Google Cloud

生成 AI 時代のデータ エンジニアリング入門

～ Google Cloud で実現する次世代のデータ エンジニアリングの第一歩 ～

Google Cloud
Data Analytics スペシャリスト
高村 哲貴

Table of Contents

スピーカー紹介

Google Cloud の Data Analytics スペシャリスト

日系 SIer 企業にて、データベース エンジニア、クラウド アーキテクトを経て、2021 年より現職。現在は、Google Cloud のデータ アナリティクス領域のスペシャリストとして、お客様のデータ活用を技術観点から支援

高村 哲貴
Google Cloud
Data Analytics スペシャリスト

01

生成 AI 時代 に
データ エンジニアリング はどう変わる？

データエンジニアの役割

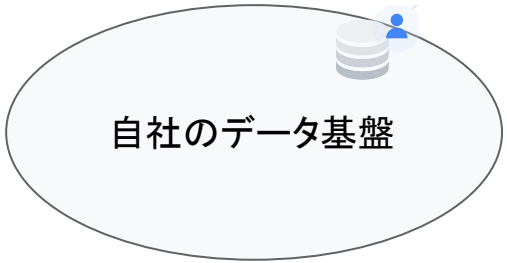
データを価値に変えるために、
ビジネスやユースケースに合わせて
データを最適な形で整理し、届ける

信頼できるデータを安全に、管理・統治する

データエンジニアリングの “イマ”

キーワード

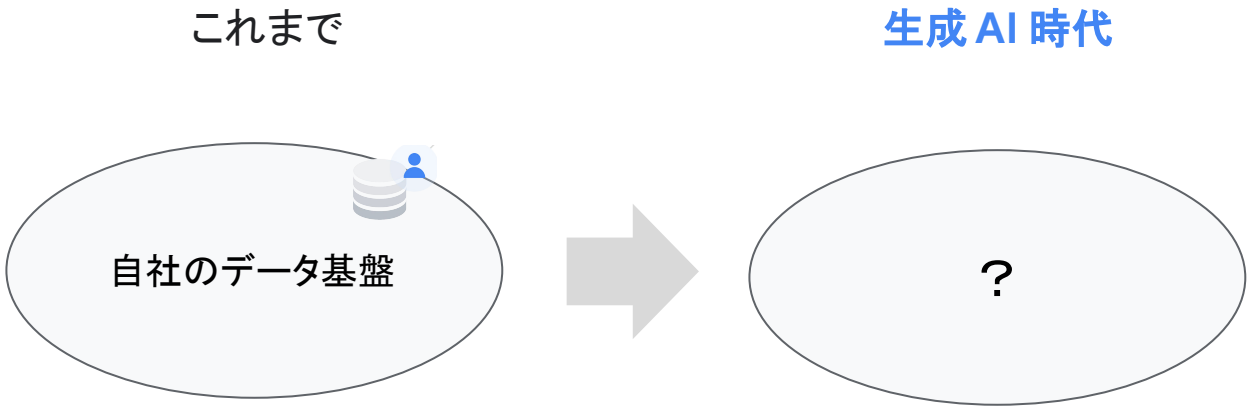
- データウェアハウス/ データレイク
- DataOps / MLOps
- データパイプライン(ETL, ELT)
- 構造化データ/ 非構造化データ
- バッチ / ストリーミング
- データメッシュ
- データ共有・交換
- データガバナンス
- データカタログ・メタデータ
- データ UI / エクスペリエンス
- データセキュリティ



限界と課題

- データが価値に繋がらない
- データサイロ
- 分断された環境やプロダクト
- データレイク が使われない
- ユーザがデータ活用するために必要な学習コスト
- 精度維持のための仕組み
- データの信頼性やプライバシーの問題

生成 AI の到来



- 生成 AI 時代に、データ基盤にはどういう変化が訪れるのか？
- データ基盤をどう進化させ、どう備えるべきなのか？

再掲

生成 AI のエンタープライズにおける活用シーン

会話、検索、クリエイティビティの 3 大ユースケースで生産性を向上

顧客接点

オンラインのやりとりを自然な会話で自動化・効率化

- 顧客サポートの自動化
- イントラのナレッジのQ&A
- ウェブサイトのナビゲーション

など

ビジネスユーザー / 分析者

複雑なデータに簡単にアクセス

- 製品・コンテンツカタログの探索
- ビジネスプロセスの自動化
- ドキュメント検索

など

クリエイティブ / エンジニア

ワンクリックでコンテンツを生成、生産性を向上

- コード自動生成と提案
- チャットによるアプリ開発の半自動化
- 画像生成による壁打ち

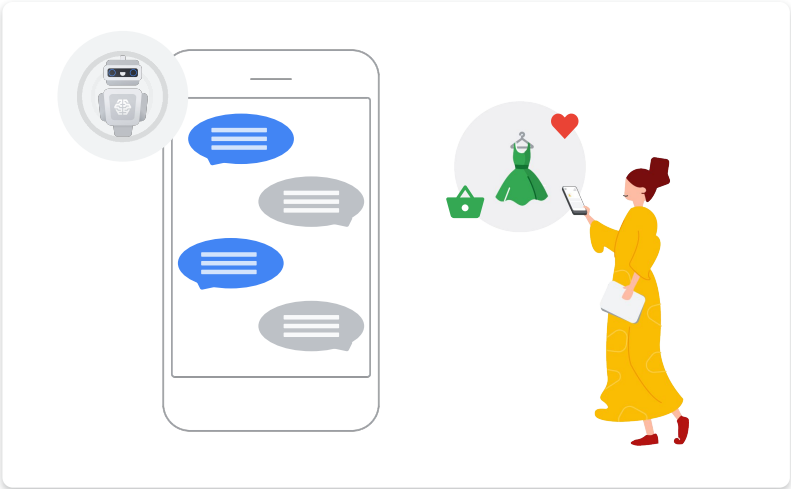
など

エンタープライズ データで活用

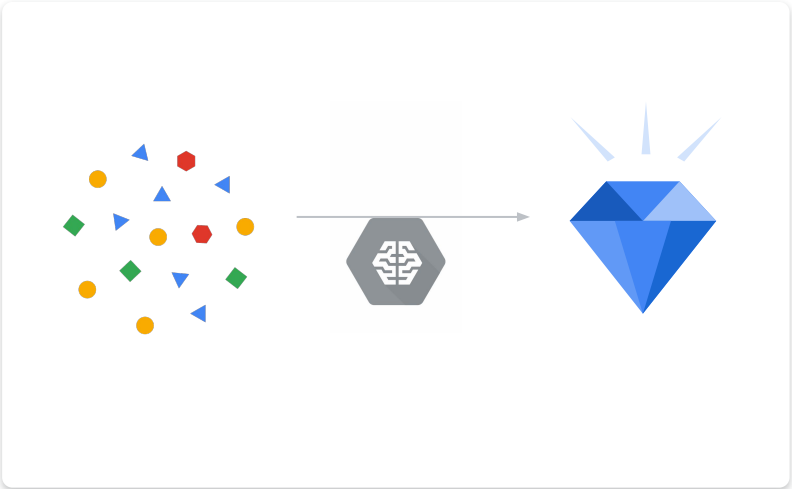
再掲

基盤モデル活用のパターン

生成 AI による
新しいユーザー体験の実現



あらゆる形式のデータを
ビジネスの価値に



“生成 AI” がデータ エンジニアリングに与えるイノベーション

本日のフォーカス

扱うデータが変わる

構造化データだけではなく、
非構造化データの重要性が高まる
新たにベクトルデータも

- 既存データのベクトル化と管理

データ活用が変わる

LLM を育てる (チューニング)、使う (プロンプト)、拡張する (グラウンディング)

- 自社データを用いた LLM のグラウンディング，セマンティック検索

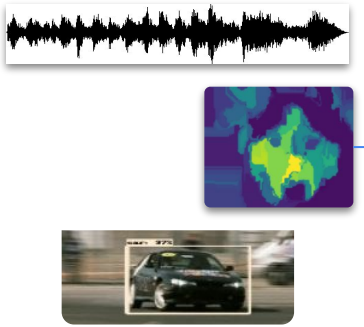
仕事の仕方が変わる

AI コーディネータ のサポートによる AIと協調したデータ分析

- 自然言語で SQL を生成、コード自動補完
- 自動的な異常検知

ベクトル化の意義と活用プロセス

データ → ディープラーニング モデル → エンベディング (ベクトル) 表現 → 非構造化データから 意味を取り出せる状態に



エンベディング モデル

[0.2, 0.5, 1.2, ..., 0.4, 0.05, 0.6]

画像、動画、テキスト、音声、時系列データ、など

事前学習された カスタムエンコーダ

データの意味構造を表す 数値のベクトル

類似する意味のデータは 距離が近い (クラスタ化)

なぜ、ベクトルデータが重要か？

① 情報アクセスの “アプローチ” が大きく変わる

事前に人力で整理したデータを キーワード検索

事前に準備

image_id	image_url	metadata
1	cat.png	[“cat”,”cute”]

Hit!

cute cat

非構造化データに対して手動で付けた
メタデータからデータを検索

セマンティック検索

事前に準備

[0, 0.4, 0] ⟷ [0, 0.3, 0]

Hit!

cute cat

ベクトル 同士の距離計算により
意味の近さが表現できる

ベクトル検索の応用範囲:
ベクトルを定義できるあらゆる用途で利用可能

文書や**画像**の内容で探す

似ている**製品**を探す

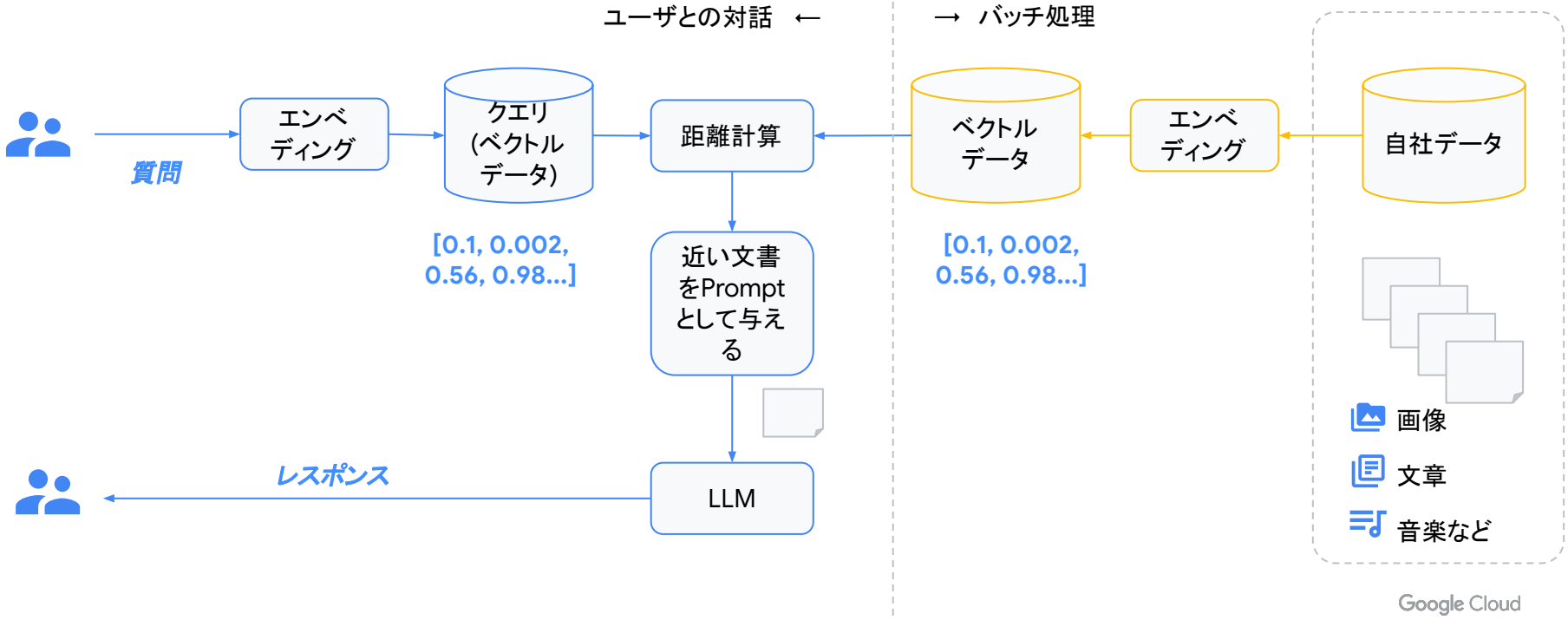
似ている**ユーザー**を探す

おすすめの**音楽**や**動画**を探す

故障しそうな **IoT デバイス**を探す

なぜ、ベクトルデータが重要か？
② 大規模言語モデル (LLM) を自社用にカスタマイズする

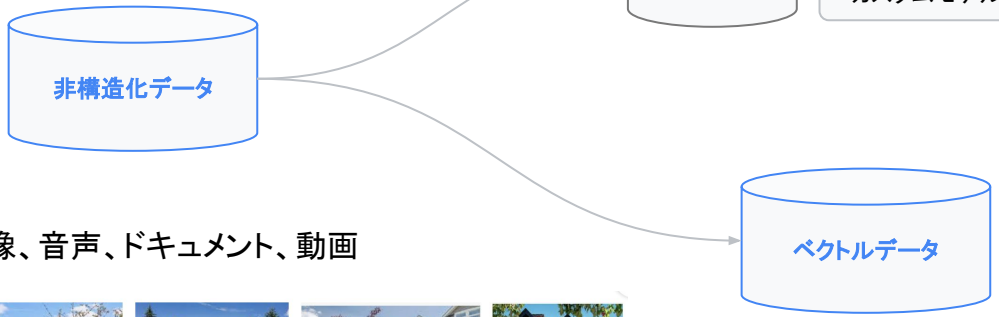
自社のデータをベクトル化し、特定分野の知識を繋いで　検索・回答する
(LLM の限界を超え、信頼性のあるサービスを構築するために　ハルシネーション問題を回避する)



なぜ、非構造化データの “管理” が重要か？

これまで、価値化が難しかった非構造化データも、
ベクトル化することでシンプル・スピーディに価値に変えられる

＜画像の場合＞



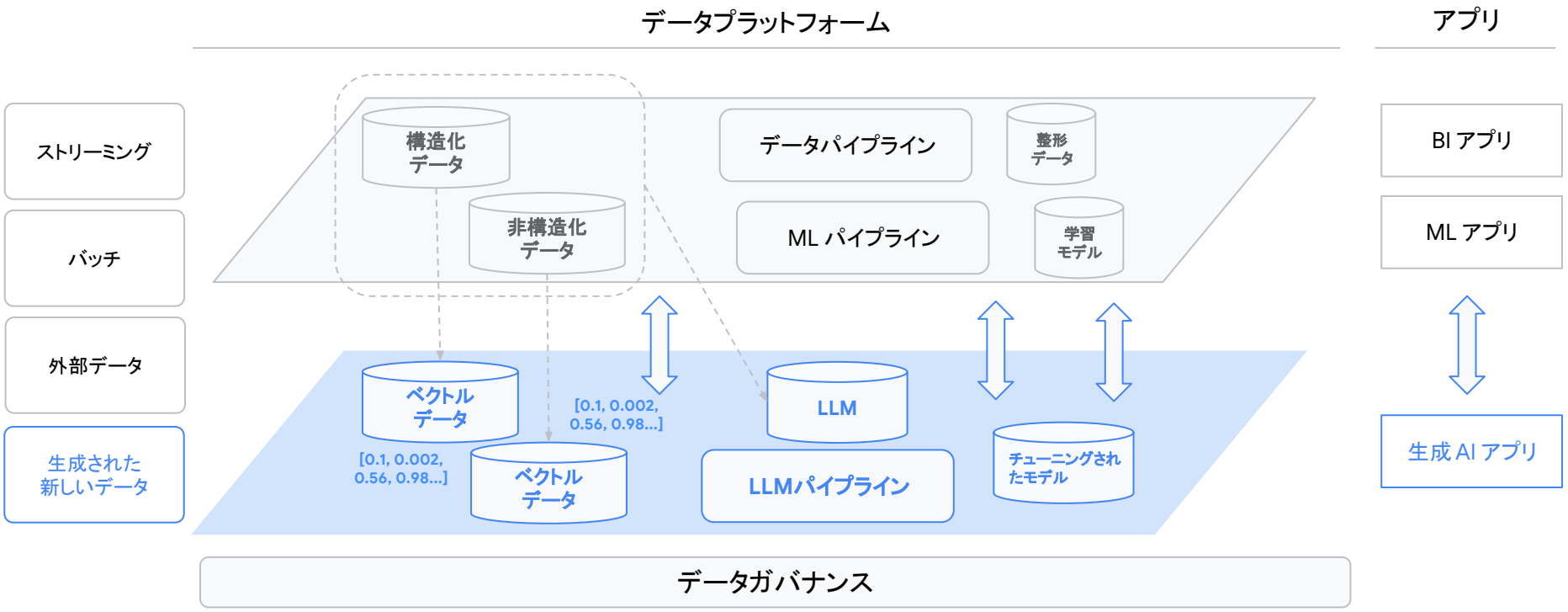
- 画像分類
- 物体検出
- セグメンテーション
- ・・・

画像、音声、ドキュメント、動画

- 検索
- クラスタリング
- レコメンデーション
- 異常検出
- 分類
- ・・・

生成 AI 時代に “データ基盤” はどう変わるか？

- 非構造化データの重要性は高まり、ベクトルデータで新しいデータ活用が促進される
- 従来のデータ活用と生成 AI を横断する**ハイブリットなデータ処理・管理・ガバナンス**が求められてくる



02

Google Cloud で実現する
生成 AI 時代のデータ エコシステム

Google Cloud の データ & AI クラウド

あらゆるユーザーの “データ & AI エクスペリエンス” をシンプルに

Any user

- **データエンジニア** — Clean, useful data
- **ML エンジニア** — Integrated intelligence
- **データサイエンティスト** — Models that work
- **デベロッパー** — Intelligent apps
- **データアナリスト** — Query and analyze
- **ビジネスユーザ** — Insights Everywhere
- **お客様** — Value

データ & AI レイクハウス

BigQuery

マルチエンジン（SQL, Spark, Python, Search, リモート関数）

マルチストレージ（マネージドストレージ GCS, BQ Omni cross-cloud）

マルチフォーマット（構造化データ,半構造化データ,非構造化データ）

マルチケイパビリティ（ガバナンス, ストリーミング, データパイプライン）

生成AI & ML

Vertex AI

ビジネスインサイト

Looker

基盤モデル（テキスト, チャット, コード, 動画, 画像, 音声）

Google Cloud インフラストラクチャ: GPUs/TPUs

GA (一般公開)

非構造化データを管理・活用 - BigQuery の非構造化データサポート

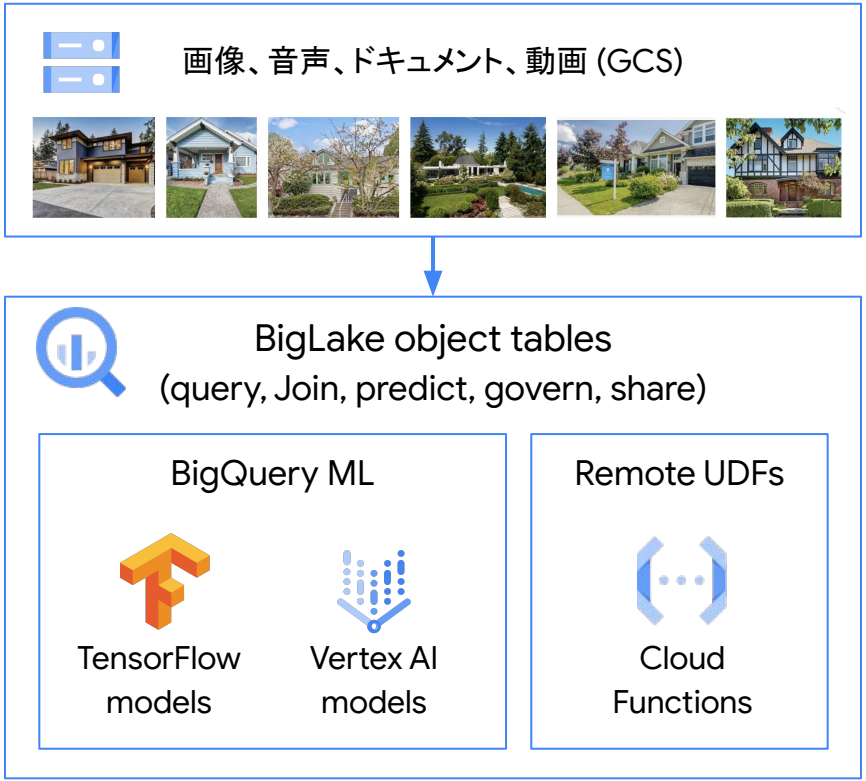
一つの基盤で非構造化データも扱える

ユーザの間口を広げる

- SQL を使えるユーザなら誰でも簡単に非構造化データをビジネス価値に

データ活用の幅が広がる

- 構造化データと同じ基盤で非構造化データ (画像、音声、ドキュメント、動画) を管理。BQML やリモート UDF と組み合わせた活用が可能



ユースケースに合わせてベクトルデータを管理・活用

pgvector support

Cloud SQL
for PostgreSQL

高可用なフルマネージドデータベース サービス (PostgreSQL, MySQL, SQL Server)

pgvector support

AlloyDB
for PostgreSQL

トップティア ワークロードのための PostgreSQL 完全互換クラウドネイティブ データベース

BigQuery

サーバーレスで費用対効果に優れたエンタープライズデータ ウェアハウス

従来のデータ基盤と共存しながら
ベクトルデータにも対応する

GA (一般公開)

Vertex AI
Matching Engine

拡張性が高くレイテンシが低いベクトル類似性マッチング（近似最近傍探索：ANN）サービスを提供

より高い性能要求、拡張性に対応
特化型ソリューション

既存データと LLM をシームレスにつなぐ

Preview

BigQuery - integrates with Vertex LLMs

SQL クエリだけで生成 AI とつなぐことができ、あらゆるデータをシームレスにビジネス価値に変える

ユーザの間口を広げる

- SQL を使える全てのユーザが生成AI を使いビジネスアイディアが実現できる

使えるデータが広がる

- BigQuery 上にあるデータのみならず、BigQuery と連携できる全てのデータソース(CloudSQL, Spanner, GCS 上の CSV ファイルや外部データソース) を移動することなく、生成AI を活用できる

1. Register the model as a remote model

```
CREATE MODEL my_project.my_company.llm_model
REMOTE WITH CONNECTION my_project.us.remote_connection_name
OPTIONS (remote_service_type = 'CLOUD_AI_LARGE_LANGUAGE_MODEL_V1')
```

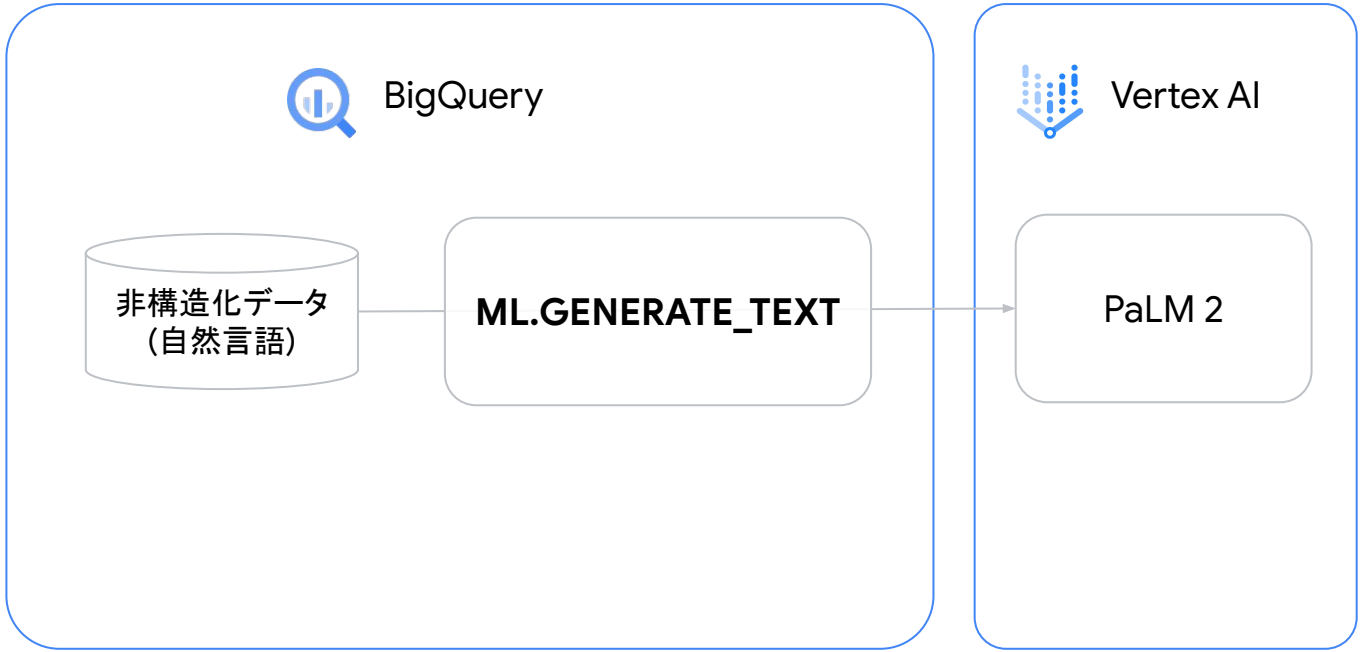
2. *Run inference*. Here's an example where users can do data enrichment by obtaining the country name for a given city name. Note that "city" is a column in the "example_table".

```
SELECT * FROM
ML.GENERATE_TEXT (
      MODEL 'my_company.llm_model',
(SELECT CONCAT ("Give the country name for city: ", city) AS prompt
FROM example_table),
STRUCT ( 0.2 AS temperature,
  1024 AS max_output_tokens,
  0.8 AS top_p,
  40 AS top_k))
```

03

[Demo]
BigQuery 上の非構造化データに対して SQL のみで LLM のテキスト生成を実行する

Demo



- サマライズ (要約)
- 感情分析
- キーワード抽出
- エンリッチメント
- 分類

04

本日のまとめ

本日のまとめ

01 | 生成 AI 時代の到来により、データ活用のニーズは変化し、非構造化データやベクトルデータを含めたハイブリッドなデータ管理が要求される

02 | Google Cloud では、BigQuery や Matching Engine など、生成 AI の ユースケース や アプリケーションの要件に合わせて最適なソリューションを提供

03 | Google Cloud を活用して 生成 AI に対応したデータ基盤に進化＆再構築し、生成 AI を競争優位に

Google Cloud Next ‘23 (8 月 29 日～31 日) の発表もご期待ください！

Thank you.

Google Cloud