Red Hat OpenShift AI Cloud Service 1

Introduction to Red Hat OpenShift AI Cloud Service

OpenShift AI is a platform for data scientists and developers of artificial intelligence and machine learning (AI/ML) applications

Last Updated: 2025-03-18

Red Hat OpenShift AI Cloud Service 1 Introduction to Red Hat OpenShift AI Cloud Service

OpenShift AI is a platform for data scientists and developers of artificial intelligence and machine learning (AI/ML) applications

Legal Notice

Copyright © 2025 Red Hat, Inc.

The text of and illustrations in this document are licensed by Red Hat under a Creative Commons Attribution–Share Alike 3.0 Unported license ("CC-BY-SA"). An explanation of CC-BY-SA is available at
http://creativecommons.org/licenses/by-sa/3.0/
. In accordance with CC-BY-SA, if you distribute this document or an adaptation of it, you must provide the URL for the original version.

Red Hat, as the licensor of this document, waives the right to enforce, and agrees not to assert, Section 4d of CC-BY-SA to the fullest extent permitted by applicable law.

Red Hat, Red Hat Enterprise Linux, the Shadowman logo, the Red Hat logo, JBoss, OpenShift, Fedora, the Infinity logo, and RHCE are trademarks of Red Hat, Inc., registered in the United States and other countries.

Linux ® is the registered trademark of Linus Torvalds in the United States and other countries.

Java ® is a registered trademark of Oracle and/or its affiliates.

XFS ® is a trademark of Silicon Graphics International Corp. or its subsidiaries in the United States and/or other countries.

MySQL ® is a registered trademark of MySQL AB in the United States, the European Union and other countries.

Node.js ® is an official trademark of Joyent. Red Hat is not formally related to or endorsed by the official Joyent Node.js open source or commercial project.

The OpenStack ® Word Mark and OpenStack logo are either registered trademarks/service marks or trademarks/service marks of the OpenStack Foundation, in the United States and other countries and are used with the OpenStack Foundation's permission. We are not affiliated with, endorsed or sponsored by the OpenStack Foundation, or the OpenStack community.

All other trademarks are the property of their respective owners.

Abstract

Red Hat OpenShift AI is a platform for data scientists and developers of artificial intelligence and machine learning applications.

Table of Contents

CHAPTER 1. OVERVIEW OF OPENSHIFT AI

Red Hat OpenShift AI is a platform for data scientists and developers of artificial intelligence and machine learning (AI/ML) applications.

OpenShift AI provides an environment to develop, train, serve, test, and monitor AI/ML models and applications on-premise or in the cloud.

For data scientists, OpenShift AI includes Jupyter and a collection of default notebook images optimized with the tools and libraries required for model development, and the TensorFlow and PyTorch frameworks. Deploy and host your models, integrate models into external applications, and export models to host them in any hybrid cloud environment. You can enhance your data science projects on OpenShift AI by building portable machine learning (ML) workflows with data science pipelines, using Docker containers. You can also accelerate your data science experiments through the use of graphics processing units (GPUs) and Intel Gaudi AI accelerators.

For administrators, OpenShift AI enables data science workloads in an existing Red Hat OpenShift or ROSA environment. Manage users with your existing OpenShift identity provider, and manage the resources available to notebook servers to ensure data scientists have what they require to create, train, and host models. Use accelerators to reduce costs and allow your data scientists to enhance the performance of their end-to-end data science workflows using graphics processing units (GPUs) and Intel Gaudi AI accelerators.

OpenShift AI has two deployment options:

- **Self-managed software** that you can install on-premise or in the cloud. You can install OpenShift AI Self-Managed in a self-managed environment such as OpenShift Container Platform, or in Red Hat-managed cloud environments such as Red Hat OpenShift Dedicated (with a Customer Cloud Subscription for AWS or GCP), Red Hat OpenShift Service on Amazon Web Services (ROSA Classic or ROSA HCP), or Microsoft Azure Red Hat OpenShift.
  For information about OpenShift AI as self-managed software on your OpenShift cluster in a connected or a disconnected environment, see Product Documentation for Red Hat OpenShift AI Self-Managed.

- A **managed cloud service**, installed as an add-on in Red Hat OpenShift Dedicated (with a Customer Cloud Subscription for AWS or GCP) or in Red Hat OpenShift Service on Amazon Web Services (ROSA Classic).

For information about OpenShift AI supported software platforms, components, and dependencies, see the Red Hat OpenShift AI: Supported Configurations Knowledgebase article.

For a detailed view of the release lifecycle, including the full support phase window, see the Red Hat OpenShift AI Cloud Service Life Cycle Knowledgebase article.

CHAPTER 2. PRODUCT FEATURES

Red Hat OpenShift AI provides several features for data scientists and IT operations administrators.

2.1. FEATURES FOR DATA SCIENTISTS

Containers

While tools such as JupyterLab already offer intuitive ways for data scientists to develop models on their machines, there are always inherent complexities involved with collaboration and sharing work. Moreover, using specialized hardware such as powerful GPUs can be very expensive when you have to buy and maintain your own. The Jupyter environment that is included with OpenShift AI lets you take your development environment anywhere you need it to be. Because all of the workloads are run as containers, collaboration is as easy as sharing an image with your team members, or even simply adding it to the list of default containers that they can use. As a result, GPUs and large amounts of memory are significantly more accessible, since you are no longer limited by what your laptop can support.

Integration with third-party machine learning tools

We have all run into situations where our favorite tools or services do not play well with one another. OpenShift AI is designed with flexibility in mind. You can use a wide range of open source and third-party tools with OpenShift AI. These tools support the complete machine learning lifecycle, from data engineering and feature extraction to model deployment and management.

Collaboration on notebooks with Git

Use Jupyter's Git interface to work collaboratively with others, and keep good track of the changes to your code.

Securely built notebook images

Choose from a default set of notebook images that are pre-configured with the tools and libraries that you need for model development. Software stacks, especially those involved in machine learning, tend to be complex systems. There are many modules and libraries in the Python ecosystem that can be used, so determining which versions of what libraries to use can be very challenging. OpenShift AI includes many packaged notebook images that have been built with insight from data scientists and recommendation engines. You can start new projects quickly on the right foot without worrying about downloading unproven and possibly insecure images from random upstream repositories.

Custom workbench images

In addition to workbench images provided and supported by Red Hat and independent software vendors (ISVs), you can configure custom workbench images that cater to your project's specific requirements.

Data science pipelines

OpenShift AI supports data science pipelines 2.0, for an efficient way of running your data science workloads. You can standardize and automate machine learning workflows that enable you to develop and deploy your data science models.

Model serving

As a data scientist, you can deploy your trained machine-learning models to serve intelligent applications in production. Deploying or serving a model makes the model's functions available as a service endpoint that can be used for testing or integration into applications. You have much control over how this serving is performed.

Optimize your data science models with accelerators

If you work with large data sets, you can optimize the performance of your data science models in OpenShift AI with NVIDIA graphics processing units (GPUs) or Intel Gaudi AI accelerators. Accelerators enable you to scale your work, reduce latency, and increase productivity.

2.2. FEATURES FOR IT OPERATIONS ADMINISTRATORS

Manage users with an identity provider

OpenShift AI supports the same authentication systems as your OpenShift cluster. By default, OpenShift AI is accessible to all users listed in your identity provider and those users do not need a separate set of credentials to access OpenShift AI. Optionally, you can limit the set of users who have access by creating an OpenShift group that specifies a subset of users. You can also create an OpenShift group that identifies the list of users who have administrator access to OpenShift AI.

Manage resources with OpenShift

Use your existing OpenShift knowledge to configure and manage resources for your OpenShift AI users.

Control Red Hat usage data collection

Choose whether to allow Red Hat to collect data about OpenShift AI usage in your cluster. Usage data collection is enabled by default when you install OpenShift AI on your OpenShift cluster.

Apply autoscaling to your cluster to reduce usage costs

Use the cluster autoscaler to adjust the size of your cluster to meet its current needs and optimize costs.

Manage resource usage by stopping idle notebooks

Reduce resource usage in your OpenShift AI deployment by automatically stopping notebook servers that have been idle for a period of time.

Implement model-serving runtimes

OpenShift AI provides support for model-serving runtimes. A model-serving runtime provides integration with a specified model server and the model frameworks that it supports. By default, OpenShift AI includes the OpenVINO Model Server runtime. However, if this runtime doesn't meet your needs (for example, if it doesn't support a particular model framework), you can add your own custom runtimes.

Install in a disconnected environment

OpenShift AI Self-Managed supports installation in a disconnected environment. Disconnected clusters are on a restricted network, typically behind a firewall and unable to reach the Internet. In this case, clusters cannot access the remote registries where Red Hat provided OperatorHub sources reside. In this case, you deploy the OpenShift AI Operator to a disconnected environment by using a private registry in which you have mirrored (copied) the relevant images.

Manage accelerators

Enable NVIDIA graphics processing units (GPUs) or Intel Gaudi AI accelerators in OpenShift AI and allow your data scientists to use compute-heavy workloads.

CHAPTER 3. TRY IT

Data scientists and developers can try OpenShift AI and access tutorials and activities in the Red Hat Developer sandbox environment.

IT operations administrators can try OpenShift AI in your own cluster with a 60-day product trial.

CHAPTER 4. GET IT

Managed cloud service

You have the following options for subscribing to OpenShift AI as a managed service:

- For OpenShift Dedicated, subscribe through Red Hat.
- For Red Hat OpenShift Service on Amazon Web Services (ROSA), subscribe through Red Hat or subscribe through the AWS Marketplace.

Self-managed software

To get Red Hat OpenShift AI as self-managed software, sign up for it with your Red Hat account team.