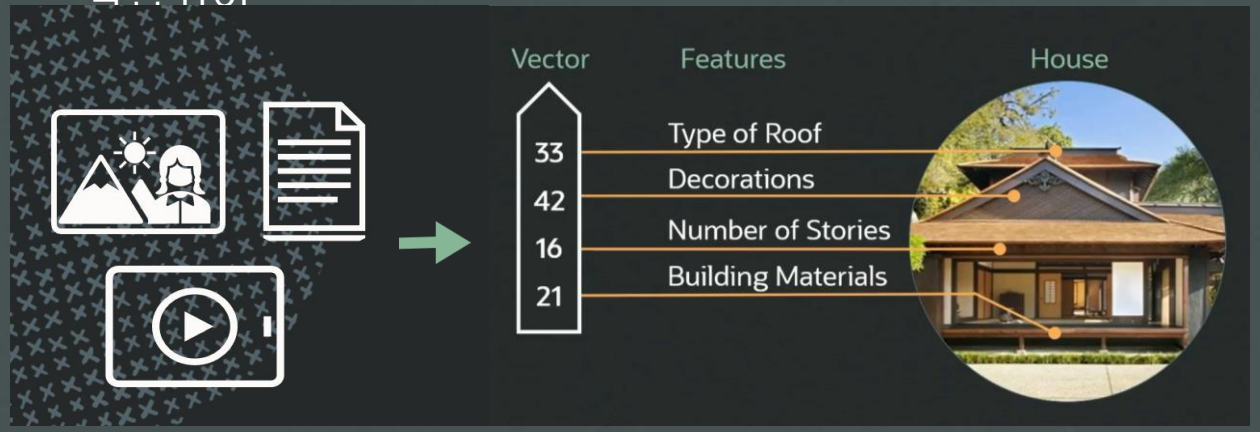
RAG 지원을 위한 Oracle DB의 전략 : Vector Type in DB 23c

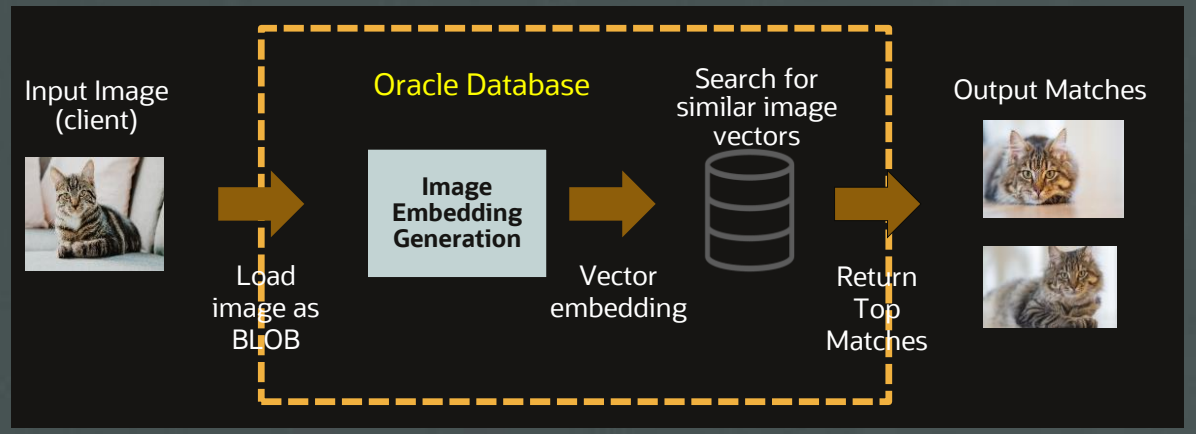
Vector Search: 오라클 DB 내에서 Vector Type 제공

- 23c에서 Vector Type 지원
- AI vector는 비정형 데이터 자체가 아니라 semantic content를 나타내는 특정 feature들을 담고 있는 데이터
- 기존 테이블 구조에 vector type을 추가하여 keyword + semantic search 수행 가능
- 기존의 RAC/Exadata/Partition/Sharding과 결합되어

Oracle DB 내에서의 Vector DB 구성의 장점

- 고객들은 이미 오라클 DB 내에 대규모의 **operational/analytical enterprise data**를 운용 중
- 별도의 Vector DB를 구성하는 것보다 이미 운영 중인 오라클 DB 안에 vector embedding을 추가하는 것이 (1) 구현의 용이성, (2) 비용 절감, (3) 성능/안정성 제공 측면에서 효율적





Vector 데이터 타입과 SQL 활용 예제

New VECTOR datatype (with underlying BLOB storage for long-term extensibility)

```
VECTOR (<optional # of dimensions>, <optional format for dimension values>)
create table my_images (id number, image BLOB, img_vec VECTOR(768, FLOAT32))
```

Insert

```
create table vec_tab(id number, dataVec VECTOR(3, 'FLOAT32'))
Insert into vec_tab values (1, TO_VECTOR('[1.1, 2.2, 3.3]')
```

Fetch

```
Select dataVec from tab-> Select FROM_VECTOR(dataVec) from tab -> '[1.1, 2.2, 3.3]'
```

Get the Top-5 Nearest Vectors to a given query

```
select id from tab order by VECTOR_DISTANCE(data_vec, :queryVec)
fetch first 5 rows only;
```

Vector Indexing 지원

Vector search에서 top-k 일치 항목을 찾기 위한 테이블의 모든 벡터와 쿼리 벡터 간의 거리 계산은 100% 정확하지만 반면에 매우 느림

새로운 벡터 인덱스는 속도를 위해 검색 정확도를 trade-off

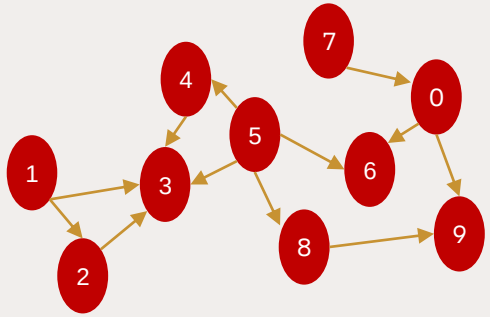
- 벡터들은 정확성을 위해 유사성을 기반으로 군집 후 연결되어짐
- 속도를 위해 정확도를 제한한 greedy search 기법 사용

Vector indexes

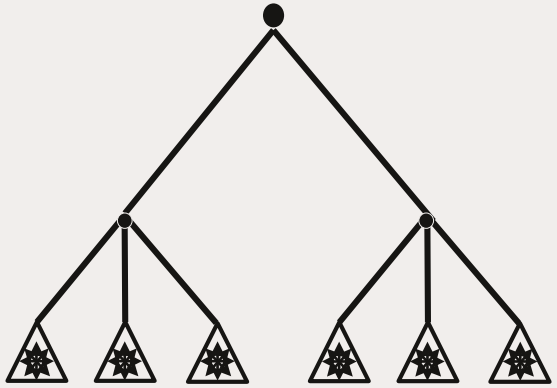
- **In-Memory Neighbor Graph Vector Index** – 노드는 벡터를 나타내고 노드 사이의 에지는 유사도를 나타내는 그래프 기반 인덱스. 정확성과 속도 모두에 있어 매우 효율적임
- **Neighbor Partition Vector Index** – 유사성을 기준으로 테이블 파티션으로 클러스터링된 벡터가 포함된 파티션 기반 인덱스. 빠르고 원활한 트랜잭션 지원을 갖춘 효율적이고 확장이 가능한 인덱스

오라클은 2가지 타입의 vector index를 모두 지원

Graph Vector Index
(**HNSW** Index)

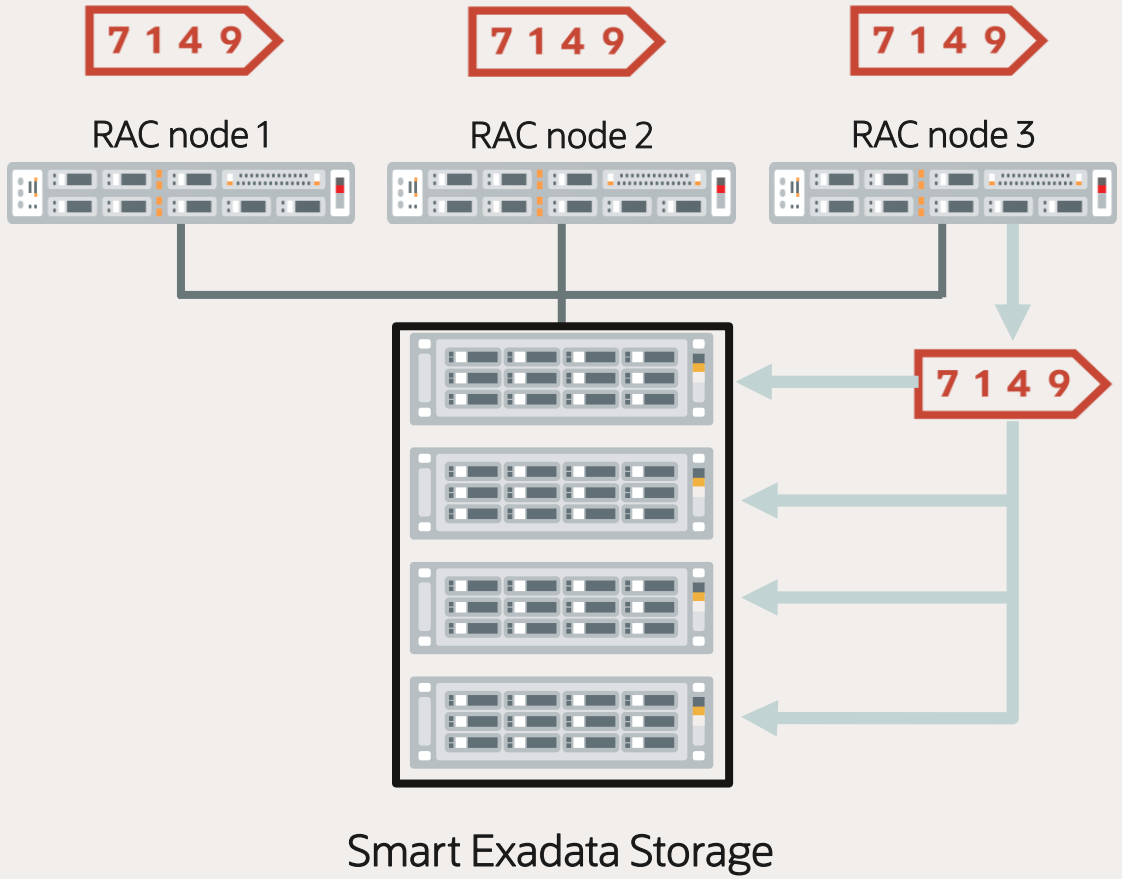


Partition Vector Index
(**IVF_FLAT** index)



오라클 AI Vector Search의 장점 | 기존 기능과의 결합을 통한 빠른 성능과 안정성 제공

Scale-Out with RAC & Exadata



RAC를 활용한 벡터 처리

Oracle은 완전한 데이터 일관성을 유지하면서 RAC 클러스터 내의 컴퓨터 전반에 걸쳐 벡터 처리를 투명하게 확장함

Exadata 기반의 빠른 Vector Search

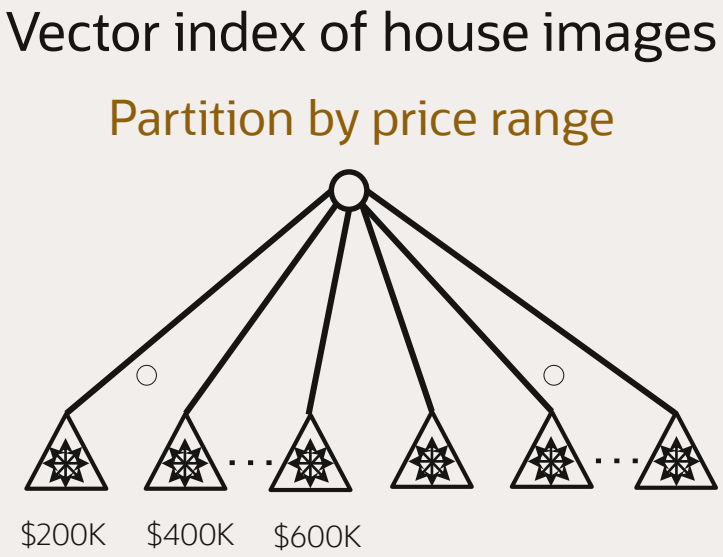
더 빠른 검색을 위해 Oracle AI 벡터 검색을 Smart Exadata Storage로 투명하게 오프로드하여 처리할 수 있음

오라클 AI Vector Search의 장점 | 기존 기능과의 결합을 통한 빠른 성능과 안정성 제공

Scale-Out with Partitioning and Sharding

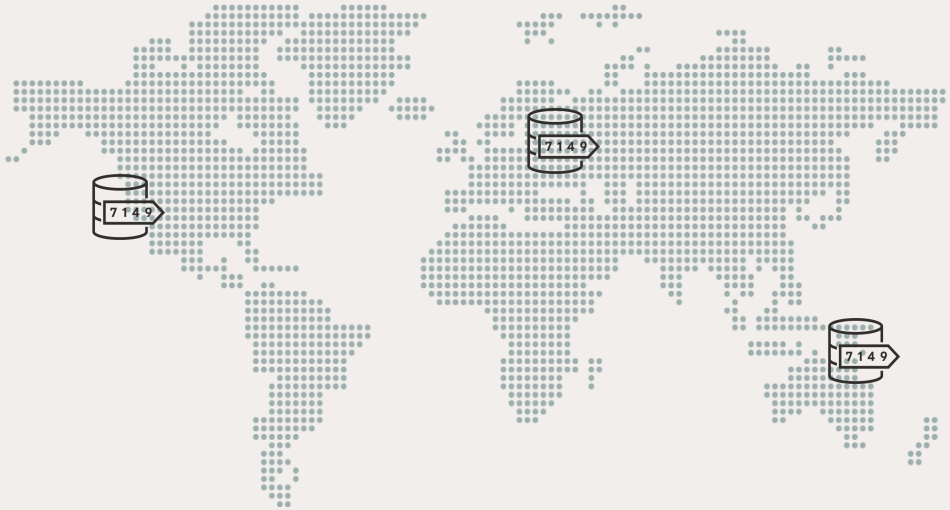
Partitioning

파티션별로 별도의 벡터 인덱스 생성



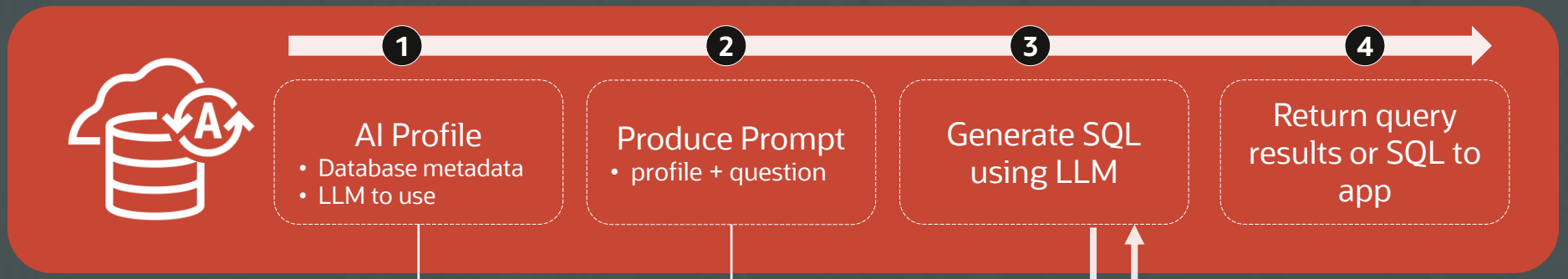
Sharding

무제한 확장 또는 데이터 주권을 위해
데이터베이스 내의 sharding 기능을 사용하여
벡터 처리를 지역적으로 분산할 수 있음



오라클 DB 23c : LLM을 이용한 SQL 생성 지원

```
SELECT AI <ACTION>
Give me the average
salary of employees in
each department;
```




```
dbms_cloud_ai.create_profile
```

```
dbms_cloud_ai.set_profile
```

ACTION

runsql	return sql result set
narrate	return a conversational result
showsql	return the generated query
chat	general AI-chat conversation

LLM을 이용한 SQL 생성 지원(23c)

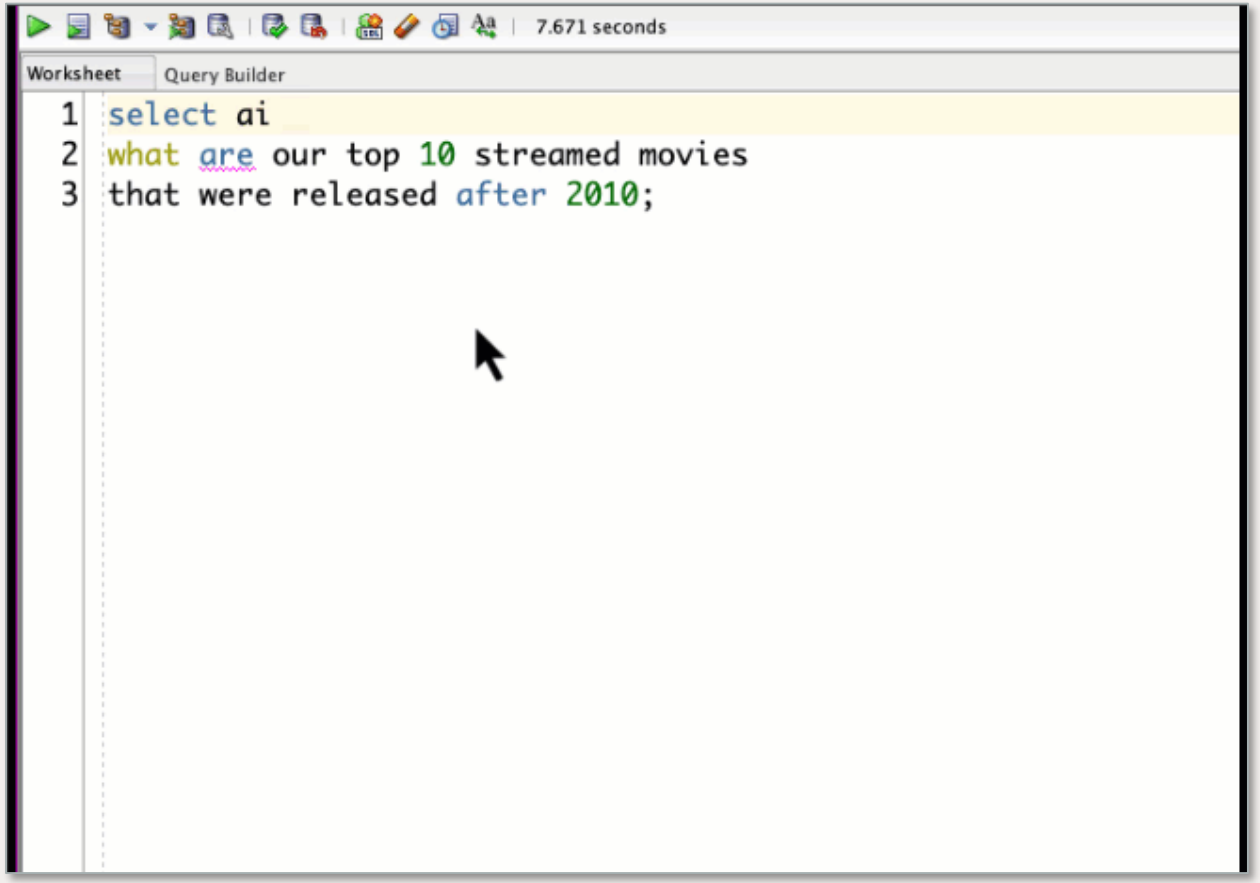
새로운 DBMS_CLOUD_AI PL/SQL 패키지를 사용하여 사용 가능

키워드 AI 및 질문과 함께 표준 SELECT 문을 사용

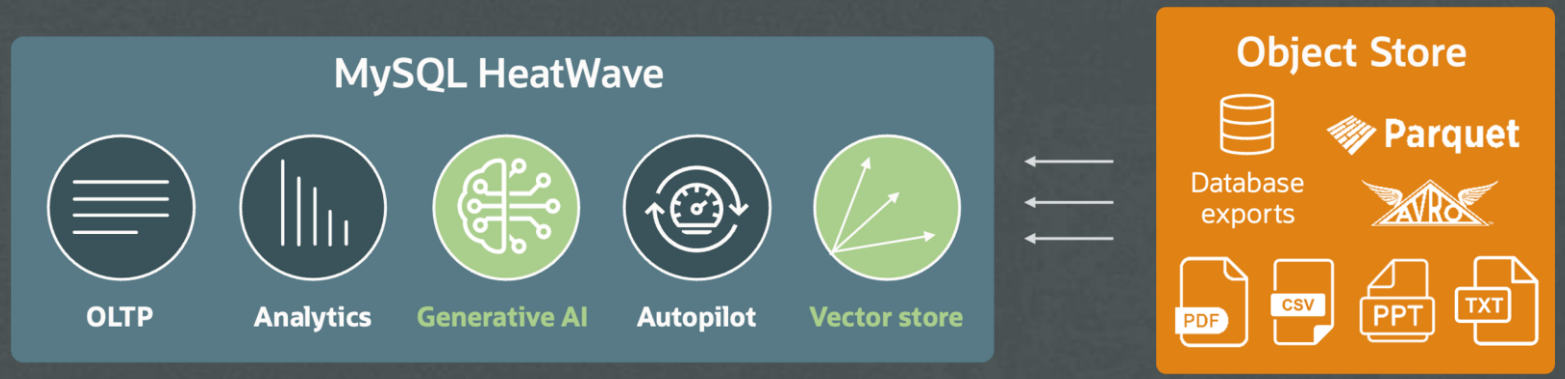
다른 SQL 결과 집합과 마찬가지로 결과를 처리

Actions

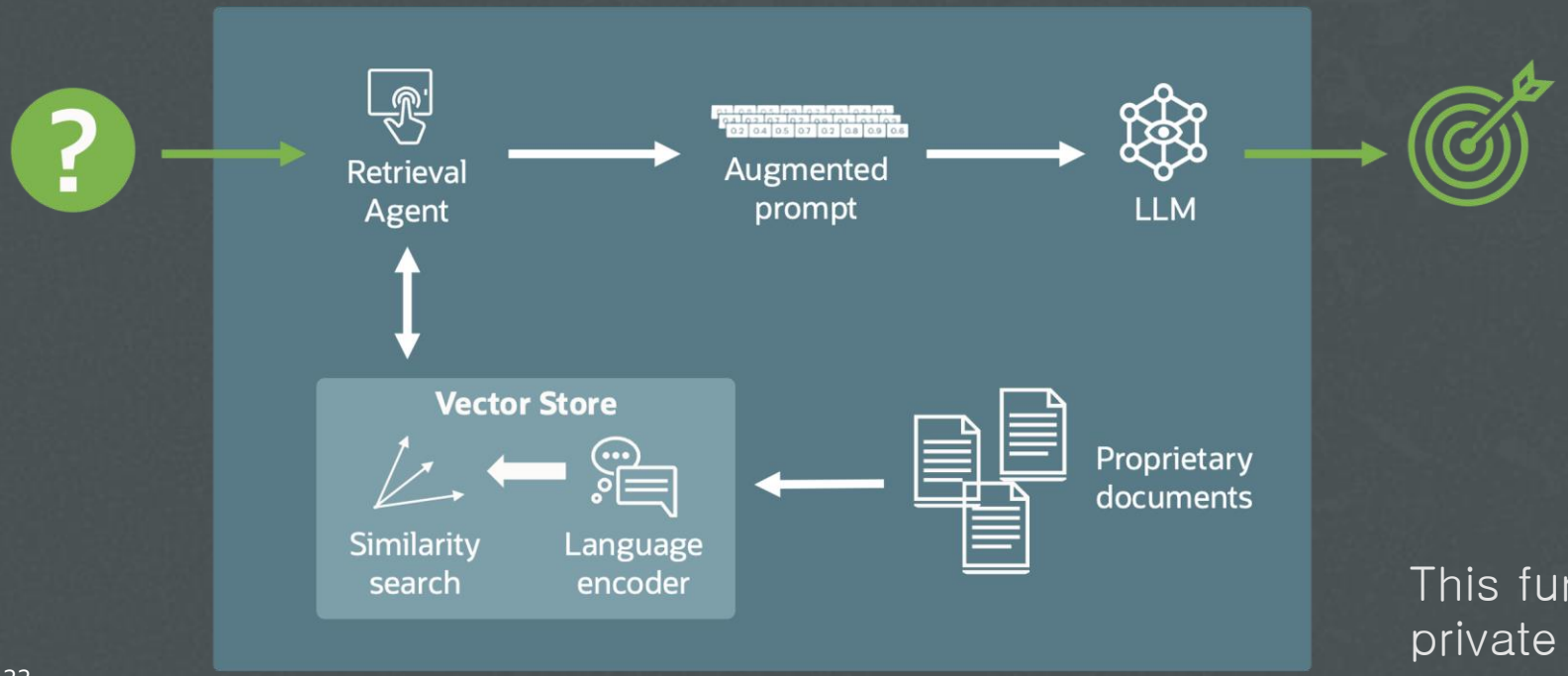
- **runsql** - SQL 결과 집합 반환(기본값)
- **showsql** – 생성된 쿼리문 반환
- **narrate** – 대화 결과 반환
- **Chat** – 일반 AI 챗



MySQL HeatWave 역시 Vector Store를 통해 생성형 AI 지원



- 사용자는 자연어로 정보를 쿼리하고 검색 가능
- RAG 지원을 위해 Vector Store 기능 제공



- Vector Store는 documen를 분석하여 embedding(Vector 형태) 을 생성한 후 저장. 이 embedding을 LLM 입력으로 사용하여 사용자 질의 시 더 정확한 Q&A 처리가 가능해짐

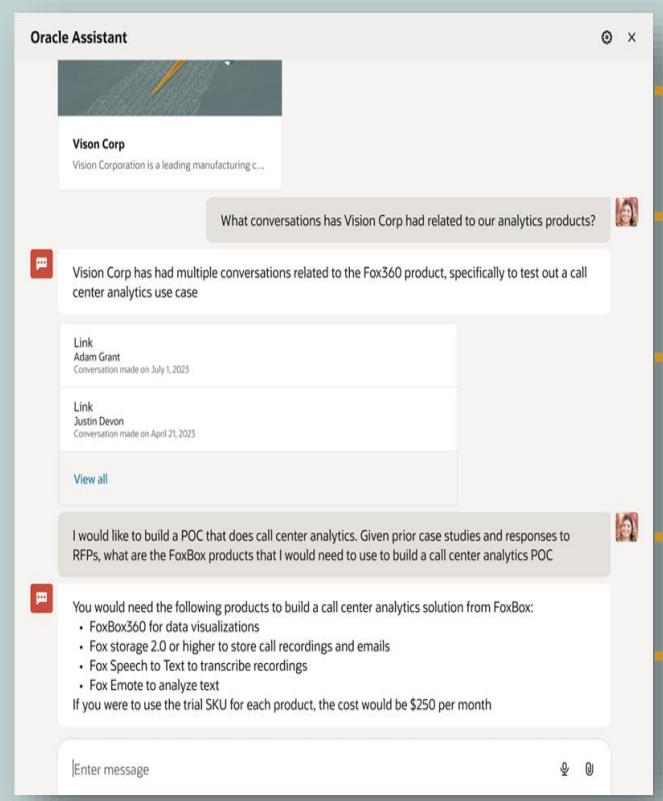
This functionality is currently available in private preview

NEW

소개 : OCI Generative AI Agents

RAG Agent : Gen AI Agents의 첫번째 서비스

RAG를 지원하기 위한 Agent 서비스



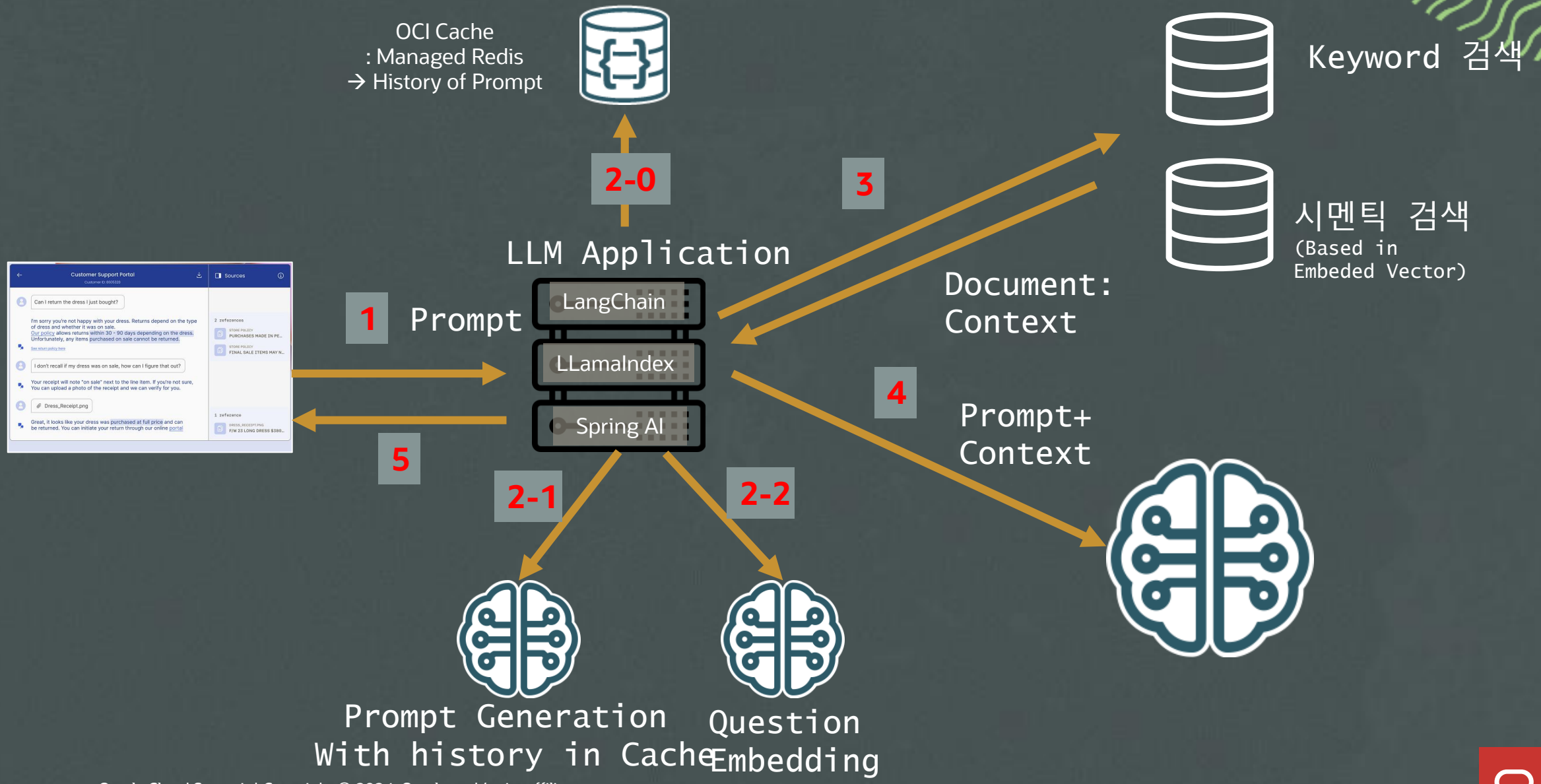
- OCI GenAI RAG Agent connected to Vision Corp's knowledge bases
- Analyst asks a natural language question
- GenAI RAG Agent responds in humanlike manner and provides links to relevant source documents
- Analyst asks a follow up question
- GenAI RAG Agent uses chat history and further information retrieval to respond

RAG Agent : Beta release – January 2024

Beta:
OCI OpenSearch

Coming Soon:
Oracle Database 23c AI Vector Search
MySQL HeatWave Vector Store

RAG(Retrieval Augmented Generation) 아키텍처: Before OCI Gen AI Agent

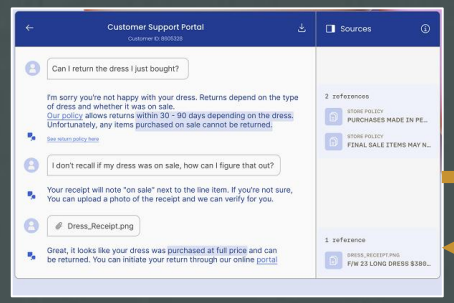


RAG(Retrieval Augmented Generation) 아키텍처: Applying OCI Gen AI Agent

OCI Cache
: Managed Redis
→ History of Prompt

2-0

LLM Application



1 Prompt

LangChain
LLamaIndex
Spring AI

3

4

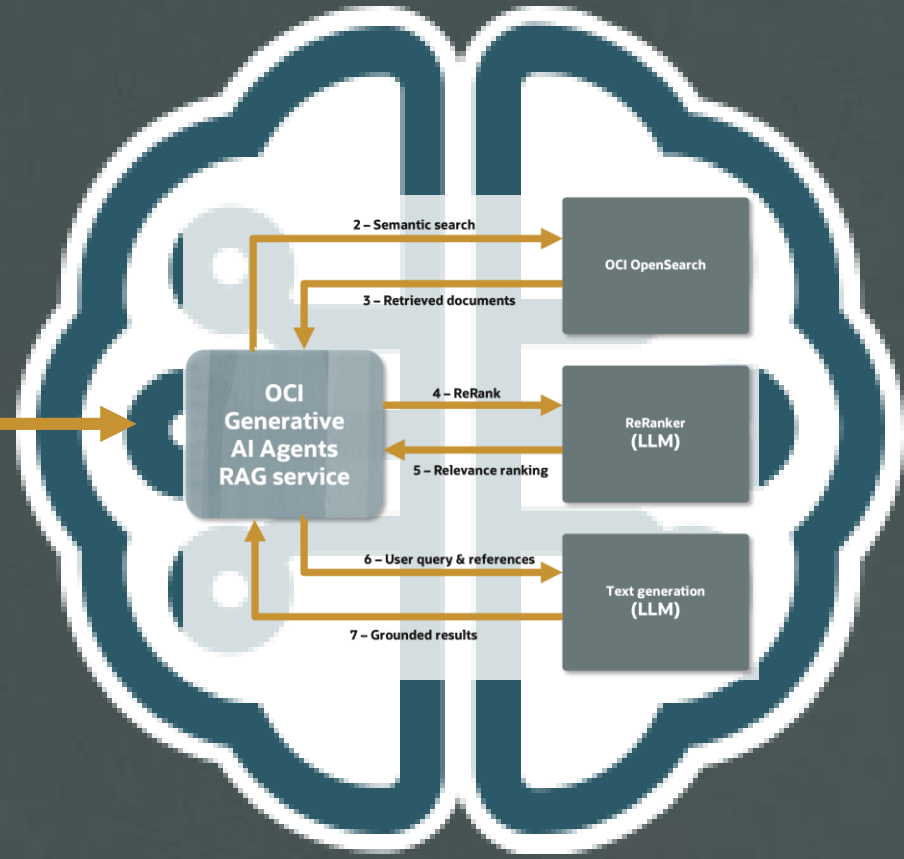
Redefined Prompt

2-1

2-2

Prompt Generation
With history in Cache

Question Embedding



NEW

OCI Generative AI Agent 로드맵

MySQL HeatWave

23c ORACLE Database

OpenSearch

RAG with OpenSearch in Beta Jan 2024

정보 수집
OCI Gen AI Agent 현재 베타 상태이며 데이터 저장소로 OCI OpenSearch 지원

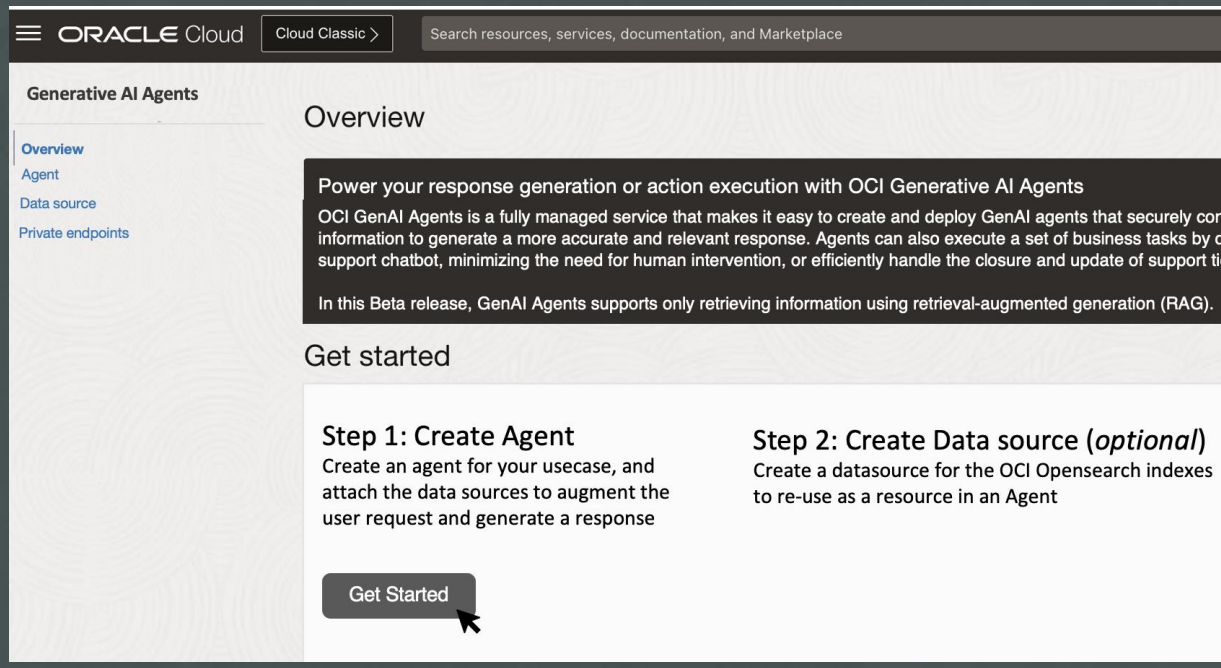
2024 상반기 Oracle Database 23c AI Vector Search와 MySQL HeatWave Vector Store 지원 예정

에이전트 작동 방식
OCI Gen AI Agent는 관련 정보를 조회하고 프롬프트의 컨텍스트에 설정하는 일련의 작업을 추상화 및 자동화

고성능 생성형 모델과 통합
OCI Gen AI Agent는 OCI Generative AI 서비스 모델(Cohere 및 Llama-2)을 기본 지원

Reasoning & Planning
OCI Gen AI Agent는 ReAct framework을 통해서 일련의 생각, 행동, 관찰을 바탕으로 추론, 계획 및 동작

멀티-턴 에이전트
과거 요청 기록을 유지, 모델 컨텍스트와 응답을 재정의하여 명확성 향상

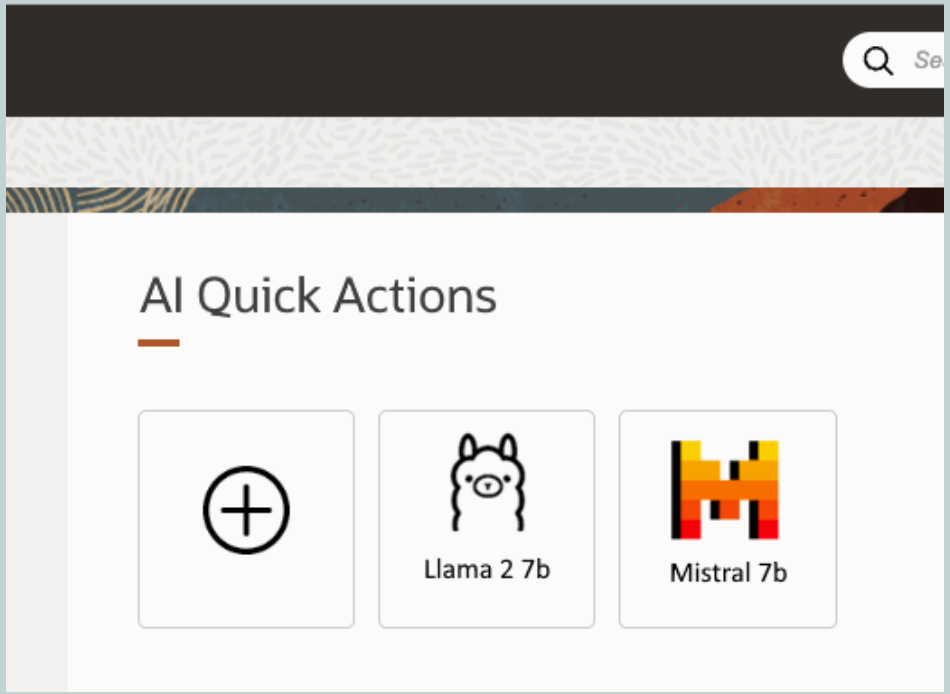


NEW

Announcing: AI Quick Actions for OCI Data Science

Quick actions:

- Deploy
- Fine tune
- Integrate
- Scale



- ✓ Data Science 노트북 환경 내의 사용자 인터페이스에서 버튼 클릭만으로 호출할 수 있는 사용 사례 모음
- ✓ Data Science 노트북과의 원활한 통합을 통해 Llama2 및 Mistral 7B와 같은 LLM에 코드 없이 액세스할 수 있음

Deployment

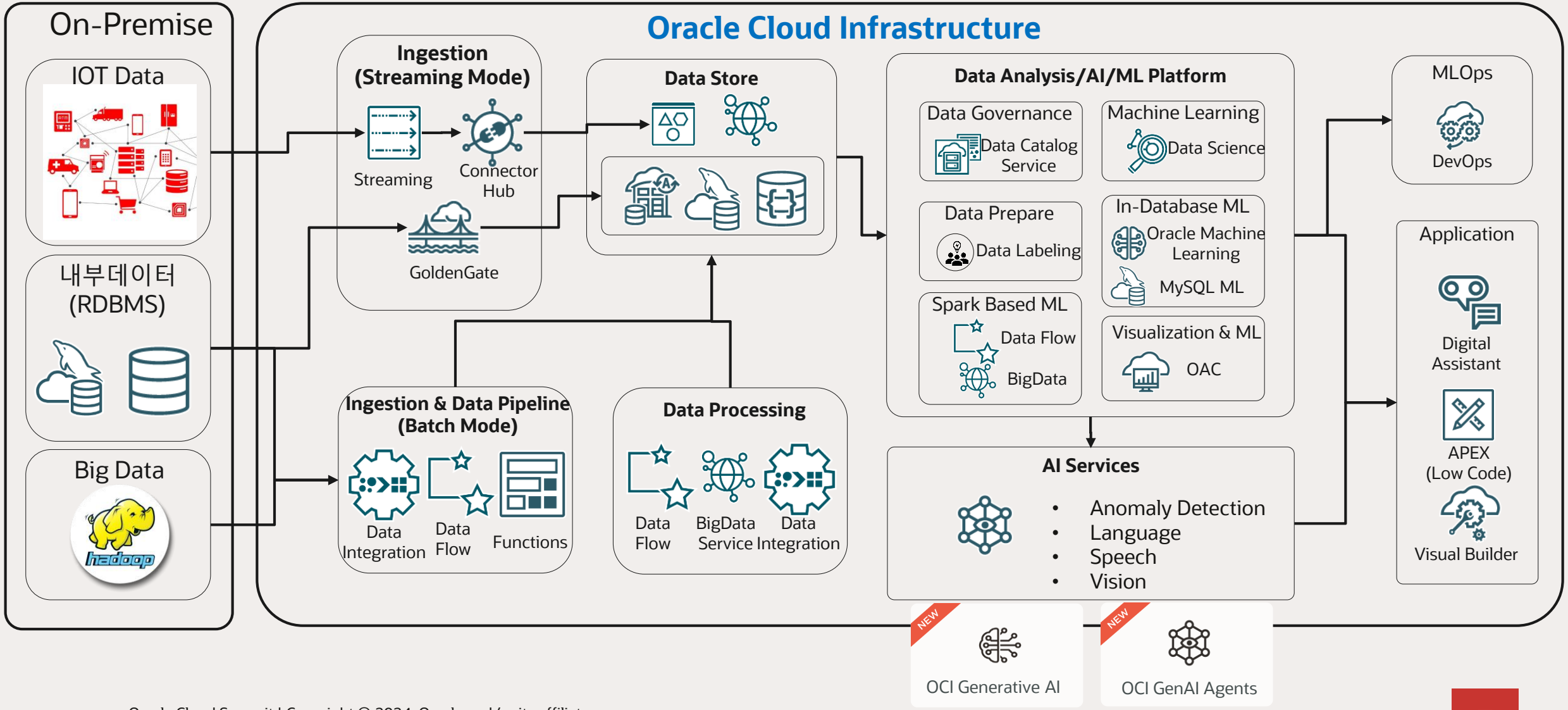
- Support for model deployment using Text Generation Inference (Hugging Face), vLLM (UC Berkeley) and Nvidia Triton serving with public examples for:
  - Llama2 7b and 13b using A10s
  - Llama2 70b using A100 and A10s via GPTQ Quantization
  - Mistral 7b
  - Jina Embedding Model using A100

Fine Tuning

- Distributed Training with PyTorch, Hugging Face Accelerate and DeepSpeed for Fine-tuning of LLM
- Mount for Object Storage and File System as a Service – enables effortless checkpointing and storage of fine-tuned weights
- Service-provided Condas, eliminates the requirement for custom docker environments

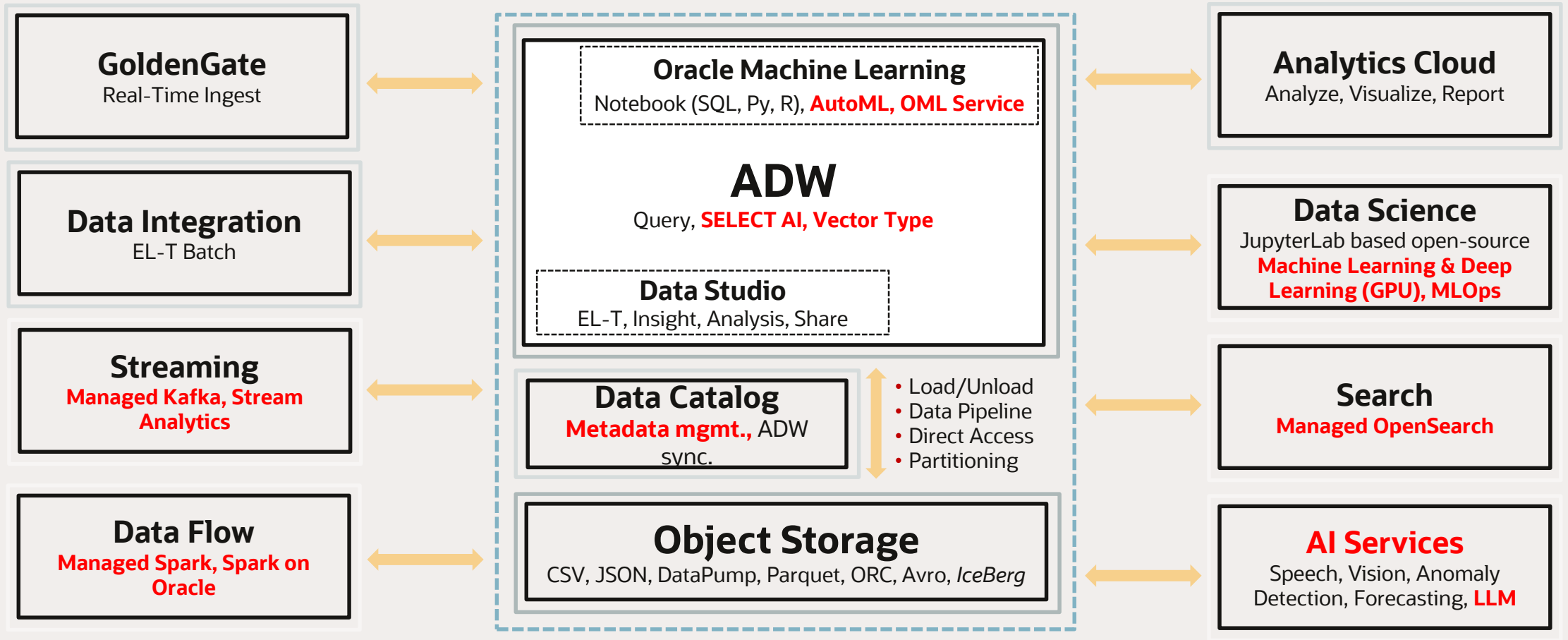
오라클은 클라우드 상에서 데이터 플랫폼을 제공해 드립니다

온프레미스 및 타 클라우드와의 실시간 연결을 통해 다양한 데이터의 수집/저장/분석/배포에 이르는 데이터 파이프라인을 제공



Modern Data Platform이 제공하는 세부적인 데이터 관리 및 AI 지원 기능

오라클은 end-to-end data pipeline 안에 ML, Gen AI, RAG를 embed 하여 data와 AI seamless integration 지원



Enterprise AI/Data Platform 구성 시 고려 사항

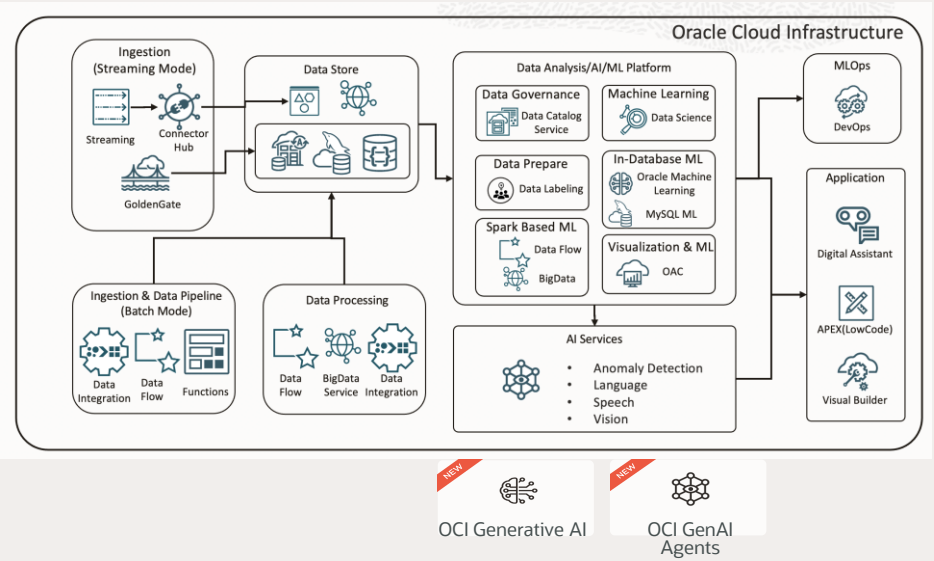
- AI 고려 사항
  - ✓ 대규모의 GPU 인프라
  - ✓ LLM 선택 및 프로세스 관리 서비스
  - ✓ 균형 잡힌 AI 시스템 설계
    - Gen AI: 자연어 입력 및 추론 지원/컨텐츠 생산, ML: 수치 예측 및 데이터 분류, DL: 멀티미디어 데이터의 탐지 및 분류
    - ML / DL / Gen AI의 balanced AI architecture 구성 필요

- Data 고려 사항
  - ✓ End-to-end data pipeline 구성
    - 데이터의 수집/저장/가공/학습/분석/Aging 까지의 모든 단계를 유연하게 지원해야 함
  - ✓ 정형/비정형 데이터를 아우르는 data lakehouse 구성
    - Data Lakehouse는 전사 핵심 분석 시스템일 뿐만 아니라 Vector DB를 통한 Gen AI 활용의 핵심 저장소이므로 비용효율적인 아키텍쳐로 설계 필요
  - ✓ AI와의 연계 방안 정립 : fine-tuning dataset 구성

- AI / Data Integration 고려 사항
  - RAG 구성
    - ✓ 기존 DB와의 연계 및 Vector DB 구성 필요
      - Data lakehouse와 Vector DB 사이의 유기적인 연계 방안 필요
      - Vector Data 증가에 대비하여 성능과 안정성을 고려한 아키텍쳐 구성 필요
    - ✓ RAG 구성 시 컨텐츠 뿐만 아니라 보안 관련 프로세스가 중요
      - Vector DB는 회사의 주요 지식과 경험을 보관할 핵심 지식 저장소
      - Vector DB의 중요 컨텐츠에 대한 세심한 접근 제어 및 권한 관리 중요
  - Agent 구성
    - ✓ 전체 AI 프로세스를 지원할 AI Agent의 구성

오라클 모던 데이터 플랫폼이 통합 AI/Data 플랫폼으로서 제공하는 가치

주요 가치 : Modern Data Platform with in-DB M/L and Gen AI engineering



- ✓ 데이터 관리의 end-to-end management platform을 제공
- ✓ On Prem과 클라우드에서 동일한 기술로 Data Lakehouse를 distributed & integrated architecture로 구축할 수 있는 유연성/통합성 제공
- ✓ Converged DB는 모든 유형의 데이터를 통합 관리할 수 있는 all-in-one DB infra 제공 ➔ Vector Type을 이용한 Vector DB 구성 지원
- ✓ Autonomous DB는 누구나 쉽게 자신만의 DB를 운영할 수 있는 편리성 제공 (관리 부담 일체 없음) ➔ NL2SQL을 통해 편리성 극대화 지원
- ✓ RAG Agent를 통해 기업 수준의 private Gen AI 인프라로 확장 가능
- ✓ In-DB M/L을 통해 데이터 입력-분석-예측 작업을 모두 Oracle DB 상에서 수행 가능 ➔ well-balanced AI 통합 플랫폼 구축 지원
- ✓ Data Science는 data scientist들이 쉽게 M/L modeling & deployment 작업을 관리할 수 있는 환경 제공 ➔ Quick Action도 지원

Summary

오라클의 생성형 AI 지원 전략과 기업 데이터 플랫폼과의 통합 방안

- 오라클의 생성형 AI 지원 전략
  - ✓ AI Infrastructure : NVIDIA와의 협력을 통한 GPU SuperCluster 지원
  - ✓ App embedding with Gen AI : Fusion Application에 Gen AI 내재화
  - ✓ Gen AI Service : SOTA LLM(Llama2,Cohere)을 활용한 fine-tuning 및 inference의 end-to-end process 지원
  - ✓ Vector type in Oracle DB 23c : Oracle DB의 Vector DB로의 활용 지원 → 기존 투자 활용, 쉬운 구축, 성능/안정성
  - ✓ RAG Agent : 기업 데이터 플랫폼과 결합된 RAG 구축 지원
- 생성형 AI와 기업 데이터 플랫폼과의 통합 방안
  - ✓ Oracle Modern Data Platform with in-DB ML and Gen AI
  - ✓ Data Platform과 생성형 AI의 유기적인 연계를 통한 기업 수준의 통합된 AI / Data platform 지원


ORACLE