Red Hat

Production AI for private and hybrid clouds

Datasheet

Highlights

Simplify the adoption of AI into your business, increase AI adoption, and provide flexibility in AI initiatives.

Establish AI/ML operational consistency across teams with a consistent user experience that empowers AI engineers, data scientists, data engineers, and DevOps teams to collaborate effectively.

Offers flexibility and consistency to build, deploy, and manage AI at scale across any hardware and hybrid cloud, addressing data constraints, privacy, security, and cost control.

Develop, train and deploy AI models and applications

Red Hat® OpenShift® AI is an MLOps platform that allows you to develop, train, and deploy AI models and applications at scale across private and hybrid cloud environments. OpenShift AI offers organizations an efficient way to deploy an integrated set of common open source and third-party tools to perform both generative AI (gen AI) and predictive AI and machine learning (AI/ML) modeling. Adopters gain a collaborative open source toolset and platform for building experimental models and serving these models to production environments in a container-ready format, consistently, across public and private cloud, on-premise, and edge environments.

As a key component of Red Hat AI, OpenShift AI provides IT operations and platform engineers a simple to manage, scalable, and security-focused environment. For data scientists and AI engineers, it provides a comprehensive, unified platform for development and deployment of AI solutions at scale.

OpenShift AI supports gen AI foundation models, letting you fine tune and serve with your private data. Workloads can be distributed across Red Hat OpenShift clusters, independent of their location. The platform is integrated with and layered on Red Hat OpenShift, simplifying AI hardware acceleration and supporting central processing unit (CPU) and graphic processing unit (GPU)-based hardware infrastructure, including NVIDIA and AMD GPUs and Intel XPUs, whether on premise or in the sovereign or public cloud.

Table 1. Features and benefits of Red Hat OpenShift AI

Features	Benefits
Model development and customization	An interactive JupyterLab interface with AI/ML libraries and workbenches. Integrates data ingestion, synthetic data generation, InstructLab toolkit, and Retrieval Augmented Generation (RAG) for private data connection.
Model training and experimentation	Organizes development files and artifacts. Supports distributed workloads for efficient training and tuning. Features experiment tracking and simplified hardware allocation.
Intelligent GPU and hardware speed	Self-service GPU access is available. Offers intelligent GPU use for workload scheduling, quota management, priority access and visibility of use through hardware profiles.
AI pipelines	Can automate model delivery and testing. Pipelines are versioned, tracked and managed to reduce user error and

	simplify experimentation and production workflows.
Optimized model serving	Serves models from various providers and frameworks via a virtual large language model (vLLM), optimized for high throughput and low latency. The llm-d distributed inference framework supports predictable and scalable performance and efficient resource management. Includes LLM compressor and access to common, optimized and validated gen AI models.
Agentic AI and gen AI user interfaces (UIs)	Speeds agentic AI workflows with core platform services. A unified application programming interface (API) layer (MCP and Llama Stack API) and dedicated dashboard experience (AI hub and gen AI studio).
Model observability and governance	Common open source tooling for lifecycle management, performance, and management. Tracks metrics, including performance, data drift and bias detection and AI guardrails or inference. Offers LLM evaluation (LM Eval) and LLM benchmarking (GuideLLM) to assist real world inference deployments.
Catalog and registry	Centralized management for predictive and gen AI models and MCP servers and their metadata, and artifacts.
Feature store	A UI for managing clean, well-defined data features for ML models, enhancing performance and accelerating workflows.
Models-as-a-service	Allows AI engineers to use models via a managed, built-in API gateway for self-service access and usage tracking (developer preview feature).
Disconnected environments and edge	Supports disconnected and air-gapped clusters for security and regulatory compliance.

In addition to the capabilities of OpenShift AI, integrated partner products include:

- Starburst for distributed data access across diverse data sets.
- HPE for data lineage and versioning.
- NVIDIA for performance management of GPUs.
- AMD for GPU acceleration.
- Intel for high performance inference on Intel hardware.
- Elastic and EDB for vector databases with Retrieval Augmented Generation (RAG) applications.

Next steps

Learn more about Red Hat OpenShift AI and watch the informative video.

facebook.com/redhatinc
twitter.com/RedHat
linkedin.com/company/red-

About Red Hat
Red Hat is the world’s leading provider of enterprise open source software solutions, using a community-powered approach to deliver reliable and high-performing Linux, hybrid cloud, container, and Kubernetes technologies. Red Hat helps customers develop cloud-native applications, integrate existing and new IT applications, and automate and manage complex environments. A trusted adviser to the Fortune 500, Red Hat provides award-winning support, training, and consulting services that bring the benefits of open

hat	innovation to any industry. Red Hat is a connective hub in a global network of enterprises, partners, and communities, helping organizations grow, transform, and prepare for the digital future.
	North America 1–888–REDHAT1 www.redhat.com **Europe, Middle East, And Africa** 00800 7334 2835 europe@redhat.com **Asia Pacific** +65 6490 4200 apac@redhat.com **Latin America** +54 11 4329 7300 info-latam@redhat.com
	Copyright © 2025 Red Hat. Red Hat, the Red Hat logo, and OpenShift are trademarks or registered trademarks of Red Hat or its subsidiaries in the United States and other countries.