ORACLE

# OCIの生成AIで
# プライベートデータをセキュアで効率よく活用

日本オラクル株式会社

2024/3/13

# 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.

リリース済

# OCI Generative AI Service
## Cohere-command、Meta Llama2

### 高性能なモデル

第三者による生成AIのベンチマーク(HELM)において、**高いスコア**を達成

(not included: GPT-4 from OpenAI)

| Company | Model | Model type | Mean win rate |
|---|---|---|---|
| cohere | **Cohere Command (52B)** | **Command** | **93.0%** |
| OpenAI | Davinci Instruct 002 | Command | 93.0% |
| OpenAI | Davinci Instruct 003 | Command | 89.8% |
| NVIDIA Microsoft | TNLG v2 (530B) *not publicly available or viable to serve given size* | Base | 85.5% |
| ANTHROPIC | Anthropic v4 (52B) | Command | 84.2% |
| AI21labs | J1 Grande v2 (17B) | Command | 80.6% |
| Aleph Alpha | Luminous Supreme (70B) | Command | 78.3% |
| cohere | **Cohere XL (52B)** | **Base** | **74%** |
| Meta | OPT (175B) | Base | 67.8% |
| OpenAI | GPT-3 Davinci (175B) | Base | 62.8% |
| AI21labs | J1-Jumbo (178B) | Base | 59.2% |
| Aleph Alpha | Luminous Extended (30B) | Command | 58.2% |
| Hugging Face | BLOOM (176B) | Base | 52.9% |

Source: Stanford's HELM benchmarks

### コンパクト

高性能ながらコンパクト。お客様による**カスタマイズが容易で早いレスポンス。**

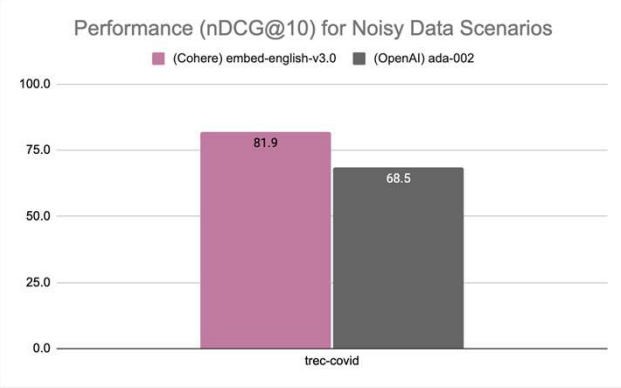
cohere 520億

**OpenAI GPT-3** 1,750億

パラメータ数

### 高品質なベクトル化

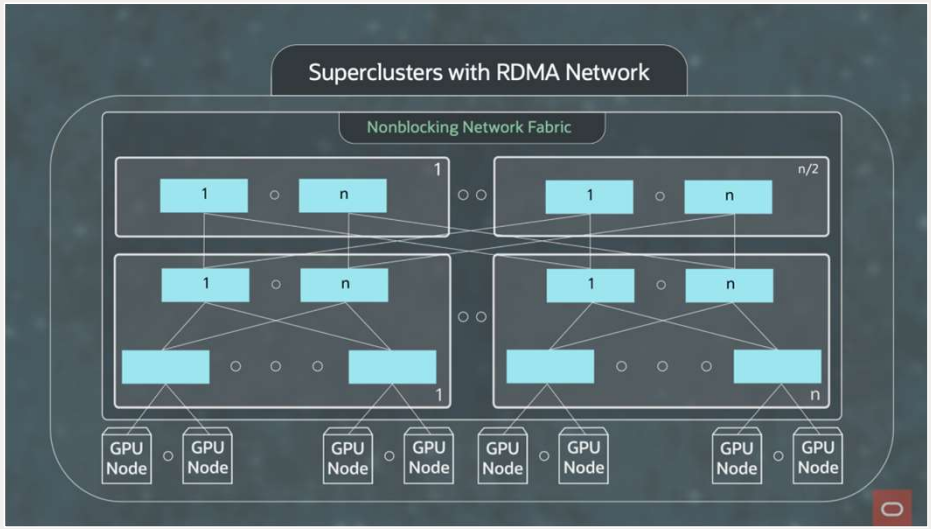
多様なデータソースから収集した"ノイズの多い"データセットに対しても、**より正確な結果を得る**ことが可能




Data Set: TREC-COVID
Source: Cohere

・Cohere社とMeta社のモデルをご提供
・オンデマンド/定額の2つの課金体系をご用意

# 高性能・低コストなクラウドAIインフラ
# 生成AI基盤として提供




- Super-Spine/Spine/Leafトポロジによるフルバイセクションバンド幅を備える低遅延の大規模RDMA(RoCEv2)ネットワークをOCIリージョン内に複数配備
- OCI Super Clusters: RDMAネットワークあたり最大 16,384ノードのCPUベアメタルもしくは最大16,384基のH100 GPUが接続され、複数テナントでクラスタを切り出して利用可能

NVIDIA

**NVIDIA、DGX Cloudの基盤に**
**Oracle Cloud Infrastructureを採用**

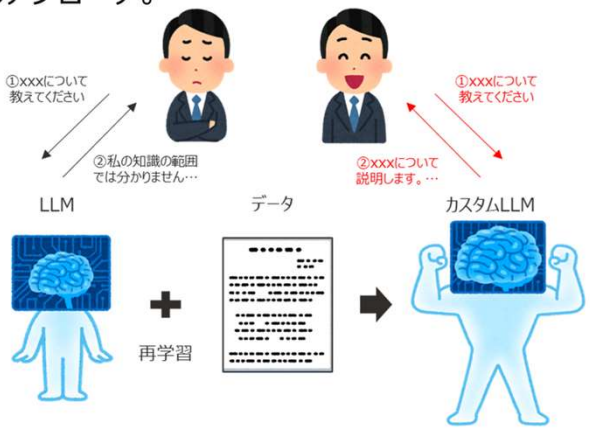
Microsoft

**マイクロソフト、Bingの対話型検索に**
**Oracle Cloud Infrastructureを活用**

# 生成AIをプライベートデータを合わせて活用する2つの選択肢
## ファインチューニングとRAG

### ファインチューニング

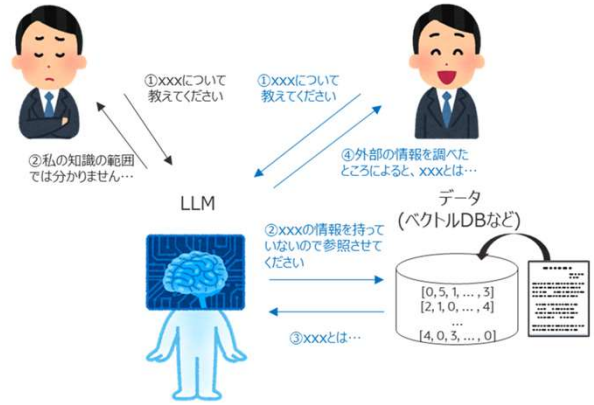
「LLMが持っていない知識」をLLMの外部のデータを用いて**追加学習**させることで、LLM自体を賢くするアプローチ。




・カスタムモデル作成にかかるコスト大
・カスタムモデルを動かすためのインフラコスト大

### RAG (Retrieval-Augmented Generation)

「LLMが持っていない知識」をLLMの外部のデータから**参照・補完**するアプローチ。

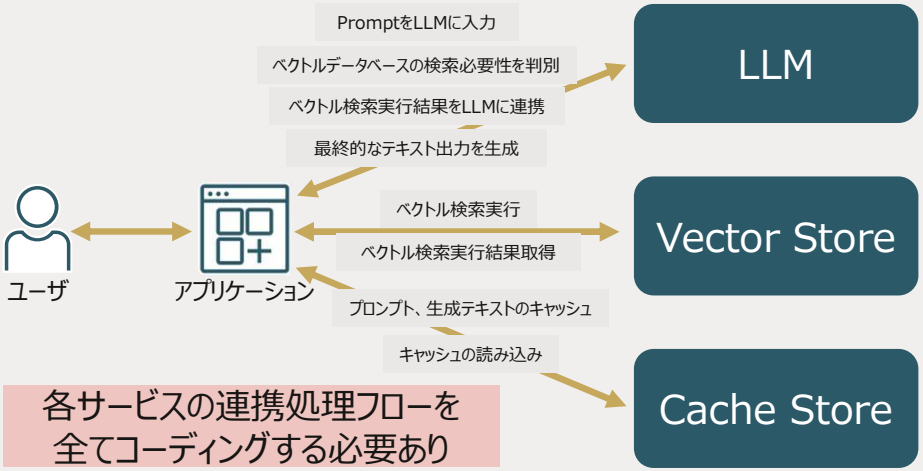



・ファンデーションモデルサービス利用による低コスト実現
・比較的容易なプライベートデータ等独自データを含めた応答
・ハルシネーション削減

限定公開中

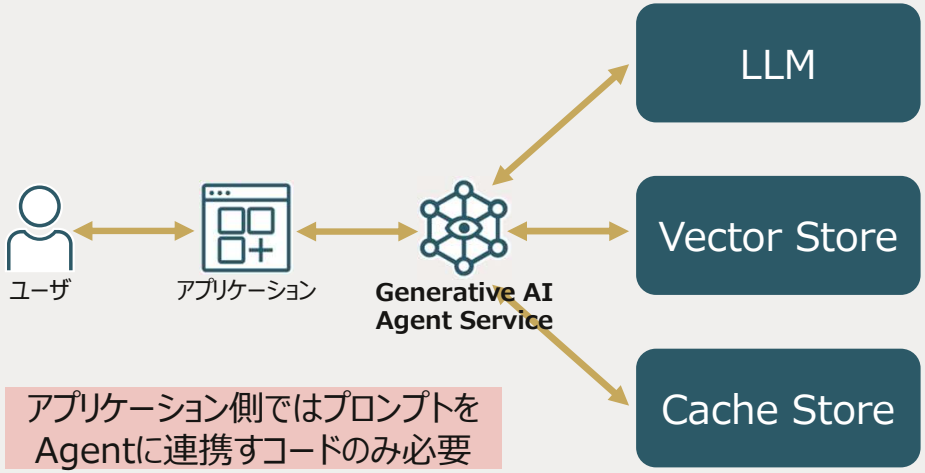
# 容易なRAG導入を提供する OCI Generative AI Agent Service
## LLMとその他サービスのオーケストレーションをGUIのみで実装

### オーケストレーションツールが無い場合




各サービスの連携処理フローを
全てコーディングする必要あり

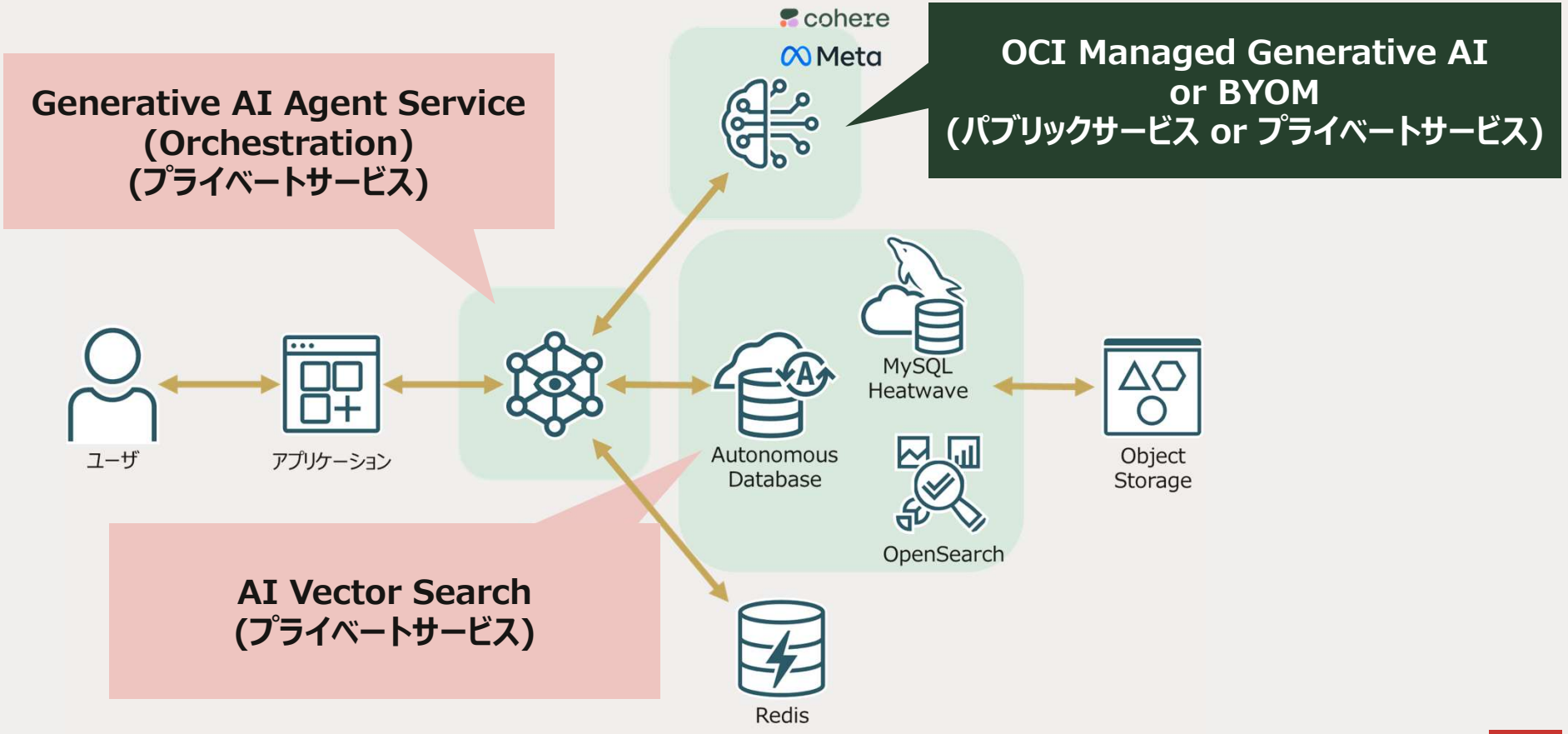
### オーケストレーションツールとして OCI Generative AI Serviceを利用した場合




アプリケーション側ではプロンプトを
Agentに連携するコードのみ必要

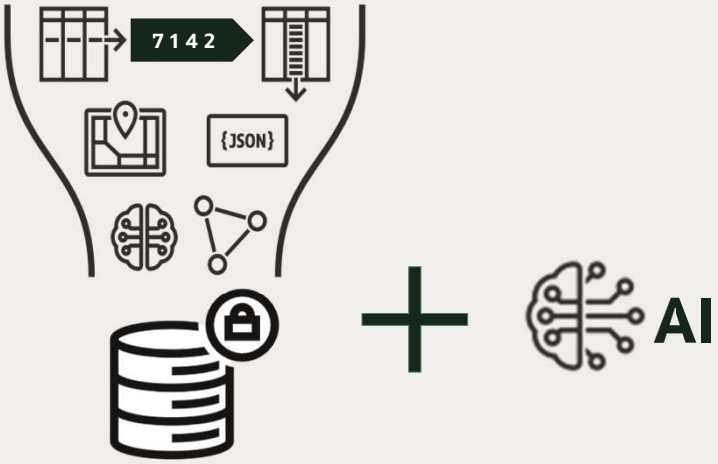
・Agent Serviceを利用することで、生成AIアプリケーションの開発を大幅に効率化
・GUIのみで実装ができるため、エンジニアの学習コストを抑え早期のプロジェクト着手を実現

# Oracleの生成AI RAG関連サービス



限定公開中

# AI Vector Search（Oracle Database 23c 追加機能）
## あらゆる生成AIをサポートする、データ・プラットフォーム




**Oracle Database 23c**
**AI Vector Search**

| | 分類 | 価格 | テキスト | ベクトル（テキスト） | 画像 | ベクトル（画像） |
|---|---|---|---|---|---|---|
| 0 | 001 | ¥1,000 | AAA…… | 0.1, 0.2, 0.6. …. |  | 0.5, 1.5, 2.6, …. |
| 1 | 002 | ¥2,000 | BBB…… | 0.8, 0.1, 0.4. …. |  | 1.0, 0.9, 1.6, …. |
| 2 | 003 | ¥3,000 | CCC…… | 0.5, 0.3, 0.9. …. |  | 0.6, 1.1, 1.3, …. |

### 1. 様々なデータを1つのデータベースで統合管理
- ベクトルデータを含む、**あらゆるデータ・タイプを一つのデータベース**に集約
- ビジネスデータとベクトルデータを組み合わせた検索を**一つのSQLで実現**
- データベース内で**データの一貫性を保ったまま、高速にベクトル変換するため、最新のデータもベクトル検索が可能**

### 2. 高速なベクトル検索
- Oracle Databaseの機能を活用した検索高速化
- パーティション化された、**ベクトル索引**
- グラフ・テクノロジーを活用した、**ベクトル検索**

### 3. 高可用性、スケーラビリティ、セキュリティ
- データ整合性を保持しながら、スケールアウトを可能に
- Oracle Database専用マシンの持つ、高性能、高可用性、堅牢なセキュリティをベクトル検索に

限定公開中

# AI Vector Search（Oracle Database 23c 追加機能）
## 1つのSQL文であらゆる情報を横断的に検索

SQL検索

AIかそれと同等の経験があって、xxに住んでいる人

xxさんと似た経歴の人は？

カメラに映った人と似ていて、住所がxxの人

| 画像ベクトル | 写真 | 文書ベクトル | 経歴書 | 名前 | 住所 |
|---|---|---|---|---|---|
| 0.1, 0.2... | | 0.9, 0.0... | | 佐藤 | 東京都... |
| 0.3, 0.7... | | 0.7, 0.1... | | 鈴木 | 大阪府... |
| 0.1, 0.1... | | 0.0, 0.8... | | 高橋 | 愛知県... |
| 0.9, 0.5... | | 0.2, 0.1... | | 田中 | 北海道... |


**Oracle Database 23c Vector Search**


ORACLE