aws

で拡張生成オプションとアーキテクチャを取得する AWS

AWS 規範ガイダンス

AWS 規範ガイダンス: で拡張生成オプションとアーキテクチャを取得する AWS

Table of Contents

で拡張生成オプションとアーキテクチャを取得する AWS

Mithil Shah、Rajeev Muralidhar、および Natacha Fort、Amazon Web Services

2024 年 10 月 (ドキュメント履歴)

生成 AI とは、シンプルなテキストプロンプトから画像、動画、テキスト、オーディオなどの新しいコンテンツやアーティファクトを作成できる AI モデルのサブセットを指します。生成 AI モデルは、幅広いサブジェクトとタスクを含む膨大な量のデータでトレーニングされます。これにより、明示的にトレーニングされていないタスクであっても、さまざまなタスクの実行における汎用性を実証することができます。単一のモデルが複数のタスクを実行できるため、これらのモデルは多くの場合、基盤モデル (FMsと呼ばれます。

生成 AI モデルの注目すべきアプリケーションの 1 つは、質問に答える能力です。ただし、これらのモデルを使用してカスタムドキュメントに基づいて質問に回答するときに発生する特定の課題があります。カスタムドキュメントには、専有情報、内部ウェブサイト、内部ドキュメント、Confluence ページ、SharePointページなどが含まれます。1 つのオプションは、取得拡張生成 (RAG) を使用することです。RAG では、基盤モデルは、レスポンスを生成する前に、トレーニングデータソース (カスタムドキュメントなど) の外部にある信頼できるデータソースを参照します。

このガイドでは、検索拡張生成 (RAG) システムなど、カスタムドキュメントからの質問に回答するために使用できる個別の生成 AI オプションについて説明します。また、Amazon Web Services () で RAG システムを構築する方法の概要についても説明しますAWS。RAG オプションとアーキテクチャを確認することで、 AWS のフルマネージドサービスとカスタム RAG アーキテクチャを選択できます。

対象者

このガイドの対象となるのは、RAG ソリューションを構築し、利用可能なアーキテクチャを確認し、各オプションの利点と欠点を理解したい生成 AI アーキテクトとマネージャーです。

目的

このガイドは以下を行う際に役立ちます。

- カスタムドキュメントからの質問への回答に使用できる生成 AI オプションを理解する
- で RAG システムのアーキテクチャオプションを確認する AWS

- 各 RAG オプションの利点と欠点を理解する
- AWS 環境の RAG アーキテクチャを選択する

カスタムドキュメントをクエリするための生成 AI オプション

多くの場合、組織には構造化データと非構造化データのさまざまなソースがあります。このガイドでは、生成 AI を使用して非構造化データから質問に回答する方法に焦点を当てます。

組織内の非構造化データは、さまざまなソースから取得できます。PDF PDFs 、テキストファイル、内部 Wiki、技術文書、公開ウェブサイト、ナレッジベースなどです。非構造化データに関する質問に回答できる基盤モデルが必要な場合は、次のオプションを使用できます。

- カスタムドキュメントやその他のトレーニングデータを使用して新しい基盤モデルをトレーニングする
- カスタムドキュメントのデータを使用して既存の基盤モデルを微調整する
- コンテキスト内学習を使用して、質問をするときに基盤モデルにドキュメントを渡す
- 取得拡張生成 (RAG) アプローチを使用する

カスタムデータを含む新しい基盤モデルをゼロからトレーニングすることは、野心的な取り組みです。BloombergGPT モデルなど、いくつかの企業が成功Bloombergしています。もう 1 つの例は、によるマルチモーダルEXAONEモデルです。このモデルはLG AI Research、6,000 億個のアートワークと 2 億 5,000 万個の高解像度イメージをテキストとともに使用してトレーニングされました。AI のコスト: 基盤モデルを構築または購入すべき (LinkedIn) によると、トレーニングにかかるMetaLlama 2コストは約 48 0 万 USD です。ゼロからモデルをトレーニングするための主な前提条件は 2 つあります。リソースへのアクセス (財務、技術、時間) と明確な投資収益率です。これが適していないと思われる場合、次のオプションは既存の基盤モデルを微調整することです。

既存のモデルを微調整するには、Amazon Titan、Mistral、Llama モデルなどのモデルを取得し、そのモデルをカスタムデータに適応させる必要があります。ファインチューニングにはさまざまな手法があり、そのほとんどはモデル内のすべてのパラメータを変更するのではなく、少数のパラメータのみを変更するものです。これは、パラメータ効率の高い微調整と呼ばれます。ファインチューニングには主に 2 つの方法があります。

- 教師ありファインチューニングでは、ラベル付きデータが使用され、新しい種類のタスク用にモデルをトレーニングするのに役立ちます。たとえば、PDF フォームに基づいてレポートを生成する場合は、十分な例を提供することで、その方法をモデルに教える必要があります。
- 教師なしファインチューニングはタスクに依存しず、基盤モデルを独自のデータに適応させます。ドキュメントのコンテキストを理解するようにモデルをトレーニングします。次に、ファイン

チューニングされたモデルは、よりカスタムなスタイルを使用してレポートなどのコンテンツを作成します。

ただし、ファインチューニングは、質疑応答のユースケースには適していない場合があります。詳細については、このガイドの「RAG とファインチューニングの比較」を参照してください。

質問すると、基盤モデルをドキュメントに渡し、モデルのコンテキスト内学習を使用してドキュメントから回答を返すことができます。このオプションは、1 つのドキュメントのアドホッククエリに適しています。ただし、このソリューションは、複数のドキュメントのクエリや、Microsoft SharePoint や Atlassian Confluence などのシステムやアプリケーションのクエリには適していません。

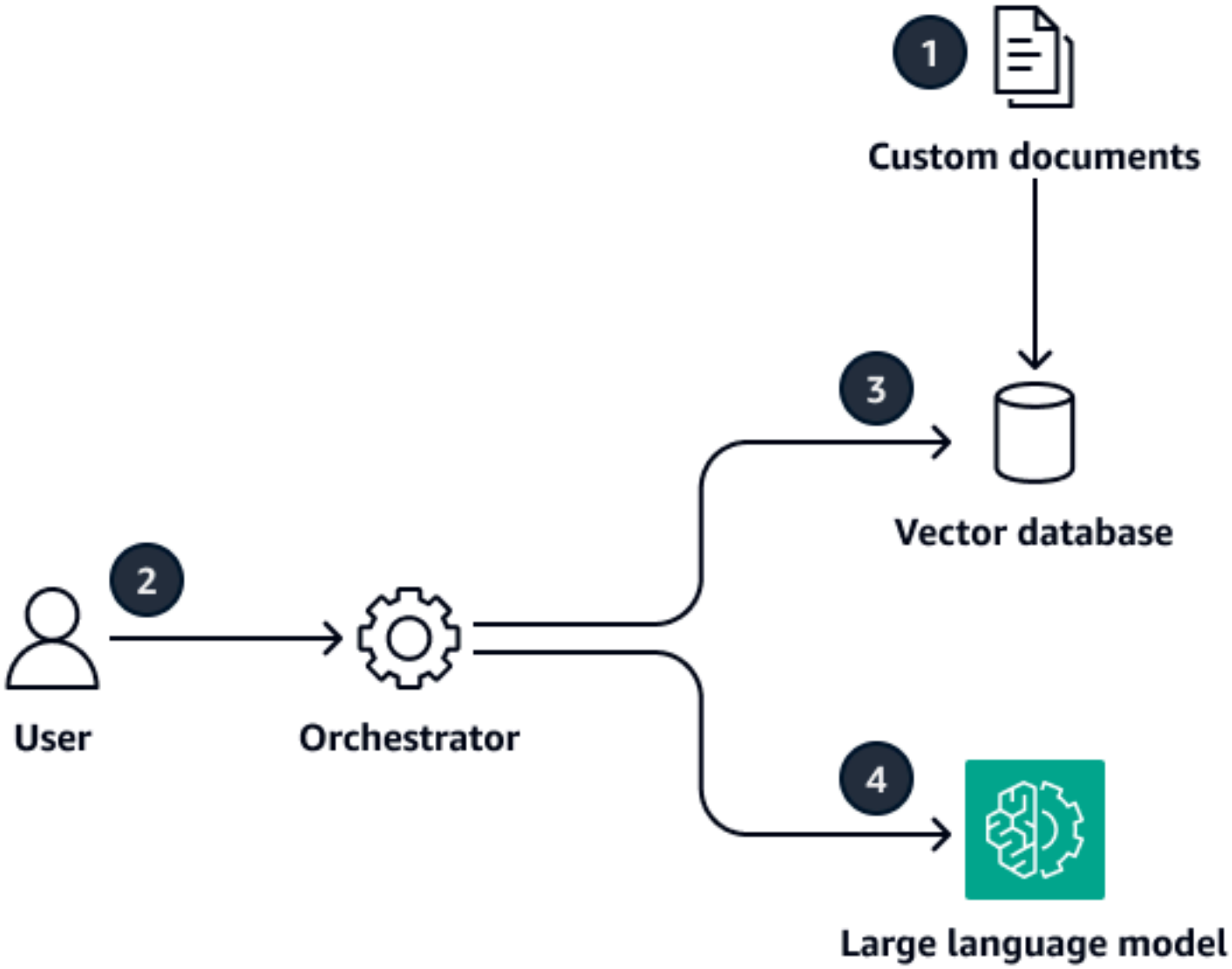
最後のオプションは、RAG を使用することです。RAG では、基盤モデルはレスポンスを生成する前にカスタムドキュメントを参照します。RAG は、モデルの機能を組織の内部ナレッジベースに拡張します。モデルを再トレーニングする必要はありません。これは、モデル出力を改善して、さまざまなコンテキストで関連性、正確性、有用性を維持するための費用対効果の高いアプローチです。

このセクションのトピック:

- 取得拡張生成について
- 取得拡張生成と微調整の比較
- 取得拡張生成のユースケース

取得拡張生成について

Retrieval Augmented Generation (RAG) は、会社の内部ドキュメントなどの外部データを使用して大規模言語モデル (LLM) を補強するために使用される手法です。これにより、特定のユースケースに対して正確で有用な出力を生成するために必要なコンテキストをモデルに提供します。RAG は、企業で LLMs を使用するための実用的で効果的なアプローチです。次の図は、RAG アプローチの仕組みの概要を示しています。



大まかに言うと、RAG プロセスは 4 つのステップです。最初のステップは 1 回実行され、残りの 3 つのステップは必要な回数だけ実行されます。

1. 埋め込みを作成して、内部ドキュメントをベクトルデータベースに取り込みます。埋め込みは、データの意味的またはコンテキスト的な意味をキャプチャするドキュメント内のテキストの数値表現です。ベクトルデータベースは基本的にこれらの埋め込みのデータベースであり、ベクトルストアまたはベクトルインデックスと呼ばれることもあります。このステップでは、データのクリーニング、フォーマット、チャンキングが必要ですが、これは 1 回限りの前払いアクティビティです。
2. 人間は自然言語でクエリを送信します。

3. オーケストレーターはベクトルデータベースで類似度検索を実行し、関連するデータを取得します。オーケストレーターは、取得したデータ (コンテキストとも呼ばれます) をクエリを含むプロンプトに追加します。
4. オーケストレーターはクエリとコンテキストを LLM に送信します。LLM は、追加のコンテキストを使用してクエリへのレスポンスを生成します。

ユーザーの観点から見ると、RAG は任意の LLM とやり取りしているように見えます。ただし、システムは問題のコンテンツについてより多くの知識を持ち、組織のナレッジベースに微調整された回答を提供します。

RAG アプローチの仕組みの詳細については、 ウェブサイトの「RAG とは AWS 」を参照してください。

本番稼働レベルの RAG システムのコンポーネント

本番稼働レベルの RAG システムを構築するには、RAG ワークフローのさまざまな側面について考える必要があります。概念的には、本番稼働レベルの RAG ワークフローには、特定の実装に関係なく、次の機能とコンポーネントが必要です。

- コネクタ — さまざまなエンタープライズデータソースをベクトルデータベースに接続します。構造化データソースの例としては、トランザクションデータベースや分析データベースなどがあります。非構造化データソースの例としては、オブジェクトストア、コードベース、Software as a Service (SaaS) プラットフォームなどがあります。データソースごとに異なる接続パターン、ライセンス、設定が必要になる場合があります。
- データ処理 — データは、PDFs、スキャンされた画像、ドキュメント、プレゼンテーション、Microsoft SharePointファイルなど、さまざまな形状と形式で提供されます。インデックス作成のためにデータを抽出、処理、準備するには、データ処理手法を使用する必要があります。
- 埋め込み — 関連性検索を実行するには、ドキュメントとユーザークエリを互換性のある形式に変換する必要があります。言語モデルを埋め込むことで、ドキュメントを数値表現に変換します。これらは基本的に基盤となる基盤モデルの入力です。
- ベクトルデータベース — ベクトルデータベースは、埋め込み、関連するテキスト、メタデータのインデックスです。インデックスは検索と取得に最適化されています。
- リトリーバー — ユーザークエリの場合、リトリーバーはベクトルデータベースから関連するコンテキストを取得し、ビジネス要件に基づいてレスポンスをランク付けします。
- 基盤モデル — RAG システムの基盤モデルは、通常 LLM です。コンテキストとプロンプトを処理することで、基盤モデルはユーザーのレスポンスを生成してフォーマットします。

- ガードレール — ガードレールは、クエリ、プロンプト、取得されたコンテキスト、LLM レスポンスが正確で、責任があり、倫理的で、幻覚やバイアスがないことを確認するように設計されています。
- オーケストレーター — オーケストレーターはend-to-endのワークフローのスケジュールと管理を担当します。
- ユーザーエクスペリエンス — 通常、ユーザーはチャット履歴の表示やレスポンスに関するユーザーのフィードバックの収集など、豊富な機能を備えた会話型チャットインターフェイスを操作します。
- ID とユーザー管理 — アプリケーションへのユーザーアクセスをきめ細かく制御することが重要です。では AWS クラウド、通常、ポリシー、ロール、アクセス許可は AWS Identity and Access Management (IAM) によって管理されます。

明らかに、RAG システムの計画、開発、リリース、管理には多大な労力がかかります。Amazon Bedrock や Amazon Q Business などのフルマネージドサービスは、未分化の重労働の一部を管理するのに役立ちます。ただし、カスタム RAG アーキテクチャでは、リトリーバーやベクトルデータベースなどのコンポーネントをより詳細に制御できます。

取得拡張生成と微調整の比較

次の表に、ファインチューニングと RAG ベースのアプローチの利点と欠点を示します。

アプローチ	利点	欠点
ファインチューニング	• 微調整されたモデルが教師なしアプローチを使用してトレーニングされている場合、組織のスタイルにより近いコンテンツを作成できます。 • 独自のデータまたは規制データに基づいてトレーニングされた微調整されたモデルは、組織が社内または業界固有のデータとコンプライアンス標準に従うのに役立ちます。	• モデルのサイズによっては、微調整に数時間から数日かかる場合があります。したがって、カスタムドキュメントが頻繁に変更される場合、これは適切なソリューションではありません。 • ファインチューニングには、低ランク適応 (LoRA) やパラメータ効率の高いファインチューニング (PEFT) などの手法を理解す

アプローチ	利点	欠点
		る必要があります。微調整にはデータサイエンティストが必要になる場合があります。 • ファインチューニングは、すべてのモデルで利用できるとは限りません。 • 微調整されたモデルは、レスポンスでソースへの参照を提供しません。 • 微調整されたモデルを使用して質問に回答すると、幻覚のリスクが高くなる可能性があります。

アプローチ	利点	欠点
RAG	• RAG を使用すると、ファインチューニングなしでカスタムドキュメントの質問応答システムを構築できます。 • RAG は数分で最新のドキュメントを組み込むことができます。 • AWS は、フルマネージド RAG ソリューションを提供します。したがって、データサイエンティストや機械学習に関する専門知識は必要ありません。 • レスポンスでは、RAG モデルが情報ソースへの参照を提供します。 • RAG はベクトル検索のコンテキストを生成された回答の基礎として使用するため、幻覚のリスクが軽減されます。	• RAG は、ドキュメント全体からの情報を要約する際にうまく機能しません。

カスタムドキュメントを参照する質疑応答ソリューションを構築する必要がある場合は、RAG ベースのアプローチから始めることをお勧めします。要約などの追加のタスクを実行するためにモデルが必要な場合は、ファインチューニングを使用します。

ファインチューニングアプローチと RAG アプローチを 1 つのモデルにまとめることができます。この場合、RAG アーキテクチャは変更されませんが、回答を生成する LLM もカスタムドキュメントと微調整されます。これにより、両方の長所が組み合わされ、ユースケースに最適なソリューションになる可能性があります。教師ありファインチューニングと RAG を組み合わせる方法の詳細については、 の「RAFT: Adapting Language Model to Domain Specific RAG research」を参照してください University of California, Berkeley。

取得拡張生成のユースケース

RAG アプローチを使用する際の一般的なユースケースは次のとおりです。

- 検索エンジン – RAG 対応の検索エンジンは、検索結果でより正確でup-to-date注目スニペットを提供できます。
- 質疑応答システム – RAG は、質疑応答システムにおける応答の品質を向上させることができます。検索ベースのモデルは、類似度検索を使用して、回答を含む関連するパッセージまたはドキュメントを検索します。次に、その情報に基づいて簡潔で関連するレスポンスを生成します。
- 小売または e コマース – RAG は、より関連性の高いパーソナライズされた製品のレコメンデーションを提供することで、e コマースのユーザーエクスペリエンスを向上させることができます。ユーザーの好みと製品の詳細に関する情報を取得して取り込むことで、RAG は顧客にとってより正確で有用なレコメンデーションを生成できます。
- 産業または製造 – 製造では、RAG は工場の工場オペレーションなどの重要な情報にすばやくアクセスするのに役立ちます。また、意思決定プロセス、トラブルシューティング、組織のイノベーションにも役立ちます。厳格な規制フレームワーク内で事業を行うメーカーの場合、RAG は、業界標準や規制機関などの内部および外部のソースから、更新された規制とコンプライアンス標準を迅速に取得できます。
- ヘルスケア – RAG は、正確でタイムリーな情報へのアクセスが重要なヘルスケア業界で可能性を秘めています。関連する医療知識を外部ソースから取得して取り込むことで、RAG は医療アプリケーションでより正確でコンテキスト対応を提供できます。このようなアプリケーションは、モデルではなく最終的に呼び出しを行う人間の臨床医がアクセスできる情報を強化します。
- 法的 – RAG は、複雑な法的文書がクエリのコンテキストを提供する合併や買収などの法的シナリオに強力に適用できます。これにより、法律専門家は複雑な規制上の問題を迅速に解決できます。

での完全マネージド型取得拡張生成オプション AWS

で検索拡張生成 (RAG) ワークフローを管理するには AWS、カスタム RAG パイプラインを使用するか、 AWS が提供するフルマネージドサービス機能の一部を使用できます。これらには RAG ベースのシステムのコアコンポーネントが多数含まれているため、フルマネージドサービスは未分化の重労働の一部を管理するのに役立ちます。ただし、これらのサービスはカスタマイズの機会が少なくなります。

フルマネージド型 はコネクタ AWS のサービス を使用して、ウェブサイト、Atlassian Confluence、Microsoft SharePoint などの外部データソースからデータを取り込みます。サポートされているデータソースは、 によって異なります AWS のサービス。

このセクションでは、 で RAG ワークフローを構築するための以下のフルマネージドオプションについて説明します AWS。

- Amazon Bedrock ナレッジベース
- Amazon Q Business
- Amazon SageMaker AI Canvas

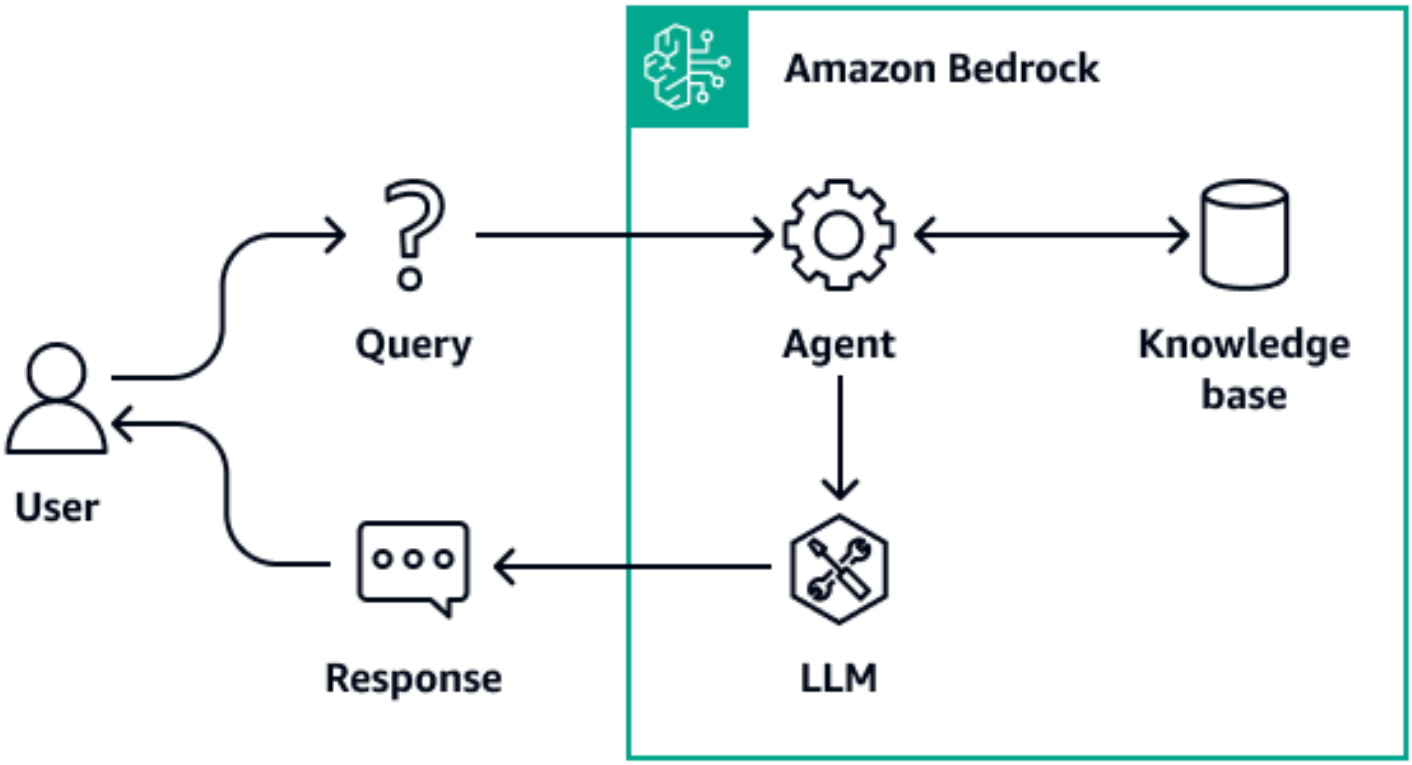
これらのオプションから選択する方法の詳細については、このガイド[で取得拡張生成オプションを選択する AWS]の「」を参照してください。

Amazon Bedrock ナレッジベース

Amazon Bedrock は、主要な AI スタートアップ企業や Amazon が提供する高パフォーマンスな基盤モデル (FM) を、統合 API を通じて利用できるようにするフルマネージド型サービスです。ナレッジベースは、取り込みから取得、プロンプト拡張まで、RAG ワークフロー全体を実装するのに役立つ Amazon Bedrock の機能です。データソースへのカスタム統合を構築したり、データフローを管理したりする必要はありません。セッションコンテキスト管理は、生成 AI アプリケーションがマルチターン会話を簡単にサポートできるように組み込まれています。

データの場所を指定すると、Amazon Bedrock のナレッジベースは内部でドキュメントを取得し、それらをテキストのブロックにチャンクし、テキストを埋め込みに変換して、選択したベクトルデータベースに埋め込みを保存します。Amazon Bedrock は埋め込みを管理および更新し、ベクトルデータベースをデータと同期させます。ナレッジベースの仕組みの詳細については、「Amazon Bedrock ナレッジベースの仕組み」を参照してください。

Amazon Bedrock エージェントにナレッジベースを追加すると、エージェントはユーザー入力に基づいて適切なナレッジベースを識別します。エージェントは関連情報を取得し、入力プロンプトに情報を追加します。更新されたプロンプトは、レスポンスを生成するためのより多くのコンテキスト情報をモデルに提供します。透明性を向上させ、幻覚を最小限に抑えるために、ナレッジベースから取得した情報はソースまで追跡可能です。



Amazon Bedrock は、RAG に対して次の 2 APIs をサポートしています。

- RetrieveAndGenerate – この API を使用してナレッジベースをクエリし、取得した情報からレスポンスを生成できます。Amazon Bedrock は内部的にクエリを埋め込みに変換し、ナレッジベースにクエリを実行し、検索結果をコンテキスト情報としてプロンプトを拡張して、LLM で生成されたレスポンスを返します。また、Amazon Bedrock は会話の短期記憶を管理し、よりコンテキストに応じた結果を提供します。
- 取得 – この API を使用して、ナレッジベースから直接取得した情報を使用してナレッジベースをクエリできます。この API から返された情報を使用して、取得したテキストの処理、関連性の評価、またはレスポンス生成のための別のワークフローの開発を行うことができます。Amazon Bedrock は内部的にクエリを埋め込みに変換し、ナレッジベースを検索して、関連する結果を返します。検索結果の上に追加のワークフローを構築できます。たとえば、

LangChainAmazonKnowledgeBasesRetrieverプラグインを使用して、RAG ワークフローを生成 AI アプリケーションに統合できます。

API を使用するためのアーキテクチャパターンの例とstep-by-stepの手順については、 APIs「ナレッジベースが Amazon Bedrock でフルマネージド RAG エクスペリエンスを提供するようになりました」(AWS ブログ記事) を参照してください。RetrieveAndGenerate API を使用してインテリジェントなチャットベースのアプリケーションの RAG ワークフローを構築する方法の詳細については、「Amazon Bedrock ナレッジベースを使用したコンテキストチャットボットアプリケーションの構築」(AWS ブログ記事) を参照してください。

ナレッジベースのデータソース

所有権を持つ独自のデータをナレッジベースに接続することができます。データソースコネクタを設定したら、データをナレッジベースと同期または最新の状態に保ち、データをクエリに使用できるようにします。Amazon Bedrock ナレッジベースは、次のデータソースへの接続をサポートします。

- Amazon Simple Storage Service (Amazon S3) – コンソールまたは API を使用してAmazon S3バケットを Amazon Bedrock ナレッジベースに接続できます。ナレッジベースは、バケット内のファイルを取り込み、インデックスを作成します。このタイプのデータソースは、次の機能をサポートしています。
  - ドキュメントメタデータフィールド – 別のファイルを含めて、Amazon S3 バケット内のファイルのメタデータを指定できます。その後、これらのメタデータフィールドを使用して、レスポンスの関連性をフィルタリングして改善できます。
  - 包含フィルターまたは除外フィルター – クロール時に特定のコンテンツを含めたり除外したりできます。
  - 増分同期 – コンテンツの変更は追跡され、前回の同期以降に変更されたコンテンツのみがクロールされます。
- Confluence – コンソールまたは API を使用して、Atlassian Confluenceインスタンスを Amazon Bedrock ナレッジベースに接続できます。このタイプのデータソースは、次の機能をサポートしています。
  - メインドキュメントフィールドの自動検出 – メタデータフィールドは自動的に検出され、クロールされます。これらのフィールドはフィルタリングに使用できます。
  - 包含コンテンツフィルターまたは除外コンテンツフィルター – スペース、ページタイトル、ブログタイトル、コメント、添付ファイル名、または拡張機能のプレフィックスまたは正規表現パターンを使用して、特定のコンテンツを包含または除外できます。

  - 増分同期 - コンテンツの変更は追跡され、前回の同期以降に変更されたコンテンツのみがクロールされます。
  - OAuth 2.0 認証、ConfluenceAPI トークンによる認証 – 認証情報は に保存されます AWS Secrets Manager。
- Microsoft SharePoint – コンソールまたは API を使用してSharePoint、インスタンスをナレッジベースに接続できます。このタイプのデータソースは、次の機能をサポートしています。
  - メインドキュメントフィールドの自動検出 – メタデータフィールドは自動的に検出され、クロールされます。これらのフィールドはフィルタリングに使用できます。
  - 包含コンテンツフィルターまたは除外コンテンツフィルター – メインページのタイトル、イベント名、ファイル名 (拡張子を含む) のプレフィックスまたは正規表現パターンを使用して、特定のコンテンツを包含または除外できます。
  - 増分同期 - コンテンツの変更は追跡され、前回の同期以降に変更されたコンテンツのみがクロールされます。
  - OAuth 2.0 認証 – 認証情報は に保存されます AWS Secrets Manager。
- Salesforce – コンソールまたは API を使用してSalesforce、インスタンスをナレッジベースに接続できます。このタイプのデータソースは、次の機能をサポートしています。
  - メインドキュメントフィールドの自動検出 – メタデータフィールドは自動的に検出され、クロールされます。これらのフィールドはフィルタリングに使用できます。
  - 包含または除外コンテンツフィルター – プレフィックスまたは正規表現パターンを使用して、特定のコンテンツを包含または除外できます。フィルターを適用できるコンテンツタイプのリストについては、Amazon Bedrock ドキュメントの「包含/除外フィルター」を参照してください。
  - 増分同期 – コンテンツの変更は追跡され、前回の同期以降に変更されたコンテンツのみがクロールされます。
  - OAuth 2.0 認証 – 認証情報は に保存されます AWS Secrets Manager。
- ウェブクローラー – Amazon Bedrock ウェブクローラーは、指定した URLs に接続してクロールします。次の機能がサポートされています。
  - クロールする複数の URL を選択する
  - `Allow` や などの標準の robots.txt ディレクティブを尊重する `Disallow`
  - パターンに一致する URLs を除外する
  - クローリングのレートを制限する
  - Amazon CloudWatch で、クロールされた各 URL のステータスを表示します。

Amazon Bedrock ナレッジベースに接続できるデータソースの詳細については、「ナレッジベースのデータソースコネクタを作成する」を参照してください。

ナレッジベースのベクトルデータベース

ナレッジベースとデータソース間の接続を設定するときは、ベクトルストアとも呼ばれるベクトルデータベースを設定する必要があります。ベクトルデータベースは、Amazon Bedrock がデータを表す埋め込みを保存、更新、管理する場所です。各データソースは、さまざまなタイプのベクトルデータベースをサポートしています。データソースで使用できるベクトルデータベースを確認するには、データソースタイプを参照してください。

Amazon Bedrock で Amazon OpenSearch Serverless にベクトルデータベースを自動的に作成する場合は、ナレッジベースを作成するときにこのオプションを選択できます。ただし、独自のベクトルデータベースを設定することもできます。独自のベクトルデータベースをセットアップする場合は、ナレッジベースの独自のベクトルストアの前提条件を参照してください。ベクトルデータベースのタイプごとに独自の前提条件があります。

データソースタイプに応じて、Amazon Bedrock ナレッジベースは次のベクトルデータベースをサポートします。

- Amazon OpenSearch Serverless
- Amazon Aurora PostgreSQL-Compatible Edition
- Pinecone (Pinecone ドキュメント)
- Redis Enterprise Cloud (Redis ドキュメント)
- MongoDB Atlas (MongoDB ドキュメント)

Amazon Q Business

Amazon Q Business は、質問への回答、概要の提供、コンテンツの生成、エンタープライズデータに基づくタスクの完了のために設定できる、フルマネージドの生成 AI を活用したアシスタントです。これにより、エンドユーザーは、引用を含むエンタープライズデータソースからアクセス許可対応のレスポンスをすぐに受け取ることができます。

主な特徴

Amazon Q Business の以下の機能は、本番稼働グレードの RAG ベースの生成 AI アプリケーションの構築に役立ちます。

- 組み込みコネクタ – Amazon Q Business は、、、Adobe Experience Manager (AEM)、Jiraおよびの コネクタなどSalesforce、40 種類以上のコネクタをサポートしていますMicrosoft SharePoint。完全なリストについては、「サポートされているコネクタ」を参照してください。サポートされていないコネクタが必要な場合は、Amazon AppFlow を使用してデータソースから Amazon Simple Storage Service (Amazon S3) にデータを取得し、Amazon Q Business を Amazon S3 バケットに接続できます。Amazon AppFlow がサポートするデータソースの完全なリストについては、「サポートされているアプリケーション」を参照してください。
- 組み込みインデックス作成パイプライン – Amazon Q Business は、ベクトルデータベース内のデータをインデックス作成するための組み込みパイプラインを提供します。 AWS Lambda 関数を使用して、インデックス作成パイプラインの前処理ロジックを追加できます。
- インデックスオプション – Amazon Q Business でネイティブインデックスを作成してプロビジョニングし、Amazon Q Business リトリーバーを使用してそのインデックスからデータを取得できます。または、事前設定された Amazon Kendra インデックスをリトリーバーとして使用できます。詳細については、「Amazon Q Business アプリケーションのリトリーバーの作成」を参照してください。
- 基盤モデル – Amazon Q Business は、Amazon Bedrock でサポートされている基盤モデルを使用します。完全なリストについては、「Amazon Bedrock でサポートされている基盤モデル」を参照してください。
- プラグイン – Amazon Q Business は、 でチケット情報とチケット作成をまとめる自動化された方法など、プラグインを使用してターゲットシステムと統合する機能を提供しますJira。設定が完了すると、プラグインはエンドユーザーの生産性を高めるのに役立つ読み取りアクションと書き込みアクションをサポートできます。Amazon Q Business は、組み込みプラグインとカスタムプラグインの 2 種類のプラグインをサポートしています。
- ガードレール – Amazon Q Business は、グローバルコントロールとトピックレベルのコントロールをサポートしています。例えば、これらのコントロールは、プロンプト内の個人を特定できる情報 (PII)、不正使用、または機密情報を検出できます。詳細については、「Amazon Q Business の管理者コントロールとガードレール」を参照してください。
- ID 管理 – Amazon Q Business を使用すると、ユーザーとその RAG ベースの生成 AI アプリケーションへのアクセスを管理できます。詳細については、「Amazon Q Business の Identity and Access Management」を参照してください。また、Amazon Q Business コネクタは、ドキュメント自体とともにドキュメントにアタッチされているアクセスコントロールリスト (ACL) 情報のインデックスを作成します。次に、Amazon Q Business はインデックスを作成する ACL 情報を Amazon Q Business User Store に保存して、ユーザーとグループのマッピングを作成し、エンドユーザーのドキュメントへのアクセスに基づいてチャットレスポンスをフィルタリングします。詳細については、「データソースコネクタの概念」を参照してください。

- ドキュメントエンリッチメント – ドキュメントエンリッチメント機能は、インデックスに取り込まれるドキュメントとドキュメント属性の両方と、それらがどのように取り込まれるかを制御するのに役立ちます。これは、次の 2 つのアプローチで実現できます。
  - 基本オペレーションの設定 – 基本オペレーションを使用して、データからドキュメント属性を追加、更新、または削除します。たとえば、PII に関連するドキュメント属性を削除することを選択して、PII データをスクラブできます。
  - Lambda 関数の設定 – 事前設定された Lambda 関数を使用して、よりカスタマイズされた高度なドキュメント属性操作ロジックをデータに対して実行します。例えば、エンタープライズデータはスキャンされたイメージとして保存される場合があります。その場合、Lambda 関数を使用して、スキャンされたドキュメントで光学文字認識 (OCR) を実行し、そこからテキストを抽出できます。次に、スキャンされた各ドキュメントは、取り込み中にテキストドキュメントとして扱われます。最後に、チャット中に、Amazon Q はスキャンされたドキュメントから抽出されたテキストデータをレスポンスの生成時に考慮します。

  ソリューションを実装するときは、両方のドキュメントエンリッチメントアプローチを組み合わせることができます。基本的なオペレーションを使用してデータの最初の解析を行い、Lambda 関数を使用してより複雑なオペレーションを行うことができます。詳細については、「Amazon Q Business でのドキュメントエンリッチメント」を参照してください。
- 統合 – Amazon Q Business アプリケーションを作成したら、 Slackや などの他のアプリケーションに統合できますMicrosoft Teams。例えば、「forAmazon Q Business 用のSlackゲートウェイをデプロイする」と「Amazon Q Business 用のMicrosoft Teamsゲートウェイをデプロイする」(AWS ブログ記事) を参照してください。

エンドユーザーのカスタマイズ

Amazon Q Business は、組織のデータソースとインデックスに保存されていない可能性のあるドキュメントのアップロードをサポートしています。アップロードされたドキュメントは保存されません。これらは、ドキュメントがアップロードされる会話でのみ使用できます。Amazon Q Business は、アップロード用に特定のドキュメントタイプをサポートしています。詳細については、「Amazon Q Business でファイルとチャットをアップロードする」を参照してください。

Amazon Q Business には、ドキュメント属性によるフィルタリング機能が含まれています。管理者とエンドユーザーの両方がこの機能を使用できます。管理者は、属性を使用してエンドユーザーのチャットレスポンスをカスタマイズおよび制御できます。例えば、データソースタイプがドキュメントに関連付けられた属性である場合、チャットレスポンスが特定のデータソースからのみ生成される

ように指定できます。または、選択した属性フィルターを使用して、エンドユーザーがチャットレスポンスの範囲を制限することを許可することもできます。

エンドユーザーは、より広範な Amazon Q Business アプリケーション環境内で、軽量で専用の Amazon Q Apps を作成できます。Amazon Q アプリを使用すると、マーケティングチーム専用のアプリなど、特定のドメインのタスクを自動化できます。

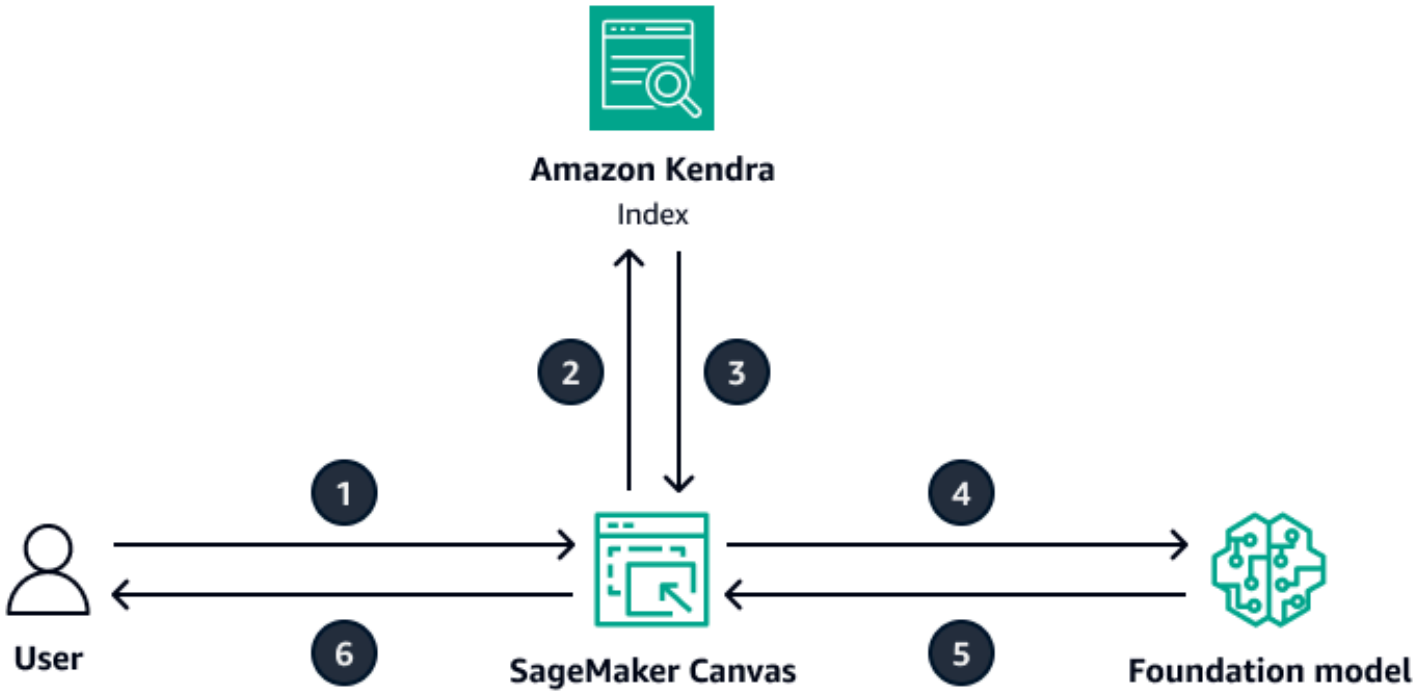
Amazon SageMaker AI Canvas

Amazon SageMaker AI Canvas は、コードを記述することなく、機械学習を使用して予測を生成するのに役立ちます。ML モデルを準備、構築、デプロイできるノーコードのビジュアルインターフェイスを提供し、統合された環境でend-to-endの ML ライフサイクルを合理化します。データ準備、モデル開発、バイアス検出、説明可能性、モニタリングの複雑さは、直感的なインターフェイスの背後で抽象化されます。ユーザーは、SageMaker AI Canvas でモデルを開発、運用、モニタリングするために、SageMaker AI または機械学習オペレーション (MLOps) の専門家である必要はありません。

SageMaker AI Canvas では、RAG 機能はノーコードのドキュメントクエリ機能を通じて提供されます。Amazon Kendra インデックスを基盤となるエンタープライズ検索として使用することで、SageMaker AI Canvas のチャットエクスペリエンスを強化できます。詳細については、「ドキュメントクエリを使用してドキュメントから情報を抽出する」を参照してください。

SageMaker AI Canvas を Amazon Kendra インデックスに接続するには、1 回限りのセットアップが必要です。ドメイン設定の一部として、クラウド管理者は SageMaker Canvas を操作するときにユーザーがクエリできる 1 つ以上の Kendra インデックスを選択できます。ドキュメントクエリ機能を有効にする方法については、Amazon SageMaker AI Canvas の使用開始」を参照してください。

SageMaker AI Canvas は、Amazon Kendra と選択した基盤モデル間の基盤となる通信を管理します。SageMaker AI Canvas がサポートする基盤モデルの詳細については、SageMaker AI Canvas の生成 AI 基盤モデル」を参照してください。次の図は、クラウド管理者が SageMaker AI Canvas を Amazon Kendra インデックスに接続した後のドキュメントクエリ機能の仕組みを示しています。



この図表は、次のワークフローを示しています:

1. ユーザーは SageMaker AI Canvas で新しいチャットを開始し、クエリドキュメントを有効にしてターゲットインデックスを選択し、質問を送信します。
2. SageMaker AI Canvas は、クエリを使用して Amazon Kendra インデックスで関連データを検索します。
3. SageMaker AI Canvas は、Amazon Kendra インデックスからデータとそのソースを取得します。
4. SageMaker AI Canvas は、Amazon Kendra インデックスから取得したコンテキストを含めるようにプロンプトを更新し、基盤モデルにプロンプトを送信します。
5. 基盤モデルは、元の質問と取得したコンテキストを使用して回答を生成します。
6. SageMaker AI Canvas は、生成された回答をユーザーに提供します。これには、レスポンスの生成に使用されたドキュメントなどのデータソースへの参照が含まれます。

でのカスタム取得拡張生成アーキテクチャ AWS

前のセクションでは、完全マネージド AWS のサービス 型の取得拡張生成 (RAG) を使用する方法について説明します。ただし、一部のユースケースでは、リトリーバーや LLM (ジェネレーターとも呼ばれます) などのシステムコンポーネントをより細かく制御する必要があります。たとえば、独自のベクトルデータベースを選択したり、サポートされていないデータソースにアクセスしたりする柔軟性が必要になる場合があります。これらのユースケースでは、カスタム RAG アーキテクチャを構築できます。

このセクションは、以下のトピックで構成されます。

- RAG ワークフローのリトリーバー
- RAG ワークフロー用のジェネレーター

このセクションでリトリーバーオプションとジェネレーターオプションを選択する方法の詳細については、このガイドの「で取得拡張生成オプションを選択する AWS」を参照してください。

RAG ワークフローのリトリーバー

このセクションでは、リトリーバーを構築する方法について説明します。Amazon Kendra などのフルマネージド型のセマンティック検索ソリューションを使用することも、 AWS ベクトルデータベースを使用してカスタムセマンティック検索を構築することもできます。

リトリーバーオプションを確認する前に、ベクトル検索プロセスの 3 つのステップを理解していることを確認してください。

1. インデックス作成する必要があるドキュメントを小さな部分に分割します。これはチャンキングと呼ばれます。
2. 埋め込みと呼ばれるプロセスを使用して、各チャンクを数学ベクトルに変換します。次に、ベクトルデータベース内の各ベクトルのインデックスを作成します。ドキュメントのインデックス作成に使用するアプローチは、検索の速度と精度に影響します。インデックス作成のアプローチは、ベクトルデータベースとそれが提供する設定オプションによって異なります。
3. ユーザークエリをベクトルに変換するには、同じプロセスを使用します。リトリーバーは、ユーザーのクエリベクトルに似たベクトルをベクトルデータベースで検索します。類似度は、ユークリッド距離、コサイン距離、ドット積などのメトリクスを使用して計算されます。

このガイドでは、以下の AWS のサービス またはサードパーティーのサービスを使用してカスタム取得レイヤーを構築する方法について説明します AWS。

- Amazon Kendra
- Amazon OpenSearch Service
- Amazon Aurora PostgreSQL と pgvector
- Amazon Neptune Analytics
- Amazon MemoryDB
- Amazon DocumentDB
- Pinecone
- MongoDB Atlas
- Weaviate

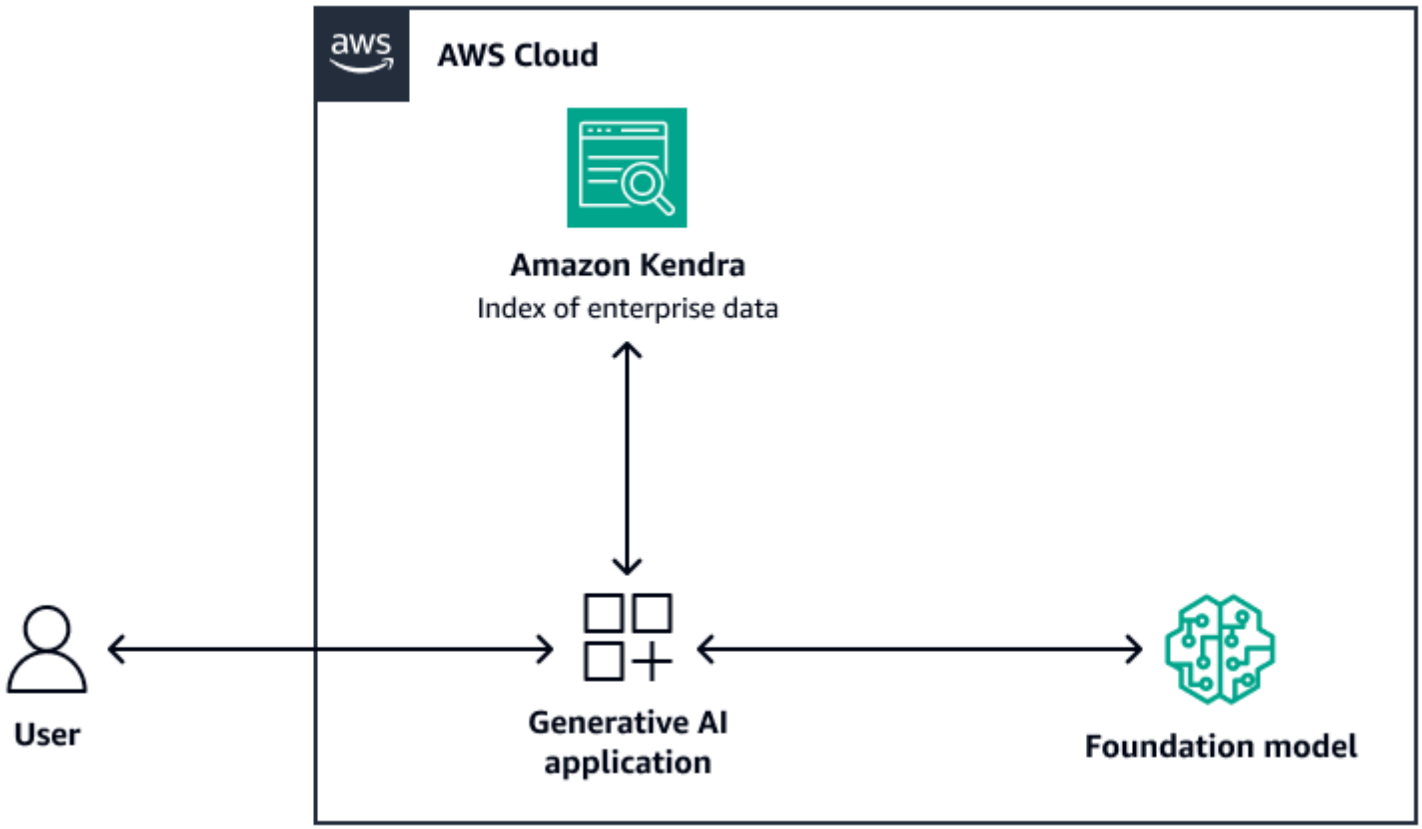
Amazon Kendra

Amazon Kendra は、自然言語処理と高度な機械学習アルゴリズムを使用して、データからの検索質問に対する特定の回答を返す、フルマネージド型のインテリジェントな検索サービスです。Amazon Kendra は、複数のソースからドキュメントを直接取り込み、正常に同期された後にドキュメントをクエリするのに役立ちます。同期プロセスにより、取り込まれたドキュメントでベクトル検索を作成するために必要なインフラストラクチャが作成されます。したがって、Amazon Kendra はベクトル検索プロセスの従来の 3 つのステップを必要としません。最初の同期後、定義されたスケジュールを使用して継続的な取り込みを処理できます。

RAG に Amazon Kendra を使用する利点は次のとおりです。

- Amazon Kendra はベクトル検索プロセス全体を処理するため、ベクトルデータベースを維持する必要はありません。
- Amazon Kendra には、データベース、ウェブサイトクローラー、Amazon S3 バケット、Microsoft SharePointインスタンス、Atlassian Confluenceインスタンスなどの一般的なデータソース用の構築済みコネクタが含まれています。Box および のコネクタなど、 AWS パートナーによって開発されたコネクタを使用できますGitLab。
- Amazon Kendra は、エンドユーザーがアクセスできるドキュメントのみを返すアクセスコントロールリスト (ACL) フィルタリングを提供します。
- Amazon Kendra は、日付やソースリポジトリなどのメタデータに基づいてレスポンスをブーストできます。

次の図は、RAG システムの取得レイヤーとして Amazon Kendra を使用するサンプルアーキテクチャを示しています。詳細については、「Amazon Kendra、、LangChain大規模言語モデルを使用してエンタープライズデータに高精度の生成 AI アプリケーションをすばやく構築する」(AWS ブログ記事) を参照してください。



基盤モデルでは、Amazon Bedrock または Amazon SageMaker AI JumpStart を介してデプロイされた LLM を使用できます。 AWS Lambda で を使用してLangChain、ユーザー、Amazon Kendra、LLM 間のフローをオーケストレーションできます。Amazon Kendra、、LangChainおよびさまざまな LLMs を使用する RAG システムを構築するには、Amazon Kendra LangChain Extensions GitHub リポジトリを参照してください。

Amazon OpenSearch Service

Amazon OpenSearch Service は、ベクトル検索を実行するために、K 最近傍 (k-NN) 検索用の組み込み ML アルゴリズムを提供します。OpenSearch Service は、Amazon EMR Serverless 用のベクトルエンジンも提供します。このベクトルエンジンを使用して、スケーラブルで高性能なベクトルストレージと検索機能を備えた RAG システムを構築できます。OpenSearch Serverless を使用して RAG システムを構築する方法の詳細については、「Amazon OpenSearch Serverless および

Amazon Bedrock Claude モデルのベクトルエンジンを使用してスケーラブルでサーバーレスな RAG ワークフローを構築する」(AWS ブログ記事) を参照してください。

ベクトル検索に OpenSearch Service を使用する利点は次のとおりです。

- OpenSearch Serverless を使用してスケーラブルなベクトル検索を構築するなど、ベクトルデータベースを完全に制御できます。
- これにより、チャンキング戦略を制御できます。
- 非メトリクススペースライブラリ (NMSLIB)、Faiss、Apache Lucene ライブラリからの近似近傍 (ANN) アルゴリズムを使用して、k-NN 検索を強化します。 https://github.com/facebookresearch/faiss https://lucene.apache.org/ユースケースに基づいてアルゴリズムを変更できます。OpenSearch Service を使用してベクトル検索をカスタマイズするためのオプションの詳細については、「Amazon OpenSearch Service のベクトルデータベース機能の説明」(AWS ブログ記事) を参照してください。
- OpenSearch Serverless は、ベクトルインデックスとして Amazon Bedrock ナレッジベースと統合されます。

Amazon Aurora PostgreSQL と pgvector

Amazon Aurora PostgreSQL 互換エディションは、PostgreSQL デプロイのセットアップ、運用、スケーリングに役立つフルマネージドのリレーショナルデータベースエンジンです。pgvector は、ベクトル類似性検索機能を提供するオープンソースの PostgreSQL 拡張機能です。この拡張機能は、Aurora PostgreSQL 互換と Amazon Relational Database Service (Amazon RDS) for PostgreSQL の両方で使用できます。Aurora PostgreSQL 互換と pgvector を使用する RAG ベースのシステムを構築する方法の詳細については、次の AWS ブログ記事を参照してください。

- Amazon SageMaker AI と pgvector を使用して PostgreSQL で AI を活用した検索を構築する
- pgvector と Amazon Aurora PostgreSQL を活用して自然言語処理、チャットボット、感情分析を行う

pgvector と Aurora PostgreSQL 互換を使用する利点は次のとおりです。

- 近傍検索と近似近傍検索をサポートしています。また、L2 距離、内部積、コサイン距離の類似度メトリクスもサポートしています。
- フラット圧縮 (IVFFlat) および階層ナビゲーション可能なスモールワールド (HNSW) インデックスを持つ反転ファイルをサポートしています。 https://github.com/pgvector/pgvector#hnsw

- ベクトル検索を、同じ PostgreSQL インスタンスで利用可能なドメイン固有のデータに対するクエリと組み合わせることができます。
- Aurora PostgreSQL 互換 は I/O 用に最適化されており、階層型キャッシュを提供します。使用可能なインスタンスメモリを超えるワークロードの場合、pgvector はベクトル検索のクエリを 1 秒あたり最大 8 回まで増やすことができます。

Amazon Neptune Analytics

Amazon Neptune Analytics は、分析用のメモリ最適化グラフデータベースエンジンです。グラフトラバーサル内の最適化されたグラフ分析アルゴリズム、低レイテンシーのグラフクエリ、ベクトル検索機能のライブラリをサポートしています。また、ベクトル類似度検索が組み込まれています。グラフの作成、データのロード、クエリの呼び出し、ベクトル類似度検索を実行する 1 つのエンドポイントを提供します。Neptune Analytics を使用する RAG ベースのシステムを構築する方法の詳細については、「ナレッジグラフを使用して Amazon Bedrock と Amazon Neptune で GraphRAG アプリケーションを構築する」(AWS ブログ記事) を参照してください。

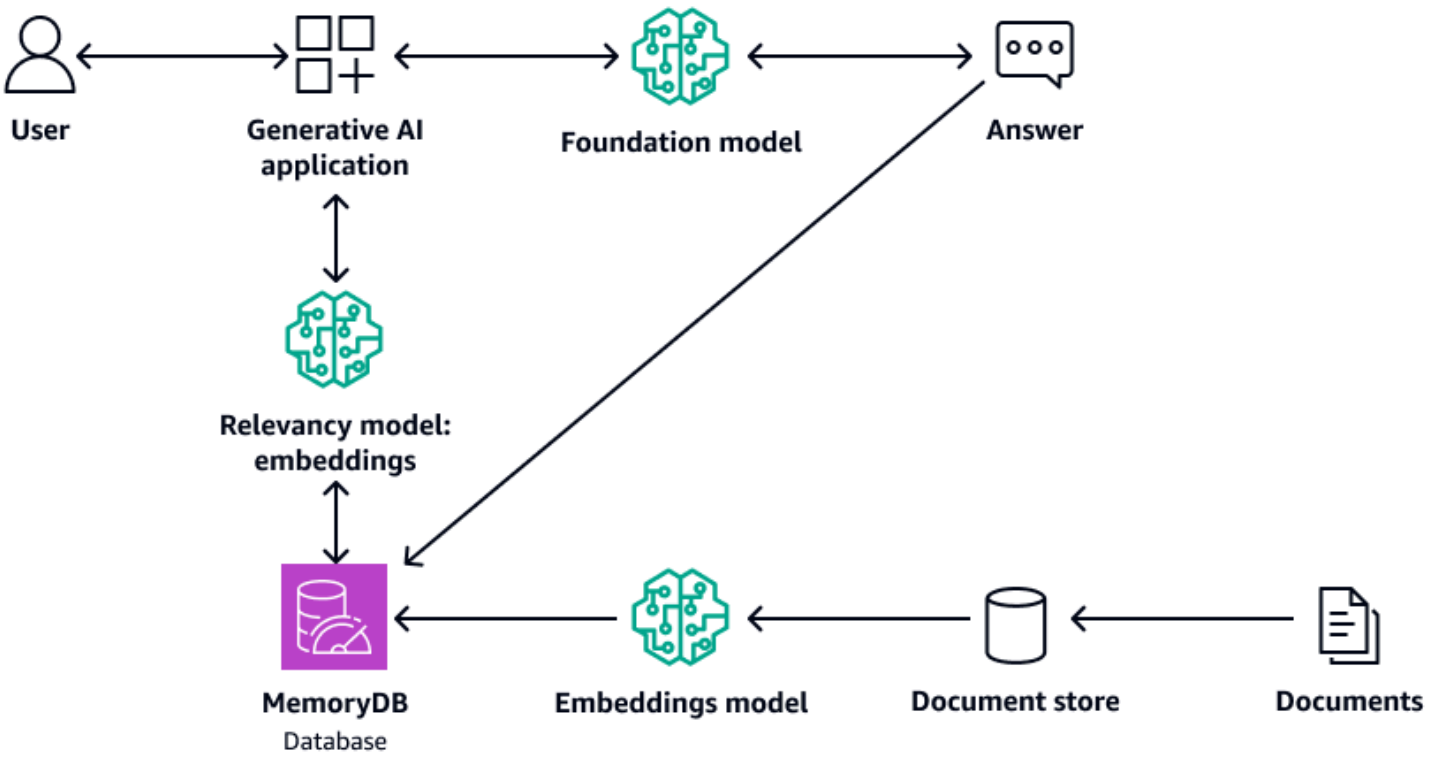
Neptune Analytics を使用する利点は次のとおりです。

- 埋め込みをグラフクエリに保存および検索できます。
- Neptune Analytics を と統合する場合LangChain、このアーキテクチャは自然言語グラフクエリをサポートします。
- このアーキテクチャは、大きなグラフデータセットをメモリに保存します。

Amazon MemoryDB

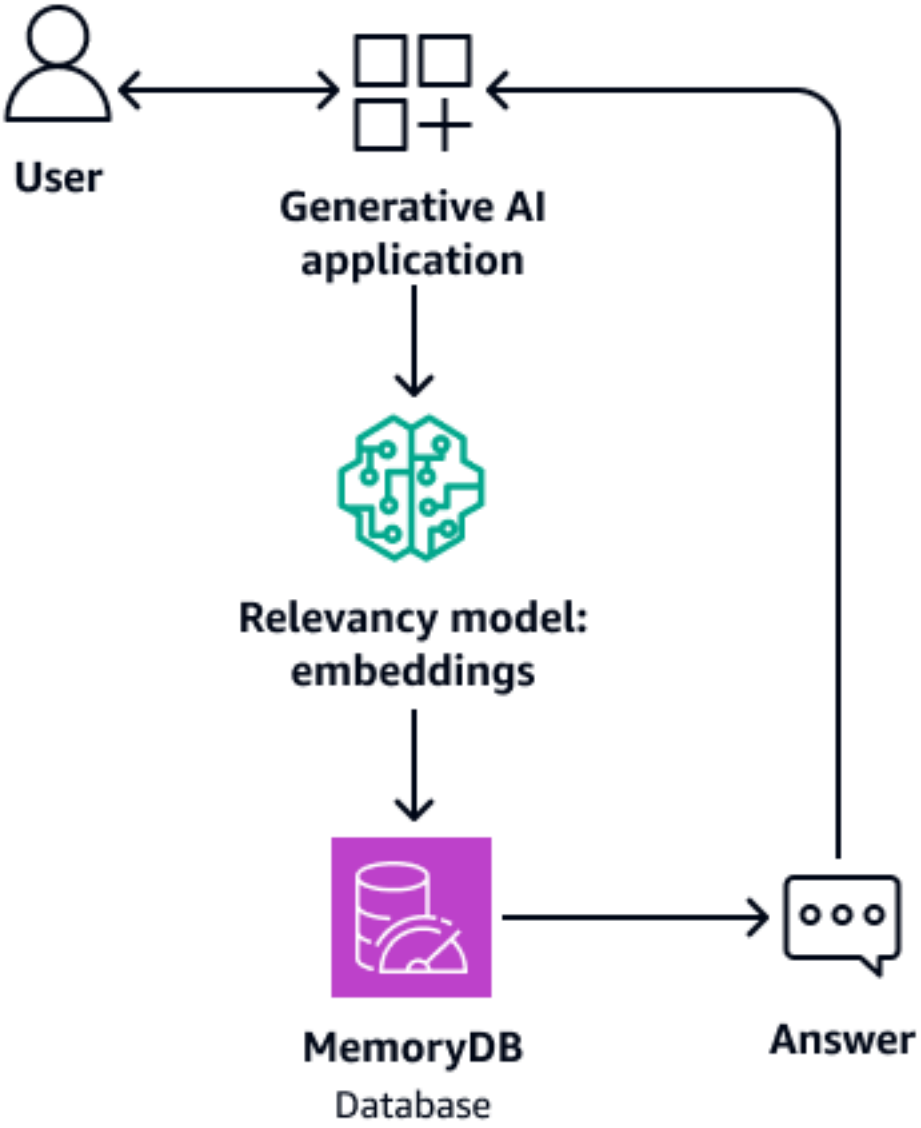
Amazon MemoryDB は、超高速のパフォーマンスを提供する耐久性の高いインメモリデータベースサービスです。すべてのデータはメモリに保存され、マイクロ秒の読み取り、1 桁ミリ秒の書き込みレイテンシー、高スループットをサポートします。MemoryDB のベクトル検索は MemoryDB の機能を拡張し、既存の MemoryDB 機能と組み合わせて使用できます。詳細については、GitHub の「LLM および RAG リポジトリを使用した質問への回答」を参照してください。

次の図は、MemoryDB をベクトルデータベースとして使用するサンプルアーキテクチャを示しています。



MemoryDB を使用する利点は次のとおりです。

- フラットインデックス作成アルゴリズムと HNSW インデックス作成アルゴリズムの両方をサポートしています。詳細については、「 AWS ニュースブログ」の「Amazon MemoryDB のベクトル検索が一般公開されました」を参照してください。
- また、基盤モデルのバッファメモリとしても機能します。つまり、以前に回答した質問は、取得および生成プロセスを繰り返すのではなく、バッファから取得されます。以下の図はこのプロセスを示しています。



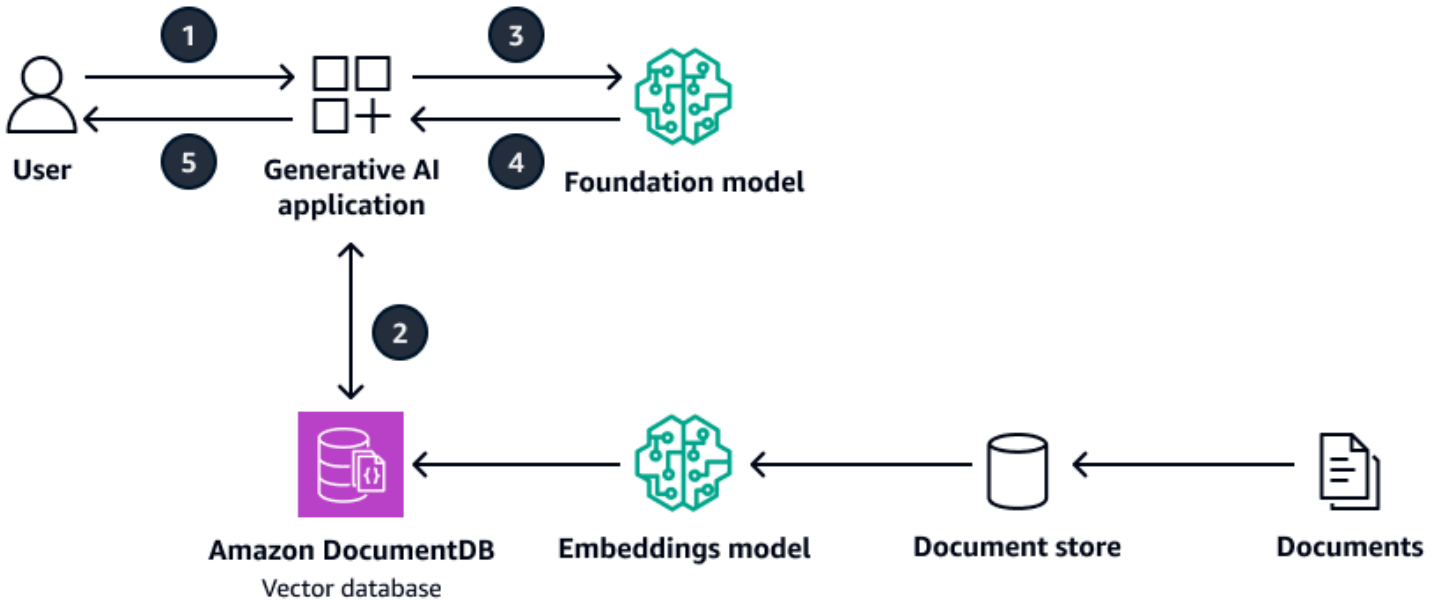
- インメモリデータベースを使用するため、このアーキテクチャはセマンティック検索に 1 桁ミリ秒のクエリ時間を提供します。
- 95～99% の再現率で 1 秒あたり最大 33,000 件のクエリを提供し、99% を超える再現率で 1 秒あたり最大 26,500 件のクエリを提供します。詳細については、AWS 「re:Invent 2023 - Ultra-low latency vector search for Amazon MemoryDB video on 」を参照してくださいYouTube。

Amazon DocumentDB

Amazon DocumentDB (MongoDB 互換) は、高速で信頼性の高いフルマネージドデータベースサービスです。クラウドで MongoDB互換データベースを簡単にセットアップ、運用、スケーリングできます。Amazon DocumentDB のベクトル検索は、JSON ベースのドキュメントデータベースの柔

軟性と豊富なクエリ機能とベクトル検索の能力を組み合わせています。詳細については、GitHub の「LLM および RAG リポジトリを使用した質問への回答」を参照してください。

次の図は、Amazon DocumentDB をベクトルデータベースとして使用するサンプルアーキテクチャを示しています。



この図表は、次のワークフローを示しています:

1. ユーザーは生成 AI アプリケーションにクエリを送信します。
2. 生成 AI アプリケーションは、Amazon DocumentDB ベクトルデータベースで類似度検索を実行し、関連するドキュメント抽出を取得します。
3. 生成 AI アプリケーションは、取得したコンテキストでユーザークエリを更新し、ターゲット基盤モデルにプロンプトを送信します。
4. 基盤モデルは、コンテキストを使用してユーザーの質問に対するレスポンスを生成し、レスポンスを返します。
5. 生成 AI アプリケーションは、ユーザーにレスポンスを返します。

Amazon DocumentDB を使用する利点は次のとおりです。

- HNSW と IVFFlat の両方のインデックス作成メソッドをサポートしています。
- ベクトルデータで最大 2,000 のディメンションをサポートし、ユークリッド、コサイン、ドット積の距離メトリクスをサポートします。

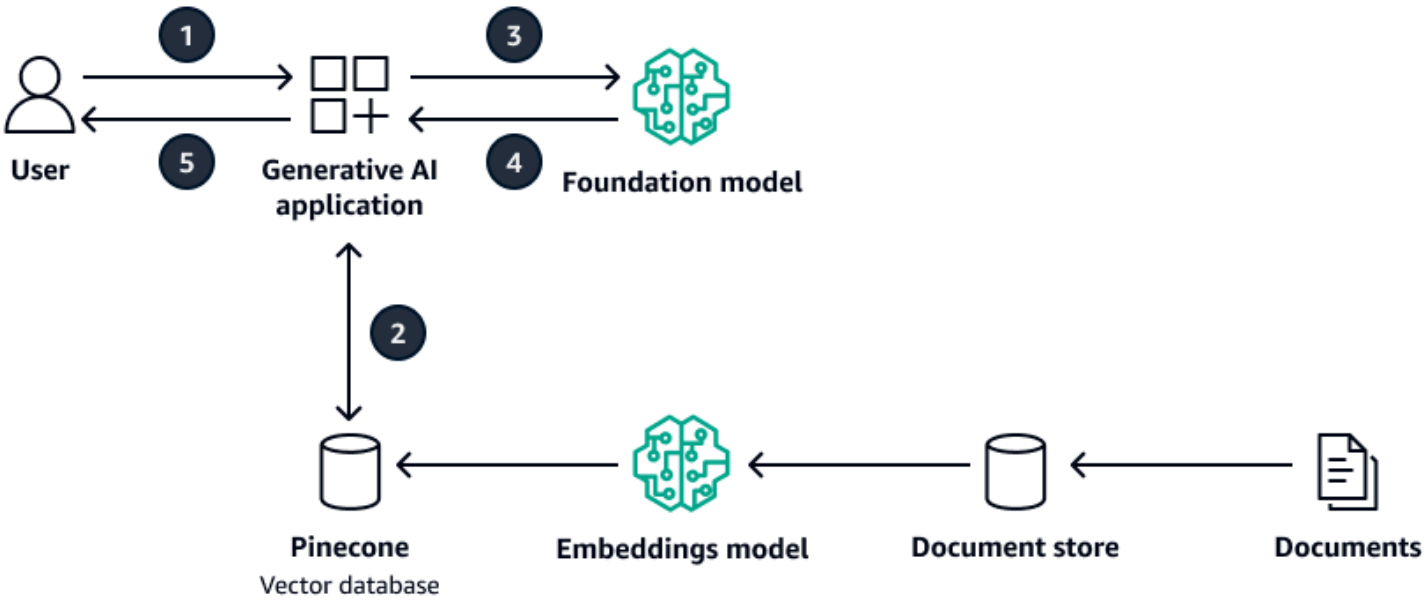
- ミリ秒の応答時間を提供します。

Pinecone

Pinecone は、本番稼働用アプリケーションにベクトル検索を追加するのに役立つフルマネージド型のベクトルデータベースです。これは、 から入手できますAWS Marketplace。請求は使用量に基づいており、料金はポッド料金にポッド数を乗算して計算されます。を使用する RAG ベースのシステムを構築する方法の詳細についてはPinecone、次の AWS ブログ記事を参照してください。

- Amazon SageMaker AI JumpStart のPineconeベクトルデータベースと Llama-2 を使用して RAG を介してハルシネーションを軽減する
- Amazon SageMaker AI Studio を使用して Llama 2、、 で RAG 質問への回答ソリューションを構築しLangChain、Pinecone迅速な実験を行う

次の図は、 をベクトルデータベースPineconeとして使用するアーキテクチャの例を示しています。



この図表は、次のワークフローを示しています:

1. ユーザーは生成 AI アプリケーションにクエリを送信します。
2. 生成 AI アプリケーションは、Pineconeベクトルデータベースで類似度検索を実行し、関連するドキュメント抽出を取得します。
3. 生成 AI アプリケーションは、取得したコンテキストでユーザークエリを更新し、ターゲット基盤モデルにプロンプトを送信します。

4. 基盤モデルは、コンテキストを使用してユーザーの質問に対するレスポンスを生成し、レスポンスを返します。
5. 生成 AI アプリケーションは、ユーザーにレスポンスを返します。

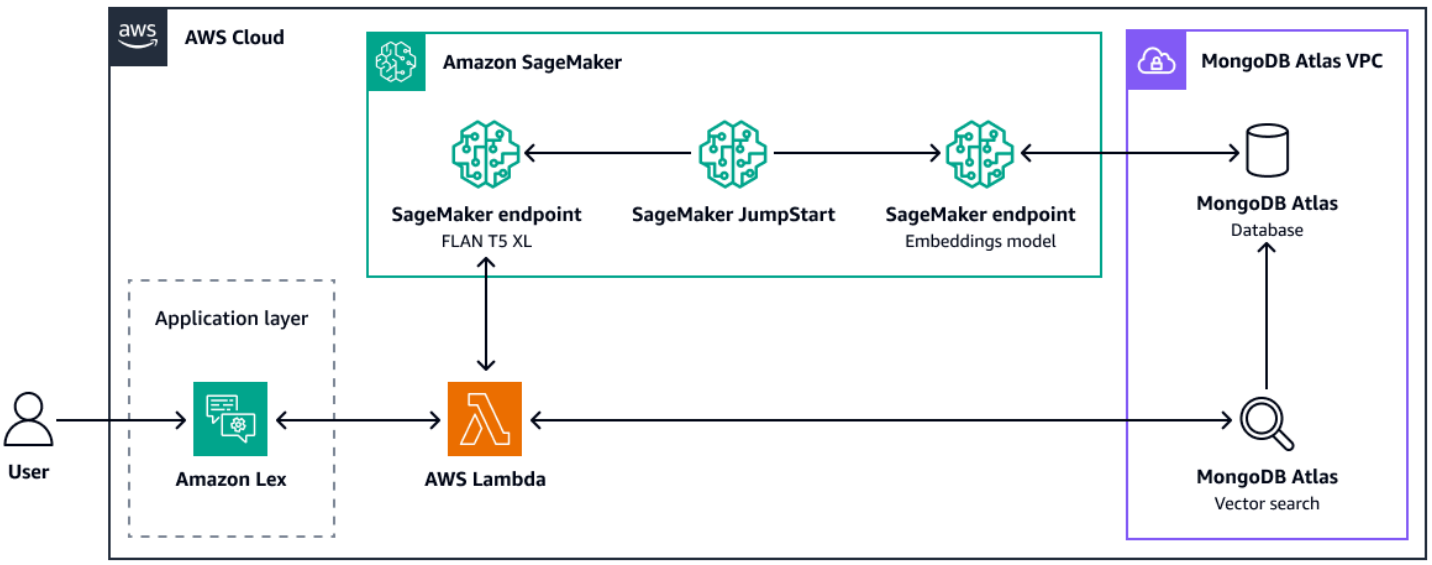
を使用する利点は次のとおりですPinecone。

- これはフルマネージド型のベクトルデータベースであり、独自のインフラストラクチャを管理するオーバーヘッドを排除します。
- フィルタリング、ライブインデックスの更新、キーワードブースト (ハイブリッド検索) の追加機能を提供します。

MongoDB Atlas

MongoDB Atlas は、デプロイのデプロイと管理の複雑さをすべて処理するフルマネージド型のクラウドデータベースです AWS。のベクトル検索MongoDB Atlasを使用して、MongoDBデータベースにベクトル埋め込みを保存できます。Amazon Bedrock ナレッジベースは、ベクトルストレージMongoDB Atlasをサポートしています。詳細については、 MongoDBドキュメントの「Amazon Bedrock ナレッジベース統合の開始方法」を参照してください。

RAG のMongoDB Atlasベクトル検索の使用方法の詳細については、「 を使用した検索拡張生成 LangChain」、Amazon SageMaker AI JumpStart」、およびMongoDB Atlas「セマンティック検索」(AWS ブログ記事) を参照してください。次の図は、このブログ記事で詳しく説明されているソリューションアーキテクチャを示しています。



MongoDB Atlas ベクトル検索を使用する利点は次のとおりです。

- の既存の実装を使用してMongoDB Atlas、ベクトル埋め込みを保存および検索できます。
- MongoDB Query API を使用して、ベクトル埋め込みをクエリできます。
- ベクトル検索とデータベースは個別にスケールできます。
- ベクトル埋め込みはソースデータ (ドキュメント) の近くに保存されるため、インデックス作成のパフォーマンスが向上します。

Weaviate

Weaviate は、テキストや画像などのマルチモーダルメディアタイプをサポートする、一般的なオープンソースの低レイテンシーベクトルデータベースです。データベースにはオブジェクトとベクトルの両方が保存され、ベクトル検索と構造化フィルタリングが組み合わされます。Weaviate と Amazon Bedrock を使用して RAG ワークフローを構築する方法の詳細については、「Build enterprise-ready generative AI solutions with Cohere foundation models in Amazon Bedrock and Weaviate vector database on AWS Marketplace (AWS ブログ記事) 」を参照してください。

を使用する利点は次のとおりですWeaviate。

- これはオープンソースであり、強力なコミュニティに支えられています。
- ハイブリッド検索 (ベクトルとキーワードの両方) 用に構築されています。
- AWS マネージド Software as a Service (SaaS) サービスまたは Kubernetes クラスターとして にデプロイできます。

RAG ワークフロー用のジェネレーター

大規模言語モデル (LLMs) は、膨大な量のデータで事前トレーニングされた非常に大規模な深層学習モデルです。これらは非常に柔軟です。LLMsは、質問への回答、ドキュメントの要約、言語の翻訳、文の完了など、さまざまなタスクを実行できます。コンテンツの作成や、検索エンジンや仮想アシスタントの使用方法が中断される可能性があります。LLMs「完」ではありませんが、比較的少ないプロンプトまたは入力数に基づいて予測を行う能力を示しています。

LLMsは RAG ソリューションの重要なコンポーネントです。カスタム RAG アーキテクチャの場合、プライマリオプションとして機能する AWS のサービス 2 つの があります。

- Amazon Bedrock は、主要な AI 企業と Amazon LLMs を統合 API を通じて使用できるようにするフルマネージドサービスです。

- Amazon SageMaker AI JumpStart は、基盤モデル、組み込みアルゴリズム、構築済みの ML ソリューションを提供する ML ハブです。SageMaker AI JumpStart を使用すると、基盤モデルを含む事前トレーニング済みのモデルにアクセスできます。独自のデータを使用して、事前トレーニング済みモデルを微調整することもできます。

Amazon Bedrock

Amazon Bedrock は、、Anthropic、Stability AI、、Meta、CohereAI21 LabsMistral AI、および Amazon の業界をリードするモデルを提供しています。詳細なリストについては、「Amazon Bedrock でサポートされている基盤モデル」を参照してください。Amazon Bedrock では、独自のデータを使用してモデルをカスタマイズすることもできます。

モデルのパフォーマンスを評価して、RAG ユースケースに最適な を決定できます。最新のモデルをテストし、どの機能や機能が最良の結果を提供し、最も安価であるかをテストすることもできます。Anthropic Claude Sonnet モデルは、幅広いタスクに優れ、高い信頼性と予測可能性を提供するため、RAG アプリケーションの一般的な選択肢です。

SageMaker AI JumpStart

SageMaker AI JumpStart は、さまざまな問題タイプに対応する事前トレーニング済みのオープンソースモデルを提供します。デプロイする前に、これらのモデルを段階的にトレーニングおよび微調整できます。Amazon SageMaker AI Studio の SageMaker AI JumpStart ランディングページから、または SageMaker AI Python SDK を使用して、事前トレーニング済みのモデル、ソリューションテンプレート、および例にアクセスできます。 Amazon SageMaker

SageMaker AI JumpStart は、コンテンツ書き込み、コード生成、質問への回答、コピー書き込み、要約、分類、情報取得などのユースケース向けのstate-of-the-art基盤モデルを提供します。JumpStart 基盤モデルを使用して独自の生成 AI ソリューションを構築し、カスタムソリューションを SageMaker AI 追加機能と統合します。詳細については、Amazon SageMaker AI JumpStart の開始方法」を参照してください。

SageMaker AI JumpStart は、ML ライフサイクルにアクセス、カスタマイズ、統合するために公開されている基盤モデルをオンボードし、維持します。詳細については、「公開されている基盤モデル」を参照してください。SageMaker AI JumpStart には、サードパーティープロバイダーの独自の基盤モデルも含まれています。詳細については、「独自の基盤モデル」を参照してください。

で取得拡張生成オプションを選択する AWS

このガイドの完全マネージド型 RAG オプションとカスタム RAG アーキテクチャのセクションでは、RAG ベースの検索ソリューションを構築するためのさまざまなアプローチについて説明しますAWS。このセクションでは、ユースケースに基づいてこれらのオプションから選択する方法について説明します。状況によっては、複数のオプションが機能することがあります。このシナリオでは、実装のしやすさ、組織で利用できるスキル、会社のポリシーと基準によって異なります。

次の順序でフルマネージド型およびカスタムの RAG オプションを検討し、ユースケースに適した最初のオプションを選択することをお勧めします。

1. 以下の場合を除き、Amazon Q Business を使用してください。
   - このサービスは で利用できず AWS リージョン、利用可能なリージョンにデータを移動することはできません。
   - RAG ワークフローをカスタマイズする特定の理由がある
   - 既存のベクトルデータベースまたは特定の LLM を使用する
2. 以下の場合を除き、Amazon Bedrock のナレッジベースを使用します。
   - サポートされていないベクトルデータベースがある
   - RAG ワークフローをカスタマイズする特定の理由がある
3. 以下の場合を除き、Amazon Kendra を任意のジェネレーターと組み合わせてください。
   - 独自のベクトルデータベースを選択する
   - チャンキング戦略をカスタマイズする
4. リトリーバーをより詳細に制御し、独自のベクトルデータベースを選択する場合:
   - 既存のベクトルデータベースがなく、低レイテンシーやグラフクエリを必要としない場合は、Amazon OpenSearch Service の使用を検討してください。
   - 既存のPostgreSQLベクトルデータベースがある場合は、Amazon Aurora PostgreSQL と pgvector オプションの使用を検討してください。
   - 低レイテンシーが必要な場合は、Amazon MemoryDB や Amazon DocumentDB などのインメモリオプションを検討してください。
   - ベクトル検索をグラフクエリと組み合わせる場合は、Amazon Neptune Analytics を検討してください。
   - 既にサードパーティーのベクトルデータベースを使用している場合や、そのデータベースから特定の利点を見つけた場合は、、PineconeMongoDB Atlas、および を検討してくださいWeaviate。

5. LLM を選択する場合:
   - Amazon Q Business を使用している場合、LLM を選択することはできません。
   - Amazon Bedrock を使用する場合は、サポートされている基盤モデルのいずれかを選択できます。
   - Amazon Kendra またはカスタムベクトルデータベースを使用する場合は、このガイドで説明されているジェネレーターのいずれかを使用するか、カスタム LLM を使用できます。

   > **Note**
   >
   > カスタムドキュメントを使用して既存の LLM を微調整し、レスポンスの精度を高めることもできます。詳細については、このガイドの「RAG とファインチューニングの比較」を参照してください。

6. 使用する Amazon SageMaker AI Canvas の既存の実装がある場合、または異なる LLMs、Amazon SageMaker AI Canvas を検討してください。

結論

このガイドでは、取得拡張生成 (RAG) システムを構築するためのさまざまなオプションについて説明します AWS。Amazon Q Business や Amazon Bedrock ナレッジベースなどのフルマネージドサービスから始めることができます。RAG ワークフローをより詳細に制御したい場合は、カスタムリトリーバーを選択できます。ジェネレーターの場合、API を使用して Amazon Bedrock でサポートされている LLM を呼び出すことも、Amazon SageMaker AI JumpStart を使用して独自の LLM をデプロイすることもできます。「RAG オプションの選択」の推奨事項を確認して、ユースケースに最適なオプションを決定します。ユースケースに最適なオプションを選択したら、このガイドに記載されているリファレンスを使用して RAG ベースのアプリケーションの構築を開始します。

ドキュメント履歴

以下の表は、本ガイドの重要な変更点について説明したものです。今後の更新に関する通知を受け取る場合は、RSS フィード をサブスクライブできます。

変更	説明	日付
初版発行	—	2024 年 10 月 28 日

AWS 規範ガイダンスの用語集

以下は、 AWS 規範ガイダンスによって提供される戦略、ガイド、パターンで一般的に使用される用語です。エントリを提案するには、用語集の最後のフィードバックの提供リンクを使用します。

数字

7 Rs

アプリケーションをクラウドに移行するための 7 つの一般的な移行戦略。これらの戦略は、ガートナーが 2011 年に特定した 5 Rs に基づいて構築され、以下で構成されています。

- リファクタリング/アーキテクチャの再設計 — クラウドネイティブ特徴を最大限に活用して、俊敏性、パフォーマンス、スケーラビリティを向上させ、アプリケーションを移動させ、アーキテクチャを変更します。これには、通常、オペレーティングシステムとデータベースの移植が含まれます。例: オンプレミスの Oracle データベースを Amazon Aurora PostgreSQL 互換エディションに移行する。
- リプラットフォーム (リフトアンドリシェイプ) — アプリケーションをクラウドに移行し、クラウド機能を活用するための最適化レベルを導入します。例: お客様のオンプレミスの Oracle データベースを AWS クラウドの Oracle 用の Amazon Relational Database Service (Amazon RDS) に移行する。
- 再購入 (ドロップアンドショップ) — 通常、従来のライセンスから SaaS モデルに移行して、別の製品に切り替えます。例: 顧客関係管理 (CRM) システムを Salesforce.com に移行する。
- リホスト (リフトアンドシフト) — クラウド機能を活用するための変更を加えずに、アプリケーションをクラウドに移行します。例: お客様のオンプレミスの Oracle データベースを AWS クラウドの EC2 インスタンス上の Oracle に移行する。
- 再配置 (ハイパーバイザーレベルのリフトアンドシフト) – 新しいハードウェアを購入したり、アプリケーションを書き換えたり、既存の運用を変更したりすることなく、インフラストラクチャをクラウドに移行できます。オンプレミスプラットフォームから同じプラットフォームのクラウドサービスにサーバーを移行します。例： Microsoft Hyper-Vアプリケーションを に移行します AWS。
- 保持 (再アクセス) — アプリケーションをお客様のソース環境で保持します。これには、主要なリファクタリングを必要とするアプリケーションや、お客様がその作業を後日まで延期したいアプリケーション、およびそれらを移行するためのビジネス上の正当性がないため、お客様が保持するレガシーアプリケーションなどがあります。

- 廃止 — お客様のソース環境で不要になったアプリケーションを停止または削除します。

A

A2A (Agent-to-Agent)

タスクの委任と状態転送をサポートするagent-to-agentコラボレーション用のステートフルプロトコル。

ABAC

「属性ベースのアクセス制御」をご覧ください。

抽象化されたサービス

「マネージドユーザー」をご覧ください。

ACID

「原子性、一貫性、分離性、耐久性 (ACID)」をご覧ください。

アクティブ/アクティブ移行

(双方向レプリケーションツールまたは二重書き込み操作を使用して) ソースデータベースとターゲットデータベースを同期させ、移行中に両方のデータベースが接続アプリケーションからのトランザクションを処理するデータベース移行方法。この方法では、1 回限りのカットオーバーの必要がなく、管理された小規模なバッチで移行できます。アクティブ/パッシブ移行よりも柔軟な方法ですが、さらに多くの作業が必要となります。

アクティブ/パッシブ移行

ソースデータベースとターゲットデータベースを同期させながら、データがターゲットデータベースにレプリケートされている間、接続しているアプリケーションからのトランザクションをソースデータベースのみで処理するデータベース移行方法。移行中、ターゲットデータベースはトランザクションを受け付けません。

[エージェント]

目標を達成するためのツールを使用して、自律的に推論、計画、アクションを実行できる AI システム。

エージェントオペレーション

AI エージェントを本番環境で大規模に構築、テスト、デプロイ、実行するための運用プラクティス。

集計関数

複数行に処理を行い、グループ全体を対象に単一の戻り値を計算する SQL 関数。集計関数の例としては、SUM や MAX などがあります。

AI

「人工知能」をご覧ください。

AIOps

「AI オペレーション」をご覧ください。

匿名化

データセット内の個人情報を完全に削除するプロセス。匿名化は個人のプライバシー保護に役立ちます。匿名化されたデータは、もはや個人データとは見なされません。

アンチパターン

繰り返し起こる問題に対して頻繁に用いられる解決策で、その解決策が逆効果であったり、効果がなかったり、代替案よりも効果が低かったりするもの。

アプリケーション制御

マルウェアからシステムを保護するために、承認されたアプリケーションのみを使用できるようにするセキュリティアプローチ。

アプリケーションポートフォリオ

アプリケーションの構築と維持にかかるコスト、およびそのビジネス価値を含む、組織が使用する各アプリケーションに関する詳細情報の集まり。この情報は、ポートフォリオの検出と分析プロセス の重要な要素であり、移行、モダナイズ、最適化するアプリケーションを特定し、優先順位を付けるのに役立ちます。

人工知能 (AI)

コンピューティングテクノロジーを使用し、学習、問題の解決、パターンの認識など、通常は人間に関連づけられる認知機能の実行に特化したコンピュータサイエンスの分野。詳細については、「人工知能 (AI) とは何ですか?」をご覧ください。

AI オペレーション (AIOps)

機械学習技術を使用して運用上の問題を解決し、運用上のインシデントと人の介入を減らし、サービス品質を向上させるプロセス。 AWS 移行戦略での AIOps の使用方法については、オペレーション統合ガイド を参照してください。

非対称暗号化

暗号化用のパブリックキーと復号用のプライベートキーから成る 1 組のキーを使用した、暗号化のアルゴリズム。パブリックキーは復号には使用されないため共有しても問題ありませんが、プライベートキーの利用は厳しく制限する必要があります。

原子性、一貫性、分離性、耐久性 (ACID)

エラー、停電、その他の問題が発生した場合でも、データベースのデータ有効性と運用上の信頼性を保証する一連のソフトウェアプロパティ。

属性ベースのアクセス制御 (ABAC)

部署、役職、チーム名など、ユーザーの属性に基づいてアクセス許可をきめ細かく設定する方法。詳細については、 AWS Identity and Access Management (IAM) ドキュメントの「 の ABAC AWS」を参照してください。

信頼できるデータソース

最も信頼性のある情報源とされるデータのプライマリーバージョンを保存する場所。匿名化、編集、仮名化など、データを処理または変更する目的で、信頼できるデータソースから他の場所にデータをコピーすることができます。

アベイラビリティーゾーン (AZ)

他のアベイラビリティーゾーンの障害から AWS リージョン 隔離され、同じリージョン内の他のアベイラビリティーゾーンへの低コストで低レイテンシーのネットワーク接続を提供する 内の別の場所。

AWS クラウド導入フレームワーク (AWS CAF)

組織がクラウドへの移行を成功させるための効率的で効果的な計画を立て AWS るための、 のガイドラインとベストプラクティスのフレームワークです。 AWS CAF は、ビジネス、人材、ガバナンス、プラットフォーム、セキュリティ、運用という 6 つの重点分野にガイダンスを整理しています。ビジネス、人材、ガバナンスの観点では、ビジネススキルとプロセスに重点を置き、プラットフォーム、セキュリティ、オペレーションの視点は技術的なスキルとプロセスに焦点を当てています。例えば、人材の観点では、人事 (HR)、人材派遣機能、および人材管理を扱うステークホルダーを対象としています。この観点から、 AWS CAF は、クラウド導入を成功させるための組織の準備に役立つ人材開発、トレーニング、コミュニケーションに関するガイダンスを提供します。詳細については、AWS CAF ウェブサイトと AWS CAF のホワイトペーパー を参照してください。

AWS ワークロード認定フレームワーク (AWS WQF)

データベース移行ワークロードを評価し、移行戦略を推奨し、作業見積もりを提供するツール。AWS WQF は AWS Schema Conversion Tool (AWS SCT) に含まれています。データベーススキーマとコードオブジェクト、アプリケーションコード、依存関係、およびパフォーマンス特性を分析し、評価レポートを提供します。

B

不正なボット

個人や組織に混乱や損害を与えることを目的としたボット。

BCP

「ビジネス継続性計画 (BCP)」をご覧ください。

動作グラフ

リソースの動作とインタラクションを経時的に示した、一元的なインタラクティブビュー。Amazon Detective の動作グラフを使用すると、失敗したログオンの試行、不審な API 呼び出し、その他同様のアクションを調べることができます。詳細については、Detective ドキュメントの「動作グラフのデータ」を参照してください。

ビッグエンディアンシステム

最上位バイトを最初に格納するシステム。「エンディアン性」もご覧ください。

二項分類

バイナリ結果 (2 つの可能なクラスのうちの 1 つ) を予測するプロセス。例えば、お客様の機械学習モデルで「この E メールはスパムですか、それともスパムではありませんか」などの問題を予測する必要があるかもしれません。または「この製品は書籍ですか、車ですか」などの問題を予測する必要があるかもしれません。

ブルームフィルター

要素がセットのメンバーであるかどうかをテストするために使用される、確率的でメモリ効率の高いデータ構造。

ブルー/グリーンデプロイ

それぞれが独立しているが、同一の環境を 2 つ作成するデプロイ戦略。現在のアプリケーションバージョンを 1 つの環境 (ブルー) で実行し、新しいアプリケーションバージョンを別の環境 (グリーン) で実行します。この戦略は、最小限の影響で迅速にロールバックするのに役立ちます。

ボット

インターネット経由で自動タスクを実行し、人間のアクティビティややり取りをシミュレートするソフトウェアアプリケーション。インターネット上の情報のインデックスを作成するウェブクローラーなど、一部のボットは有用または有益です。悪質なボットと呼ばれる他のボットの中には、個人や組織を混乱させたり、損害を与えたりすることを意図したものもあります。

ボットネット

マルウェアに感染しており、ボットハーダーまたはボットオペレーターと呼ばれる単一の当事者によって制御されているボットのネットワーク。ボットネットは、ボットとその影響力を拡大する仕組みとして、非常によく知られています。

ブランチ

コードリポジトリに含まれる領域。リポジトリに最初に作成するブランチは、メインブランチといいます。既存のブランチから新しいブランチを作成し、その新しいブランチで機能を開発したり、バグを修正したりできます。機能を構築するために作成するブランチは、通常、機能ブランチと呼ばれます。機能をリリースする準備ができたら、機能ブランチをメインブランチに統合します。詳細については、「ブランチの概要」(GitHub ドキュメント) を参照してください。

ブレークグラスアクセス

例外的な状況では、承認されたプロセスを通じて、ユーザーが AWS アカウント 通常アクセス許可を持たない にすばやくアクセスできるようにします。詳細については、 AWS Well-Architected ガイダンスの「ブレークグラス手順の実装」インジケータを参照してください。

ブラウンフィールド戦略

環境の既存インフラストラクチャ。システムアーキテクチャにブラウンフィールド戦略を導入する場合、現在のシステムとインフラストラクチャの制約に基づいてアーキテクチャを設計します。既存のインフラストラクチャを拡張している場合は、ブラウンフィールド戦略とグリーンフィールド戦略を融合させることもできます。

バッファキャッシュ

アクセス頻度が最も高いデータが保存されるメモリ領域。

ビジネス能力

価値を生み出すためにビジネスが行うこと (営業、カスタマーサービス、マーケティングなど)。マイクロサービスのアーキテクチャと開発の決定は、ビジネス能力によって推進できます。詳細については、AWSでのコンテナ化されたマイクロサービスの実行ホワイトペーパーの「ビジネス機能を中心に組織化」セクションを参照してください。

ビジネス継続性計画 (BCP)

大規模移行など、中断を伴うイベントが運用に与える潜在的な影響に対処し、ビジネスを迅速に再開できるようにする計画。

C

CAF

「AWS クラウド導入フレームワーク」を参照してください

カナリアデプロイ

エンドユーザーへのバージョンリリースを、時間をかけて段階的に行うこと。確信が持てたら新規バージョンをデプロイして、現在のバージョン全体を置き換えます。

CCoE

「Cloud Center of Excellence」を参照してください。

CDC

「変更データキャプチャ」を参照してください。

変更データキャプチャ (CDC)

データソース (データベーステーブルなど) の変更を追跡し、その変更に関するメタデータを記録するプロセス。CDC は、ターゲットシステムでの変更を監査またはレプリケートして同期を維持するなど、さまざまな目的に使用できます。

カオスエンジニアリング

障害や破壊的なイベントを意図的に導入して、システムの耐障害性をテストすること。AWS Fault Injection Service (AWS FIS) を使用して、 AWS ワークロードにストレスを与え、その応答を評価する実験を実行できます。

CI/CD

「継続的インテグレーションと継続的デリバリー」を参照してください。

分類

予測を生成するのに役立つ分類プロセス。分類問題の機械学習モデルは、離散値を予測します。離散値は、常に互いに区別されます。例えば、モデルがイメージ内に車があるかどうかを評価する必要がある場合があります。

シチズンデベロッパー

専門的な技術スキルを持たないノーコード/ローコードプラットフォームを使用して AI アプリケーションを作成するビジネスユーザー。

クライアント側の暗号化

ターゲットが AWS のサービス 受信する前に、ローカルでデータを暗号化します。

Cloud Center of Excellence (CCoE)

クラウドのベストプラクティスの作成、リソースの移動、移行のタイムラインの確立、大規模変革を通じて組織をリードするなど、組織全体のクラウド導入の取り組みを推進する学際的なチーム。詳細については、 AWS クラウド エンタープライズ戦略ブログの CCoE 投稿を参照してください。

クラウドコンピューティング

リモートデータストレージと IoT デバイス管理に通常使用されるクラウドテクノロジー。クラウドコンピューティングは、一般的に、エッジコンピューティングに接続されています。

クラウド運用モデル

IT 組織において、1 つ以上のクラウド環境を構築、成熟、最適化するために使用される運用モデル。詳細については、「クラウド運用モデルの構築」を参照してください。

導入のクラウドステージ

組織が、 AWS クラウドへの移行時に通常実行する 4 つの段階。

- プロジェクト — 概念実証と学習を目的として、クラウド関連のプロジェクトをいくつか実行する
- 基礎固め — お客様のクラウドの導入を拡大するための基礎的な投資 (ランディングゾーンの作成、CCoE の定義、運用モデルの確立など)
- 移行 — 個々のアプリケーションの移行
- 再発明 — 製品とサービスの最適化、クラウドでのイノベーション

これらのステージは、 AWS クラウド エンタープライズ戦略ブログのブログ記事「クラウドファーストへのジャーニー」と「導入のステージ」で Stephen Orban によって定義されました。移行戦略との関連性については、 AWS「移行準備ガイド」を参照してください。

CMDB

「構成管理データベース (CMDB)」を参照してください。

コードリポジトリ

ソースコードやその他の資産 (ドキュメント、サンプル、スクリプトなど) が保存され、バージョン管理プロセスを通じて更新される場所。一般的なクラウドリポジトリには、GitHub や Bitbucket Cloud があります。コードの各バージョンはブランチと呼ばれます。マイクロサービスの構造では、各リポジトリは 1 つの機能専用です。1 つの CI/CD パイプラインで複数のリポジトリを使用できます。

コールドキャッシュ

空である、または、かなり空きがある、もしくは、古いデータや無関係なデータが含まれているバッファキャッシュ。データベースインスタンスはメインメモリまたはディスクから読み取る必要があり、バッファキャッシュから読み取るよりも時間がかかるため、パフォーマンスに影響します。

コールドデータ

めったにアクセスされず、通常は過去のデータです。この種類のデータをクエリする場合、通常は低速なクエリでも問題ありません。このデータを低パフォーマンスで安価なストレージ階層またはクラスに移動すると、コストを削減することができます。

コンピュータビジョン (CV)

機械学習を使用してデジタルイメージやビデオといった、ビジュアル形式の情報を分析および抽出する AI の分野。例えば、Amazon SageMaker AI では、CV 用の画像処理アルゴリズムを利用できます。

設定ドリフト

ワークロードにおいて、設定が想定した状態から変化すること。これによって、ワークロードが非準拠になる可能性がありますが、この状態は、徐々に生じ、意図的なものではありません。

構成管理データベース (CMDB)

データベースとその IT 環境 (ハードウェアとソフトウェアの両方のコンポーネントとその設定を含む) に関する情報を保存、管理するリポジトリ。通常、CMDB のデータは、移行のポートフォリオの検出と分析の段階で使用します。

コンフォーマンスパック

コンプライアンスチェックとセキュリティチェックをカスタマイズするためにアセンブルできる AWS Config ルールと修復アクションのコレクション。YAML テンプレートを使用して、コンフォーマンスパックを AWS アカウント および リージョンの単一のエンティティとしてデプロイ

することも、組織全体にデプロイすることもできます。詳細については、 AWS Config ドキュメントの「コンフォーマンスパック」を参照してください。

継続的インテグレーションと継続的デリバリー (CI/CD)

ソフトウェアリリースプロセスのソース、ビルド、テスト、ステージング、本番の各ステージを自動化するプロセス。CI/CD は一般的にパイプラインと呼ばれます。プロセスの自動化、生産性の向上、コード品質の向上、配信の加速化を可能にします。詳細については、「継続的デリバリーの利点」を参照してください。CD は継続的デプロイ (Continuous Deployment) の略語でもあります。詳細については「継続的デリバリーと継続的なデプロイ」を参照してください。

CV

「コンピュータビジョン」を参照してください。

D

保管中のデータ

ストレージ内にあるデータなど、常に自社のネットワーク内にあるデータ。

データ分類

ネットワーク内のデータを重要度と機密性に基づいて識別、分類するプロセス。データに適した保護および保持のコントロールを判断する際に役立つため、あらゆるサイバーセキュリティのリスク管理戦略において重要な要素です。データ分類は、 AWS Well-Architected フレームワークのセキュリティの柱のコンポーネントです。詳細については、「データ分類」を参照してください。

データドリフト

実稼働データと ML モデルのトレーニングに使用されたデータとの間に有意な差異が生じたり、入力データが時間の経過と共に有意に変化したりすることです。データドリフトは、ML モデル予測の全体的な品質、精度、公平性を低下させる可能性があります。

転送中のデータ

ネットワーク内 (ネットワークリソース間など) を活発に移動するデータ。

データメッシュ

非一元的で分散型のデータ所有権を持つとともに、一元的な管理およびガバナンスを行えるアーキテクチャフレームワーク。

データ最小化

厳密に必要なデータのみを収集し、処理するという原則。でデータ最小化を実践 AWS クラウドすることで、プライバシーリスク、コスト、分析のカーボンフットプリントを削減できます。

データ境界

AWS 環境内の一連の予防ガードレール。信頼された ID のみが、期待されるネットワークから信頼されたリソースにアクセスできるようにします。詳細については、「 でのデータ境界の構築 AWS」を参照してください。

データの前処理

raw データをお客様の機械学習モデルで簡単に解析できる形式に変換すること。データの前処理とは、特定の列または行を削除して、欠落している、矛盾している、または重複する値に対処することを意味します。

データ出所

データの生成、送信、保存の方法など、データのライフサイクル全体を通じてデータの出所と履歴を追跡するプロセス。

データ件名

データを収集、処理している個人。

データウェアハウス

分析などのビジネスインテリジェンスをサポートするデータ管理システム。データウェアハウスには、一般的に、大量の履歴データが含まれており、多くの場合、それらはクエリや分析に使用されます。

データベース定義言語 (DDL)

データベース内のテーブルやオブジェクトの構造を作成または変更するためのステートメントまたはコマンド。

データベース操作言語 (DML)

データベース内の情報を変更 (挿入、更新、削除) するためのステートメントまたはコマンド。

DDL

「データベース定義言語」を参照してください。

ディープアンサンブル

予測のために複数の深層学習モデルを組み合わせます。ディープアンサンブルを使用して、より正確な予測を取得したり、予測の不確実性を推定したりできます。

深層学習

人工ニューラルネットワークの複数層を使用して、入力データと対象のターゲット変数の間のマッピングを識別する機械学習サブフィールド。

多層防御

一連のセキュリティメカニズムとコントロールをコンピュータネットワーク全体に層状に重ねて、ネットワークとその内部にあるデータの機密性、整合性、可用性を保護する情報セキュリティの手法。この戦略を採用するときは AWS、 AWS Organizations 構造の異なるレイヤーに複数のコントロールを追加して、リソースの安全性を確保します。たとえば、多層防御アプローチでは、多要素認証、ネットワークセグメンテーション、暗号化を組み合わせることができます。

委任管理者

では AWS Organizations、互換性のあるサービスが AWS メンバーアカウントを登録して組織のアカウントを管理し、そのサービスのアクセス許可を管理できます。このアカウントを、そのサービスの委任管理者と呼びます。詳細、および互換性のあるサービスの一覧は、 AWS Organizations ドキュメントの「AWS Organizationsで使用できるサービス」を参照してください。

トラブルシューティング

アプリケーション、新機能、コードの修正をターゲットの環境で利用できるようにするプロセス。デプロイでは、コードベースに変更を施した後、アプリケーションの環境でそのコードベースを構築して実行します。

開発環境

「環境」を参照してください。

検出管理

イベントが発生したときに、検出、ログ記録、警告を行うように設計されたセキュリティコントロール。これらのコントロールは副次的な防衛手段であり、実行中の予防的コントロールをすり抜けたセキュリティイベントをユーザーに警告します。詳細については、「 AWSでのセキュリティコントロールの実装」の「検出的コントロール」を参照してください。

開発バリューストリームマッピング (DVSM)

ソフトウェア開発ライフサイクルのスピードと品質に悪影響を及ぼす制約を特定し、優先順位を付けるために使用されるプロセス。DVSM は、もともとリーンマニュファクチャリング・プラクティスのために設計されたバリューストリームマッピング・プロセスを拡張したものです。ソフトウェア開発プロセスを通じて価値を創造し、動かすために必要なステップとチームに焦点を当てています。

デジタルツイン

建物、工場、産業機器、生産ラインなど、現実世界のシステムを仮想的に表現したものです。デジタルツインは、予知保全、リモートモニタリング、生産最適化をサポートします。

ディメンションテーブル

スタースキーマにおいて、ファクトテーブルの定量データに関するデータ属性が含まれる小さいテーブル。ディメンションテーブルの属性は、通常、テキストフィールド、またはテキストのように扱える個別の数値で示されます。これらの属性は、一般的に、クエリの制約、フィルタリング、結果セットのラベル付けに使用されます。

ディザスタ

ワークロードまたはシステムが、導入されている主要な場所でのビジネス目標の達成を妨げるイベント。これらのイベントは、自然災害、技術的障害、または意図しない設定ミスやマルウェア攻撃などの人間の行動の結果である場合があります。

ディザスタリカバリ (DR)

ディザスタによるダウンタイムとデータ損失を最小限に抑えるための戦略とプロセス。詳細については、 AWS Well-Architected フレームワークの「 でのワークロードのディザスタリカバリ AWS: クラウドでのリカバリ」を参照してください。

DML

「データベース操作言語」を参照してください。

ドメイン駆動型設計

各コンポーネントが提供している変化を続けるドメイン、またはコアビジネス目標にコンポーネントを接続して、複雑なソフトウェアシステムを開発するアプローチ。この概念は、エリック・エヴァンスの著書、Domain-Driven Design: Tackling Complexity in the Heart of Software (ドメイン駆動設計:ソフトウェアの中心における複雑さへの取り組み) で紹介されています (ボストン: Addison-Wesley Professional、2003)。strangler fig パターンでドメイン駆動型設計を使用す

る方法の詳細については、「コンテナと Amazon API Gateway を使用して、従来の Microsoft ASP.NET (ASMX) ウェブサービスを段階的にモダナイズ」を参照してください。

DR

「ディザスタリカバリ」を参照してください。

ドリフト検出

ベースライン設定からの偏差を追跡します。たとえば、 AWS CloudFormation を使用してシステムリソースのドリフトを検出したり、 を使用して AWS Control Tower 、ガバナンス要件への準拠に影響するランディングゾーンの変更を検出したりできます。

DVSM

「開発バリューストリームマッピング」を参照してください。

E

EDA

「探索的データ分析」を参照してください。

EDI

「電子データ交換」を参照してください。

エッジコンピューティング

IoT ネットワークのエッジにあるスマートデバイスの計算能力を高めるテクノロジー。クラウドコンピューティングと比較すると、エッジコンピューティングは通信レイテンシーを短縮し、応答時間を改善できます。

電子データ交換 (EDI)

組織間で行う、ビジネスドキュメントの自動交換。詳細については、「電子データ交換とは」を参照してください。

暗号化

人間が読み取り可能なプレーンテキストデータを暗号文に変換するコンピューティング処理。

暗号化キー

暗号化アルゴリズムが生成した、ランダム化されたビットからなる暗号文字列。キーの長さは決まっておらず、各キーは予測できないように、一意になるように設計されています。

エンディアン

コンピュータメモリにバイトが格納される順序。ビッグエンディアンシステムでは、最上位バイトが最初に格納されます。リトルエンディアンシステムでは、最下位バイトが最初に格納されます。

エンドポイント

「サービスエンドポイント」を参照してください。

エンドポイントサービス

仮想プライベートクラウド (VPC) 内でホストして、他のユーザーと共有できるサービス。を使用してエンドポイントサービスを作成し AWS PrivateLink 、他の AWS アカウント または AWS Identity and Access Management (IAM) プリンシパルにアクセス許可を付与できます。これらのアカウントまたはプリンシパルは、インターフェイス VPC エンドポイントを作成することで、エンドポイントサービスにプライベートに接続できます。詳細については、Amazon Virtual Private Cloud (Amazon VPC) ドキュメントの「エンドポイントサービスを作成する」を参照してください。

エンタープライズリソースプランニング (ERP)

エンタープライズの主要なビジネスプロセス (会計、MES、プロジェクト管理など) を自動化および管理するシステム。

エンベロープ暗号化

暗号化キーを、別の暗号化キーを使用して暗号化するプロセス。詳細については、 AWS Key Management Service (AWS KMS) ドキュメントの「エンベロープ暗号化」を参照してください。

環境

実行中のアプリケーションのインスタンス。クラウドコンピューティングにおける一般的な環境の種類は以下のとおりです。

- 開発環境 — アプリケーションのメンテナンスを担当するコアチームのみが使用できる、実行中のアプリケーションのインスタンス。開発環境は、上位の環境に昇格させる変更をテストするときに使用します。このタイプの環境は、テスト環境と呼ばれることもあります。
- 下位環境 — 初期ビルドやテストに使用される環境など、アプリケーションのすべての開発環境。
- 本番環境 — エンドユーザーがアクセスできる、実行中のアプリケーションのインスタンス。CI/CD パイプラインでは、本番環境が最後のデプロイ環境になります。

- 上位環境 — コア開発チーム以外のユーザーがアクセスできるすべての環境。これには、本番環境、本番前環境、ユーザー承認テスト環境などが含まれます。

エピック

アジャイル方法論で、お客様の作業の整理と優先順位付けに役立つ機能カテゴリ。エピックでは、要件と実装タスクの概要についてハイレベルな説明を提供します。たとえば、 AWS CAF セキュリティエピックには、ID とアクセスの管理、検出コントロール、インフラストラクチャセキュリティ、データ保護、インシデント対応が含まれます。 AWS 移行戦略のエピックの詳細については、プログラム実装ガイドを参照してください。

ERP

「エンタープライズリソース計画」を参照してください。

探索的データ分析 (EDA)

データセットを分析してその主な特性を理解するプロセス。お客様は、データを収集または集計してから、パターンの検出、異常の検出、および前提条件のチェックのための初期調査を実行します。EDA は、統計の概要を計算し、データの可視化を作成することによって実行されます。

F

ファクトテーブル

スタースキーマの中央にあるテーブル。ビジネスオペレーションに関する定量的データが保存されます。一般的に、ファクトテーブルは、2 種類の列で構成されます。1 つは測定値が含まれる列、もう 1 つはディメンションテーブルへの外部キーが含まれる列です。

フェイルファスト

開発ライフサイクルを短縮するために、頻繁かつ段階的にテストを行う哲学であり、アジャイルアプローチでは、この考え方がきわめて重要です。

障害分離境界

では AWS クラウド、障害の影響を制限し、ワークロードの耐障害性を高めるのに役立つアベイラビリティーゾーン AWS リージョン、コントロールプレーン、データプレーンなどの境界。詳細については、「AWS 障害分離境界」を参照してください。

機能ブランチ

「ブランチ」を参照してください。

特徴量

お客様が予測に使用する入力データ。例えば、製造コンテキストでは、特徴量は製造ラインから定期的にキャプチャされるイメージの可能性もあります。

特徴量重要度

モデルの予測に対する特徴量の重要性。これは通常、Shapley Additive Deskonations (SHAP) や積分勾配など、さまざまな手法で計算できる数値スコアで表されます。詳細については、「 を使用した機械学習モデルの解釈可能性 AWS」を参照してください。

機能変換

追加のソースによるデータのエンリッチ化、値のスケーリング、単一のデータフィールドからの複数の情報セットの抽出など、機械学習プロセスのデータを最適化すること。これにより、機械学習モデルはデータの恩恵を受けることができます。例えば、「2021-05-27 00:15:37」の日付を「2021 年」、「5 月」、「木」、「15」に分解すると、学習アルゴリズムがさまざまなデータコンポーネントに関連する微妙に異なるパターンを学習するのに役立ちます。

数ショットプロンプト

LLM に、タスクと望ましい出力を示す例を少数提示した後に、類似のタスクを実行させること。この手法は、プロンプトに記述された例 (ショット) からモデルが学習する「インコンテキスト学習」の一種です。数ショットプロンプトは、特定のフォーマット、推論、専門知識が必要なタスクに効果的です。「ゼロショットプロンプト」も参照してください。

FGAC

「きめ細かなアクセス制御」を参照してください。

きめ細かなアクセス制御 (FGAC)

複数の条件を使用してアクセス要求を許可または拒否すること。

フラッシュカット移行

変更データのキャプチャによる継続的なデータ複製を利用して、段階的なアプローチではなく、可能な限り短時間でデータを移行するデータベース移行方法。目的はダウンタイムを最小限に抑えることです。

FM

「基盤モデル」を参照してください。

基盤モデル (FM)

大規模な深層学習ニューラルネットワークであり、一般化およびラベル付けされていないデータからなる大規模データセットでトレーニングされています。FM により、言語理解、テキストおよび画像生成、自然言語での会話といった、一般的な各種タスクを実行できます。詳細については、「基盤モデルとは何ですか?」を参照してください。

FM ゲートウェイ

基盤モデルへのアクセスを制御および正規化する一元化された仲介者。LLM ゲートウェイとも呼ばれます。

G

生成 AI

AI モデルのサブセット。大量のデータでトレーニングされており、シンプルなテキストプロンプトを使用して、画像、動画、テキスト、オーディオなどの新しいコンテンツやアーティファクトを作成できます。詳細については、「生成 AI とは何ですか?」を参照してください。

ジオブロッキング

「地理的制限」を参照してください。

地理的制限 (ジオブロッキング)

特定の国のユーザーがコンテンツ配信にアクセスできないようにするための、Amazon CloudFront のオプション。アクセスを許可する国と禁止する国は、許可リストまたは禁止リストを使って指定します。詳細については、CloudFront ドキュメントの「コンテンツの地理的ディストリビューションの制限」を参照してください。

Gitflow ワークフロー

下位環境と上位環境が、ソースコードリポジトリでそれぞれ異なるブランチを使用する方法。Gitflow ワークフローは古いと見なされている方法であり、トランクベースのワークフローは推奨されている新しい方法です。

ゴールデンイメージ

システムまたはソフトウェアのスナップショットであり、システムまたはソフトウェアの新規インスタンスをデプロイするテンプレートとして使用されます。製造の例で言えば、ゴールデンイメージを使用すると、複数のデバイスにソフトウェアをプロビジョニングして、デバイス製造オペレーションの速度、スケーラビリティ、生産性を向上させることができます。

グリーンフィールド戦略

新しい環境に既存のインフラストラクチャが存在しないこと。システムアーキテクチャにグリーンフィールド戦略を導入する場合、既存のインフラストラクチャ (別名ブラウンフィールド) との互換性の制約を受けることなく、あらゆる新しいテクノロジーを選択できます。既存のインフラストラクチャを拡張している場合は、ブラウンフィールド戦略とグリーンフィールド戦略を融合させることもできます。

ガードレール

組織単位 (OU) 全般のリソース、ポリシー、コンプライアンスを管理するのに役立つ概略的なルール。予防ガードレールは、コンプライアンス基準に一致するようにポリシーを実施します。これらは、サービスコントロールポリシーと IAM アクセス許可の境界を使用して実装されます。検出ガードレールは、ポリシー違反やコンプライアンス上の問題を検出し、修復のためのアラートを発信します。これらは AWS Config、、Amazon GuardDuty AWS Security Hub CSPM、、 AWS Trusted Advisor Amazon Inspector、およびカスタム AWS Lambda チェックを使用して実装されます。

ガードレール (AI)

エージェントの入力と出力をフィルタリング、検証、制約して、責任ある安全な AI 動作を確保するのに役立つ安全メカニズム。

H

HA

「高可用性」を参照してください。

異種混在データベースの移行

別のデータベースエンジンを使用するターゲットデータベースへお客様の出典データベースの移行 (例えば、Oracle から Amazon Aurora)。異種間移行は通常、アーキテクチャの再設計作業の一部であり、スキーマの変換は複雑なタスクになる可能性があります。AWS は、スキーマの変換に役立つ AWS SCTを提供します。

高可用性 (HA)

課題や災害が発生した場合に、介入なしにワークロードを継続的に運用できること。HA システムは、自動的にフェイルオーバーし、一貫して高品質のパフォーマンスを提供し、パフォーマンスへの影響を最小限に抑えながらさまざまな負荷や障害を処理するように設計されています。

ヒストリアンのモダナイゼーション

製造業のニーズによりよく応えるために、オペレーションテクノロジー (OT) システムをモダナイズし、アップグレードするためのアプローチ。ヒストリアンは、工場内のさまざまなソースからデータを収集して保存するために使用されるデータベースの一種です。

ホールドアウトデータ

機械学習モデルのトレーニング用データセットから保留される、ラベル付き履歴データの一部。ホールドアウトデータを使用すると、モデル予測をホールドアウトデータと比較して、モデルのパフォーマンスを評価できます。

ヒューhuman-in-the-loop (HitL)

エージェント???の実行が重要な決定時点で人間によるレビューと承認のために一時停止するワークフローパターン。

同種データベースの移行

お客様の出典データベースを、同じデータベースエンジンを共有するターゲットデータベース (Microsoft SQL Server から Amazon RDS for SQL Server など) に移行する。同種間移行は、通常、リホストまたはリプラットフォーム化の作業の一部です。ネイティブデータベースユーティリティを使用して、スキーマを移行できます。

ホットデータ

リアルタイムデータや最近の翻訳データなど、頻繁にアクセスされるデータ。通常、このデータには高速なクエリ応答を提供する高性能なストレージ階層またはクラスが必要です。

ホットフィックス

本番環境の重大な問題を修正するために緊急で配布されるプログラム。緊急性が高いため、通常の DevOps のリリースワークフローからは外れた形で実施されます。

ハイパーケア期間

カットオーバー直後、移行したアプリケーションを移行チームがクラウドで管理、監視して問題に対処する期間。通常、この期間は 1～4 日です。ハイパーケア期間が終了すると、アプリケーションに対する責任は一般的に移行チームからクラウドオペレーションチームに移ります。

I

IaC

「Infrastructure as Code」を参照してください。

ID ベースのポリシー

AWS クラウド 環境内のアクセス許可を定義する 1 つ以上の IAM プリンシパルにアタッチされたポリシー。

アイドル状態のアプリケーション

90 日間の平均的な CPU およびメモリ使用率が 5～20% のアプリケーション。移行プロジェクトでは、これらのアプリケーションを廃止するか、オンプレミスに保持するのが一般的です。

IIoT

「インダストリアル IoT」を参照してください。

イミュータブルインフラストラクチャ

既存インフラストラクチャの更新、パッチ適用、変更などを行わずに、本番環境ワークロードに使用する新規インフラストラクチャをデプロイするモデル。本質的に、イミュータブルインフラストラクチャは、ミュータブルインフラストラクチャよりも一貫性、信頼性、予測性に優れています。詳細については、 AWS Well-Architected フレームワークにある「イミュータブルインフラストラクチャを使用してデプロイする」のベストプラクティスを参照してください。

インバウンド (受信) VPC

AWS マルチアカウントアーキテクチャでは、アプリケーションの外部からネットワーク接続を受け入れ、検査し、ルーティングする VPC。AWS Security Reference Architecture では、アプリケーションとより広範なインターネット間の双方向のインターフェイスを保護するために、インバウンド、アウトバウンド、インスペクションの各 VPC を使用してネットワークアカウントを設定することを推奨しています。

増分移行

アプリケーションを 1 回ですべてカットオーバーするのではなく、小さい要素に分けて移行するカットオーバー戦略。例えば、最初は少数のマイクロサービスまたはユーザーのみを新しいシステムに移行する場合があります。すべてが正常に機能することを確認できたら、残りのマイクロサービスやユーザーを段階的に移行し、レガシーシステムを廃止できるようにします。この戦略により、大規模な移行に伴うリスクが軽減されます。

インダストリー 4.0

2016 年に Klaus Schwab 氏が提唱した用語で、接続、リアルタイムデータ、オートメーション、分析、AI/ML の進歩による、ビジネスプロセスのモダナイズを意味します。

インフラストラクチャ

アプリケーションの環境に含まれるすべてのリソースとアセット。

Infrastructure as Code (IaC)

アプリケーションのインフラストラクチャを一連の設定ファイルを使用してプロビジョニングし、管理するプロセス。IaC は、新しい環境を再現可能で信頼性が高く、一貫性のあるものにするため、インフラストラクチャを一元的に管理し、リソースを標準化し、スケールを迅速に行えるように設計されています。

インダストリアル IoT (IIoT)

製造、エネルギー、自動車、ヘルスケア、ライフサイエンス、農業などの産業部門におけるインターネットに接続されたセンサーやデバイスの使用。詳細については、「インダストリアル IoT (IIoT) デジタルトランスフォーメーション戦略の構築」を参照してください。

インスペクション VPC

AWS マルチアカウントアーキテクチャでは、VPC (同一または異なる 内 AWS リージョン)、インターネット、オンプレミスネットワーク間のネットワークトラフィックの検査を管理する一元化された VPCs。AWS Security Reference Architecture では、アプリケーションとより広範なインターネット間の双方向のインターフェイスを保護するために、インバウンド、アウトバウンド、インスペクションの各 VPC を使用してネットワークアカウントを設定することを推奨しています。

IoT

インターネットまたはローカル通信ネットワークを介して他のデバイスやシステムと通信する、センサーまたはプロセッサが組み込まれた接続済み物理オブジェクトのネットワーク。詳細については、「IoT とは」を参照してください。

解釈可能性

機械学習モデルの特性で、モデルの予測がその入力にどのように依存するかを人間が理解できる度合いを表します。詳細については、「 を使用した機械学習モデルの解釈可能性 AWS」を参照してください。

IoT

「IoT」を参照してください。

IT 情報ライブラリ (ITIL)

IT サービスを提供し、これらのサービスをビジネス要件に合わせるための一連のベストプラクティス。ITIL は ITSM の基盤を提供します。

IT サービス管理 (ITSM)

組織の IT サービスの設計、実装、管理、およびサポートに関連する活動。クラウドオペレーションと ITSM ツールの統合については、オペレーション統合ガイドを参照してください。

ITIL

「IT 情報ライブラリ」を参照してください。

ITSM

「IT サービス管理」を参照してください。

L

ラベルベースアクセス制御 (LBAC)

強制アクセス制御 (MAC) の実装で、ユーザーとデータ自体にそれぞれセキュリティラベル値が明示的に割り当てられます。ユーザーセキュリティラベルとデータセキュリティラベルが交差する部分によって、ユーザーに表示される行と列が決まります。

ランディングゾーン

ランディングゾーンは、スケーラブルで安全な、適切に設計されたマルチアカウント AWS 環境です。これは、組織がセキュリティおよびインフラストラクチャ環境に自信を持ってワークロードとアプリケーションを迅速に起動してデプロイできる出発点です。ランディングゾーンの詳細については、「安全でスケーラブルなマルチアカウント AWS 環境のセットアップ」を参照してください。

大規模言語モデル (LLM)

大量のデータで事前トレーニングされた深層学習 AI モデル。LLM では、質問への回答、ドキュメントの要約、他言語へのテキスト翻訳、文を完成させるなど、さまざまなタスクを実行できます。詳細については、「大規模言語モデル (LLM) とは何ですか?」を参照してください。

大規模な移行

300 台以上のサーバの移行。

LBAC

「ラベルベースアクセス制御」を参照してください。

最小特権

タスクの実行には必要最低限の権限を付与するという、セキュリティのベストプラクティス。詳細については、IAM ドキュメントの「最小特権アクセス許可を適用する」を参照してください。

リフトアンドシフト

「7 Rs」を参照してください。

リトルエンディアンシステム

最下位バイトを最初に格納するシステム。「エンディアン性」もご覧ください。

LLM

「大規模言語モデル」を参照してください。

下位環境

「環境」を参照してください。

M

機械学習 (ML)

パターン認識と学習にアルゴリズムと手法を使用する人工知能の一種。ML は、モノのインターネット (IoT) データなどの記録されたデータを分析して学習し、パターンに基づく統計モデルを生成します。詳細については、「機械学習」を参照してください。

メインブランチ

「ブランチ」を参照してください。

マルウェア

コンピュータのセキュリティやプライバシーを侵害するように設計されたソフトウェア。マルウェアは、コンピュータシステムの中断、機密情報の漏洩、不正アクセスを招く可能性があります。マルウェアの例には、ウイルス、ワーム、ランサムウェア、トロイの木馬、スパイウェア、キーロガーなどがあります。

マネージドサービス

AWS のサービス はインフラストラクチャレイヤー、オペレーティングシステム、プラットフォーム AWS を運用し、エンドポイントにアクセスしてデータを保存および取得します。

マネージドサービスの例として、Amazon Simple Storage Service (Amazon S3) と Amazon DynamoDB が挙げられます。このサービスは、抽象化されたサービスとも呼ばれます。

製造実行システム (MES)

生産プロセスを追跡、モニタリング、文書化、制御するソフトウェアシステムであり、工場では、これによって、原材料から製品を完成させます。

MAP

「Migration Acceleration Program」を参照してください。

MCP

「モデルコンテキストプロトコル」を参照してください。

モデルコンテキストプロトコル (MCP)

エージェントツーツール通信のステートレスプロトコル。

MCP サーバー

Model Context Protocol を通じて 1 つ以上のツールを公開するサービス。

メカニズム

ツールを作成してその導入を推進し、導入結果を調べて調整を行うための包括的なプロセス。メカニズムとは、運用中にそれ自体を強化し改善するサイクルを意味します。詳細については、AWS 「 Well-Architected フレームワーク」の「メカニズムの構築」を参照してください。

メンバーアカウント

組織の一部である管理アカウント AWS アカウント 以外のすべて AWS Organizations。 アカウントが 組織のメンバーになることができるのは、一度に 1 つのみです。

MES

「製造実行システム」を参照してください。

Message Queuing Telemetry Transport (MQTT)

発行/サブスクライブのパターンに基づく、軽量のマシンツーマシン (M2M) 通信プロトコルであり、リソースに限りのある IoT デバイスに使用されます。

マイクロサービス

明確に定義された API を介して通信し、通常は小規模な自己完結型のチームが所有する、小規模で独立したサービスです。例えば、保険システムには、販売やマーケティングなどのビジネス機能、または購買、請求、分析などのサブドメインにマッピングするマイクロサービスが含まれ

る場合があります。マイクロサービスの利点には、俊敏性、柔軟なスケーリング、容易なデプロイ、再利用可能なコード、回復力などがあります。詳細については、AWS 「サーバーレスサービスを使用したマイクロサービスの統合」を参照してください。

マイクロサービスアーキテクチャ

各アプリケーションプロセスをマイクロサービスとして実行する独立したコンポーネントを使用してアプリケーションを構築するアプローチ。これらのマイクロサービスは、軽量 API を使用して、明確に定義されたインターフェイスを介して通信します。このアーキテクチャの各マイクロサービスは、アプリケーションの特定の機能に対する需要を満たすように更新、デプロイ、およびスケーリングできます。詳細については、「 でのマイクロサービスの実装 AWS」を参照してください。

Migration Acceleration Program (MAP)

組織がクラウドに移行するための強力な運用基盤を構築し、移行の初期コストを相殺するのに役立つコンサルティングサポート、トレーニング、サービスを提供する AWS プログラム。MAP には、組織的な方法でレガシー移行を実行するための移行方法論と、一般的な移行シナリオを自動化および高速化する一連のツールが含まれています。

大規模な移行

アプリケーションポートフォリオの大部分を次々にクラウドに移行し、各ウェーブでより多くのアプリケーションを高速に移動させるプロセス。この段階では、以前の段階から学んだベストプラクティスと教訓を使用して、移行ファクトリー チーム、ツール、プロセスのうち、オートメーションとアジャイルデリバリーによってワークロードの移行を合理化します。これは、AWS 移行戦略 の第 3 段階です。

移行ファクトリー

自動化された俊敏性のあるアプローチにより、ワークロードの移行を合理化する部門横断的なチーム。移行ファクトリーチームには、通常、運用、ビジネスアナリストおよび所有者、移行エンジニア、デベロッパー、およびスプリントで作業する DevOps プロフェッショナルが含まれます。エンタープライズアプリケーションポートフォリオの 20～50% は、ファクトリーのアプローチによって最適化できる反復パターンで構成されています。詳細については、このコンテンツセットの移行ファクトリーに関する解説とCloud Migration Factory ガイドを参照してください。

移行メタデータ

移行を完了するために必要なアプリケーションおよびサーバーに関する情報。移行パターンごとに、異なる一連の移行メタデータが必要です。移行メタデータの例としては、ターゲットサブネット、セキュリティグループ、 AWS アカウントなどがあります。

移行パターン

移行戦略、移行先、および使用する移行アプリケーションまたはサービスを詳述する、反復可能な移行タスク。例: AWS Application Migration Service を使用して Amazon EC2 への移行をリホストします。

Migration Portfolio Assessment (MPA)

オンラインツール。これによって、 AWS クラウドに移行するビジネスケースの検証に必要な情報を得られます。MPA は、詳細なポートフォリオ評価 (サーバーの適切なサイジング、価格設定、TCO 比較、移行コスト分析) および移行プラン (アプリケーションデータの分析とデータ収集、アプリケーションのグループ化、移行の優先順位付け、およびウェーブプランニング) を提供します。MPA ツール (ログインが必要) は、すべての AWS コンサルタントと APN パートナーコンサルタントが無料で利用できます。

移行準備状況評価 (MRA)

AWS CAF を使用して、組織のクラウド準備状況に関するインサイトを取得し、長所と短所を特定し、特定されたギャップを埋めるためのアクションプランを構築するプロセス。詳細については、移行準備状況ガイドを参照してください。MRA は、AWS 移行戦略の第一段階です。

移行戦略

ワークロードを AWS クラウドに移行するために使用するアプローチ。詳細については、この用語集の 7 Rs エントリと、「組織を動員して大規模な移行を加速する」を参照してください。

ML

「機械学習」を参照してください。

モダナイゼーション

古い (レガシーまたはモノリシック) アプリケーションとそのインフラストラクチャをクラウド内の俊敏で弾力性のある高可用性システムに変換して、コストを削減し、効率を高め、イノベーションを活用します。詳細については、「AWS クラウドでのアプリケーションのモダナイズ戦略」を参照してください。

モダナイゼーション準備状況評価

組織のアプリケーションのモダナイゼーションの準備状況を判断し、利点、リスク、依存関係を特定し、組織がこれらのアプリケーションの将来の状態をどの程度適切にサポートできるかを決定するのに役立つ評価。評価の結果として、ターゲットアーキテクチャのブループリント、モダナイゼーションプロセスの開発段階とマイルストーンを詳述したロードマップ、特定された

ギャップに対処するためのアクションプランが得られます。詳細については、「AWS クラウドでのアプリケーションのモダナイゼーションの準備状況を評価する」を参照してください。

モノリシックアプリケーション (モノリス)

緊密に結合されたプロセスを持つ単一のサービスとして実行されるアプリケーション。モノリシックアプリケーションにはいくつかの欠点があります。1 つのアプリケーション機能エクスペリエンスの需要が急増する場合は、アーキテクチャ全体をスケーリングする必要があります。モノリシックアプリケーションの特徴を追加または改善することは、コードベースが大きくなると複雑になります。これらの問題に対処するには、マイクロサービスアーキテクチャを使用できます。詳細については、「モノリスをマイクロサービスに分解する」を参照してください。

MPA

「Migration Portfolio Assessment」を参照してください。

MQTT

「Message Queuing Telemetry Transport」を参照してください。

多クラス分類

複数のクラスの予測を生成するプロセス (2 つ以上の結果の 1 つを予測します)。例えば、機械学習モデルが、「この製品は書籍、自動車、電話のいずれですか?」 または、「このお客様にとって最も関心のある商品のカテゴリはどれですか?」と聞くかもしれません。

ミュータブルなインフラストラクチャ

本番ワークロードに使用する既存のインフラストラクチャを更新および変更するためのモデル。Well-Architected AWS フレームワークでは、一貫性、信頼性、予測可能性を向上させるために、イミュータブルインフラストラクチャの使用をベストプラクティスとして推奨しています。

O

OAC

「オリジンアクセス制御」を参照してください。

OAI

「オリジンアクセスアイデンティティ」を参照してください。

OCM

「組織変更管理」を参照してください。

オフライン移行

移行プロセス中にソースワークロードを停止させる移行方法。この方法はダウンタイムが長くなるため、通常は重要ではない小規模なワークロードに使用されます。

OI

「オペレーション統合」を参照してください。

Ola

「オペレーショナルレベルアグリーメント」を参照してください。

オンライン移行

ソースワークロードをオフラインにせずにターゲットシステムにコピーする移行方法。ワークロードに接続されているアプリケーションは、移行中も動作し続けることができます。この方法はダウンタイムがゼロから最小限で済むため、通常は重要な本番稼働環境のワークロードに使用されます。

OPC-UA

「Open Process Communications - Unified Architecture」を参照してください。

Open Process Communications - Unified Architecture (OPC-UA)

産業オートメーション用のマシンツーマシン (M2M) 通信プロトコル。OPC-UA により、相互運用の際に、データ暗号化、認証、認可の各スキームを標準化できます。

オペレーショナルレベルアグリーメント (OLA)

サービスレベルアグリーメント (SLA) をサポートするために、どの機能的 IT グループが互いに提供することを約束するかを明確にする契約。

運用準備状況レビュー (ORR)

質問と関連するベストプラクティスのチェックリスト。インシデントや起こり得る障害を理解、評価、防止したり、その範囲を縮小したりする際に役立ちます。詳細については、 AWS Well-Architected フレームワークの「Operational Readiness Reviews (ORR)」を参照してください。

運用テクノロジー (OT)

産業オペレーション、機器、インフラストラクチャを制御するために物理環境と連携させるハードウェアおよびソフトウェアシステム。製造分野では、Industry 4.0 への変革を進める上で、OT と情報技術 (IT) システムの統合に焦点が当てられています。

オペレーション統合 (OI)

クラウドでオペレーションをモダナイズするプロセスには、準備計画、オートメーション、統合が含まれます。詳細については、オペレーション統合ガイドを参照してください。

組織の証跡

組織 AWS アカウント 内のすべての のすべてのイベント AWS CloudTrail をログに記録する によって作成された証跡 AWS Organizations。証跡は、組織に含まれている各 AWS アカウント に作成され、各アカウントのアクティビティを追跡します。詳細については、CloudTrail ドキュメントの「組織の証跡の作成」を参照してください。

組織変更管理 (OCM)

人材、文化、リーダーシップの観点から、主要な破壊的なビジネス変革を管理するためのフレームワーク。OCM は、変化の導入を加速し、移行問題に対処し、文化や組織の変化を推進することで、組織が新しいシステムと戦略の準備と移行するのを支援します。 AWS 移行戦略では、クラウド導入プロジェクトに必要な変化のスピードにより、このフレームワークは人材アクセラレーションと呼ばれます。詳細については、OCM ガイドを参照してください。

オリジンアクセス制御 (OAC)

Amazon Simple Storage Service (Amazon S3) コンテンツを保護するための、CloudFront のアクセス制限の強化オプション。OAC は AWS リージョン、すべての S3 バケット、 AWS KMS (SSE-KMS) によるサーバー側の暗号化、S3 バケットへの動的 PUT および DELETEリクエストをサポートします。

オリジンアクセスアイデンティティ (OAI)

CloudFront の、Amazon S3 コンテンツを保護するためのアクセス制限オプション。OAI を使用すると、CloudFront が、Amazon S3 に認証可能なプリンシパルを作成します。認証されたプリンシパルは、S3 バケット内のコンテンツに、特定の CloudFront ディストリビューションを介してのみアクセスできます。OAC も併せて参照してください。OAC では、より詳細な、強化されたアクセス制御が可能です。

ORR

「運用準備状況レビュー」を参照してください。

OT

「運用テクノロジー」を参照してください。

アウトバウンド (送信) VPC

AWS マルチアカウントアーキテクチャでは、アプリケーション内から開始されたネットワーク接続を処理する VPC。AWS Security Reference Architecture では、アプリケーションとより広範なインターネット間の双方向のインターフェイスを保護するために、インバウンド、アウトバウンド、インスペクションの各 VPC を使用してネットワークアカウントを設定することを推奨しています。

P

アクセス許可の境界

ユーザーまたはロールが使用できるアクセス許可の上限を設定する、IAM プリンシパルにアタッチされる IAM 管理ポリシー。詳細については、IAM ドキュメントのアクセス許可の境界を参照してください。

個人を特定できる情報 (PII)

直接閲覧した場合、または他の関連データと組み合わせた場合に、個人の身元を合理的に推測するために使用できる情報。PII の例には、氏名、住所、連絡先情報などがあります。

PII

「個人を特定できる情報」を参照してください。

プレイブック

クラウドでのコアオペレーション機能の提供など、移行に関連する作業を取り込む、事前定義された一連のステップ。プレイブックは、スクリプト、自動ランブック、またはお客様のモダナイズされた環境を運用するために必要なプロセスや手順の要約などの形式をとることができます。

PLC

「プログラマブルロジックコントローラー」を参照してください。

PLM

「製品ライフサイクル管理」を参照してください。

ポリシー

次の操作を可能にするオブジェクト: アクセス許可を定義する (ID ベースのポリシーを参照)。アクセス条件を指定する (リソースベースのポリシーを参照)。 AWS Organizations の組織における全アカウントにアクセス許可の上限を定義する (サービスコントロールポリシーを参照)。

多言語の永続性

データアクセスパターンやその他の要件に基づいて、マイクロサービスのデータストレージテクノロジーを個別に選択します。マイクロサービスが同じデータストレージテクノロジーを使用している場合、実装上の問題が発生したり、パフォーマンスが低下する可能性があります。マイクロサービスは、要件に最も適合したデータストアを使用すると、より簡単に実装でき、パフォーマンスとスケーラビリティが向上します。

ポートフォリオ評価

移行を計画するために、アプリケーションポートフォリオの検出、分析、優先順位付けを行うプロセス。詳細については、「移行の準備状況の評価」を参照してください。

述語

`true` または `false` を返すためのクエリ条件。一般的に、`WHERE` 句に記述されます。

述語プッシュダウン

データベースクエリを最適化する手法。これによって、転送前にクエリ内のデータをフィルタリングします。この手法を取ると、リレーショナルデータベースから取得し処理する必要のあるデータの量が減少するため、クエリのパフォーマンスが向上します。

予防的コントロール

イベントの発生を防ぐように設計されたセキュリティコントロール。このコントロールは、ネットワークへの不正アクセスや好ましくない変更を防ぐ最前線の防御です。詳細については、「AWSでのセキュリティコントロールの実装」の「予防的コントロール」を参照してください。

プリンシパル

アクションを実行し AWS 、リソースにアクセスできる のエンティティ。このエンティティは通常、、IAM AWS アカウントロール、またはユーザーのルートユーザーです。詳細については、IAM ドキュメントの「ロールに関する用語と概念」にあるプリンシパルを参照してください。

プライバシーバイデザイン

開発プロセス全体を通してプライバシーが考慮されているシステムエンジニアリングのアプローチ。

プライベートホストゾーン

1 つ以上の VPC 内のドメインとそのサブドメインへの DNS クエリに対し、Amazon Route 53 がどのように応答するかに関する情報を保持するコンテナ。詳細については、Route 53 ドキュメントの「プライベートホストゾーンの使用」を参照してください。

プロアクティブコントロール

非準拠リソースのデプロイ防止を目的としたセキュリティコントロール。このコントロールにより、プロビジョニング前にリソースをスキャンします。コントロールに準拠していないリソースは、プロビジョニングされません。詳細については、 AWS Control Tower ドキュメントの「コントロールリファレンスガイド」および「 セキュリティコントロールの実装」の「プロアクティブコントロール」を参照してください。 AWS

製品ライフサイクル管理 (PLM)

製品の設計、開発、発売から、成長、成熟、衰退、廃棄に至る、製品のライフサイクル全体を通してデータとプロセスを管理すること。

本番環境

「環境」を参照してください。

プログラマブルロジックコントローラー (PLC)

製造分野で使用される、信頼性と適応性に優れたコンピュータであり、これによって、マシンをモニタリングするとともに、製造プロセスを自動化します。

プロンプトチェイニング

1 つの LLM プロンプトによる出力を次のプロンプトの入力に使用して、より良いレスポンスを生成します。この手法を使用すると、複雑なタスクをサブタスクに分割したり、事前レスポンスを繰り返し改良または拡張したりできます。これによって、モデルのレスポンスの精度と関連性が向上し、粒度の高いパーソナライズされた結果を得られます。

仮名化

データセット内の個人識別子をプレースホルダー値に置き換えるプロセス。仮名化は個人のプライバシー保護に役立ちます。仮名化されたデータは、依然として個人データとみなされます。

発行/サブスクライブ (pub/sub)

マイクロサービス間の非同期通信を可能にするパターン。これにより、スケーラビリティと応答性を向上させます。例えば、マイクロサービスベースの MES の場合、マイクロサービスは、他のマイクロサービスがサブスクライブ可能なチャネルにイベントメッセージを発行できます。このシステムでは、発行サービスの変更なしに、新規マイクロサービスを追加できます。

Q

クエリプラン

手順などの一連のステップであり、SQL リレーショナルデータベースシステムのデータにアクセスするために使用されます。

クエリプランのリグレッション

データベースサービスのオプティマイザーが、データベース環境に特定の変更が加えられる前に選択されたプランよりも最適性の低いプランを選択すること。これは、統計、制限事項、環境設定、クエリパラメータのバインディングの変更、およびデータベースエンジンの更新などが原因である可能性があります。

R

RACI マトリックス

「実行責任者、説明責任者、協業先、報告先 (RACI)」を参照してください。

RAG

「検索拡張生成」を参照してください。

ランサムウェア

決済が完了するまでコンピュータシステムまたはデータへのアクセスをブロックするように設計された、悪意のあるソフトウェア。

RASCI マトリックス

「実行責任者、説明責任者、協業先、報告先 (RACI)」を参照してください。

RCAC

「行と列のアクセス制御」を参照してください。

リードレプリカ

読み取り専用に使用されるデータベースのコピー。クエリをリードレプリカにルーティングして、プライマリデータベースへの負荷を軽減できます。

リアーキテクト

「7 Rs」を参照してください。

目標復旧時点 (RPO)

最後のデータリカバリポイントからの最大許容時間です。これにより、最後の回復時点からサービスが中断されるまでの間に許容できるデータ損失の程度が決まります。

目標復旧時間 (RTO)

サービスの中断から復旧までの最大許容遅延時間。

リファクタリング

「7 Rs」を参照してください。

リージョン

地理的エリア内の AWS リソースのコレクション。各 AWS リージョン は、耐障害性、安定性、耐障害性を提供するために、他の とは独立しています。詳細については、「アカウントが使用できる AWS リージョン を指定する」を参照してください。

リグレッション

数値を予測する機械学習手法。例えば、「この家はどれくらいの値段で売れるでしょうか?」という問題を解決するために、機械学習モデルは、線形回帰モデルを使用して、この家に関する既知の事実 (平方フィートなど) に基づいて家の販売価格を予測できます。

リホスト

「7 Rs」を参照してください。

リリース

デプロイプロセスで、変更を本番環境に昇格させること。

再配置

「7 Rs」を参照してください。

リプラットフォーム

「7 Rs」を参照してください。

再購入

「7 Rs」を参照してください。

回復性

中断に抵抗または中断から回復するアプリケーションの機能。 AWS クラウドでの回復力を計画する際には、一般的に、高可用性とディザスタリカバリが考慮されます。詳細については、「AWS クラウド の耐障害性」を参照してください。

リソースベースのポリシー

Amazon S3 バケット、エンドポイント、暗号化キーなどのリソースにアタッチされたポリシー。このタイプのポリシーは、アクセスが許可されているプリンシパル、サポートされているアクション、その他の満たすべき条件を指定します。

実行責任者、説明責任者、協業先、報告先 (RACI) に基づくマトリックス

移行活動とクラウド運用に関わるすべての関係者の役割と責任を定義したマトリックス。マトリックスの名前は、マトリックスで定義されている責任の種類、すなわち責任 (R) 、説明責任 (A) 、協議 (C) 、情報提供 (I) に由来します。サポート (S) タイプはオプションです。サポートが含まれる場合は RASCI マトリックスと呼ばれ、含まれない場合は RACI マトリックスと呼ばれます。

レスポンシブコントロール

有害事象やセキュリティベースラインからの逸脱について、修復を促すように設計されたセキュリティコントロール。詳細については、「 AWSでのセキュリティコントロールの実装」の「レスポンシブコントロール」を参照してください。

保持

「7 Rs」を参照してください。

廃止

「7 Rs」を参照してください。

検索拡張生成 (RAG)

生成 AI の技術。これにより、LLM では、レスポンスの生成前に、トレーニングデータソースの外部にある信頼できるデータソースが参照されます。例えば、RAG モデルによって、組織のナレッジベースまたはカスタムデータのセマンティック検索を実行できる場合があります。細については、「RAG (検索拡張生成) とは何ですか?」を参照してください。

ローテーション

定期的にシークレット情報を更新して、攻撃者が認証情報にアクセスするのをより困難にするプロセス。

行と列のアクセス制御 (RCAC)

アクセスルールが定義された、基本的で柔軟な SQL 表現の使用。RCAC は行権限と列マスクで構成されています。

RPO

「目標復旧時点」を参照してください。

RTO

「目標復旧時間」を参照してください。

ランブック

特定のタスクを実行するために必要な手動または自動化された一連の手順。これらは通常、エラー率の高い反復操作や手順を合理化するために構築されています。

S

SAML 2.0

多くの ID プロバイダー (IdP) が使用しているオープンスタンダード。この機能を使用すると、フェデレーティッドシングルサインオン (SSO) が有効になるため、ユーザーは にログイン AWS マネジメントコンソール したり AWS API オペレーションを呼び出したりでき、組織内のすべてのユーザーを IAM で作成する必要はありません。SAML 2.0 ベースのフェデレーションの詳細については、IAM ドキュメントの「SAML 2.0 ベースのフェデレーションについて」を参照してください。

SCADA

「監視制御とデータ取得」を参照してください。

SCP

「サービスコントロールポリシー」を参照してください。

シークレット

暗号化された形式で保存する AWS Secrets Managerパスワードやユーザー認証情報などの機密情報または制限付き情報。シークレット値とそのメタデータで構成されます。シークレット値には、バイナリ、1 つの文字列、複数の文字列を指定できます。詳細については、Secrets Manager ドキュメントの「Secrets Manager シークレットの概要」を参照してください。

セキュリティバイデザイン

開発プロセス全体を通してセキュリティが考慮されているシステムエンジニアリングのアプローチ。

セキュリティコントロール

脅威アクターによるセキュリティ脆弱性の悪用を防止、検出、軽減するための、技術上または管理上のガードレール。セキュリティコントロールには、主に 4 つの種類があります。4 つとは、予防、検出、レスポンシブ、プロアクティブです。

セキュリティ強化

アタックサーフェスを狭めて攻撃への耐性を高めるプロセス。このプロセスには、不要になったリソースの削除、最小特権を付与するセキュリティのベストプラクティスの実装、設定ファイル内の不要な機能の無効化、といったアクションが含まれています。

Security Information and Event Management (SIEM) システム

セキュリティ情報管理 (SIM) とセキュリティイベント管理 (SEM) のシステムを組み合わせたツールとサービス。SIEM システムは、サーバー、ネットワーク、デバイス、その他ソースからデータを収集、モニタリング、分析して、脅威やセキュリティ違反を検出し、アラートを発信します。

セキュリティレスポンスの自動化

セキュリティイベントへの自動レスポンスまたは自動修復を目的として、事前定義およびプログラムされたアクション。これらの自動化は、セキュリティのベストプラクティスを実装するのに役立つ検出的または応答的な AWS セキュリティコントロールとして機能します。自動レスポンスアクションの例には、VPC セキュリティグループの変更、Amazon EC2 インスタンスへのパッチ適用、認証情報の更新などがあります。

サーバー側の暗号化

送信先にあるデータを、 AWS のサービス が受信する によって暗号化します。

サービスコントロールポリシー (SCP)

AWS Organizationsの組織内の、すべてのアカウントのアクセス許可を一元的に管理するポリシー。SCP は、管理者がユーザーまたはロールに委任するアクションに、ガードレールを定義したり、アクションの制限を設定したりします。SCP は、許可リストまたは拒否リストとして、許可または禁止するサービスやアクションを指定する際に使用できます。詳細については、 AWS Organizations ドキュメントの「サービスコントロールポリシー」を参照してください。

サービスエンドポイント

のエントリポイントの URL AWS のサービス。ターゲットサービスにプログラムで接続するには、エンドポイントを使用します。詳細については、「AWS 全般のリファレンス」の「AWS のサービス エンドポイント」を参照してください。

サービスレベルアグリーメント (SLA)

サービスのアップタイムやパフォーマンスなど、IT チームがお客様に提供すると約束したものを明示した合意書。

サービスレベルインジケータ (SLI)

エラー率、可用性、スループットといった、サービスパフォーマンス面の指標。

サービスレベル目標 (SLO)

サービスレベルインジケータによって測定され、サービスの状態を表すターゲットメトリクス。

責任共有モデル

クラウドのセキュリティとコンプライアンス AWS について と共有する責任を説明するモデル。AWS はクラウドのセキュリティを担当しますが、お客様はクラウドのセキュリティを担当します。詳細については、「責任共有モデル」を参照してください。

シャドウ AI

組織内の管理対象チャネルの外部で構築または使用される認可されていない AI アプリケーション。

SIEM

「Security Information and Event Management システム」を参照してください。

単一障害点 (SPOF)

特定のアプリケーションを構成する単一の重要なコンポーネントで発生し、システム稼働に支障をきたす可能性のある障害。

SLA

「サービスレベルアグリーメント」を参照してください。

SLI

「サービスレベルインジケータ」を参照してください。

SLO

「サービスレベルの目標」を参照してください。

スプリットアンドシードモデル

モダナイゼーションプロジェクトのスケーリングと加速のためのパターン。新機能と製品リリースが定義されると、コアチームは解放されて新しい製品チームを作成します。これにより、お

客様の組織の能力とサービスの拡張、デベロッパーの生産性の向上、迅速なイノベーションのサポートに役立ちます。詳細については、「AWS クラウドでのアプリケーションをモダナイズするための段階的アプローチ」を参照してください。

SPOF

「単一障害点」を参照してください。

スタースキーマ

データベースの編成構造を意味し、1 つの大きいファクトテーブルにトランザクションデータまたは測定データが保存され、1 つ以上の小さいディメンションテーブルにデータ属性が保存されます。この構造は、データウェアハウスやビジネスインテリジェンスを用途とするように設計されています。

strangler fig パターン

レガシーシステムが廃止されるまで、システム機能を段階的に書き換えて置き換えることにより、モノリシックシステムをモダナイズするアプローチ。このパターンは、宿主の樹木から根を成長させ、最終的にその宿主を包み込み、宿主に取って代わるイチジクのつるを例えています。そのパターンは、モノリシックシステムを書き換えるときのリスクを管理する方法として Martin Fowler により提唱されました。このパターンの適用方法の例については、「コンテナと Amazon API Gateway を使用して、従来の Microsoft ASP.NET (ASMX) ウェブサービスを段階的にモダナイズ」を参照してください。

サブネット

VPC 内の IP アドレスの範囲。サブネットは、1 つのアベイラビリティーゾーンに存在する必要があります。

監視制御とデータ取得 (SCADA)

製造分野において、ハードウェアとソフトウェアを使用して物理アセットと本番運用をモニタリングするシステム。

対称暗号化

データの暗号化と復号に同じキーを使用する暗号化のアルゴリズム。

合成テスト

ユーザーとのやり取りをシミュレートして、起こり得る問題を検出したり、パフォーマンスをモニタリングしたりすることで、システムをテストします。Amazon CloudWatch Synthetics を使用すると、こうしたテストを作成できます。

システムプロンプト

コンテキスト、指示、ガイドラインなどを提示して、LLM に動作を指示する手法。システムプロンプトは、コンテキストを設定して、ユーザーとやり取りするルールを確立するのに有用です。

T

タグ

AWS リソースを整理するためのメタデータとして機能するキーと値のペア。タグは、 リソースの管理、識別、整理、検索、フィルタリングに役立ちます。詳細については、「AWS リソースのタグ付け」を参照してください。

ターゲット変数

監督された機械学習でお客様が予測しようとしている値。これは、結果変数のことも指します。例えば、製造設定では、ターゲット変数が製品の欠陥である可能性があります。

タスクリスト

ランブックの進行状況を追跡するために使用されるツール。タスクリストには、ランブックの概要と完了する必要のある一般的なタスクのリストが含まれています。各一般的なタスクには、推定所要時間、所有者、進捗状況が含まれています。

テスト環境

「環境」を参照してください。

トレーニング

お客様の機械学習モデルに学習するデータを提供すること。トレーニングデータには正しい答えが含まれている必要があります。学習アルゴリズムは入力データ属性をターゲット (お客様が予測したい答え) にマッピングするトレーニングデータのパターンを検出します。これらのパターンをキャプチャする機械学習モデルを出力します。そして、お客様が機械学習モデルを使用して、ターゲットがわからない新しいデータでターゲットを予測できます。

tool

エージェントが外部システムでオペレーションを実行するために呼び出すことができる関数または API。

トランジットゲートウェイ

VPC とオンプレミスネットワークを相互接続するために使用できる、ネットワークの中継ハブ。詳細については、 AWS Transit Gateway ドキュメントの「トランジットゲートウェイとは」を参照してください。

トランクベースのワークフロー

デベロッパーが機能ブランチで機能をローカルにビルドしてテストし、その変更をメインブランチにマージするアプローチ。メインブランチはその後、開発環境、本番前環境、本番環境に合わせて順次構築されます。

信頼されたアクセス

ユーザーに代わって AWS Organizations およびそのアカウントで組織内でタスクを実行するために指定したサービスにアクセス許可を付与します。信頼されたサービスは、サービスにリンクされたロールを必要なときに各アカウントに作成し、ユーザーに代わって管理タスクを実行します。詳細については、 ドキュメントの「 を他の AWS のサービス AWS Organizations で使用する AWS Organizations 」を参照してください。

チューニング

機械学習モデルの精度を向上させるために、お客様のトレーニングプロセスの側面を変更する。例えば、お客様が機械学習モデルをトレーニングするには、ラベル付けセットを生成し、ラベルを追加します。これらのステップを、異なる設定で複数回繰り返して、モデルを最適化します。

ツーピザチーム

2 枚のピザを分け合えることができるくらい小さな DevOps チーム。ツーピザチームの規模では、ソフトウェア開発におけるコラボレーションに最適な機会が確保されます。

U

不確実性

予測機械学習モデルの信頼性を損なう可能性がある、不正確、不完全、または未知の情報を指す概念。不確実性には、次の 2 つのタイプがあります。認識論的不確実性は、限られた、不完全なデータによって引き起こされ、弁論的不確実性 は、データに固有のノイズとランダム性によって引き起こされます。

未分化なタスク

ヘビーリフティングとも呼ばれ、アプリケーションの作成と運用には必要だが、エンドユーザーに直接的な価値をもたらさなかったり、競争上の優位性をもたらしたりしない作業です。未分化なタスクの例としては、調達、メンテナンス、キャパシティプランニングなどがあります。

上位環境

「環境」を参照してください。

V

バキューミング

ストレージを再利用してパフォーマンスを向上させるために、増分更新後にクリーンアップを行うデータベースのメンテナンス操作。

バージョンコントロール

リポジトリ内のソースコードへの変更など、変更を追跡するプロセスとツール。

VPC ピアリング

プライベート IP アドレスを使用してトラフィックをルーティングできる、2 つの VPC 間の接続。詳細については、Amazon VPC ドキュメントの「VPC ピア機能とは」を参照してください。

脆弱性

システムのセキュリティを脅かすソフトウェアまたはハードウェアの欠陥。

W

ウォームキャッシュ

頻繁にアクセスされる最新の関連データを含むバッファキャッシュ。データベースインスタンスはバッファキャッシュから、メインメモリまたはディスクからよりも短い時間で読み取りを行うことができます。

ウォームデータ

アクセス頻度の低いデータ。この種類のデータをクエリする場合、通常は適度に遅いクエリでも問題ありません。

ウィンドウ関数

現在のレコードに何らかの形で関連している行のグループに計算を実行する SQL 関数。ウィンドウ関数は、移動平均を計算したり、現在の行の相対位置に基づいて他の行の値にアクセスするといったタスクの処理に役立ちます。

ワークロード

ビジネス価値をもたらすリソースとコード (顧客向けアプリケーションやバックエンドプロセスなど) の総称。

ワークストリーム

特定のタスクセットを担当する移行プロジェクト内の機能グループ。各ワークストリームは独立していますが、プロジェクト内の他のワークストリームをサポートしています。たとえば、ポートフォリオワークストリームは、アプリケーションの優先順位付け、ウェーブ計画、および移行メタデータの収集を担当します。ポートフォリオワークストリームは、これらの設備を移行ワークストリームで実現し、サーバーとアプリケーションを移行します。

WORM

「Write-Once-Read-Many」を参照してください。

WQF

「AWS ワークロード資格フレームワーク」を参照してください

Write-Once-Read-Many (WORM)

データを 1 回のみ書き込むことで、データの削除や変更を防ぐストレージモデル。承認済みユーザーは、必要な回数だけデータを読み取ることができますが、変更することはできません。このデータストレージインフラストラクチャは、イミュータブルと見なされます。

Z

ゼロデイエクスプロイト

ゼロデイ脆弱性を悪用した攻撃 (一般的にマルウェアによる)。

ゼロデイ脆弱性

実稼働システムにおける未解決の欠陥または脆弱性。脅威アクターは、このような脆弱性を利用してシステムを攻撃する可能性があります。開発者は、よく攻撃の結果で脆弱性に気付きます。

ゼロショットプロンプト

LLM にタスク実行の手順は提示するが、実行のガイドとして役立つ例 (ショット) は提示しない方法。LLM は、事前トレーニング済みの知識を使用してタスクを処理する必要があります。ゼロショットプロンプトの有効性は、タスクの複雑さとプロンプトの品質によって異なります。「数ショットプロンプト」も参照してください。

ゾンビアプリケーション

平均 CPU およびメモリ使用率が 5% 未満のアプリケーション。移行プロジェクトでは、これらのアプリケーションを廃止するのが一般的です。

翻訳は機械翻訳により提供されています。提供された翻訳内容と英語版の間で齟齬、不一致または矛盾がある場合、英語版が優先します。