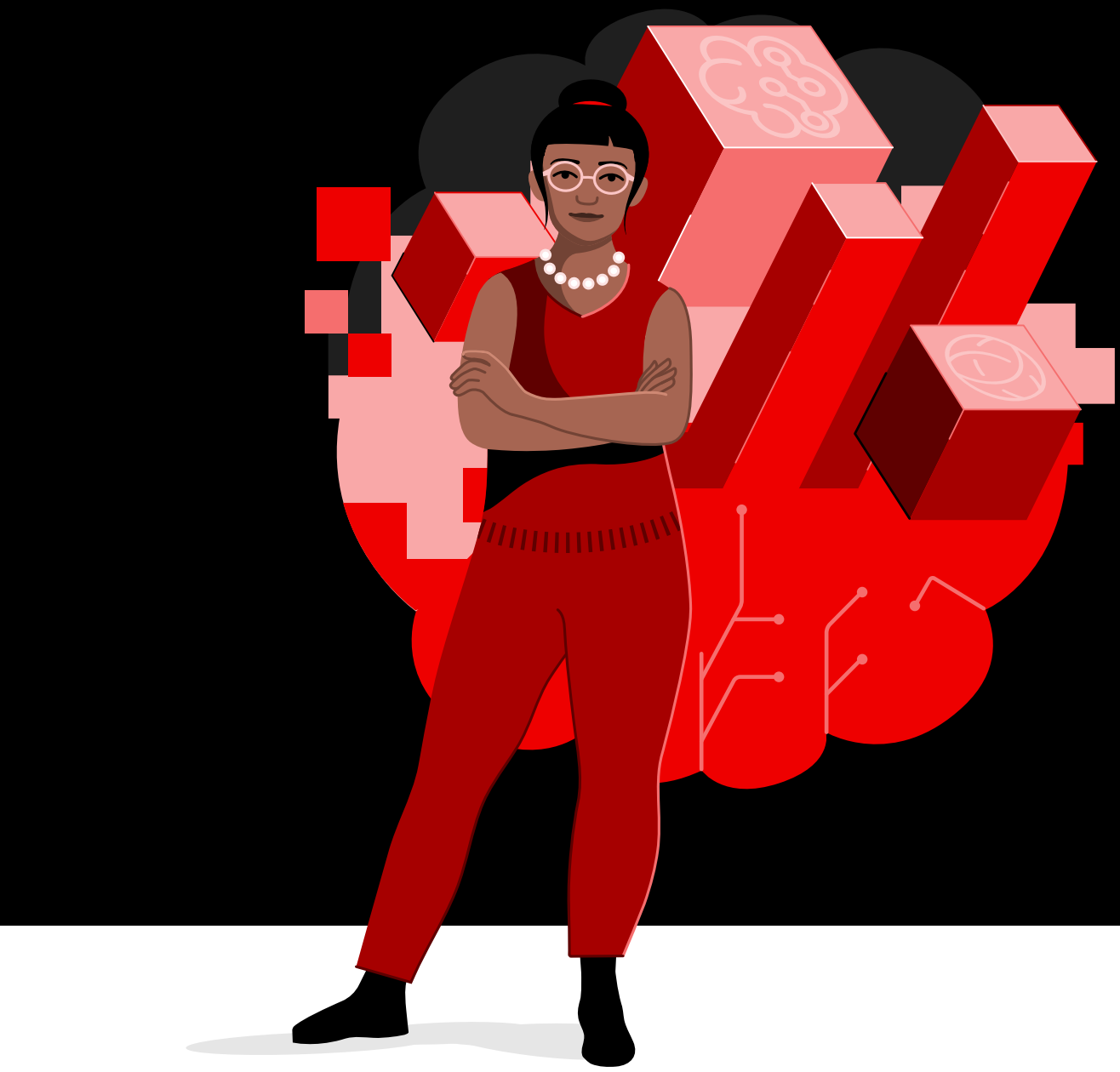
Red Hat



生成 AI の基盤を構築する際の

主な検討事項

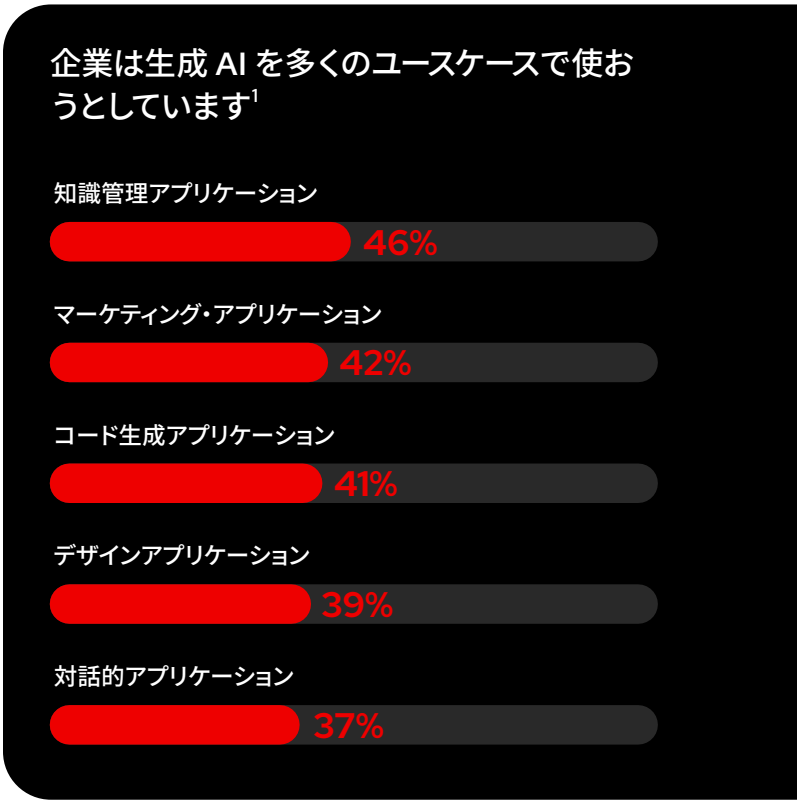
目次

ビジネスイノベーションの新たな可能性を探る

生成人工知能 (AI) は、革新的な製品の創造、プロセスの最適化、急激な変化を続ける市場における競争力の獲得を求める組織にとって強力なツールです。ディープラーニングとニューラルネットワークにおける進歩を基に、生成 AI は単にデータを処理するだけでなく新しいオリジナルなコンテンツを生成できるようになり、予測 AI の枠を超えた機能を提供します。生成 AI によって人間と機械の協働の形は新しいものへと変わり、問題解決に対する新しいアプローチを生み出しているほか、さまざまな業界で大きなビジネス上の利益をもたらしています。

世界中の組織が、生成 AI テクノロジーを使って新しい革新的なアプリケーションを構築しています。実際、現在 39% の組織が生成 AI テクノロジーに投資を行っており、さらに 37% が潜在的なユースケースを探っています。[1] 数多く存在する生成 AI のユースケースの一部をご紹介します。

- **複雑なシナリオに関する予測の生成：**生成 AI で履歴データの分析、パターンの識別、正確な予測を生成させて、戦略的計画の立案やリスク管理に役立てることができます。
- **パーソナライズしたマーケティングの開発：**生成 AI でデータを分析して顧客の好みや行動を把握し、パーソナライズしたマーケティング資料 (E メール、広告、プロモーションなど) を生成させて、エンゲージメントやコンバージョン率の最大化に役立てることができます。
- **カスタマーサービスの自動化とパーソナライズ：**生成 AI をインテリジェントなチャットボットや仮想アシスタントの基盤として使用すると、顧客の要望や行動に自動で応答し、パーソナライズされた効率的なカスタマーサービスを提供できます。



1 IDC Web カンファレンス紀要、「Unlocking Business Success with Generative AI」、Document #US50789223。2023 年 6 月。

生成 AI がもたらす新たな課題

生成 AI のメリットもデメリットもまだ見え始めてきたばかりですが、多くの企業はこの新しいテクノロジーに今投資したいと考えています。そうした企業が明確な倫理ガイドラインを作成し、フレームワークを構築し、政府および業界の規則を順守し、潜在的な問題を検知して修正するためには、生成 AI に関連する問題を理解することが助けになります。

- **データのプライバシー：** 生成 AI モデルのトレーニングに機密データや個人データを使用するとプライバシーに関する懸念が発生し、個人のプライバシーの保護に関する疑問につながります。
- **データの所有権：** プロプライエタリーなモデル、あるいはプロプライエタリーなデータでトレーニングされたモデルを使用すると、データの所有権の問題が発生し、訴訟につながるおそれがあります。
- **バイアスと公平性：** 生成 AI ツールの応答は、有害なステレオタイプやヘイトスピーチなど、人間のバイアスをそのまま反映することが示されています。
- **倫理的な使用：** 生成 AI モデルは合成コンテンツやディープフェイクを作成することもできるので、プライバシー侵害や偽情報の流布など悪意のあるアクティビティに使用されてしまう可能性があります。
- **説明可能性と解釈可能性：** 生成 AI ツールの内部処理には透明性がないので、モデルからの出力を解釈、理解、説明するのが難しく、不正確あるいは虚偽の情報が出力された場合の説明責任が存在しません。
- **想定外の結果：** 生成 AI はその性質上自律性を備えていますが、そのために現実的な損失につながる想定外の結果を人や企業にもたらす可能性があります。
- **規制に関する課題：** 生成 AI テクノロジーは日進月歩で進化しているので規制の枠組みの制定が追いつかず、責任ある倫理的な使用を徹底させるためのガイドラインの作成と施行が困難になる可能性があります。
- **エネルギー消費：** AI モデルのトレーニングには非常に高負荷な演算処理が伴い、大量の電力を必要とするので、環境に与える影響やサステナビリティに関する懸念が生じます。

この e ブックでは、生成 AI のイニシアチブを支える信頼できるインフラストラクチャ基盤を構築するための主な検討事項を見ていきます。

生成 AI に備える

IDC は「Unlocking Business Success with Generative AI」で、生成 AI のイニシアチブに対する準備として以下のアクションを推奨しています。[2]

- ビジネスニーズを満たす優先度の高いユースケースのために**すばやく実験を重ねられる環境を構築する**
- 悪意のある行動を抑制し責任ある使用を推進する**企業ポリシーを作成する**
- **生成 AI が従業員に与える影響を評価**し、プロアクティブに変革管理を行う
- **AI インフラストラクチャを任せられる信頼できるテクノロジーベンダー**およびサービスプロバイダーとパートナー関係を結ぶ
- 雇用、トレーニング、専門家サービスのサポートにより**適切なエンジニアリングスキルを確保する**

2 IDC Web カンファレンス紀要、「Unlocking Business Success with Generative AI」、Document #US50789223。2023 年 6 月。

生成 AI の基盤を構築する際の検討事項

生成 AI のテクノロジー基盤として何を選ぶかは、どれだけ容易に導入できるか、また全体的にどれだけの成功を収めることができるかに大きく影響します。この章では、生成 AI の基盤に関する主な検討事項について取り上げます。

検討事項 1：実績のあるツールセットで構築する

生成 AI モデルに基づくアプリケーションの開発は複雑なタスクになる可能性があります。オープンソースプロジェクトや商用ソリューションをベースとした言語、フレームワーク、ランタイムを備えた適切なツールセットを使用することで、モデルのチューニングを加速させ、アプリケーションの開発とデプロイを単純化できます。

革新的な AI ソリューションを迅速かつ効率的に開発するための、お客様自身が選ぶツールセットを提供する AI 基盤を選択しましょう。インタラクティブなインタフェースで探索的データサイエンス、トレーニング、チューニングを行うことができれば、コラボレーションが単純化されます。事前統合されたツールセットとセルフサービス機能が備わっていれば、複数環境をまたいで可搬性と一貫性を維持しつつ IT 運用を最適化するのに役立ちます。

検討事項 2：モデルを迅速にファインチューニングする

生成 AI のトレーニングはコストと時間のかかるプロセスなので、ほとんどの組織は汎用データで事前トレーニングされた基盤モデルを使用して AI ソリューションを構築します。データサイエンティストはその基盤モデルに対してドメイン限定のさまざまなデータを使用して、特定のタスクを実行できるように調整します。ただし、ファインチューニングも、強力なプロセッサーと分散ハイブリッドクラウド・インフラストラクチャを要する非常に高負荷な演算処理となる場合があります。

AI プラットフォームを選ぶ際には、あらゆるモデルサイズ、データ量、および処理時間のトレーニングをハイブリッドクラウド環境のどこにでもデプロイできる分散ワークロード管理およびオーケストレーション機能を備えたものを探しましょう。オンサイトのデータセンターで基盤モデルのファインチューニングを行う選択肢があれば、制限付きモデルの技術的および規制上の要件に対するコンプライアンスを単純化できます。一括トレーニング機能が備わっていれば、ファインチューニングのワークロードをプリエンプティブに実行でき、リソースの共有と管理が容易になります。

ファインチューニング以外の方法

基盤モデルをより迅速に、より効率的にチューニングする方法が研究されています。**Retrieval-augmented generation (RAG、検索拡張生成)** は社内データベース、企業イントラネット、インターネットなどの外部ソースからファクトを取得し、生成 AI モデルに最も正確かつ最新の情報を提供するための AI フレームワークです。

プロンプトチューニングでは、AI モデルは望ましい結論に向けてガイドするキューやフロントエンドプロンプト (追加の単語や AI によって生成された数字など) を受け取ります。これにより、限られたデータしか持たない組織も基盤モデルを限定されたタスク用にカスタマイズできます。

検討事項 3：効率的にモデルを提供する

生成 AI ソリューションで優れたユーザーエクスペリエンスを提供することは、IT 運用チームにとって困難な課題になる可能性があります。アプリケーションに対する需要が変動する場合、スケーラブルなインフラストラクチャと自動化された管理が必要になります。モデルを効率的にデプロイするには、パフォーマンスを監視して迅速に前のバージョンに戻せる能力が必要です。また、AI ソリューションは大量のデータを処理するので、環境全体で厳格なセキュリティ標準を適用することも欠かせません。

オンサイト・インフラストラクチャ、パブリッククラウドのリソース、エッジデバイスを含む、ハイブリッドクラウドの全体に生成 AI モデルとアプリケーションをデプロイおよびスケーリングできるプラットフォームの導入を検討してください。生成 AI モデルをオンサイトまたは分離された環境から提供できる選択肢があれば、一般利用可能なモデルの再トレーニングにプロプライエタリーなデータが使用されることがないようにできます。また、カナリアロールアウトや説明可能性ツールをサポートするものであれば、モデルの応答の一貫性と信頼性を高めるのに役立ちます。

検討事項 4：ライフサイクル管理を自動化する

継続的インテグレーション/継続的デリバリー (CI/CD) パイプラインにより、生成 AI ソリューションを自動でデプロイおよび管理できます。迅速で漸進的な変更によってモデルやアプリケーションの再トレーニングや更新を行うことで、開発のスピードアップやモデルのパフォーマンス向上を実現できます。ただし、AI パイプラインはデータ抽出、トレーニング、ファインチューニング、検証、再トレーニングなどの追加の段階を含むことが多いので、通常の CI/CD ワークフローより複雑なものになります。

Tekton や Jenkins などの CI/CD ツールをベースとして AI パイプラインを構築し、既存の DevOps ワークフローに統合して生成 AI モデルをすばやく効率的に開発、トレーニング、監視、再トレーニングできる基盤を選びましょう。ArgoCD のような **GitOps** 継続的デリバリーツールを使えば、複雑な AI ソリューションのデプロイをコードとして定義し、自動化して、一貫性をもってモデルやアプリケーションを提供することができます。

生成 AI 向けのコンテナ

コンテナや **Kubernetes** といったテクノロジーは、アジャイルなデプロイ、管理、およびスケーラビリティを実現し、生成 AI ソリューションのクラウドネイティブ開発を加速します。オンサイトデータセンター、パブリッククラウド、エッジデバイスの全体で、オンデマンドに環境をプロビジョニングすることが可能です。物理および仮想インフラストラクチャ上で自動的にコンテナインスタンスを作成、デプロイ、スケーリング、管理できます。また、オープンソースおよび商用サプライヤーの堅牢なエコシステムから提供されるコンポーネントやデータストアを生成 AI に統合することもできます。**AI 向けのコンテナのメリット**をご覧ください。

検討事項 5：一貫性をもってモデルを監視する

生成 AI モデルは人々やビジネスに実質的で大きな影響を与えることができます。モデルの動作を追跡することで、判断や理由付けを分析し、望ましくないパフォーマンスを特定し、問題のある動作を即座に報告することが可能です。この情報に基づいて効果的にモデルのガバナンスを行うことで、プロダクション環境に配置したモデルがバイアスのない、公平で正確な情報を利用して応答するようにできます。

生成 AI モデルの調査、保守、修正に役立つ、バイアスやデータドリフトに関するメトリクス、異常検知、ポイントごとの説明可能性を提供する一元的な管理機能を備えた AI 基盤を選びましょう。プロダクション環境で継続的な自動監視を行う機能があれば、企業のモデルガバナンス標準へのコンプライアンスを強化することができます。また、ユーザーフレンドリーなインタフェースを備えたツールや、人間が読めて技術専門性が高くないレポートを生成する機能は、責任のあるモデルの使用と保守を促進します。

生成 AI モデルの主な概念

- **バイアス：**特定のグループに対する肩入れ、ステレオタイプに同調する応答など、生成された出力の公平性、インクルーシブ性、倫理性に影響を与える動作パターンがモデルに存在すること
- **データドリフト：**時間の経過とともにトレーニングデータの統計的特性が変化し、モデルのパフォーマンスを低下させたり、正確性や関連性の低い応答を生成したりすること
- **異常検知：**通常とは異なる、あるいはトレーニング中のサンプルと乖離した動作を特定し、報告するプロセス
- **ポイントごとの説明可能性：**透明性が重要なアプリケーションに可視性を提供する、モデルがなぜその応答を出力したかを把握するための機能

検討事項 6：パートナーエコシステムを活用する

生成 AI ソリューションが適切に革新的なユーザーエクスペリエンスを提供するためには、複数の統合されたコンポーネントが必要です。信頼できるベンダーで構成されたコラボレーティブなエコシステムが提供するテクノロジーを適切に組み合わせることで、アプリケーション開発の迅速化、バイアスおよびデータドリフトの課題への対処、ソリューション全体の一貫性のある信頼できるパフォーマンスの実現が可能になります。

生成 AI モデルおよびアプリケーションの開発とデプロイのための完全なソリューションを提供する、広範な認定パートナーエコシステムを持つプラットフォームベンダーを探しましょう。データの統合や準備からモデルのトレーニングや提供まで、コンポーネントの幅広いセレクションから選択できれば、より迅速に、効率よく AI ソリューションを開発してデプロイするのに役立ちます。また、相互運用性に実績のある認定済みソリューションを選択することで、IT サポートリクエストを減らし、生産性を高めることができます。

検討事項 7：プラットフォームのエキスパートと連携する

生成 AI ソリューションを効率的にデプロイおよび管理するためには、専門的な知識と経験が必要です。スケーラビリティの要件、信頼性に関する懸念、既存システムとの統合により、プロダクション環境へのデプロイの複雑性が高まることもあります。コンピューティング・リソースの使用効率が悪ければ不必要なコストがかかります。また、セキュリティ標準、プライバシーポリシー、AI 規制フレームワークに対するコンプライアンス違反があった場合、意図しない結果につながることがあります。

生成 AI ソリューションの構築に関する包括的なサポートとガイダンスを提供できるエキスパートのチームを持つベンダーを選びましょう。たとえば、専任のエンジニアがいれば、お客様の AI プロジェクトを加速するツール、リソース、知識を活用した、プラットフォーム全体に対するサポートが提供されます。エキスパートのコンサルタントがいれば、デプロイに関する課題の解決、インフラストラクチャの効率性の最適化、AI ソリューション全体の相互運用性の確保について支援を受けられます。また、プロフェッショナル・トレーニング・サービスが提供されていれば、新しい生成 AI プロジェクトをすばやく開始するのに必要な知識や技能を習得するのに役立ちます。

生成 AI にはコラボレーションが必須

生成 AI プロジェクトを生成させるためには、広範な部署からの人員で構成されたチームの構築が鍵となります。[3]

- **ビジネスリーダー：**ソリューションのユーザー、またはソリューションによる影響を受ける人々の代表
- **AI スペシャリスト：**AI モデルのチューニング、保守、更新を担当
- **データサイエンティスト：**正確でバイアスを含まないトレーニングデータを前処理し、モデルに提供
- **倫理およびコンプライアンス担当者：**生成 AI イニシアチブが規制を順守するよう監督
- **IT 運用スペシャリスト：**ソリューションを既存のインフラストラクチャと統合し、セキュリティポリシーを適用

3 Kearney、「Standing up tiger teams to tackle generative AI complexity」、2023 年 11 月。

柔軟でオープンな基盤で迅速にイノベーションを実現

Red Hat は、生成 AI に関するお客様の目標達成を支援する完全なテクノロジー・ポートフォリオ、実証済みの専門知識、戦略的パートナーシップを提供します。迅速な導入のためのサービスとトレーニングだけでなく、生成 AI モデルとアプリケーションを開発し、デプロイするための基盤を提供します。

Red Hat® OpenShift® は、クラウドネイティブのイノベーションを実現する、統合されたエンタープライズ向けアプリケーション・プラットフォームです。オンデマンドのコンピュートリソース、ハードウェア・アクセラレーションのサポート、オンサイト/パブリッククラウド/エッジ環境間の一貫性により、成功するために必要な速度と柔軟性がもたらされます。Red Hat OpenShift を使用すると、データサイエンティスト、データエンジニア、開発者が迅速にインテリジェントなアプリケーションを開発するためのセルフサービスのプラットフォームを構築できます。コラボレーション機能は、チームがコンテナ化されたモデリング結果を作成し、同僚や開発者と一貫した方法で共有することを可能にします。

Red Hat OpenShift AI は Red Hat OpenShift をベースとしており、先進的な AI ソリューションのワークロードおよびパフォーマンスに関する要求を満たすと同時に、モデルとアプリケーションの構築、トレーニング、ファインチューニング、デプロイ、および監視のための包括的なプラットフォームを提供します。NVIDIA、インテル、Starburst、Anaconda、IBM、Run:ai、Pachyderm などのパートナーが提供する主要な認定済み製品を統合したコラボレーティブで一貫性のある環境で、実験からプロダクションへと迅速に移行することができます。Red Hat OpenShift AI は、テクノロジーエコシステムと併せて、ハイブリッドクラウド全体で革新的な生成 AI ソリューションの開発とデプロイを加速させるコンポーネントと機能を提供します。

IBM watsonx.ai AI studio は、インテリジェントなアプリケーションが必要とする生成 AI 機能を備えた厳選されたモデルとデプロイの選択肢を提供します。AI ソリューションのパフォーマンスと効率性を高められるよう、モデル (オープンソース、サードパーティ、IBM 開発の基盤モデルなど) はワークロードのある場所ならどこにでもデプロイできます。**IBM 開発の基盤モデル**をお客様に関連のあるデータでトレーニングした場合、お客様の生成 AI ソリューションは事業ドメインの機微を理解し、競争上の優位性を提供します。

Red Hat Ansible® Lightspeed with IBM watsonx Code Assistant は、より効率的に自動化コンテンツを作成、導入、保守できるように設計された生成 AI サービスです。IBM watsonx Code Assistant に接続された Red Hat Ansible Lightspeed により、自然言語プロンプトを使って自動化のアイデアを Ansible コードに変換できるようになります。これを使うことで、生産性を高め、組織全体で自動化にアクセスしやすくすることができます。

生成 AI を使い始める

生成 AI はオリジナルコンテンツを作るための強力なツールであり、人とアプリケーションやテクノロジーとのインタラクションのあり方を一新します。

Red Hat は、テクノロジー、専門知識、パートナーシップを通じて、お客様のチームが AI アプリケーションと ML モデルを透明性と制御性をもって構築し、デプロイするための共通の基盤を提供します。さらに、独自の AI ツールとプラットフォームを使用して他のオープンソースソフトウェアの実用性を高めています。また、パートナーとの統合により、Red Hat OpenShift AI のようなオープンソース・プラットフォームと連携するよう構築された、信頼できる AI ツールのエコシステムとお客様をつなぎます。

Red Hat OpenShift AI の詳細をご覧ください。無料でお試しいただくこともできます。

Red Hat コンサルティングを活用して、より迅速に開始する

Red Hat のエキスパートと連携し、AI/ML プロジェクトをスピーディに立ち上げましょう。Red Hat は、AI/ML の迅速な導入を支援するコンサルティングおよびトレーニングサービスを提供しています。

- AI/ML サービスの詳細：
  red.ht/aiml-consulting
- 無料のディスカバリー・セッションを予約：
  red.ht/consulting-ja

1057715_0324_KVM

Red Hat