Red Hat
OpenShift AI

ハイブリッドクラウドにおける AI モデルのためのオープン・プラットフォーム

ハイライト

実績のあるアプリケーション・プラットフォーム基盤で AI 推論とエージェント型 AI を運用化および拡張します。

データサイエンティスト、データエンジニア、アプリケーション開発者、DevOps チームを強化する一貫したユーザーエクスペリエンスにより、チーム全体で AI/ML の運用効率を向上させます。

オンプレミス、クラウド、またはエッジで AI/ML ワークロードを構築、トレーニング、デプロイ、監視することで、ハイブリッドクラウドの柔軟性を実現します。

インテリジェント・アプリケーションと生成 AI の導入

人工知能 (AI)、機械学習 (ML)、ディープラーニング (DL) は、多種多様な企業や業界におけるアプリケーションのモダナイゼーションの取り組みに大きな影響を与えています。イノベーションを起こし、データから戦略的価値と新たな知見を導き出すことの必要性が、AI 対応のクラウドネイティブ・アプリケーション、MLOps および GenAIOps 手法の使用拡大を促進しています。このすばらしい新世界における課題は複雑なものになる可能性があります。プロダクションに移行する際のモデルコストの急増、複雑なカスタマイズ、デプロイの厳格な制約、イノベーションのペースに後れを取らないために必要な運用などです。企業に必要なのは、推論コストを削減し、スケーリングと監視を単純化し、絶え間ない変化に適応するソリューションです。

Red Hat® AI は、ハイブリッドクラウド環境全体でエンタープライズ AI ソリューションの開発とデプロイを加速させます。AI/ML ライフサイクル全体を管理する包括的なプラットフォームとして機能し、MLOps 機能と GenAIOps 機能を提供します。Red Hat AI は特に次の 4 つの主な柱に注力しています。

- 高速で柔軟かつ効率的な推論によって効率を向上させる
- モデルをデータに接続する作業を単純化する
- エージェント型 AI のイノベーションを加速する
- 柔軟性と一貫性を維持しながら AI をハイブリッドクラウド全体に拡張する

業界をリードするハイブリッドクラウド・アプリケーション・プラットフォームである Red Hat OpenShift 上に構築された Red Hat OpenShift® AI は、Red Hat AI ポートフォリオの主力製品です。この AI プラットフォームは、AI エンジニア、データサイエンティスト、開発者に、生成モデルや予測モデル、AI 対応アプリケーションを大規模に構築およびデプロイするための強力な AI/ML 基盤を提供します。組織にとっては、選択したツールによる実験、コラボレーション、市場投入時間の短縮が可能になり、これらすべてを 1 つの共通プラットフォーム内で行うことができます。Red Hat OpenShift AI は、データサイエンティストや開発者が求めるセルフサービス環境と、エンタープライズ IT が求める信頼性を兼ね備えています。

迅速な開発、トレーニング、テスト、デプロイ

Red Hat OpenShift AI は、オープンソース・テクノロジーを使用して構築された柔軟でスケーラブルな MLOps プラットフォームであり、実験、モデル提供、革新的なアプリケーションの実現のための、信頼性と一貫性に優れた運用機能を提供します。OpenShift AI は AI 対応アプリケーションの提供を加速し、組織が初期のパイロットから堅牢な運用デプロイへと、高い制御性でより迅速に移行できるよう支援します。

このプラットフォームは、予測および生成 AI モデルの構築、トレーニング、チューニング、デプロイ、監視のためのツールを備えた統合ユーザーインターフェース (UI) エクスペリエンスを提供します。ソブリン AI とプライベート AI 用の制御および保護されたフットプリントを提供することに重点を置きながら、モデルをハイブリッドクラウド環境にデプロイできます。このアプローチにより、機密データと AI モデルを、指定された地理的境界または組織境界内に維持し、厳格な規制およびコンプライアンス要件を満たすことができます。

Red Hat
OpenShift AI

生成 AI アナリストの予測

「組織がワークロードとデータ要件をハイブリッド・インフラストラクチャの選択に適合させようとしているため、AI は 2026 年においてデジタル・インフラストラクチャの予算を左右する非常に重要な要素になると予想されています。意思決定者の 90% が、2026 年までに AI がデジタル・インフラストラクチャの予算とテクノロジーの選択の重要な推進力になると考えています」[1]

AI 導入の単純化

Red Hat OpenShift AI は Red Hat OpenShift のアドオンとして、オープンソース・コミュニティと堅牢な AI エコシステムを組み合わせることにより、AI 導入を広めて AI イニシアチブの信頼性を高めるように設計されたプラットフォームを提供します。これにより、組織に適した AI/ML テクノロジーを選択するための柔軟性と自由度が向上します。ユーザーは、予測モデルを構築するか外部の生成 AI モデルから始めて、プラットフォームで提供される複数のモデルサーバーの 1 つを使用し、検索拡張生成 (RAG) でモデルを強化することができます。このプラットフォームは、vLLM で効率的に動作する、最適化および検証済みのサードパーティモデル (Llama、Mistral、DeepSeek、Granite など) に迅速にアクセスできます。これらのモデルは、Hugging Face の Red Hat AI リポジトリで入手可能です。カタログを使用してこれらのモデルを調べ、独自のモデルを追加することができます。

チーム間の運用の一貫性を向上させる

Red Hat OpenShift AI が提供する一貫したユーザーエクスペリエンスにより、データサイエンティスト、AI エンジニア、開発者、DevOps チームは効率的に協力してタイムリーに AI ソリューションを提供できます。コラボレーティブなワークフローへのセルフサービスアクセス、グラフィック・プロセッシング・ユニット (GPU) のアクセラレーション、最適化された運用が実現し、ハイブリッドクラウド環境やネットワークエッジ全体で大規模な AI ソリューションを一貫して提供することができます。

IT 運用チームは、優れたガバナンスとセキュリティを実現しながら、少ない労力でスケールアップやスケールダウンが可能な実績あるプラットフォーム上で、単純化された構成とより多くの自動化ワークフローを活用できます。

ハイブリッドクラウドの柔軟性を獲得

Red Hat OpenShift AI は、クラウド、オンプレミスのデータセンター、エアギャップ環境など、さまざまな環境で AI/ML ワークロードをトレーニング、デプロイ、監視し、規制、セキュリティ、およびデータの要件を満たすことができます。このプラットフォームは、NVIDIA、AMD、インテルなどのベンダーが提供する複数の AI アクセラレーターと互換性があります。この機能は拡張して GPU-as-a-Service 環境を構築することが可能です。これにより、組織は GPU リソースを一元的に管理、パーティション分割、スケジュールし、使用状況に関する詳細な可観測性も得ることができます。

生成 AI とエージェント型 AI

生成 AI プロジェクトでは、プラットフォームエンジニア向けのダッシュボード・エクスペリエンスである AI hub (開発者プレビュー) などのコンポーネントを通じて専用のユーザーエクスペリエンスが提供され、カタログ、レジストリ、およびモデルのデプロイが統合されます。これを使用してモデルと MCP サーバーをセットアップおよびデプロイできます。gen AI studio (開発者プレビュー) は AI アセットエンドポイントと AI プレイグラウンドを提供します。これによって AI エンジニアやアプリケーション開発者は、デプロイされたモデルや MCP サーバーにアクセスし、実験、比較、評価、テストを行うことができます。

OpenShift AI は、統合された API レイヤーと柔軟でスケーラブルな基盤を提供することで、エージェント型 AI を加速させます。Llama Stack API と MCP (テクニカルプレビュー) のサポートには、Llama Stack API のエンタープライズグレードの実装が含まれており、さまざまな AI 機能への単一の標準化されたエントリーポイントとなります。

また、追加のツールとして LLM 評価 (LM Eval) と LLM ベンチマークが含まれており、これらは実際の推論デプロイを支援します。LLM compressor は、組織のカスタムモデルのサイズを縮小するアルゴリズムを提供します。このアルゴリズムでは、Hugging Face の Red Hat AI リポジトリにある、検証済みで最適化されたモデルを作成するために Red Hat が使用しているのと同様の方法が使用されています。

1 IDC Tech Supplier、「AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025」、Doc #US53418426、2025 年 10 月。(クライアントのログインが必要)

Red Hat
OpenShift AI

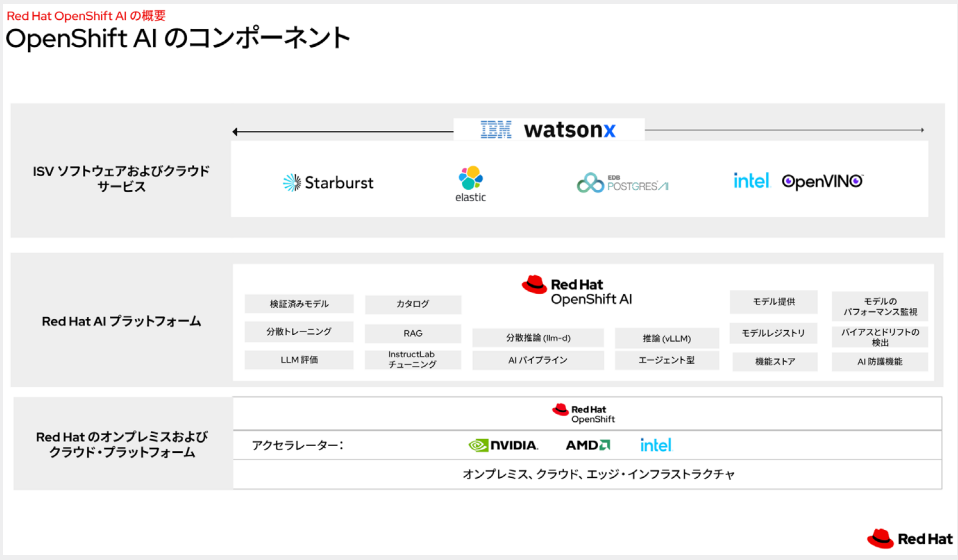

図 1. Red Hat OpenShift AI とオプションのパートナーコンポーネント

強固な基盤を実現するために Red Hat OpenShift AI が提供するその他の中心的なツールと機能には、以下のものがあります。

- **モデルの構築とカスタマイズ：**すぐに使える安全に構築されたノートブックイメージと、一般的な Python ライブラリが提供され、データサイエンティストは JupyterLab UI を使用して探索的データサイエンスを行うことができます。生成 AI プロジェクトの場合、OpenShift AI によって検索拡張生成 (RAG) と分散型 InstructLab トレーニングが可能になり、スキルや知識をより効率的に生成 AI モデルに提供するためのモデルアライメントツールが提供されます。
- **モデルの提供：**Red Hat OpenShift AI は、モデル提供のコアエンジンとして KServe を使用するさまざまなフレームワークを提供し、予測機械学習や基盤モデルのプロダクション環境へのデプロイを単純化します。最大限のスケーラビリティを必要とする LLM 向けに、OpenShift AI は vLLM ランタイムによる並列処理を提供します。llm-d は、パイプラインをモジュール式サービスに分離することで LLM 推論を最適化するためのフレームワークを提供し、スマートな自動スケーリングと効率的なリクエストルーティングをサポートします。
- **AI パイプライン：**Red Hat OpenShift AI はパイプライン・コンポーネントを提供し、AI タスクをパイプラインにオーケストレーションしたり、グラフィカルなフロントエンドを使用してパイプラインを構築できます。組織は、データの準備、モデルの構築、モデルの提供といったプロセスを連鎖させることができます。
- **モデル監視：**Red Hat OpenShift AI は、運用指向のユーザーがモデルサーバーとデプロイされたモデルの運用とパフォーマンスのメトリクスを監視するのに役立ちます。ユーザーは、すぐに使える視覚化機能にアクセスしてパフォーマンスや運用のメトリクスを確認したり、他の可観測性サービスとデータを統合したりすることができます。
- **ワークロードの分散：**分散ワークロードにより、モデルのトレーニング、チューニング、提供とともにデータ処理を高速化できます。この機能は、最適なノード利用とともに、ジョブの実行の優先順位付けと分散をサポートします。高度な GPU サポートは、基盤モデルのワークロード要求に対応するのに役立ちます。

- **AI 防護機能、バイアスとドリフトの検出：**Red Hat OpenShift AI は、データサイエンティストや AI エンジニアがトレーニングデータに基づいてモデルが公平で偏りがないかを監視するだけでなく、実際のデプロイ時における公平性も監視できるツールを提供します。AI 防護機能は、重要な安全性制御を実装するカスタマイズ可能なフレームワークを提供し、プロダクションでの使用におけるモデルの透明性、公平性、信頼性を確保します。ドリフト検出ツールには、デプロイされた ML モデルの入力データ分布が含まれており、モデル推論に使用されるライブデータがモデルのトレーニングに使用されたデータから大きく逸脱したときに、それを検出できます。
- **カタログとレジストリ：**Red Hat OpenShift AI は、内部モデルカタログと厳選されたカタログを提供し、プラットフォームエンジニアはここから最適化された生成 AI モデルを見つけ、比較、評価できます。また、データサイエンティストや AI エンジニアが予測型 AI モデルおよび生成 AI モデル、メタデータ、モデルアーティファクトを共有、変更、デプロイ、追跡するのに役立つ中央レジストリも提供します。
- **機能ストア：**ML モデルのためのクリーンで明確に定義されたデータ機能を管理し、パフォーマンスを向上させ、ワークフローを高速化します。

AI ライフサイクル全体のためのツール

Red Hat OpenShift は、組織がモデルを正常にトレーニングおよびデプロイしてプロダクションに移行できるようにするためのサービスとソフトウェアを提供します (図 2 を参照)。

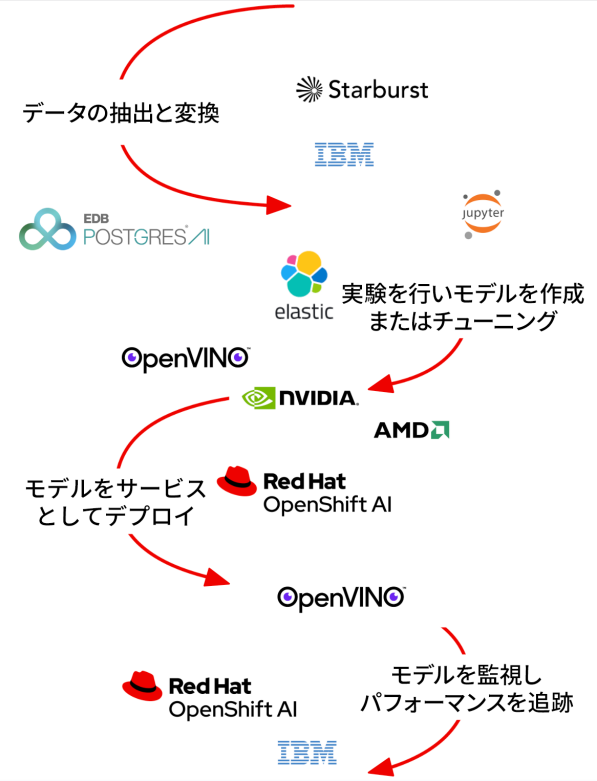

図 2. 典型的なモデル運用化ライフサイクル

Red Hat OpenShift AI

Red Hat OpenShift AI ダッシュボードは、すべてのアプリケーションとドキュメントを検出してそれらにアクセスするための中心的な場所を提供し、導入を容易にします。スマート・スタート・チュートリアルは、一般的なコンポーネントと統合パートナーソフトウェアに関するベストプラクティスのガイダンスを提供します。ダッシュボードから直接利用できるため、データサイエンティストがより迅速に学習し、開始することができます。以下のセクションでは、Red Hat OpenShift AI と統合されたテクノロジー・パートナー・ツールについて説明します。ツールによっては、テクノロジーパートナーからの追加ライセンスが必要になります。

Starburst

Starburst

Starburst は、組織のチームがビジネス機能を改善するためにデータを迅速かつ容易に活用できるようにすることで、分析を加速します。Starburst はセルフマネージド製品またはフルマネージドサービスとして提供され、データアクセスを民主化し、データ利用者に包括的な知見を提供します。Starburst は、プレミア超並列処理 (MPP) 構造化クエリ言語 (SQL) エンジンであるオープンソースの Trino (旧称 PrestoSQL) に基づいて構築されています。Trino の専門家によって構築および運用されている Starburst を使用すると、組織のデータを移動することなく、さまざまなデータセットがどこにあっても自由に調べることができます。

Starburst は、Red Hat OpenShift が提供するスケーラブルなクラウドストレージとコンピューティングのサービスと統合し、すべてのエンタープライズデータをクエリする、安定したセキュリティ重視の効率的かつコスト効率の高い方法を実現します。そのメリットには以下のようなものがあります。

- **自動化：**Starburst および Red Hat OpenShift Operator は、クラスタの自動設定、自動チューニング、自動管理を提供します。
- **高可用性 (HA) と段階的なスケールダウン：**Red Hat OpenShift ロードバランサーは、Trino コーディネーターなどのサービスを常時オンの状態に保つことができます。
- **弾力性に優れたスケーラビリティ：**Red Hat OpenShift は、クエリの負荷に基づいて Trino ワーカークラスタを自動的にスケーリングできます。

Hewlett Packard Enterprise

HPE Machine Learning Data Management Software

組織は、ノートパソコンでの実験から重要なエンタープライズデプロイまで、あらゆるものを容易にするデータ管理ソリューションを必要としています。HPE Machine Learning Data Management Software (旧称 Pachyderm) により、データサイエンスチームは、自動的なデータバージョン管理が保証するデータリネージュによって、コンテナ化されたデータ駆動型 ML パイプラインを構築および拡張できます。現実世界のデータサイエンスの問題を解決するように設計されている HPE Machine Learning Data Management Software により、チームは再現性を保証しながら ML ライフサイクルを自動化および拡張できるデータ基盤を手に入れることができます。HPE Machine Learning Data Management Software には、非構造化データからデータウェアハウス、自然言語処理、動画および画像の抽出、変換、ロード (ETL)、金融サービス、ライフサイエンスなどさまざまなユースケースがあり、以下を提供します。

- データ変更を追跡するための高性能の方法を提供する、自動化されたデータバージョン管理
- コンピューティングコストを削減しながらデータ処理を高速化する、データ駆動型のコンテナ化パイプライン
- ML ライフサイクルのアクティビティとアセットに固定レコードを提供する、不変のデータリネージュ
- 有向非巡回グラフ (DAG) の直感的な視覚化を実現し、デバッグと再現性を支援するコンソール
- バージョン管理データへのポイント・アンド・クリック・インタフェース用の JupyterLab Mount Extension による Jupyter ノートブックのサポート
- 組織内のさまざまなチームにまたがって HPE Machine Learning Data Management Software を大規模にデプロイおよび管理するための堅牢なツールによるエンタープライズ管理

Red Hat OpenShift AI

NVIDIA による AI ソリューションのデプロイの加速

AI/ML アプリケーションがビジネスの成功にとってますます重要になるにつれ、複雑なワークロードを処理し、ハードウェアの利用を最適化し、スケーラビリティを提供できるプラットフォームが企業にとって必要となっています。スケーラブルなデータ処理、データ分析、ML トレーニング、推論はすべて、リソースを大量に消費する計算タスクです。NVIDIA ソフトウェアにより、GPU の並列処理機能を利用して、エンドツーエンドのデータサイエンスのあらゆる側面を高速化することが可能になります。

NVIDIA NIM は、Red Hat OpenShift 環境内の NVIDIA GPU の管理とパフォーマンスを強化し、AI アプリケーションが NVIDIA の AI ソフトウェアとハードウェアの可能性をフルに活用できるようにします。NVIDIA NIM と Red Hat OpenShift AI の統合により、リソース割り当ての改善、効率の向上、AI ワークロード実行の生産性向上が実現します。

Intel OpenVINO ツールキット

インテル OpenVINO ツールキットは、インテル・プラットフォーム上での高性能 DL 推論アプリケーションの開発とデプロイを加速します。このツールキットを使用すると、OpenVINO 開発ツールのエコシステムを使用して、ニューラル・ネットワーク・モデルの導入、最適化、チューニングを実質的に行い、包括的な AI 推論を実行することができます。

- **モデル：**ソフトウェア開発者は独自の DL モデルを柔軟に使用できます。市場投入時間を短縮するために、インテルと Hugging Face のコラボレーションによる OpenVINO ツールキットを通じて利用可能な、事前トレーニング済みで最適化されたモデルを使用することもできます。OpenVINO は、PyTorch、ONNX、TensorFlow などの一般的なモデル形式をサポートしています。
- **最適化：**OpenVINO ツールキットは、利便性とパフォーマンスを向上させるために複数の方法でモデルを変換することができ、ソフトウェア開発者がより高速かつ効率的な AI モデルの実行を実現できるよう支援します。開発者はモデルの変換を省略し、PyTorch、ONNX、TensorFlow、TensorFlow Lite、JAX、PaddlePaddle の形式から直接推論を実行することができます。OpenVINO IR への変換により最適なパフォーマンスが提供され、OpenVINO のニューラルネットワーク圧縮フレームワークで利用できる重み圧縮および量子化機能を使用してさらに最適化できます。また、これらの機能により、ストレージとランタイムのフットプリントも削減されます。
- **デプロイ：**OpenVINO ランタイム推論エンジンは、アプリケーションに統合して推論プロセスを高速化するように設計されているアプリケーション・プログラミング・インタフェース (API) です。「一度書けばどこにでもデプロイできる」アプローチにより、CPU (中央処理装置)、GPU、NPU、FPGA など、さまざまなインテル製ハードウェア上で推論タスクを効率的に実行できます。OpenVINO GenAI 拡張ライブラリは、生成 AI ワークロードのデプロイを単純化し、多くの場合、必要なコードはわずか 3 - 5 行に削減されます。OpenVINO Model Server は、エージェント型およびモデル提供型のシナリオ向けの複数の機能を提供し、開発の労力をさらに削減します。

EDB

EDB Postgres AI は、トランザクション、分析、AI ワークロードを処理するために設計された強力でインテリジェントなプラットフォームであり、データがオンプレミスにあるかクラウド環境にあるかに関係なく、比類のない柔軟性を提供します。エンタープライズ Postgres データベース・ソリューションの世界的リーダーである EDB は、データおよび AI のオープンでエンタープライズグレードのソブリン・プラットフォームを提供し、AI プロジェクトをプロダクションに投入するスピードを 3 倍にします。Red Hat OpenShift AI と統合された EDB Postgres AI により、ユーザーは検索拡張生成 (RAG) のための堅牢な AI ナレッジベースを構築し、AI データ、モデル、アプリケーションを、どこにでもデプロイ可能なフルスタックのソブリン AI プラットフォームに統合することができます。このように運用のコアデータを AI 対応アセットに変換することで効率が最大 30% 向上し、非構造化データを含めたプライベートデータの使用を単純化して、モデルの出力を組織のナレッジベースにグラウンディングすることができます。

Red Hat OpenShift AI

Elastic

Elastic Search AI Platform (ELK Stack 上に構築[2]) は、検索の精度と AI のインテリジェンスの両方を兼ね備えています。ユーザーは LLM のプロトタイプ作成と統合をより迅速に行うことができ、スケーラブルでコスト効率の高いアプリケーションを構築するために生成 AI を活用することができます。Elastic Search AI Platform により、ユーザーは革新的な検索拡張生成 (RAG) アプリケーションを構築し、可観測性の問題をプロアクティブに解決し、複雑なセキュリティ脅威に対処することができます。Elasticsearch は、オンプレミス、好みのクラウドプロバイダー、エアギャップ環境など、お客様のアプリケーションのある場所にデプロイすることができます。

Elastic は単一の簡単な API 呼び出しによって、Red Hat OpenShift AI、Hugging Face、Cohere、OpenAI などのエコシステムからの組み込みモデルと統合できます。このアプローチにより、RAG ワークロードのハイブリッド推論を管理するためのクリーンなコードが実現します。これには以下のような機能が含まれます。

- 多様なデータセットを検索レイヤーに取り込むためのチャンク化、コネクター、Web クローラー
- 多言語ベクトル検索を可能にする、Elastic Learned Sparse EncodeR (ELSER)、組み込みの ML モデル、E5 組み込みモデルを使用したセマンティック検索
- 文書およびフィールドレベルのセキュリティによる、組織のロールベースのアクセス制御 (RBAC) にマッピングされた権限と資格の実装

Elastic Search AI Platform では、開発者の世界的なコミュニティの一員として、インスピレーションやサポートを得ることができます。Elastic のコミュニティについては Slack、ディスカッション・フォーラム、ソーシャルメディアでご覧いただけます。

まとめ

Red Hat OpenShift AI を使用することで実験とコラボレーションが可能になり、結果的に AI を活用したアプリケーションの導入を加速できます。データサイエンティストや AI エンジニアは、Red Hat OpenShift AI を使用して、モデルを構築してハイブリッドクラウド全体にデプロイできる柔軟性を得ることができます。IT 運用部門やプラットフォームエンジニアは MLOps および GenAIOps 機能からメリットを得られるため、モデルをより迅速にプロダクションにデプロイすることができます。開発者、AI エンジニア、データサイエンティスト向けのセルフサービス (GPU へのアクセスを含む) により、エンタープライズ IT がすでに使用し完全に信頼しているアプリケーション・プラットフォームでのイノベーションが加速します。Red Hat OpenShift AI は今後も、堅牢なパートナーエコシステムに支えられながら、包括的で信頼できる一貫したプラットフォームを提供し、効率的な推論、エージェント型 AI、スケーラブルなハイブリッドクラウド運用において独自の差別化要因をもたらします。

詳細はこちら

Red Hat OpenShift AI にアクセスして、今すぐ始めましょう。

Red Hat について

Red Hat は、受賞歴のあるサポート、トレーニング、コンサルティング・サービスをお客様に提供し、複数の環境にわたる標準化、クラウドネイティブ・アプリケーションの開発、複雑な環境の統合、自動化、セキュリティ保護、運用管理を支援します。

fb.com/RedHatJapan
twitter.com/RedHatJapan
linkedin.com/company/red-hat

jp.redhat.com
#2876428_1125

2 ELK スタックは、Elasticsearch、Kibana、Beats、Logstash で構成されています。

アジア太平洋
+65 6490 4200
apac@redhat.com

オーストラリア
1 800 733 428

インド
+91 22 3987 8888

インドネシア
001 803 440 224

日本
03 4590 7472

韓国
080 708 0880

マレーシア
1 800 812 678

ニュージーランド
0800 450 503

シンガポール
800 448 1430

中国
800 810 2100

香港
800 901 222

台湾
0800 666 052