非エンジニア向け

AI 時代を見据えた データ基盤の進化

Google Cloud Next Tokyo

國友　貴文

Takafumi Kunitomo

Google Cloud
データ アナリティクス
セールス スペシャリスト

Agenda

01 **データ基盤の変遷**

・データ基盤の役割変遷と主要技術
・生成 AI の登場によって、データ基盤に求められる要件は何が変わったか？

02 **生成 AI はデータ基盤の何を変えたか？**

・“ AI ” とは何か？
・Gemini が変えるデータ基盤の進化は何か？

03 **まとめ**

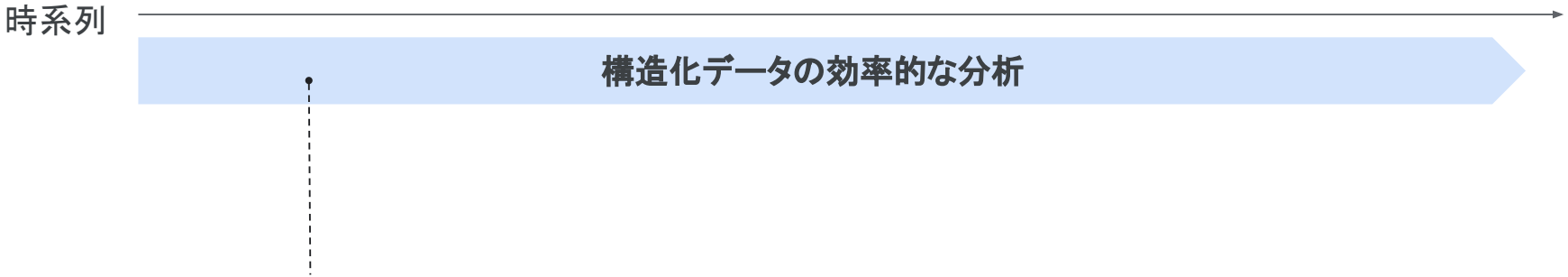
01. データ基盤の変遷

データ基盤の進化やニーズの変遷は、
取り扱うデータの種類や量 と**利用目的の変化**
と密接な関係性がある

データ基盤の役割変遷と主要技術

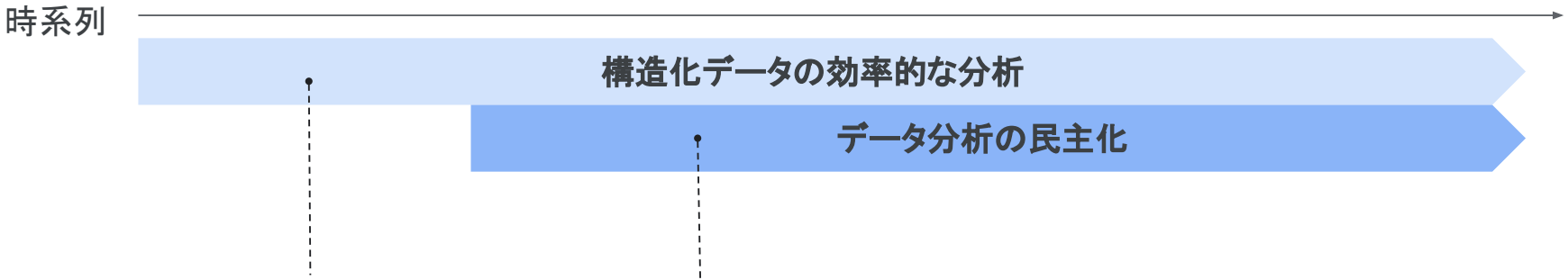
時代	主な役割	代表的な技術・概念	データタイプ	主な利用者
1990 年代	**経験・勘からの脱却、データ マイニングの本格化**	データ マイニング、RDB	構造化データ	専門家、一部のビジネス ユーザー
2010 年代前半	**大量データの効率的処理、ビッグデータ活用**	Hadoop, NoSQL, データウェアハウス	構造化データ 半構造化データ	データ サイエンティスト、IT 部門
2010 年代後半	**データの民主化、全社的データ活用**	データレイク, セルフサービス BI, データカタログ	構造化データ 半構造化データ 非構造化データ	ビジネス ユーザー、データ サイエンティスト

データ基盤の役割変遷と主要技術



- サイロ化されたデータの統合
- **PB 級の大量データの処理** スピード
- 柔軟な**スケーラビリティ** の実現
- 機械学習との統合による**”構造化データ”の将来予測**

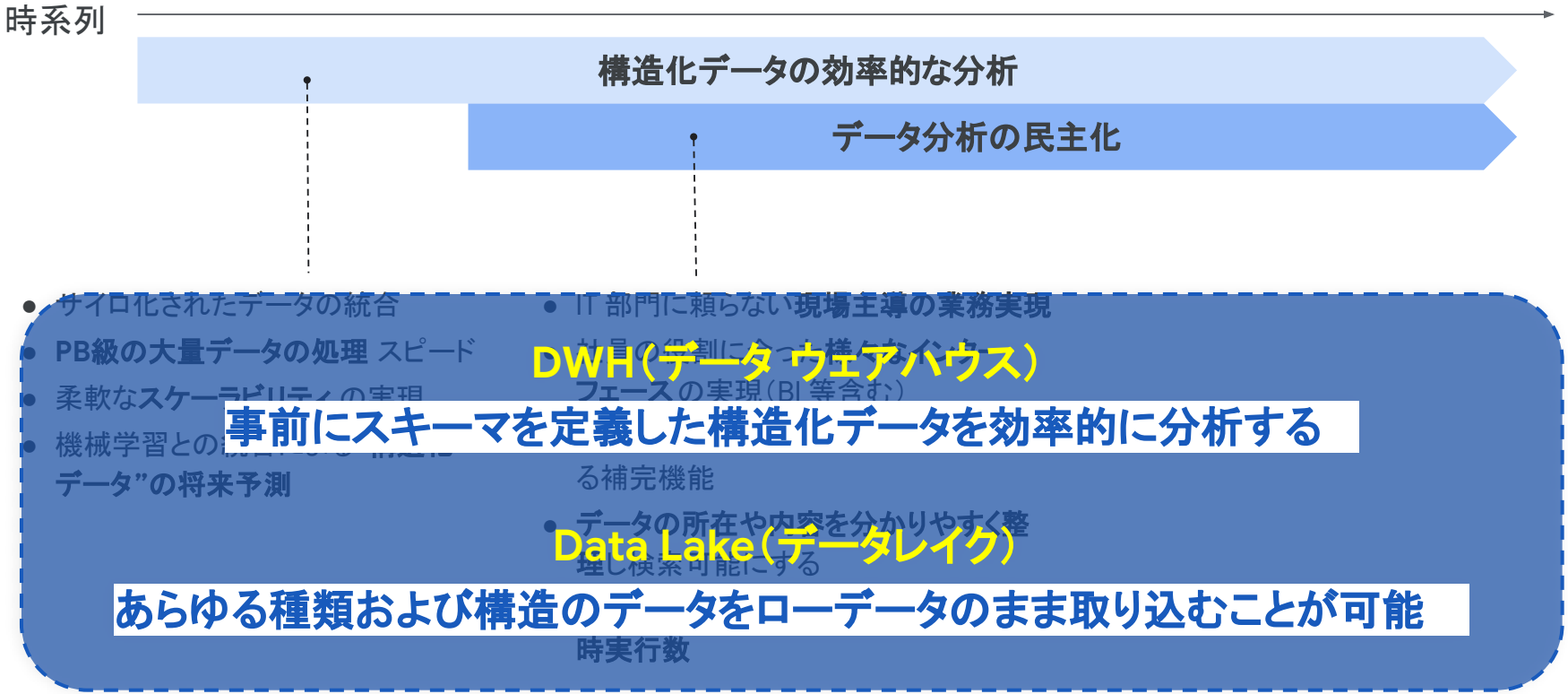
データ基盤の役割変遷と主要技術



- サイロ化されたデータの統合
- **PB 級の大量データの処理** スピード
- 柔軟な**スケーラビリティ** の実現
- 機械学習との統合による**”構造化データ”の将来予測**

- IT 部門に頼らない**現場主導の業務実現**
- 社員の役割に合った**様々なインターフェース**の実現（BI 等含む）
- IT リテラシーに依らない **UI、UX** を実現する補完機能
- **データの所在や内容を分かりやすく整理**し検索可能にする
- 多数の社員が接続しても処理可能な**同時実行数**

データ基盤の役割変遷と主要技術



そして、2020 年代に入ると、

生成 AI が登場する

生成 AI の登場によって、データ基盤に求められる要件は何が変わったか？

注：様々な要件はあるが、特に重要な内容を掲載

❶ 生成 AI のサービス / アプリが利用するデータ基盤としての役割

データ分析に携わるメンバーが利用する基盤ではなく、
生成 AI を用いたサービスが必要とするデータを取り扱う基盤 として、
より重要な役割を担うことに

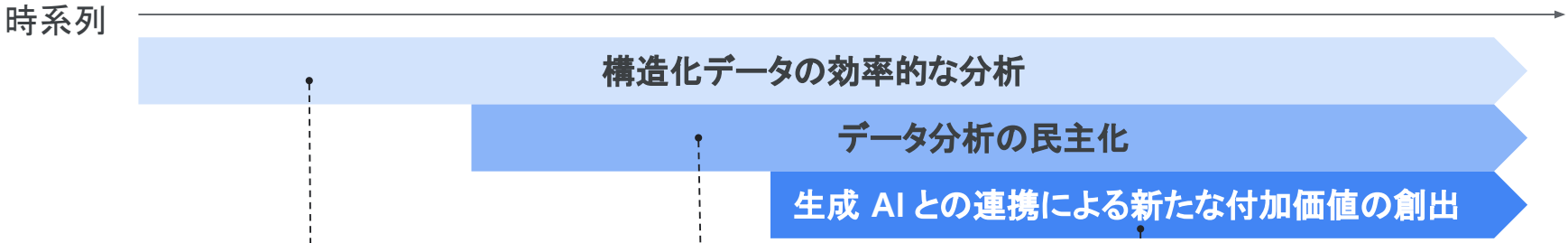
❷ 非構造化データの処理が必須要件に

生成 AI は、従来の構造化データだけでなく、テキスト、画像、音声、動画といった
「非構造化データ」を主要な入力・出力形式として扱う 。
これらの膨大な非構造化データを効率的に格納、管理、そして意味的な類似性に
基づいて照会するための「ベクトル データベース」が重要な技術に

❸ リアルタイム処理の重要性が向上

生成 AI の応用領域が、チャットボットやコンテンツ生成だけでなく、リアルタイムでの意思決定を伴うアプリケーション（例：配車アプリの価格決定 等）へと広がるにつれて、
データ基盤には「即時性」が強く求められる ように

データ基盤の役割変遷と主要技術



- サイロ化されたデータの統合
- **PB 級の大量データの処理** スピード
- 柔軟な**スケーラビリティ** の実現
- 機械学習との統合による**”構造化データ”の将来予測**

- IT 部門に頼らない**現場主導の業務実現**
- 社員の役割に合った**様々なインターフェース**の実現（BI 等含む）
- IT リテラシーに依らない **UI、UX** を実現する補完機能
- **データの所在や内容を分かりやすく整理**し検索可能にする
- 多数の社員が接続しても処理可能な**同時実行数**

- **生成 AI が必要とする様々なデータ** との連携親和性の実現
- 増え続ける**非構造化データの処理**
- **リアルタイム処理の実現**
- 単なる分析業務の枠を超えた**”新たな業務 / 付加価値 ”の創出**
- **MLOpsへの対応**
- **Data Agent を通じた自律的な分析の実現**

生成 AI の登場によって、データ基盤に求められる要件は何が変わったか？

注：様々な要件はあるが、特に重要な内容を掲載

しかし、従来の DWH や データレイクでは、生成 AI の時代において、

いくつかの課題も見られるように・・・

生成 AI の登場によって、データ基盤に求められる要件は何が変わったか？

注：様々な要件はあるが、特に重要な内容を掲載

しかし、従来の DWH や データレイクでは、生成 AI の時代において、
いくつかの課題も見られるように・・・

1

機能面での課題

・データレイクにトランザクション処理がサポートされていない
・データレイクを DWH へのデータとりこみの起点とした際、データレイクと DWH でデータが 2 重持ちに

2

パフォーマンス 面での課題

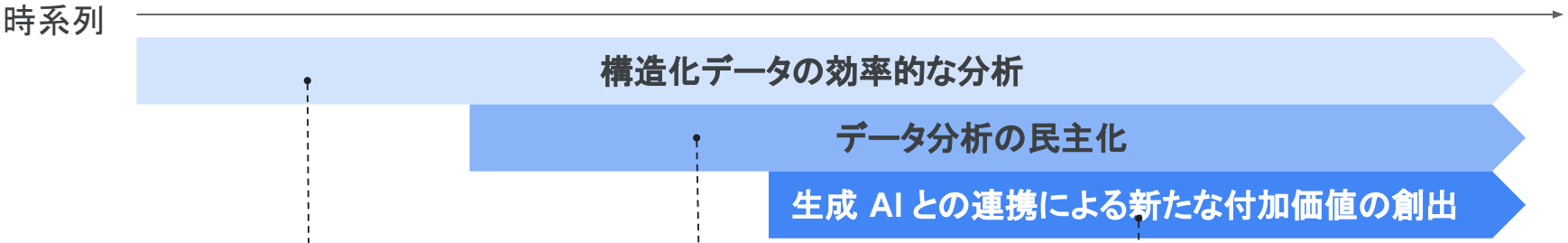
・データ可視化（BI）のように高いレスポンスが求められるワークロードに、データレイクの性能が追いつかない

3

管理・ガバナンスの課題

・様々なデータをデータレイクに入れた結果、分析に必要なデータが簡単に見つからない
・テーブルとファイルのアクセス制御方式の違いによるデータ ガバナンスの不整合

データ基盤の役割変遷と主要技術



- サイロ化されたデータの統合
- **PB 級の大量データの処理** スピード
- 柔軟な**スケーラビリティ** の実現
- 機械学習との統合による**”構造化データ”の将来予測**

- IT 部門に頼らない**現場主導の業務実現**
- 社員の役割に合った**様々なインターフェース**の実現（BI 等含む）
- IT リテラシーに依らない **UI、UX** を実現する補完機能
- **データの所在や内容を分かりやすく整理**し検索可能にする
- 多数の社員が接続しても処理可能な**同時実行数**

- **生成 AI が必要とする様々なデータ** との連携親和性の実現
- **リアルタイム処理の実現**
- 単なる分析業務の枠を超えた**”新たな業**

AI Lakehouse
（レイクハウス）
DWH とデータレイクの強みを
うまく補完するという特徴がある

データ基盤の役割変遷と主要技術

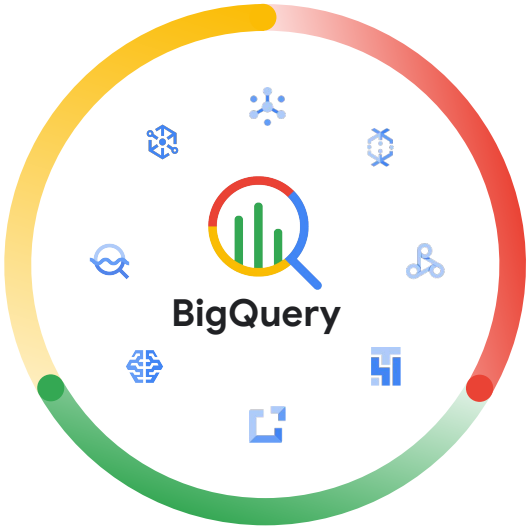
時代	主な役割	代表的な技術・概念	データタイプ	主な利用者
1990 年代	経験・勘からの脱却、データ マイニングの本格化	データ マイニング、RDB	構造化データ	専門家、一部のビジネス ユーザー
2010 年代前半	大量データの効率的処理、ビッグデータ活用	Hadoop, NoSQL, データウェアハウス	構造化データ 半構造化データ	データ サイエンティスト、IT 部門
2010 年代後半	データの民主化、全社的データ活用	データレイク, セルフサービス BI, データカタログ	構造化データ 半構造化データ 非構造化データ	ビジネス ユーザー、データ サイエンティスト
2020 年代以降 (生成 AI による変化)	生成 AI 機能の基盤、AI サービス支援	**ベクトル データベース , RAG, MLOps, ストリーミング パイプライン (リアルタイム処理)**	構造化データ 半構造化データ 非構造化データ	AI 開発者、ビジネス ユーザー、顧客

AI 時代には、新しいデータプラットフォームが必要

BigQuery なら、DWH かデータレイクか、どちらかを選ぶ必要はありません。Google のデータクラウドは、両方の長所を提供しながら、生成 AI の力を活用できます

	BigQuery
ユースケースの特性	**構造化データと非構造化データの力を組み合わせて** 、何が起こったのかという質問に答え、将来を予測するための AI / ML モデルを構築
データタイプ	構造化、非構造化、ストリーミング データ
データへのアクセス	自然言語、SQL、プログラミング言語
データ ガバナンスとセキュリティ	統合

BigQuery を中心としたデータ分析の簡素化と統合



データから AI まで全ての統合プラットフォーム

統一された エクスペリエンス	データ統合	統合エンジン
Single product UX	Structured / unstructured	SQL
New GenAI powered experiences	Iceberg / Delta / Hudi	Spark
Collaborative workflows	GCS	Python
	Cross-cloud (AWS, Azure)	Remote functions
	Unified governance	Business intelligence

上記を下支えする AI / ML / LLM の統合

Auto ML

BQML

LLM (Gemini)

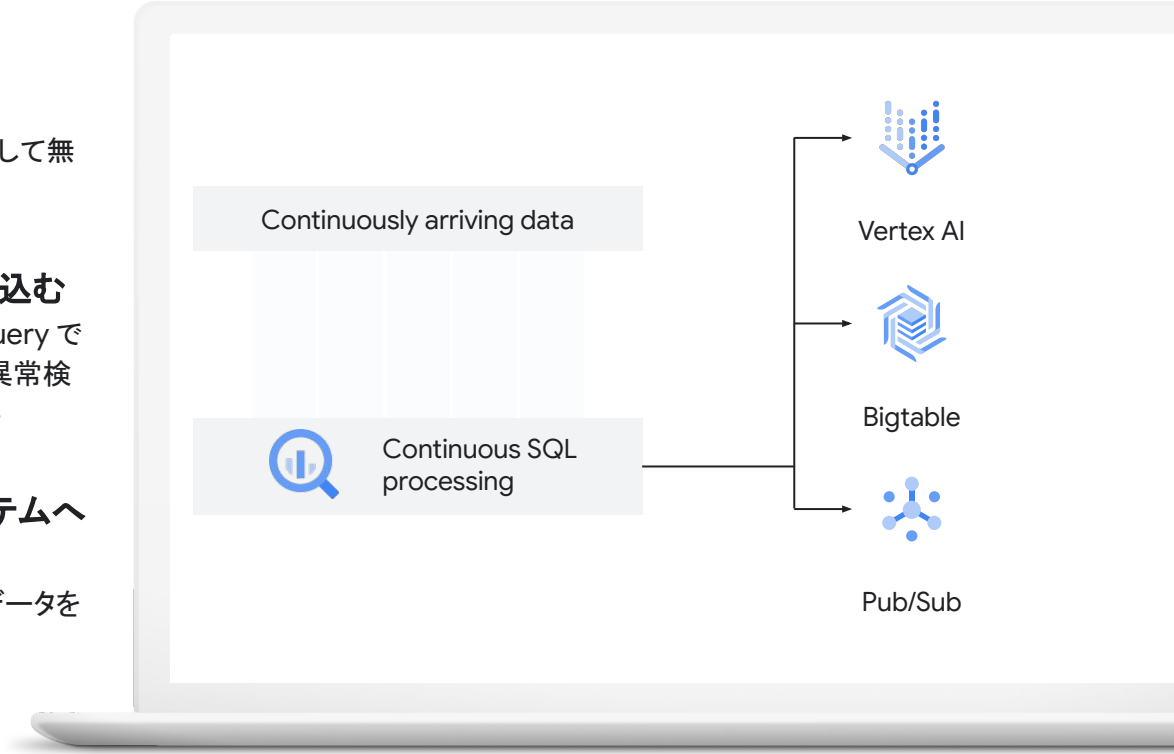
BigQuery で リアルタイム ストリーム 分析

Continuous queries

Continuous analytical processing
SQL を使用して、受信データのストリームに対して無制限のサーバーレス分析を実行

深い洞察を得るために Gen AI を組み込む
ML と Vertex AI モデル機能を使用して BigQuery で自動化パイプラインを開発し、リアルタイムの異常検出、予測、感情分析、推奨アルゴリズムを提供

BigQuery からオペレーショナル システムへのリバース ETL
BigQuery から Pub/Sub または Bigtable にデータをプログラムでリアルタイムに変換および複製

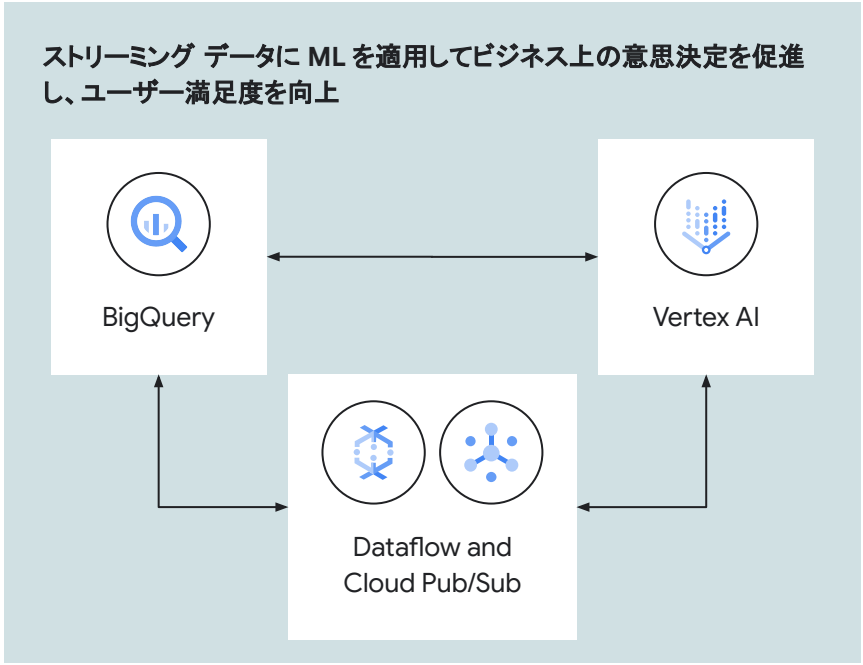


Streaming ML for GenAI

バッチ機能/予測からストリーミング機能とオンライン予測

不正検出、製品推奨、動的価格設定、 IoT ヘルスモニタリング、インテリジェント監査、セキュリティにおけるリアルタイムの意思決定を強化

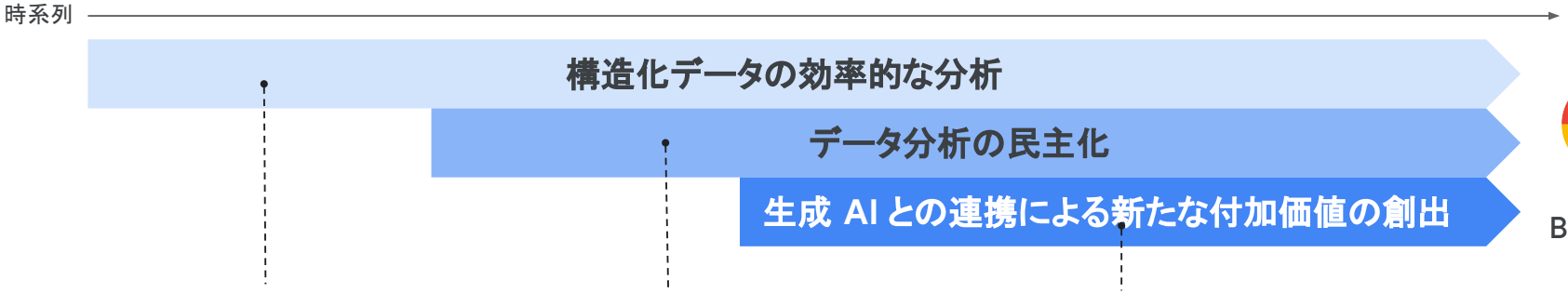
- リアルタイム ML を可能にするリアルタイム機能処理
- ストリーミング データをヒストリカルデータで強化する
- オペレーティング メッセージングとデータベースを統合して予測をアクティブ化
- Vertex AI 機能と統合して高度な ML 機能を実現
- ユースケースに適したサイズ : GPU ベースのローカル予測と Vertex Online 予測の統合を選択



あらゆる種類のデータとワークロードに対応する BigQuery

- **すべてのデータのための単一の統合アクセス層** ：構造化データ、非構造化データ、運用ストリームなど、あらゆるデータが対象
- **あらゆるデータ形式とストレージをサポート** ：S3、Azure Storage、Iceberg、Delta Lake、Hudiなど、多様なストレージに対応
- **SQL、オープンソースエンジン、 AI/ML のための単一の共有メタストア** ：すべてのデータアクセスを一元管理
- **あらゆるデータに対する大規模なガバナンスの実現:** データの整合性とセキュリティを確保

データ基盤の役割変遷と主要技術



- サイロ化されたデータの統合
- **PB 級の大量データの処理** スピード
- 柔軟な**スケーラビリティ** の実現
- 機械学習との統合による**”構造化データ”の将来予測**

- IT 部門に頼らない**現場主導の業務実現**
- 社員の役割に合った**様々なインターフェース**の実現（BI 等含む）
- IT リテラシーに依らない **UI、UX** を実現する補完機能
- **データの所在や内容を分かりやすく整理**し検索可能にする
- 多数の社員が接続しても処理可能な**同時実行数**

- **生成 AI が必要とする様々なデータ** との連携親和性の実現
- 増え続ける**非構造化データの処理**
- 単なる分析業務の枠を超えた**”新たな業務 / 付加価値”の創出**
- **Data Agent を通じた自律的な分析の実現**

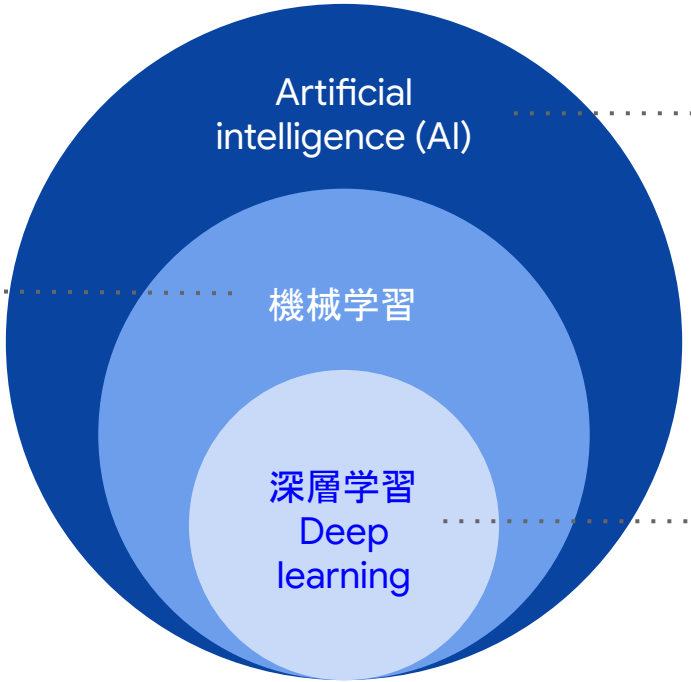
02. 生成 AI はデータ基盤の何を変えたか？

“ AI ” とは何か？

The capability of a machine to imitate intelligent human behavior, such as:

- Reasoning
- Learning
- Natural language processing

機械が人間の知的な行動を模倣する能力。
例:
推論、学習、自然言語処理



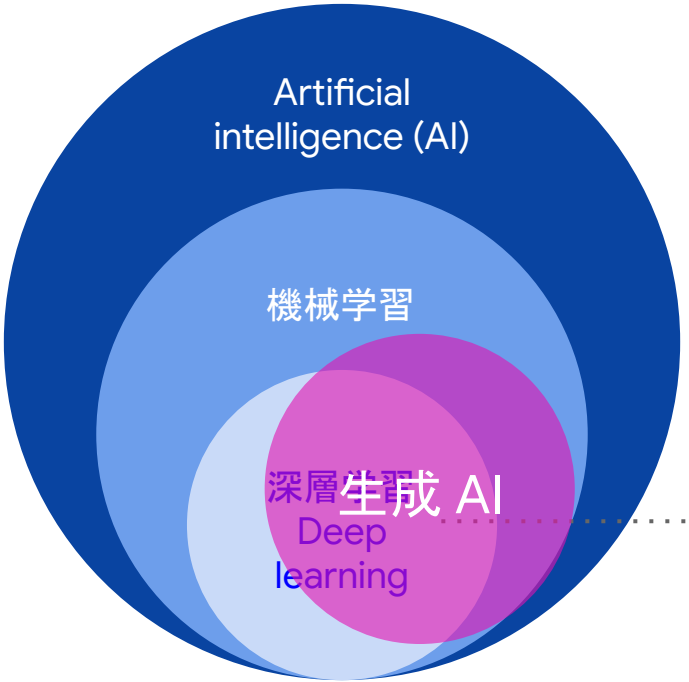
The capability of a machine to “learn” — by using data to improve performance on a specific task — without being explicitly programmed

明示的にプログラムすることなく、データを使用して特定のタスクのパフォーマンスを向上させる機械の「学習」能力

A class of machine learning algorithms that use a cascade of nonlinear processing units to learn in supervised or unsupervised ways

非線形処理ユニットのカスケードを使用して、教師ありまたは教師なしの方法で学習する機械学習アルゴリズムのクラス

“ AI ” とは何か？



生成 AI は
Deep Learning
のサブセット

大規模言語モデル (LLMs)
もまた、
Deep Learning
のサブセット

Gemini

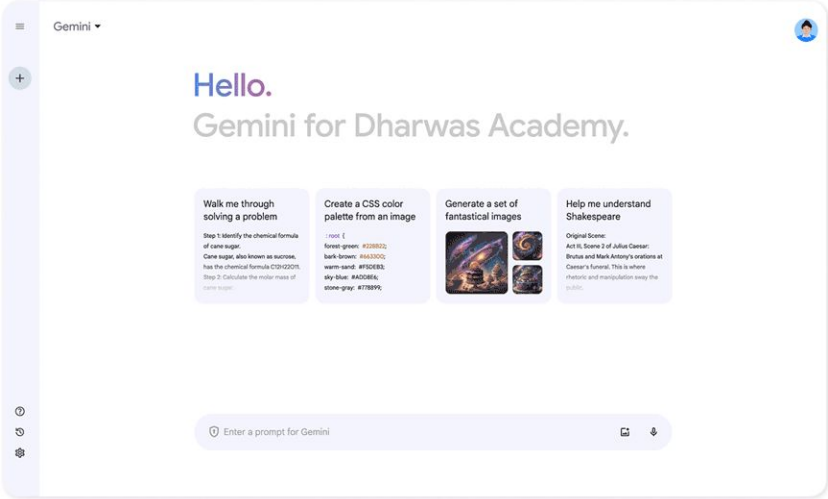
Gemini とは?

あなたの創造力を引き出し、
生産性を高める、Google の AI です

Gemini を動かしているのは ?

Gemini は、Google の最上位 AI モデルの名称でもあり、
テキストだけでなく画像や動画なども一緒に処理できる、マルチ モーダルモデルの大規模言語モデル
(LLM, Large Language Model) です

※ LLM とは、ユーザーのプロンプトとこれまでに生成されたテキストに基づいて、次に来るであろう単語を予測できる生成AI です

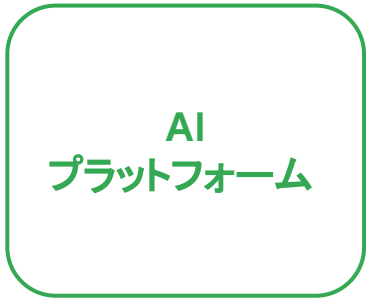


gemini.google.com

Gemini が変えるのは何か？

Gemini が変えるデータ基盤の進化は何か？

データ基盤
“ 自体に ”
生成 AI が組み込
まれる



例
・SQL を使わなくても、**自然言語でデータの取扱い**ができる（機能名 : Gemini in BQ）
・データの型が異なっても **意味を理解して整形 / 加工** を実施する（機能名 : Data preparation）
・**内包された MLモデル** によって、高度な予測が可能になる（機能名 : AutoML、BQML 等）

Gemini エコシステム

Gemini モデルを以下の形態で、コンシューマ及び企業のユーザーに提供

Gemini モデルファミリー　Ultra　Pro　Flash　Nano

単独のサービスとして

Gemini アプリ

Gemini

Gemini.google.com

NotebookLM

EXPERIMENTAL

各種 Google サービス・製品に組み込み

Google Workspace

Gemini 搭載

Android　Pixel　chromebook

開発者向け：サービス構築における API 活用

Google AI Studio

Google Cloud

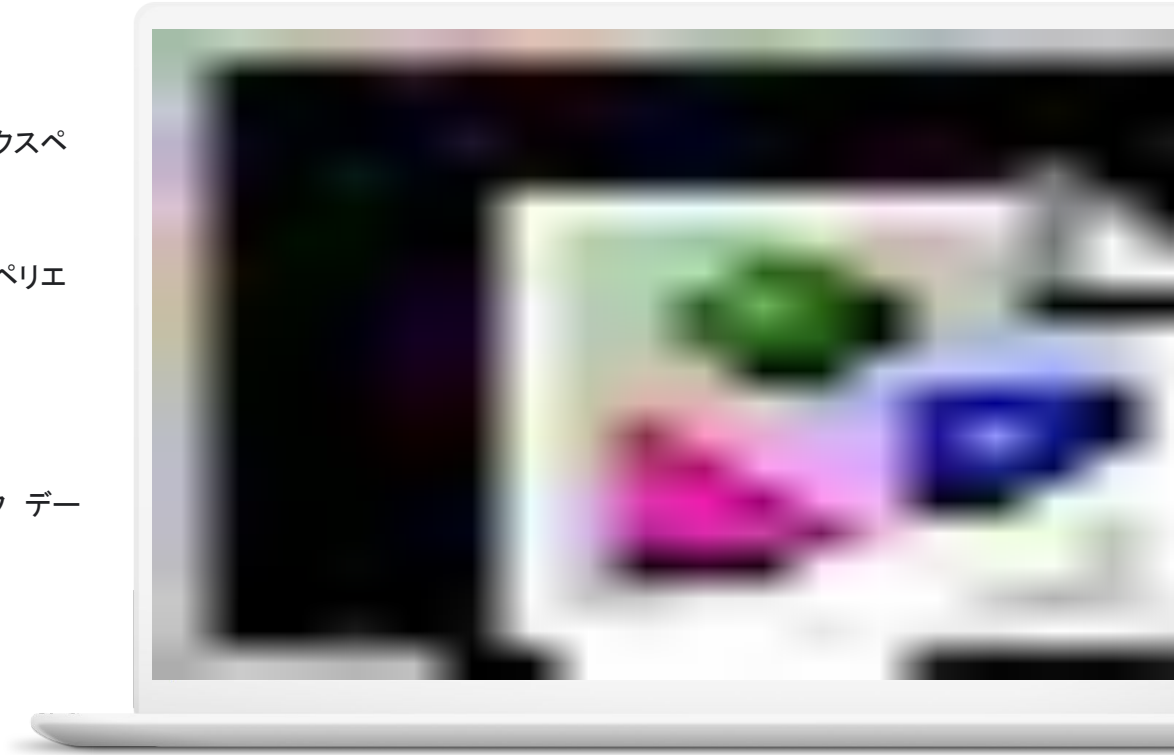
Vertex AI

注: 一部サービスのみ例示

ビジュアルと AI によるデータ検出、モデリング、分析

BigQuery Data Canvas

- データ探索と視覚化のための GenAI 中心のエクスペリエンス
- インタラクティブかつガイド付きのユーザー エクスペリエンス
- BigQuery Studio に統合
- Dataplex カタログでサポートされるセマンティック データ検出
- 自動化された Python ノートブックの生成
- データ アナリスト向けの組み込みコラボレーション

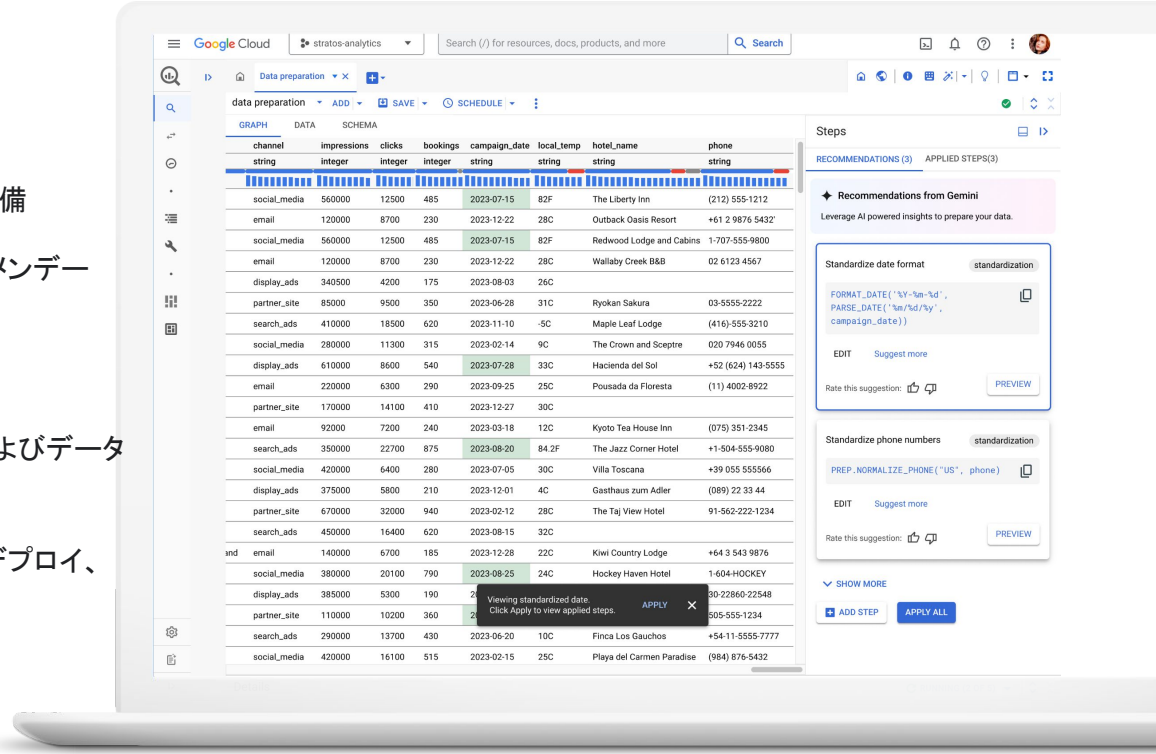


BigQuery data preparation

生成 AI アシストで進化した直感的な画面操作可能なデータ プリパレーション

BigQuery Data Preparation

- BigQuery でのデータのクレンジング、変換、準備
- 生成 AI によりコンテキストを認識した変換レコメンデーションにより、データ開発を加速
- 合理化された取り込みと変換で時間を節約
- 自動化されたスキーマ マッチング、ビジネスおよびデータ品質ルールの識別と検証
- 生成された SQL パイプラインを BigQuery にデプロイ、オーケストレーション、モニタリング



Gemini が変えるデータ基盤の進化は何か？

データ基盤

AI
プラットフォーム

“ 生成 AI を用いて ”
新しいサービスを
創り出す

例
・生成 AI モデルを活用した、従来とは異なる顧客体験 / 業務プロセスを実現するアプリケーション / サービス
・非構造化データから何かを判断し、意思決定につなげていくアプリケーション

Vertex AI

AI モデル及びエージェント開発のためのエンド ツー エンドのプラットフォーム

Agent Builder：エージェントの構築を担うレイヤー

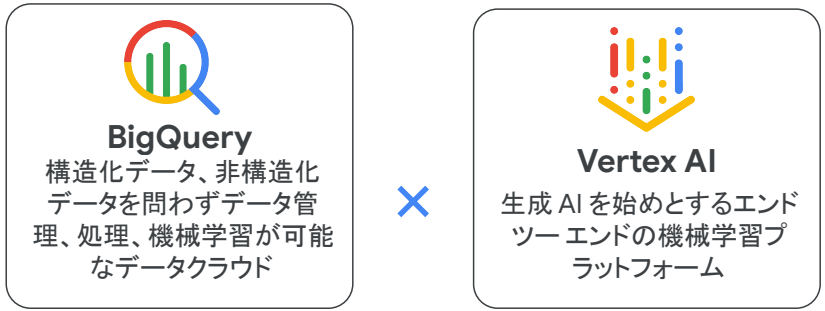
Model Builder：生成 AI モデルのカスタマイズを担うレイヤー

Model Garden：生成 AI モデルを提供するレイヤー

Google モデル | パートナーモデル | オープンモデル

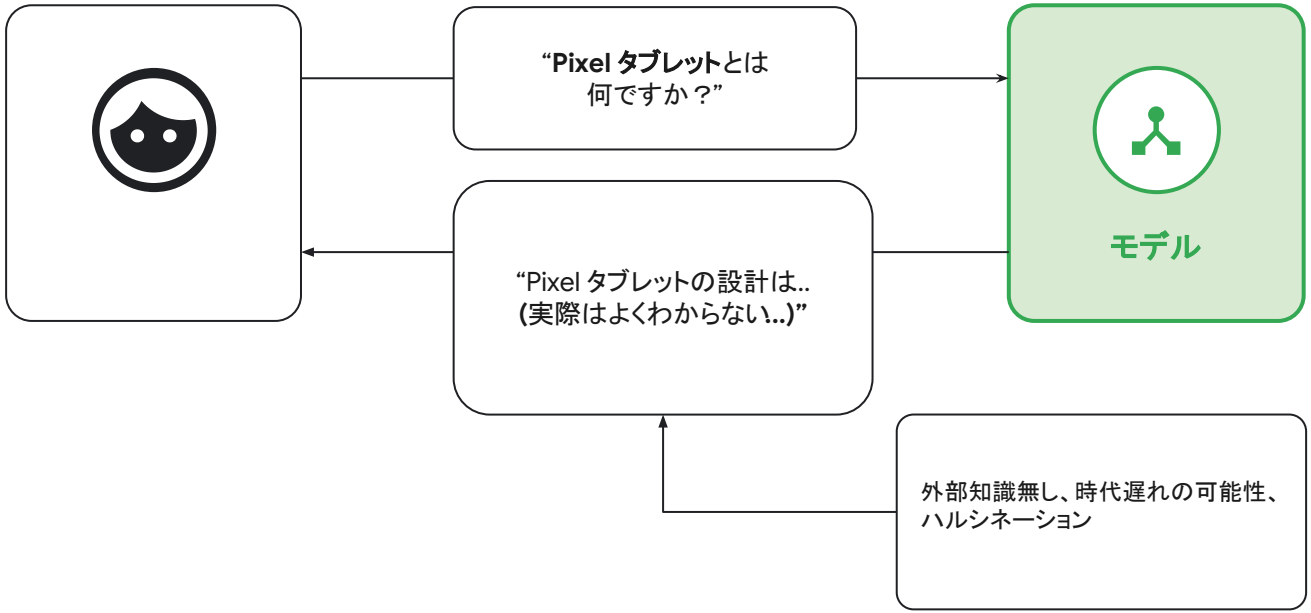
- Gemini
  Imagen / Veo
  など
- Claude
  (Anthropic)
  など
- Llama
  DeepSeek
  など

データ基盤と AI プラットフォームの関係



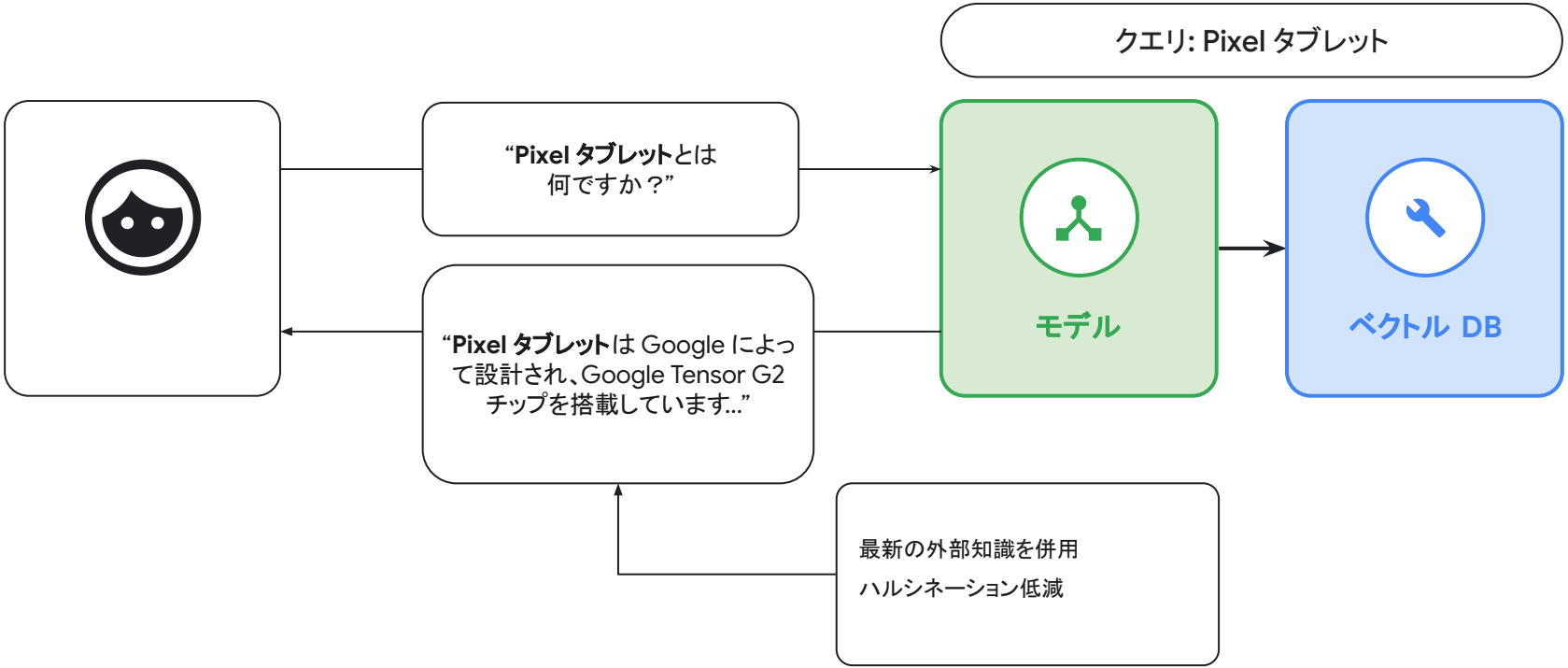
LLM アプリケーションの簡単な歴史

初期にはモデルのみが存在



LLM アプリケーションの簡単な歴史

...その後、検索拡張生成（RAG）が登場



生成 AI の仕組みを作るために必要となるデータは何が考えられる？？

サービスは様々であるため、記載の内容全てが必要というわけではありませんが、必要となる可能性がある内容を記載しております。

Fine Tuning / Supervised Tuning といった際の **追加学習用データ** や、プロンプト エンジニアリングで使った **インプットデータ**、RLHF の仕組みでフィードバックされたデータ 等

LLM が知識として持っていないが、利用する必要がある **拡張データ**

LLM が理解できるように処理（ Embedding）された際の **結果データ**

サービス・企業が保有する商品情報やサービス仕様といった **個別データ**
（構造化・非構造化の両方が対象として考えられる）

個別の Recommendation をするための **ログデータ** や、生成 AI が正確にデータの中身を把握できるようにするための **メタデータ**

生成 AI の仕組みを作るために必要となるデータは何が考えられる？？

サービスは様々であるため、記載の内容全てが必要というわけではありませんが、必要となる可能性がある内容を記載しております。

Fine Tuning / Supervised Tuning といった際の **追加学習用データ** や、プロンプト エンジニアリングで使った **インプットデータ**、RLHF の仕組みでフィードバックされたデータ 等

LLM が知識として持っていないが、利用する必要がある **拡張データ**

LLM が理解できるように処理（Embedding）された際の **結果データ**

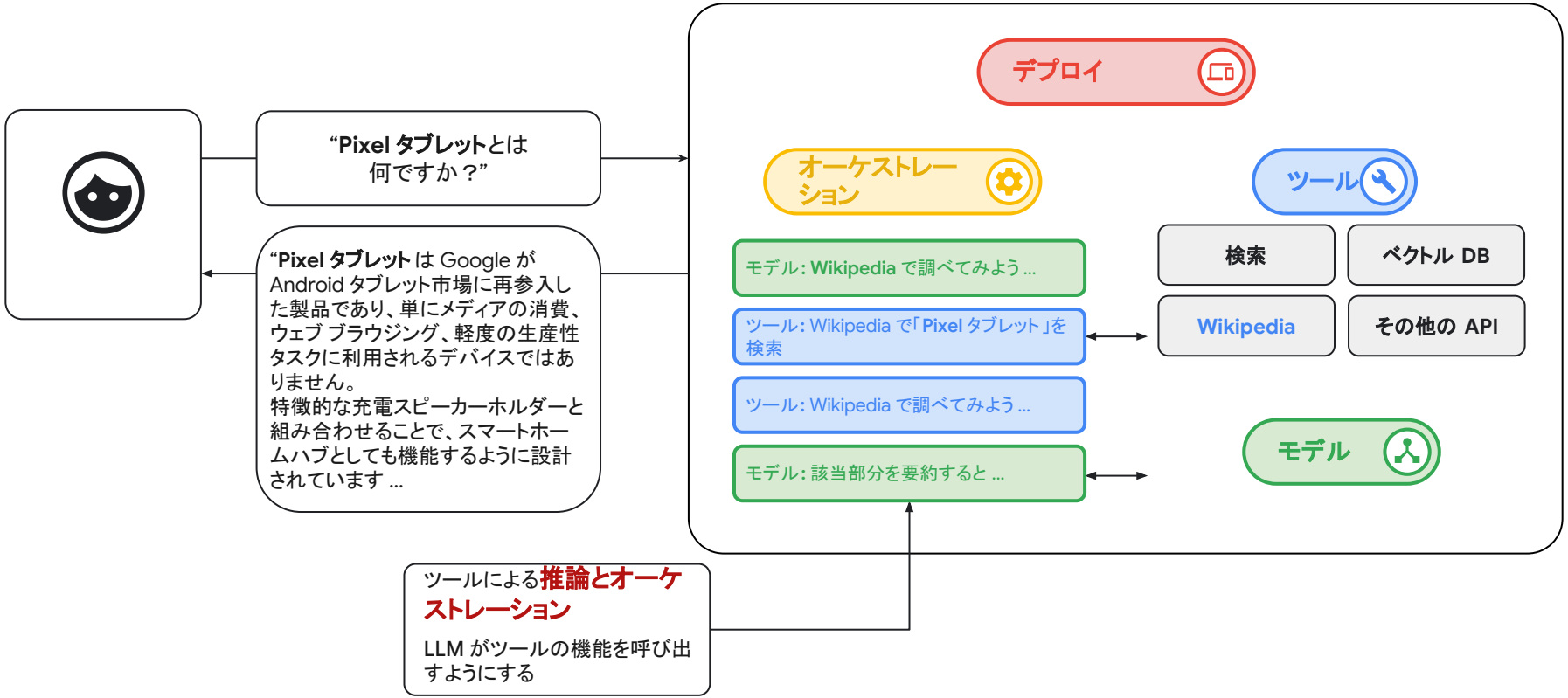
サービス・企業が保有する商品情報やサービス仕様といった **個別データ**
（構造化・非構造化の両方が対象として考えられる）

個別の Recommendation をするための **ログデータ** や、生成 AI が正確にデータの中身を把握できるようにするための **メタデータ**

これらの “ データ ” が【効果的に使える状態になっていること】が重要

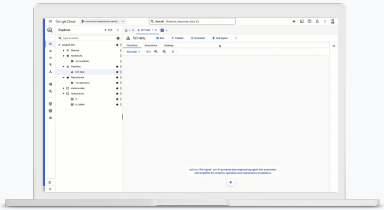
LLM アプリケーションの簡単な歴史

...そして推論とオーケストレーション を備えた生成 AI エージェントへと進化



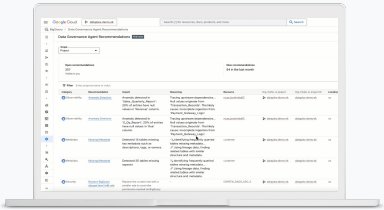
自律的に判断する Agent が搭載されたデータ基盤の進化

データ エージェント ファミリー



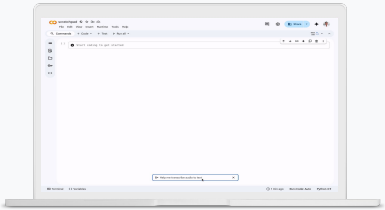
Data Engineering Agent

データ準備、データ取り込み、データ クリーニング、ブロンズ→シルバー→ゴールドの自動化



Data Governance Agent

データ モデリング、メトリック定義、メタデータ生成



Data Science Agent

Colab ノートブック エージェント、探索的データ分析 (EDA)、データ可視化、予測モデリング

Conversational Analytics Agent

Data QnA, フォーキャスト,

マルチモーダル検索

データ基盤と AI プラットフォームの関係

AI プラットフォーム側で作られたサービスやアプリケーションは

“ 単体 “ では価値を生み出せない 。

データ基盤との “ 連携の親和性 “ が成り立って初めて価値を創出する

03. まとめ

お客様との接点の多様化に伴って求められるデータ＆AI 基盤高度化

データの種類・連携方法の多様化により、データ基盤に求められる要件が高度化しているため、**”高度化された AI lakehouse としてのデータ基盤 ”**と**”サービスを実現する AI Platform”**、そしてそれらの**シームレスな連携** が必須となる