推論用 AI POD

概要

世界中のあらゆる業界の企業が、AI を活用してビジネスを変革し、顧客満足度を高め、競争上の優位性を獲得しようとしています。生成 AI アプリケーションの導入は複雑なプロセスであり、慎重な計画、モデルとインフラストラクチャの評価、実行を必要とします。これは成功するチャンスであると同時に、失敗する可能性も秘めています。

多くの組織は、高コストで複雑なポイント AI ソリューションのリスクに対処しながら、AI プロジェクトへの投資を成功に導く戦略を策定するのに苦労しています。インフラストラクチャのニーズは、AI モデルのタイプとサイズによって大きく異なる可能性があります。Cisco® は、現在のビジネスのニーズと IT のニーズを両立させながら、将来の拡張性も視野に入れ、AI 関連のインフラストラクチャへの投資を適切な規模に調整するよう支援します。

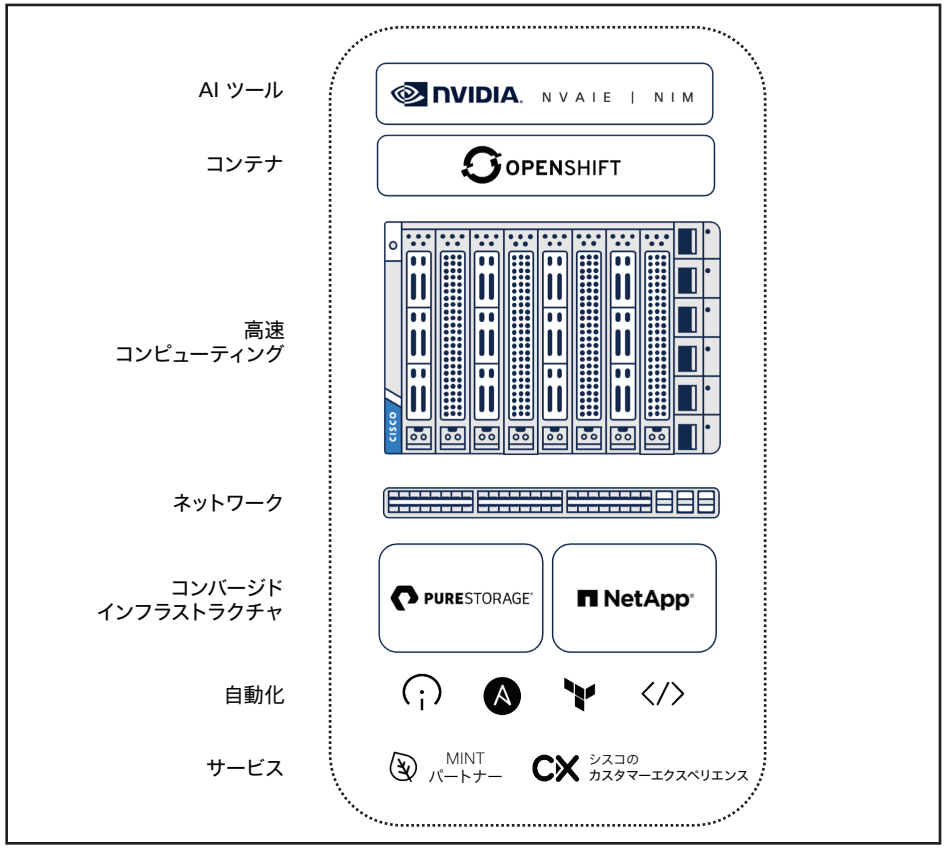

図 1.　推論用 AI POD ソリューション

AI 推論とは

AI 推論では、事前に訓練されたモデル (GPT-4、Claude 3、Llama 3 など) を使用して新しいデータを分析し、そのデータに基づいて推論や最も可能性の高い結果を生成します。このプロセスは、チャットボット、コーディング支援、画像認識などのアプリケーションで広く使用されています。従来の AI モデルは、一般的な知識に関する質問には効果的ですが、トレーニングに含まれていない特定のデータ (企業独自のデータなど) を必要とするクエリには適さない可能性があります。

そこで登場するのが、検索拡張生成(RAG) です。RAG は、元のモデルのトレーニングに使用されなかった外部データソースを組み込むことで、AI 推論の精度と関連性を高めます。RAG はモデルをドメイン固有のデータに結び付け、より正確で関連性の高い出力を生成できるようにします。たとえば、ある国の人口データに基づいてトレーニングされた保険モデルを考えてみましょう。顧客固有のデータを追加することで、このモデルはより正確でビジネスに関連したインサイトを提供できます。

メリット

- パフォーマンス アシュアランスとシームレスな拡張性を備えた AI 対応インフラストラクチャを自信を持って導入し、システムが高度な AI ワークロードに対応できるようにします。
- インフラストラクチャ、ソフトウェア、AI ツールセットのフルスタック検証を活用することで、AI モデルの導入を加速し、推論が実稼働に対応するまでの時間を短縮します。
- AI 導入アーキテクチャをクラス最高レベルの単一サポートで運用して、オペレーションを合理化し、AI システム全体の信頼性を高めます。

詳細情報

- シスコの推論用 AI POD の詳細については、データシートを参照してください。
- シスコのデータセンター向け AI ネイティブ インフラストラクチャの詳細については、Cisco.com にアクセスしてください。
- Cisco UCS X シリーズ モジュラーシステムの詳細については、https://www.cisco.com/jp/go/ucsx を参照してください。

エキスパートによるコンサルティングを予約して、AI 対応インフラストラクチャの導入に着手しましょう

AI 対応インフラストラクチャによるネットワークおよびコンピューティング インフラストラクチャの近代化に関するエキスパートによるガイダンスでは、テクノロジー、製品、Cisco Validated Design を組み合わせ、サステナビリティの実現に向けた取り組みを推進しつつ、AI ワークロードをサポートおよび拡張する方法をご紹介します。

ご予約はこちら

機能

シスコは 20 年以上にわたって検証済みデザインを開発し、提供してきました。Cisco Validated Design(CVD)は、IT インフラストラクチャの効果的な導入と管理に役立つ、包括的で厳密にテストされたガイドラインです。詳細な実装ガイド、ベストプラクティス、実際の使用例が含まれており、多くの場合、シスコのテクノロジーパートナー製品が組み込まれています。CVD は、Cisco Technical Assistance Center (TAC) のサポートを受けながら、導入リスクの軽減、パフォーマンスの最適化、拡張性の確保を実現します。このサポートと統合により、お客様はビジネス目標を達成するための信頼できる効率的な手段を得ることができます。

シスコの推論用 AI POD は、エッジ推論、RAG、大規模推論向けの CVD ベースのソリューションです。一元管理と自動化により、迅速な導入を実現します。このソリューションは、パフォーマンステストが実施されています。また、実際のモデルシミュレーションのベンチマークテストで直線的な拡張性が実証されており、データセットのサイズが変化しても一貫したパフォーマンスが得られます。シスコの推論用 AI POD は、インフラストラクチャの各レイヤで独立した拡張性を備え、DC またはエッジ AI の導入に最適です。POD の CPU と GPU の数が異なる 4 つの構成があります。

構成に関係なく、以下のものがすべて含まれます。

- Cisco UCS X シリーズ モジュラーシステム
  - Cisco UCS X9508 シャーシ
  - Cisco UCS X シリーズ M7 コンピューティング ノード
  - Cisco UCS X440p PCIe ノード (Nvidia GPU 搭載)
  - Cisco UCS 9108 インテリジェント ファブリック モジュール
  - Cisco UCS 6536 ファブリック インターコネクトまたは Cisco UCS ファブリック インターコネクト 9108 100G
  - Cisco UCS X9416 X-Fabric モジュール
- Cisco Intersight®
- シスコ サービス
- Nvidia NVAIE サブスクリプション
- NVIDIA HPC-X ソフトウェアツールキット
- RedHat OpenShift ライセンス

オプションのストレージは、NetApp (FlexPod) および Pure Storage (FlashStack) からも入手できます。どちらも、開発者やデータサイエンティストがさまざまなデータ管理タスクを実行できるように DataOps ツールキットを提供しています。

C45-4661663-01 11/24