Informatica

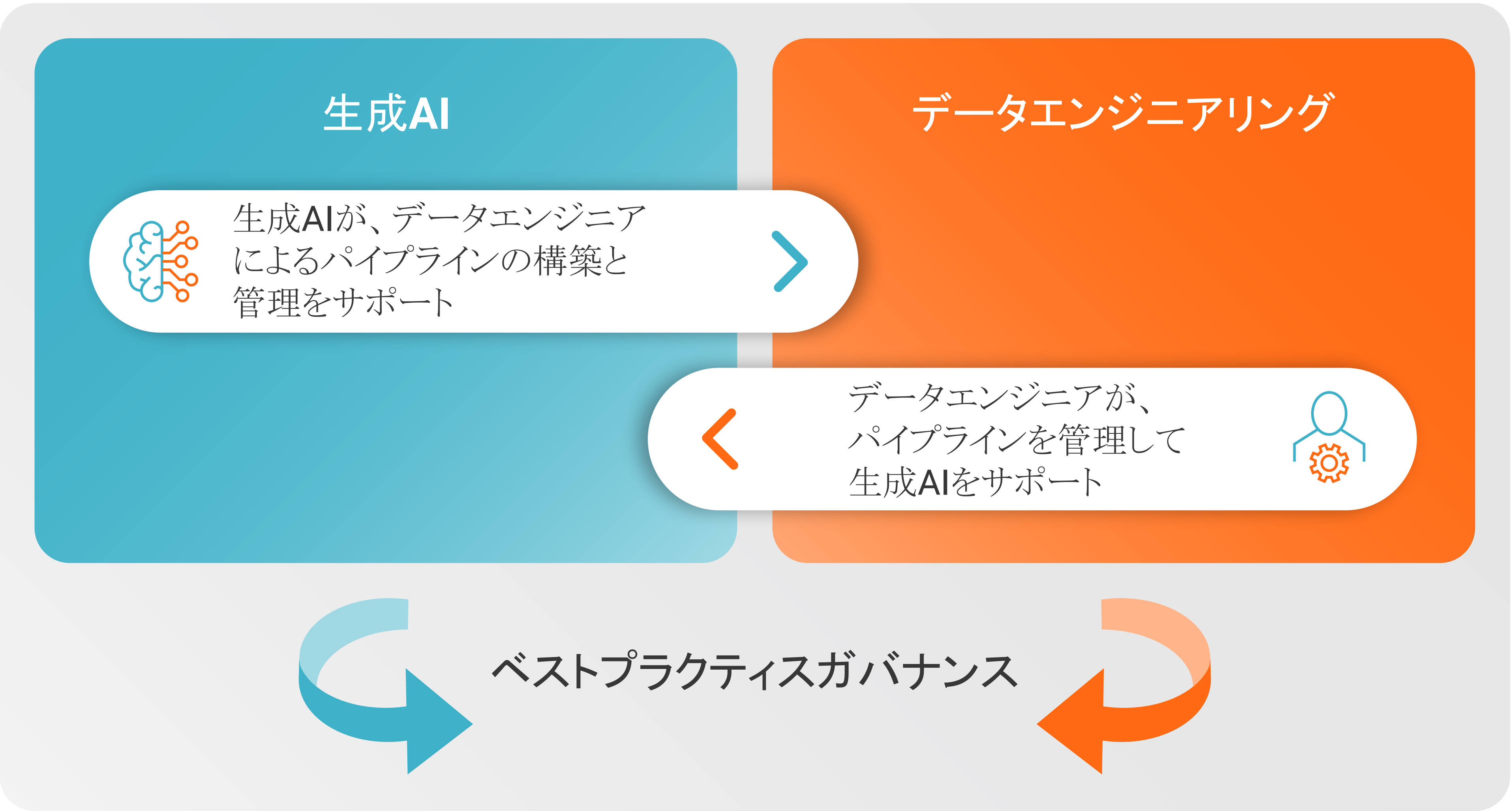
インフォグラフィック

生成AIとデータエンジニアリングの相乗効果～問題解決のためのソリューション

データエンジニアリングには生成AIが不可欠であり、生成AIにはデータエンジニアリングが不可欠です。なぜでしょうか。生成AIはデータエンジニアリングの生産性を高め、データエンジニアリングは生成AIのイノベーションを促進するからです。

生成AIとデータエンジニアリングの相乗効果を高めるための秘訣をご覧ください。

生成AIとデータエンジニアリングの融合



生成AIがデータエンジニアリングに役立つ理由

課題

データエンジニアはアナリティクス用のデータを提供するためのパイプラインの設計、テスト、実装、監視、最適化を行います。SaaSアプリケーションやモバイルアプリ、IoTセンサー、データプラットフォーム、分析ツール、業務担当者などの急増により、データインジェスチョン／変換タスクの管理が複雑化しており、各種コンポーネントをシームレスに統合することが困難になっています。

ソリューション

言語モデル（LM）プラットフォームと高度な機能を備えたパイプラインツールにより、この課題に対応できます。このようなパイプラインツールは、データエンジニアが入力した自然言語プロンプトに基づいて、データパイプラインのための初期コードを生成して、コードのデバッグ方法を提案し、カタログ化のためにパイプラインと関連データセットを文書化します。また、LMはデータ品質チェックのためのルールを提案したり、パイプラインを設計するための各種アーキテクチャアプローチを評価したりします。LMを通じてデータエンジニアリングに関連する複雑で面倒なタスクを自動化、高速化、簡素化することで、時間を大幅に短縮できます。

データエンジニアリングが生成AIに役立つ理由

課題

各企業は、LM（またはLMに接続するアプリケーションプログラミングインターフェイス（API））、対話型ユーザーインターフェイス（UI）、LMの出力結果に基づいてタスクを実行する追加機能を搭載した独自の生成AIアプリケーションを構築しています。生成AIアプリケーションから有用な出力結果を得るためには、有用な入力に加え、非構造化データオブジェクトを数値ベクトルに変換して、ユーザーのプロンプトをエンリッチ化し、LMの微調整を支援するパイプラインが必要です。

ソリューション

データエンジニアは、ETL（抽出、変換、ロード）またはELT（抽出、ロード、変換）の各段階で構成されるパイプラインを構築することで、この課題に対応できます。抽出、ロード、変換、再ロードの一連の作業により、自社の分野に特化したデータを生成AIアプリケーションで利用できるようになります。

抽出とロード
パイプラインが、関連するテキストをアプリケーションやファイルから抽出し、プラットフォーム（Databricks Lakehouse、Snowflake Data Cloudなど）のランディングゾーンにロードします。生成AIの精度を高めるためには、このテキストがマスターデータと一致していること、そして品質基準を満たしていることが重要です。

変換
パイプラインが、データを変換してLMで利用できる状態にし、テキストを数値トークンに変換して、それらのトークンを「チャンク」としてグループ化し、各チャンクの意味と相互関係を表すベクトルを作成します。

ロード
パイプラインが、埋め込み内容をベクトルデータベース（Pinecone、Weaviateなど）またはベクトル対応プラットフォーム（Databricks、MongoDBなど）にロードします。

データチームは、ベクトル化データを利用して、次の2つの方法を通じて生成AIアプリケーションをサポートできます。

1. 検索拡張生成（RAG）を実装する。RAGにより、ベクトルデータベース内で関連コンテンツを検索し、これをユーザープロンプトに追加することで、LMからより質の高い結果を得られるようになります。
2. LMを微調整して、パラメーターをベクトル化テキストに合わせて調整する。

生成AIアプリケーションの応答精度を強化するためには、RAGとLMの微調整が不可欠です。これにより、ビジネスコンテキストをより正確に反映させて、データハルシネーションのリスクを軽減させることができます。

生成AIとデータエンジニアリングを融合させて、データ戦略を再定義。

詳細はこちら

出典：『Achieving Fusion: How GenAI and Data Engineering Help One Another』、2024年

Where data & AI come to LIFE™

Informatica

IN20-4792-0524