ソリューション概要

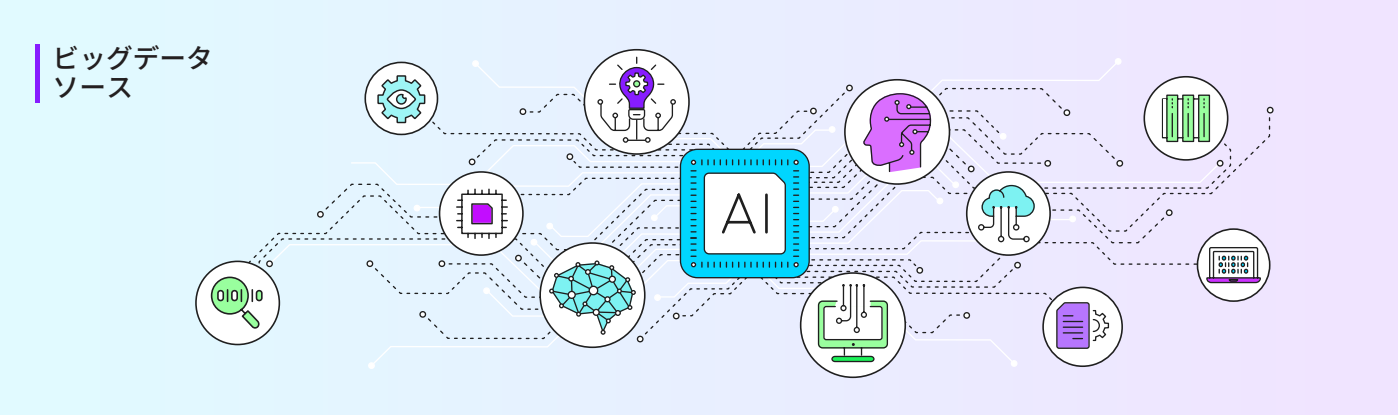
AIワークロードにおけるInfiniBox®の活用

検索拡張生成(RAG):機械学習(ML)と人工知能(AI)の導入を促進する新しい「キラーアプリ」

現代の企業は組織規模を問わず、膨大な量の多様なデータセットを蓄積しています。その目的は、それらのデータを機械学習や予測モデリング、その他の高度な分析に活用することで、ビジネス上の問題を解決し、実用的なインサイトを取得することです。

しかしそこで課題となるのが、AI大規模言語モデル(LLM)を実行するために必要となる大量の計算リソースとエネルギー(電力と冷却)です。このような課題のため、多くの企業はこの分野への投資に踏み切れず、代わりにクラウドプロバイダーのAIサービスの採用を選んでいます。そうしたなか、小規模言語モデル(SLM)と呼ばれるより小型でコンパクトなモデルも登場し、広まりを見せています。必要とされるパラメーターと計算リソースが少ないSLMは、実行とトレーニングに高額なコストがかかるLLMに比べてより実用的でカスタマイズが容易であり、特に初めてのAIプロジェクトにおいてそのようなコスト確保が難しい企業様にとって有力な選択肢です。

LLMとSLMは通常、公開データに基づいてトレーニングされているため、特定分野の情報や企業の非公開データを学習する機会が不足しています。これらの制約はしばしば、「ハルシネーション」と呼ばれる不正確で信頼性の低い回答の生成につながります。これは、ChatGPTなどの公開されている生成AIチャットボットの多くで認められている現象であり、それらのサービスでは誤情報が生成される可能性についての警告が表示されます。ハルシネーションはもっともらしい回答のように見えることが多いですが、実際には間違っており、企業の顧客体験にさまざまな形で影響する可能性があります。



InfiniBoxによる検索拡張生成(RAG)でハルシネーションを解決

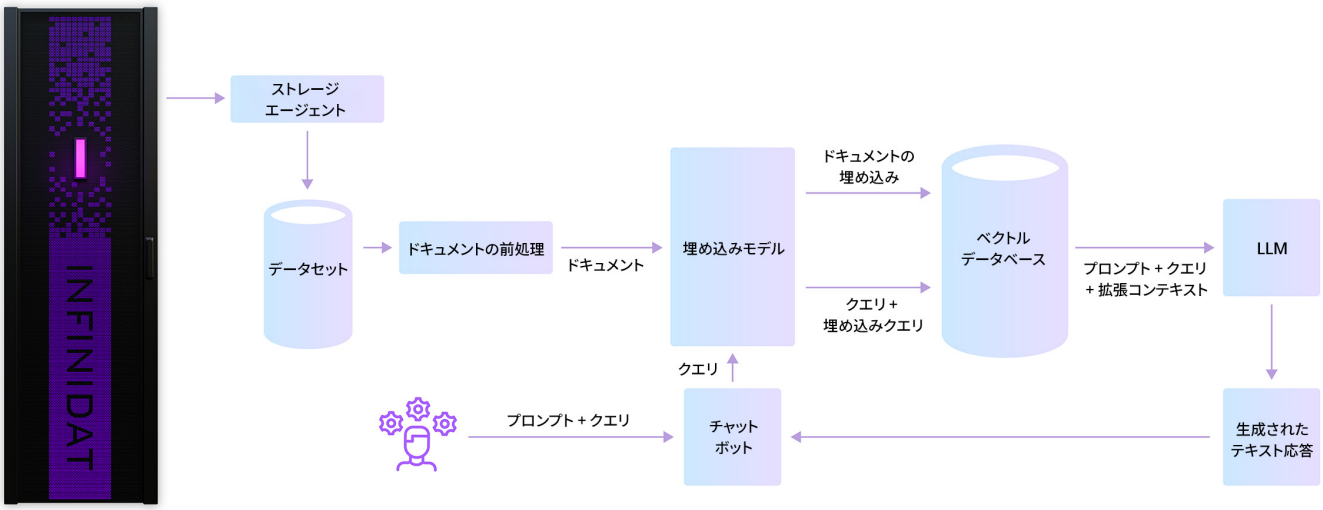
不正確なデータを使用すると、誤った情報に基づいてビジネス上の意思決定を下したり、顧客に誤った情報を提供したりすることになり、企業の評判が損なわれるほか、業務の非効率化につながりかねません。

生成AI(GenAI)は強力ですが、その信頼性を高めそれを維持するためには、信頼できる最新の正確なデータで継続的にAIを強化する必要があります。検索拡張生成(RAG)は、この課題を克服してAIの精度を向上させる高価値なツールとして登場しました。RAGは生成AIのパワーと企業の非公開データを組み合わせて、新たなインサイトの取得とワークフローの効率化を可能にします。外部ソースからの情報をLLMとSLMに統合することで、企業はより正確で目的に特化した成果物を得られます。

ではどこから取り組みを始めればよいのか、と不安に感じられるかもしれませんが、RAGの導入は必ずしも困難ではありません。おそらく多くの企業は、自社インフラストラクチャ内の既存のシステムやストレージ内に、利用することのできるデータを保有しているでしょう。個々のAIワークロードにおいて、AIはInfiniBoxにとって他のワークロードと同じワークロードの1つにすぎないことは特記すべき点です。InfiniBoxは、さまざまな最新のワークロードに対してスケーラブルで一貫性のあるI/Oを提供する、革新的な設計を備えたエンタープライズストレージソリューションです。

RAGワークフローは、既存のオープンソース製品と企業のオンプレミスのデータセンターにすでに存在するデータから簡単に作成できます。下の図は、Infinidatの開発者が作成した、プロセスの概要を示すRAGワークフローアーキテクチャです。

InfinidatのRAGワークフローアーキテクチャ[1]



InfinidatのRAGワークフローアーキテクチャはKubernetesクラスター上で実行されます。オンプレミスのデータを使用してRAGを実行したいけれど利用可能なGPUリソースがないという企業は、クラウドを活用した高速で便利なソリューションを利用できます。当社のアプローチでは、RAGパイプラインを実行するための基盤としてKubernetesクラスターを使用することで、優れた可用性、拡張性、リソース効率性を実現します。AWS Terraformの採用により、RAGシステムのセットアップが大幅に簡素化され、すべての自動化を1つのコマンドで実行できるようになっています。また、オンプレミスのInfiniBoxとInfuzeOS™ Cloud Edition間で同じコアコードが実行されるため、レプリケーションが非常にスムーズです。10分以内に、RAGシステムがフルに機能してInfuzeOS Cloud Edition上のデータを処理できるようになります。

RAGパイプラインの開発は本質的に反復的なプロセスであり、正確性を維持するためには更新と保守が必要です。最新の進歩を継続的に取り入れ、RAGパイプラインを常時改良することで、企業はAIから得られるインサイトの精度と実用性を大幅に向上させ、この新しいテクノロジーがもたらす競争優位性を最大限に高めることができます。

RAGにおけるInfiniBoxの強み

Infinidatをご利用中のお客様の場合、データセンターのInfiniBoxには、言語モデルへの埋め込みに必要なドキュメントを作成するためのデータセットの「ビッグデータ」レポジトリがすでにある可能性もあります。埋め込み(ベクトル)を格納するデータベースの選択は、RAGシステムにおいて非常に重要です。幸いなことに、Oracle、Postgres、MongoDB、DataStax Enterpriseなどの広く使用されているデータベースエンジンの現在のバージョンでは、ベクトルデータの格納と取得がサポートされているため、RAGに対応しています。現在の当社のお客様の多くは、Oracleを利用するミッションクリティカルなアプリケーションでInfiniBoxを使用しています。

InfuzeOSが提供する強力なデータサービスレイヤーにより、すべてのインフラストラクチャコンポーネントをソフトウェアで制御できるようになります。これは非常に強力な機能です。InfuzeOSは、革新的なニューラルキャッシュ(Neural Cache)を利用して、データの配置を動的かつ予測的に最適化します。最もアクティブなデータ(ホットデータ)をDRAMキャッシュに配置して、SSDまたはHDDからのデータ取得を最小限に抑えます。InfuzeOSにより、

Fortune Business Insightsは、ビッグデータテクノロジー市場の規模は2023年の3,494億ドルから、2032年までに1兆1,943億5,000万ドルに成長すると予測しています。

オンプレミスおよびクラウド環境全体のさまざまな最新のワークロードに対して、大規模で継続的に最適化されたパフォーマンス、最高クラスの低レイテンシ、効果的な運用サポートが確保されます。

この強みがRAGワークロードにもたらすのは、高次元ベクトルがデータベースに保存される埋め込みの段階でのInfiniBoxによる非常に低いレイテンシと超高速のパフォーマンスであり、これによりプロセスが円滑に進みます。

ユーザーが(ChatGPTなどに)質問すると、そのクエリはベクトルデータベース内の既存の埋め込みと同じ領域内に存在する埋め込みに変換されます。類似性検索により、ベクトルデータベースはクエリに最も近いベクトルをすばやく識別して応答します。ここでもInfiniBoxの超低レイテンシにより、生成AIワークロードに対する迅速な応答が実現します。

さらに、ベクトルデータベースをInfiniBoxに保存することで他のメリットも得られます。データの取り込みには多くのリソースと時間がかかりますが、結果はベクトルデータベースに保存されます。このデータベースはスナップショットやレプリケーションが可能で、他のRAGワークフローの展開にすぐにプラグインできるデータベースとして利用できるため、他の展開でのリソース要件が緩和されます。

データが前処理され、効率的かつ迅速に取得できるデータストレージシステム(InfiniBox)に保存されると、大規模言語モデル(LLM)による処理に移ります。LLMは一貫性のある文脈に適した応答を生成します。LLMはデータベースから関連データを取得した後、取得した知識とその理解を組み合わせて包括的な回答を生成します。

最後に、InfuzeOS Cloud Edition for AWS and Azureがオンプレミスと同様のクラウドベースのストレージ操作を提供します。ストレージ管理者はデータセンターと同じInfuzeOSの強力な機能をクラウドで利用して、データセットをAIリソースに迅速にレプリケーションできます。このようなストレージの標準化により、すべての環境におけるストレージ管理のコストを削減でき、複雑さが軽減されます。

RAGにおけるInfiniBoxのメリットは主に、Neural Cacheなどのイノベーションによって、このワークロードに最適なパフォーマンスを誇るストレージプラットフォームを提供できる点にあります。Infinidatのソリューションは数の制限なくInfiniBoxプラットフォームを網羅できるほか、NFSなどのファイルベースのプロトコルを介してサードパーティのストレージソリューションへの拡張も可能です。また、効率的なAIインフラストラクチャに不可欠な要素である、100%の可用性とサイバーリカバリの保証も重要な強みです。

生成AIと検索拡張生成(RAG)の活用は、AIモデルの精度と適切性を高めるうえで大いに役立ちます。企業は、InfiniBoxアーキテクチャへの既存の投資と上記のRAGワークフローを活用して、既存のプライベートデータセットから高速で応答性の高いAIモデルを作成し、新たなインサイトを入手してワークフローを効率化できます。

[1] RAGワークフローの設計および開発:Andrew Wang

jp_sales@infinidat.com

SB-AIRAGA-241105-JP | © 2026 Infinidat, a Lenovo company

Lenovo INFINIDAT