Red Hat
OpenShift AI

Develop, train, test, and deploy ML across the hybrid cloud

Highlights

Rapidly deliver intelligent applications on premise or in the public cloud.

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

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

Gain hybrid cloud flexibility with the ability to build, train, deploy, and monitor AI/ML workloads in the cloud, on-premise, or at the edge—close to where data is located.

Open Data Hub is a blueprint for building an AI-as-a-Service (AIaaS) platform on Red Hat OpenShift. It inherits from upstream efforts such as Jupyter and Kubeflow, and is the foundation for Red Hat OpenShift AI.

Accelerating artificial intelligence and machine learning deployments

Artificial intelligence, machine learning, and deep learning (AI/ML/DL) have rapidly become critical for businesses and organizations with adoption of generative AI (gen AI) increasing rapidly. Gartner now estimates that 85% of enterprises will have used gen AI application programming interfaces (APIs) or deployed gen AI-enabled applications by 2026.[1] Deploying these technologies, however, can be complicated. As data scientists strive to build their models, they often encounter a lack of alignment between rapidly evolving tools. These gaps can negatively impact productivity and collaboration among data scientists, software developers, and IT operations. Scaling AI/ML deployments can be resource-limited and administratively complex while requiring expensive graphics processing unit (GPU)-resources for hardware acceleration and distributed workloads for gen AI. Popular cloud platforms offer scalability and attractive toolsets, but those same tools often lock users in, limiting architectural and deployment choices.

Based on the open source Open Data Hub project, Red Hat® OpenShift® AI[2] lets data scientists rapidly train, test, serve and monitor ML/DL models including gen AI. Users can immediately focus on their modeling and application development without waiting for infrastructure provisioning. Available as an add-on to Red Hat OpenShift, either as a fully managed cloud service or as a self-managed software product, OpenShift AI combines Red Hat components, open source software, and technology partner offerings with the flexibility to develop and serve models on-premise, in a cloud, or on edge infrastructure.

Red Hat OpenShift AI

OpenShift AI offers organizations an efficient way to deploy an integrated set of common open source and third-party tools to perform AI/ML modeling. The platform represents an alternative to prescriptive and opinionated AI/ML suites available from individual cloud providers. Adopters gain a collaborative open source toolset and a platform for building experimental models without worrying about the infrastructure or lock-in from public cloud-specific tools. They can then extend that base platform with partner tools to gain increased capabilities. Models can be served to production environments in a container-ready format, consistently, across hybrid cloud and edge environments. OpenShift AI provides IT operations with an environment that is simple to manage, with straightforward configurations on a proven, scalable, and security-focused platform.

OpenShift AI supports popular gen AI foundation models, letting you prompt-tune, fine tune, and serve these pretrained models for your unique use cases and with your own data. You can even distribute workloads across multiple Red Hat OpenShift clusters, independent of their location. The platform makes it simpler to exploit AI hardware acceleration too, supporting central processing unit (CPU) and graphic processing unit (GPU)-based hardware infrastructure including Nvidia GPUs and Intel XPUs—all without the need to stand up and manage your own data science platform.

[1] *Gartner press release . "Gartner Says More Than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026." 11 Oct. 2023.*

[2] *Formerly Red Hat OpenShift Data Science*

Red Hat Consulting offers the OpenShift AI pilot engagement to help organizations get started on their OpenShift AI journey and integrate it with their existing enterprise.

For organizations looking to move beyond model experimentation to develop strategies for deploying models to production, Red Hat Consulting also offers an MLOps Foundation consulting service.

Upstream open source and commercial technology partner tools

Red Hat OpenShift AI provides a subset (Table 1) of the tools found in the upstream Open Data Hub project. Organizations can develop, test, and deploy models across any cloud environment, fully managed, and self-managed Red Hat OpenShift and centrally monitor their performance. Red Hat provides regular updates to open source tools (e.g., Jupyter, Pytorch, and Tensorflow), removing integration, testing and maintenance burden. The offering also integrates several AI/ML technology partner offerings (Table 1). Additional commercial technology partner offerings can also be added from more than 30 AI technology partners who have certified their products on Red Hat OpenShift.

Table 1. Red Hat OpenShift AI ecosystem

AI/ML modeling and visualization tools	JupyterLab UI with prebuilt notebook images and common Python libraries and packages; TensorFlow; PyTorch, CUDA; Kubeflow notebook controller for managing multiple notebook sessions, Anaconda (Professional is optional); AI Tools from Intel
Data engineering	Starburst (Galaxy and Enterprise are optional); Pachyderm (optional)
Data ingestion and storage	Red Hat AMQ (optional add-on); Amazon Simple Storage Service (S3)
GPU support	NVIDIA (with GPU operator), Intel XPUs (including Intel Xeon processors, Habana Gaudi, and Intel Data Center GPU Flex Series)
Model serving and monitoring	Model serving (KServe with user interface), model monitoring, OpenShift Source-to-Image (S2I), Red Hat OpenShift API Management (optional add-on), Intel Distribution of the OpenVINO toolkit
Data science pipelines	Data science pipelines (Kubeflow Pipelines) chain together processes like data preparation, build models, and serve models

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