TD SYNNEX

OpenShift AI

Scaling your AI workload with OpenShift AI

Red Hat

Authorized Distributor

Sensitivity: Public

Date	Author	Version
14.05.2025	Sejid Canoski	1.0

Sejid Canoski

Technical Presales Consultant / Architect

Sejid Canoski is a solutions architect at TD SYNNEX. He designs and validates partner solutions based on IBM Security, IBM Power and IBM Storage products.

He also teaches IBM Security and IBM Power at the TD SYNNEX Academy.

When he is not busy with his main job he loves to tinker with IBM Cyber Resiliency solutions and watsonx.ai / watsonx.governance.

TD SYNNEX Germany GmbH & Co. OHG
Kistlerhofstraße 75
81379 Munich

E-Mail: Sejid.canoski@tdsynnex.com
Tel.: +49 (0) 89 4700 – 1812
Mobil: +49 (0) 176 200 83730

LinkedIn   Slack   IBM TechXchange Community

Todays Agenda

Scaling AI workloads with OpenShift AI

- Recap of last sessions
- Challenges with enterprise AI
- Breakdown OpenShift AI

TD SYNNEX

The AI journey

Frequently asked questions for the AI journey

Where to start?

How to support?

How to scale?

There is no one way.

Different adoption strategies can require different approaches. Every starting point is unique.

Flexibility matters.

From where you inference or train your model, to ensuring control over data and costs.

Optimize your use cases.

Expanding AI footprint and evolving technology is difficult to manage.

Will it be a challenging journey?

Efficiency

Accessibility

Flexibility

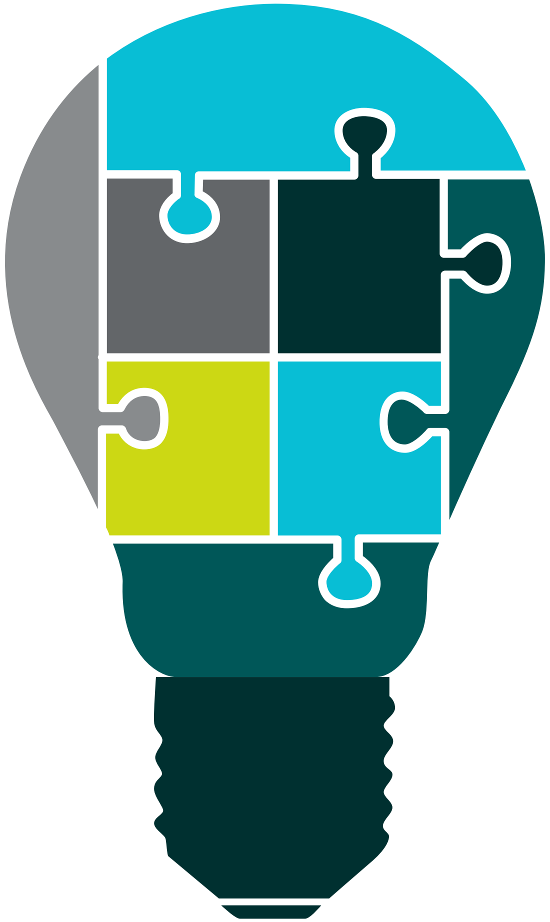
TD SYNNEX

Necessary steps for the transformative journey

Create a clear strategy on business value

Leadership commitment

Operating models that scale



In-house talent bench

Technology that empowers innovation

Change management

What is RHEL AI?

Granite family models

Open source-licensed LLMs, distributed under the Apache-2.0 license, with complete transparency on training datasets.

InstructLab model alignment tools

Scalable, cost-effective solution for enhancing LLM capabilities and making AI model development open and accessible to all users.

Optimized bootable model runtime instances

Granite models & InstructLab tooling packaged as a bootable RHEL image, including Pytorch/runtime libraries and hardware optimization (NVIDIA, Intel and AMD).

Enterprise support, lifecycle & indemnification

Trusted enterprise platform, 24x7 production support, extended model lifecycle and model IP indemnification by Red Hat.

TD SYNNEX

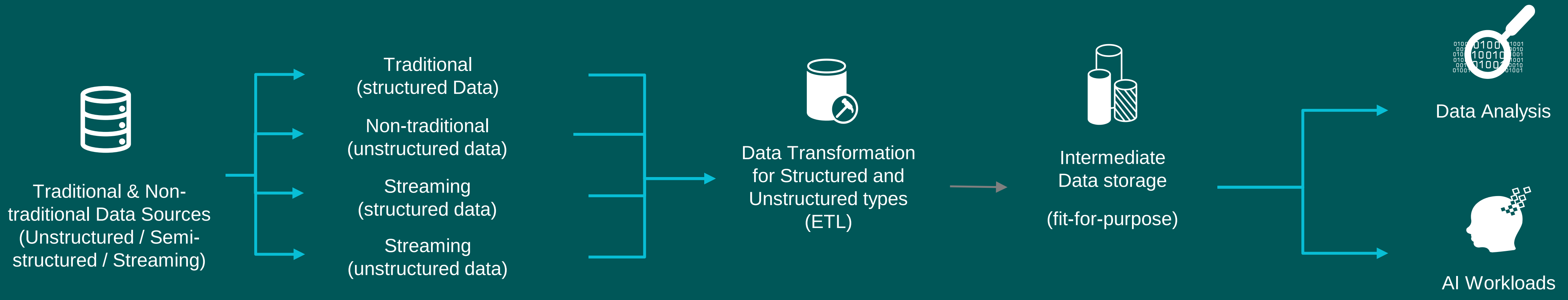
Red Hat Enterprise Linux AI

Foundation Model Platform

Seamlessly develop, test, and run Granite family large language models (LLMs) for enterprise applications.

TD SYNNEX

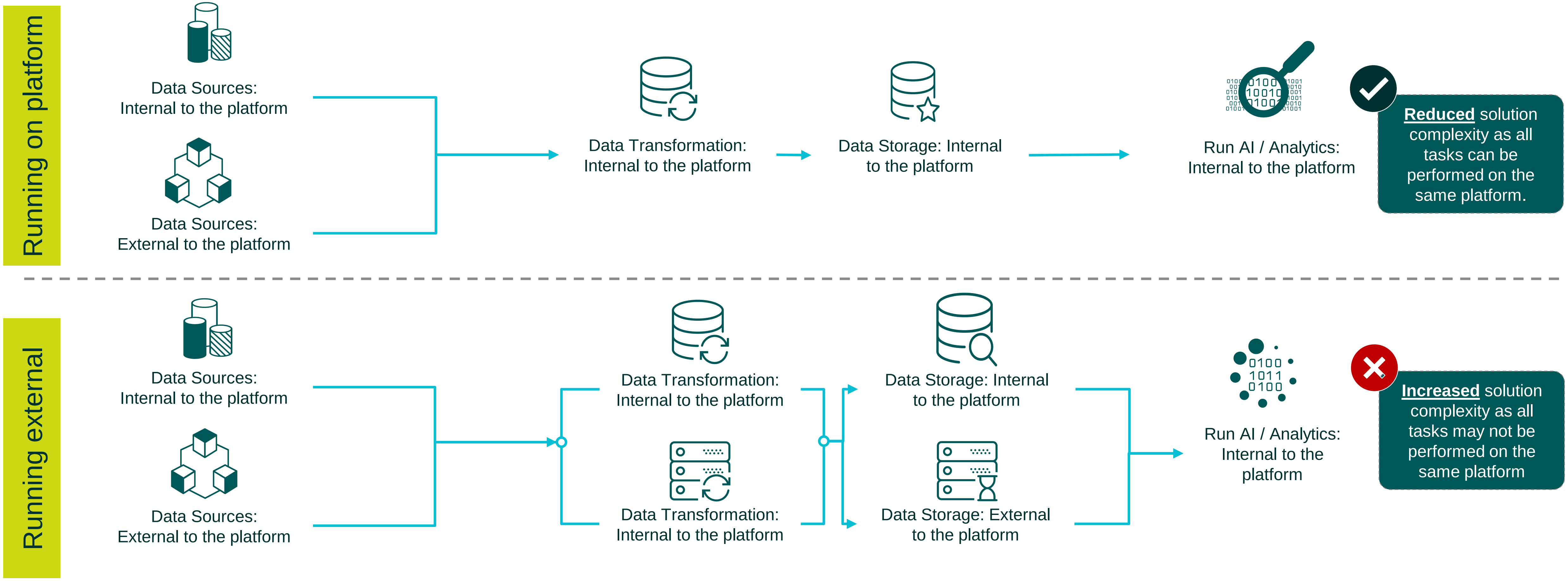
Accelerate AI Efficiently by simplifying solution architecture



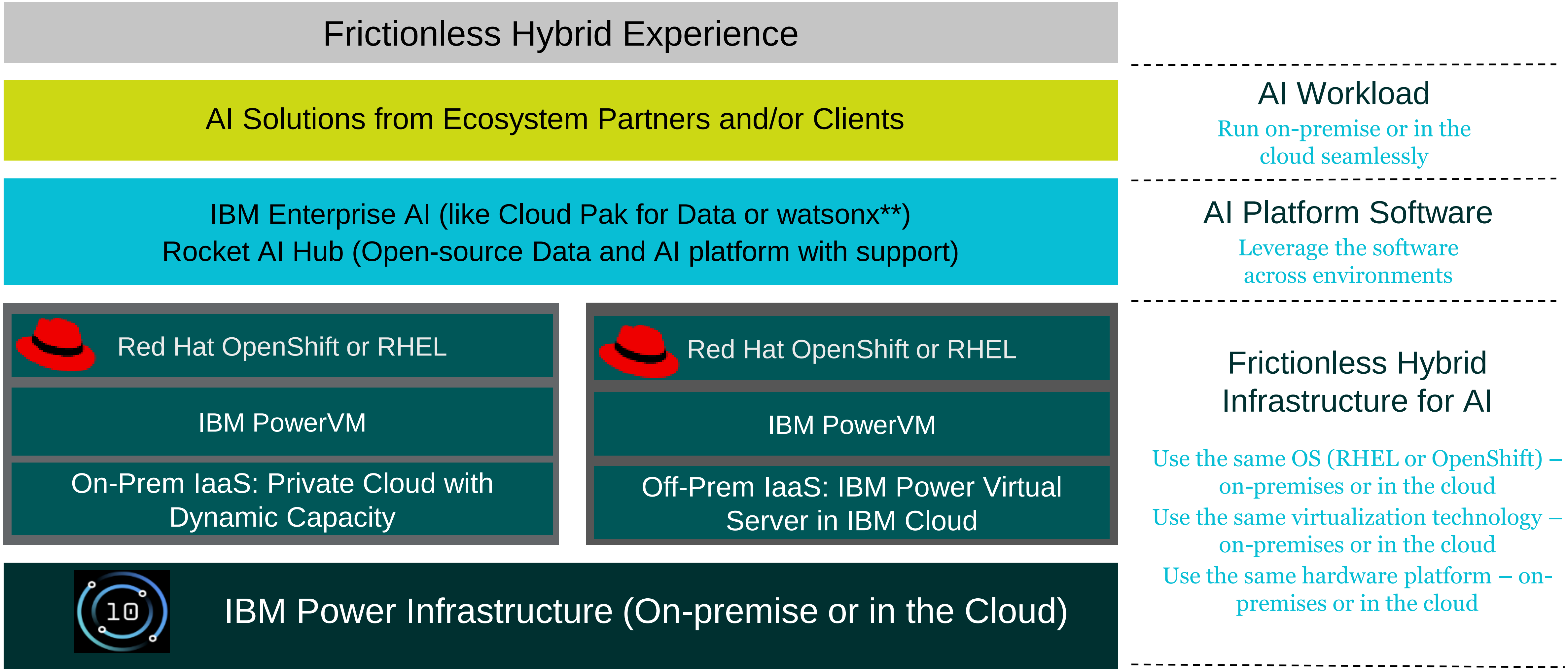
Typical AI Workflow

TD SYNNEX

Reduce complexity: run on workloads on the same platform hardware accelerated



Orchestrate AI Flexibly with frictionless hybrid infrastructure



TD SYNNEX

The power of community that Red Hat OpenShift AI and RHEL AI offer

Red Hat OpenShift AI

RHEL AI

Red Hat OpenShift AI	**Model serving & monitoring** Deploy models anywhere with consistent cloud-to-edge production deployment, and monitoring capabilities	**Resource optimization and management** Create model pipelines for tuning and alignment, validation and delivery; Share and optimize resources across teams
RHEL AI	**Foundation model runtime engine** Develop, deploy, and test open Granite language/code models, supported & indemnified from Red Hat and IBM	**Developer-friendly model toolkit** Bring rapid LLM experimentation to developers via supported InstructLab CLI

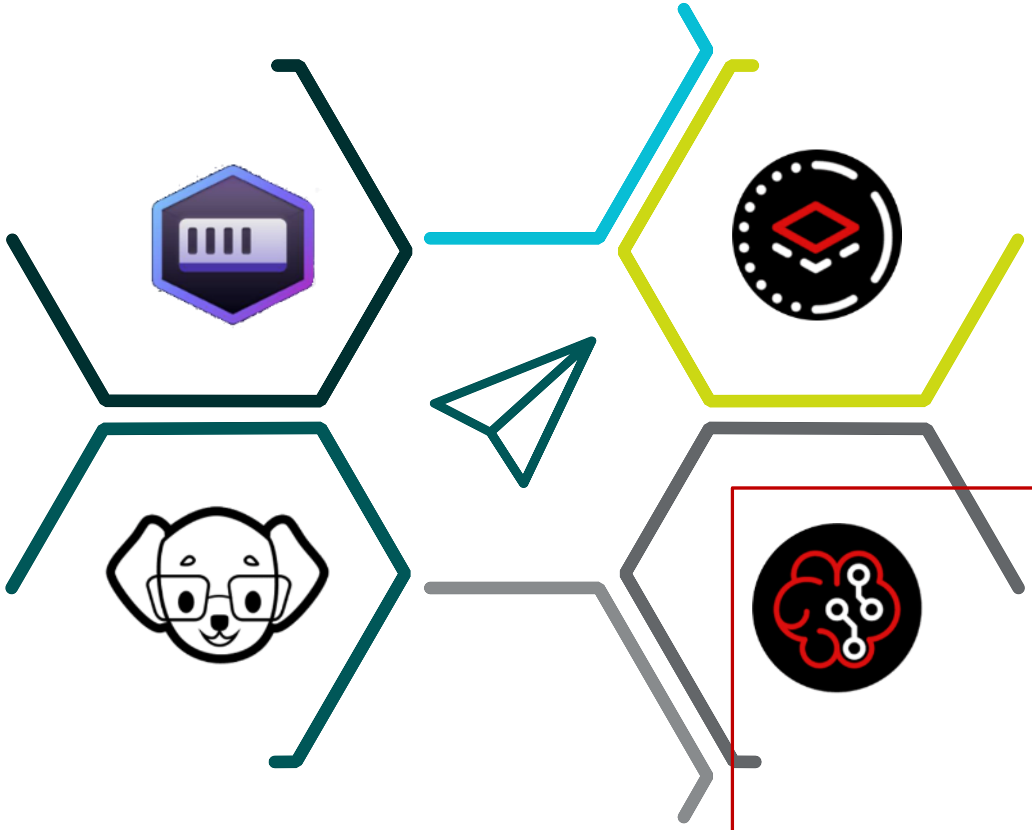
Runs on any infrastructure

Physical · Virtual · Private · Public

Shortly summarized …

Podman Desktop

As unified tools, Podman, Podman Desktop, and Podman AI Lab Extensions offer a local development AI model experimentation platform that includes a graphic user interface (GPU), making it ideal for novices and experts alike.



RHEL AI

Once the training data or user is validated on local hardware, integrating and training large amounts of data into the model necessitates more advanced hardware. Red Hat Enterprise Linux (RHEL) AI offers a scalable solution with InstructLab software, transitioning from local commodity hardware to cloud or private data center server solutions equipped with advanced server-grade hardware, accelerators, and Red Hat Enterprise Linux operating systems.

Instructlab

InstructLab helps customers understand AI model augmentation using standard laptop/desktop hardware via an easy-to-use command-line tool that delivers a hands-on approach to automated model fine tuning via the Large-Scale Alignment for ChatBots (LAB) methodology and using synthetic data generation to reduce the manual effort of creating and validating model training data.

OpenShift AI

Red Hat OpenShift AI serves as a comprehensive machine learning operations platform, potentially being the first choice for enterprise customers while representing the final stage of progression. It includes tools for automating data ingestion pipelines, managing AI training and serving, and monitoring, providing IT teams with a fully supported, scalable hybrid cloud solution for excelling in generative AI solutions.

TD SYNNEX

Common Technical Challenges in AI Platform implmentation

AI models and approaches

Models vary in size, breadth, and depth, and not all are created equal. Models can be multimodal (i.e., capable of processing different types of inputs), or they can be paired together to solve different use cases. Models are available as proprietary and open models, as well as models delivered as a service (e.g. OpenAI GPT4, Google Gemini), and delivered as open software (e.g. Meta Llama, Mistral Al, Databricks DBRX).

Generative and predictive AI applications

Most organizations have already implemented predictive AI solutions in-house that are powering critical applications, which need to be maintained. More than ever, organizations are searching for solutions that allow them to develop and run both generative AI and predictive AI applications in parallel while being able to react to demand, market dynamics, and customers needs.

Cloud vs. On-premise trade-offs

Both predictive and generative AI models require large data sets, skilled personnel, and specialty hardware. These constraints will vary depending on where the organization trains or deploys its models.

AI software-as-a-service (SaaS) solutions provide quick access to ready-to-use, highly performant models and on-demand hardware resources, which can speed time-to-market and at the same time require organizations to be flexible on data migration and accessibility. For organizations that prioritize data gravity and governance, on-premise solutions could offer an advantage over cloud offerings.

Red Hat provides support for a variety of AI model types, including foundation models, large language models (LLMs), generative models, and machine learning models.

Red Hat AI portfolio

Red Hat
Enterprise Linux AI

Foundation model platform for developing, testing and running Granite family LLMs

- Provides a simplified approach to get started with generative AI that includes open-source models
- Makes Ai accessible to developers and domain experts with little data science expertise
- Provides the ability to do training & inference on individual production server deployments

Red Hat
OpenShift AI

Integrated MLOps platform for model lifecycle management at scale anywhere

- Provides support for both generative and predictive AI models with *bring-your-own-model* (BYOM) approach
- Includes distributed compute, collaborative workflows, model serving and monitoring
- Offers enterprise MLOps capabilities and the ability to scale across hybrid-clouds
- Includes Red Hat Enterprise Linux AI, including the Granite family models

High-level technical features of RHOAI

Model Development Tooling

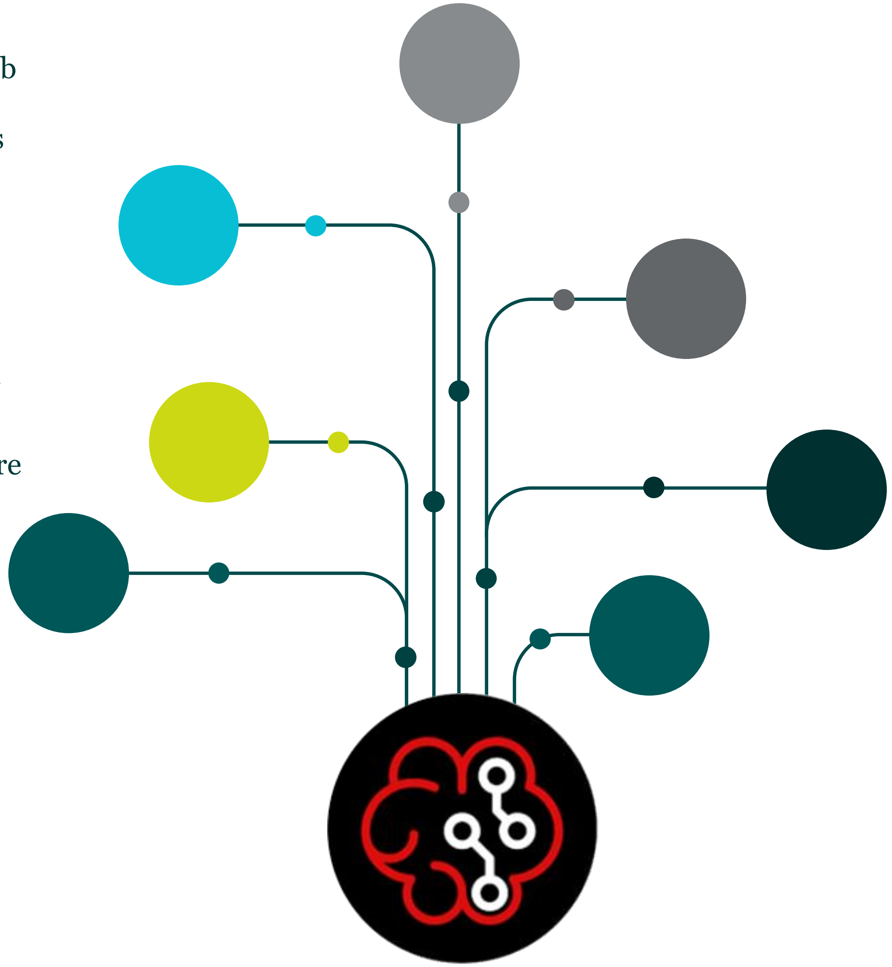
The interactive, collaborative interface based on JupyterLab enhances productivity and innovation by allowing data scientists to work together seamlessly across various stages of model development, from initial exploration to final deployment.

Data and Model Pipelines

The visual editor simplifies the creation and automation of data science pipelines, facilitating data exploration, model training, validation, and storage, while supporting the entire lifecycle of data science projects from data ingestion to model deployment.

Superior Support

Our dedicated engineering teams offer comprehensive support from the operating system to individual tools, ensuring expert assistance at every project stage and enhancing user experience by providing peace of mind and confidence in the platform.



Model Serving

Frameworks for model serving routing simplify the deployment of predictive machine learning or foundation models to production environments, ensuring efficient and effective deployment while providing robust infrastructure for managing model serving at scale.

Model Monitoring

Centralized monitoring capabilities to track models' performance and accuracy. The solution includes out-of-the-box monitoring for key performance and operations metrics for models deployed in production. This ensures that models remain reliable and performant over time.

Extended AI Capabilities

A variety of technology partners are collaborating with Red Hat to certify operability with Red Hat OpenShift AI, ensuring users have access to a wide range of compatible tools and technologies, thereby enhancing the platform's flexibility and extensibility for building comprehensive AI solutions.

Distributed Workload Capabilities

Red Hat OpenShift AI offers distributed workload capabilities on Red Hat OpenShift with minimal friction, providing a seamless experience for data scientists, developers, and ITOps, enabling efficient resource utilization and scalability, and ensuring effective management of complex data science and machine learning tasks.

RHOA simplifies by…

- …supporting both predictive and generative AI models.
- … providing AI-enabled application development and deployment.
- … providing hardware acceleration compatibility.
- … providing access to certified partner products and services.

RHOAI supports operational consistency by ...

- …deploying across various cloud environments.
- …establishing collaborative workflows.
- …providing the ability to run AI at scale across multiple clusters.
- …providing multi-cloud and multi-vendor support.
- …providing tools for managing the entire lifecycle of AI models.

RHOAI creates hybrid cloud flexibility …

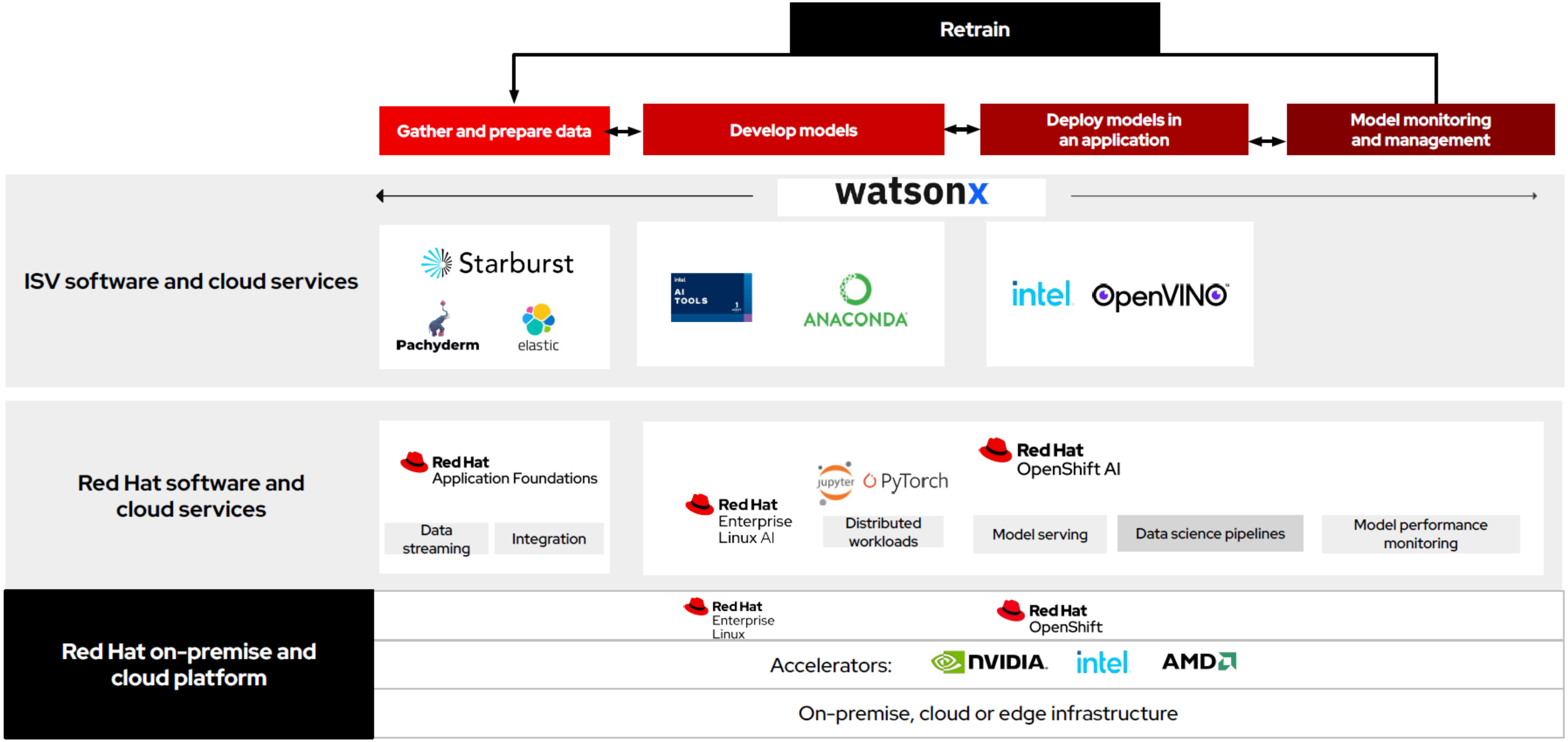
…through distributed compute capabilities.

…with model serving and monitoring tools.

…with regulatory, security, and data requirements.

…with cost management.

Red Hat OpenShift AI components



Red Hat OpenShift AI: Deployment Options

Self-managed software

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.

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).

Red Hat OpenShift AI: Supported Configurations

Red Hat OpenShift AI Self-Managed

You install OpenShift AI Self-Managed by installing the Red Hat OpenShift AI Operator and then configuring the Operator to manage standalone components of the product. RHOAI Self-Managed is supported on OpenShift Container Platform running on an **x86_64**, **ppc64le**, and **s390x** architectures.
This includes the following providers:

- Bare Metal
- Hosted control planes on Bare Metal
- IBM Cloud
- Red Hat OpenStack
- Amazon Web Services
- Google Cloud Platform
- Microsoft Azure
- VMware vSphere
- Oracle Cloud
- IBM Power (Technology Preview)
- IBM Z (Technology Preview)

https://access.redhat.com/articles/rhoai-supported-configs

TD SYNNEX

Applicable Operators

Operator: *Red Hat OpenShift AI* (RHOAI) (old name *Red Hat OpenShift Data Science*)
There are additional operators that you may need to install depending on which features and components you want to utilize

Red Hat OpenShift Serverless Operator

provides a collection of APIs that enables containers, microservices and functions to run "serverless". Required if you want to install the Single-model serving platform component.

Red Hat OpenShift Service Mesh Operator

provides an easy way to create a network of deployed services that provides discovery, load balancing, service-to-service authentication, failure recovery, metrics, and monitoring. Required if you want to install the Single-model serving platform component

Red Hat Authorino Operator

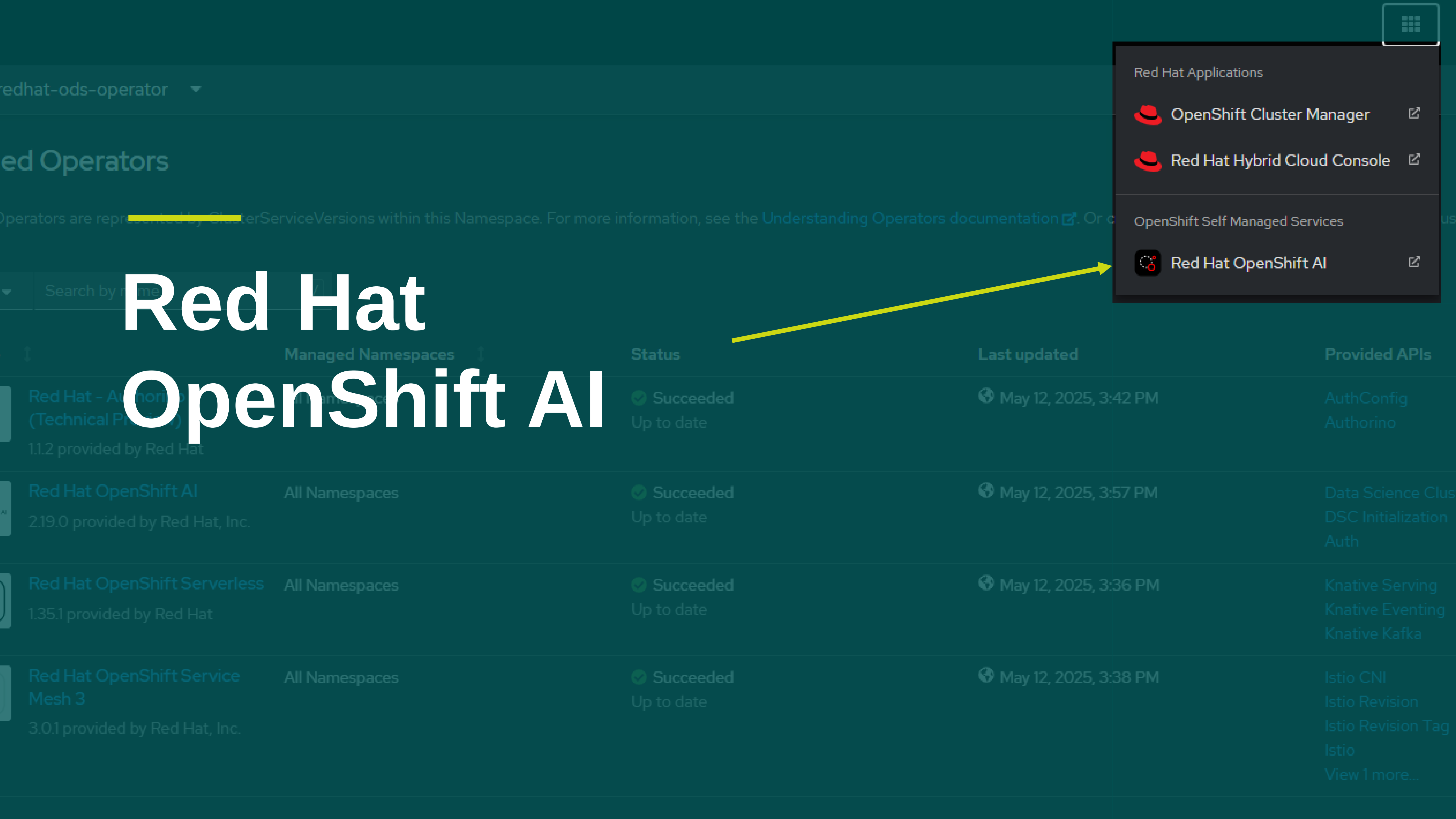
is an open source, Kubernetes-native external authorization service to protect APIs. Required to support enforcing authentication policies in Red Hat OpenShift AI.

Node Feature Discovery Operator (NFD)

a prerequisite for the NVIDIA GPU Operator.

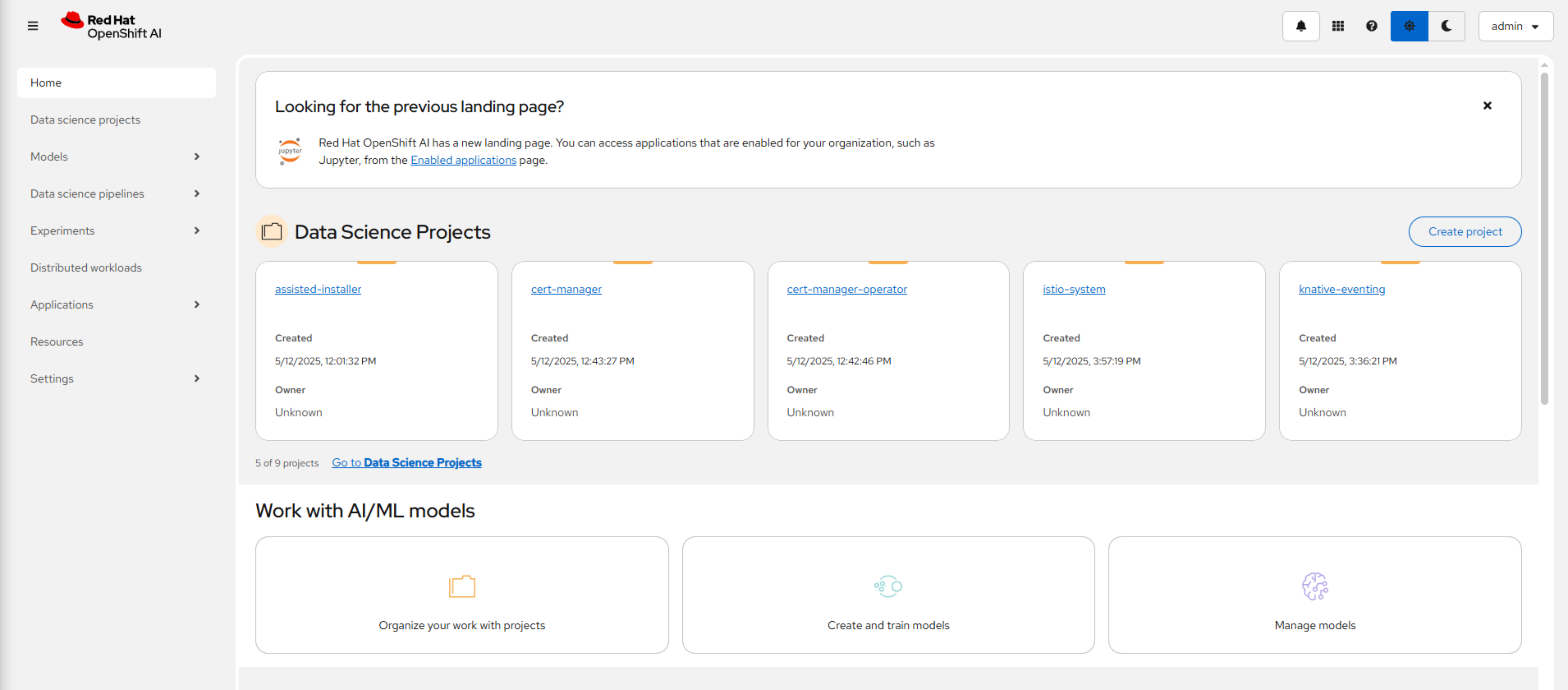
NVIDIA GPU Operator

required for GPU support in Red Hat OpenShift AI.


Red Hat OpenShift AI
redhat-ods-operator
ed Operators
Red Hat Applications
OpenShift Cluster Manager
Red Hat Hybrid Cloud Console
OpenShift Self Managed Services
Red Hat OpenShift AI
Managed Namespaces
Status
Last updated
Provided APIs
Succeeded
Up to date
May 12, 2025, 3:42 PM
AuthConfig
Authorino
1.1.2 provided by Red Hat
Red Hat OpenShift AI
2.19.0 provided by Red Hat, Inc.
All Namespaces
Succeeded
Up to date
May 12, 2025, 3:57 PM
DSC Initialization
Auth
Red Hat OpenShift Serverless
1.35.1 provided by Red Hat
All Namespaces
Succeeded
Up to date
May 12, 2025, 3:36 PM
Knative Serving
Knative Eventing
Knative Kafka
Red Hat OpenShift Service Mesh 3
3.0.1 provided by Red Hat, Inc.
All Namespaces
Succeeded
Up to date
May 12, 2025, 3:38 PM
Istio CNI
Istio Revision
Istio Revision Tag
Istio
View 1 more...

TD SYNNEX

Red Hat OpenShift AI – Dashboard



Data Science Projects

Data Science Projects > ollama-model

ollama-model

Testin ollama-model with OpenShift AI.

Actions

Overview | Workbenches | Pipelines | Models | Cluster storage | Connections | Permissions | Settings

Train models

Workbenches

A workbench is an isolated area where you can work with models in your preferred IDE, such as a Jupyter notebook. You can add accelerators and data connections, create pipelines, and configure cluster storage in your workbench.

Create a workbench →

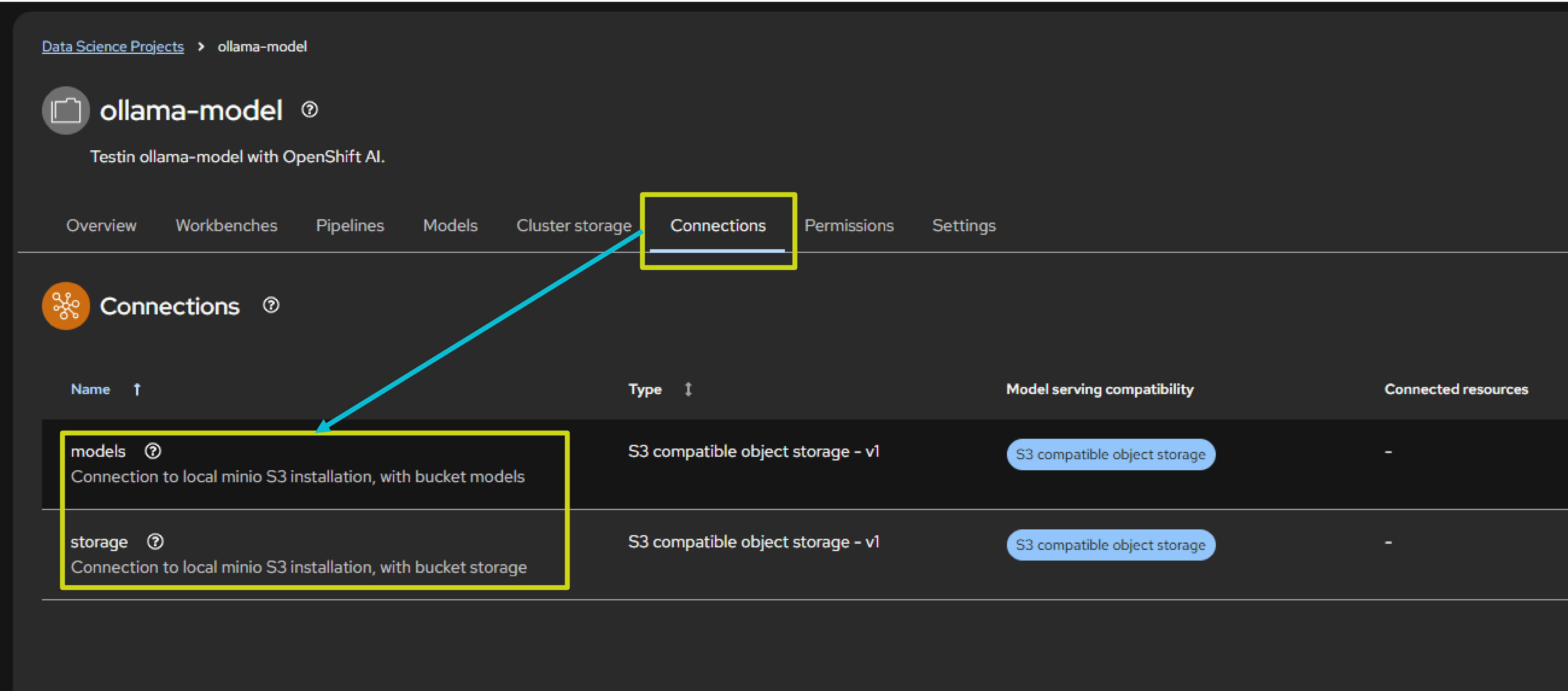
Pipelines

Pipelines are platforms for building and deploying portable and scalable machine-learning (ML) workflows. You can import a pipeline or create one in a workbench. Before you can work with pipelines, you must first configure a pipeline server in your project.

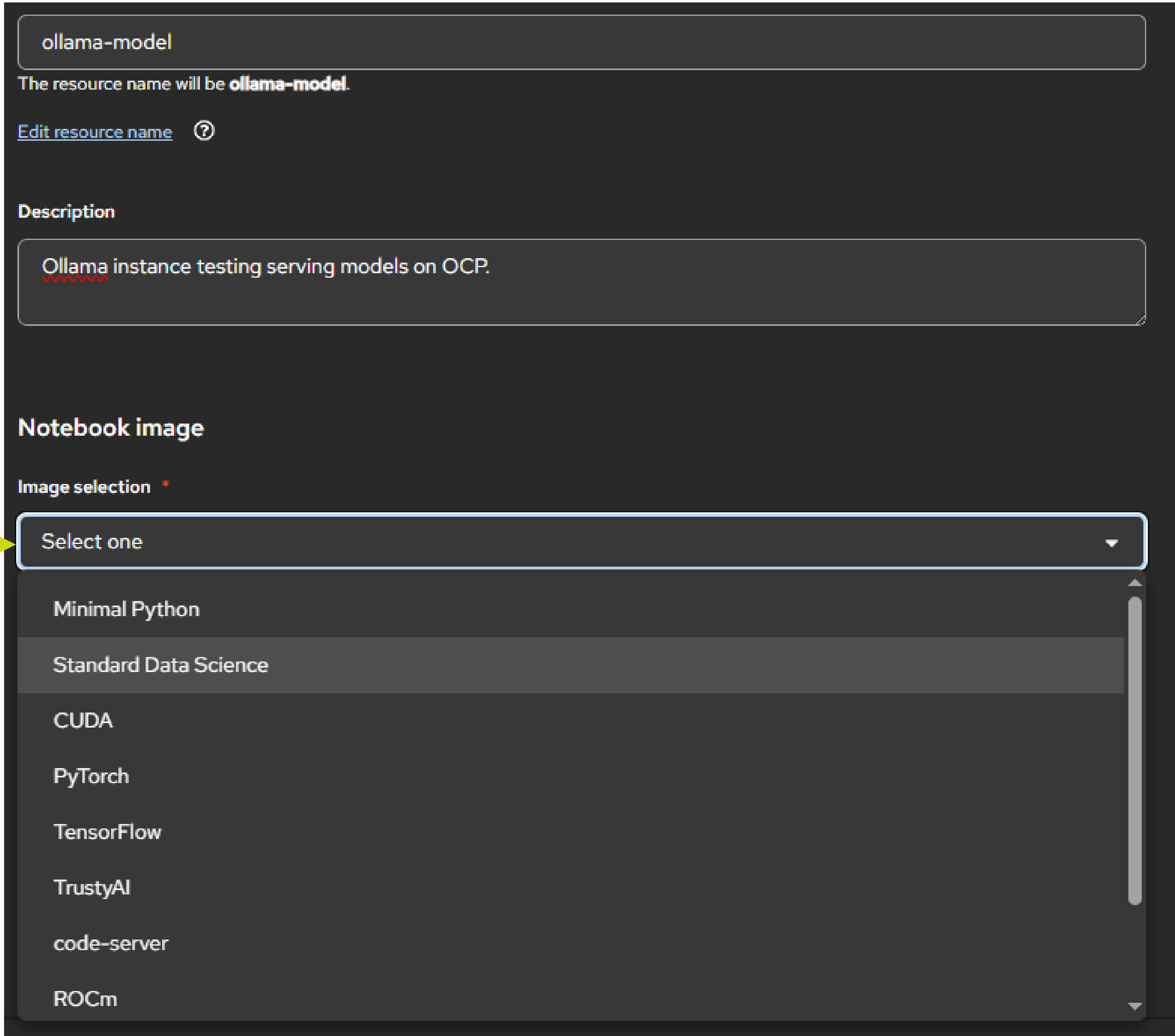
Configure pipeline server

Serve models

Connections and storage



Supported notebook images



Serving models

Data Science Projects > ollama-model

ollama-model

Actions

Testin ollama-model with OpenShift AI.

Overview | Workbenches | Pipelines | **Models** | Cluster storage | Connections | Permissions | Settings

Select the model serving type to be used when deploying from this project.

Single-model serving platform

Each model is deployed on its own model server. Choose this option when you want to deploy a large model such as a large language model (LLM).

Select single-model

Multi-model serving platform

Multiple models can be deployed on one shared model server. Choose this option when you want to deploy a number of small or medium-sized models that can share the server resources.

Select multi-model

You can change the model serving type before the first model is deployed from this project. After deployment, switching types requires deleting all models and servers.

Takeaways

TD SYNNEX

Collaborative tools

RHOAI offers interactive, collaborative tools for data science model development, training, tuning, and serving, along with centralized monitoring capabilities that track the performance and accuracy of models in production.

Distributed workloads

RHOAI supports distributed workloads across multiple environments, enabling organizations to scale their AI operations efficiently.

Simplification

RHOAI simplifies AI deployment with model serving frameworks and tools that integrate seamlessly with existing infrastructure

Optimization

RHOAI includes features that optimize resource usage, helping organizations manage costs effectively

TD SYNNEX

Red Hat

Authorized Distributor

The Power of us

Connect. Empower. Transform.

Questions & Answers

TD SYNNEX

Red Hat

Authorized Distributor

The Power of us

Connect. Empower. Transform.