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
生成 AI データ エンジニアリングの第一歩 ～

Google Cloud
データ アナリティクス スペシャリスト
饗庭 秀一郎

Table of Contents

生成 AI 時代に データ エンジニアリングはどう変わる？	01
Google Cloud で実現する生成 AI 時代の AI レイクハウス	02
Demo	03

スピーカー紹介

饗庭 秀一郎

Google Cloud

データ アナリティクス 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、ユーザーに合わせて
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

“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

非構造化データを活用

従来の構造化データだけではなく、非構造化データを活用してビジネスを差別化

- 非構造化データ (ドキュメント、音声、画像など) を使えるデータに
- 構造化データと共にガバナンスとアクセシビリティ

基盤モデルを用いた AI/ML の活用

基盤モデルを活用して、プロンプトでタスクを指示できるので様々なユーザが活用しやすい

- LLM をつかってデータをエンリッチメント
- 生成 AI アプリの実装

AI アシスタントで 生産性向上とデータ民主化

自然言語で AI アシスタントとやり取りし、データ分析に必要な知識やスキルをサポート

- 自然言語で Google Cloud に関することを質問
- 自然言語から SQL を生成、コード自動補完

データ エンジニアリングに求められること

あらゆる種類のデータを分析可能に

- **非構造化データ**を分析可能に
- 非構造化データの**データマネジメントとガバナンス**

データがあるところで AI を活用

- **データのあるところで素早く** AI/ML 処理を実行可能に
- AI を活用したい**ユーザーに適したツールとデータアクセス**を提供

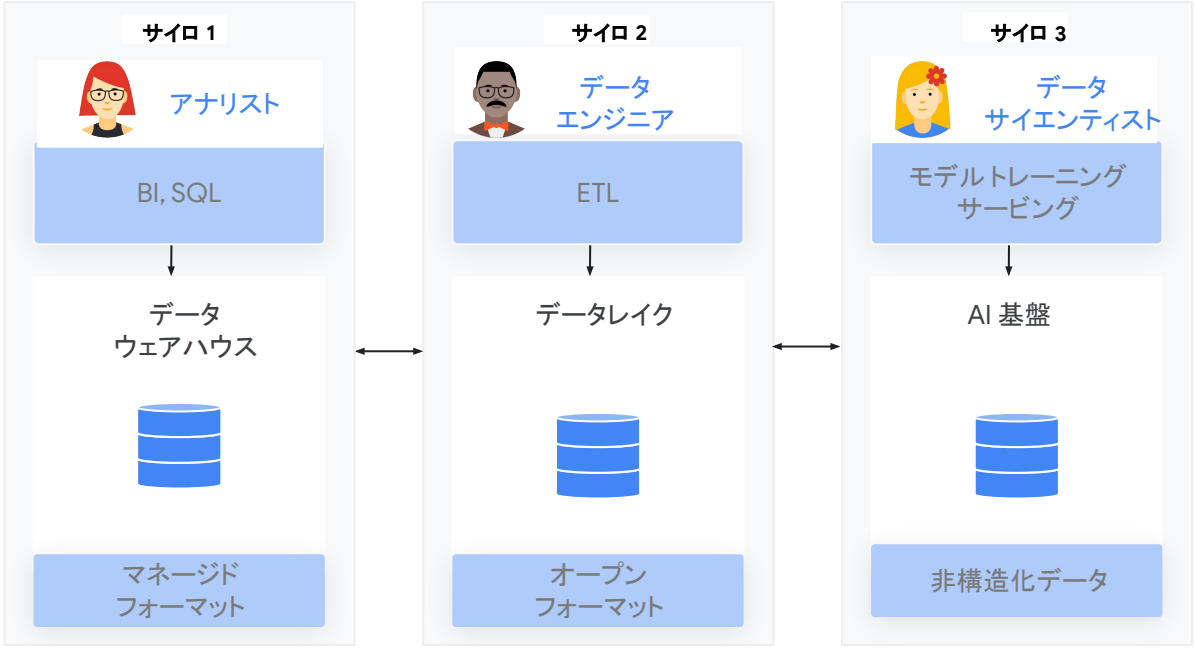
AI アシストで効率的かつ民主的に

- **データ民主化の啓蒙活動**に AI アシストを利用
- AI を活用して**誰でも欲しいデータによりアクセスできる**環境を構築

サイロ化されたデータ基盤の課題

データをプラットフォーム間で移動や複製させることによる課題

- データ移動・コピーのコストとデータパイプラインの複雑化
- データが分散し、かつ追跡しづらくなることで、ガバナンスが低下
- 様々なツールのプラットフォーム毎の管理や権限などのセキュリティリスク
- 使い慣れない、使えないツールによる生産性の低下



02

Google Cloud で実現する
生成 AI 時代の AI レイクハウス

Google Cloud の AI レイクハウス ソリューション

データ ウェアハウスとデータレイクの両方の利点を持つ**統合されたデータ基盤**

様々なユースケースとペルソナ
エンジニアからビジネスユーザまで
BI ユースケースから AI/ML ユースケースまで

AI をデータのあるところで
データに AI を持ってくることでセキュアかつ効率的に従来の ML から最新の 生成 AI (LLM やエンベディングス) まで構築・運用

統合されたガバナンス
構造化/非構造化、データから ML モデルまで一元的に管理

いかなるデータも
構造化 / 半構造化 / 非構造化、フォーマット、ストレージ サービスに関わらず

BigQuery

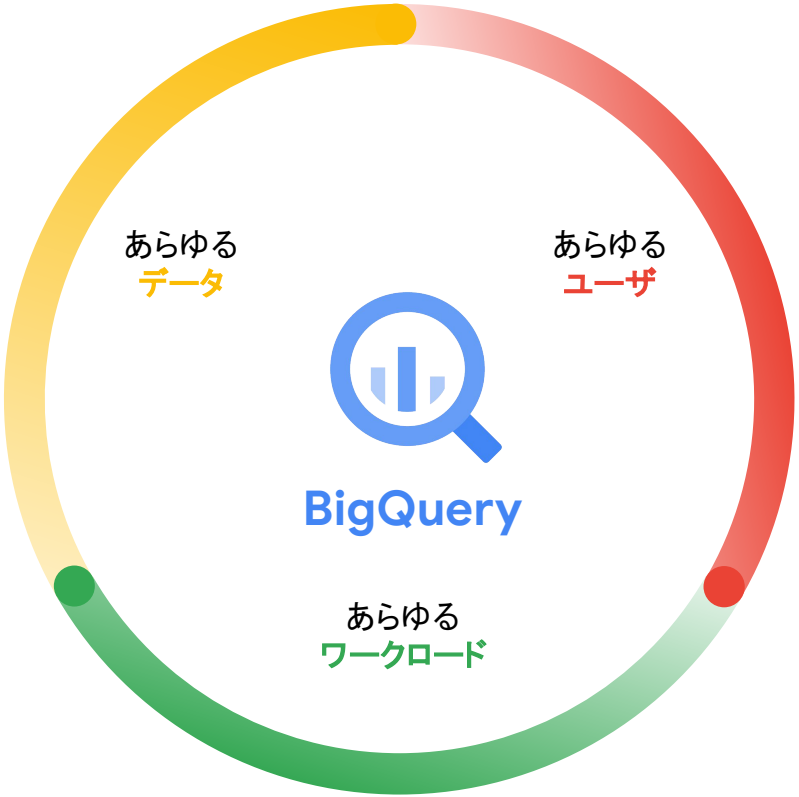
PB/EB 級のデータを扱えるサーバレスなデータウェアハウスソリューション

マルチエンジン
SQL、Spark、Python、remote 関数

マルチストレージ
マネージド、Cloud Storage、
他クラウド ストレージ

マルチフォーマット
構造化、半構造化、非構造化、
オープンフォーマット

マルチ機能
ガバナンス、ストリーミング、
データパイプライン

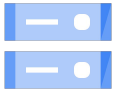


あらゆる種類のデータを分析可能に

SQL で非構造化データの分析

画像、音声、
ドキュメントなど

Cloud Storage



オブジェクトテーブル

メタデータを自動収集しテーブルにすることでアクセス権限なども従来のテーブル同様に管理可能

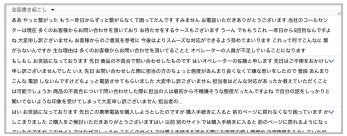
SQL で BigQuery ML
推論エンジン利用

```
SELECT
  transcripts AS `会話書き起こし`,
  ml_transcribe_result,
  uri
FROM
  ML.TRANSCRIBE(
    # 音声学習モデルの利用
    MODEL `unstructured_demo.bqml-cloud-speech2text`,
    # 音声ファイルデータの Object table の指定
    TABLE unstructured_demo.call_center_conversations,
    recognition_config => JSON '{"model": "long"}' )
ORDER BY
  uri DESC;
```

```
SELECT
  REPLACE( STRING(ml_annotate_image_result.full_text_annotation.text),
  *
FROM
  ML.ANNOTATE_IMAGE( MODEL `unstructured_demo.bqml-cloud-ai-vision`,
    TABLE `shua-demo-401705.unstructured_demo.classic_movie_posters`,
    STRUCT(['TEXT_DETECTION'] AS vision_features))
```

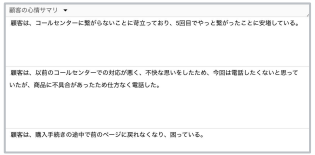
```
# テキスト化した会話を大規模言語モデルを使い要約し、さらに顧客が何に不満を持っているかを分類する。
WITH
  summary_t AS (
  SELECT
    `会話書き起こし`, ml_generate_text_llm_result AS summary, uri
  FROM
    ML.GENERATE_TEXT (MODEL `unstructured_demo.bqml-cloud-llm`,
      (SELECT
        concat ("""コールセンターでの会話内容のtextで顧客がどう感じているかを自然な文章で30文字程度で
                text : \"""",
                `会話書き起こし`) || "\"" AS prompt,
        `会話書き起こし`, uri
      FROM `unstructured_demo.transcribed_results`), STRUCT (TRUE AS flatten_json_output)
  reason_t AS (
```

音声から文字書き起こし



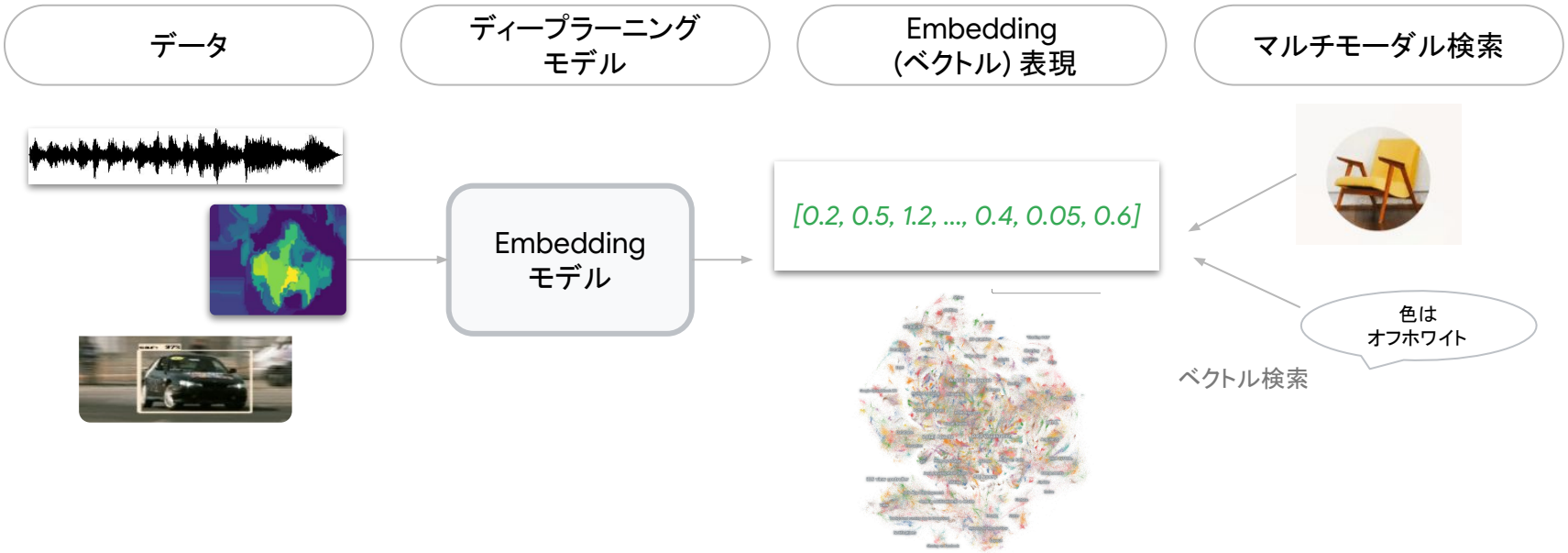
画像からテキスト抽出

LLM 用いて文章を要約



あらゆる種類のデータを分析可能に

エンベディングとベクトルによる非構造化データの活用



あらゆる種類のデータを分析可能に

BigQuery でエンベディング活用 - ベクトル検索

ML.GENERATE_TEXT と ML.DISTANCE

1. テキストをエンベディングに変換

TREES DOWN NEAR THE INTERSECTION OF HIGHWAY 43 AND HIGHWAY 187. (BMX)

→

[-0.01059811282902956, -0.00997710507363081, -0.039084006100893021, 0.014862567186355591, -0.029290193691849709, -0.021420236676931381, "0.0075757815502583981", "-0.0038812912534922361",...

2. ベクトル間の距離を計算

TREES DOWN NEAR THE INTERSECTION OF HIGHWAY

↓

Row	検索対象文	ベクトル間の距離
1	TREES DOWN NEAR THE INTERSECTION OF HIGHWAY 43 AND HIGHWAY 187. (BMX)	0.081762477527...
2	TREES DOWN ON HWY 33 SOUTH OF THE INTERSECTION OF HWY 36. (HUN)	0.123115501710...
3	TREES DOWN NEAR THE INTERSECTION OF LEE RD 440 AND LEE RD 179. ALSO REPORT OF TREES DOWN JUST TO THE EAST NEAR THE INTERSECTION OF LEE RD 240 AND LEE RD 212. TIME ESTIM (BMX)	0.141185000401...

エンベディング化とベクトル検索のクエリ

```
# 0. エンベディングのためのモデルを生成 (PaLM embedding API)
CREATE OR REPLACE MODEL
 text.embedding_model REMOTE
WITH CONNECTION `project.region.connection` OPTIONS
(remote_service_type=
"CLOUD_AI_TEXT_EMBEDDING_MODEL_V1")

# 1. テキストをエンベディングに変換
SELECT * FROM
 ML.EMBED_TEXT(MODEL text.embedding_model,
              (SELECT comments AS content
                 FROM text.wind_reports))

# 2. ベクトル間の距離を計算
SELECT content,
 ML.DISTANCE(
  (SELECT text_embedding FROM text.semantic_queries),
  text_embedding,
  'COSINE') AS distance
 FROM text.embeddings_wind_reports
```

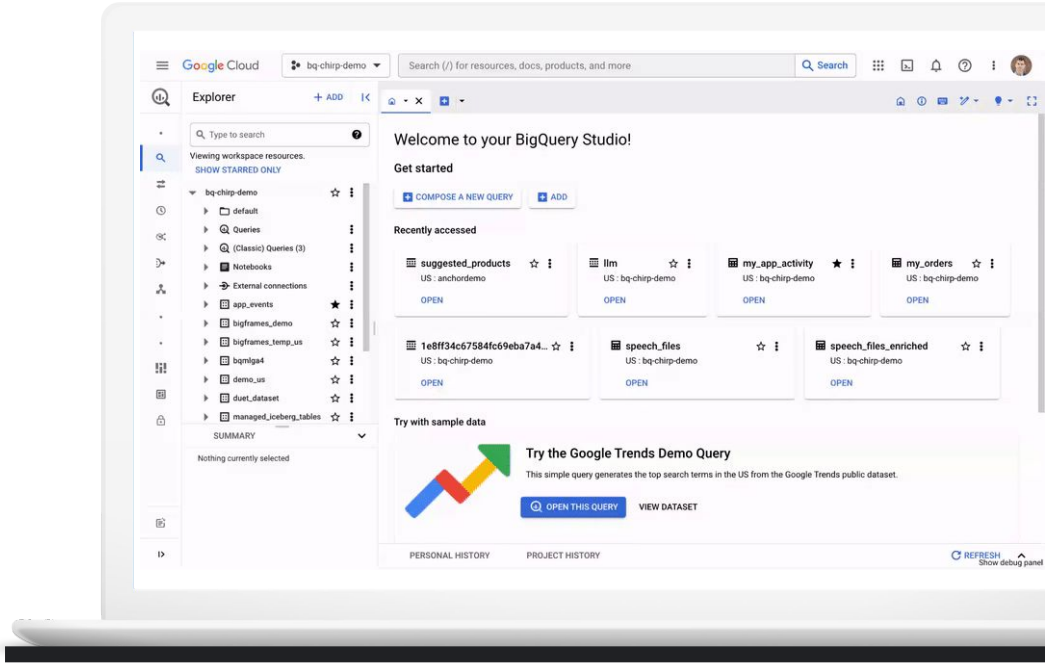
データがあるところで AI を活用

Preview

BigQuery Studio

全てのデータ プラクティショナーのデータ分析や AI のワークフローを加速

- BigQuery の画面で**ノートブック**が利用可能に SQL に加えて、**Python** での開発が可能
- テーブルのスキーマ情報に加えて、**データ プロファイリング**、**データ品質**、**リネージ**が利用可能
- **AI によるチャット**、**コードアシスタント**により生産性を最大化
- **BigQuery DataFrames** により、Pandas / scikit-learn の書き心地で BigQuery のパワーを活用



データがあるところで AI を活用

BigQuery ML
AI をデータのあるところへ

SQL で機械学習処理を実行
多くのモデルをサポートしており、外部のモデルのインポートや Google が事前学習させたモデルを推論エンジンを使って利用することも可能

LLM も利用可能
GENERATE_TEXT
GENERATE_TEXT_EMBEDDING

Preview
New!

Preview

BigQuery で LLM 活用

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 (Cloud SQL, Spanner, Cloud Storage 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))
```

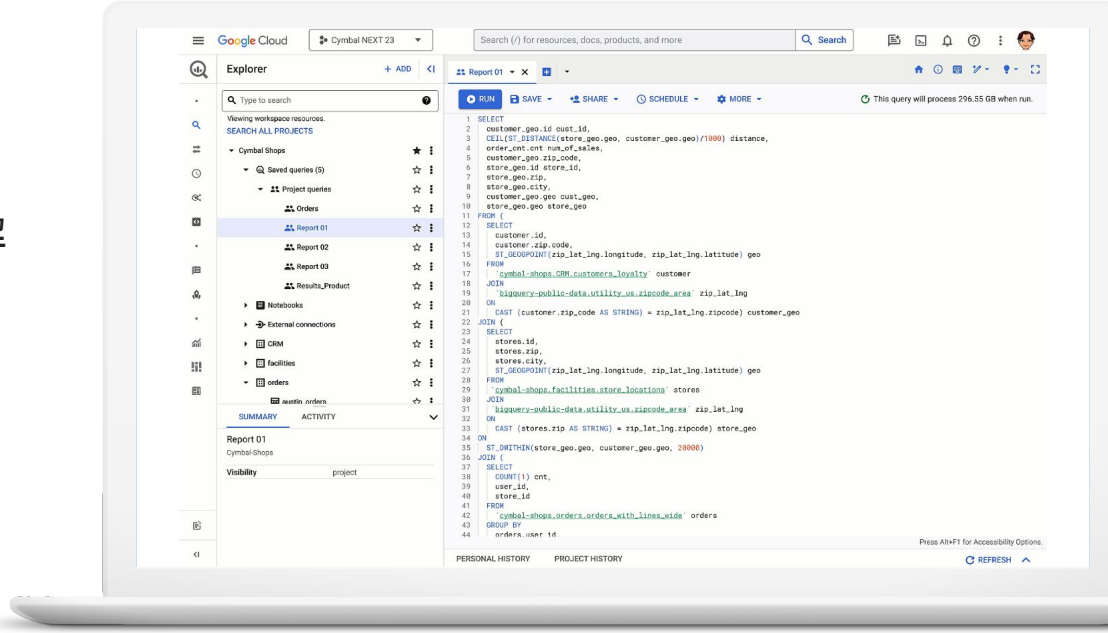
AI アシストで効率的かつ民主的に

Preview

Duet AI in BigQuery

統合された 生成 AI アシスタントですべてのデータ プラクティショナーのデータ分析を加速

- 自然言語からSQL を自動生成
- SQL の処理内容を自然言語で解説・サマリ
- SQL コード補完
- チャットとの統合



Coming soon

Duet AI in Dataplex

データ インサイトを民主化

グローバルかつAI 活用したサーチ

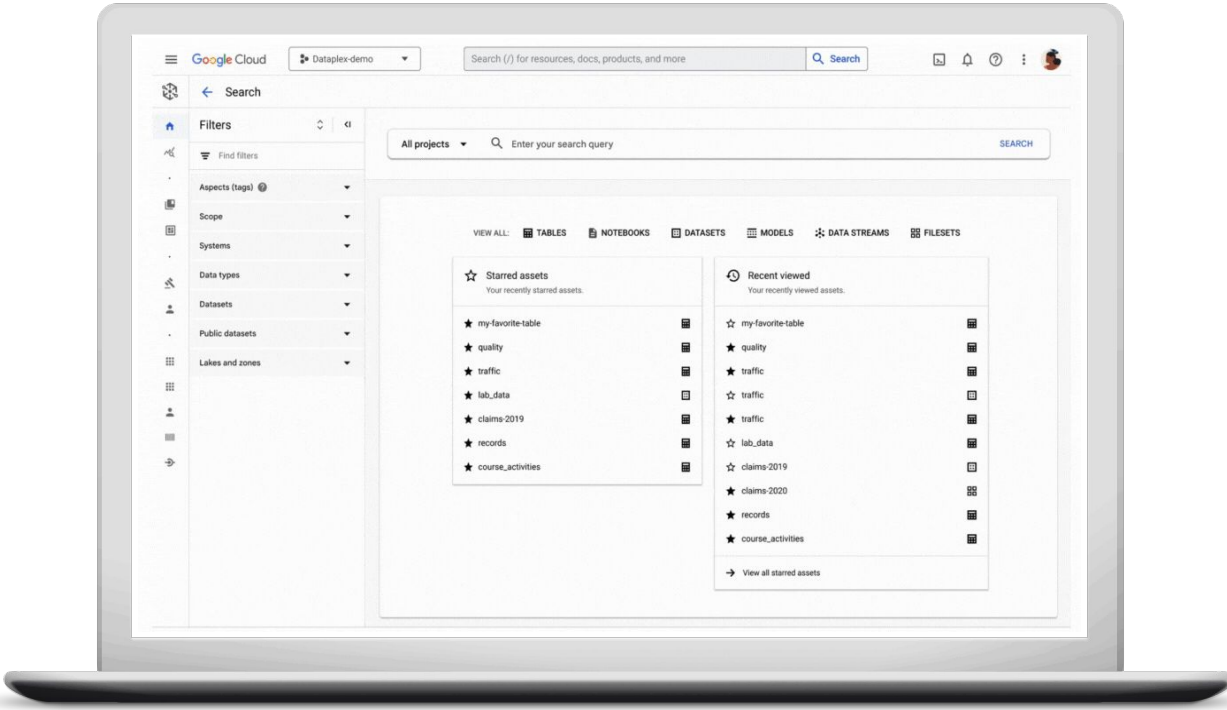
好きな言語でほしいデータセットを検索

AI を活用したデータインサイト

AI-powered data insights

キュレートされた分析の一覧表示。データに詳しくない人もデータ分析が始めやすく。

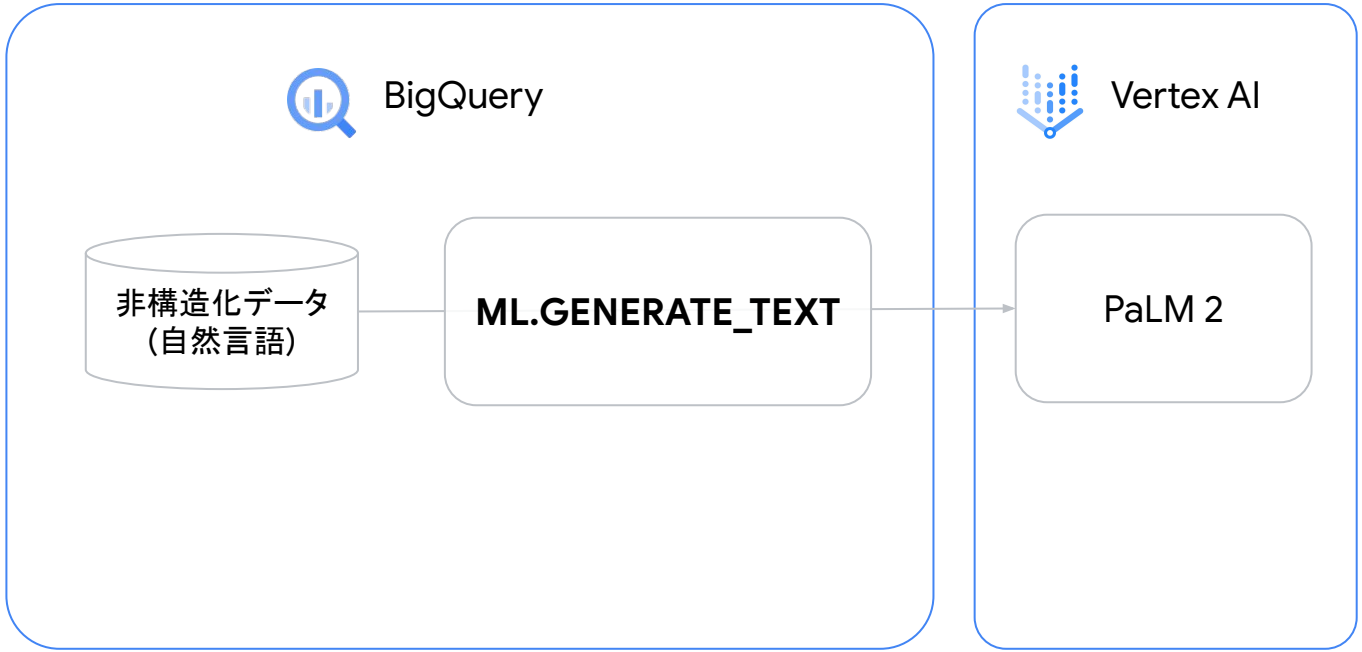
メタデータから質問を自動生成。データやクエリを検索するカタログ画面から、クエリをワンクリックで実行可能。



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

Thank you.

Google Cloud