Red Hat
OpenShift AI

An open platform for AI models in the hybrid cloud

Highlights

Lower costs when moving from experiment to production by scaling and automating infrastructure.

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 fostering the expanding use of AI-enabled cloud-native applications and MLOps methodologies. At the same time, this brave new world can be complex, with implications for everyone—from developers to data scientists to operations staff. Operationalizing AI/ML is not trivial—often requiring months while generative AI (gen AI) innovation progresses daily. This mismatch can result in project failures that put your business at risk amid myriad challenges:

- Matching the pace of AI progress can be daunting, from keeping rapidly evolving tools and application services current and consistent to provisioning hardware resources like graphic processing units (GPUs), and scaling AI-enabled applications.
- Organizations must mitigate risks when investing in AI and still gain value, especially when using sensitive data to inform models and applications.
- Maintaining different platforms for application developers and data scientists can complicate collaboration and hinder the speed of development.
- AI-enabled application deployment must happen at scale and close to data generation points.

Built on Red Hat® OpenShift®, a leading hybrid cloud application platform and part of the Red Hat AI portfolio, Red Hat OpenShift AI gives data scientists and developers a powerful AI/ML platform for building and deploying intelligent applications. Organizations can experiment with a choice of tools, collaborate, and accelerate time to market—all within 1 common platform. Red Hat OpenShift AI combines the self-service environment that data scientists and developers want with the confidence that enterprise IT demands.

Having a trusted foundation reduces friction throughout the lifecycle. Red Hat OpenShift AI offers a robust platform, a broad ecosystem of popular certified tools, and familiar workflows for deploying models into production. With these advantages, teams can collaborate with less friction and get AI-enabled applications into the market more efficiently, ultimately delivering greater value for the business.

Red Hat
OpenShift AI

Rapidly develop, train, test, and deploy

Red Hat OpenShift AI is a flexible, scalable AI platform with tools to build, deploy, and manage AI-enabled applications. It is built with open source technologies and provides trusted, operationally consistent capabilities for teams to experiment, serve models, and deliver innovative apps. Red Hat OpenShift AI accelerates the delivery of intelligent applications, helping ML models move from early pilots into intelligent applications with greater speed—on a shared, consistent platform.

When asked which technology areas would receive increased funding in 2025, 84% selected AI , compared with 73% the previous year.[1]

Red Hat OpenShift AI 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 on premise or in all major public clouds, gaining the flexibility of running workloads wherever needed, with no commercial cloud lock-in. Red Hat OpenShift AI is based on the community Open Data Hub project and common open source projects such as Jupyter, Pytorch, vLLM, and Kubeflow.

Lower costs to scale in production

As an add-on to Red Hat OpenShift, OpenShift AI provides a platform designed to handle the most demanding workloads. OpenShift AI lowers the ongoing training, serving, and infrastructure costs for generative and predictive AI projects as you move from experiment to production by simplifying resource provisioning and automating multiple tasks through data pipelines. OpenShift AI reduces the costs of model inferencing by using optimized serving engines and runtimes, such as vLLM, and scaling the underlying infrastructure as the workload demands.

Data scientists can use their familiar tools and frameworks or access a growing technology partner ecosystem for deeper AI/ML expertise—without being burdened with a prescriptive toolchain. Rather than waiting for IT to provision necessary resources, they get on-demand infrastructure with 1 click rather than an IT ticket.

Reduce operational complexity

Red Hat OpenShift AI provides a consistent user experience that empowers data scientists, data engineers, application engineers, and DevOps teams to collaborate to deliver timely AI solutions effectively. It offers self-service access to collaborative workflows, acceleration through access to GPUs, and streamlined operations. Organizations can deliver AI solutions consistently at scale across hybrid cloud environments and at the edge.

Because OpenShift AI is an add-on offering to Red Hat OpenShift, IT operations can provide more straightforward configurations to data scientists and application developers on a stable and proven platform that they can scale up or down with low effort. IT worries less about governance and security, with no need to chase down rogue cloud-platform accounts.

Gain hybrid cloud flexibility

Red Hat OpenShift AI provides the ability to train, deploy, and monitor AI/ML workloads in a cloud environment, in on-premise datacenters, or at the network edge, close to where data is generated or located. This flexibility allows AI strategies to evolve, moving operations to the cloud or to the edge as required by the business. Organizations can train and deploy models and AI-enabled applications wherever they need to meet relevant regulatory, security, and data requirements, including air-gapped and disconnected environments.

1 *Gartner Research. "2025 Planning Guide for Analytics and Artificial Intelligence." 14 Oct. 2024.*

Red Hat
OpenShift AI

Red Hat OpenShift AI

Figure 1 illustrates how the model operation lifecycle integrates OpenShift AI as a common platform, expanding the capabilities of the leading application platform, Red Hat OpenShift. Organizations build on a proven hybrid cloud platform, offered as either self-managed as traditional software or as a managed cloud service, yielding tremendous flexibility. The self-managed version can run anywhere Red Hat OpenShift runs, either on-premise or on any of the 3 major public cloud environments. The cloud service version is available on Red Hat OpenShift Service on AWS and Red Hat OpenShift Dedicated (on AWS or Google Cloud Platform).

Our solutions also allow for expanding AI capabilities through collaboration with a rich ecosystem of dozens of AI software and Software-as-a-Service (SaaS) partners. Red Hat OpenShift AI is flexible and composable, allowing customers to assemble an end-to-end AI/ML platform that fits their specific needs.

For those just getting started with gen AI models, OpenShift AI includes the components of Red Hat Enterprise Linux® AI—a foundational model platform to develop, test, and run Granite family large language models (LLMs) to power enterprise applications. In addition to the Granite models provided with Red Hat OpenShift AI, the framework supports models from HuggingFace, Stability AI, and other model repositories.

Red Hat OpenShift AI is a foundational component in IBM watsonx.ai, providing fundamental AI tooling and services for gen AI workloads. Watsonx offers an enterprise studio for AI builders to deliver gen AI applications with low code and no code requirements, user-friendly workflows for model development, and access to a library of IBM foundation models and curated open source models. Red Hat OpenShift and OpenShift AI are embedded technical prerequisites for watsonx software.

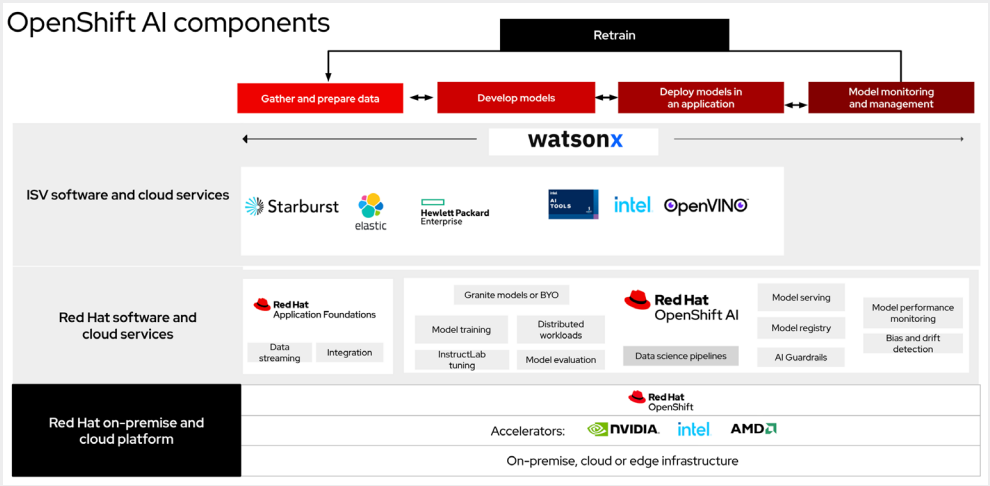

Figure 1. Red Hat OpenShift AI components

Core tools and capabilities provided with Red Hat OpenShift AI offer a solid foundation:

- **Model building and fine tuning.** Data scientists can conduct exploratory data science in a JupyterLab UI, offering out-of-the-box securely built notebook images with common Python libraries and packages, including TensorFlow, PyTorch, and CUDA. In addition, organizations can provide their own custom notebook images, allowing them to create and collaborate on notebooks while organizing work in projects and workbenches.
- **Model serving.** Red Hat OpenShift AI provides a variety of frameworks for model serving routing to simplify the deployment of predictive machine learning or foundation models to production environments regardless of their compute resource requirements. For generative AI workloads, OpenShift AI provides vLLM-powered model inferencing, offering industry-leading performance and efficiency across the most popular open source large language models (LLMs). The solution also empowers customers to bring and utilize their preferred runtimes, ensuring flexibility and control.
- **Data science pipelines.** Red Hat OpenShift AI also includes a data science pipelines component that lets you orchestrate data science 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. Data scientists can use 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 utilization. Advanced GPU support helps handle the workload demands of foundation models.
- **Bias and drift detection.** Red Hat OpenShift AI provides tools to help data scientists monitor whether models are fair and unbiased based on the training data but also for fairness during real-world deployments. 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.
- **AI Guardrails (in tech preview).** AI Guardrails support includes input detectors to safeguard the types of interactions a user can request and output detectors to help safety-check the model outputs. AI Guardrails help filter out hateful, abusive, or profane speech, personally identifiable information, competitive information or other domain specific constraints. We provide a set of detectors, and customers can add their own.
- **Model evaluation.** During the model exploration and development phase, the LM evaluation (LM-Eval) component provides important information on the quality of the model. LM-Eval allows data scientists to benchmark the performance of LLM models across a variety of tasks, such as logical reasoning, mathematical reasoning, adversarial natural language, and many others. The benchmarks we provide are based on industry standards.

- **Model registry.** Red Hat OpenShift AI provides a central place to view and manage registered models, helping data scientists share, version, deploy and track predictive and gen AI models, metadata and model artifacts.

In addition to Jupyter Notebook images and a Jupyter spawner to deploy prepackaged or an organization's custom images to data science teams, OpenShift AI also includes the Git plug-in to JupyterLab, speeding integration with Git directly from the user interface. Other common analytics packages are provided as a part of the product to simplify operation and make getting started more straightforward with the right tools for your project, including Pandas, scikit-learn, and NumPy. RStudio server (tested and verified) and VS Code Server are also offered as alternate IDEs to JupyterLab giving data scientists more choice. The project's UI allows data scientists to create their own workspaces to organize and share their notebook images and artifacts as projects and collaborate with other users.

For gen AI projects, OpenShift AI enables distributed InstructLab training as a preview feature. InstructLab is a key component of Red Hat Enterprise Linux AI, and provides model alignment tooling to help organizations more efficiently tune smaller language models with their private data even if their teams lack AI expertise. Red Hat OpenShift AI also supports efficient fine-tuning of large language models (LLMs) with LoRa/QLoRA to reduce the computational overhead and memory footprint as well as support for embeddings to make it easier to integrate text info with vector databases required for RAG.

The requirement to serve both predictive and gen AI models demands a flexible model-serving approach. Red Hat OpenShift AI supports multiple model-serving frameworks, and you can choose the provided multi-model and single-model inference servers or your own custom inference server. The model-serving UI is directly integrated into the Red Hat OpenShift AI dashboard as well as the projects workspace. Underlying cluster resources can scale up or down as your workload requires. For LLMs that require maximum scalability, Red Hat OpenShift AI offers parallelized serving across multiple nodes with vLLM runtimes providing the ability to handle multiple requests in real-time.

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 (Figure 2). In addition to OpenShift AI, this process is integrated with Red Hat Application Foundations, which includes streams for Apache Kafka for real-time data and event streaming, Red Hat 3scale for API Management, and Red Hat build of Apache Camel for data integration.

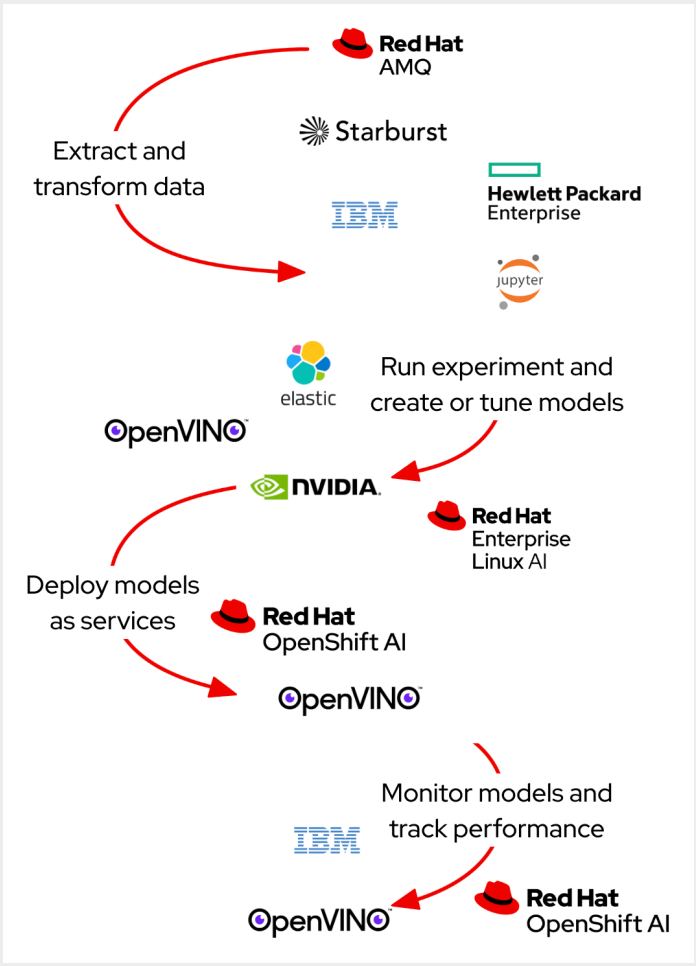

Figure 2. A typical model operationalization lifecycle

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 more comprehensive insights to data consumers. Starburst is built on open source Trino (formerly known as PrestoSQL), the premiere massively parallel processing (MPP) SQL engine. Built and operated by Trino experts and the creators of Presto, Starburst gives you the freedom to interrogate diverse data sets wherever they exist without having to move your data.

Starburst integrates with the scalable cloud storage and computing services Red Hat OpenShift provides, yielding a more 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 graceful 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 a guaranteed data lineage provided by automatic data versioning. Engineered to solve real-world data science problems, Pachyderm provides the data foundation that allows teams to automate and scale their ML lifecycle while guaranteeing reproducibility. With use cases that range from unstructured data to data warehouses, natural language processing, video and image ETL (extract, transform, and load), 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 all 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 all 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 accelerated computing hardware and software platforms power the new era of computing

As AI/ML applications become increasingly critical to business success, organizations require platforms that can handle complex workloads, optimize hardware utilization, and provide scalability. Scalable data processing, data analytics, machine learning training, and inferencing all represent highly resource-intensive computational tasks that are well suited for accelerated computing. NVIDIA AI Enterprise software streamlines the deployment and deployment of production-grade AI solutions. NVIDIA NIM, a part of NVIDIA AI Enterprise, is a set of easy-to-use microservices designed for secure, reliable deployment of high-performance AI model inferencing. NVIDIA NIM enhances the management and performance of AI models within the Red Hat OpenShift environment, allowing AI applications to use the full potential of NVIDIA accelerated computing and NVIDIA AI Enterprise software. The combination of NVIDIA accelerated computing, NVIDIA AI Enterprise, and Red Hat OpenShift AI allows for better resource allocation, greater efficiency, and faster AI workload execution.

Intel OpenVINO toolkit

The Intel Distribution of the OpenVINO toolkit accelerates the development and deployment of high-performance DL inference applications on Intel platforms. The toolkit lets you adopt, optimize, and tune virtually any neural network model 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.
- **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 TensorFlow, TensorFlow Lite, ONNX, or PaddlePaddle formats. Conversion to OpenVINO IR provides optimal performance, reducing the time it takes for the 1st inference and saving storage space. The Neural Network Compression Framework can provide further improvements.
- **Deploy.** The 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, and accelerators.

AI Tools from Intel®

AI Tools from Intel (formerly Intel AI Analytics Toolkit) give data scientists, AI developers, and researchers familiar Python tools and frameworks to accelerate end-to-end data science and analytics pipelines on Intel architectures. The components use Intel's oneAPI libraries for low-level compute optimizations. This toolkit maximizes performance from preprocessing through ML and provides interoperability for efficient model development.

Using AI Tools from Intel, you can:

- Deliver high-performance DL training on Intel XPUs and integrate faster inferencing into your AI development workflow with Intel-optimized DL frameworks for TensorFlow and PyTorch, including pretrained models and low-precision tools.

- Achieve drop-in acceleration for data preprocessing and ML workflows with compute-intensive Python packages, Modin, scikit-learn, and XGBoost, optimized for Intel.
- Gain direct access to analytics and AI optimizations from Intel to ensure your software works together uninterrupted.

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.

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 gain the flexibility of using Red Hat OpenShift AI as a cloud-based add-on service managed by Red Hat or as a self-managed software offering, simplifying tasks no matter where they build their models. IT operations benefit from MLOps capabilities, allowing models to deploy into production more rapidly. Self-service for developers and data scientists, including access to GPUs, advances innovation on an application platform already used and fully trusted by enterprise IT. Unlike competing approaches, data scientists can choose tooling with no restrictive toolchain, yielding new data insights without forcing arbitrary limitations.

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
#1847825_0325

Copyright © 2025 Red Hat, Inc. Red Hat, the Red Hat logo, and OpenShift are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries. Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries. All other trademarks are the property of their respective owners.