2025年4月 | eBook

Enterprise AI on Nutanix

最も素早いのは、一番シンプルな道筋



NUTANIX

目次

エンタープライズ AI を迅速に導入 3

AI に関する問題点 4

AI 向けに Nutanix Cloud Platform を選ぶ理由 6

Nutanix GPT-in-a-Box 7

業界パートナーシップ 8

AI のユースケース 9

Nutanix を使い始める 10



エンタープライズ AI への道のりを迅速化

生成 AI の台頭により、組織は自社の AI 計画を再考し、その予定を加速するようになりました。

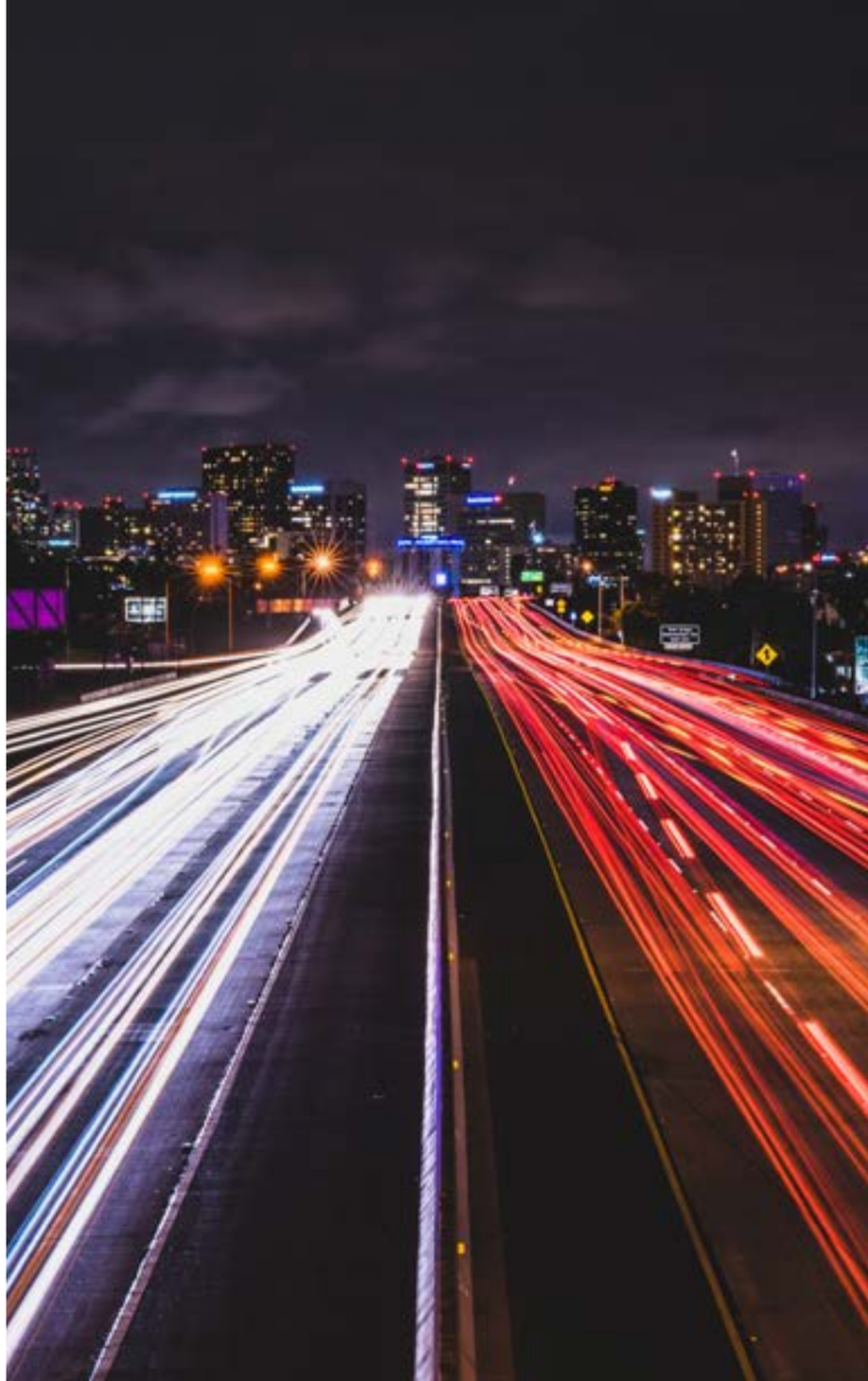
しかし、650 名の IT 意思決定者、DevOps およびプラットフォームエンジニアリング専門家を対象とした調査では、
多くの組織が生成 AI にどう取り組めばよいのか分からずにいる点が明らかになりました。主な懸念事項は次のとおりです。

- **IT インフラ不足:** 回答者の91% は、AI ワークロードをサポートするには IT インフラの改善が必要だと回答しています。
- **データセキュリティおよび回復力の確保:** データセキュリティ、回復力、スケーラビリティは、エンタープライズ AI が抱える主な課題となっています。
- **スキル不足:** すべての組織が、今後 12 ヶ月間にわたる AI 関連のイニシアチブを支援するにあたって追加のスキルが必要になると回答しています。

AI をデプロイする前に、次の 3 つの一般的なアプローチを検討してみましょう:

- **クラウドのみに依存する。**AI の実験および AI アプリの開発・本番環境にクラウドを利用することは選択肢の一つですが、費用は高額になりがちです。クラウドリソースの獲得をめぐって他の企業と競合することになり、自社データをリスクに晒すことになりかねません。
- **データセンターとエッジにカスタム AI スタックを構築する。**このアプローチには、膨大な専門知識と時間を要するため、社内専門家または専任のパートナーがいる場合にのみ検討に値します。それ以外の場合では、AI の進捗状況は遅々として進まず、失敗も頻繁に起きる可能性が高いでしょう。
- **ターンキーソリューションを選択する。**ターンキーソリューションでは、経験の浅いチームでも、柔軟性を高めるクラス最高のハードウェアと AI ソフトウェアをすぐに使用できるようになります。

この eBook では、チームが AI に備えて計画する過程で直面する問題点を検証し、Nutanix Cloud Platform と Nutanix GPT-in-a-Box がオンプレミス、クラウド、エッジでの AI の本番環境への道筋をシンプルにして、より短時間でより良い結果を生み出す方法について解説します。



AI に関する問題点

大多数の組織 **(85%)** は、AI アプリを開発するために既存の AI モデルを購入するか、オープンソースの AI モデルの活用を検討しており、自社によるモデル開発を計画しているのはわずか **10%** でした。

多くの企業にとって、AI モデルを一から設計して、トレーニングする必要は一切ありません。次のいずれかで、事前トレーニング済みの基礎モデルをプライベートデータでライセンスし、拡張するのがより容易になります：

- **ファインチューニング:** 自社のニーズに合わせて、プライベートデータを利用してモデルをファインチューニングします。
- **RAG (検索拡張生成):** モデル外部のデータを使用して、自社に合わせてカスタマイズされた結果を提供できます。

どちらの方法でも、AI の実験から本番環境へのデプロイへの移行過程では、以下の領域で重大な課題に直面することになります。

複雑さ

生成 AI を活用するには、コンテナ化された大規模言語モデル (LLM) をデプロイし、これらのファインチューニングまたは RAG の実装、そしてMLOps を利用した本番環境への度重なるデプロイを行える機能が必要です。

これには、ファインチューニングと推論に適したインフラが必要になる場合があります。推論は、レイテンシを減らして応答性を高めるために、AI サービスの消費者に近いエッジで実行されることがよくあります。

- エッジでの推論により、リモートインフラ管理の面で課題をもたらします。
- データセンターまたはエッジでも、推論とトレーニングにはコンピューティング、GPU、そしてストレージリソースの適切な組み合わせが必要であり、これにより需要の変化に合わせて迅速に拡張しながら、過剰プロビジョニングを回避できます。

AI 運用を万全にするには、発生する問題の根本原因が果たしてモデル、インフラ、または MLOps プロセスの脆弱性によるものなのか判断できるインテリジェントツールが必要です。

コンプライアンス

AI チームは IT チームとは別に編成される事が多く、ミッションクリティカルな回復力、データ管理、そして個人識別情報 (PII) の保護を適切に考慮せずにソリューションの開発に重点を置くことがよくあります。

生成 AI やその他の AI ベースのアプリケーションを、セキュリティ、データ保護、回復力、その他の Day 2 オペレーションのニーズに関する IT ポリシーへのコンプライアンスを損なうことなくデプロイして運用するのは困難な場合があります。

すべてのトレーニングデータは、氏名、住所、電話番号、社会保障番号、金融機関情報、クレジットカード番号などの PII が一切含まれることのないようスクラブして確実に消去する必要があります。

コスト

クラウドは AI 運用にいくつかの利点をもたらしますが、オンプレミスで AI モデルを実行する場合と比べて、クラウドで AI モデルを実行する場合のコストは大幅に高くなる可能性があります。

その一方で、AI インフラは大量の電力を消費することで有名です。データセンターまたはエッジで AI ワークロードを実行する場合は、電力、冷却、物理的なスペース、その他の Day 2 の考察事項を考慮する必要があります。

ガバナンス

AI には、複数のガバナンス面での課題が存在します。

- **データ管理とデータ主権:** エッジ、コア、クラウドにまたがる一貫したポリシーベースのデータ保護とセキュリティは、簡単には確保できません。エッジで生成されたデータは、さらなるトレーニングを必要とするケースがよくあります。別々の場所にデータを移行して管理するには、シンプルで反復可能かつ自動化された手段が欠かせません。

  時には、データ主権の規制違反に該当することのないよう、エッジで生成されたデータを確実な場所に保管・保護する必要もあります。このデータを利用してトレーニングをする必要がある場合、モデルをデータのある場所に移動させてトレーニングできるようにプロビジョニングする必要があるかもしれません。

  こうした場所が複数存在する場合、モデルを各場所で連続してトレーニングする、連邦型アプローチが必要となる可能性があります。

- **MLOps:** AI チームは、各モデルバージョンに関する完全な情報を正確に追跡・維持できなくてはいけません。これには、トレーニングに使用した厳密なデータ、そしてモデルをデプロイした時期と場所が含まれます。

  これは、トレーニングのイテレーションの頻度、そしてデータセットの数と複雑さに応じて、従来のソフトウェア開発におけるコードのリビジョンの追跡よりも著しく複雑になります。

セキュリティとデータプライバシー

AI では、セキュリティ面でのさらなる懸念が伴います。

- **データ漏洩:** 機密または専有情報を含んだ企業情報を、クラウド上で稼働する非主権型 AI サービスに供給したいとう強い衝動に駆られることがあるかもしれません。

  これには、未公開レポートの文面を Copilot、Gemini、または ChatGPT に供給して要約の作成を頼んだり、OpenAI の Whisper モデルを利用して社内会議の議事録の文字起こしを作成するなど、一見無害な行為が挙げられます。

  しかし、モデルが使用した後のデータはどうなるのでしょうか？Microsoft、Google、AWS、OpenAI は信頼できるかもしれませんが、生成アプリおよびサービスの膨大なエコシステムの台頭はリスクを高めます。

- **データポイズニング:** LLM への悪意のあるデータの注入は、仮にそのモデルを自社でコントロールしている場合でも、ハルシネーションやバイアスと呼ばれる不正確な生成コンテンツの発生につながりかねません。

- **プロンプトインジェクション:** AI のガードレールを回避するために悪意のあるプロンプトを直接挿入して、LLM の反応を乗っ取ります。

- **セキュリティツール:** 現在のセキュリティツールでは MLOps とアライメントを取ることができず、アップグレードまたは完全に新しいツールが必要となるケースです。

Nutanix Cloud Platform は、革新的なソフトウェアと独自のハイパーコンバージドインフラストラクチャー (HCI) を駆使して、こうした課題にオンプレミス、エッジ、パブリッククラウドで対処できます。





AI 向けに Nutanix Cloud Platform を選ぶ理由

Nutanixは、インフラストラクチャーの複雑さを軽減し、ハイブリッド・マルチクラウド運用を可能にすることに、全力を投じています。当社の実績のある HCI をベースに構築された Nutanix Cloud Platform (NCP) は、エッジからコアデータセンター、さらにはパブリッククラウドに至るあらゆる場所で、AI ニーズを満たす、俊敏で回復力に優れたインフラを提供します。

AI 運用をシンプルにする

運用の複雑さも、AI の本番環境をデプロイする道のりで大きな課題の一つとなります。リソース面での制約から、多くの組織は AI の実験をクラウドで行っています。

しかし、クラウドで行われた作業をデータセンターに移行したり、エッジで運用化するには、どうしたらよいのでしょうか?

NCP はインフラの複雑さと制限を解消し、AI/ML イニシアティブを迅速に開始できます。Nutanix では、エッジ、データセンター、クラウド環境にまたがるアプリケーション、ワークロード、データのシームレスなモビリティを実現できます。

NCP はあらゆる場所で動作するため、チームはあらゆる環境で効率的に作業を進めることができます。

エッジでの NCP

エッジの場合、複雑なインフラをリモートにデプロイ・管理する必要があるため、特に難点となっています。コンピューティング、ネットワーク、ストレージを組み合わせた、実証済みのHCI 設計を備えた NCP は、以下の独自の方法でエッジデプロイメントのニーズを満たします。

- コンパクトなフットプリント
- 簡単なリモート管理とリモートアプリデプロイメント
- 高度なデータ保護およびセキュリティ
- 制限のない運用
- データサービスのフルスイート

Nutanix GPT-in-a-box

多くの企業は生成 AI の活用に苦労しています。特に当てはまるのが、データ主権、ガバナンス、プライバシー上の懸念からパブリッククラウドで実行できないユースケースです。

Nutanix GPT-in-a-Box は、完全なコントロールを保ちながら、LLM やその他の AI モデルのファインチューニングおよび実行を可能にする、すぐに利用可能なターンキープライベート AI を提供します。

Nutanix GPT-in-a-Box は、生成 AI 開発で直面する複雑さ、拡張、およびセキュリティ上の課題に対処します。フルスタックでソフトウェアデファインドかつ AI 対応の GPT-in-a-Box は、NCP 上で実行することでデプロイメントをシンプルにし、AI の取り組みを迅速化します。

GPT-in-a-Box には以下のメリットがあります。

- **フルスタックの AI プラットフォーム。**生成 AI をデプロイしたあらゆる場所で、希望するハードウェア、CPU や GPU、VM やコンテナ、LLM や AI フレームワークを利用できます。
- **AI 向けのビルトインデータサービス。**ファイル、ブロック、オブジェクト向けの安全かつ拡張性の高いデータサービスと、統合されたスナップショットおよびディザスタリカバリコントロールを提供します。
- **スケーラブルな AI ワークロード。**クラウド運用を統合する一貫性を備えたプラットフォームサービスによって、より少ない労力でスケーラブルな AI をあらゆる場所に提供できます。
- **すぐに使える LLM でデプロイして適応:** エッジからクラウドまで素早くデプロイして、価値創出までの時間を加速させる、検証済みの生成 AI モデルを利用できます。
- **AI アプリ向けにセキュアな API をデプロイして運用:** セキュアなロールベースの API を簡単に作成して、アプリを AI モデルに接続します。

Nutanix GPT-in-a-Box では、AI トレーニングのデータおよびモデルを社内で保管することで、セキュリティ、プライバシー、コンプライアンスの要件を満たしながら、IT コストの最適化を実現できます。

GPT-in-a-Box には、生成 AI モデルの実行に取りかかる上で必要なものがすべて揃っています。あとは、希望する基盤モデルを提供するだけです。

GPU および CPU の柔軟なオプション

推論、トレーニング、または両方を行う場合でも、適切な GPU と CPU は AI で成功を収めるために不可欠な要素です。

GPU

Nutanix は、多様なニーズに応えられるよう NVIDIA GPU のフルレンジに対応しています。NCP は GPU パススルーに対応しているため、仮想化またはコンテナ化環境でも GPU を効率的に利用できます。

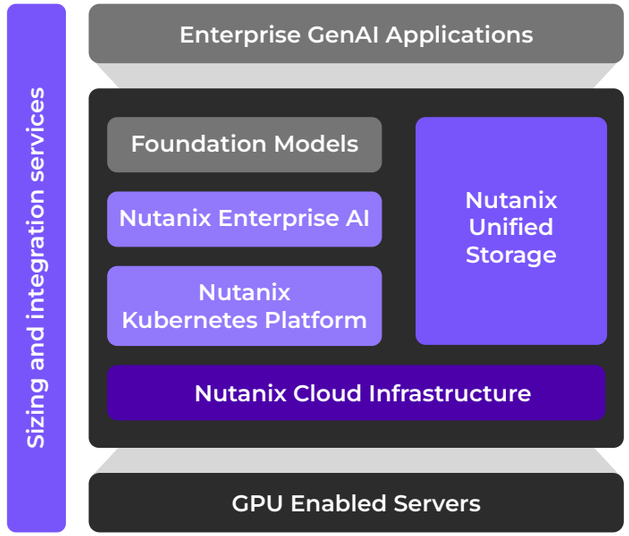
GPU パススルーは、VM 上で実行中のアプリに対して GPU リソースへの直接アクセスを提供します。Nutanix は GPU に関するクラスタ全体のビューを提供するため、利用可能な GPU を VM に割り当てることが可能になります。単一の VM に対して複数の GPU を割り当てることが可能です。パススルーの場合、任意の時点で 1 つの VM のみが GPU を利用できます。

CPU

多くの組織は、GPU が AI にとって不可欠であるという印象を抱いている一方で、最新世代の CPU にはトレーニングと推論を加速させる改良が含まれているため、GPU なしでも一部の AI 運用のサポートに適しています。

また、適切な CPU は GPU を最大限に活かすために必要となる並列処理、メモリ容量、そして帯域幅も提供できます。

利用可能なサービスには、計画ワークショップ、スタックデザインワークショップ、そしてスタックデプロイメントが含まれます。



AI 対応のバンドルサービスは、大手オープンソース AI および MLOps フレームワークを利用して、厳選された LLM 一式のデプロイに適したインフラのサイジングおよび構成に役立ちます。

業界パートナーシップ

Nutanix は、業界トップの AI 企業と提携することでフルスタックソリューションを提供しています。

NVIDIA とのパートナーシップを通じて、Nutanix AHV ハイパーバイザーは本番環境グレードの AI ソリューションの開発・デプロイメントを合理化するクラウドネイティブソフトウェアプラットフォームである NVIDIA AI Enterprise の実行が認定されています。

これには AI エージェント、コンピュータービジョン、スピーチ AI などが含まれています。NVIDIA AI Enterprise の認定により、こうした機能がすべて、NCP 上で利用可能かつ期待通りに動作することが保証されます。

NCP は NVIDIA AI Enterprise との相性が良く、俊敏性、効率性、拡張性を高める環境を構築できます。NVIDIA GPU のアクセラレーティッドコンピューティングは、AI ワークロードが GPU リソースに直接アクセスして、レイテンシの削減とパフォーマンスの加速を必ず実現できるようにします。

Nutanix は NVIDIA 以外に、Intel そして AMD とも緊密なパートナーシップを結んでいます。その結果、これらのテクノロジーリーダーによる AI イノベーションを綿密に追跡し、当社のプラットフォームにサポートを追加できるようになります。



AI のユースケース

多くの企業にとって、インテリジェントチャットボット、サポートコパイロット、スマートドキュメント処理などの生成 AI のユースケースは最優先事項となっています。Nutanix は、あらゆる種類の AI ユースケースに対応した複数のオプションを提供できます。

NVIDIA AI Enterprise と NVIDIA NIM

Nutanix は NVIDIA AI Enterprise を実行できることが認定されているため、コンピュータービジョン、スピーチ、言語理解、分子生成を含む多様なユースケースとドメインに及ぶ幅広いモデルをサポートしています。

NVIDIA AI Enterprise の顧客にとって、GPT-in-a-Box は生成 AI 向けに最適化された一式のクラウドネイティブマイクロサービスである NVIDIA NIM の展開を容易にします。

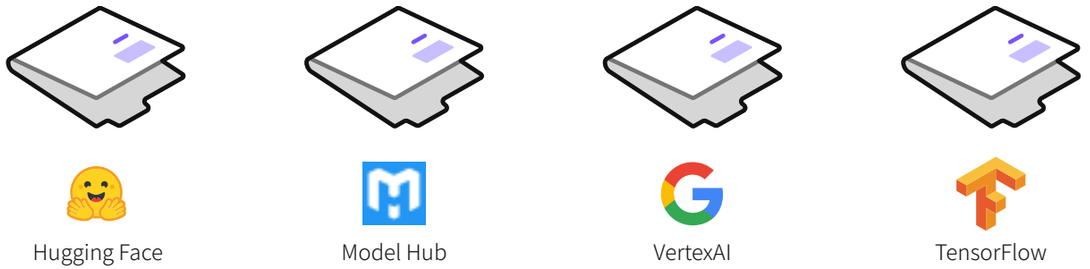
Hugging Face やその他モデルハブ

Nutanix は、モデルデプロイメントを迅速化するために Hugging Face とのパートナーシップを締結しました。これにより、Hugging Face の LLM を GPT-in-a-Box と簡単に統合できるようになります。検証済み AI LLM の検索、ダウンロード、デプロイに関するシームレスなワークフローを全面的サポートの下で提供します。

また、GPT-in-a-Box では、ユーザーが希望するモデルが未検証かつサポート対象外であってもアップロードおよびデプロイが可能です。

モデルハブ

使用に関してパーミッシブライセンスが適用される、オープンソースモデルのリポジトリ



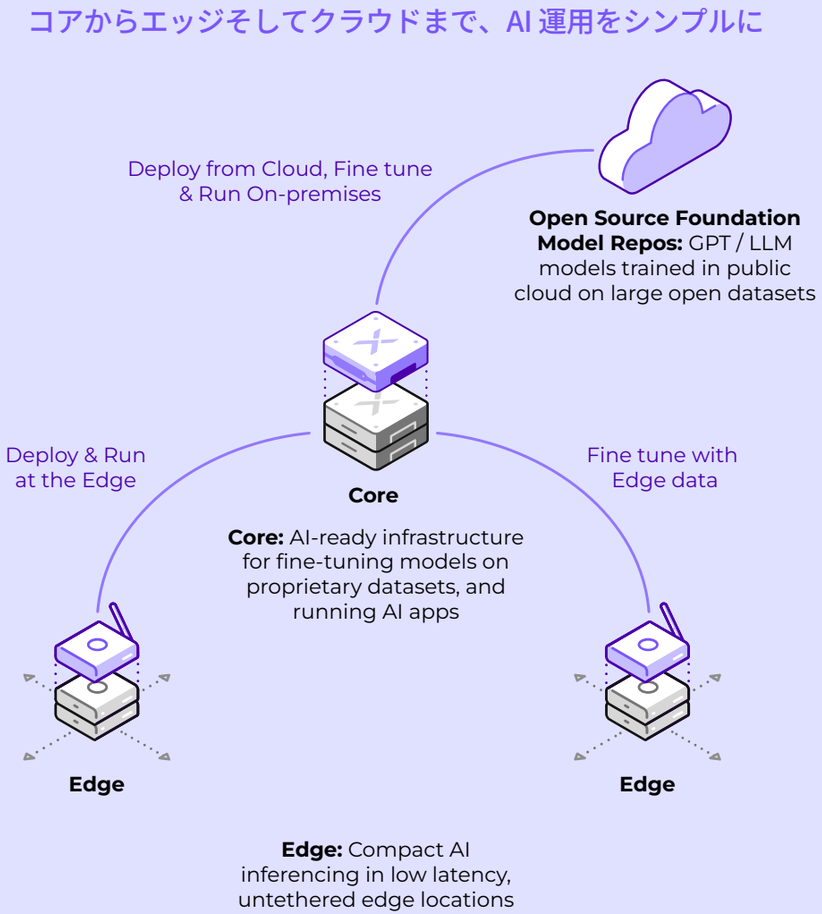
Nutanix を使い始める

Nutanix は AI に関する難点を排除して、生成 AI やその他の AI アプリを本番環境にデプロイするための工程を無事進められるように全力で取り組んでいます。

NCP はエッジからコアそしてクラウドに至る AI 運用を簡素化し、データサービスを簡素化して、自信を持って前に進めるようデータ保護とセキュリティを提供します。GPT-in-a-Box のターンキーソリューションと組み合わさることで、Nutanix はより簡単に 生成AI に取り組めるようにします。

Nutanix Enterprise AI ソリューションの詳細に関しては、AI ソリューションページにアクセスして、ぜひAI テストドライブをお試しください。

テストドライブ



NUTANIX

contact-jp@nutanix.com | nutanix.com/ja | @nutanixjapan

©2025 Nutanix, Inc.All rights reserved.本文書に記載された、Nutanix、Nutanixのロゴ、およびNutanixのその他全ての製品、提供機能、およびサービス名は、米国およびその他の国においてNutanix, Inc. の登録商標または商標となります。本文書に記載された、その他のブランド名は、識別目的のみに使用されており、それぞれの所有者の商標となります。

AI-EnterpriseAIonNutanix-eBook-FY25Q3-v2_ja-JP_100625