

本番利用に適した AI 環境構築のための 最重要事項



目次



データは重要なビジネス
資産

3 ページ



本番利用に適した AI プ
ラットフォームの構築

6 ページ



コンテナとコンテナ・
オーケストレーション

9 ページ



アプリケーション管理と
GenAIOps

11 ページ



ハイブリッドクラウド・
プラットフォーム

13 ページ



モデルのカスタマイズと
アライメント

14 ページ



大規模な AI 推論

15 ページ



AI の安全性

17 ページ



AI のためのオープンで
柔軟な基盤の構築

18 ページ



認定 AI/ML パートナー
エコシステムで選択肢と
柔軟性を得る

20 ページ



パートナーのハイライト

21 ページ



成功事例

22 ページ



データを最大限に活用す
る準備はできていますか？

23 ページ

データは重要な ビジネス資産

1010
11011

エンタープライズ AI 市場の状況

生成 AI は、多くの組織にとって実験から日常的に使用されるツールになりました。

さまざまなチームが、コンテンツの要約、コードやコンテンツの作成支援、より自然な方法でのデータ操作のために使用しています。エンタープライズ規模でリーダーが生成 AI に期待していることは、単にアドホックな質問に答えたり面白いミームを作ったりするだけでなく、顧客、従業員、オペレーション全体の結果を向上させることです。

生成 AI は既存のデータとアプリケーションを基盤として、組織が以下のことを実現するのに役立ちます。



大量の非構造化コンテンツを、検索可能で再利用可能な知識に転換する



開発者、アナリスト、ライターがコード、レポート、コンテンツをより迅速に作成および改良できるように支援する



チャンネルをまたいで顧客と従業員のデジタル・エクスペリエンスをパーソナライズする



明確なポリシーに従った、定型的な意思決定やワークフローを自動化する



開発、運用、ビジネスの各チームの生産性を向上させる

最近の業界調査によると、このような変化はすでに進行しています。IDC の報告によると、調査対象の組織の半数以上がすでに生成 AI で強化されたアプリケーションやサービスをプロダクション環境で実行しており、2025 年から 2029 年までの間、AI に対する支出は毎年約 30% 増加し、2029 年には約 1 兆 3,000 億米ドルに達すると予測されています。¹ ほとんどの企業にとって、生成 AI は中核的な製品およびサービスの一部になりつつあります。

¹ IDC ホワイトペーパー、「[Agentic AI to Dominate IT Budget Expansion Over Next Five Years, Exceeding 26% of Worldwide IT Spending, and \\$1.3 Trillion in 2029, According to IDC](#)」、2025 年 8 月 26 日。

同時に、組織は次のステップであるエージェント型 AI に目を向けています。エージェント型 AI では、生成 AI を単一のチャットボットまたはアシスタントとして扱うのではなく、ツールの呼び出し、アプリケーションとの対話、マルチステップタスクの調整を実行できる AI エージェントを使用します。実際、このアプローチにより、顧客のセルフサービスや IT 運用から複雑なビジネスワークフローまで、ソフトウェアの構築方法や運用方法が変革される可能性があります。

IDC のレポートによると、半数以上の組織がエージェント型 AI の概念実証や初期のユースケースをすでに実施しており、AI 対応アプリケーションのほぼ 3 分の 1 が 2026 年末までにエージェント型 AI を使用するようになるとのことです。² 企業は今やエージェント型 AI を、前進するための戦略的な手段として捉えています。

このような価値を実現するには、AI を実行する方法や場所に柔軟性が必要です。

現在、多くの組織が、パブリッククラウドと専用のオンプレミス環境を組み合わせたハイブリッド AI インフラストラクチャの計画を立てています。IDC によると、最も一般的なデジタル・インフラストラクチャ戦略は、パブリッククラウドとオンプレミス・インフラストラクチャを組み合わせたハイブリッドとなっており、ほとんどの意思決定者が、AI ワークロードにはハイブリッド・デプロイメントが必要だと考えているとのことです。³



ハイブリッドのオープン・プラットフォームにより、組織では次のことが可能になります。



機密データとモデルを組織の管理下に維持する



データのプライバシーと主権に関する要件を満たす



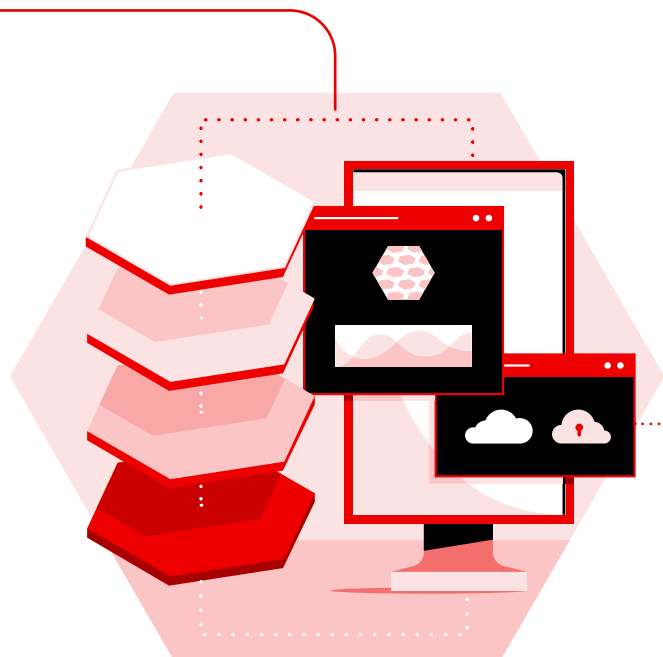
さまざまなハードウェアオプションから選択する



幅広いオープンソースモデルから選択する



必要なときにクラウド規模のメリットを活用する



この e ブックでは、本番利用に適した AI プラットフォームを構築するための基本ステップ、組織がその過程で直面する重要な検討事項、そして Red Hat® AI Enterprise が提供する統合ソリューションがどのようにその構築を支援するかについて説明します。

² IDC ホワイトペーパー、「Agentic AI Impact on Digital Infrastructure Strategies」、Document #US53418526、2025 年 10 月 (購入が必要です)

³ IDC ホワイトペーパー、「AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025」、Document #US53418426、2025 年 10 月 (購入が必要)

各業界での AI ユースケース



医療

- ・ 臨床効率の向上
- ・ 診断の速度と精度の向上
- ・ 患者の治療効果の向上



通信

- ・ 顧客の行動に関する知見の取得
- ・ カスタマーエクスペリエンスの向上
- ・ 5G ネットワークのパフォーマンス最適化



保険

- ・ 請求処理の自動化
- ・ 利用ベースの保険サービスの提供
- ・ リスク計算の支援



金融サービス

- ・ カスタマーサービスのパーソナライズ
- ・ リスク分析の向上
- ・ 詐欺やマネーロンダリングの検出



自動車

- ・ 自動運転のサポート
- ・ メンテナンスの必要性の予測
- ・ サプライチェーンの強化



エネルギー

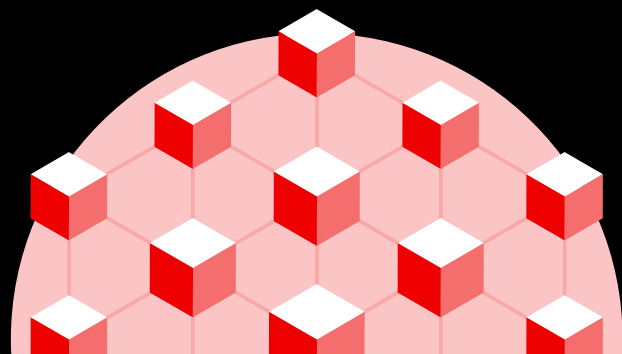
- ・ メンテナンス要件の予測
- ・ 現場の運用および安全の最適化
- ・ 油層のシミュレーションと予測の迅速化

エンタープライズ AI の構成要素

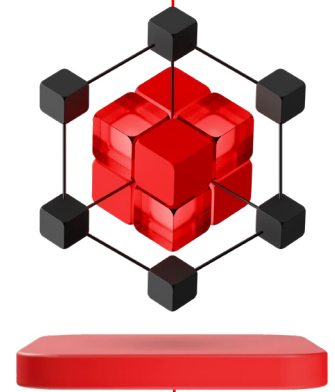
この e ブックでは、エンタープライズ・アーキテクチャにおいてさまざまな種類の AI がどのように連携するかを説明します。

- ・ **生成 AI**: 大規模言語モデル (LLM) を使用して、データとプロンプトからテキスト、コード、その他のコンテンツを生成します。これにより、チームはより迅速に作業を行い、より簡単に実験できるようになります。
- ・ **予測型 AI**: 履歴データとリアルタイムデータを使用して、需要、リスク、機器の健全性など、将来の結果を推定します。これにより、組織はより早く、自信を持って行動に移せます。
- ・ **エージェント型 AI**: ツールの呼び出し、アプリケーションへの接続、1 つの質問に答えるだけでなく目標に向けた多段階のワークフローの調整が可能な AI エージェントを使用します。

- ・ **AI 推論**: プロダクションランタイムのフェーズであり、モデルが学習した内容を実世界の新しいデータに適用して、予測、推奨事項、またはアクションを返します。推論はハイブリッド環境、つまりオンプレミス、クラウド、またはエッジで実行できます。



本番利用に適した AI プラットフォーム の構築



生成 AI を活用したアプリケーションと AI エージェントの構築は反復プロセスであり、単に AI モデルを作成するだけではありません。AI ライフサイクルの主なステップは次のとおりです。

- 1 ユースケースを定義し、AI イニシアチブのビジネス目標を設定し、ステークホルダーやリーダーの賛同を得る
- 2 モデルの実験およびデプロイメント・プラットフォームを実行する場所（オンプレミスまたはクラウド）を選択する
- 3 ニーズに最適な AI モデルを選択するオープンモデルを選んでロックスインを回避する
- 4 検索拡張生成（RAG）を使用して、選択したモデルをカスタマイズする、またはプロプライエタリーなデータに合わせて調整する
- 5 モデルを推論サーバーにデプロイする
- 6 生成 AI を活用したアプリケーションやワークロードを構築する
- 7 作業環境が整ったら、エージェント型 AI を通じてワークフローを拡張および自動化する
- 8 セキュリティを重視した方法で、大規模にモデルを監視および管理する



オープンで適応性の高い AI アーキテクチャは、このプロセスをより効率的に実行するのに役立ちます。このアーキテクチャには、いくつかの主要なテクノロジーと機能が必要です。

- **最先端のオープンウェイトモデルへのアクセス**：これが組織の出発点となります。
- **GenAIOps および DevOps ツール**：これを使用して、AI エンジニア、データサイエンティスト、機械学習 (ML) エンジニア、アプリケーション開発者は、AI モデル、AI エージェント、AI を活用したアプリケーションを作成、デプロイ、管理できます。
- **ファインチューニングや RAG 機能などのモデル・チューニング・ツールへのアクセス**：エンタープライズのプライベートデータを使用してモデルをカスタマイズし、ドメイン固有のユースケースに合わせて調整できます。
- **推論ランタイム**：最適なパフォーマンス、スループット、レイテンシーを実現できます。
- **AI エージェントの基本コンポーネント**：プロダクションでの実装を管理、制御、保護します。
- **コンピューティング、ストレージ、ネットワーク・アクセラレーター**：データの準備、モデルのカスタマイズ、推論タスクを高速化します。
- **インフラストラクチャ・エンドポイント**：オンサイト、仮想、エッジ、およびプライベート、パブリック、ハイブリッドの各クラウド環境において、AI 運用のすべての段階でリソースを提供します。



この eブックでは、効果的な AI アーキテクチャを構築するための重要な考慮事項について説明します。

推論は AI のプロダクションランタイムです。モデルは、API を取得してコンテンツを提供しなければ、ユーザーにとって役立つことはしません。そのコンテンツは推論を通じて提供されます。

Chris Wright (クリス・ライト)

Red Hat CTO⁴

⁴ Ron Miller, 「[Red Hat's CTO sees AI as next step for company's open approach](#)」、Fastforward、2025 年 11 月 11 日。

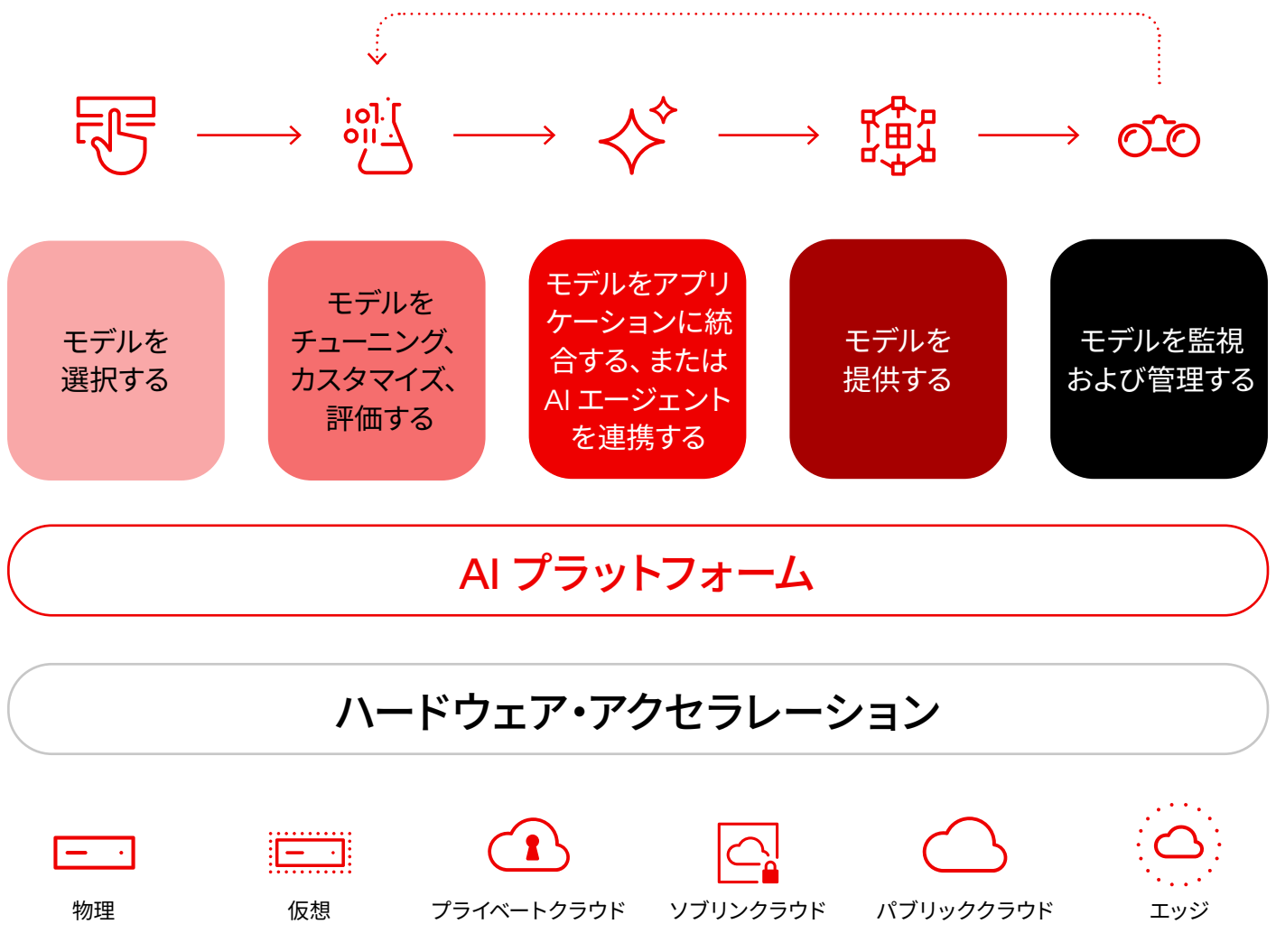


図 1. AI アーキテクチャのコンポーネント

AI のデプロイの課題

エンタープライズ組織は、競争力を提供する AI ソリューションを選択、構築、提供しなければならないというプレッシャーにさらされています。AI デプロイメントの運用化と拡張には、複数の課題が立ちほだかります。

- モデルのコスト:** 大規模モデルを実行し、推論を大規模に実行すると、高コストとなる可能性があります。組織は、モデルと推論を最適化してコンピューティングコストを抑制しながら、正確で応答性の高いアプリケーションを提供する必要があります。
- アライメントの複雑さ:** モデルのトレーニングとチューニング、RAG パイプラインの作成は複雑であり、グラフィックス・プロセッシング・ユニット (GPU) を大量に消費しま

す。組織はエンタープライズデータのカスタマイズを単純化し、対象分野の専門家や開発者を関与させて、実験からプロダクションへと迅速に移行させることができます。

- 制御と一貫性:** 事前にパッケージ化された AI サービスでは、ハードウェア、データ、ガバナンスの制御が制限されます。ハイブリッドアプローチを選択することで、データの所有権、ライフサイクル、デプロイメントの規模を維持しながら、モデルとインフラストラクチャを選択できます。

これらの課題に対処するには、環境全体でのモデルの最適化、カスタマイズ、ガバナンスのための一貫したツールを提供するオープンなハイブリッド AI プラットフォームが必要です。

コンテナとコンテナ・オーケストレーション



コンテナ

コンテナはソフトウェアの基本単位であり、あらゆる依存関係とともにアプリケーションをパッケージ化します。コンテナを使用することでアプリケーションのビルドプロセスは単純化され、変更を加えることなくアプリケーションを異なる環境にデプロイできるようになります。



AI にとって重要な理由

AI エンジニアやアプリケーション開発者が最大限に生産性を高めるためには、自分の好みのツールとリソースを利用できることが必要です。同時に、IT 運用チームは、リソースが最新で、コンプライアンスに準拠し、安全に使用されるようにしなくてはなりません。

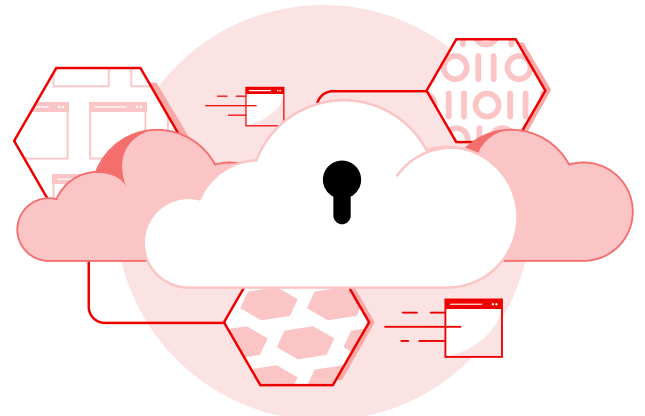
コンテナは、多くの場合、LLM のデプロイや生成 AI を活用するアプリケーションのデプロイに最適な選択肢です。モデルサーバー、依存関係、構成が反復可能なユニットにパッケージ化されるので、プロダクション環境のロールアウト、スケーリング、更新がより管理しやすくなるからです。

コンテナを使用すると、ハイブリッド環境全体に一貫した方法でさまざまな AI ツールをデプロイできます。チームは、変更を追跡して透明性を実現するバージョン管理機能を使用して、コンテナイメージを繰り返し変更し、共有することができます。同時に、プロセスの分離とリソース制御により、脅威からの保護が向上します。



ベストプラクティスと推奨事項

統合されたセキュリティ機能が含まれ、環境間でのコンテナのデプロイ、管理、移動方法を効率化する、柔軟で可用性の高いコンテナ・プラットフォームを選択しましょう。幅広いテクノロジーと統合できるオープンソース・プラットフォームを選択すれば、柔軟性と選択肢を増やすことができます。



コンテナ・オーケストレーション

コンテナのオーケストレーションにより、環境全体におけるコンテナの作成、デプロイ、ライフサイクルが管理されます。



AI にとって重要な理由

コンテナを導入したら、それらを効率的にデプロイ、管理、スケーリングする方法が必要です。コンテナ・オーケストレーション・エンジンにより、コンテナのライフサイクルを一貫した方法で管理できます。これらのツールは通常、オンサイト、エッジ、クラウド環境全体で、コンピュータ、ストレージ、ネットワークのリソースへのアクセスを一元化します。また、ワークロード・スケジューリング、マルチテナンシー制御、クォータの適用を統合します。

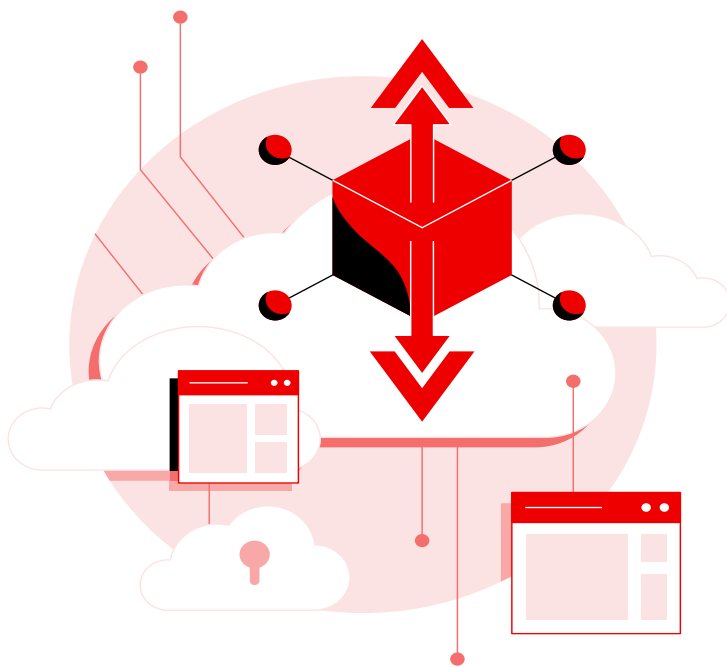


ベストプラクティスと推奨事項

優れたオープンソース・テクノロジーをベースに構築し、プロプライエタリーなクラウドへのロックインを回避するには、Kubernetes ベースのコンテナ・オーケストレーション環境を選択します。AI ワークロードを一貫して管理できるように、強力なマルチテナンシー制御、ロールベースのアクセス、ポリシー管理を提供するプラットフォームを探しましょう。Operator の幅広いエコシステムと統合機能でオプションに優先順位を付けることで、ハイブリッド環境全体で AI サービスをデプロイ、スケーリング、管理する方法を標準化できます。



すべての AI デプロイメントのうち、基盤となるコンピューティング環境としてコンテナ・テクノロジーを使用する割合は、2027 年までに 75% を超えると予測されています (2024 年の 50% 未満から上昇)。⁵



⁵ Gartner, 「Magic Quadrant for Container Management」、2024 年 9 月 10 日。

アプリケーション管理と GenAIOps



AI ワークロードのライフサイクル管理

AI ワークロードのライフサイクル管理では、AI のユースケースを強化するツールやサービスのデプロイ、スケーリング、管理の方法に焦点が当てられます。



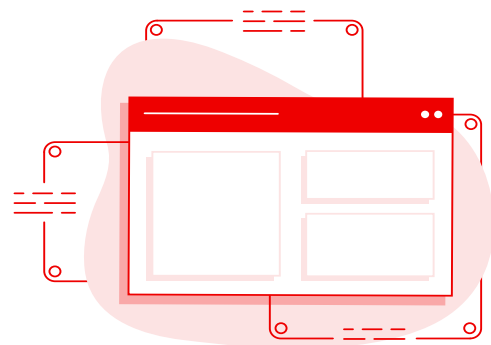
AI にとって重要な理由

AI 環境は本質的に複雑です。AI ワークロードのライフサイクル管理コンポーネント（ノートブック、ワークベンチ、パイプライン、モデル提供エンドポイントなど）は、コンテナ化して、制御と管理をシンプルにする必要があります。IT 運用チームは、構成、プロビジョニング、更新などの一般的なライフサイクルタスクを自動化して、精度を向上させ、手作業を削減できます。データサイエンティスト、AI エンジニア、アプリケーション開発者は、IT 部門でチケットを発行しなくても、カタログから事前承認済みの AI 環境をリクエストできます。また、自動化により、スタッフは繰り返しのタスクではなく、より価値の高い戦略的作業に時間を費やすことができます。



ベストプラクティスと推奨事項

効果的な AI ワークロードのライフサイクル管理は、一般的に使用される AI および ML ライブラリを含む、厳選された AI ワークベンチとノートブックのイメージから始まります。これにより、チームはアドホックな環境ではなく、サポートされた安全なベースラインから開始できます。組織は、Git の統合を備えたブラウザベースのノートブック環境を提供する必要があります。これにより、チームは実験に共同で取り組み、コードやモデルの変更を長期にわたって追跡できます。



GenAIOps および MLOps プラクティス

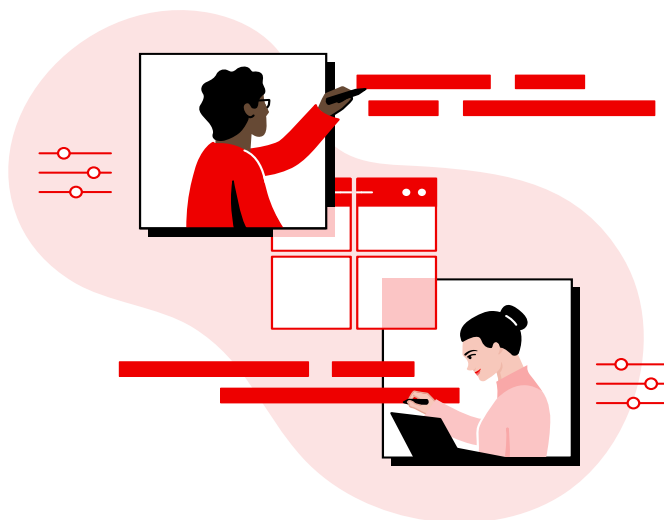
GenAIOps および MLOps プラクティスは、AI を大規模に運用するために必要なツール、プラットフォーム、プロセスを統合します。



AI にとって重要な理由

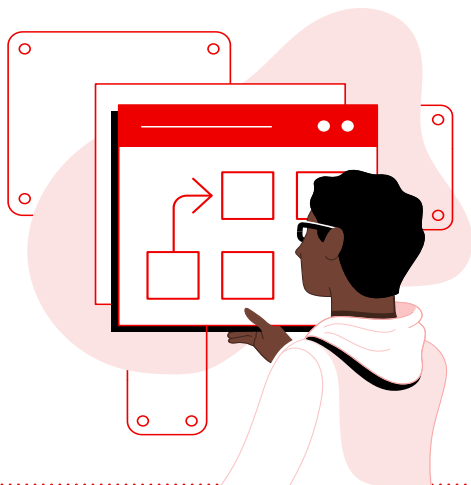
組織は、AI モデル（およびそれらを使用するアプリケーション）を迅速かつ効率的に開発してデプロイする必要があります。このような取り組みを成功させるためには、チーム間の連携が不可欠です。

DevOps と同様に、GenAIOps および MLOps アプローチは、AI エンジニア、アプリケーション開発者、IT 運用間のコラボレーションを促進し、生成 AI モデル、AI エージェント、AI を活用したアプリケーションの作成、トレーニング、デプロイ、管理を加速します。自動化（多くの場合、継続的インテグレーション/継続的デリバリー（CI/CD）パイプラインの形式）により、モデルおよびアプリケーション開発のライフサイクルを短縮するために、迅速で段階的な反復型の変更を行うことが可能になります。



GenAIOps および MLOps プラクティス

GenAIOps および MLOps はテクノロジーだけの問題ではなく、人とプロセスが重要な役割を果たします。GenAIOps および MLOps プラクティスを AI ライフサイクル全体に適用しましょう。[Kubeflow](#) などのオープンソース・テクノロジーとともに、プラットフォームとツールの自動化を使用して、CI/CD パイプラインとワークフローを構築しましょう。



ハイブリッドクラウド・プラットフォーム



ハイブリッドクラウド・プラットフォームは、オンサイト、エッジ、およびクラウド環境で、AI の開発、デプロイ、管理を行うための基盤になります。また、最初からソブリン AI とプライベート AI 用に設計できるため、パブリッククラウドで実行するワークロードと、自身で制御するオンプレミスまたはプライベートクラウド環境に配置するワークロードを選択できます。



AI にとって重要な理由

AI のモデル、エージェント、ソフトウェア、アプリケーションには、開発とデプロイのためのスケーラブルなインフラストラクチャが必要です。一貫性のあるハイブリッドクラウド・プラットフォームを使用すると、インフラストラクチャのあらゆる部分で、AI モデルとアプリケーションの開発、チューニング、テスト、デプロイ、管理を同じ方法で行うことができ、柔軟性が高まります。

また、ソブリン AI およびプライベート AI 戦略もサポートしており、機密データやモデルを特定のリージョンやオフライン環境に保存して、データの保存場所、プライバシー、コンプライアンスの要件を満たしながら、必要に応じてパブリッククラウド・サービスに接続できます。さらに、セルフサービス機能により、IT 制御を維持しながら、リソース提供を高速化できます。

そして、一貫性のあるプラットフォームが、サードパーティベンダー、オープンソース・コミュニティ、および使用するカスタムツールによるテクノロジー統合の基盤を提供します。

ベストプラクティスと推奨事項

ハードウェア・アクセラレーション、AI とアプリケーション開発ツールの広範なエコシステム、統合された GenAIOps と運用管理機能をサポートする、セキュリティ重視のプラットフォームを選択しましょう。

データの局所性、モデルの配置、アクセスのための強力なポリシー制御を検討しましょう。これにより、ソブリンおよびプライベート AI ワークロードをオンプレミスまたはプライベートクラウドで実行しながら、必要に応じてパブリッククラウドに接続できます。

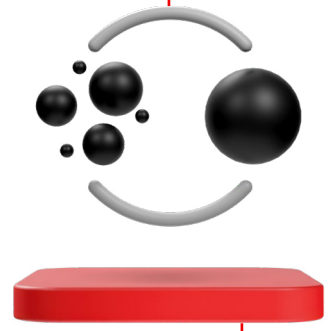
オープンソース・プラットフォームを選択すると、統合の機会が増えて柔軟性が向上し、コミュニティ主導の開発を通じて迅速なイノベーションを促進できます。また、IT 制御を維持しながらリソース提供を迅速化できるセルフサービス機能も得られます。

パブリッククラウドと専用のオンプレミス・インフラストラクチャのハイブリッドの組み合わせは、最も一般的なデジタル・インフラストラクチャ・アーキテクチャ戦略です。³

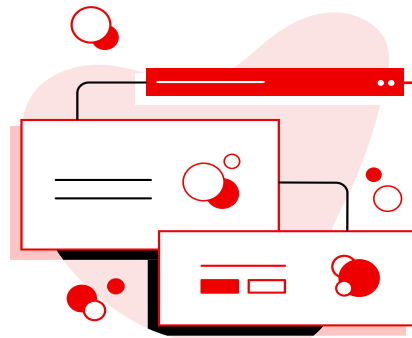


³ IDC ホワイトペーパー、「AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025」、Document #US53418426、2025 年 10 月 (購入が必要)

モデルのカスタマイズとアライメント



AI を活用した先進的なアプリケーションには、組織固有のデータ、ワークフロー、ビジネス上の制約を反映するモデルが必要です。最先端のモデルやオープンなモデルとプロプライエタリーな情報とのアライメントにより、汎用的な対応から、ドメインを認識した正確な結果へと移行できます。



AI にとって重要な理由

生成 AI とエージェント型 AI は、用語、データ、実際のコンテキストを理解するモデルに依存します。

プライベートデータをモデルの基礎としてアライメントを行うことで、正確性と関連性を維持できます。これにより、推論コストが削減され、不必要なサイズオーバーが回避されるため、効率が向上します。ビジネスロジック、安全性ルール、コンプライアンス要件をモデルの動作に直接実装できるため、ガバナンスと制御が強化されます。また、データの進化に合わせてモデルのアップデート、再トレーニング、バージョン管理を行うための一貫したプロセスが実現するため、スケーラビリティもサポートされます。

カスタマイズによってソブリン AI およびプライベート AI 戦略もサポートされるため、組織はデータのレジデンシー、プライバシー、規制要件を満たすために、制御された環境内のみでモデルをトレーニングして提供することができます。



ベストプラクティスと推奨事項

1つの手法に依存するのではなく、ニーズに応じて RAG、ファインチューニング、プロンプトエンジニアリング、およびポリシーのレイヤーから始まるモジュール式のワークフローを導入しましょう。オープンモデルを使用してロックインを回避し、モデルを透過的にファインチューニング、量子化、評価する機能を維持します。モデルに実際のビジネスコンテキストとデータの正確性を反映させるために、各分野の専門家を参加させます。量子化、蒸留、効率的なランタイムなどの手法を適用してコストとレイテンシーを制御することで、早期にモデルを推論用に最適化します。さらに、バージョンデータセット、トレーニングの実行、モデルの重み、評価指標を使用して、再現性と強力なガバナンスを維持します。

大規模な AI 推論



プロダクションで AI を実行するには、高速で効率的、かつ信頼性の高い推論が必要です。モデルのトレーニングやアライメントが完了すると、推論のフェーズとなります。このフェーズでは、新しいデータを処理したり、予測を返したり、コンテンツを生成したり、アプリケーションやワークフロー内でアクションをトリガーしたりします。

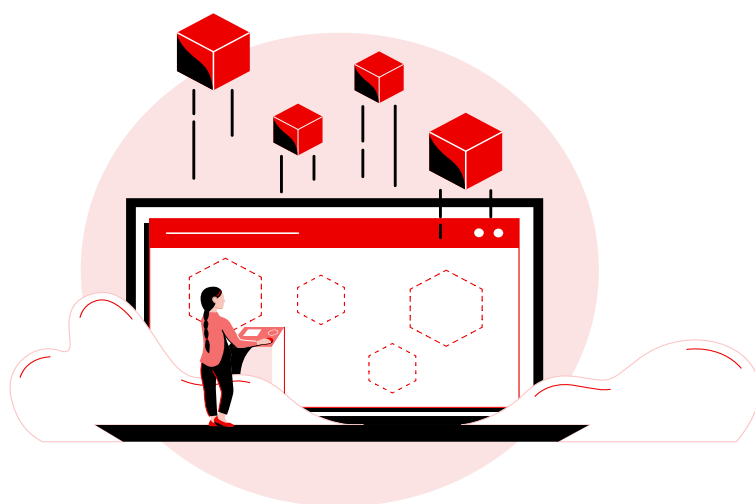
組織における生成 AI やエージェント型 AI の導入が進む中で、特にアプリケーションがシングルクエリのインタラクションから、AI エージェントが実行するマルチステップの継続的なタスクへと移行するにつれて、推論はコストとパフォーマンスの重要な要素となっています。



AI にとって重要な理由

推論は、ユーザーエクスペリエンス、アプリケーション・パフォーマンス、運用コストに直接影響を与えます。生成 AI およびエージェント型 AI のワークロードは多くの場合、データセンターからパブリッククラウド、エッジサイトまで、多数の環境にわたって迅速な応答、並列リクエスト、一貫したスループットを必要とします。

効率的な推論ランタイムは、GPU と CPU (中央処理装置) のコスト削減、インタラクティブなタスクのレイテンシーの改善、ツールを呼び出す AI エージェントをスケーリングするニーズのサポート、アプリケーション・プログラミング・インタフェース (API) の使用、マルチステップのワークフローの調整に役立ちます。また、推論を最適化することで、組織は予測可能なパフォーマンスを維持しながら、オンプレミスまたはプライベートクラウドで推論を機密データの近くで実行できるようになるため、ソブリン AI およびプライベート AI 戦略もサポートされます。



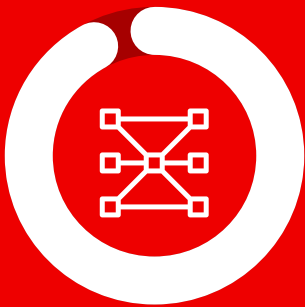
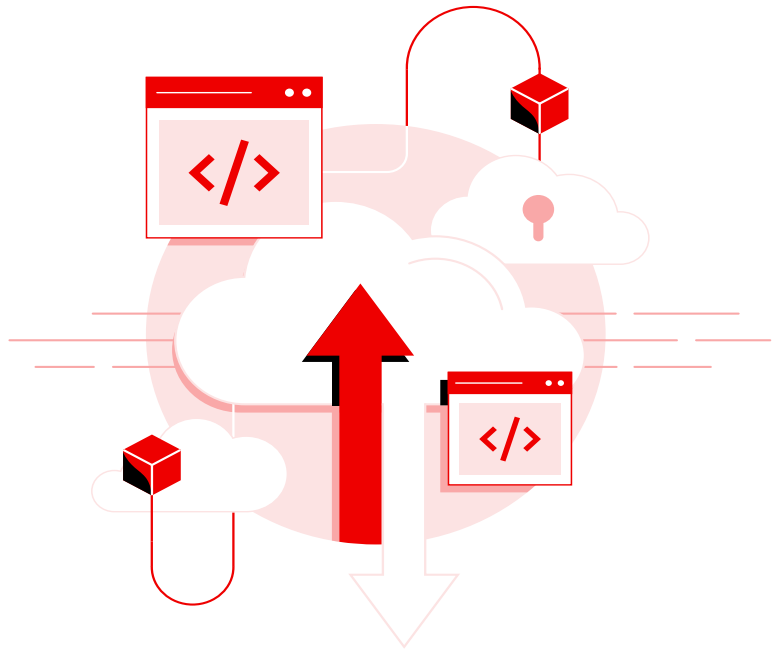


ベストプラクティスと推奨事項

LLM、マルチモーダルモデル、予測モデル、エージェント型ワークロードなど、モデルタイプとデプロイ環境に適した、最適化された推論ランタイムを選択しましょう。インタラクティブな LLM とエージェントベースの推論の予測不可能な需要に対応するには、水平方向にも垂直方向にも動的なスケーリングをサポートするランタイムとインフラストラクチャを優先します。

量子化、蒸留、モデル最適化などのアプローチに関する技術を使用するか、それらの専門知識を持つベンダーと提携して、コストを削減し、レイテンシーを改善します。こうした最適化を、広く採用されているテクノロジー（高スループットの LLM 推論用の vLLM など）や、先進的な分散推論フレームワーク（推論プロセスを分離して各フェーズを個別にスケーリングする llm-d など）と組み合わせましょう。

推論をコンテナ内にデプロイして依存関係をパッケージ化すれば、ハイブリッド環境間で一貫してスケーリングできます。推論エンドポイントをデータとアプリケーションの近くに配置することで、特にソブリン AI やプライベート AI のシナリオでは、データの移動を減らし、制御性を維持できます。そして、モデルのパフォーマンスを経時的に監視し、データの分布が変化したらバージョンを更新して、精度と信頼性を大規模に維持します。



90%

2026 年までに AI がデジタル・インフラストラクチャの予算およびテクノロジーの選択に影響を与える重要な要因になると考えている意思決定者の割合。³

³ IDC ホワイトペーパー、「AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025」、Document #US53418426、2025 年 10 月 (購入が必要)

AI の安全性



AI システムは、信頼性が高く予測可能な方法で、組織のポリシーに準拠して動作する必要があります。エンタープライズが実験からプロダクションへと移行するにつれて、AI の安全性が中心的な要件となります。提案を行うだけでなくアクションを実行できる生成 AI、エージェント型 AI、自律型ワークフローをデプロイする場合はなおさらです。



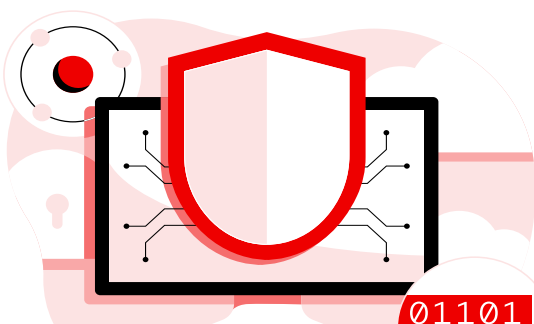
AI にとって重要な理由

安全性は、AI モデルとエージェントを定義された境界内に保ち、ビジネス、法律、倫理の要件を維持することに重点を置いています。不正確な出力、モデルドリフト、安全でないデータ処理、または意図しないアクションは、運用上の実質的なリスクを生じさせる可能性があります。また、生成 AI およびエージェント型システムからは、ハルシネーション、未承認のツールの実行、権限の昇格、マルチステップタスク間での一貫性のない推論など、安全性に関する新たな考慮事項が生じます。強力な安全プラクティスは、信頼の維持、機密データの保護、有害なアクションや元に戻すことができないアクションの防止に役立ちます。規制業界において安全管理は、ハイブリッド環境およびオンプレミス環境でのコンプライアンスと監査の対応に不可欠です。



ベストプラクティスと推奨事項

ポリシーベースの防護機能、コンテンツフィルター、エージェント型ワークフローのツール実行制御など、安全性に対する階層化されたアプローチを採用しましょう。モデルを定期的に検証およびテストして、ドリフトや精度の低下を検出します。機密性の高いワークロードをプライベートまたはオンプレミス環境で実行し、ソブリン AI およびプライベート AI 戦略に合わせて、データ露出とモデル動作の制御を維持します。モデル評価フレームワークを使用して、バイアス、堅牢性、信頼性を監視します。モデルとデータを強化するために、それらを標準コンテナ (OCI) 準拠のレジストリに格納してセキュアなサプライチェーンを提供するツールを使います。また、LLM 推論用の vLLM など、広く採用されているテクノロジーや、llm-d などの先進的な分散テクノロジーが、コストの削減と AI プロジェクトのデプロイの拡張に役立ちます。そして、モデル、データセット、ポリシーをバージョン管理し、文書化します。これにより、意思決定を追跡し、AI ライフサイクル全体にわたって一貫したガバナンスを管理できます。



AIのためのオープンで柔軟な基盤の構築



Red Hat AI Enterprise は、ハイブリッドクラウド環境全体で効率的かつコスト効果の高い AI モデル、エージェント、アプリケーションを開発およびデプロイするための統合 AI プラットフォームであり、Red Hat AI ポートフォリオの一部です。

AI モデルとアプリケーションのライフサイクルを統合し、エンタープライズグレードの機能を備えたすぐに使える開発環境を提供することで、運用効率の向上、提供の迅速化、リスクの低減を実現します。

このプラットフォームは Red Hat OpenShift を活用した、テスト済みかつサポート付きのフル AI スタックであり、相互運用性を強化し、ビジネス継続性を維持します。モデルチューニング、高性能推論、エージェント型 AI ワークフロー管理などのコア機能が含まれています。これにより、データロケーション要件を満たしながら、あらゆるモデルをサポートし、あらゆるハードウェアを使用し、どこにでもデプロイできる柔軟性が得られます。Red Hat AI Enterprise はさまざまなハイブリッド環境でサポートされているため、チームは自信を持って容量、GPU、将来の AI プロジェクトを計画できます。

Red Hat AI Enterprise には、Red Hat が IBM、NVIDIA、Google、AMD などのコラボレーションによって立ち上げたオープンソースの llm-d プロジェクトのテクノロジーが組み込まれています。llm-d は、推論のプリフィルフェーズとデコードフェーズを分離し、それぞれに異なるスケーリングを可能にすることでコスト効率を向上させます。推論対応のロードバランサーがトークンキューに基づいてリクエストをルーティングすることで、応答時間を短縮し、場合によってはプリフィルのワークロードを CPU に転送することもあります。



価値実現までの時間を短縮

事前構成済みのツール、自動デプロイ、組み込みの可観測性により、エンタープライズ対応の AI スタックを任意のインフラストラクチャにデプロイできます。つまり、開発者と AI エンジニアは、AI を活用したクラウドネイティブなエージェント型アプリケーションの構築と提供に集中できます。



運用効率の向上

コードのコミットからモデルのデプロイを通じた AI パイプラインワークフローの確立まで、ワークフローを効率化および自動化します。つまり、IT 運用チームはインテリジェントなリソース割り当てと統合ライフサイクル管理によって、一貫したパフォーマンスを実現し、既存のインフラストラクチャからより多くの価値を得ることができます。



リスクの低減

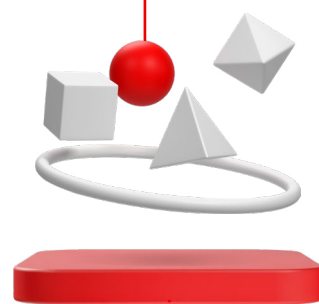
あらゆるモデル、あらゆるハードウェア、およびハイブリッドクラウド環境の相互運用性を強化する、テスト済みで完全にサポートされた統合 AI スタックが、エンタープライズ AI の導入リスクを軽減します。この基盤を使用すれば、データレジデンシーや規制要件に対処し、自信を持って AI を拡張できます。





図 2. 統合 AI プラットフォームのコンポーネント

認定 AI/ML パート ナーエコシステムで選 択肢と柔軟性を得る



AI ツールとテクノロジーの状況は急速に進化し続けており、IT 環境内で安定性と信頼性を維持しながら進歩に対応することが困難になっています。

Red Hat AI Enterprise は、NVIDIA、AMD、インテル、および AI テクノロジーパートナーとの連携を通じて、ハイブリッドクラウドで拡張可能なエンドツーエンドのエンタープライズ AI プラットフォームを実現し、より迅速なデプロイメント、効率性の向上、ハイブリッドクラウドのサポートを提供します。Red Hat の検証および認定プログラムによってハードウェアを最大限に活用することができ、ワークロード管理が最適化されるので GPU を効率的に使用でき、パフォーマンスとお客様の価値が最大化されます。

[Hugging Face エコシステム](#)と[モデルコンテキストプロトコル \(MCP\)](#) サーバーカタログに Red Hat が参加することで、お客様は、増え続ける検証済みモデルのライブラリや、Red Hat AI Enterprise と一貫して動作する事前統合済みのツールにアクセスできます。同時に、複数のアクセラレーター・プロバイダーとのパートナーシップにより、ハイブリッド環境全体で GPU や専用ハードウェアを活用できます。連携に信頼性があるため、ニーズに最適なパートナー、モデル、ツール、テクノロジーを自信を持って選択できます。それらはエキスパートのサービス、サポート、トレーニングによって支えられており、AI ワークフローの構築と拡張の成功を支援します。



成功事例



ターキッシュ エアラインズ

ターキッシュ エアラインズは Red Hat AI を使用して運用をモダナイズし、航空業界において AI を活用したイノベーションの先駆者となっています。オープンでスケーラブルな AI プラットフォームで標準化することで、同社はモデル開発を加速し、旅客サービスを向上させ、運用上の意思決定を効率化しました。これは、ハイブリッド AI が世界最大級の航空ネットワークの1つをどのように変革できるかを示すものです。

[詳細はこちら](#)

DenizBank

DenizBank は Red Hat AI を使用して、同行のデジタルバンキング・エコシステム全体で AI イノベーションを加速させています。オープンでスケーラブルなプラットフォームで AI インフラストラクチャをモダナイズすることで、同行は実験を迅速化し、モデルの信頼性を向上させ、よりスマートなカスタマーエクスペリエンスを提供しています。これは、ハイブリッド AI によって、金融機関が厳格なセキュリティポスチャとガバナンスを維持しながら、より迅速に行動できることの実証となっています。

[詳細はこちら](#)

AGESIC

ウルグアイのデジタル政府機関である AGESIC は、Red Hat AI を使用して 180 以上の公共機関で AI を標準化し、拡張しています。ハイブリッド AI プラットフォームは、MLOps プラクティスをサポートし、セキュリティを強化し、市民向けサービスを向上させる AI アプリケーションの構築、デプロイ、管理を支援します。

[詳細はこちら](#)



データを最大限に 活用する準備はでき ていますか？



AI はビジネスのほぼあらゆる側面を変革しています。Red Hat は、組織のビジネス目標をサポートするインテリジェント・アプリケーションの開発と提供を高速化する、本番利用に適した AI 環境の構築を支援します。

Red Hat AI Enterprise が AI 向け統合プラットフォームの構築にどのように役立つかをご覧ください。

