ジェネレーティブAIとNetAppの価値

NetApp artificial intelligence solutions

NetApp
August 18, 2025

目次

ジェネレーティブAIとNetAppの価値

生成型人工知能 (AI) の需要は業界全体に混乱をもたらし、ビジネスの創造性と製品のイノベーションを促進しています。

概要

多くの組織が生成 AI を使用して、新しい製品機能を構築し、エンジニアリングの生産性を向上させ、より優れた結果と消費者エクスペリエンスを提供する AI 搭載アプリケーションのプロトタイプを作成しています。Generative Pre-trained Transformers (GPT) などの生成 AI は、ニューラル ネットワークを使用して、テキスト、オーディオ、ビデオなど多様な新しいコンテンツを作成します。大規模言語モデル (LLM) には極めて大きな規模と膨大なデータセットが関係するため、企業が AI ソリューションを設計する前に、オンプレミス、ハイブリッド、マルチクラウドの導入オプションの魅力的なデータ ストレージ機能を活用し、データのモビリティ、データ保護、ガバナンスに関連するリスクを軽減する堅牢な AI インフラストラクチャを設計することが重要です。このホワイト ペーパーでは、これらの考慮事項と、生成 AI モデルのトレーニング、再トレーニング、微調整、推論のための AI データ パイプライン全体でシームレスなデータ管理とデータ移動を可能にする対応するNetApp AI 機能について説明します。

概要

最近では、2022 年 11 月に GPT-3 のスピンオフである ChatGPT がリリースされて以来、ユーザーのプロンプトに応じてテキスト、コード、画像、さらには治療用タンパク質を生成するために使用される新しい AI ツールが大きな注目を集めています。これは、ユーザーが自然言語を使用してリクエストを行うことができ、AI がユーザーのリクエストを反映したニュース記事や製品の説明などのテキストを解釈して生成したり、既存のデータでトレーニングされたアルゴリズムを使用してコード、音楽、音声、視覚効果、3D アセットを生成したりすることを意味します。その結果、安定的な拡散、幻覚、プロンプトエンジニアリング、価値の調整などのフレーズが AI システムの設計で急速に登場しています。これらの自己教師型または半教師型の機械学習 (ML) モデルは、クラウド サービス プロバイダーやその他の AI 企業ベンダーを通じて、事前トレーニング済みの基礎モデル (FM) として広く利用できるようになり、さまざまな業界のさまざまな企業で、幅広い下流の NLP (自然言語処理) タスクに採用されています。マッキンゼーなどの調査アナリスト企業は、「生成 AI が生産性に与える影響は、世界経済に数兆ドルの価値をもたらす可能性がある」と主張しています。企業が AI を人間の思考パートナーとして再考するとともに、FM が同時に企業や機関が生成 AI で何ができるかまで範囲を広げるなか、膨大な量のデータを管理する機会は今後も増え続けるでしょう。このドキュメントでは、オンプレミス、ハイブリッド、またはマルチクラウド環境の両方でNetApp の顧客に価値をもたらすNetApp の機能に関連する生成 AI と設計コンセプトの概要を説明します。

では、顧客が **AI** 環境で**NetApp**を使用するとどのようなメリットがあるのでしょうか**?** NetApp は、データとクラウドの急速な増加、マルチクラウド管理、AI などの次世代テクノロジーの導入によって生じる複雑さに組織が対応できるよう支援します。 NetApp は、AI ワークロードに最適化された高性能とバランスの取れたインテリジェントなデータ管理ソフトウェアとストレージ インフラストラクチャにさまざまな機能を統合しました。 LLM のような生成 AI ソリューションでは、インテリジェンスを強化するために、ソース データセットをストレージからメモリに何度も読み取って処理する必要があります。 NetApp は、エッジからコア、クラウドまでのエコシステム全体にわたるデータ モビリティ、データ ガバナンス、データ セキュリティ テクノロジーのリーダーであり、大規模な AI ソリューションを構築するエンタープライズ顧客にサービスを提供しています。 NetApp は、強力なパートナー ネットワークを活用して、最高データ責任者、AI エンジニア、エンタープライズ アーキテクト、データ サイエンティストを支援し、AI モデルのトレーニングと推論におけるデータ準備、データ保護、戦略的データ管理の責任を果たすための自由に流れるデータ パイプラインを設計し、AI/ML ライフサイクルのパフォーマンスと拡張性を最適化しています。ディープラーニング データ パイプライン用のNetApp ONTAP AI、ストレージ エンドポイント間でデータをシームレスかつ効率的に転送するためのNetApp SnapMirror 、データ フローがバッチからリアルタイムに移行し、データ エンジニアリングが適切なタイミングでNetAppれる場合にリアルタイム レンダリングを実現するNetApp FlexCacheなどの

NetApp のデータ テクノロジーと機能は、リアルタイム生成 AI モデルの導入に価値をもたらします。あらゆるタイプの企業が新しい AI ツールを導入するにつれ、エッジからデータ センター、クラウドに至るまで、拡張可能で責任ある説明可能な AI ソリューションが求められるデータの課題に直面しています。ハイブリッドおよびマルチクラウドのデータ オーソリティとして、 NetAppは、生成 AI モデルのトレーニング (事前トレーニング)、微調整、コンテキストベースの推論、LLM のモデル減衰モニタリングのためのデータ パイプラインとデータ レイクの構築のあらゆる側面を支援するパートナー ネットワークと共同ソリューションの構築に取り組んでいます。

生成 **AI** とは何ですか**?**

生成 AI は、コンテンツの作成方法、新しいデザイン コンセプトの生成方法、斬新な構成の探索方法を変えています。これは、テキスト、コード、画像、オーディオ、ビデオ、合成データなどの新しいコンテンツを生成できる、Generative Adversarial Network (GAN)、Variational Autoencoders (VAE)、Generative Pre-Trained Transformers (GPT) などのニューラル ネットワーク フレームワークを示しています。 OpenAI の Chat-GPT、Google の Bard、Hugging Face の BLOOM、Meta の LLaMA などのトランスフォーマーベースのモデルは、大規模言語モデルの多くの進歩を支える基盤技術として登場しました。同様に、OpenAI の Dall-E 、Meta の CM3leon、Google の Imagen は、テキストから画像への拡散モデルの例であり、顧客に前例のないレベルのフォトリアリズムを提供して、新しい複雑な画像をゼロから作成したり、データセットの拡張と、テキストと視覚のセマンティクスをリンクするテキストから画像への合成を使用して既存の画像を編集して高品質のコンテキスト認識画像を生成したりします。デジタル アーティストは、NeRF (Neural Radiance Field) などのレンダリング テクノロジと生成 AI を組み合わせて、静的な 2D 画像を没入型の 3D シーンに変換し始めています。一般に、LLMは次の4つのパラメータによって特徴付けられます: (1) モデルのサイズ（通常は数十億のパラメータ単位）; (2) トレーニングデータセットのサイズ; (3) トレーニングのコスト、および (4) トレーニング後のモデルパフォーマンス。 LLM も主に 3 つのトランスフォーマー アーキテクチャに分類されます。 (i) エンコーダーのみのモデル。例：BERT（Google、2018）、（ii）エンコーダー・デコーダー例：BART（Meta、2020）、および（iii）デコーダーのみのモデル。例：LLaMA（Meta、2023）、PaLM-E（Google、2023）。ビジネス要件に応じて、企業がどのアーキテクチャを選択するかに関係なく、トレーニング データセット内のモデル パラメータの数 (N) とトークンの数 (D) によって、通常、トレーニング (事前トレーニング) または LLM の微調整のベースライン コストが決まります。

エンタープライズユースケースと下流の**NLP**タスク

さまざまな業界の企業は、ビジネス運営、販売、マーケティング、法務サービスのために、AI が既存のデータから新しい形の価値を抽出して生み出す可能性をますます発見しています。 IDC (International Data Corporation) の世界規模の生成 AI の使用事例と投資に関する市場情報によると、ソフトウェア開発と製品設計における知識管理が最も大きな影響を受け、次いでマーケティングのストーリーライン作成と開発者向けのコード生成が影響を受けます。ヘルスケア分野では、臨床研究組織が医学の新たな境地を開拓しています。ProteinBERT のような事前トレーニング済みモデルには、遺伝子オントロジー (GO) アノテーションが組み込まれており、医薬品のタンパク質構造を迅速に設計できます。これは、創薬、バイオインフォマティクス、分子生物学における重要なマイルストーンとなります。バイオテクノロジー企業は、肺組織に不可逆的な瘢痕化を引き起こす肺疾患である肺線維症（IPF）などの疾患の治療を目的とした、AIが発見した生成薬のヒト臨床試験を開始した。

図1: 生成AIを推進するユースケース

[図1: 生成AIを推進するユースケース]

生成 AI によって促進される自動化の導入の増加は、多くの職業における仕事の需要と供給にも変化をもたらしています。マッキンゼーによれば、米国の労働市場（下の図）は急速な変化を遂げており、AIの影響を考慮するとこの傾向は今後も続く可能性がある。

出典：マッキンゼー・アンド・カンパニー

[図2：出典：マッキンゼー・アンド・カンパニー]

生成AIにおけるストレージの役割

LLM は主にディープラーニング、GPU、コンピューティングに依存しています。ただし、GPU バッファがいっぱいになると、データをストレージにすぐに書き込む必要があります。一部の AI モデルはメモリ内で実行できるほど小さいですが、特に数十億のトークンや数百万の画像が含まれる場合、LLM では大規模なデータセットへの高速アクセスを提供するために高い IOPS と高スループットのストレージが必要です。 LLM の一般的な GPU メモリ要件では、10 億のパラメータを持つモデルをトレーニングするために必要なメモリは、32 ビットの完全精度で最大 80 GB になる可能性があります。その場合、70 億から 700 億のパラメータの規模に及ぶ LLM ファミリである Meta の LLaMA 2 では、70x80、約 5600 GB または 5.6 TB の GPU RAM が必要になる可能性があります。さらに、必要なメモリの量は、生成するトークンの最大数に正比例します。例えば、最大512トークン（約380語）の出力を生成したい場合は、"512 MB" 。重要ではないと思われるかもしれませんが、より大きなバッチを実行したい場合は、その分が加算されていきます。そのため、メモリ内で LLM をトレーニングまたは微調整する組織にとっては非常にコストがかかり、ストレージが生成 AI の基礎となります。

LLMへの3つの主なアプローチ

ほとんどの企業にとって、現在の傾向に基づくと、LLM を導入するアプローチは 3 つの基本シナリオに要約できます。最近の"ハーバード・ビジネス・レビュー"記事: (1) LLM をゼロからトレーニング (事前トレーニング) する – コストがかかり、専門的な AI/ML スキルが必要。(2) 企業データを使用して基礎モデルを微調整する – 複雑だが可能。(3) 検索拡張生成 (RAG) を使用して、企業データを含むドキュメントリポジトリ、API、ベクターデータベースを照会する。これらはそれぞれ、さまざまな種類の問題を解決するために使用される実装において、労力、反復速度、コスト効率、およびモデルの精度の間でトレードオフがあります (下の図)。

図3: 問題の種類

[図3: 問題の種類]

基礎モデル

基礎モデル (FM) はベースモデルとも呼ばれ、大規模な自己監督を使用して大量のラベルなしデータでトレーニングされた大規模な AI モデル (LLM) であり、通常は幅広い下流の NLP タスクに適応します。トレーニング データは人間によってラベル付けされていないため、モデルは明示的にエンコードされるのではなく、自動的に生成されます。これは、モデルが明示的にプログラムされなくても、独自のストーリーや物語を生成できることを意味します。したがって、FM の重要な特徴は均質化であり、これは多くの領域で同じ方法が使用されることを意味します。しかし、パーソナライゼーションと微調整のテクニックにより、最近登場した製品に統合された FM は、テキストの生成、テキストから画像への変換、テキストからコードへの変換が優れているだけでなく、ドメイン固有のタスクの説明やコードのデバッグにも役立ちます。たとえば、OpenAI の Codex や Meta の Code Llama などの FM は、プログラミング タスクの自然言語による記述に基づいて、複数のプログラミング言語でコードを生成できます。これらのモデルは、Python、C#、JavaScript、Perl 、Ruby、SQL など 12 を超えるプログラミング言語に精通しています。ユーザーの意図を理解し、ソフトウェア開発、コードの最適化、プログラミング タスクの自動化に役立つ、目的のタスクを実行する特定のコードを生成します。

微調整、ドメイン特異性、再トレーニング

データの準備とデータの前処理に続く LLM 展開における一般的な方法の 1 つは、大規模で多様なデータセットでトレーニングされた事前トレーニング済みモデルを選択することです。微調整の文脈では、これはオープンソースの大規模言語モデル、例えば"メタのラマ2"700億のパラメータと2兆個のトークンでトレーニングされました。事前トレーニング済みのモデルを選択したら、次のステップはドメイン固有のデータでそれを微調

整することです。これには、モデルのパラメータを調整し、特定のドメインとタスクに適応するために新しいデータでトレーニングすることが含まれます。たとえば、金融業界に役立つ幅広い金融データについてトレーニングされた独自の LLM である BloombergGPT などです。特定のタスク向けに設計およびトレーニングされたドメイン固有のモデルは、通常、その範囲内では高い精度とパフォーマンスを備えていますが、他のタスクまたはドメイン間での転送性は低くなります。ビジネス環境とデータが一定期間にわたって変化すると、テスト時のパフォーマンスと比較して、FM の予測精度が低下し始める可能性があります。このとき、モデルの再トレーニングまたは微調整が重要になります。従来の AI/ML におけるモデルの再トレーニングとは、展開された ML モデルを新しいデータで更新することを指し、通常は発生する 2 種類のドリフトを排除するために実行されます。 （1）概念のドリフト - 入力変数と目標変数の関係が時間の経過とともに変化すると、予測したい内容の記述が変化するため、モデルは不正確な予測を生成する可能性があります。 （2）データドリフト - 入力データの特性が変化したときに発生します。たとえば、顧客の習慣や行動が時間の経過とともに変化し、モデルがそのような変化に対応できなくなります。同様に、再トレーニングは FM/LLM にも適用されますが、コストが非常に高くなる可能性がある (数百万ドル単位) ため、ほとんどの組織では検討できない可能性があります。これは現在活発に研究されており、LLMOps の分野ではまだ発展途上です。そのため、再トレーニングの代わりに、微調整された FM でモデルの減衰が発生した場合、企業は新しいデータセットを使用して再度微調整すること (はるかに安価) を選択できます。コストの観点から、Azure-OpenAI Services のモデル価格表の例を以下に示します。各タスク カテゴリごとに、顧客は特定のデータセットでモデルを微調整および評価できます。

出典: Microsoft Azure

[出典: Microsoft Azure]

迅速なエンジニアリングと推論

プロンプト エンジニアリングとは、モデルの重みを更新せずに LLM と通信して目的のタスクを実行するための効果的な方法を指します。 AI モデルのトレーニングと微調整が NLP アプリケーションにとって重要であるのと同様に、トレーニングされたモデルがユーザーのプロンプトに応答する推論も同様に重要です。推論のシステム要件は、一般的に、LLM から GPU にデータを供給する AI ストレージ システムの読み取りパフォーマンスに大きく依存します。これは、最適な応答を生成するために、保存されている数十億のモデル パラメータを適用する必要があるためです。

LLMOps、モデル監視、Vectorstores

従来の機械学習オペレーション (MLOps) と同様に、大規模言語モデルオペレーション (LLMOps) でも、運用環境での LLM の管理のためのツールとベスト プラクティスを備えたデータ サイエンティストと DevOps エンジニアのコラボレーションが必要です。ただし、LLM のワークフローと技術スタックはいくつかの点で異なる可能性があります。たとえば、LangChain などのフレームワークを使用して構築された LLM パイプラインは、ベクターストアやベクター データベースなどの外部埋め込みエンドポイントへの複数の LLM API 呼び出しを連結します。ダウンストリーム コネクタ (ベクター データベースなど) に埋め込みエンドポイントとベクターストアを使用することで、データの保存方法とアクセス方法が大きく進歩しました。ゼロから開発される従来の ML モデルとは対照的に、LLM は、より具体的なドメインでのパフォーマンスを向上させるために新しいデータで微調整された FM から始まるため、転移学習に依存することがよくあります。したがって、LLMOps がリスク管理とモデル減衰監視の機能を提供することが重要です。

生成AI時代のリスクと倫理

「ChatGPT – 洗練されているが、それでも意味不明な情報を吐き出す。」– MIT Tech Review。ゴミを入れればゴミしか出てこないというのは、コンピューティングにおいて常に難しい問題です。生成型 AI との唯一の違いは、ゴミを非常に信頼できるものにして、不正確な結果をもたらすことに優れていることです。 LLM は、構築している物語に合うように事実を捏造する傾向があります。したがって、生成 AI を AI と同等のコストを削減する絶好の機会と捉えている企業は、ディープフェイクを効率的に検出し、バイアスを減らし、リスクを低減して、システムの誠実性と倫理性を維持する必要があります。エンドツーエンドの暗号化と AI ガー

ドレールによるデータ モビリティ、データ品質、データ ガバナンス、データ保護をサポートする堅牢な AI インフラストラクチャを備えた自由に流れるデータ パイプラインは、責任ある説明可能な生成 AI モデルの設計において重要です。

顧客シナリオと**NetApp**

図3: 機械学習/大規模言語モデルのワークフロー

[図3: 機械学習/大規模言語モデルのワークフロー]

トレーニングですか、それとも微調整ですか**?** (a) LLM モデルを最初からトレーニングするか、事前トレーニング済みの FM を微調整するか、RAG を使用して基盤モデル外のドキュメント リポジトリからデータを取得してプロンプトを拡張するか、(b) オープンソースの LLM (例: Llama 2) または独自の FM (例: ChatGPT 、Bard、AWS Bedrock) を活用するかという問題は、組織にとって戦略的な決定です。それぞれのアプローチには、コスト効率、データ重力、運用、モデルの精度、LLM の管理の間でトレードオフがあります。

NetAppは企業として、社内の業務文化や製品の設計およびエンジニアリングへの取り組みに AI を取り入れています。たとえば、NetApp の自律型ランサムウェア保護は、AI と機械学習を使用して構築されています。ファイル システムの異常を早期に検出し、脅威が運用に影響する前に特定できるようにします。 2 番目に、NetApp は売上や在庫の予測などのビジネス オペレーションに予測 AI を使用し、チャットボットを使用してコール センターの製品サポート サービス、技術仕様、保証、サービス マニュアルなどで顧客を支援します。 3 番目に、 NetApp は、NetApp ONTAP AI、NetApp SnapMirror、NetApp FlexCache などの NetApp 製品と機能を使用して、需要予測、医療用画像処理、感情分析などの予測 AI ソリューションや、製造業における産業用画像異常検出用の GAN や銀行および金融サービスにおけるマネー ロンダリング防止および詐欺検出などの生成 AI ソリューションを構築NetAppにサービスを提供する製品とソリューションを通じて、AINetAppSnapMirrorONTAPとNetApp /NetAppFlexCacheに顧客価値をもたらします。

NetAppの機能

チャットボット、コード生成、画像生成、ゲノムモデル表現などの生成 AI アプリケーションにおけるデータの移動と管理は、エッジ、プライベート データ センター、ハイブリッド マルチクラウド エコシステムにまたがって行われます。たとえば、ChatGPT などの事前トレーニング済みモデルの API を介して公開されるエンドユーザー アプリから、乗客が航空券をビジネス クラスにアップグレードするのを支援するリアルタイム AI ボットは、乗客情報がインターネット上で公開されていないため、単独ではそのタスクを達成できません。 API は、ハイブリッドまたはマルチクラウド エコシステム内に存在する可能性のある航空会社からの乗客の個人情報とチケット情報にアクセスする必要があります。同様のシナリオは、1 対多の生物医学研究機関が関与する創薬全体にわたって臨床試験を実行するために LLM を使用するエンドユーザー アプリケーションを介して薬剤分子と患者データを共有する科学者にも当てはまる可能性があります。 FM または LLM に渡される機密データには、PII、財務情報、健康情報、生体認証データ、位置データ、通信データ、オンライン行動、法的情報などが含まれる場合があります。このようなリアルタイム レンダリング、プロンプト実行、エッジ推論のイベントでは、エンド ユーザー アプリからオープン ソースまたは独自の LLM モデルを経由してストレージ エンドポイントにデータが移動され、さらにオンプレミスまたはパブリック クラウド プラットフォーム上のデータ センターにデータが移動されます。このようなすべてのシナリオにおいて、大規模なトレーニング データセットとそのようなデータの移動に依存する LLM を伴う AI 操作にとって、データのモビリティとデータ保護は非常に重要です。

図4: 生成AI - LLMデータパイプライン

[図4: 生成AI-LLMデータパイプライン]

NetApp のストレージ インフラストラクチャ、データ、クラウド サービスのポートフォリオは、インテリジェントなデータ管理ソフトウェアによって強化されています。

データ準備: LLM 技術スタックの最初の柱は、従来の ML スタックからほとんど変更されていません。 AI パイプラインでのデータ前処理は、トレーニングや微調整の前にデータを正規化およびクレンジングするために必要です。このステップには、Amazon S3 層の形式、またはファイル ストアやNetApp StorageGRIDなどのオブジェクト ストアなどのオンプレミスのストレージ システムの形式で存在する場所に関係なくデータを取り込むためのコネクタが含まれます。

- NetApp ONTAP* は、データセンターとクラウドにおける NetApp の重要なストレージ ソリューションを支える基盤テクノロジーです。 ONTAPには、サイバー攻撃に対する自動ランサムウェア保護、組み込みデータ転送機能、オンプレミス、ハイブリッド、NAS、SAN、オブジェクト、ソフトウェア定義ストレージ (SDS) の LLM 展開におけるマルチクラウドのさまざまなアーキテクチャに対応するストレージ効率機能など、さまざまなデータ管理および保護機能が含まれています。
- ディープラーニング モデルのトレーニング用NetApp ONTAP AI*。 NetApp ONTAP は、 ONTAPストレージ クラスタとNVIDIA DGX コンピューティング ノードを使用するNetApp顧客向けに、NFS over RDMA を使用したNVIDIA GPU Direct Storage をサポートします。コスト効率の高いパフォーマンスを提供し、ソース データセットをストレージからメモリに何度も読み取って処理することでインテリジェンスを促進し、組織が LLM へのアクセスをトレーニング、微調整、スケーリングできるようにします。
- NetApp FlexCache* は、ファイルの配布を簡素化し、アクティブに読み取られたデータのみをキャッシュするリモート キャッシュ機能です。これは、LLM のトレーニング、再トレーニング、微調整に役立ち、リアルタイム レンダリングや LLM 推論などのビジネス要件を持つ顧客に価値をもたらします。
- NetApp SnapMirror * は、任意の 2 つのONTAPシステム間でボリューム スナップショットを複製するONTAP機能です。この機能は、エッジのデータをオンプレミスのデータセンターまたはクラウドに最適に転送します。 SnapMirror は、顧客がエンタープライズ データを含む RAG を使用してクラウドで生成 AI を開発する場合に、オンプレミス クラウドとハイパースケーラー クラウド間でデータを安全かつ効率的に移動するために使用できます。変更のみを効率的に転送することで、帯域幅を節約し、レプリケーションを高速化します。これにより、FM または LLM のトレーニング、再トレーニング、微調整の操作中に重要なデータ移動機能が提供されます。
- NetApp SnapLock* は、データセットのバージョン管理のために、 ONTAPベースのストレージ システムに不変のディスク機能をもたらします。マイクロコア アーキテクチャは、FPolicy Zero Trust エンジンを使用して顧客データを保護するように設計されています。 NetApp は、攻撃者が特にリソースを消費する方法で LLM と対話する場合に、サービス拒否 (DoS) 攻撃に抵抗することで、顧客データの可用性を確保します。
- NetApp Cloud Data Sense* は、企業のデータセット内に存在する個人情報を識別、マッピング、分類し、ポリシーを制定し、オンプレミスまたはクラウドでのプライバシー要件を満たし、セキュリティ体制を改善して規制に準拠するのに役立ちます。
- NetApp BlueXP* Cloud Data Sense を活用した分類。顧客は、データ資産全体のデータを自動的にスキャン、分析、分類、操作し、セキュリティ リスクを検出し、ストレージを最適化し、クラウドの導入を加速できます。統合コントロール プレーンを介してストレージ サービスとデータ サービスを組み合わせ、顧客は計算に GPU インスタンスを使用し、コールド ストレージ階層化とアーカイブおよびバックアップにハイブリッド マルチクラウド環境を使用できます。
- NetAppファイルとオブジェクトの二重性*。 NetApp ONTAP は、NFS と S3 のデュアルプロトコル アクセスを可能にします。このソリューションにより、顧客はNetApp Cloud Volumes ONTAPの S3 バケットを介して Amazon AWS SageMaker ノートブックの NFS データにアクセスできるようになります。これにより、NFS と S3 の両方からデータを共有し、異種データ ソースに簡単にアクセスする必要がある顧客に柔軟性が提供されます。たとえば、ファイル オブジェクト バケットにアクセスして、SageMaker 上の Meta の Llama 2 テキスト生成モデルなどの FM を微調整します。
- NetApp Cloud Sync* サービスは、クラウドまたはオンプレミスの任意のターゲットにデータを移行するためのシンプルで安全な方法を提供します。 Cloud Sync は、オンプレミスまたはクラウド ストレージ、NAS、オブジェクト ストア間でデータをシームレスに転送および同期します。
- NetApp XCP* は、あらゆるデバイスからNetAppへ、またNetAppからNetAppへのデータ移行を高速かつ確

実に実行できるクライアント ソフトウェアです。 XCP は、Hadoop HDFS ファイル システムからONTAP NFS、S3、またはStorageGRIDに大量のデータを効率的に移動する機能も提供し、XCP ファイル分析はファイル システムの可視性を提供します。

- NetApp DataOps Toolkit* は、データ サイエンティスト、DevOps、データ エンジニアが、高性能スケールアウト NetApp ストレージを基盤とするデータ ボリュームや JupyterLab ワークスペースのプロビジョニング、クローン作成、スナップショット作成などのさまざまなデータ管理タスクをほぼ瞬時に実行できるようにするNetAppライブラリです。

NetApp の製品セキュリティ。 LLM は応答の中で機密データを誤って公開する可能性があるため、LLM を活用した AI アプリケーションに関連する脆弱性を研究する CISO にとっては懸念事項となります。 OWASP (Open Worldwide Application Security Project) で概説されているように、LLM 内のデータ ポイズニング、データ漏洩、サービス拒否、プロンプト インジェクションなどのセキュリティ問題は、データの露出から攻撃者による不正アクセスまで、企業に影響を及ぼす可能性があります。データ ストレージ要件には、構造化データ、半構造化データ、非構造化データの整合性チェックと不変のスナップショットが含まれる必要があります。データセットのバージョン管理には、 NetAppスナップショットおよびSnapLock が使用されています。厳格なロールベースのアクセス制御 (RBAC)、安全なプロトコル、業界標準の暗号化を導入し、保存中のデータと転送中のデータの両方を保護できます。 Cloud Insightsと Cloud Data Sense を組み合わせることで、脅威の原因をフォレンジック的に特定し、復元するデータの優先順位を決定するのに役立つ機能が提供されます。

* DGX BasePOD を搭載したONTAP AI*

NVIDIA DGX BasePOD を搭載したNetApp ONTAP AI リファレンス アーキテクチャは、機械学習 (ML) および人工知能 (AI) ワークロード向けのスケーラブルなアーキテクチャです。 LLM の重要なトレーニング フェーズでは、通常、データはデータ ストレージからトレーニング クラスターに定期的にコピーされます。このフェーズで使用されるサーバーは、GPU を使用して計算を並列化し、膨大な量のデータを必要とします。生の I/O 帯域幅のニーズを満たすことは、高い GPU 使用率を維持するために非常に重要です。

* NVIDIA AI Enterprise を搭載したONTAP AI*

NVIDIA AI Enterprise は、 NVIDIA認定システムを搭載した VMware vSphere で実行できるようにNVIDIAによって最適化、認定、サポートされている、エンドツーエンドのクラウド ネイティブな AI およびデータ分析ソフトウェア スイートです。このソフトウェアは、最新のハイブリッド クラウド環境での AI ワークロードのシンプルかつ迅速な導入、管理、スケーリングを可能にします。 NetAppと VMware を搭載したNVIDIA AI Enterprise は、シンプルで使い慣れたパッケージでエンタープライズ クラスの AI ワークロードとデータ管理を提供します。

1Pクラウドプラットフォーム

完全に管理されたクラウド ストレージ オファリングは、Microsoft Azure ではAzure NetApp Files (ANF) として、AWS ではAmazon FSx for NetApp ONTAP (FSx ONTAP) として、Google ではGoogle Cloud NetApp Volumes (GNCV) としてネイティブに利用できます。 1P は、パブリッククラウドでデータセキュリティを強化し、可用性の高い AI ワークロードを実行できるようにする、マネージド型の高性能ファイルシステムです。AWS SageMaker、Azure-OpenAI サービス、Google の Vertex AI などのクラウドネイティブ ML プラットフォームを使用して LLM/FM を微調整できます。

NetAppパートナー ソリューション スイート

NetApp は、コアとなるデータ製品、テクノロジー、機能に加え、強力な AI パートナー ネットワークとも緊密に連携し、顧客に付加価値を提供しています。

- AI システムのNVIDIA Guardrails* は、AI テクノロジーの倫理的かつ責任ある使用を保証するための安全策として機能します。 AI 開発者は、特定のトピックに関する LLM 対応アプリケーションの動作を定義し、不要なトピックに関するディスカッションに参加できないようにすることができます。オープンソース ツールキットである Guardrails は、信頼性が高く、安全で、セキュリティ保護された LLM 会話システムを構築するために、LLM を他のサービスにシームレスかつ安全に接続する機能を提供します。

Domino Data Lab は、AI 導入のどの段階であっても、高速、安全、かつ経済的な Generative AI の構築と製品化のための多用途のエンタープライズ グレードのツールを提供します。 Domino のエンタープライズ MLOps プラットフォームを使用すると、データ サイエンティストは好みのツールとすべてのデータを使用し、どこからでも簡単にモデルをトレーニングおよび展開し、リスクとコストを効果的に管理できます。これらはすべて 1 つのコントロール センターから実行できます。

エッジ **AI** 向け **Modzy**。 NetAppと Modzy は提携して、画像、音声、テキスト、表など、あらゆる種類のデータに大規模な AI を提供しています。 Modzy は、AI モデルを展開、統合、実行するための MLOps プラットフォームであり、データ サイエンティストにモデル監視、ドリフト検出、説明可能性の機能を提供し、シームレスな LLM 推論のための統合ソリューションを提供します。

Run:AI とNetAppは提携して、AI ワークロードのオーケストレーションを簡素化する Run:AI クラスタ管理プラットフォームを使用したNetApp ONTAP AI ソリューションの独自の機能を実証しました。 GPU リソースを自動的に分割および結合し、Spark、Ray、Dask、Rapids の統合フレームワークが組み込まれているため、データ処理パイプラインを数百台のマシンに拡張できるように設計されています。

まとめ

生成 AI は、モデルが大量の高品質データでトレーニングされた場合にのみ効果的な結果を生成できます。 LLM は目覚ましい成果を達成してきましたが、その限界、設計上の課題、およびデータのモビリティとデータ品質に関連するリスクを認識することが重要です。 LLM は、異種のデータ ソースからの大規模かつ多様なトレーニング データセットに依存します。モデルによって生成された不正確な結果や偏った結果は、企業と消費者の両方を危険にさらす可能性があります。これらのリスクは、データ品質、データ セキュリティ、およびデータ モビリティに関連するデータ管理の課題から潜在的に生じる LLM の制約に対応する可能性があります。 NetApp は、急速なデータ増加、データ モビリティ、マルチクラウド管理、AI の導入によって生じる複雑さに組織が対応できるよう支援します。大規模な AI インフラストラクチャと効率的なデータ管理は、生成 AI などの AI アプリケーションの成功を定義する上で非常に重要です。コスト効率、データ ガバナンス、倫理的な AI プラクティスを管理しながら、企業のニーズに応じて拡張する能力を損なうことなく、すべての展開シナリオをカバーすることが、お客様にとって重要です。 NetApp は、お客様の AI 導入を簡素化し、加速させるために常に取り組んでいます。

著作権に関する情報

商標に関する情報