Red Hat
OpenShift AI

An open platform for AI models in the hybrid cloud

Highlights

Operationalize and scale AI inference and agentic AI on a proven app platform foundation.

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

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

Embrace intelligent applications and generative AI

Artificial intelligence (AI), machine learning (ML), and deep learning (DL) are having a profound influence on application modernization efforts across diverse businesses and industries. The need to innovate and derive strategic value and new insights from data is expanding the use of AI-enabled cloud-native applications, MLOps, and GenAIOps methodologies. Challenges in this brave new world can be complex– ranging from rapidly increasing model costs when moving into production, complex customization, rigid deployment constraints, and the operations needed to keep up with the pace of innovation. Enterprises require solutions that lower inference costs, simplify scaling and monitoring, and adapt to constant change.

Red Hat® AI accelerates the development and deployment of enterprise AI solutions across hybrid cloud environments. It serves as a comprehensive platform for managing the entire AI/ML lifecycle, offering MLOps and GenAI Ops capabilities. Red Hat AI focuses specifically on 4 key pillars:

- Increasing efficiency with fast, flexible, and efficient inferencing;
- Simplifying the experience for connecting models to data;
- Accelerating Agentic AI innovation; and
- Ensuring flexibility and consistency when scaling AI across the hybrid cloud.

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

Rapidly develop, train, test, and deploy

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

The platform offers an integrated user interface (UI) experience with tooling for building, training, tuning, deploying, and monitoring predictive and gen AI models. You can deploy models to hybrid cloud environments, with a specific emphasis on providing a controlled and protected footprint for sovereign and private AI. This approach makes certain 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 a very important factor driving digital infrastructure budgets in 2026 as organizations work 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 enhance trust in AI initiatives by combining open source communities with a robust AI ecosystem. This offers an increase in flexibility and freedom to select the right AI/ML technology for your organization. Users can build their predictive models or start with an external gen AI model, then enhance it with retrieval-augmented generation (RAG) using one of several model servers provided in the platform. The platform offers quick access to optimized and validated third-party models, such as Llama, Mistral, DeepSeek and Granite, that run efficiently on vLLM, available on the Red Hat AI repository on Hugging Face. The catalog allows users to explore these models and add their own.

Improve operational consistency across teams

Red Hat 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, graphic 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 better governance and security.

Gain hybrid cloud flexibility

Red Hat OpenShift AI allows training, deployment, and monitoring AI/ML workloads across various environments—cloud, on-premise datacenters, or environments air-gapped—to meet regulatory, security, and data requirements. The platform is compatible with multiple AI accelerators from vendors like NVIDIA, AMD, and Intel. This capability can be expanded to create a GPU-as-a-service environment, which allows organizations to centrally manage, partition, and schedule GPU resources, while also providing detailed observability into their use.

Gen AI and agentic

For gen AI projects, dedicated user experiences are offered through components like AI hub (Developer Preview), a dashboard experience for platform engineers, consolidating the catalog, registry, and model deployments to set up and deploy models and MCP servers. Gen AI studio (Developer Preview) provides AI asset endpoints and an AI playground where AI engineers and application developers can access, experiment, compare, evaluate, and test deployed models and MCP servers.

OpenShift AI accelerates agentic AI by providing a unified API layer and a flexible, scalable foundation. The Llama Stack API and MCP (Tech Preview) support includes an enterprise-grade implementation of the Llama Stack API, offering a single, standardized entry point for various AI capabilities.

Additional tools include LLM evaluation (LM Eval) and LLM benchmarking to assist real world inference deployments. LLM compressor provides algorithms to reduce the size of an organization's custom models using similar methods that Red Hat uses to create validated and optimized models in the Red Hat AI repository on Hugging Face.

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

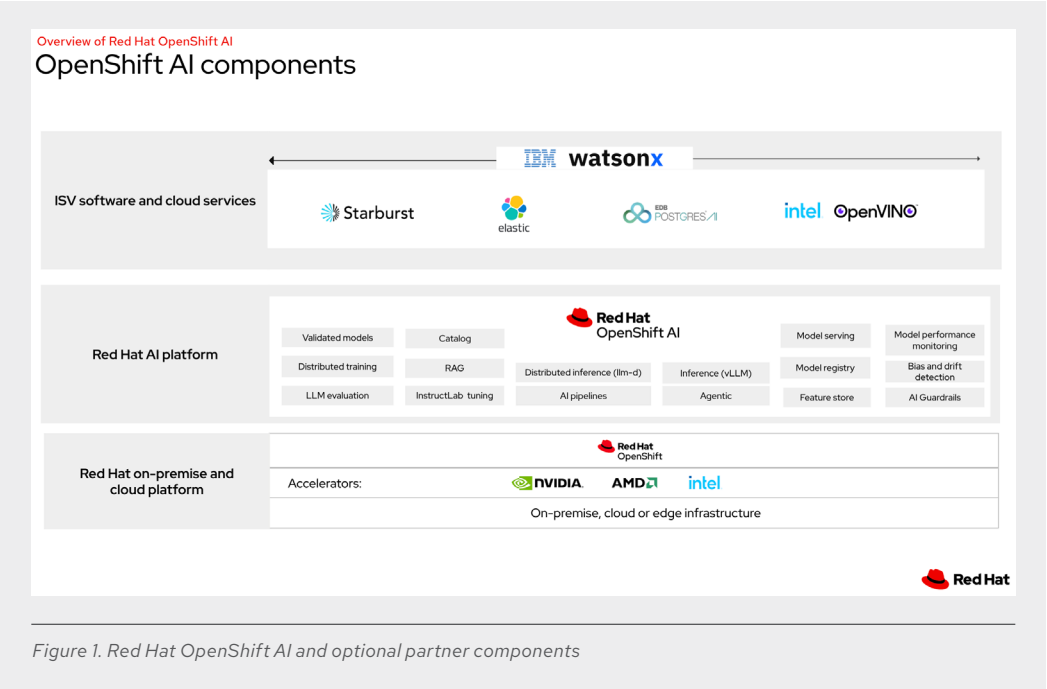

Figure 1. Red Hat OpenShift AI and optional partner components

Several other core tools and capabilities provided with Red Hat OpenShift AI offer a solid foundation:

- **Model building and customization.** Data scientists can conduct exploratory data science in a JupyterLab UI, offering out-of-the-box securely built notebook images with common Python libraries. For gen AI projects, OpenShift AI enables Retrieval Augmented Generation (RAG) and distributed InstructLab training, providing model alignment tooling to contribute skills and knowledge to genAI models more efficiently.
- **Model serving.** Red Hat OpenShift AI provides a variety of frameworks using KServe as the core engine for model serving to simplify the deployment of predictive machine learning or foundation models to production environments. For LLMs requiring maximum scalability, OpenShift AI offers parallelized serving with vLLM runtimes. LIm-d offers a framework for optimizing LLM inference by disaggregating the pipeline into modular services, that supports smart autoscaling and efficient request routing.
- **AI pipelines.** Red Hat OpenShift AI offers a pipelines component that lets you orchestrate AI tasks into pipelines and build pipelines using a graphical front end. Organizations can chain together processes like data preparation, build models, and serve models.
- **Model monitoring.** Red Hat OpenShift AI helps Ops-oriented users monitor operations and performance metrics for model servers and deployed models. Users can access out-of-the-box visualizations for performance and operations metrics or integrate data with other observability services.
- **Distributed workloads.** Distributed workloads allow teams to accelerate data processing along with model training, tuning, and serving. This capability supports prioritization and distribution of job execution along with optimal node use. Advanced GPU support helps handle the workload demands of foundation models.

- **AI guardrails, bias and drift detection.** Red Hat OpenShift AI provides tools to help data scientists and AI engineers monitor whether models are fair and unbiased based on the training data but also for fairness during real-world deployments. AI guardrails provide a customizable framework implementing crucial safety controls, helping ensure that models are transparent, fair, and reliable for use in production. Drift detection tools include input data distributions for deployed ML models to detect when the live data used for model inference significantly deviates from the data upon which the model was trained.
- **Catalog and registry.** Red Hat OpenShift AI provides an internal model catalog and a curated catalog where Platform Engineers can discover, compare, and evaluate optimized gen AI models. It also provides a central registry helping data scientists and AI engineers share, modify, deploy, and track predictive and gen AI models, metadata and model artifacts.
- **Feature store.** Manage clean, well-defined data features for ML models, enhancing performance and accelerating workflows.

Tools for the complete AI lifecycle

Red Hat OpenShift provides the services and software to let organizations successfully train and deploy their models and move them to production (see Figure 2).

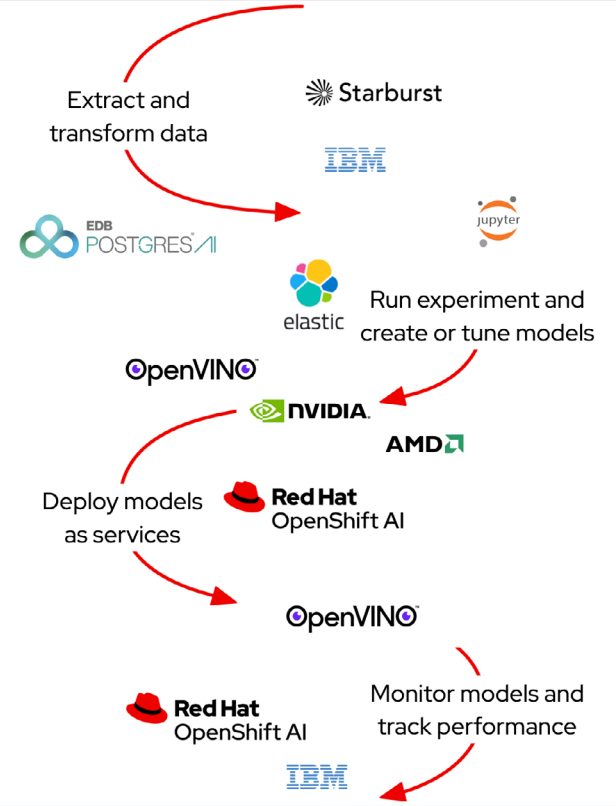

Figure 2. A typical model operationalization lifecycle

The Red Hat OpenShift AI dashboard provides a central place to discover and access all applications and documentation, easing adoption. Smart start tutorials offer best-practice 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 accelerates analytics by making it fast and easy for your teams to capitalize on your data to improve how the business functions. 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 premiere massively parallel processing (MPP) Structured Query Language (SQL) engine. Built and operated by Trino experts, Starburst gives you the freedom to interrogate diverse data sets wherever they exist without needing to move your 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 your enterprise data. Benefits include:

- **Automation.** Starburst and Red Hat OpenShift operators provide auto-configuration, auto-tuning, and auto-management of 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.

Hewlett Packard
Enterprise

HPE Machine Learning Data Management Software

Organizations need data management solutions that facilitate everything from laptop experiments to critical enterprise deployments. HPE Machine Learning Data Management Software (formerly known as Pachyderm) allows data science teams to build and scale containerized, data-driven ML pipelines with an assured data lineage provided by automatic data versioning. Engineered to solve real-world data science problems, HPE Machine Learning Data Management Software provides the data foundation that allows teams to automate and scale their ML lifecycle with reproducibility. With use cases that range from unstructured data to data warehouses, natural language processing, video and image extract, transform, and load (ETL), financial services, and life science, HPE Machine Learning Data Management Software provides:

- Automated data versioning that gives teams a high-performance way to keep track of data changes.
- Data-driven containerized pipelines that speed up data processing while lowering compute costs.
- An immutable data lineage that provides a fixed record for activities and assets in the ML lifecycle.
- A console that provides an intuitive visualization of your directed acyclic graph (DAG) and aids with debugging and reproducibility.
- Jupyter notebook support with a JupyterLab Mount Extension for a point-and-click interface to versioned data.

- Enterprise administration with robust tools for deploying and administering HPE Machine Learning Data Management Software at scale across different teams within the organization.

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 makes it possible to 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, allowing AI applications to 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 DL inference applications on Intel platforms. The toolkit lets you 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 convert models for better convenience and 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 IR provides optimal performance, which can be optimized further by using weights compression and quantization features available in OpenVINO's Neural Network Compression Framework. The same features also reduce the storage and runtime footprint.
- **Deploy.** OpenVINO Runtime Inference Engine is an application programming interface (API) designed to be integrated into your applications to accelerate the inference process. Its "write once, deploy anywhere" approach allows you to efficiently run inference tasks on various Intel hardware, including central processing units (CPUs), GPUs, NPU and FPGAs.. OpenVINO GenAI extension library simplifies deployment of gen AI workloads, in many cases reducing the code needed to just 3 to 5 lines. OpenVINO Model Server offers multiple features for Agentic and model serving scenarios, reducing development effort even further.

EDB

EDB Postgres AI is a powerful and intelligent platform designed to handle transactional, analytical, and AI workloads, offering unparalleled flexibility whether data resides on-premise or in any cloud landscape. As a global leader in enterprise Postgres database solutions, EDB provides an open, enterprise-grade sovereign data and AI platform that helps accelerate AI projects into production up to 3 times faster. Integrating with Red Hat OpenShift AI, EDB Postgres AI allows users to build robust AI knowledge bases for Retrieval-Augmented Generation (RAG), unifying AI data, models, and

applications into a full-stack sovereign AI platform deployable 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

The Elastic Search AI Platform (built on the ELK Stack[2]) 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. The Elastic Search AI Platform allows users to build transformative retrieval augmented generation (RAG) applications, proactively resolve observability issues, and address complex security threats. Elasticsearch can be deployed where your applications are: on-premise, on your chosen cloud provider, or in air-gapped environments.

Elastic integrates with embedding models from the ecosystem including Red Hat OpenShift AI, Hugging Face, Cohere, OpenAI, and others 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 your 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 your organization's role-based access control (RBAC).

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

Conclusion

With Red Hat OpenShift AI, organizations can experiment, collaborate, and ultimately accelerate their AI-powered application journey. Data scientists and AI engineers gain the flexibility of using Red Hat OpenShift AI to build and deploy models across the hybrid cloud. IT operations and platform engineers benefit from MLOps and GenAIOps capabilities, allowing models to deploy into production more rapidly. Self-service for developers, AI engineers, and data scientists, including access to GPUs, advances innovation on an application platform already used and fully trusted by enterprise IT. Red Hat OpenShift AI continues to deliver a comprehensive, trusted, and consistent platform, offering unique differentiators in efficient inference, agentic AI, and scalable hybrid cloud operations, backed by a robust partner ecosystem.

Learn more

Get started today by visiting us at Red Hat OpenShift AI.

2 *The ELK stack consists of Elasticsearch, Kibana, Beats, and Logstash.*

About Red Hat

Red Hat helps customers standardize across environments, develop cloud-native applications, and integrate, automate, secure, and manage complex environments with award-winning support, training, and consulting services.

facebook.com/redhatinc
@RedHat
linkedin.com/company/red-hat

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

redhat.com
#2876428_1125