Red Hat
OpenShift AI

An open platform for AI models in the hybrid cloud

Highlights

Accelerate AI innovation and reduce the operational cost of developing and providing AI solutions.

Advance AI/ML operational efficiency across teams with a consistent user experience that empowers data scientists, AI engineers, application developers, and DevOps teams.

Gain hybrid cloud flexibility by building, training, deploying, and observing AI/ML workloads on premise, in a cloud, at the edge, or in disconnected environments.

Operationalize and scale AI on a proven foundation

Predictive AI, gen AI, and agentic AI are having a profound influence on application modernization efforts across diverse businesses and industries. The need to innovate and derive strategic insights from data is expanding the use of AI-powered, cloud-native applications; machine learning operations (MLOps); generative AI operations (GenAIOps); and agent operations (AgentOps) methodologies. This brave new world offers complex challenges, from soaring model costs in production and inflexible deployment rules to the operational demands of safely governing, monitoring, and scaling autonomous agent systems. Enterprises require solutions that lower inference costs, simplify scaling, deliver end-to-end observability and governance, and adapt to constant change.

Red Hat® AI speeds AI innovation and reduces the operational cost of developing and deploying AI solutions across hybrid cloud environments. It provides cost-effective solutions with optimized models and efficient inference, simplifies integration with private data, and delivers agentic AI on a scalable, flexible platform.

Red Hat OpenShift® AI—built on Red Hat OpenShift, a leading hybrid cloud application platform—is a key product offering in the Red Hat AI portfolio. The AI platform gives AI engineers, data scientists, and developers a reliable AI/ML foundation for building and deploying predictive models and gen AI-powered applications at scale. Organizations can experiment with a choice of tools, collaborate, and speed time to market with 1 common platform. Red Hat OpenShift AI combines the self-service environment that data scientists and developers want with the confidence and governance that enterprise IT demands.

Rapidly develop, train, test, and deploy

Red Hat OpenShift AI is an MLOps and GenAIOps platform built with open source technologies, providing trusted and operationally consistent capabilities for teams to experiment, serve models, and build innovative applications. OpenShift AI accelerates the delivery of AI-enabled applications, helping organizations move from early pilots into operationally reliable deployments with greater speed and control.

The platform offers an integrated user interface (UI) experience with tooling for building, training, tuning, deploying, and managing predictive and gen AI models. Organizations can deploy models to hybrid cloud environments with a specific emphasis on providing a controlled and protected environment for sovereign and private AI. This approach ensures that sensitive data and AI models remain within designated geographic or organizational boundaries, meeting strict regulatory and compliance requirements.

Red Hat OpenShift AI

Gen AI Analyst forecast

AI is expected to be an important factor in digital infrastructure budgets in 2026 as organizations aim to match workload and data requirements to hybrid infrastructure choices.

90% of decision makers believe AI will be an important driver of their digital infrastructure budget and technology choices through 2026.[1]

Simplify AI adoption

As an add-on to Red Hat OpenShift, OpenShift AI provides a platform designed to increase AI adoption and trust by combining open source communities with a reliable AI ecosystem. This offers increased flexibility and freedom to select the right AI/ML technology for each organization. Users can build their predictive models or start with an external gen AI model, then enhance it with retrieval-augmented generation (RAG) using 1 of several model servers provided in the platform. The platform offers access to optimized and validated third-party models—such as Llama, Mistral, DeepSeek, Qwen, Kimi and Granite—that run efficiently on vLLM (available on the Red Hat AI repository on Hugging Face). The catalog lets users explore these models and add their own. The OpenShift AI dashboard provides a central place to discover and access all applications and documentation, which simplifies adoption.

Ensure operational consistency across teams

OpenShift AI provides a consistent user experience that empowers data scientists, AI engineers, developers, and DevOps teams to collaborate effectively to deliver timely AI solutions. It offers self-service access to collaborative workflows, graphics processing unit (GPU) acceleration, and streamlined operations, providing a consistent delivery of AI solutions at scale across hybrid cloud environments and at the network edge.

IT operations benefit from simplified configurations and more automated workflows on a proven platform that can scale up or down with low effort, while providing reliable governance and security controls.

Gain hybrid cloud flexibility

With OpenShift AI, organizations can train, deploy, and manage AI/ML workloads across various clouds, on-premise datacenters, or air-gapped environments to meet regulatory, security, and data requirements. The platform is compatible with multiple AI accelerators from vendors like NVIDIA, AMD, Intel, IBM, Google, and Amazon Web Services (AWS). This capability can expand to a GPU-as-a-Service environment, so organizations can centrally manage, partition, and schedule GPU resources, while also providing detailed observability into their use.

Gen AI and agentic AI

For gen AI projects, organizations can get dedicated user experiences through components like AI hub, which provides a dashboard experience for platform engineers. AI hub consolidates catalog, registry, and model deployments, allowing teams to set up and deploy models, as well as discover and deploy verified Model Context Protocol (MCP) servers (developer preview). The experience also features Open Container Initiative (OCI)-compliant storage, artifact signing (tech preview), Hugging Face integration, and embedded performance insights to help govern how agents interact with internal systems.

Gen AI studio provides a hands-on environment to discover and interact with models, experiment in a playground environment, tune hyperparameters, and quickly prototype gen AI applications. It speeds development by adding side-by-side chat comparisons (tech preview), centralized prompt management and versioning, vector store integrations (tech preview), and an embedded MLflow user UI for end-to-end agentic traceability (tech preview).

1 IDC Tech Supplier. "AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025." Doc #US53418426, October 2025. (login required)

Red Hat OpenShift AI

OpenShift AI accelerates agentic AI by providing a unified application programming interface (API) layer and a flexible, scalable foundation. The platform's MCP support acts as a standardized integration layer to govern how agents interact with external tools. OpenShift AI operationalizes agentic workflows with AgentOps capabilities, including embedded MLflow for end-to-end agent traceability and tool-use logging, EvalHub for scientifically benchmarking and scoring AI agents (tech preview), and automated adversarial vulnerability scanning to proactively catch prompt injections and ensure agent safety before deployment (tech preview).

Additional tools include EvalHub (tech preview), which advances large language model (LLM) evaluation and benchmarking to support evaluation-driven development. EvalHub provides a unified interface to scientifically validate the performance, safety, and adversarial resilience of models, RAG pipelines, and agentic workflows prior to real-world deployment. To assist with efficient inference, tools like LLM Compressor and speculative decoding provide algorithms to reduce the size of custom models and increase response speeds by 2–3 times without quality loss, using methods similar to the one Red Hat uses to create validated and optimized models.

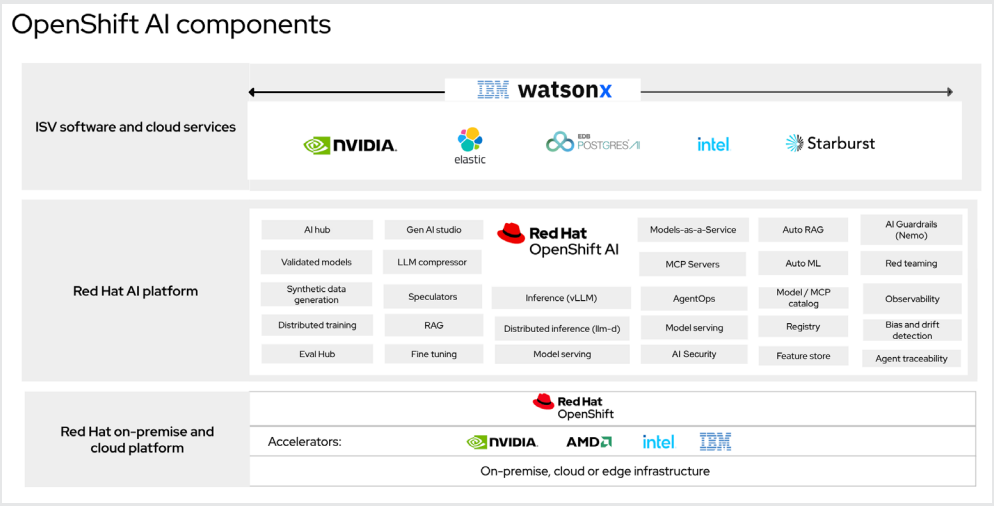

Figure 1. OpenShift AI and optional partner components

Several other core tools and capabilities provided with OpenShift AI offer a solid foundation.

Table 1. Features and benefits of OpenShift AI

Features	Benefits
Model development and customization	Accelerate AI development using self-service notebooks and integrated development environments (IDEs) preloaded with curated AI/ML libraries. Reduce the time to creating the 1st model by integrating data ingestion, synthetic data, InstructLab, and RAG. AutoRAG and AutoML (tech previews) automate optimization so teams can focus on delivering business value.

Red Hat OpenShift AI

Model training and experimentation	Cut training time and cost of running distributed workloads across GPU clusters with intelligent hardware allocation and experiment tracking. Versioned artifacts and reproducible workflows keep teams aligned, eliminating repeated work.
Intelligent GPU and hardware speed	Maximize GPU use and control costs with intelligent workload scheduling, quota enforcement, and priority-based access across NVIDIA, AMD, and other accelerator hardware. Hardware profiles give platform teams real-time visibility into GPU use, allowing data scientists to provision accelerators on demand without requiring operational intervention.
AI pipelines	Eliminate manual handoffs and reduce human error with automated, versioned AI pipelines. Each tracked run lets teams reproduce, audit, and optimize workflows from experimentation to production without relying on organizational knowledge.
Optimized model serving	Serve LLMs at production scale with high throughput and low latency using vLLM, and deploy predictive ML models using out-of-the-box and custom runtime servers. Achieve cost-efficient distributed inference with the llm-d framework for predictable, scalable performance. Reduce serving cost through LLM Compressor quantization and use a curated catalog of optimized, validated gen AI models to accelerate time to production.
Agentic AI and gen AI UIs	Speeds agentic AI workflows with expanding focus on AgentOps and connecting agents to core platform services. Provides a unified API layer combining MCP and agentic APIs (the Open Responses API), and a dedicated dashboard experience (AI hub and gen AI studio). MLflow integration provides end-to-end agent traceability and observability, logging LLM calls and tool use for comprehensive visibility.
Model observability and governance	Monitor model health by continuously tracking performance, data drift, and bias in real time, allowing proactive intervention before quality issues reach users. Pair runtime guardrails with LM-Eval and GuideLLM benchmarking to validate models against real-world inference conditions, and capture audit trails through MLflow for compliance evidence for governance and regulatory requirements.
Evaluation	Prevent costly production failures with EvalHub (tech preview), a unified evaluation control plane to scientifically benchmark, score, and assess models, RAG pipelines, and AI agents before and during deployment. Built-in, domain-specific evaluation collections replace ad-hoc manual testing with reproducible, standardized evaluation suites.
Catalog and registry	Govern AI assets from a central registry including predictive and gen AI models, MCP servers, metadata, and deployment artifacts. A curated ecosystem of validated models reduces onboarding time while metadata management ensures traceability and compliance across hybrid cloud deployments.

Feature store	Reduce data preparation time with a centralized feature store providing consistent, reusable datasets. Shared definitions eliminate redundant feature engineering and training-serving skew, accelerating production-ready models.
Models-as-a-service	Provides AI engineers with self-service API access to approved models via a managed, built-in gateway. Use tracking gives administrators visibility into consumption patterns for showback, quota enforcement, and cost accountability.
AI safety and security	Proactively catch jailbreaks, prompt injections, and toxic outputs before production with automated, adversarial vulnerability scanning powered by Garak and NeMo Guardrails. Synthetic data generation (developer preview) creates tailored adversarial test datasets, validating guardrails against realistic threat scenarios and supporting risk documentation required for AI regulations.
Disconnected environments and edge	Deploy portable AI workloads across disconnected, air-gapped, and edge environments to meet strict data sovereignty and regulatory compliance.

Tools for the complete AI lifecycle

Red Hat OpenShift provides the capabilities for organizations to successfully train and deploy their models and move them to production.

The Red Hat OpenShift AI dashboard provides a central place to access applications and documentation, easing adoption. Smart-start tutorials offer optimal guidance for common components and integrated partner software and are available directly from the dashboard to help data scientists learn and get started in less time. The following sections describe the technology partner tools integrated with Red Hat OpenShift AI. Some tools will require an additional license from the technology partner.

Starburst

Starburst

Starburst speeds up analytics by helping teams use data to improve business operations. Delivered as a self-managed product or a fully managed service, Starburst democratizes data access, bringing comprehensive insights to data consumers. Starburst is built on open source Trino (formerly known as PrestoSQL), the premier massively parallel processing (MPP) Structured Query Language (SQL) engine. Built and operated by Trino experts, Starburst gives organizations the freedom to interrogate diverse datasets wherever they exist without needing to move data.

Starburst integrates with the scalable cloud storage and computing services Red Hat OpenShift provides, yielding a stable, security-focused, efficient, and cost-effective way to query all enterprise data. Benefits include:

- **Automation.** Starburst and Red Hat OpenShift operators autoconfigure, autotune, and automanage clusters.

- **High availability and gradual scaledown.** The Red Hat OpenShift load balancer can keep services like the Trino coordinator in an always-on state.
- **Elastic scalability.** Red Hat OpenShift can automatically scale the Trino worker cluster based on query load.

NVIDIA accelerates deployment of AI solutions

As AI/ML applications become increasingly critical to business success, organizations require platforms that can handle complex workloads, optimize hardware use, and provide scalability. Scalable data processing, data analytics, ML training, and inferencing all represent highly resource-intensive computational tasks. NVIDIA software helps accelerate all aspects of end-to-end data science by taking advantage of the parallel processing capabilities of GPUs.

NVIDIA NIM enhances the management and performance of NVIDIA GPUs within the Red Hat OpenShift environment, so AI applications can use the full potential of NVIDIA's AI software and hardware. The integration of NVIDIA NIM and Red Hat OpenShift AI allows for better resource allocation, greater efficiency, and more productive AI workload execution.

Intel OpenVINO toolkit

The Intel OpenVINO toolkit accelerates the development and deployment of high-performance deep learning (DL) inference applications on Intel platforms. The toolkit lets organizations adopt, optimize, and tune neural network models virtually and run comprehensive AI inferencing using the OpenVINO ecosystem of development tools.

- **Model.** Software developers have the flexibility to use their own DL models. For time-to-market advantage, they can also use pretrained and preoptimized models available through Intel's collaboration with Hugging Face for the OpenVINO toolkit. OpenVINO supports PyTorch, ONNX, TensorFlow, and other popular model formats..
- **Optimize.** The OpenVINO toolkit offers several ways to conveniently convert models for better performance, helping software developers achieve faster and more efficient AI model execution. Developers can skip model conversion and run inference directly from PyTorch, ONNX, TensorFlow, TensorFlow Lite, JAX, or PaddlePaddle formats. Conversion to OpenVINO Intermediate Representation (IR) provides optimal performance, which can be further improved by using weights compression and quantization features available in OpenVINO's Neural Network Compression Framework. The same features also reduce storage and runtime.
- **Deploy.** OpenVINO Runtime Inference Engine is an API designed to be integrated into applications to speed up the inference process. Its "write once, deploy anywhere" approach allows teams to efficiently run inference tasks on various Intel hardware, including central processing units (CPUs), GPUs, neural processing units (NPUs), and field-programmable gate arrays (FPGAs). The OpenVINO GenAI extension library simplifies deployment of gen AI workloads, in many cases reducing the code needed to just 3–5 lines. OpenVINO Model Server offers multiple features for agentic and model serving scenarios, reducing development effort even further.

Red Hat OpenShift AI

EDB

EDB Postgres AI is an intelligent platform designed to handle transactional, analytical, and AI workloads, offering flexibility whether data resides on-premise or in any cloud. As a global leader in enterprise Postgres database solutions, EDB provides an open, enterprise-grade sovereign data and AI platform that helps bring AI projects into production up to 3 times faster. Integrating with Red Hat OpenShift AI, EDB Postgres AI allows users to build AI knowledge bases for RAG, unifying AI data, models, and applications into a full-stack sovereign AI platform that can be deployed anywhere. This transformation of core operational data into an AI-ready asset can boost efficiency by up to 30% and can simplify the use of private data, including unstructured data, to ground model outputs in an organization's knowledge base.

Elastic

Elastic's Search AI Platform is built on the Elastic Stack, which comprises Elasticsearch, Kibana, Beats, and Logstash. It combines the precision of search and the intelligence of AI, letting users prototype and integrate with LLMs faster and engage gen AI to build scalable, cost-effective applications. With Elastic's Search AI Platform, users can build transformative RAG applications, proactively resolve observability issues, and address complex security threats. Elasticsearch can be deployed wherever applications are—on-premise, on any cloud, or in air-gapped environments.

Elastic integrates with embedding models from the ecosystem including Red Hat OpenShift AI, Hugging Face, Cohere, and OpenAI via a single, straightforward API call. This approach ensures clean code for managing hybrid inference for RAG workloads, with features that include:

- Chunking, connectors, and web crawlers for ingesting diverse datasets into the search layer.
- Semantic search with Elastic Learned Sparse EncodeR (ELSER), the built-in ML model, and the E5 embedding model, enabling multilingual vector search.
- Document and field-level security, implementing permissions and entitlements that map to an organization's role-based access control (RBAC).

With Elastic's Search AI Platform, organizations are part of a worldwide community of developers where inspiration and support are never far away. Find the Elastic community on Slack, on discussion forums, or on social media.

Conclusion

With Red Hat OpenShift AI, organizations accelerate their journey from predictive and generative AI to autonomous agentic workflows. Teams gain the flexibility to build and deploy models and AI agents across hybrid cloud environments. IT operations and platform engineers benefit from MLOps, GenAIOps, and AgentOps capabilities that support rapid, governed deployments with end-to-end observability. Internal MaaS capabilities provide developers and data scientists with low-toil, self-service API access to approved models and GPUs. Ultimately, OpenShift AI delivers a trusted, consistent foundation offering differentiators in efficient inference, agentic lifecycle management, and scalable hybrid cloud operations.

Learn more

Get started today by visiting Red Hat OpenShift AI.

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

facebook.com/redhat
x.com/RedHat
linkedin.com/company/red-hat

redhat.com
#4112257_0526