Session Report

企業がリスクを乗り越えて
生成 AI ソリューションを活用するには

生成 AI のもたらすインパクトとエンタープライズ利用における検討ポイント

Google Cloud
カスタマー エンジニア
遠山 雄二

Google Cloud

Google Cloud

セッション レポート概要

生成 AI には大きなケイパビリティがあるがゆえに、多くのユースケースが検討されています。ただし、エンタープライズでの利用時には一定のリスクが伴います。ここでは、検討すべきポイントと必要な対応について、システム構成の観点から説明します。

プレゼンター紹介

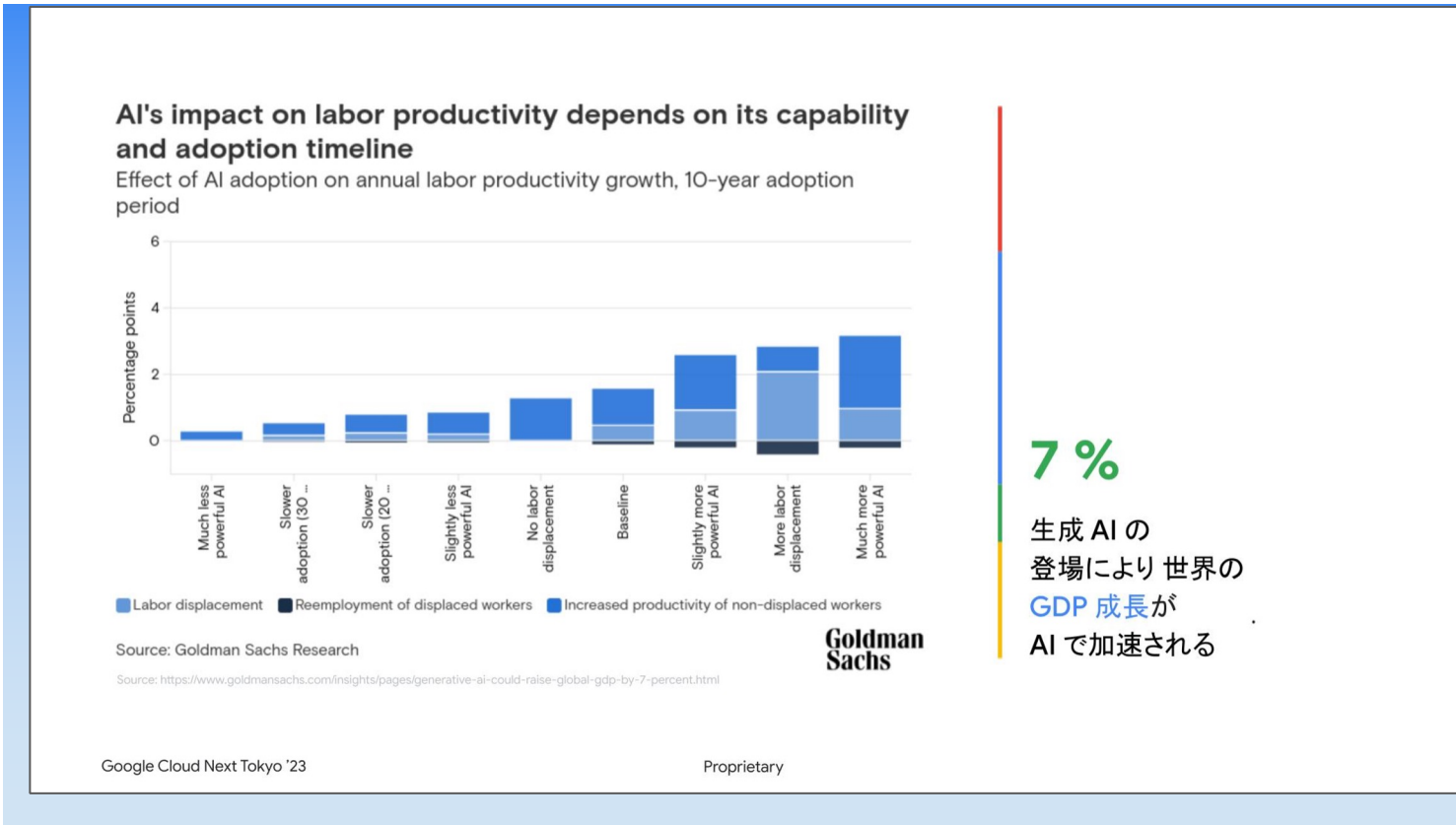
Google Cloud
カスタマー エンジニア
遠山 雄二

技術知識を生かしてビジネス課題を解決することに強い情熱を持っており、Google Cloud では業種/技術問わず、フルスタックで案件を支援しています。最近は生成 AI のエンター プライズ利用に関する、多くのお客様の課題を解決できるよう努めています。

目次

生成 AI のもたらすインパクトとリスク

ゴールドマン・サックスは、生成 AI が世界の GDP 成長を 7% 促すと予測しています。しかし、生成 AI 活用には、もっともらしい誤情報を生み出す「ハルシネーション」を始めとするさまざまなリスクが存在します。企業が生成 AI を利用するためには、AI 倫理を考慮したサービスを構築する必要があります。



生成 AI が世界の GDP 成長を加速させる

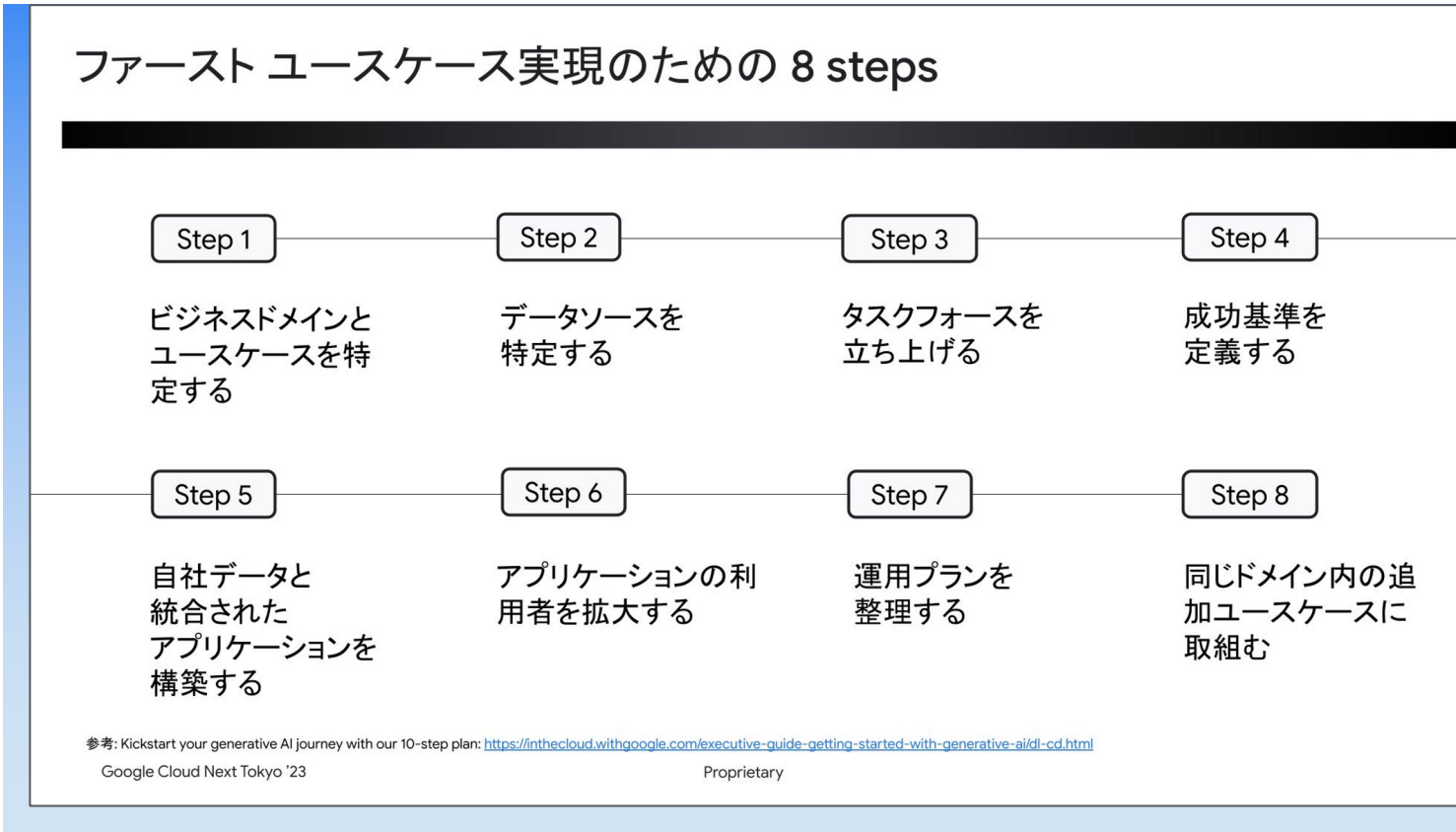
Google Cloud でも「AI Principles」という原理・原則を制定しています。その取り組みの一環として、画像生成 AI ツールの「Imagen（イマージェン）」で生成された画像をメタデータから特定できる機能を開発したり、大規模言語モデル（LLM）の「PaLM」で不適切な生成文章をフィルターにかける機能を展開したりしました。

このように、企業が生成 AI を利用する際には、リスクに適切に対処する方法を用意する必要があります。

エンタープライズでの生成 AI 検討ステップ

企業が生成 AI ソリューションを導入する際は、どのようなステップを踏むべきでしょうか。今回は、特にシステム構成面でどのような考慮が必要になるかを解説します。

ファースト ユースケースを実現するためには、下図のような 8 つのステップを経て取り組みを進めると良いでしょう。



ファースト ユースケースを実現する 8 ステップ

この 8 ステップは私が携わった生成 AI プロジェクトを踏まえ、共通点をポイントとして抜粋・整理したものです。必ずしも正解とは限らないので、状況に応じて参考にしてください。以下、それぞれのステップの詳しい内容を説明します。

(1) ビジネス ドメインとユースケースを特定する

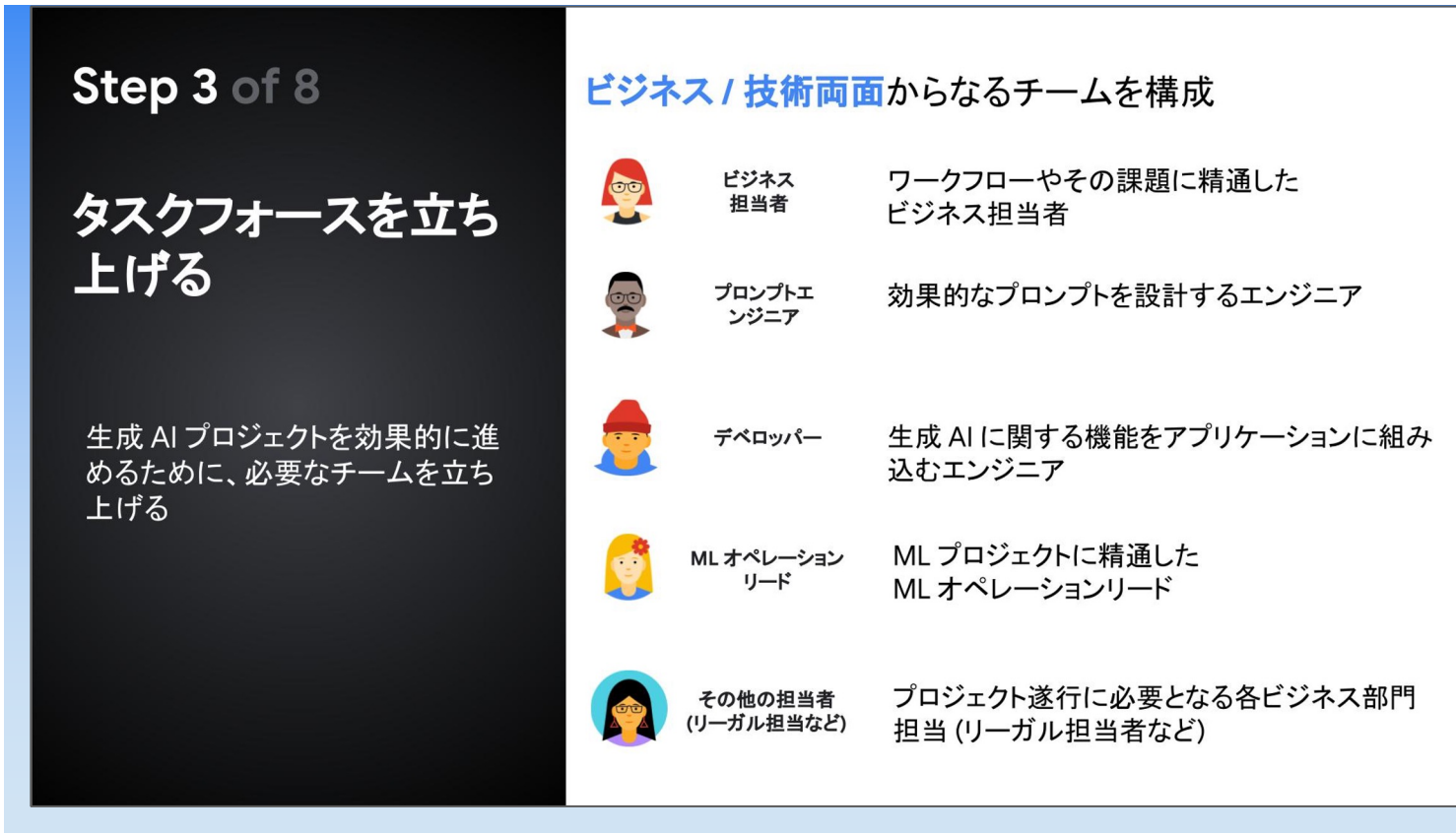
「ビジネス インパクト」と「フィジビリティ（実現可能性）」の観点から、ビジネス ドメインとユースケースを特定します。繰り返し作業や問い合わせ対応は生成 AI の得意とする分野で、使えば業務を大幅に効率化できますが、導入には一定量のデータが必要です。必要な量のデータを用意できるか、生成 AI を利用する時のリスクを許容できるかを検討します。

(2) データソースを特定する

特定したユースケースを実現するために、どこにある何のデータが必要か、また、どのようなガバナンスを設けるべきかを確認します。生成 AI がアクセスするデータの種類や格納場所を整理し、セキュリティ事故を防ぐために、データのアクセス権管理の方法や機密情報の有無なども整理しましょう。システム構築時の重要なインプット情報となるので、しっかりと整理することが大切です。

(3) タスクフォースを立ち上げる

プロジェクトを進める上で必要なメンバーを集めたタスクフォースを立ち上げます。ポイントは、ビジネス・技術両面に知見のあるチームになるようにする点です。最低限必要なメンバーは、ユースケースに詳しいビジネス担当者と各エンジニアです。法的検討が必要な場合は法務担当を含める場合もあるでしょう。ユースケースに応じてメンバーを選びます。



ユースケースに即したメンバーを選定

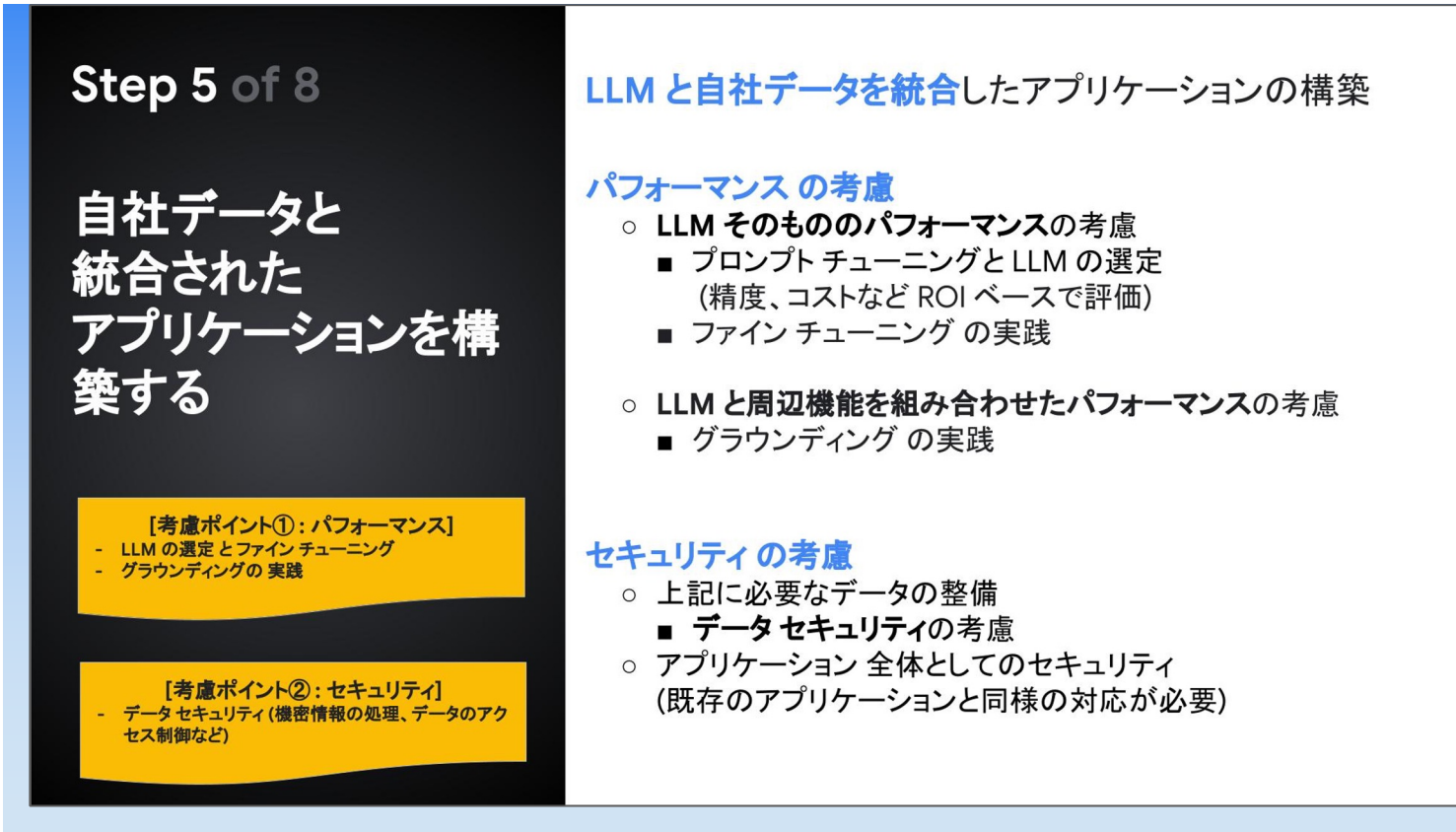
(4) 成功基準を定義する

ビジネス・技術両面から KPI を検討することが大切です。よく用いられる KPI は生成 AI の出力精度ですが、費用対効果が見合わない場合は残念な結果になりかねません。参考となる KPI は、生産性や顧客満足度、コスト削減効果、工数削減効果、エラーレート、学習時間・コスト、スケーラビリティ、規約・コンプライアンス準拠、ヒューマン インザループ指標などです。

(5) 自社データと統合されたアプリケーションを提案する

今回は、自社のドキュメント データなどに基づき質問に回答する「情報検索ソリューション」の構築を例に考えます。LLM に自社要件に基づいた回答をさせるには、「ファイン チューニング」と「グラウンディング」の機能を実装することが重要です。グラウンディングは、ハルシネーションを防ぐ上でも有効とされています。

また、生成 AI が機密情報を学習する事故が起きないよう、データ セキュリティを考慮することも大切です。

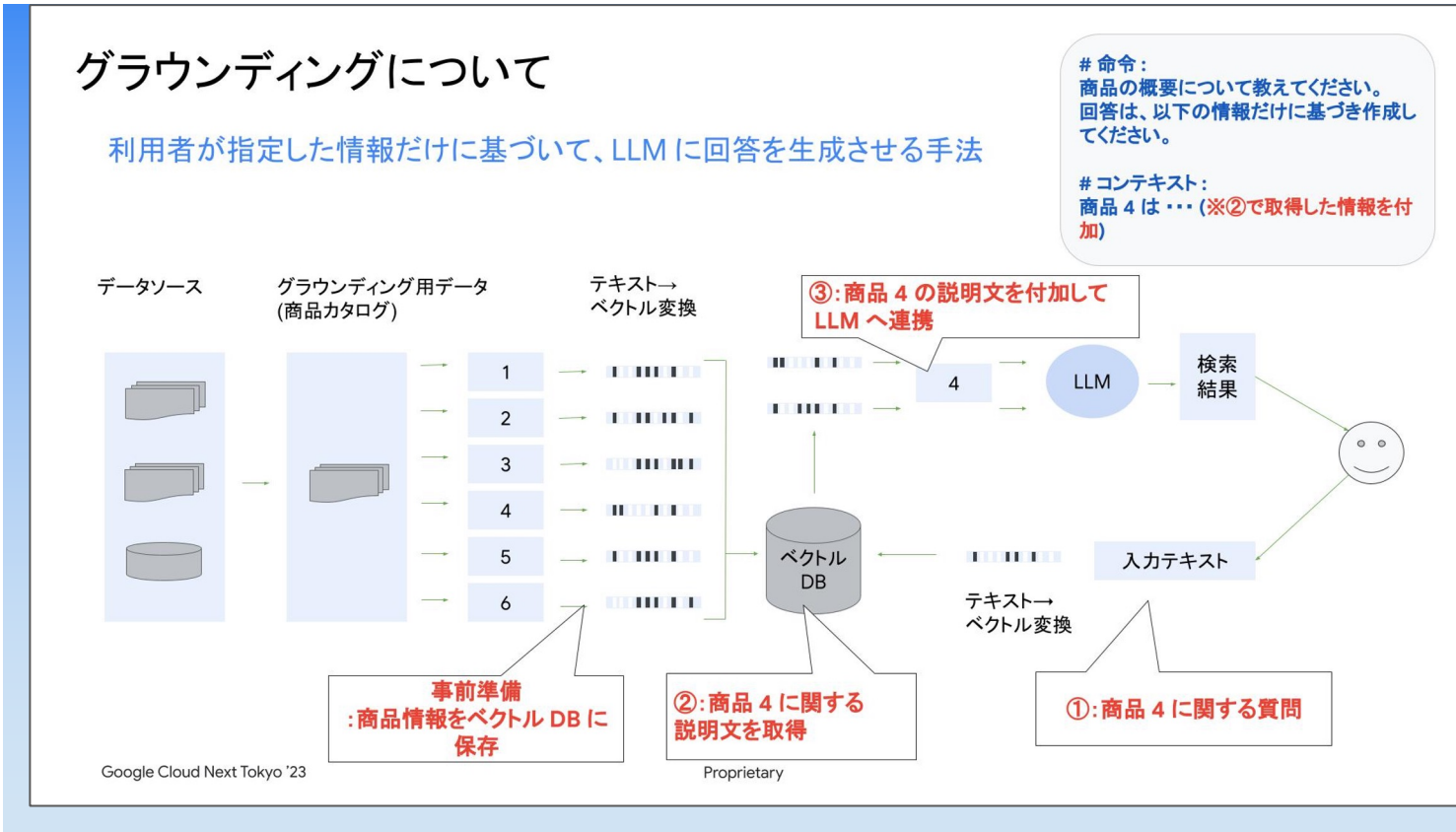


パフォーマンスとセキュリティ面の考慮が必要

補足：グラウンディングとは？

グラウンディングとは、LLM に指定した情報に基づいた回答をさせる手法です。代表的な構成として、ベクトル検索技術を利用する RAG と呼ばれるアーキテクチャについて解説します。ここでは、自社商品のカタログデータに基づき、エンドユーザーからの質問に回答する情報検索ソリューションの構築を例に考えましょう。

ユーザーから「商品 4」に関する質問を受けた際の動きは下図の通りです。事前準備の段階では、商品カタログをベクトル化し、ベクトル DB に保存しておきます。



グラウンディングを活用した検索の仕組みと実際の流れ

まず、質問文をベクトルに変換し、ベクトル DB に保存してある距離の近いベクトルを取得します。ここでは、商品カタログに記載された商品 4 の情報を取得します。そして、商品 4 の説明文を利用して LLM へ問い合わせします。

ポイントは「以下の情報だけに基づき作成してください」と命令し、コンテキストにベクトル DB で取得した情報を当てはめる点です。Google Cloud でグラウンディングを実現する方法は後述します。

(6) アプリケーションの利用者を拡大する

ユーザーからのフィードバックを得るステップです。社内のユーザーに積極的にアプリケーションを共有し、ユーザー インタビューやサーベイを実施したり、ワークショップなどを開催してフィードバックを得ます。フィードバックを基に、アプリケーションの改善ポイントを探りましょう。

(7) 運用プランを整理する

アプリケーションの中長期的な運用を見据え、モニタリング戦略など、運用方針を整理します。

システム構成面では、オブザーバビリティを考慮する必要があります。継続的に価値を得るには、アプリケーションの利用状況を確認し、改善する仕組みを検討することが重要です。

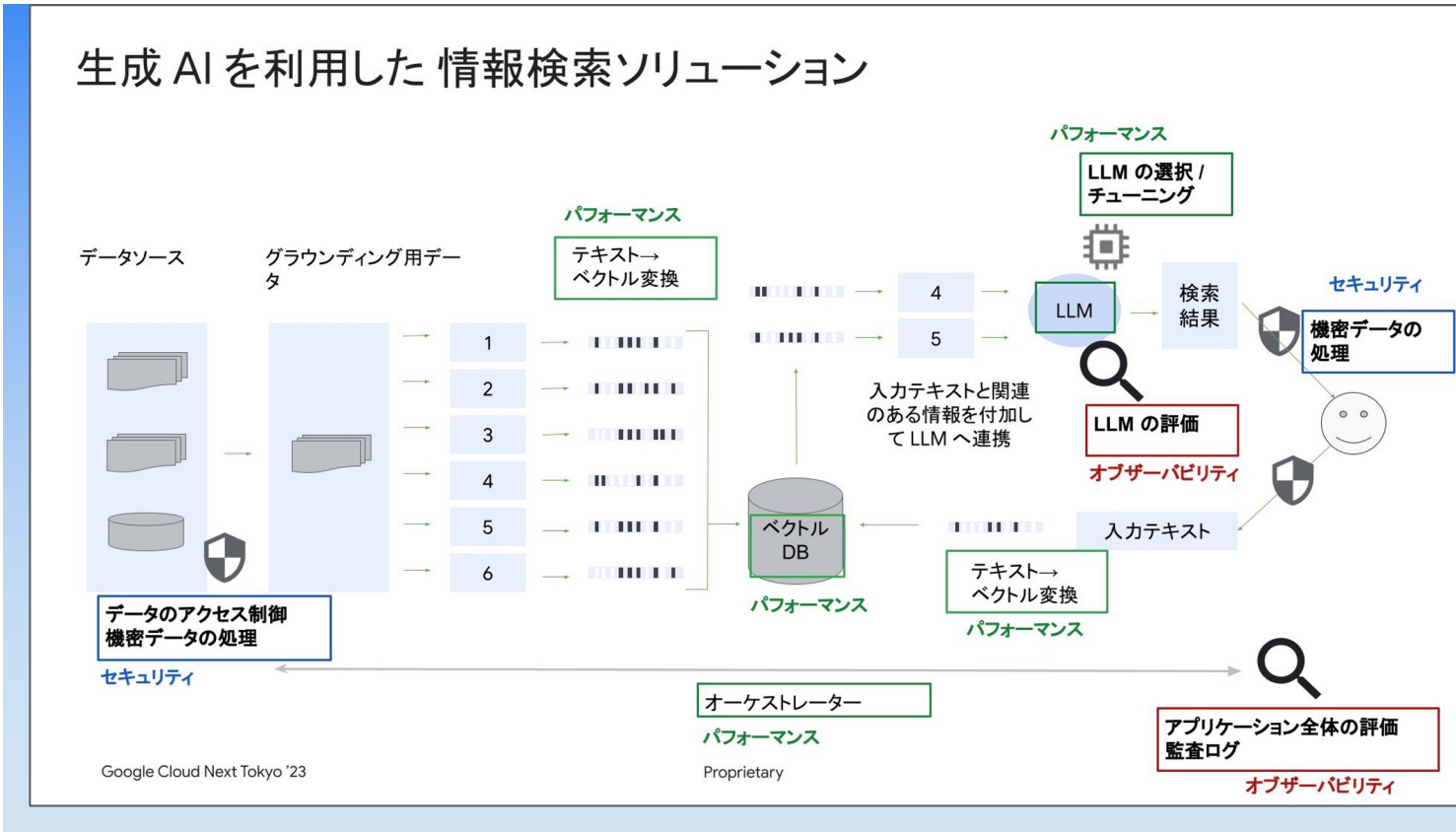
監視対象とする項目は、KPI やフィードバックをもとに決定します。代表例として、LLM のアウトプット品質、アプリケーション全体の品質、監査情報の取得が挙げられます。

(8) 同じドメイン内の追加ユースケースに取り組む

最初のユースケースを踏まえ、次のユースケースを検討します。ユースケースの横展開のほか、最初に得られたデータを活用して LLM をファイン チューニングして強化するなどの選択肢が考えられるでしょう。また、全社展開や外部公開に向けた CCoE（Cloud Center of Excellence）の立ち上げ、ガイドラインの整備も必要になるかもしれません。

システム構成において考慮すべき 3 つのポイント

システム構成を検討する時は、パフォーマンス、セキュリティ、オブザーバビリティが主に考慮すべきポイントとなります。今回は、Google Cloud における情報検索アプリケーションの構築を例に、3 つのポイントをどのように実装するかを解説します。



システム構成における 3 つのポイントと関連する要素

パフォーマンスに関連する要素

(1) ベクトル変換 API

ベクトル変換に特化した基盤モデル「Embeddings API」は、テキストだけでなくマルチモーダルな対応が可能です。後述する「LangChain」など OSS プロジェクトと統合し、簡単にアプリケーションに組み込めます。また、マルチリンガル モデルもリリースされているので、さまざまな環境で利用可能です。

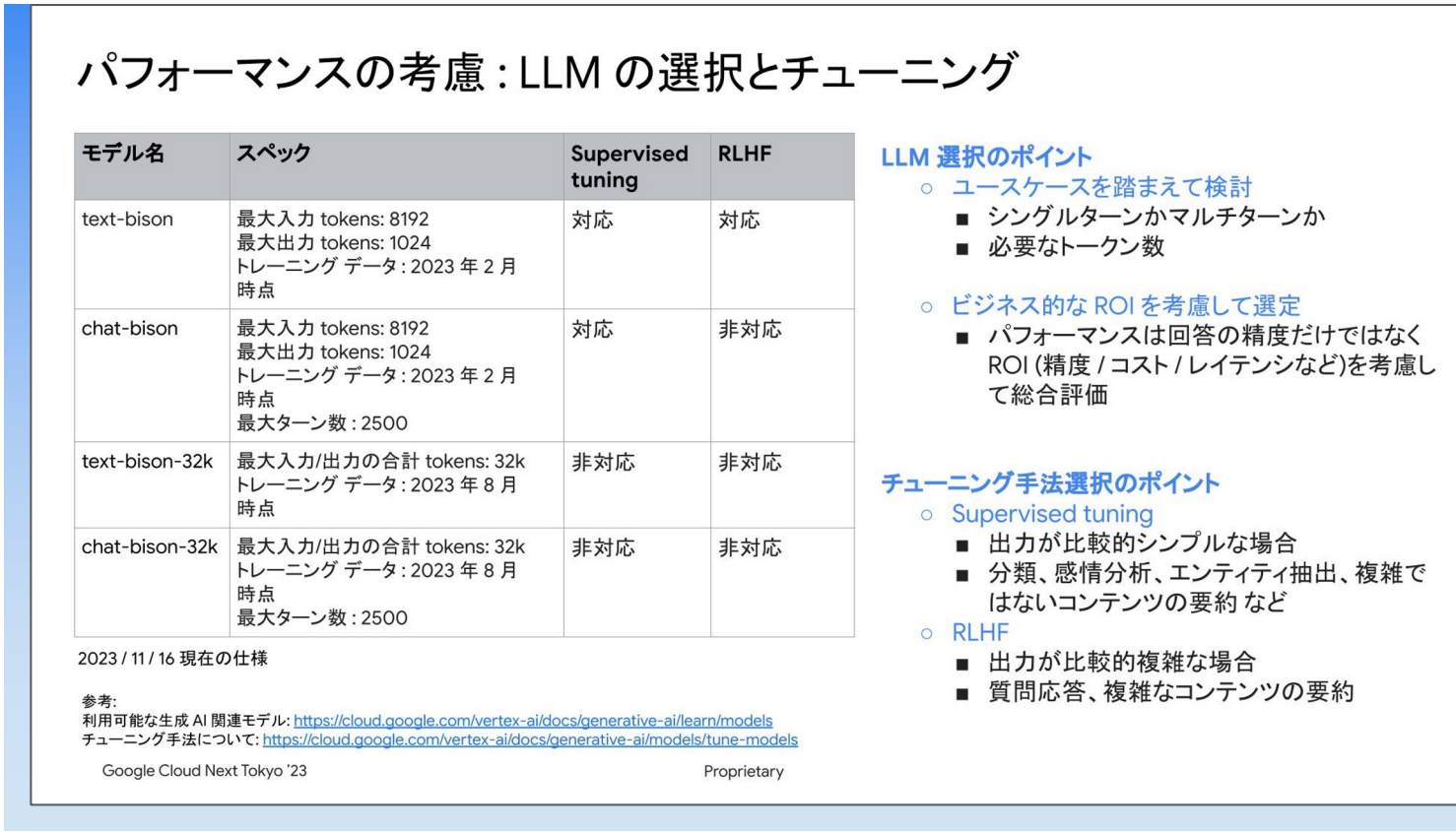
(2) ベクトル DB

Google Cloud のベクトル DB は、既存の DB 機能の一環としてベクトルデータを保存する「Cloud SQL for PostgreSQL」、「AlloyDB for PostgreSQL」、「BigQuery」と、ベクトルデータの保存に特化したデータベースとして利用する「Vertex AI Vector Search」に大別されます。従来のデータ基盤と併用するなら前者を、より高性能な要求があるなら後者を選択しましょう。

(3) LLM の選択とチューニング

LLM を選択する時のポイントは、ユースケースと ROI を考慮することです。シングルターンとマルチターンのどちらにするか、また、必要な入出力トークン数などをベースに検討を進めるのがおすすめです。

Google Cloud のチューニング手法は下図にまとめています。



Google Cloud が提供するチューニング手法

ただ、個人的にはグラウンディングが指定したデータに基づいた回答を行わせるために、チューニングは LLM の出力フォーマットを整えるために利用するのが有効と考えます。

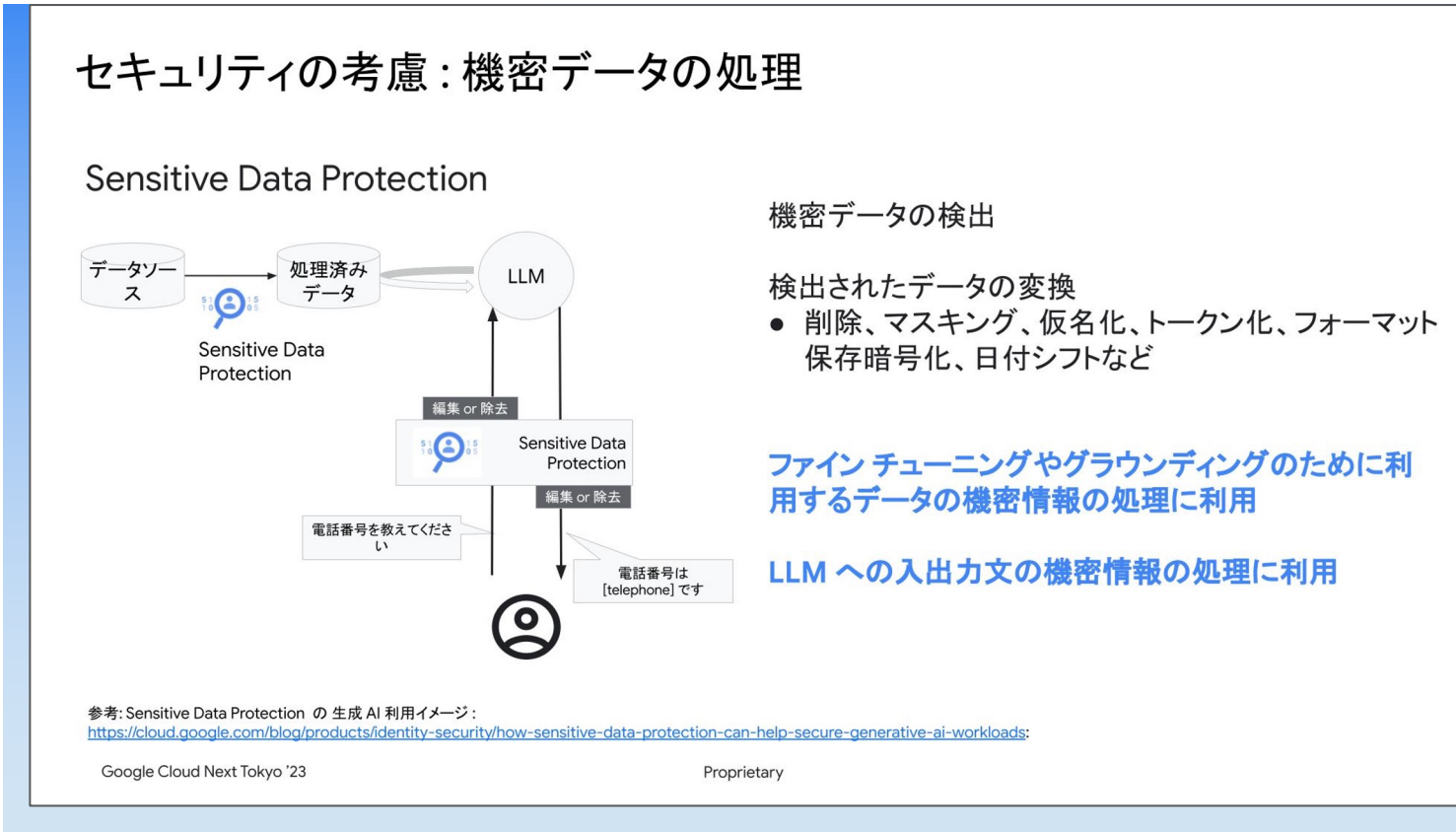
(4) LMM をアプリケーション実装に落とし込むフレームワーク（LangChain）

LangChain では、Retrieval や Agents など、LLM を用いたデータ処理を実装する際に必要となる典型的な機能を個別モジュールとして提供します。PaLM 2 や Vertex AI Vector Search など、Google Cloud との連携用モジュールも利用可能です。

セキュリティに関連する要素

(1) 機密データの処理

一般的に機密データの処理では、機密データの場所を整理し、削除する流れとなりますが、ここで「Sensitive Data Protection」を活用できます。ファイン チューニングやグラウンディング目的で利用する機密データを処理する、あるいは Sensitive Data Protection の API を使って LLM への入出力文書で機密データを処理する方法が考えられます。



Sensitive Data Protection による機密データ処理

また、PaLM を用いてプロンプト内で機密データを処理したり、ベクトル DB を使って対処する方法も選択できます。ベクトル DB を使う場合、事前に危険性のあるアタック プロンプトをベクトル化し、ベクトル DB に保存します。その上で、入出力プロンプトとアタック プロンプトの類似度をベクトルデータとして計算し、一定の類似度が認められる場合に回答をフィルターします。

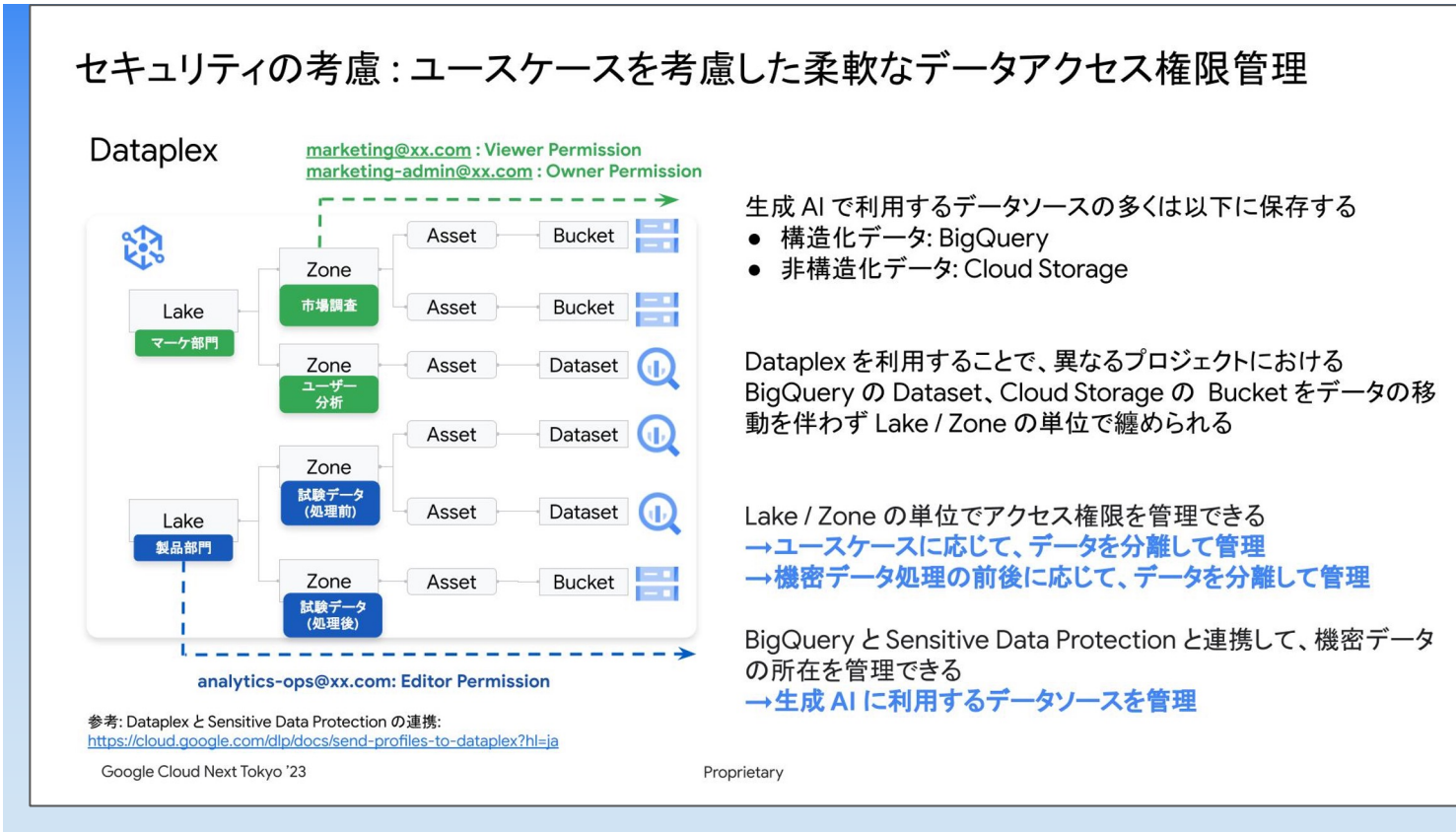
このような構成をとることで、機密情報の処理だけでなく、「プロンプト インジェクション」というプロンプトを用いた攻撃手法に対応できる可能性があります。

(2) データのアクセス制御

生成 AI 用途で BigQuery や「Cloud Storage」に保存したデータへのアクセス管理を効率良く行い機密情報の所在を確認するサービスとして「Dataplex」を活用できます。

Dataplex は格納したデータの移動を伴わず、論理的にまとめてアクセス制御することが可能です。ユースケースごとにデータをまとめる、あるいは、機密データの処理前・処理後でデータを分け、処理後のデータに限定してアプリケーションからのアクセス権限を付与するといった使い方ができます。

さらに、Dataplex と Sensitive Data Protection は連携可能です。連携させると BigQuery 上の機密データの所在を検出できるため、生成 AI に利用するデータの管理が容易になるでしょう。



Dataplex によるデータ制御

オブザーバビリティに関連する要素

(1) LLM/アプリケーション全体の評価、監査ログの取得

一般に LLM は、従来の機械学習モデルとは異なり評価が難しく、人による主観的な評価に限らず、客観的な指標による評価を採用することも重要です。「Vertex AI Model Evaluation」はこれを支援します。

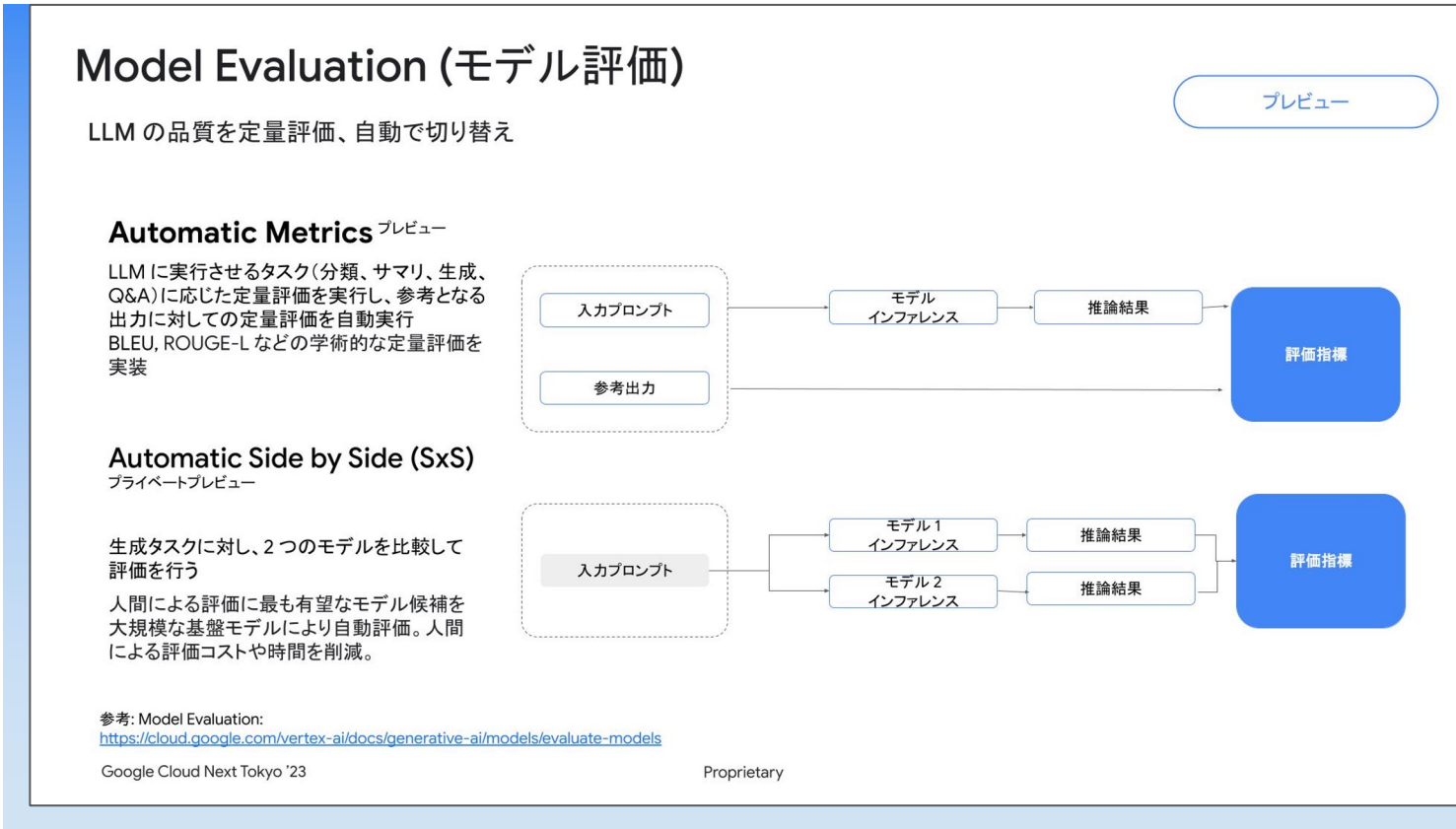
アプリケーション全体の評価を実現するには、アプリケーション ログを取得し目的に応じた形で保存・可視化する必要があります。BigQuery や Looker Studio などが役立ちます。

監査ログは、利用者情報などを保存する必要がありますが、「Cloud Audit Logs」を用いれば取得できます。

補足：2 つの評価メトリクスを提供する Vertex AI Model Evaluation

「Vertex AI Model Evaluation」では、LLM に対する 2 つの評価メトリクスを提供しています。1 つ目は Automatic Metrics で、モデルの出力と、事前に用意した参考出力の類似度を BLEU や ROUGE-L など、学術的な指標を使って評価します。

2 つ目は Automatic Side by Side（SxS）で、2 つのモデルの出力を比較し、評価を得ることが可能です。一定のパフォーマンスが得られるモデルと、検討中のモデルを比較することで、検討中モデルの妥当性を確認できます。



Vertex AI Model Evaluation の LLM に対する評価メトリクス

マネージド サービスの利用も 1 つの選択肢

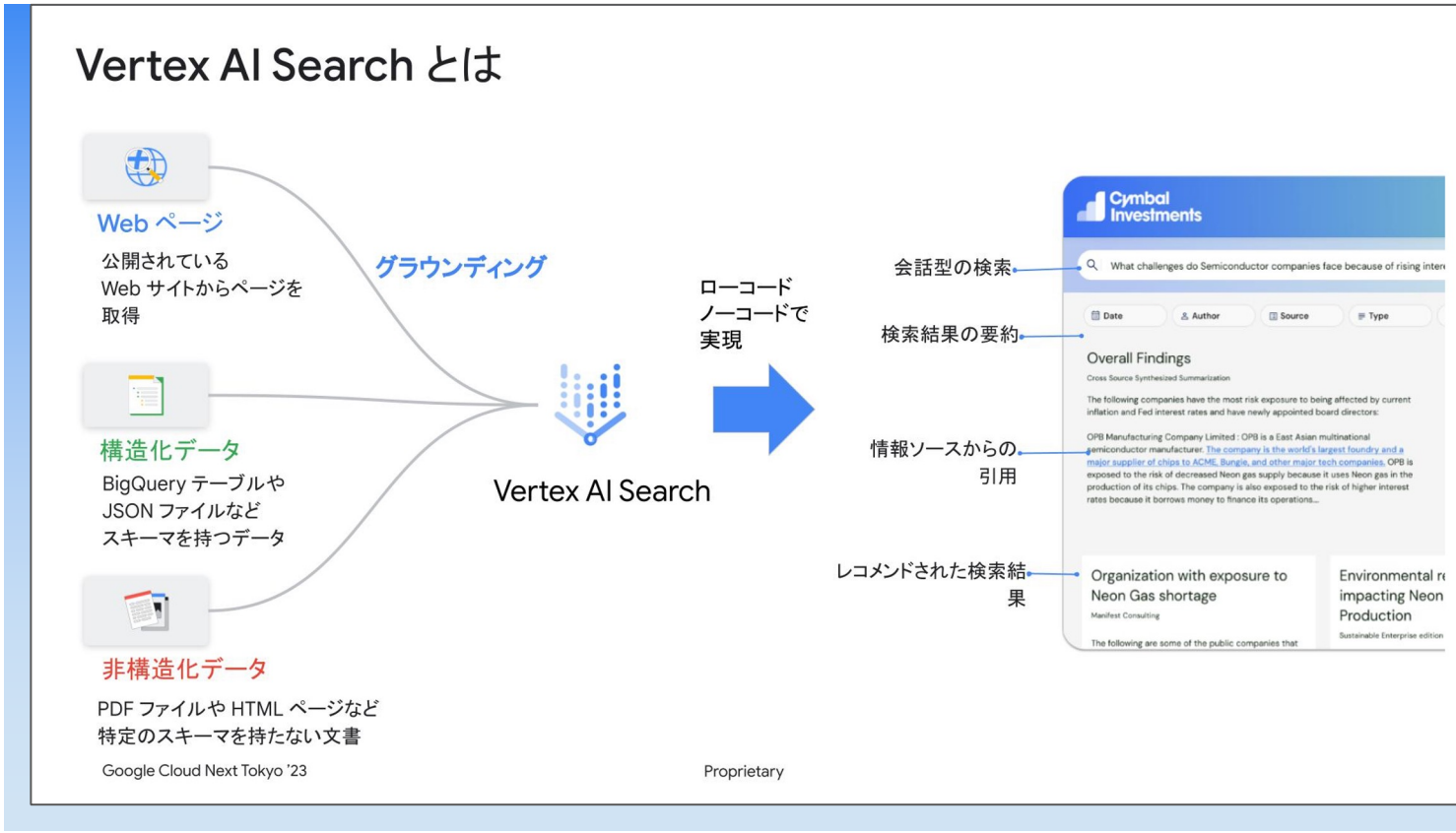
情報検索ソリューションをスクラッチ的に開発する方法を解説しましたが、Google Cloud では「Vertex AI Search」というマネージド サービスを選択することも可能です。

Vertex AI Search とは？

Vertex AI Search は、ユーザーがデータを指定すると、グラウンディングのためのアーキテクチャをマネージドで構築し、簡単に自社データに基づいた生成 AI の検索ソリューションを実現できる機能です。

グラウンディングで利用できるデータソースは、Web ページ、構造化データ、非構造化データがサポートされており、コンソール上で指定するだけで、ローコード・ノーコードで開発できます。

また、エンタープライズ レベルの使用用途に耐える機能を有します。エンタープライズ グレードのアクセス制御を管理し、カスタム エンベディングで自社の文脈での検索体験を実現可能です。



構造化、非構造化データなどから情報を取得

Vertex AI Search は、社内ポータルなどのサイトからバックエンド機能として呼び出して利用できます。フロントエンドのアプリケーションから API を呼び出すためのコードが生成されるため、簡単に実装可能です。

Vertex AI Search の典型的なユースケースは、リサーチ業務の効率化、プロダクト カタログの検索性向上、Web サイトのナビゲーション、社内文書検索の 4 つです。いわゆる生成 AI を利用した検索ソリューションという位置付けですが、自然言語でより柔軟な問い合わせを可能とし、多くのユースケースに対応できます。

Vertex AI Search はさまざまなユースケースに柔軟に対応

今回は、エンタープライズにおける生成 AI ソリューション実現のために、検討ステップや考慮すべきポイントを解説しました。Google Cloud では、実装に必要となるサービスや機能を多岐にわたり提供しています。生成 AI ソリューションの利用を検討される際には、ぜひお役立てください。

参照リンク

1. Google Cloud のジェネレーティブ AI の概要
2. 生成 AI のもたらすインパクトとエンタープライズ利用における検討ポイント アーカイブ動画視聴ページ

製品、サービスに関するお問い合わせ

goo.gl/CCZL78

Google Cloud の詳細については、上記 URL もしくは QR コードからアクセスしていただくか、同ページ「お問い合わせ」よりお問い合わせください。

Google Cloud