

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

고성능 사전 학습 모델
Llama2 및 Cohere LLM을 FM로 제공

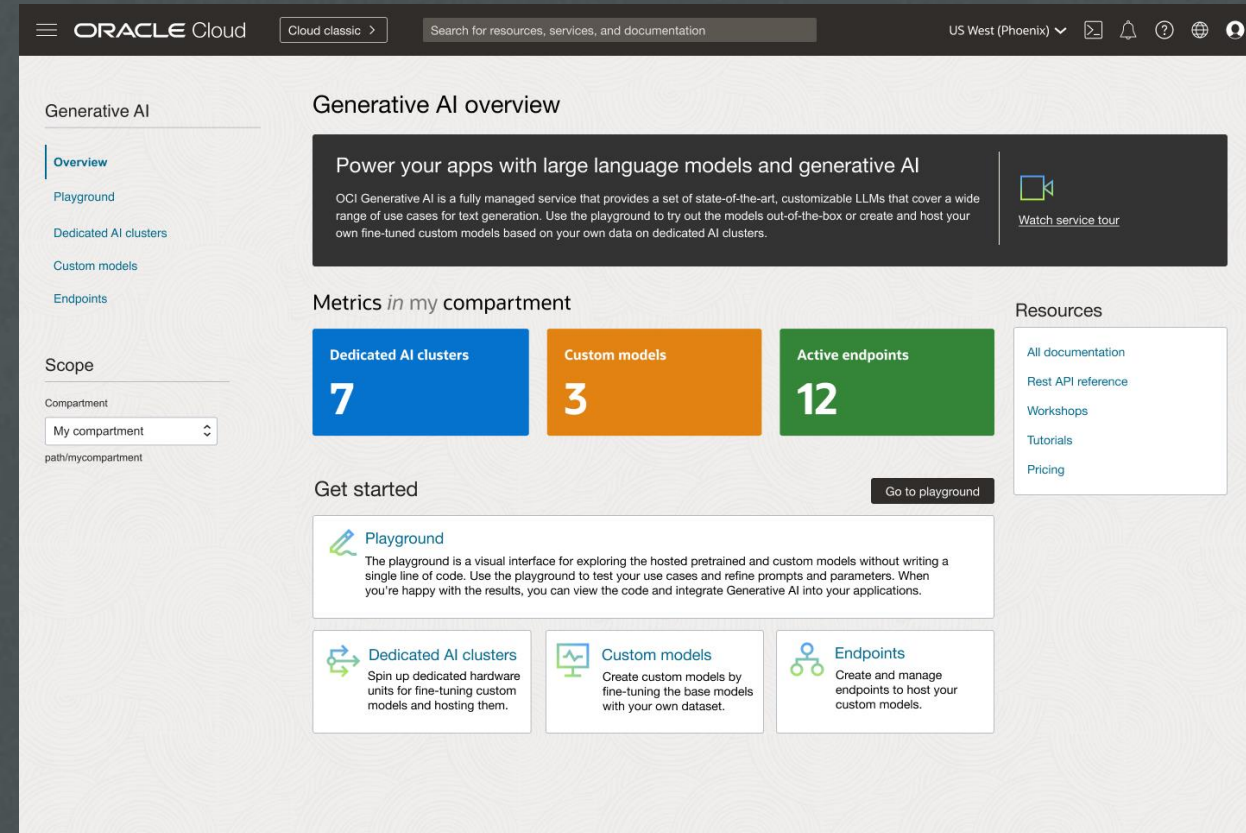
Fine-tuning 지원
자체 데이터로 사용자 전용 모델로 Fine-Tuning하여
특정 문제를 해결하거나 특정 도메인을 최적화

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

GA(Generally Available) : 1/24



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



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 요약물을 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)

생성

✓ cohere.command v15.6
cohere.command-light v15.6
meta.llama-2-70b-chat

요약

cohere.command v15.6

포함

cohere.embed-english-light-v2.0
cohere.embed-english-light-v3.0
cohere.embed-english-v3.0
cohere.embed-multilingual-light-v3.0
cohere.embed-multilingual-v3.0

OCI Generative AI 서비스: Playground

Generation: Document Transformation

The screenshot shows the OCI Generative AI Playground interface. The left sidebar has tabs for Overview, Playground (selected), Dedicated AI clusters, Custom models, Endpoints, and Scope. The main area is titled "Generative AI Playground" and includes instructions: "To get started, choose a model and a preset prompt example. Then, refine the prompts and parameters to fit your use cases. See [model types](#) for more information."

The "Model" dropdown is set to "cohere.command". The "Example" dropdown is set to "Rewrite instructions with steps". The "Input" field contains the text: "Rewrite the following steps as a numbered list: First, reboot the computer to make sure you have a clean memory footprint. Next, click the main menu, and select settings. In the search box of the settings app, type 'updates' and click search. Select the search results titled 'Software updates'. In the Software updates screen, click the 'Check for updates' link. If no updates are available you're done. If updates are available for your computer, click 'Proceed with this update' link. Read the update description and confirm the installation by clicking 'Confirm'. The installation might take several minutes and might require the computer to restart several times during the installation. If restart happens, you are asked to login to the computer after each restart. After the process completes, you get a summary of the installed software."

The "Output" field shows the transformed text: "1. Reboot the computer to make sure you have a clean memory footprint
2. Click the main menu, and select settings
3. In the search box of the settings app, type 'updates' and click search
4. Select the search results titled 'Software updates'
5. In the Software updates screen, click the 'Check for updates' link
6. If no updates are available, you're done
7. If updates are available for your computer, click 'Proceed with this update' link
8. Read the update description and confirm the installation by clicking 'Confirm'
9. The installation might take several minutes and might require the computer to restart several times during the installation
10. If restart happens, you are asked to login to the computer after each restart
11. After the process completes, you get a summary of the installed software"

Summarization

The screenshot shows the OCI Generative AI Playground interface. The left sidebar has tabs for Overview, Playground (selected), Dedicated AI clusters, Custom models, Endpoints, and Scope. The main area is titled "Generative AI Playground" and includes instructions: "To get started, choose a model and a preset prompt example. Then, refine the prompts and parameters to fit your use cases. See [model types](#) for more information."

The "Model" dropdown is set to "cohere.command". The "Example" dropdown is set to "Summarize a blog post". The "Input" field contains the text: "트랜스포머(Transformer)는 버트(BERT) 및 GPT-3와 같이 오늘날 널리 사용되는 언어 모델의 종류이다. 처음에는 자연어처리(NLP)를 위해 개발되었지만 컴퓨터 비전, 의료 발견 등 한 산업에서 활동해 혁신을 가속하고 있다. 예를 들어, 은행 금융 서비스 및 보험 회사에서는 고객을 유치하고 비용을 절감하며, 수익을 높이고 변화하는 규칙과 규정 등을 준수해야 한다. 이에 솔루션은 기업의 정보 검색, 의료 분석, 고객 서비스 등과 관련된 프로세스 자동화 및 응용 프로그램을 효율적으로 수행하고 그에 따른 위험을 완화하도록 돕는다. 트랜스포머 모델은 문장 속 단어와 같은 순차 데이터 내의 관계를 추적해 맥락과 의미를 학습하는 신경망을 의미한다. 양방향(attention) 또는 셀프어텐션(self-attention)이라 불리며, 인공 지능을 구축하는 수학적 기법을 응용해 서로 떨어져 있는 데이터 요소를 의미와 관계에 따라 비교하게 할 수 있는 부분까지 감지한다. 구글(Google)의 2017년 논문(Attention is All You Need)에 처음 등장한 트랜스포머는 지금까지 개발된 모델 중 가장 사용과 강력한 성능을 보여주며, '트랜스포머 AI'라 불리는 혁신적인 생태계의 혁신을 주도하고 있다. 여기에 구글은 2020년에 대규모 이미지 인식 위한 트랜스포머(API) 모델을 발표(Transfomers for Image Recognition at Scale-다윈)하기도 했다."

The "Output" field shows the summarized text: "Transformer is the core of today's popular language models, such as BERT and GPT-3. It was originally developed for natural language processing, but is now being used in various fields, such as computer vision and drug discovery. Transformers can be used to automate compliance processes, improve customer service, and reduce risk. It is a self-attention model that learns the relationship between sequential data, such as words in a sentence, and their meanings. Google's 2017 paper 'Attention is All You Need' introduced the first transformer, which has since become one of the most popular and powerful machine learning models. Google has also developed a transformer for image recognition (API) and announced it in 2020."

Document Embedding(D: 1024)

The screenshot shows the OCI Generative AI Playground interface. The left sidebar has tabs for Overview, Playground (selected), Dedicated AI clusters, Custom models, Endpoints, and Scope. The main area is titled "Generative AI Playground" and includes instructions: "To get started, choose a model and a preset prompt example. Then, refine the prompts and parameters to fit your use cases. See [model types](#) for more information."

The "Model" dropdown is set to "cohere.embed-english-light-v2.0". The "Example" dropdown is set to "Generate a list of sentences or phrases to generate embeddings (maximum of 96 inputs)". The "Input" field contains a list of sentences: "1. 방탄소년단은 BTS라고도 불리고, Hve 소속이다.", "2. 콘세라원은 4세대 대표 K-Pop 걸그룹이다.", "3. 콘세라원은 Hve 소속이다.", "4. 제원은 프세라임 라이다.", "5. 수지는 넷플릭스 드라마 미워나의 주연 배우다.", "6. 아이유는 여석재에 출연한 가수 겸 연기자다.", "7. 수지는 미워나에서 온화한 걸그룹 스타를 연기했다."

The "Output" field shows the generated embeddings as a grid of numbers. The "Character count" is 159. The "Tokens limit for each input" is 512.

Code Generation

The screenshot shows the OCI Generative AI Playground interface. The left sidebar has tabs for Overview, Playground (selected), Dedicated AI clusters, Custom models, Endpoints, and Scope. The main area is titled "Generative AI Playground" and includes instructions: "To get started, choose a model and a preset prompt example. Then, refine the prompts and parameters to fit your use cases. See [model types](#) for more information."

The "Model" dropdown is set to "cohere.embed-english-light-v2.0". The "Example" dropdown is set to "Generate a list of sentences or phrases to generate embeddings (maximum of 96 inputs)". The "Input" field contains a list of sentences: "1. 방탄소년단은 BTS라고도 불리고, Hve 소속이다.", "2. 콘세라원은 4세대 대표 K-Pop 걸그룹이다.", "3. 콘세라원은 Hve 소속이다.", "4. 제원은 프세라임 라이다.", "5. 수지는 넷플릭스 드라마 미워나의 주연 배우다.", "6. 아이유는 여석재에 출연한 가수 겸 연기자다.", "7. 수지는 미워나에서 온화한 걸그룹 스타를 연기했다."

The "Output" field shows the generated code snippet for integrating the model into an application. The code is in Python and includes comments for setup, authentication, and usage.



OCI Generative AI 서비스: 전용 모델

Create dedicated AI cluster

Dedicated AI clusters can take a few minutes to create. After a cluster is in an active state, you can use it for fine-tuning or hosting workloads.

Compartment

genaiusers (root)

Name *Optional*

HR_Query

Description *Optional*

Chat agent to provide responses to all the organizational HR questions.

Cluster type

Hosting - beta

Fine-tuning - beta

☒

 I agree that I will use this dedicated AI cluster only for beta testing. At the end of my beta testing, I will delete this dedicated AI cluster, or it will be deleted on my behalf if I do not take action.

Create

Cancel

ORACLE Cloud

Search resources, services, documentation, and Marketplace

US Midwest (Chicago)

B

Bucket Information

Tags

General

Features

Resources

Objects

Namespace: axk4z7krhqtx

Compartment: [genaiusers](#)

Created: Thu, Aug 31, 2023, 16:08:56 UTC

ETag: 093eab27-5c55-446e-80e1-0012d028819e

OCID: ...xycuh76q [Show](#) [Copy](#)

Usage

Approximate Object Count: 4 objects

Approximate Size: 86.42 KiB

Uncommitted Multipart Uploads Approximate Count: 0 uploads

Uncommitted Multipart Uploads Approximate Size: 0 bytes

Features

Default Storage Tier: Standard

Visibility: Private

Encryption Key: Oracle managed key [Assign](#)

Auto-Tiering: Disabled [Edit](#)

Emit Object Events: Disabled [Edit](#)

Object Versioning: Disabled [Edit](#)

Objects

Metrics

Pre-Authenticated Requests

Work Requests

Lifecycle Policy Rules

Replication Policy

Upload

More Actions

Search by prefix

	Name	Last Modified	Size	Storage Tier
<input type="checkbox"/>	<input type="checkbox"/> small_data_dup_records.jsonl	Thu, Aug 31, 2023, 16:15:01 UTC	21.83 KiB	Standard
<input type="checkbox"/>	<input type="checkbox"/> small_data_dup_records_1.jsonl	Thu, Aug 31, 2023, 16:27:55 UTC	21.83 KiB	Standard
<input type="checkbox"/>	<input type="checkbox"/> small_data_wrongid_records.jsonl	Thu, Aug 31, 2023, 16:45:18 UTC	21.83 KiB	Standard



OCI Generative AI 서비스: LLM 모델 Fine-Tuning

ORACLE Cloud

Search resources, services, documentation, and Marketplace

US Midwest (Chicago)

Create model

1 Model definition

2 Fine-tuning configuration

3 Data selection

Fine-tuning configuration

Define the model type, dedicated AI cluster type and hyperparameters for this specific model.

1 Models of different categories have different cluster hardware requirements for fine-tuning. The dedicated AI cluster drop-down list is filtered to show clusters that are compatible in size with the requirements of the selected base model.

Base model

cohere.command-light.15.6

Fine-tuning method

T-Few

Dedicated AI cluster in **genaiusers (root)**

generativeaidedicatedaicluster20240115214815

Create a new dedicated AI cluster

Advanced options

Hide hyperparameters

Total training epochs

3

Enter 1 or a higher integer.

Learning rate

0.01

Enter a number that's between 0 and 1.0.

Training batch size

16

Enter 1 or a higher integer. Higher numbers use more GPU to give better results.

Early stopping patience

6

Enter 1 or a higher integer for the grace period of the evaluation cycle before the early stopping stops the training. Enter 0 to disable early stopping.

Early stopping threshold

0.01

Enter 1 or a higher integer for the grace period of the evaluation cycle before the early stopping stops the training. Enter 0 to disable early stopping.

Log model metrics interval in steps

10

Enter an integer between 1 and the total number of training steps to enable logging of the training loss, and enter 0 to disable logging.

Restore defaults

Previous

Next

Cancel

ORACLE Cloud

Search resources, services, documentation, and Marketplace

US Midwest (Chicago)

Create model

1 Model definition

2 Fine-tuning configuration

3 Data selection

Data selection

1 Ensure that each file contains at least 32 examples. Learn about [training data requirements](#)

Choose bucket in **genaiusers (root)**

afshin-test

Training file

small_data_dup_records.jsonl

If your file is not listed, you might not have "read" permission to the selected bucket.

Data preview

This section shows how the data is ingested for the first few records of the training file. If the data records aren't parsing correctly, check the file format and ensure that the data meets the [requirements](#).

Warning: only found 7 examples (25 incorrectly encoded, 0 duplicated), at least 32 valid examples required.

Refresh Data

prompt	completions
I am a salesman at Oracle who wants to start with Oracle's previous featured success stories to find new business opportunities. Use Case: I plan to speak to a prospect in the IT consulting industry. This prospect is a growing IT consultancy firm based in Asia-Pacific, providing tailored solutions to clients across various industries. I am going to speak to their head of operations. Related Oracle Success Story: Inoapps turns to Oracle Cloud ERP and HCM to reimagine business operations. Description: Inoapps is a global IT consultancy and Oracle Platinum Partner that delivers innovative solutions to clients across various industries. Inoapps wanted to transform its own business operations and enhance its employee experience by moving to a cloud-based platform. Inoapps chose Oracle Cloud ERP and Oracle Cloud HCM to replace its legacy systems and streamline its finance, procurement, project management, and human resources processes. With Oracle Cloud, Inoapps has improved its visibility and control over its financial performance, reduced its operational costs and risks, increased its productivity and efficiency, and empowered its employees with self-service capabilities and career development tools. ##	Sales pitch: Subject: Streamline Your IT Consultancy's Operations with Oracle Cloud ERP and HCM Dear [Head of Operations], As a growing IT consultancy firm, the efficiency and effectiveness of your operations play a critical role in your success. I recently learned about Inoapps, a global IT consultancy and Oracle Platinum Partner, that transformed its business operations and enhanced its employee experience using Oracle Cloud ERP and Oracle Cloud HCM. By moving to Oracle Cloud, Inoapps replaced its legacy systems, streamlining finance, procurement, project management, and human resources processes. This resulted in improved visibility and control over financial performance, reduced operational costs and risks, increased productivity and efficiency, and empowered employees with self-service capabilities and career development tools. I believe that your IT consultancy firm can benefit from implementing Oracle Cloud ERP and HCM, just as Inoapps did. I would like to propose a meeting to discuss how these cloud-based solutions can help you optimize your operations and create a better working environment for your employees. Please let me know when you are available for a conversation, and I will arrange a meeting with our product specialists to provide more insights into the benefits of Oracle Cloud ERP and HCM for your organization. Looking forward to discussing this opportunity with you. Best regards, [Your Name] [Your Title] [Your Contact Information] END
I am a salesman at Oracle who wants to start with Oracle's previous featured success stories to find new business opportunities. Use Case: I plan to speak to a prospect in the food and beverage industry. This prospect is a rapidly expanding restaurant chain in Latin America with ambitious growth plans. I am going to speak to their head of IT. Related Oracle Success Story: Oracle Cloud provides key ingredients to sustaining growth at Madero Restaurants. Description: Madero Restaurants is a leading casual dining chain in Brazil, with more than 200 locations and 15,000 employees. Madero wanted to support its rapid growth and expansion plans by modernizing its IT infrastructure and applications. Madero selected Oracle Cloud Infrastructure (OCI), Oracle Cloud ERP, Oracle Cloud HCM, and Oracle Cloud CX to create a unified and scalable cloud platform. With Oracle Cloud, Madero has improved its operational efficiency and agility.	Sales pitch: Subject: Scale Your Restaurant Chain with Oracle Cloud: Enhanced Efficiency, Customer Experience, and Workforce Enablement Dear [Head of IT], I understand that your restaurant chain is experiencing rapid growth in Latin America, and I wanted to share a success story from Madero Restaurants, a leading casual dining chain in Brazil. Madero modernized its IT infrastructure and applications with Oracle Cloud to support its expansion plans and improve its business operations. By implementing Oracle Cloud Infrastructure (OCI), Oracle Cloud ERP, Oracle Cloud HCM, and Oracle Cloud CX, Madero created a unified and scalable cloud platform that improved operational efficiency and agility, optimized supply chain and inventory management, enhanced customer service and loyalty, and enabled its workforce with mobile and collaborative tools. I believe that your restaurant chain can achieve similar success by leveraging Oracle Cloud solutions. I would like to propose a meeting to discuss how these innovative technologies can help you streamline your operations.

Previous

Submit

Cancel

11 Oracle Cloud Summit | Copyright © 2024, Oracle and/or its affiliates

OCI Generative AI 서비스: LLM 모델 Fine-Tuning

ORACLE Cloud

Search resources, services, documentation, and Marketplace

US Midwest (Chicago)

Overview

Endpoints

Acme_Company_HR_QA

View in playground

Add tags

Edit

Delete

Endpoint Information

Tags

Compartment OCID: ...jm3kucc7lotk2oqep47q

Model name: generativeaimodel20230911152851

OCID: ...ktexjey43qu2ficnd2wa

Model version: v20230911152851

Description:

Dedicated AI Cluster: ocid1.generativeaidedicatedcluster.oc1.us-chicago-1.amaaaaaacqy6p4qastho3j5jnhvxpqnv2b6hrcwzplwua2xguz7ro7kgma

State: Creating

Lifecycle details:

ORACLE Cloud

Search resources, services, documentation, and Marketplace

US Midwest (Chicago)

Edit

Add tags

Delete

Custom models can take several hours to fine-tune and become active. After a custom model is in an active state, you must create an endpoint for that model to host it for inference.

Model details

Fine-tuning configuration

Tags

Compartment: ...tk2oqep47q

Base model: ocid1.generativeaimodel.oc1.us-chicago-1.amaaaaapi24rzaalcbnpexqvkxpfdcqskcupap27qtr7ew5ierjpbih2a

OCID: ...yqdpoc66zq

Capability: Text generation

Description: -

Category: Small

State: Active

Dedicated AI cluster: ocid1.generativeaidedicatedcluster.oc1.us-chicago-1.amaaaaaacqy6p4qastho3j5jnhvxpqnv2b6hrcwzplwua2xguz7ro7kgma

Lifecycle details: Created Model

Dataset: sales_pitch_generation_train.jsonl

Created: Fri, 15 Sep 2023 17:42:24 UTC

Model performance

Accuracy

Loss

0.73396

1.09664

ORACLE Cloud

Search resources, services, documentation, and Marketplace

US Midwest (Chicago)

Generative AI

Generative AI Playground

Overview

Playground

Dedicated AI clusters

Custom models

Endpoints

Model

cohere.command

View model details

Example

Generate an email

View code

As a corporate vice president, generate an email congratulating a team that has just shipped a new cloud service. Emphasize the great positive impact the new service will have on the productivity of their customers.

Hi Team,

I am so proud to announce that we have just shipped our new cloud service! This has been a huge undertaking and you all have worked so hard to make it happen. I want to thank you from the bottom of my heart for your dedication and hard work.

This new cloud service is going to have a huge positive impact on the productivity of our customers. It will allow them to work from anywhere in the world and access their files and data from any device. This is going to be a game changer for our company and for our customers.

Thank you again for all your hard work and dedication. I am so proud to be a part of such a great team.

Best,

[Your Name]

Parameters

Maximum output tokens

Temperature

Top p

Top k

Stop sequences

Frequency penalty

Presence penalty

Show likelihoods

Generate

Regenerate

Clear

Character count - 876 | Token limit - 4000

Copyright © 2023, Oracle and/or its affiliates. All rights reserved.



OCI Generative AI 서비스: LLM 모델 관리

ORACLE Cloud

Search resources, services, documentation, and Marketplace

Generative AI

Overview

Playground

Dedicated AI clusters

Custom models

Endpoints

Scope

Compartment

genaiusers (root)

Generative AI overview

Power your apps with large language models and generative AI

OCI Generative AI is a fully managed service that provides a set of state-of-the-art, customizable LLMs that cover a wide range of use cases for text generation. Use the playground to try out the models out-of-the-box or create and host your own fine-tuned custom models based on your own data on dedicated AI clusters.

Metrics in genaiusers (root) Compartment

Dedicated AI clusters

7

Custom models

6

Endpoints

10

Get started

Playground

The playground is a visual interface for exploring the hosted pretrained and custom models without writing a single line of code. Use the playground to test your use cases and refine prompts and parameters. When you're happy with the results, you can view the code and integrate Generative AI into your applications.

Go to playground

Dedicated AI clusters

Spin up dedicated hardware units for fine-tuning custom models and hosting them.

Custom models

Create custom models by fine-tuning the base models with your own dataset.

Endpoints

Create and manage endpoints to host your custom models.

ORACLE Cloud

Search resources, services, documentation, and Marketplace

US Midwest (Chicago)

custom-cohere-light-endpoint

View in playground

Edit

Add tags

Move endpoint

Delete

Endpoint Information

Tags

Compartment OCID: ...gm3kuc70k92oqep47q

Model name: cohere.command-light

OCID: ...yluztfo7behwtoria

Model version: 15.6

Description:

Created on: Mon, 15 Jan 2024 21:50:41 UTC

State: Active

Created by: laleh.haghshehass@oracle.com

Lifecycle details: Created Endpoint

Content moderation: Enabled

Dedicated AI Cluster: ...oid1.generativeai.dedicatedcluster.oc1.us-chicago-1.amaaasasay6p6kamtakstsch3k2c4qabxtkum2457v2vof44c3a

Resources

Endpoint metrics

Start time

End time

Quick selects

Reset charts

Work requests

Interval

Statistic

Count

Total processing time

Number of calls

Service Errors Count

Client Errors Count

Total number of input

Total number of output

13 Oracle Cloud Summit | Copyright © 2024, Oracle and/or its affiliates

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**

"What's the policy?"

Prompt

+

Chat History

Enhanced Prompt

"What's my corporate 401k policy?"

Embedding Model

embedding
embedding

Similarity Search



Vector Database

Vector ID
Matches

Augmented Prompt



Private
content

Fetch docs for matching IDs



Relational Database

LLM

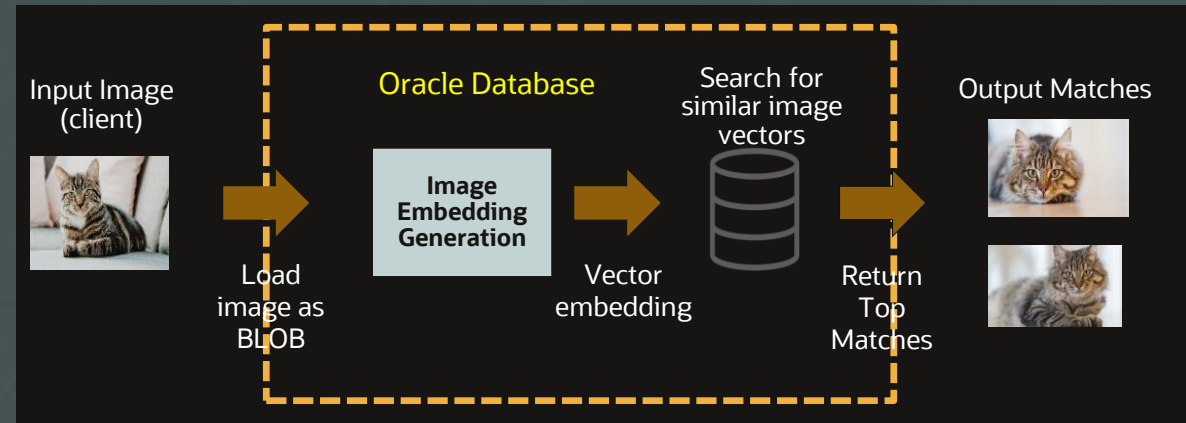
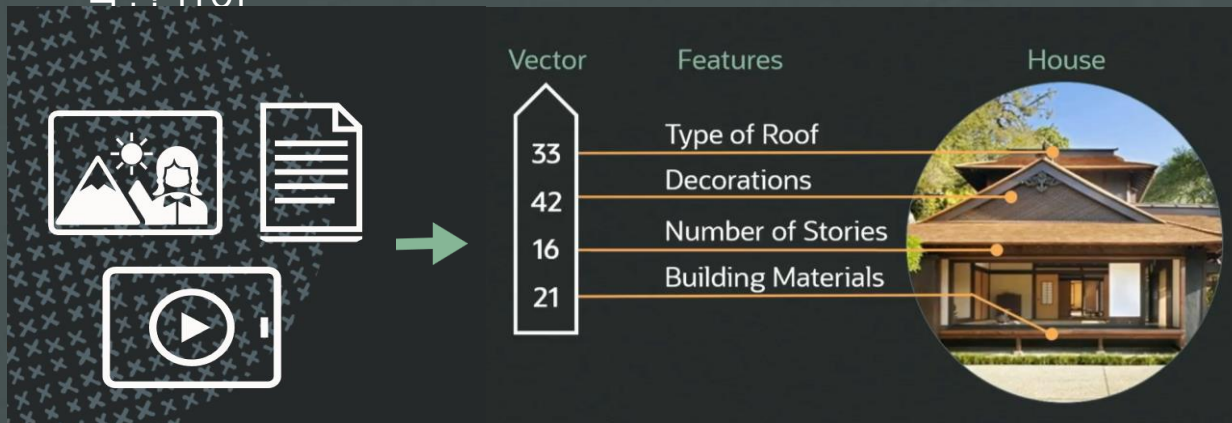
RESPONSE

Enhanced prompt + policy details +
scenario examples + known concerns +
expected policy changes...

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

Vector Search: 오라클 DB 내에서 Vector Type 제공, Oracle DB 내에서의 Vector DB 구성의 장점

- 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
- 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**

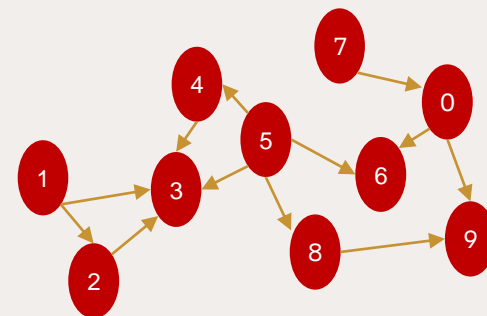
- 벡터들은 정확성을 위해 유사성을 기반으로 군집 후 연결되어짐
- 속도를 위해 정확도를 제한한 greedy search 기법 사용

Vector indexes

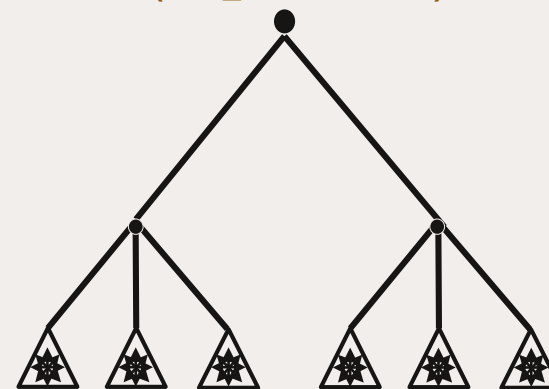
- **In-Memory Neighbor Graph Vector Index** – 노드는 벡터를 나타내고 노드 사이의 에지는 유사도를 나타내는 그래프 기반 인덱스. 정확성과 속도 모두에 있어 매우 효율적임
- **Neighbor Partition Vector Index** – 유사성을 기준으로 테이블 파티션으로 클러스터링된 벡터가 포함된 파티션 기반 인덱스. 빠르고 원활한 트랜잭션 지원을 갖춘 효율적이고 확장이 가능한 인덱스

오라클은 2가지 타입의 vector index를 모두 지원

Graph Vector Index
(HNSW Index)

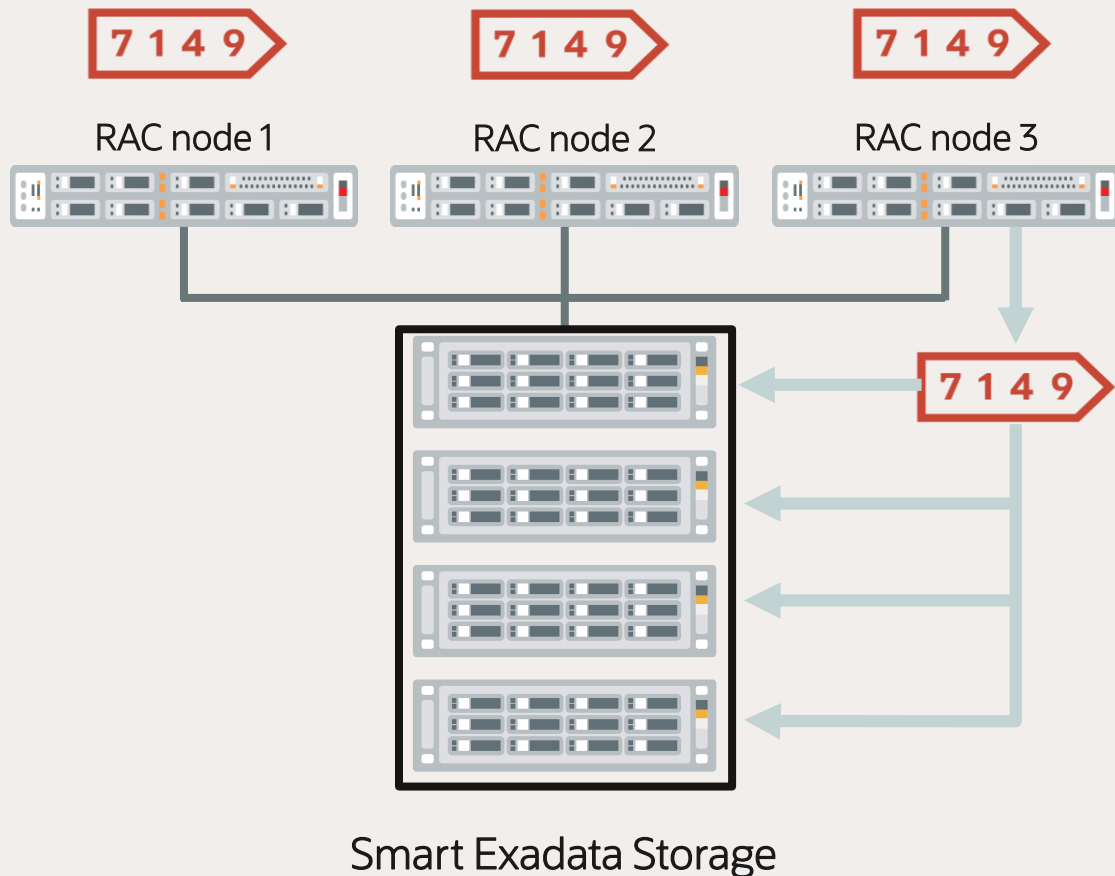


Partition Vector Index
(IVF_FLAT index)



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

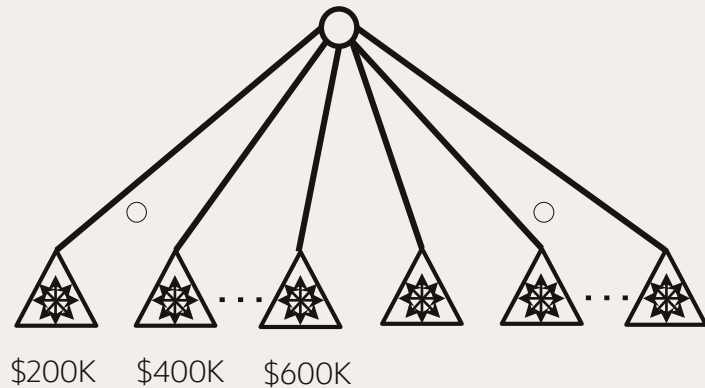
Scale-Out with Partitioning and Sharding

Partitioning

파티션별로 별도의 벡터 인덱스 생성

Vector index of house images

Partition by price range



Sharding

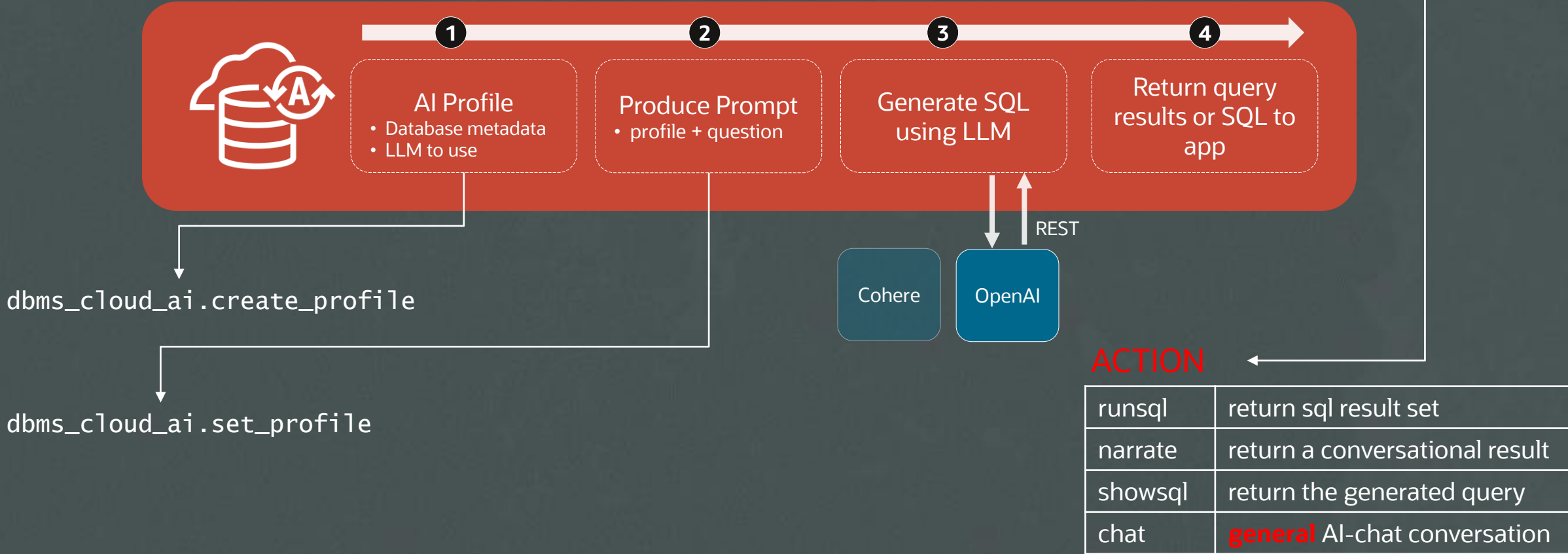
무제한 확장 또는 데이터 주권을 위해
데이터베이스 내의 sharding 기능을 사용하여
벡터 처리를 지역적으로 분산할 수 있음



오라클 DB 23c : LLM을 이용한 SQL 생성 지원



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



LLM을 이용한 SQL 생성 지원(23c)

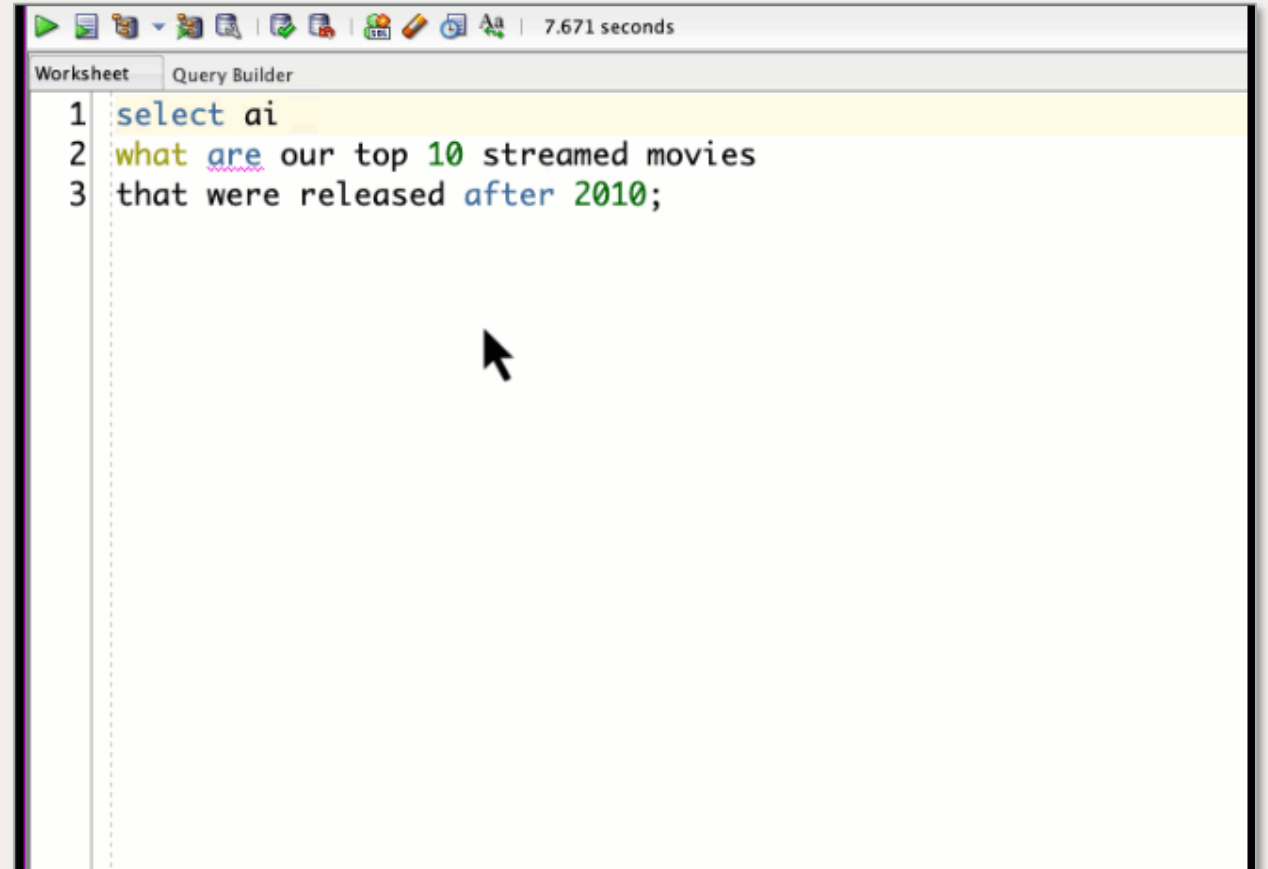
새로운 DBMS_CLOUD_AI PL/SQL 패키지를 사용하여 사용 가능

키워드 AI 및 질문과 함께 표준 SELECT 문을 사용

다른 SQL 결과 집합과 마찬가지로 결과를 처리

Actions

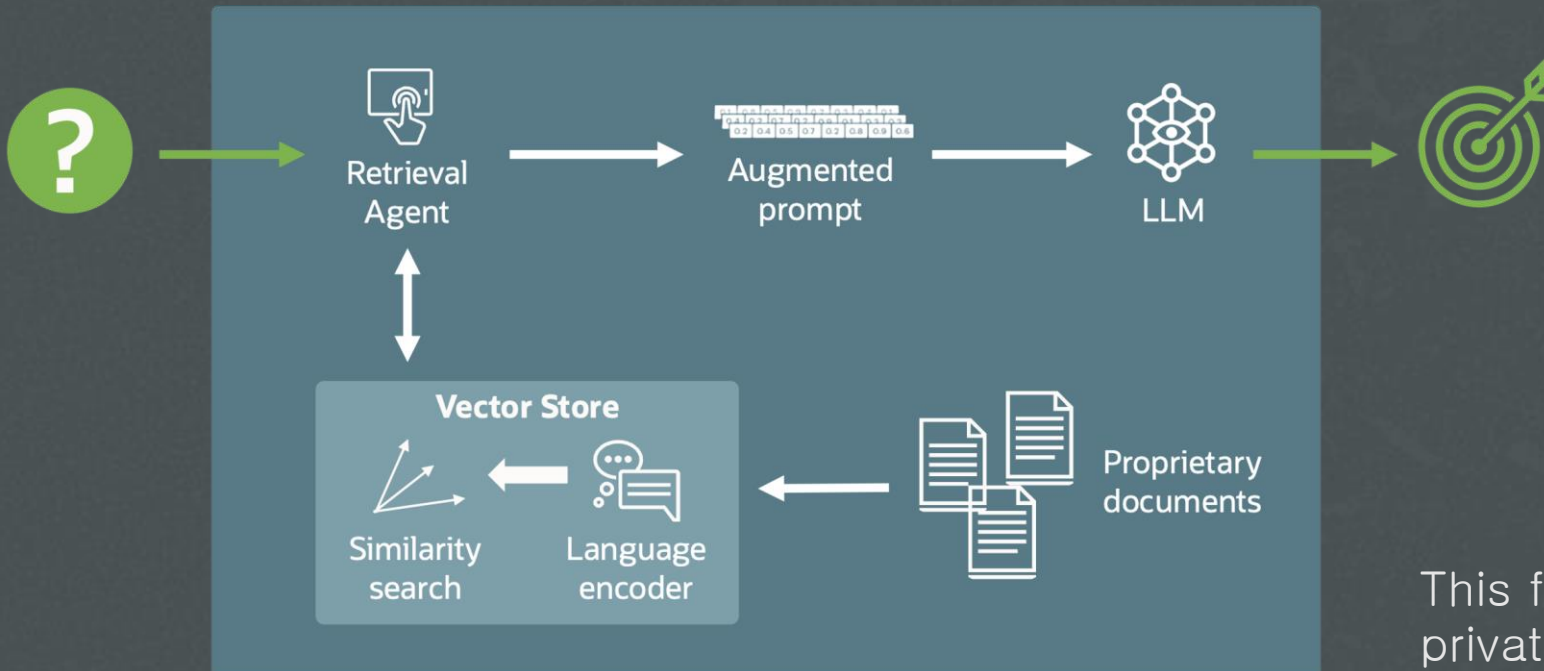
- **runsql** - SQL 결과 집합 반환 (기본값)
- **showsql** - 생성된 쿼리문 반환
- **narrate** - 대화 결과 반환
- **Chat** - 일반 AI 챗



MySQL HeatWave 역시 Vector Store를 통해 생성형 AI 지원



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



- Vector Store는 document를 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

소개 : OCI Generative AI Agents

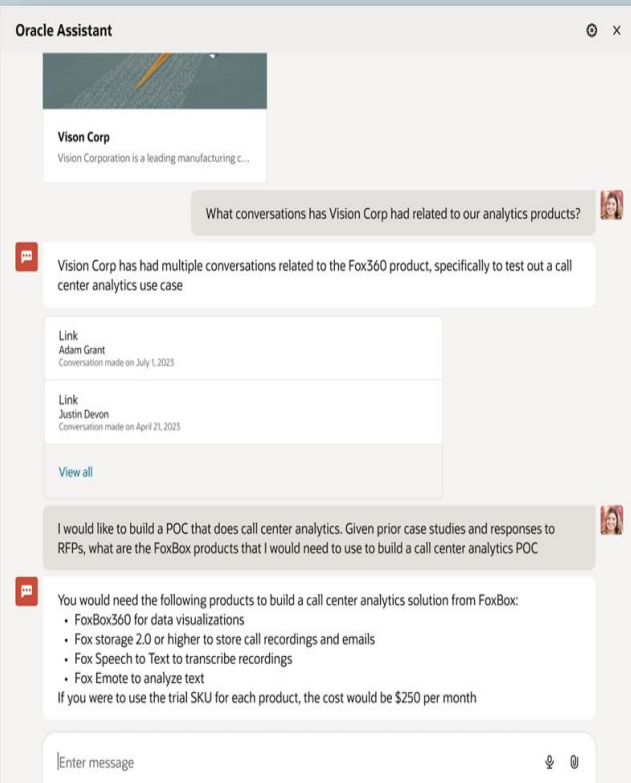
RAG Agent : Gen AI Agents의 첫번째 서비스

RAG를 지원하기 위한 Agent 서비스

RAG Agent : Beta release – January 2024

Beta:
OCI OpenSearch

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



OCI GenAI RAG Agent connected to Vision Corp's knowledge bases

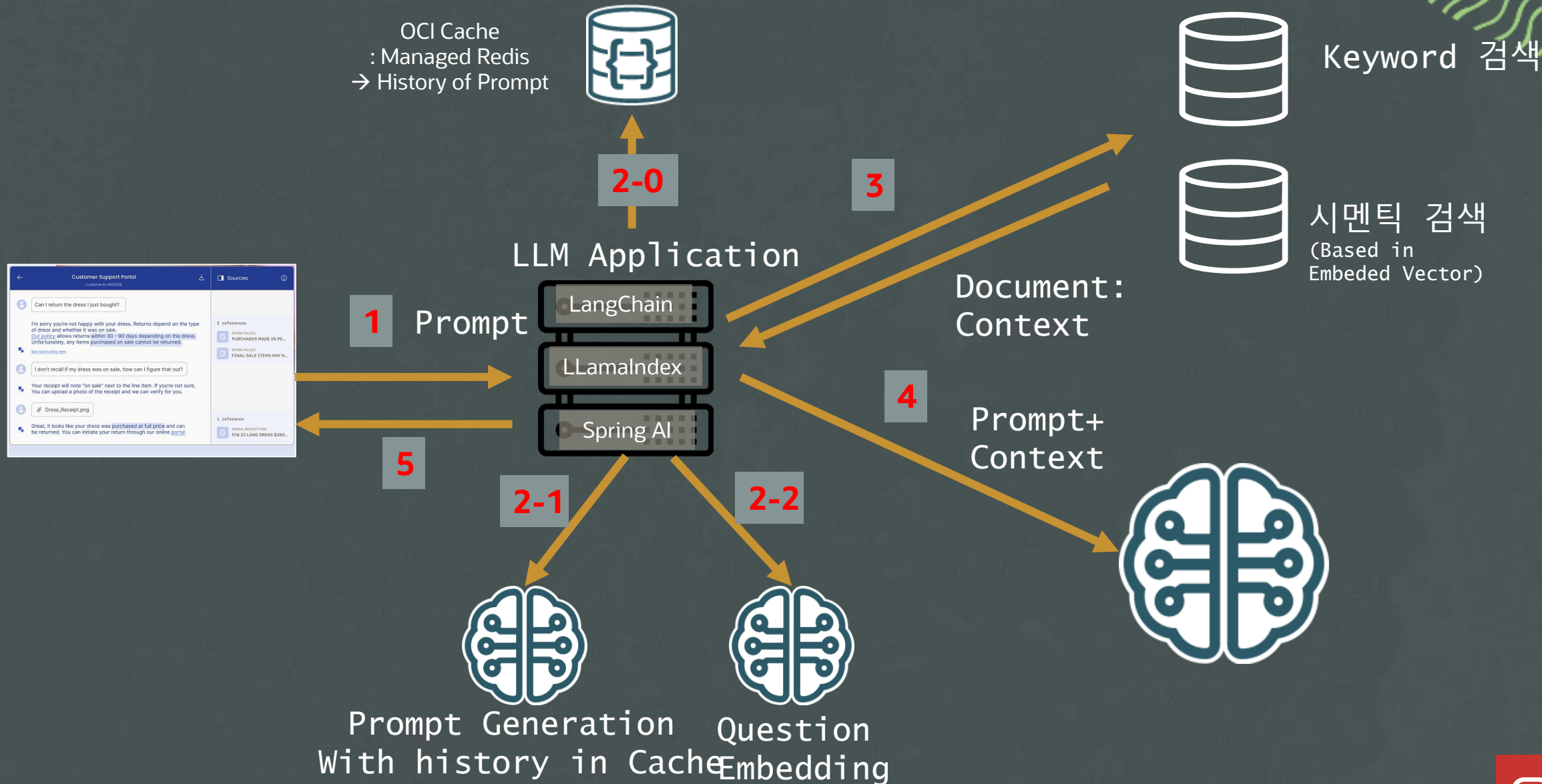
Analyst asks a natural language question

GenAI RAG Agent responds in humanlike manner and provides links to relevant source documents

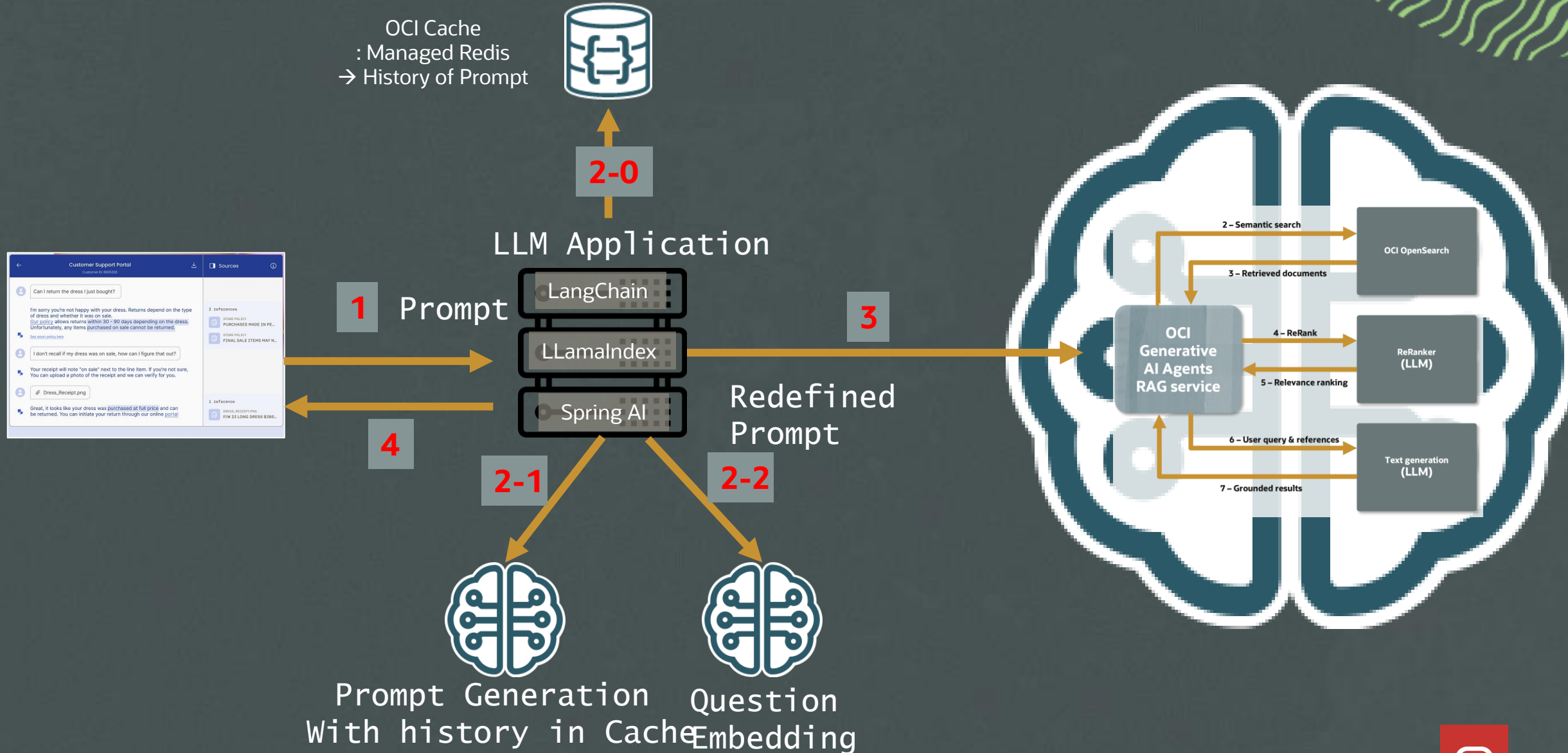
Analyst asks a follow up question

GenAI RAG Agent uses chat history and further information retrieval to respond

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



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



OCI Generative AI Agent 로드맵

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

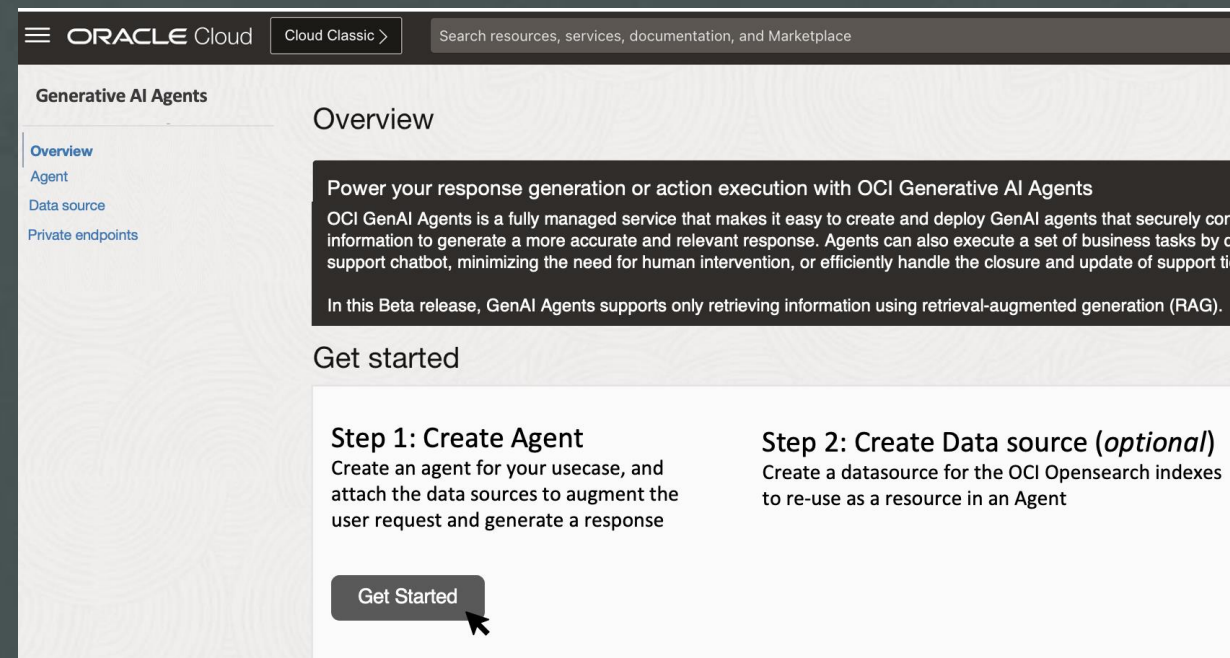
OCI Gen AI Agent는 OCI Generative AI 서비스 모델(Cohere 및 Llama-2)을 기본 지원

Reasoning & Planning

OCI Gen AI Agent는 ReAct framework를 통해서 일련의 생각, 행동, 관찰을 바탕으로 추론, 계획 및 동작

멀티-턴 에이전트

과거 요청 기록을 유지, 모델 컨텍스트와 응답을 재정의하여 명확성 향상



Announcing: AI Quick Actions for OCI Data Science

Quick actions:

- Deploy
- Fine tune
- Integrate
- Scale

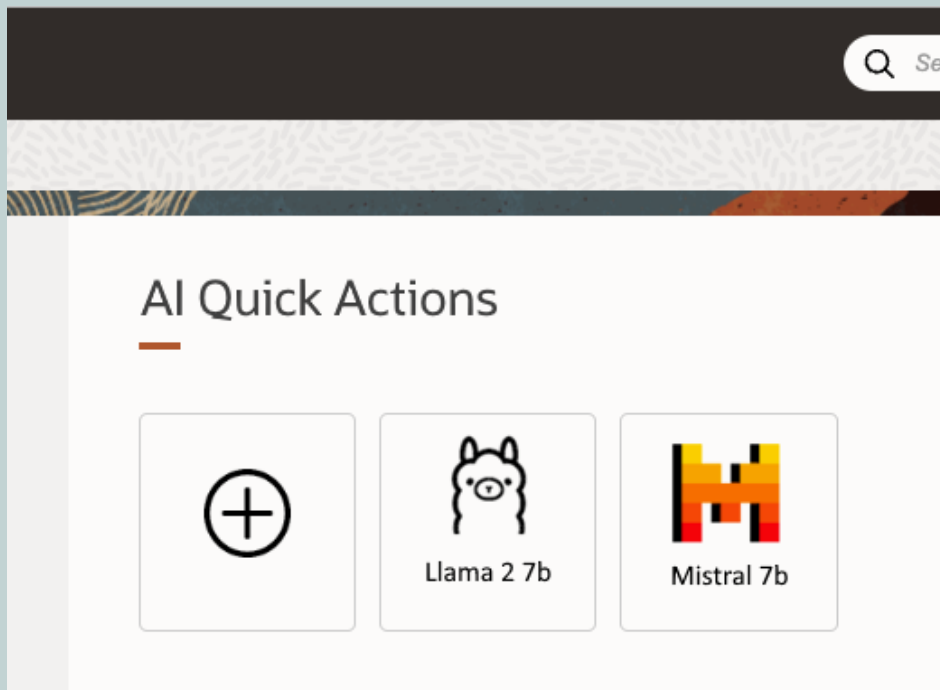
- ✓ Data Science 노트북 환경 내의 사용자 인터페이스에서 버튼 클릭만으로 호출할 수 있는 사용 사례 모음
- ✓ Data Science 노트북과의 원활한 통합을 통해 Llama2 및 Mistral 7B와 같은 LLM에 코드 없이 액세스할 수 있음

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

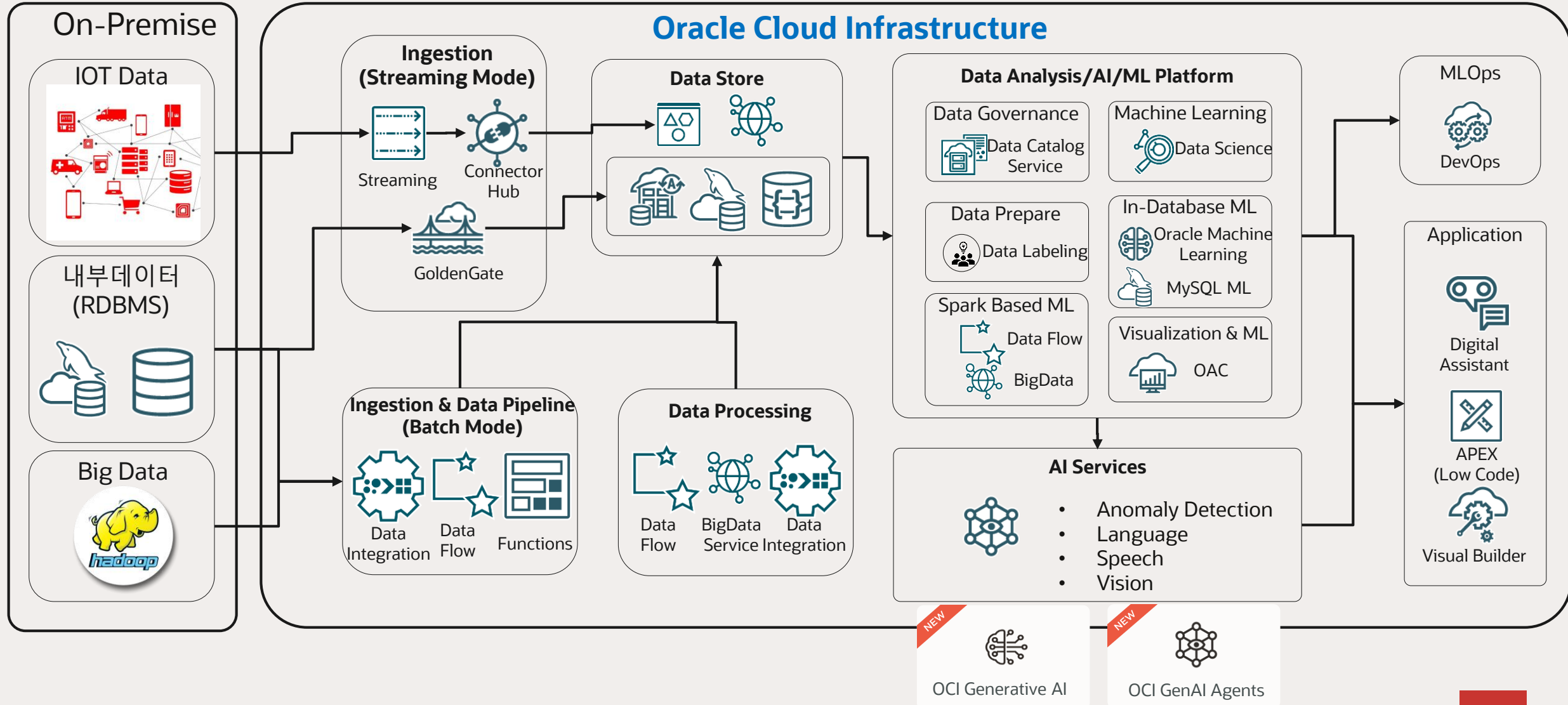
Fine Tuning

- 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



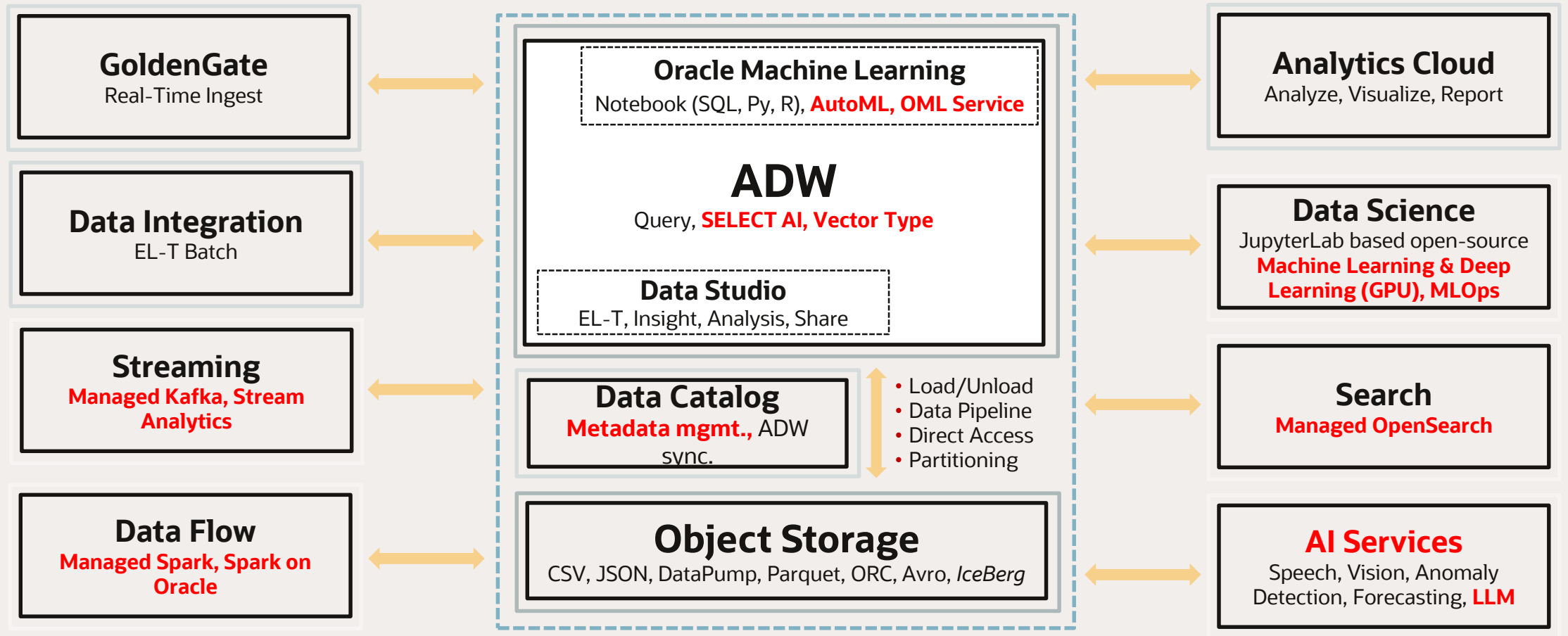
오라클은 클라우드 상에서 데이터 플랫폼을 제공해 드립니다

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

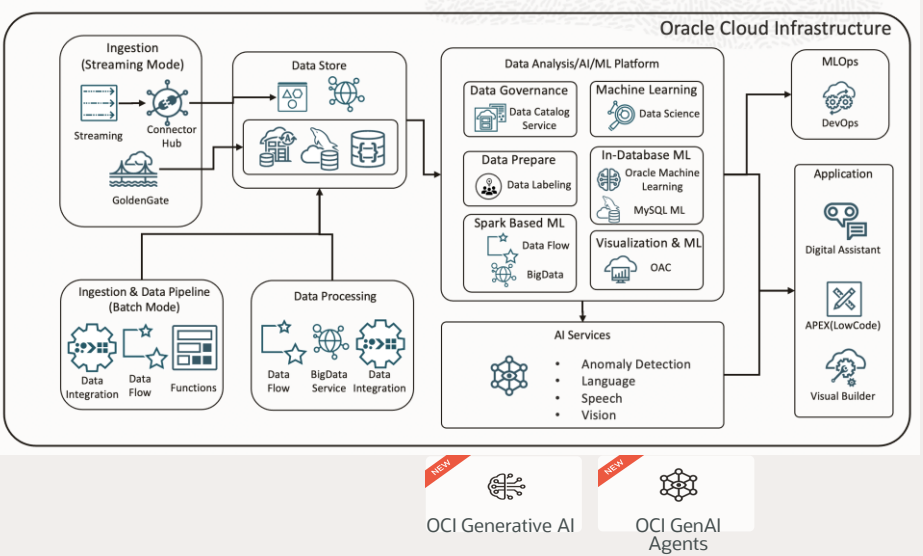


Enterprise AI/Data Platform 구성 시 고려 사항

- 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 아키텍처로 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 아키텍처 구성 필요
 - ✓ RAG 구성 시 콘텐츠 뿐만 아니라 보안 관련 프로세스가 중요
 - Vector DB는 회사의 주요 지식과 경험을 보관할 핵심 지식 저장소
 - Vector DB의 중요 콘텐츠에 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