



Harnessing Red Hat AI on ROSA (AWS)

Practical Life Cycle Tips for Intelligent Applications

Yury Titov

Sr. EMEA Black Belt for Managed Cloud Services

Introduction



Yury Titov

- ▶ former senior EMEA Architect
- ▶ present: senior BlackBelt for Managed Cloud Services
- ▶ always: open source dude

Agenda

- ▶ **How the AI/ML landscape is evolving: market opportunities & challenges**
- ▶ **AI Application Examples vs intelligent Application?**
- ▶ **Challenges of Operationalizing AI ?**
- ▶ **Team topologies and operationalizing models**
- ▶ **Red Hat OpenShift AI - key features and walkthrough**
- ▶ **Why application platforms? Gartner two speed architecture**
- ▶ **Where to start?**
- ▶ **Conclusion**

How the AI/ML landscape is evolving



AI is becoming a part of our everyday lives



Chat GPT



Stable **Diffusion**

watson**X** Code Assistant

IBM Granite Models



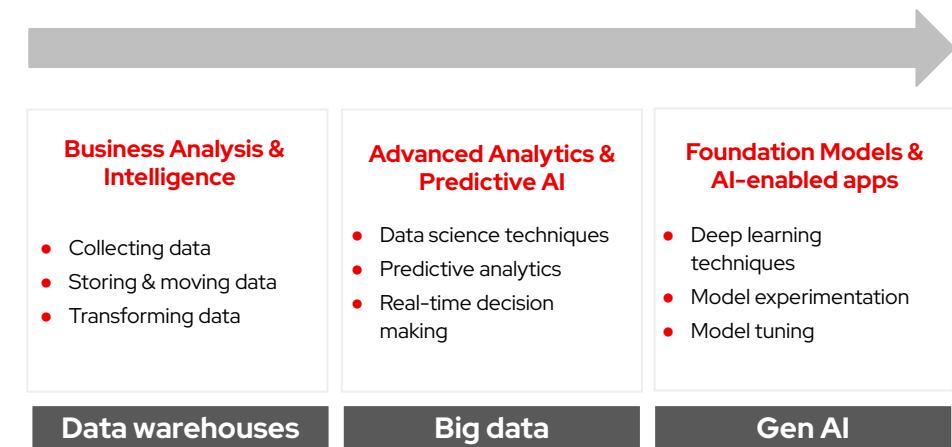
GitHub
Copilot



AI has undergone significant evolution

The evolution of AI: from Business Intelligence to Generative AI

- ▶ Predictive AI runs businesses today
- ▶ Foundation models provide a shortcut for realizing the value of AI



Intelligent Applications?



Examples of intelligent applications

- ▶ **Recommendation engines**

Netflix, Amazon, etc.

- ▶ **Virtual assistant**

Siri, Alexa, etc.

- ▶ **Detecting fraudulent activity**

Money laundering, spam, hacking, insurance

- ▶ **Quantifying risks and making smart decisions**

Insurance, loans

- ▶ **Pattern detection**

Images, videos: how many cars, humans, etc.

- ▶ **Analyze specialized data**

Seismic data for oil and gas

- ▶ **Teach AI to play video games**

AI opponents

- ▶ **Text analysis**

Summarization, accuracy, offensive, plagiarism detection

- ▶ **Medical**

Tumour detection

- ▶ **Customer retention**

Predict who's about to leave

Generative AI Application Examples

- ▶ **Text Generation**

Content creation, chatbots, etc.

- ▶ **Code Generation**

Automate and supplement code development

- ▶ **Image Creation**

Create new images for art, design, games, etc.

- ▶ **Music development**

Create original music based on existing styles

- ▶ **Medical applications**

Suggest new molecules for drug development

- ▶ **Data augmentation (synthetic data)**

Create additional training data for model development

- ▶ **Anomaly detection**

Detect outliers in new data

- ▶ **Content personalization**

Personalize content like product recommendations

- ▶ **Language translation and summarization**

Translate text or summarize long passages

- ▶ **Compliance**

Analyze contracts or other documents for compliance

10

Operationalize AI with Red Hat OpenShift AI



What is Machine Learning?

Machine learning can solve business problems

Artificial Intelligence (AI)

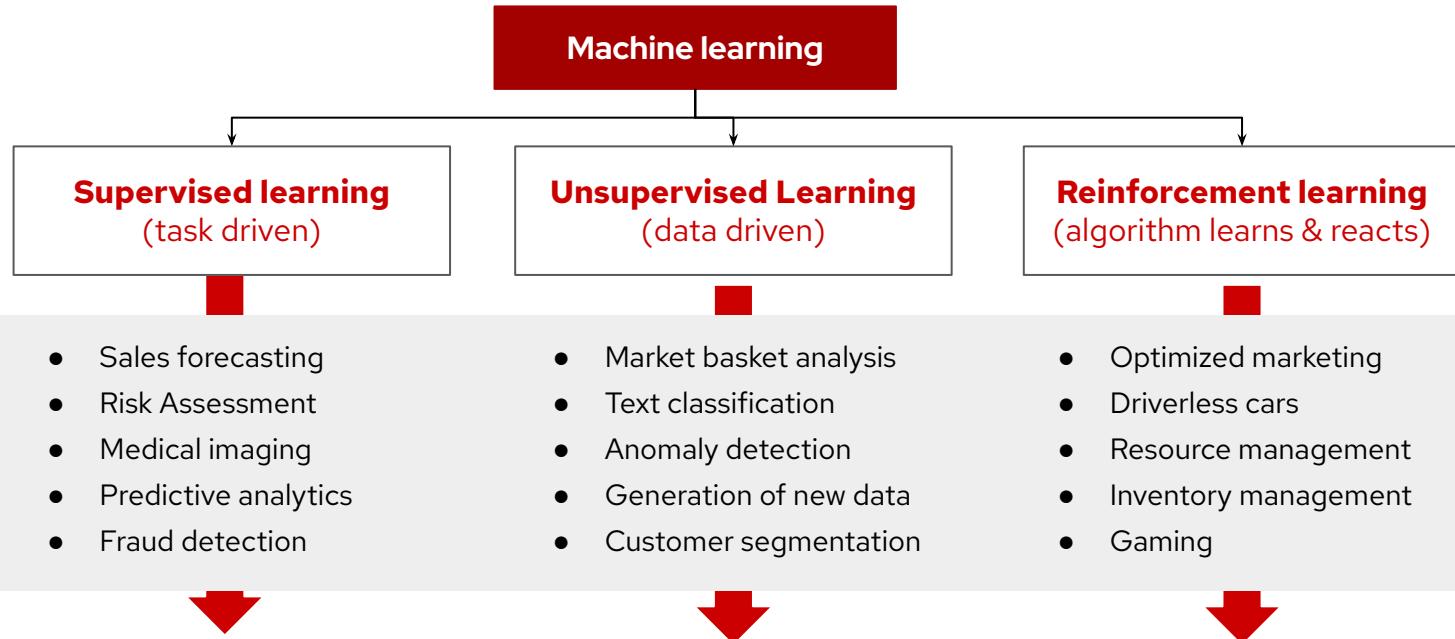
Machine Learning (ML)

Machine Learning is a subset of AI:

- ▶ This technique **empowers computers to learn from data** and improve their performance with time.
 - Training statistical models to extract knowledge and patterns from data
 - Training is done using supervised or unsupervised learning
 - ML models can make accurate predictions and decisions

What are the different types of Machine Learning?

Each type conglomerates a variety of common algorithms

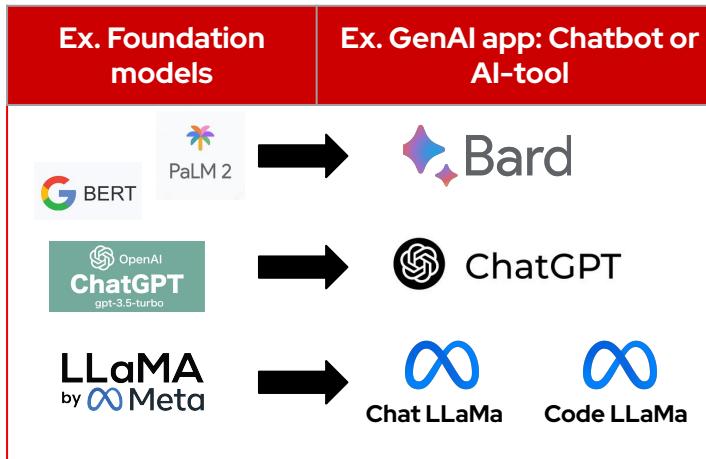


Generative AI applications are powered by foundation models

Foundation models allow developing specialized AI-enabled applications

Benefits of foundation models:

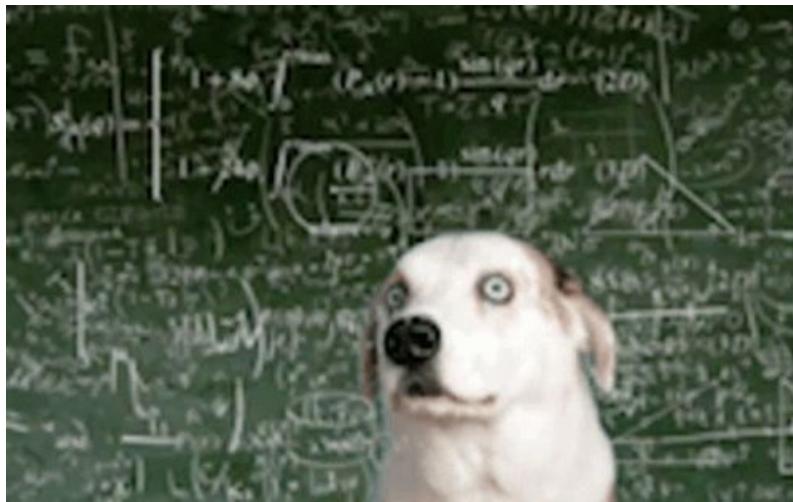
- **Time to value** - alleviates the cost of compute and people
- **Accuracy** - increases with the amount of data use during training
- **Accessibility** - makes advanced AI capabilities available to non-experts
- **Versatility** - offers support for a wide range of tasks and applications



Most common Gen AI applications:

- Text summarization
- Text generation (including code)
- Sentiment analysis
- Classification
- Conversational questions and answers

It's not magic.
It's math.



All of the amazing things that AI and Generative AI can do all comes down to mathematical computation.

- Compute intensive
- Storage intensive
- There are no small workloads
- Quota attainment

Way to production

It all starts with some models...



Real Life View of Technical Teams on AI*

*gathered from real life experience in EMEA ;)



Legacy
Monolith



Modern
Microservices

Real Life View of Technical Teams on AI*

*gathered from real life experience in EMEA ;)



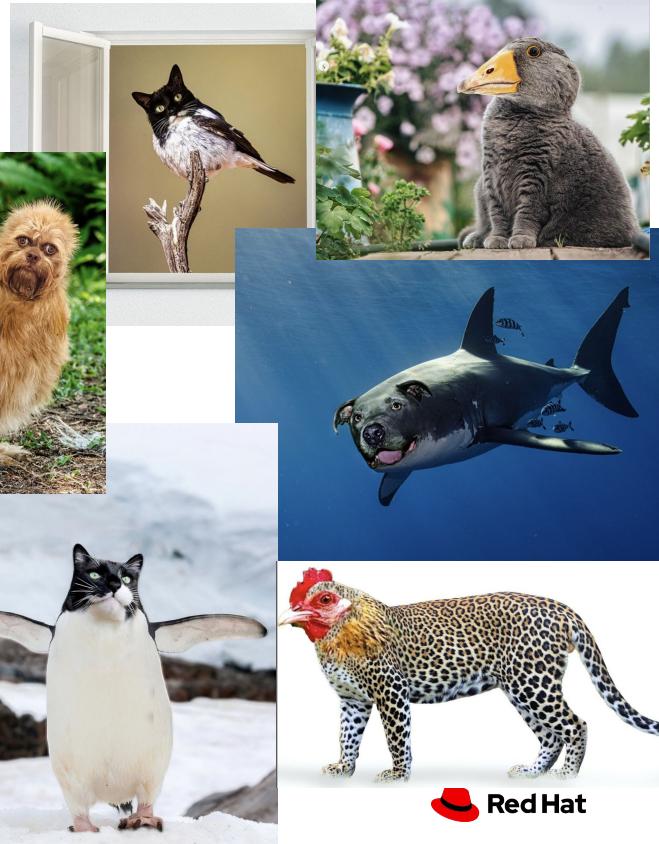
Legacy
Monolith



Modern
Microservices



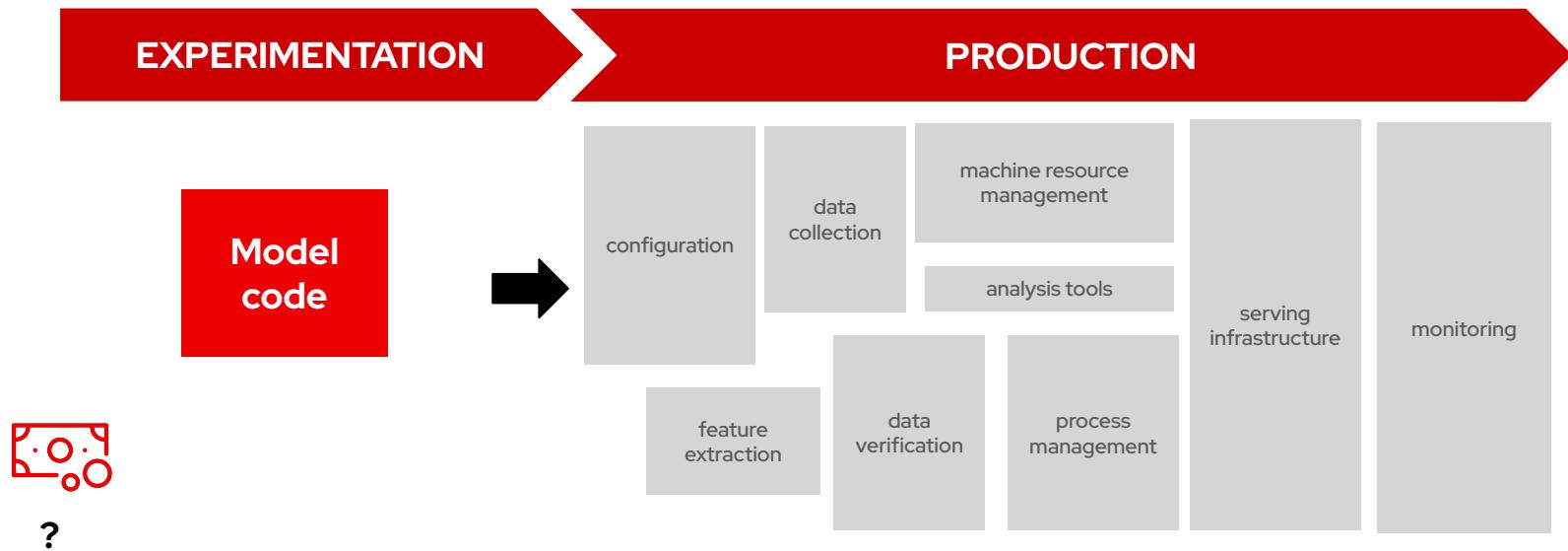
AI



Red Hat

Poorly designed systems lead to failed ML projects

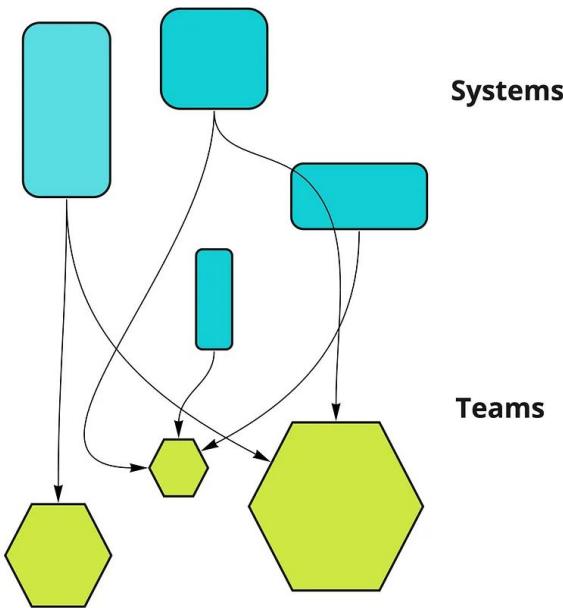
Lack of focus on end-to-end system builds technical debt



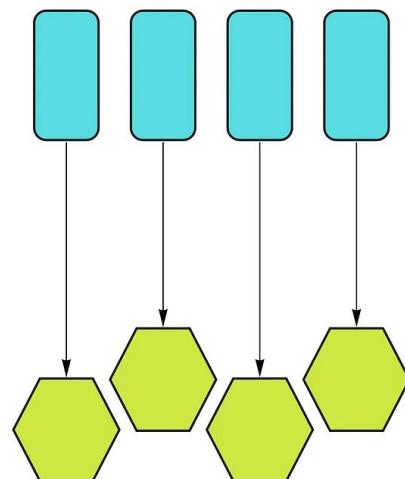
Technical debt is a barrier to production

Expected Outcome

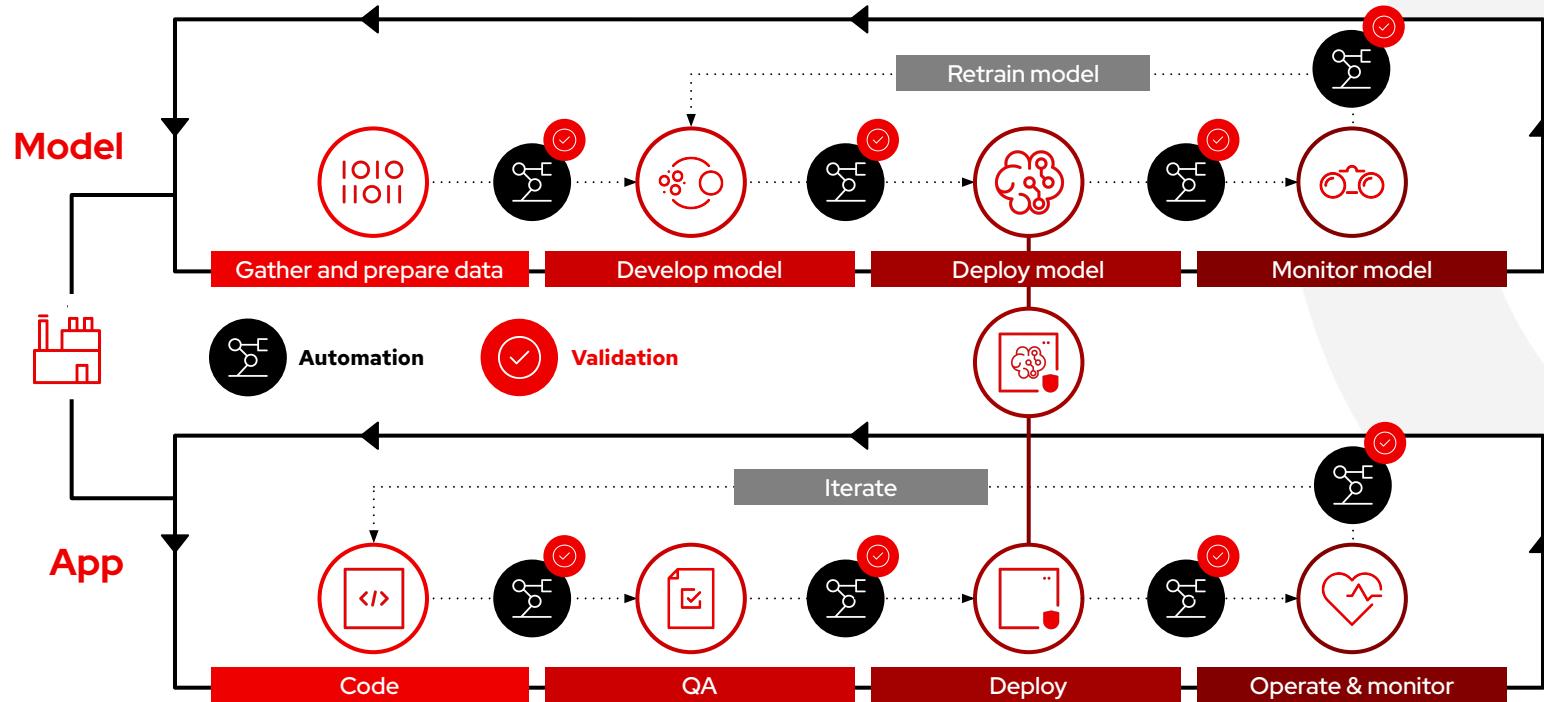
Typical mapping between systems and teams



Mapping between systems and teams in a Team-First approach

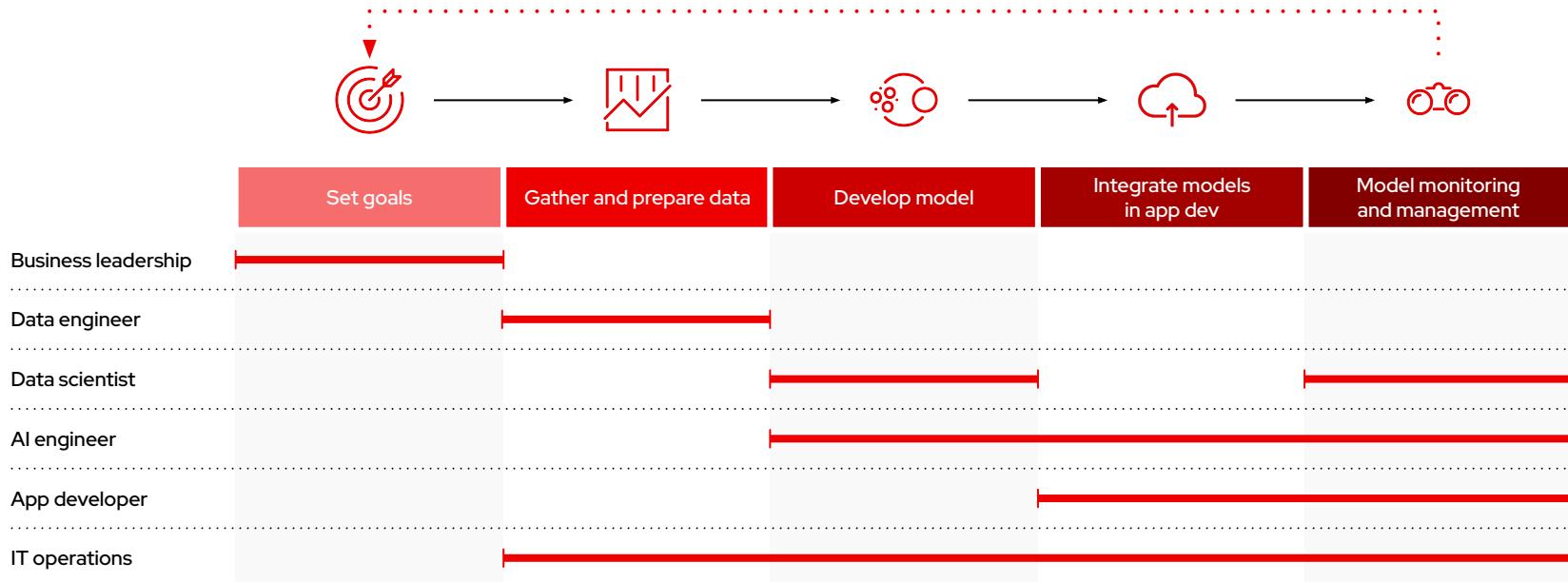


Lifecycle for operationalizing (containerised) models



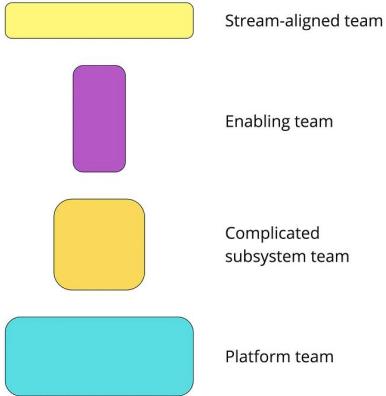
Operationalizing AI/ML requires collaboration

Every member of your team plays a critical role in a complex process

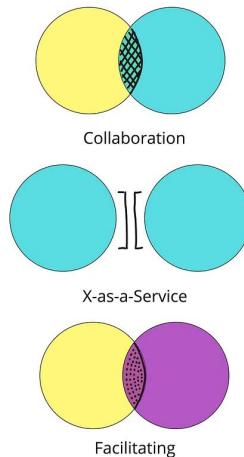


Teams

4 team types



3 interaction modes



1. Stream-aligned teams

aligned to a single delivery stream, such as a product or service (what others might call a “product team” or a “feature team”).

2. Enabling teams

specialists in a particular domain that guide stream-aligned teams

3. Complicated-subsystem teams

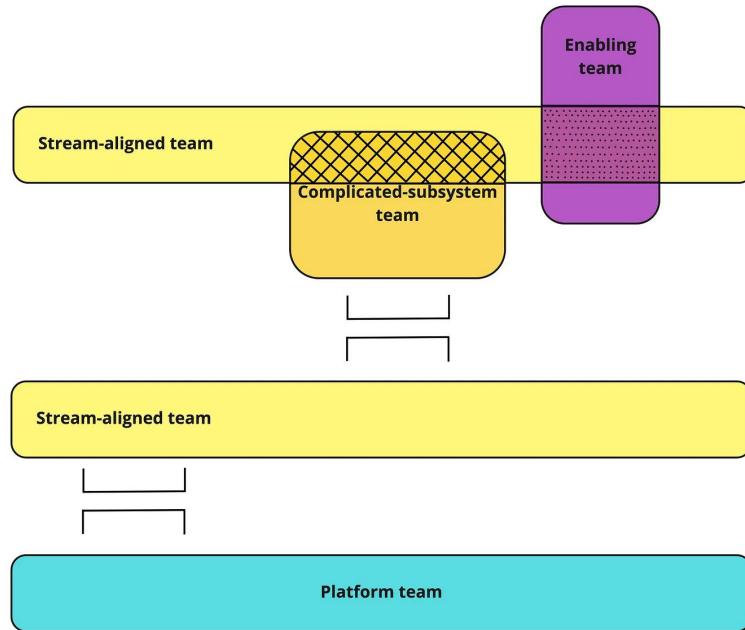
maintain a particularly complex subsystem, such as an ML model

4. Platform teams

provide internal services like deployment platforms or data services

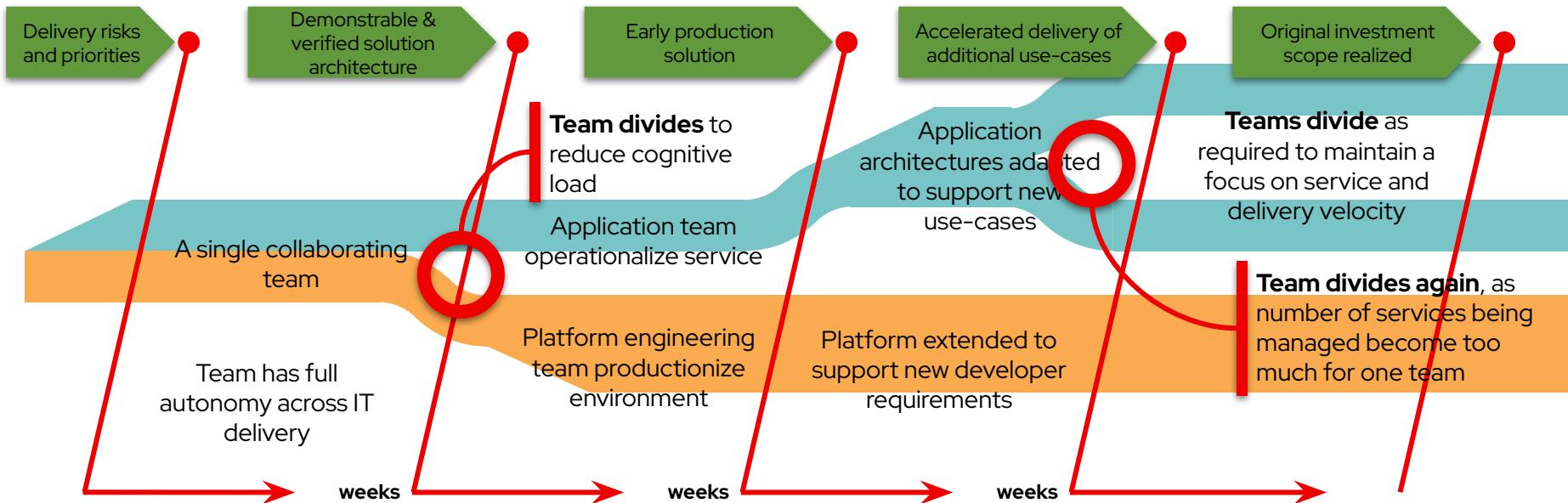


Typical interactions between different teams



Red Hat recommends an evolutionary approach to organisational change

Organisational change is seeded through delivery of specific services, and designed to scale as required



Team Topologies: Organizing Business and Technology Teams for Fast Flow, Pias & Skelton
ISBN: 9781942788812

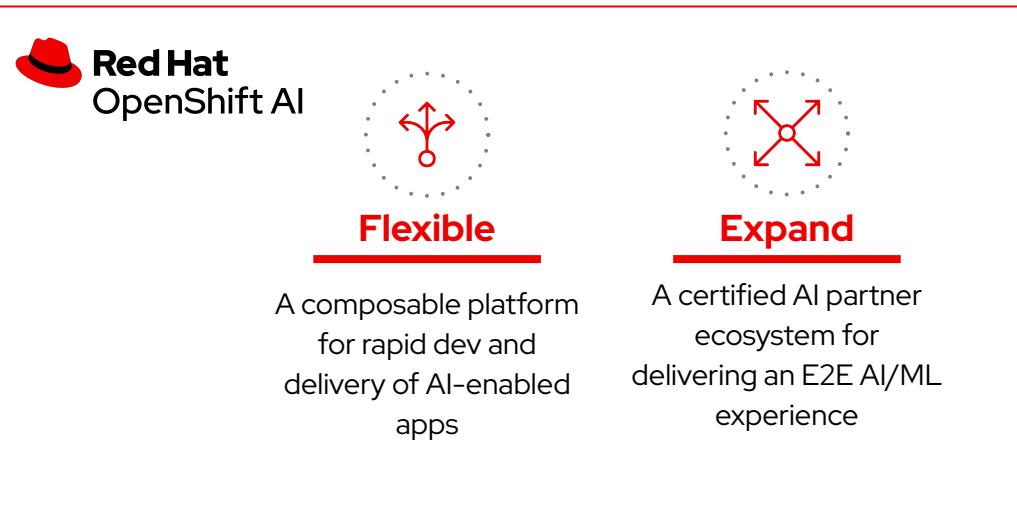
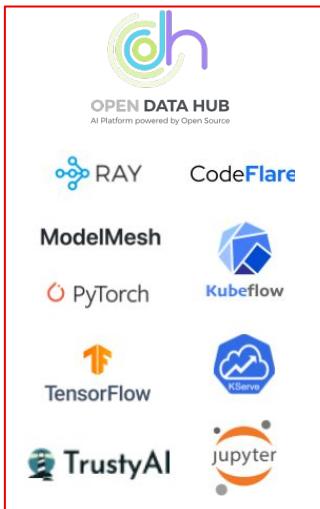
Red Hat's approach is informed by, and aligns with, Team Topologies

Version number here V00000



Simplify AI adoption

Designed to increase AI adoption and enhance trust in AI initiatives



Red Hat OpenShift AI - Key features

Model development

Interactive, collaborative UI for exploratory data science, and model training, tuning and serving

Model serving

Model serving routing for deploying models to production environments

Model monitoring

Centralized monitoring for tracking models performance and accuracy

Data & model pipelines

Visual editor for creating and automating data science pipelines

Distributed workloads

Seamless experience for efficient data processing, model training, tuning and serving

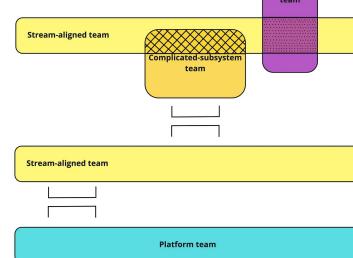
Build an AI platform for E2E AI lifecycle management



Trusted, comprehensive and consistent hybrid application platform for managing the entire application lifecycle

Open hybrid AI/ML platform, built on top of OpenShift, to create and deliver AI-enabled apps securely at scale across hybrid-clouds

Best-of-breed AI technologies from a certified partner ecosystem to complement or extend Red Hat's AI capabilities



Everything in RHOAI has an OpenShift representation

Models and model servers Deploy model Single-model serving enabled

Model name	Serving runtime	Inference endpoint	Status
My Model	OpenVINO Model Server

Framework: onnx-1

Model server replicas: 2

Model server size: Small

1 CPUs, 4Gi Memory requested
2 CPUs, 8Gi Memory limit

Accelerator: None

Number of accelerators: 0



Project: test

InferenceServices > InferenceService details

IS my-model

Details YAML

```
1 apiVersion: serving.kserve.io/v1beta1
2 kind: InferenceService
3 > metadata: ...
6 spec:
7   predictor:
8     maxReplicas: 2
9     minReplicas: 2
10    model:
11      modelFormat:
12        name: onnx
13        version: '1'
14      name: ''
15      resources: {}
16      runtime: my-model
17      storage:
18        key: aws-connection-abc
19        path: mymodel/v01
20 > status: ...
61
```

UI to Yaml

GitOps

The diagram illustrates the GitOps workflow for model deployment, showing the flow of changes from code to deployment.

GitHub Repository (Left): A screenshot of a GitHub repository page for `insurance-claim-processing`. The `model.yaml` file is open, showing the YAML configuration for the InferenceService. The code block contains annotations and labels for the model.

Deployment Interface (Middle): A screenshot of a deployment interface for an InferenceService named `my-model`. The interface shows the `YAML` configuration for the model, which is a copy of the `model.yaml` file from the GitHub repository. The interface also displays the `Details` of the model, including its serving runtime (OpenVINO Model Server) and inference endpoint (onnx-1).

Deployment Interface (Right): A screenshot of a deployment interface for a model named `My Model`. The interface shows the `YAML` configuration for the model, which is a copy of the `model.yaml` file from the GitHub repository. The interface displays the `Details` of the model, including its serving runtime (OpenVINO Model Server), inference endpoint (onnx-1), and model server size (Small). The interface also shows the `Framework` (onnx), `Model server replicas`, and `Model server size` (Small). The interface also shows the `Accelerator` (None) and `Number of accelerators` (0).

Sync Flow: Two red arrows labeled `Sync` indicate the flow of changes from the GitHub repository to the deployment interface. One arrow points from the GitHub repository to the deployment interface, and another arrow points from the GitHub repository to the deployment interface.

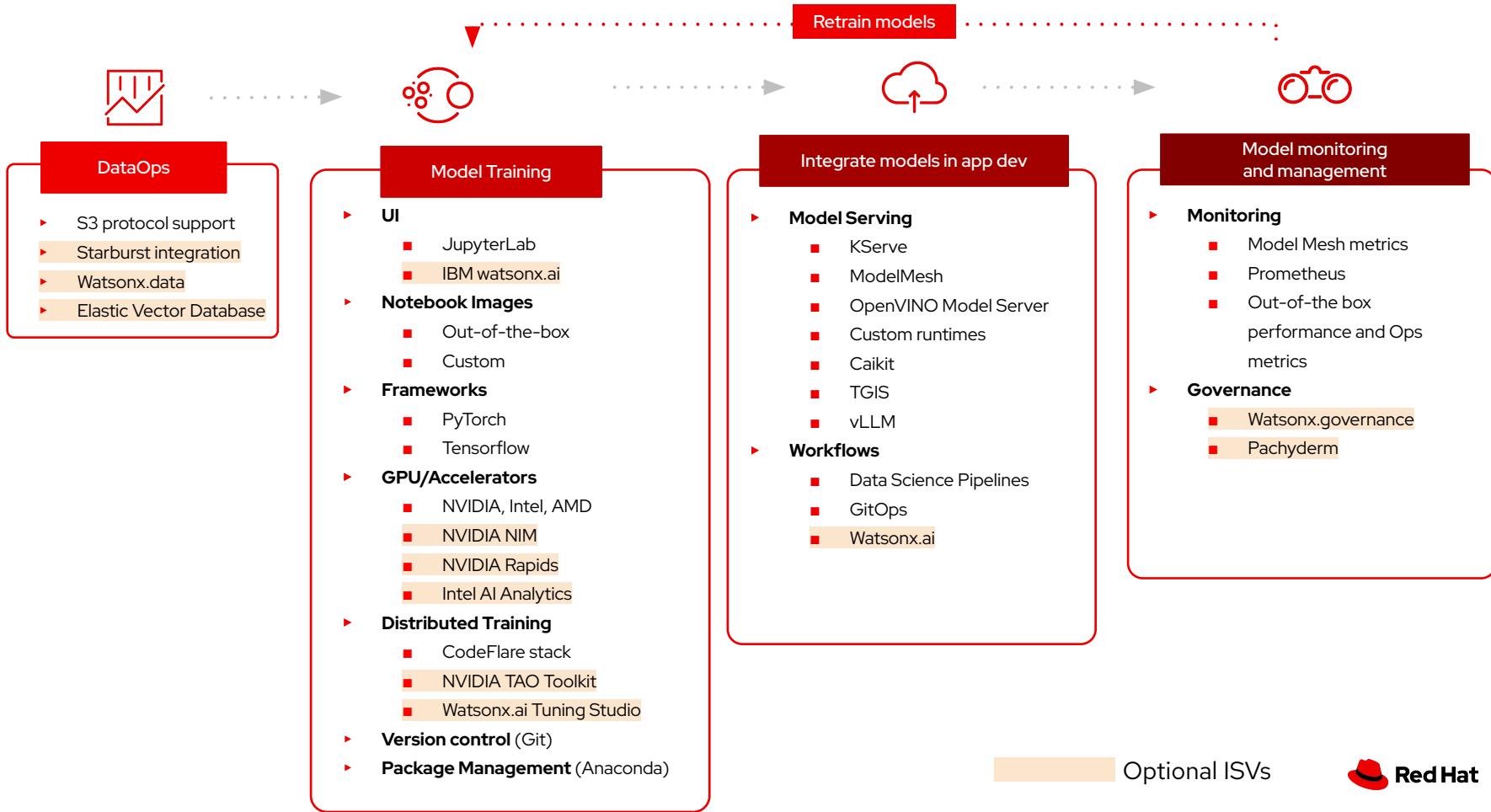
```
apiVersion: serving.kserve.io/v1beta1
kind: InferenceService
metadata:
  annotations:
    openshift.io/display-name: img-det
    serving.kserve.io/deploymentMode: Model
  labels:
    name: "img-det"
    opendatahub.io/dashboard: 'true'
  name: "img-det"
spec:
  predictor:
    model:
      modelFormat:
        name: onnx
        version: '1'
      runtime: ovms
      storage:
        key: aws-connection-minio
        path: accident/
```

Model name	Serving runtime	Inference endpoint	Status
My Model	OpenVINO Model Server	onnx-1	Framework
			Model server replicas
			Model server size
			Small
			1 CPUs, 4Gi Memory requested
			2 CPUs, 8Gi Memory limit
			Accelerator
			None
			Number of accelerators
			0

What is Red Hat OpenShift AI (RHOAI) solving

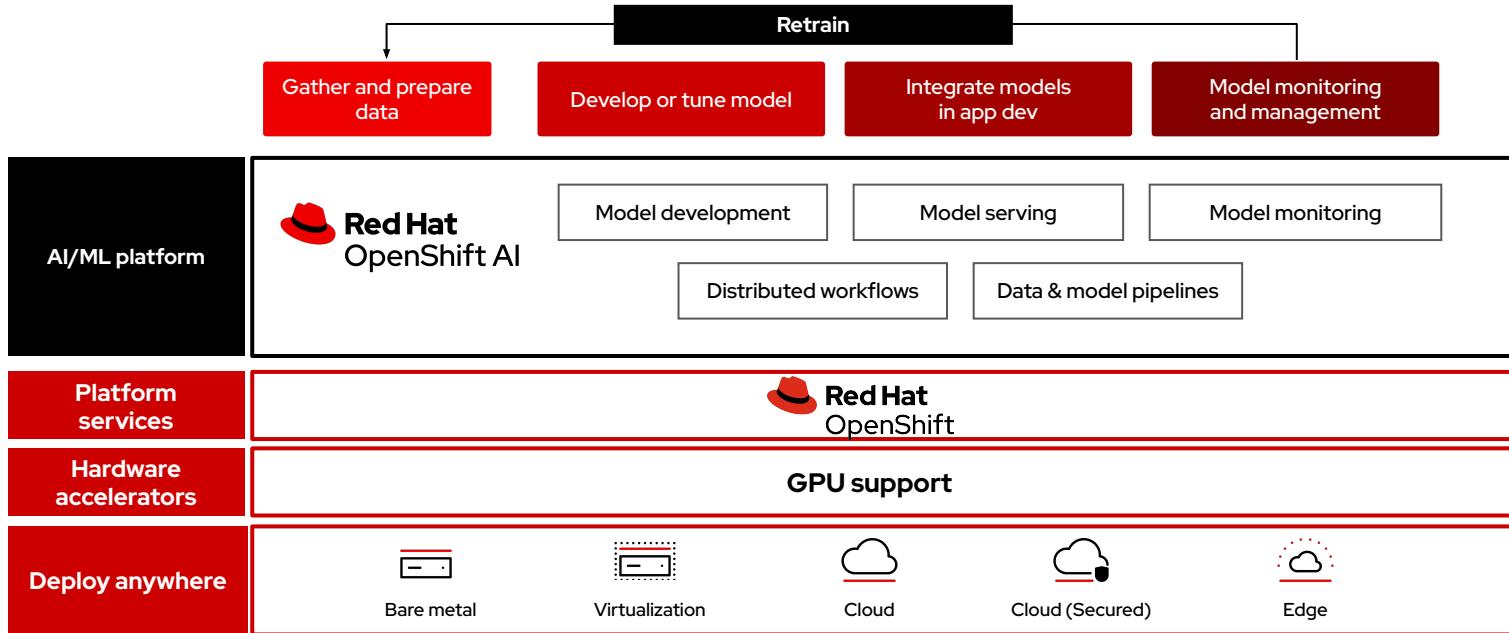
- **MLOps**
 - RHOAI helps you build out an enterprise grade AI and MLOps platform to create and deliver GenAI and predictive models by providing supported AI tooling on top of OpenShift.
 - It's based on OpenShift, a container based application platform that efficiently scales to handle workload demands of AI operations and models.
 - You can run your AI workloads across the hybrid cloud, including edge and disconnected environments.
- **Unified app platform**
 - OpenShift supports the end-to-end application lifecycle. RHOAI extends OpenShift to AI models, getting them into AI models and getting them into production with OpenShift best practices.
 - Seamless collaboration across multiple personas including IT Ops, Data scientists and application developers by providing a unified platform.
- **Extensibility**
 - RHOAI is built to be modular, allowing for a customizable AI/ML stack where you can plug in partners or open source software and technologies where needed to build out an MLOps platform that fits your organization.
- **No vendor lock-in**
 - Thanks to being modular and able to run across the hybrid cloud, you have the freedom to migrate and extend as needed, allowing you to keep up with the speed of AI innovation.

Product (and Open Source Projects) Overview



Red Hat OpenShift AI

Red Hat's AI/ML platform





Red Hat OpenShift AI

Dashboard Application

Data Science Projects

Admin Features

Model Registry

Model Development & Training

Workbenches

- Minimal Python
- PyTorch
- CUDA
- Standard Data Science
- TensorFlow
- VSCode
- RStudio
- TrustyAI

CodeFlare SDK

ISV images

Custom images

Distributed workloads

KubeRay

CodeFlare

Data and model Pipelines

Model Serving

Serving Engines

Kserve

ModelMesh

Serving Runtimes

OVMS (built-in)

Caikit/TGIS (built-in)

Custom

Model Monitoring

Performance metrics

Model explainers

Quality metrics

OpenShift Operators

OpenShift GitOps



OpenShift Pipelines



OpenShift ServiceMesh



OpenShift Serverless

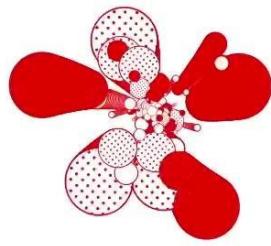


Prometheus



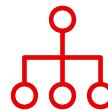
Object Storage





Red Hat OpenShift cloud services

A turnkey application platform with management and support from Red Hat and leading cloud providers



Accelerate time to value

Quickly build, deploy, and run applications that scale as needed.

Operational efficiency

Enhance operational consistency, efficiency and security with proactive management and support.

Focus on innovation

Simplify operations so your teams can refocus on innovation, not managing infrastructure.

Hybrid cloud flexibility

Deliver a consistent experience on premises and in the cloud.

Flavors of RHOAI

Supported deployment options		
Options available	Self-managed RHOAI	Cloud Service RHOAI
Bare metal	✓	
Virtual	✓	
Private cloud	✓	
Red Hat OpenShift on AWS (ROSA)	✓	✓
Azure Red Hat OpenShift (ARO)	✓	(future)
IBM Cloud	✓	
OSD-GCP/OSD-AWS	✓	✓
Edge	(future)	

Flavors of RHOAI

Supported deployment options		
Options available	Self-managed RHOAI	Cloud Service RHOAI
Bare metal	✓	
Virtual	✓	
Private cloud	✓	
Red Hat OpenShift on AWS (ROSA)	✓	✓
Azure Red Hat OpenShift (ARO)	✓	(future)
IBM Cloud	✓	
OSD-GCP/OSD-AWS	✓	✓
Edge	(future)	

Build and run a platform *versus* using a turnkey cloud service



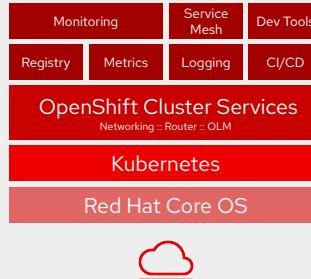
The Parts



xKS + 'native' services



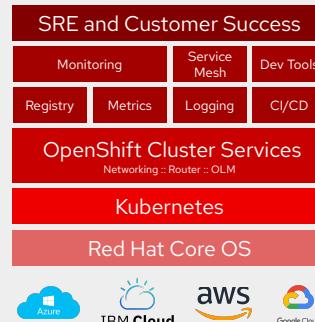
The Assembled Car



- Application Platform -
Self-managed Red Hat OpenShift



The Car & Pit Crew



- Turnkey Application Platform -
Red Hat OpenShift cloud services

"Batteries Included"

... but swappable

Individual components can be swapped out

Eg.

- Using AWS CloudWatch for logging on AWS
- Use specific cloud services or ISV offerings

Move from 24x7 operations to 9-5 innovation

End-to-End support for your entire application platform



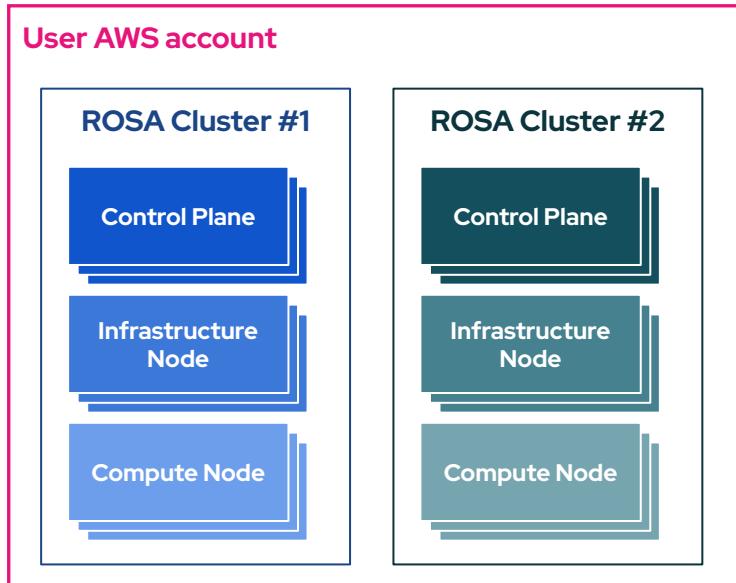
- ▶ OpenShift cloud services includes **full support for worker nodes**
 - Zero downtime upgrades,
 - proactive monitoring
 - automated patching
 - **Compliance and certifications** extend to worker nodes
- ▶ **99.95% financially backed SLA**
- ▶ **24x7 joint support** from Red Hat and cloud provider
- ▶ Automation and **Day 2 Operations** by global SREs

¹ ["The Total Economic Impact™ of Red Hat OpenShift Cloud Services by Forrester, "Jan. 2022.](#)

ROSA Variants

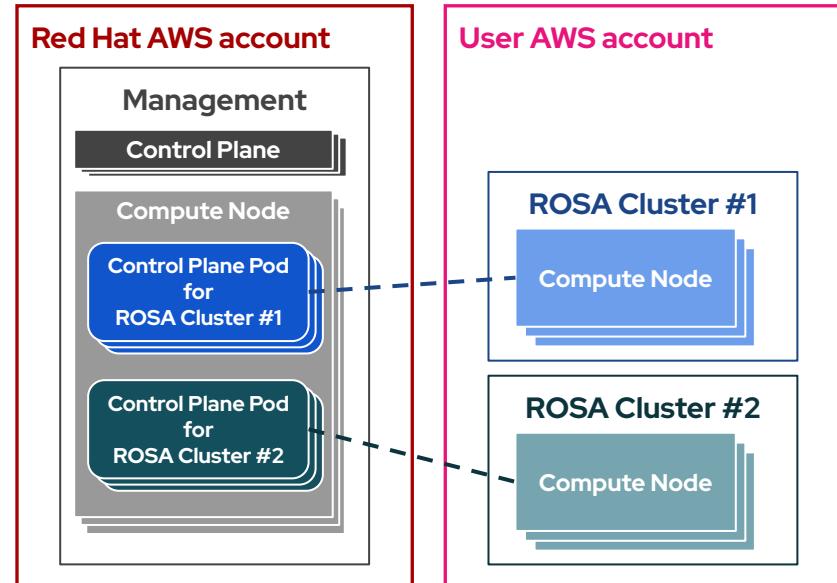
Generally Available

ROSA Classic



1 or 3 AZs per MachinePool

ROSA with Hosted Control Plane (HCP)

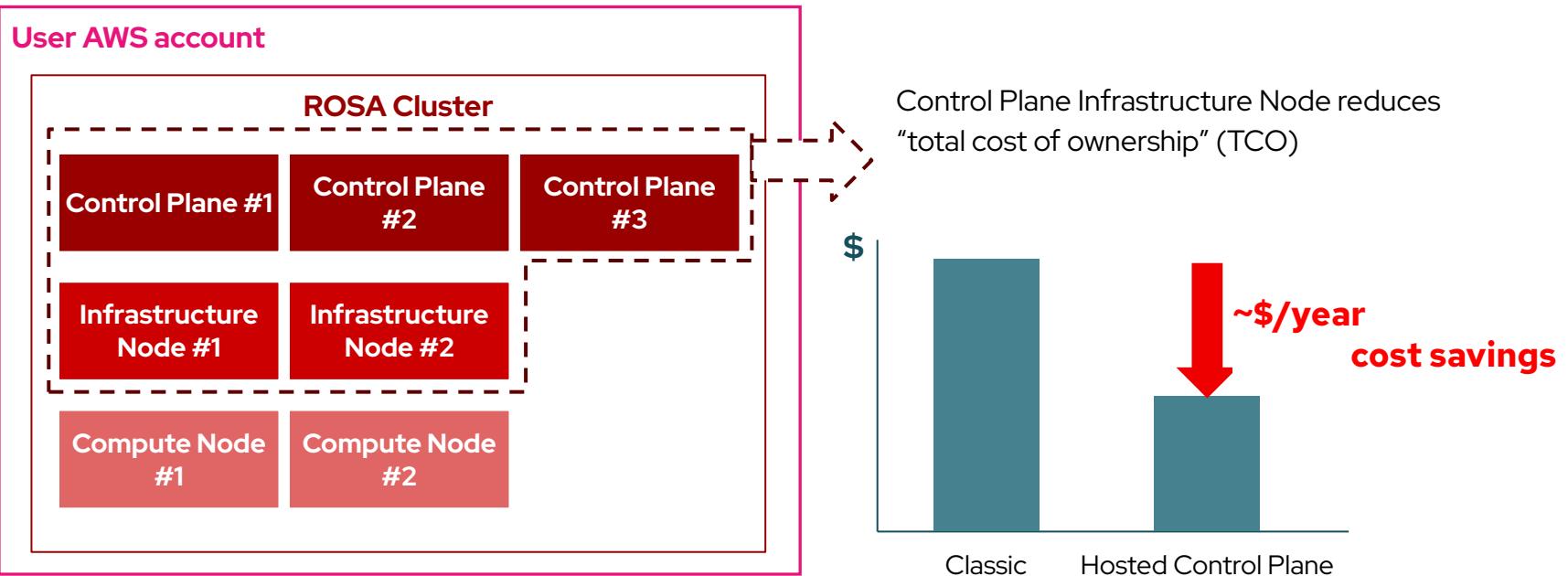


Cluster per Region (each 3 AZs)



Reduce AWS infrastructure costs

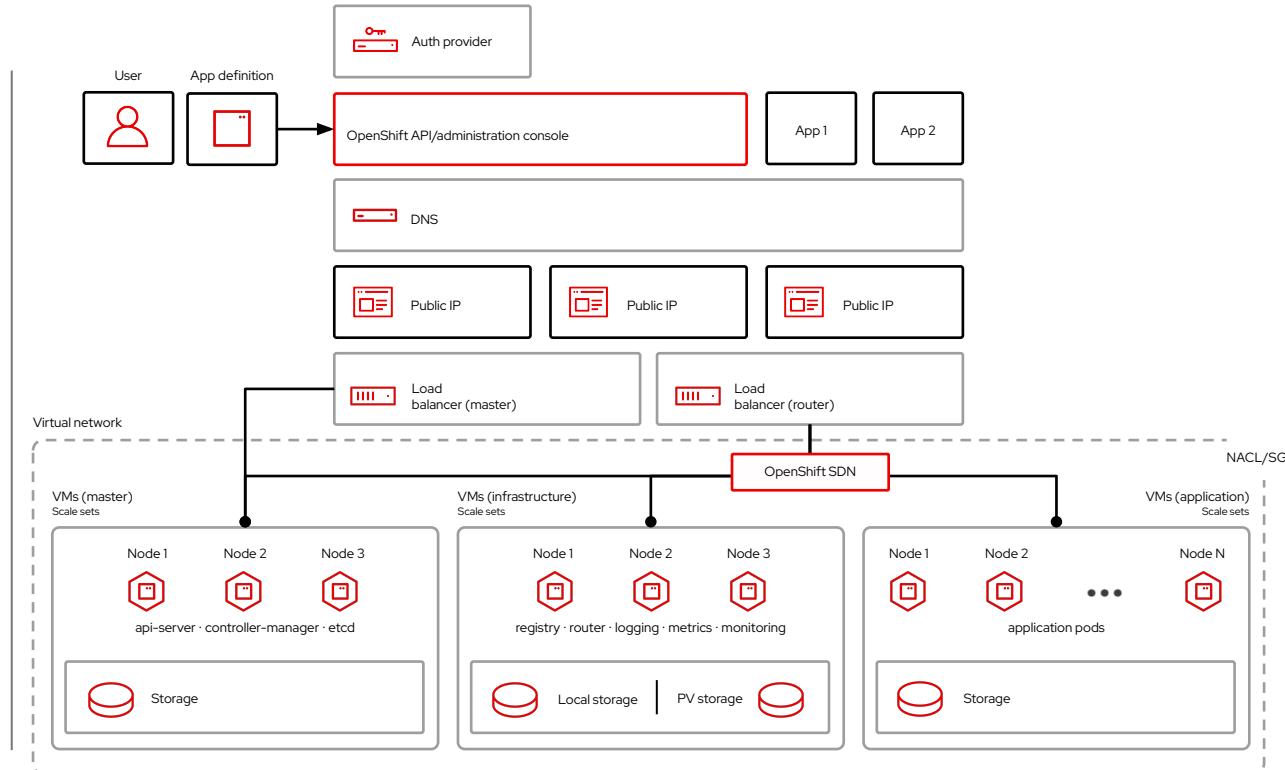
Minimum configuration of ROSA Classic



Complexity of running your own Kubernetes Cluster

Responsibilities	
User management	<input type="checkbox"/>
Project and quota management	<input type="checkbox"/>
Application life cycle	<input type="checkbox"/>
Cluster creation	<input type="checkbox"/>
Cluster management	<input type="checkbox"/>
Monitoring and logging	<input type="checkbox"/>
Network configuration	<input type="checkbox"/>
Software and security updates	<input type="checkbox"/>
Platform support	<input checked="" type="checkbox"/>

Customer Cloud provider & Red Hat



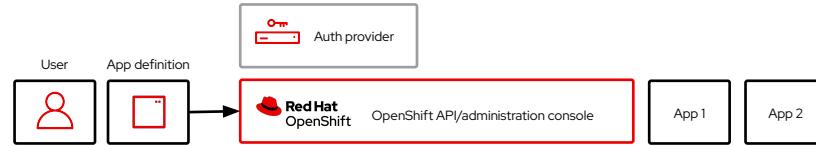
Simplify with fully managed clusters

Red Hat OpenShift cloud services

Responsibilities	
User management	
Project and quota management	
Application life cycle	
Cluster creation	
Cluster management	
Monitoring and logging	
Network configuration	
Software and security updates	
Platform support	

 Customer

 Cloud provider and Red Hat



Let Red Hat & AWS...

Manage all your clusters.

Secure your nodes.

Monitor and operate your VMs.

Manage environment patches.

You...

Compliance

Red Hat OpenShift Service on AWS follows common industry best practices for security and controls. The certifications are outlined in the following table.

Table 1. Security and control certifications for Red Hat OpenShift Service on AWS

Compliance	Red Hat OpenShift Service on AWS (ROSA)	Red Hat OpenShift Service on AWS (ROSA) with hosted control planes (HCP)
HIPAA Qualified	Yes	No
ISO 27001	Yes	Yes
ISO 27017	Yes	Yes
ISO 27018	Yes	Yes
PCI DSS	Yes	Yes
SOC 1 Type 2	Yes	Yes
SOC 2 Type 2	Yes	Yes
SOC 3	Yes	Yes
FedRAMP High ^[1]	Yes (GovCloud requisite)	No

https://docs.openshift.com/roса/roса_architecture/roса_policy_service_definition/roса-policy-process-security.html

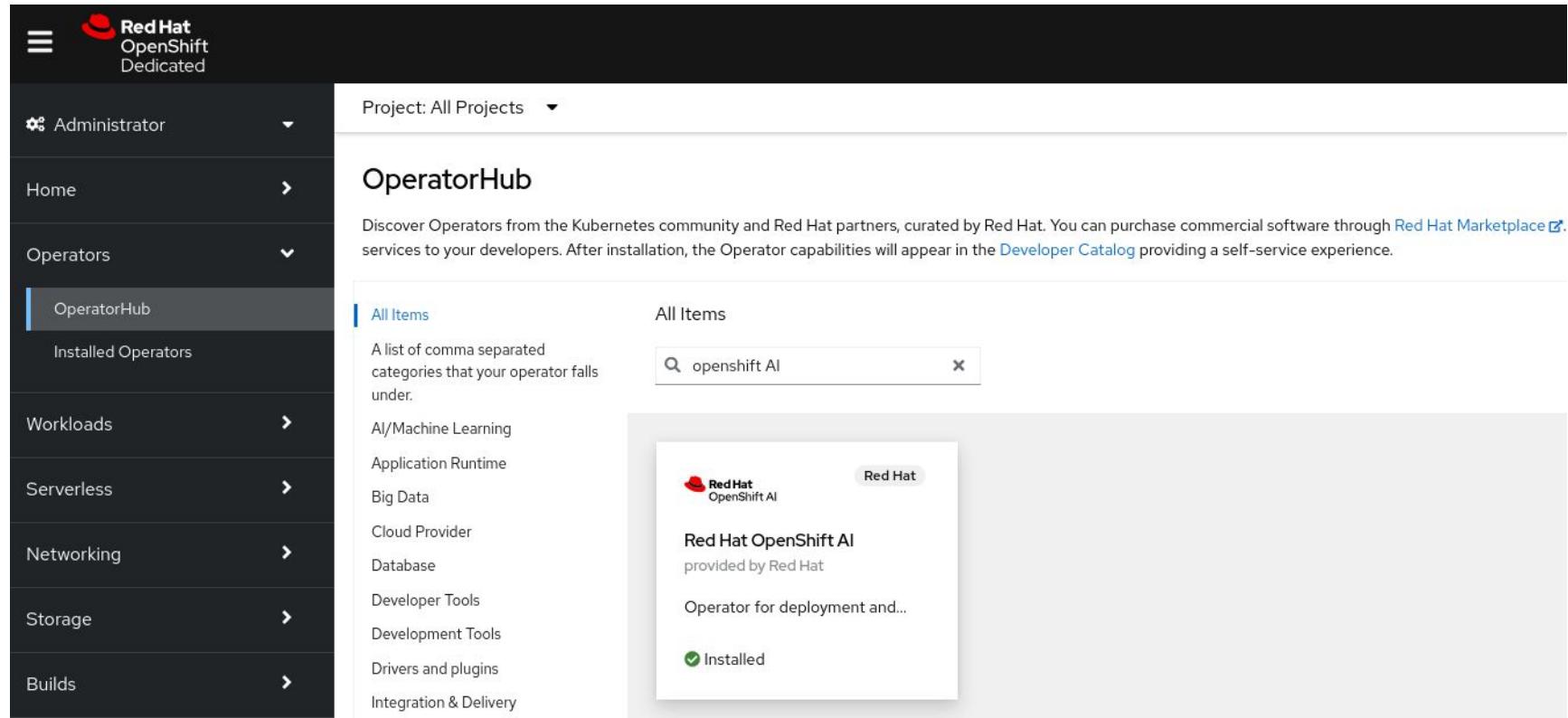


Where do we start?

45



Install RHOAI



The screenshot shows the Red Hat OpenShift Dedicated interface. The left sidebar has a navigation menu with 'Administrator' at the top, followed by 'Home', 'Operators' (which is expanded to show 'OperatorHub' and 'Installed Operators'), 'Workloads', 'Serverless', 'Networking', 'Storage', and 'Builds'. The 'OperatorHub' section is currently selected. The main content area has a heading 'OperatorHub' and a sub-section 'All Items'. A search bar contains the text 'openshift AI'. A search result card for 'Red Hat OpenShift AI' is displayed, showing the Red Hat logo, the name 'Red Hat OpenShift AI' with 'provided by Red Hat', a description 'Operator for deployment and...', and a status indicator 'Installed' with a green checkmark.

Project: All Projects

OperatorHub

Discover Operators from the Kubernetes community and Red Hat partners, curated by Red Hat. You can purchase commercial software through [Red Hat Marketplace](#). services to your developers. After installation, the Operator capabilities will appear in the [Developer Catalog](#) providing a self-service experience.

All Items

All Items

openshift AI

Red Hat OpenShift AI

provided by Red Hat

Operator for deployment and...

Installed

Everything in RHOAI has an OpenShift representation

Models and model servers Deploy model Single-model serving enabled

Model name	Serving runtime	Inference endpoint	Status
My Model	OpenVINO Model Server

Framework: onnx-1

Model server replicas: 2

Model server size: Small

1 CPUs, 4Gi Memory requested
2 CPUs, 8Gi Memory limit

Accelerator: None

Number of accelerators: 0



Project: test

InferenceServices > InferenceService details

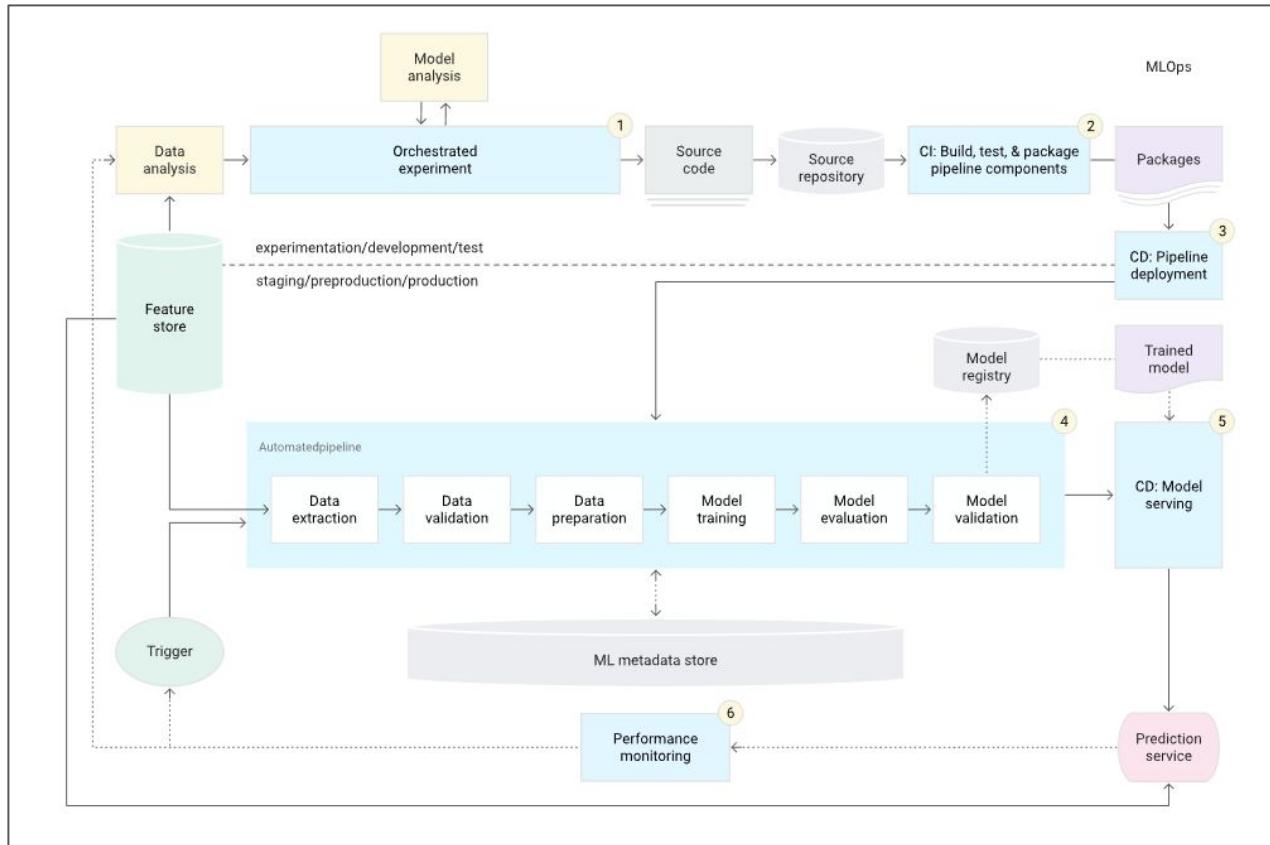
IS my-model

Details YAML

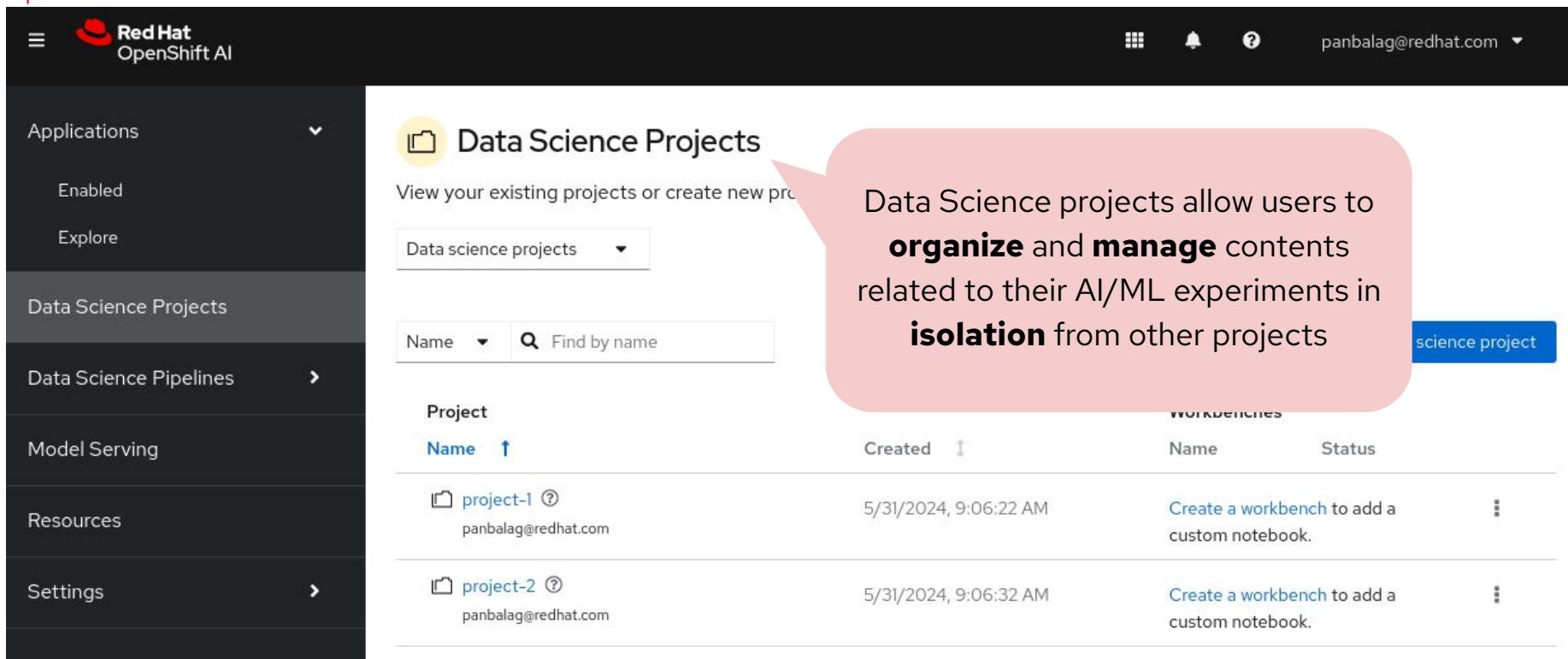
```
1 apiVersion: serving.kserve.io/v1beta1
2 kind: InferenceService
3 > metadata: ...
6 spec:
7   predictor:
8     maxReplicas: 2
9     minReplicas: 2
10    model:
11      modelFormat:
12        name: onnx
13        version: '1'
14      name: ''
15      resources: {}
16      runtime: my-model
17      storage:
18        key: aws-connection-abc
19        path: mymodel/v01
20 > status: ...
61
```

MLOps Automation

Mature MLOps Flow



Data Science Projects



Red Hat OpenShift AI

Applications

- Enabled
- Explore

Data Science Projects

Data Science Pipelines

Model Serving

Resources

Settings

Data Science Projects

View your existing projects or create new projects.

Data science projects

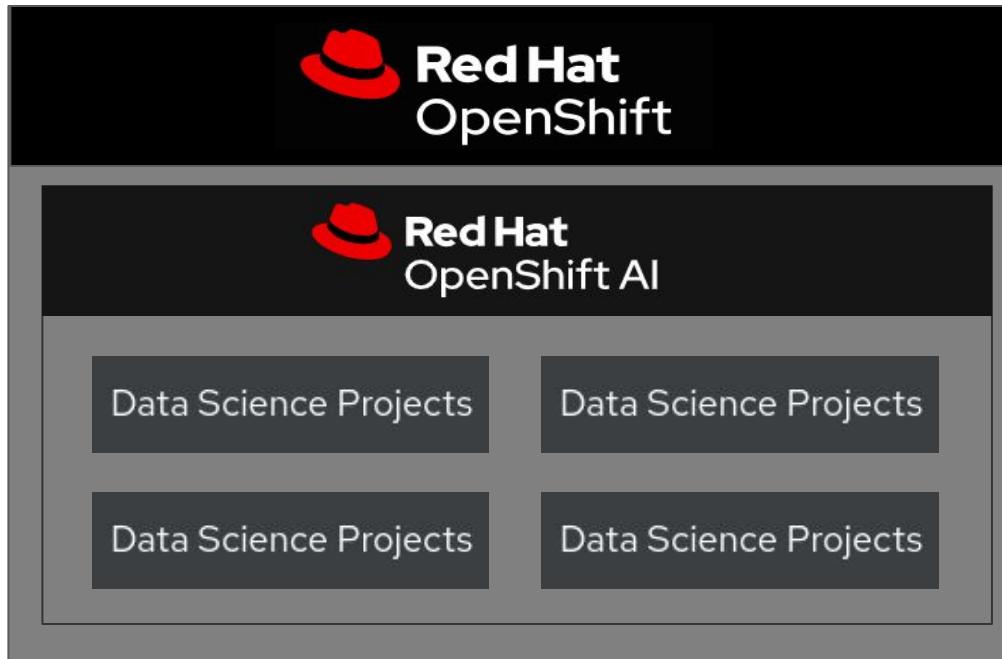
Name Find by name

Project	Created	Workbenches	
Name		Name	Status
project-1 ⓘ	5/31/2024, 9:06:22 AM	Create a workbench	⋮
panbalag@redhat.com		to add a custom notebook.	
project-2 ⓘ	5/31/2024, 9:06:32 AM	Create a workbench	⋮
panbalag@redhat.com		to add a custom notebook.	

Data Science projects allow users to **organize** and **manage** contents related to their AI/ML experiments in **isolation** from other projects

science project

Data Science Projects



- Multiple data science projects.
- Isolation from other projects
- Created by admins or users
- User/Group access privileges

Collaborate within a project

- Users that create a data science project
 - become an admin of that project
 - can give access to a project to any user or group
- Users with access permissions can access all resources in the project, modify them, and create new ones.
- Limiting user level access to data science projects needs to be handled at an OpenShift level at the moment

Collaborate between projects

- Due to isolation of data science projects, resources need to be explicitly exposed in order to be shared between projects.
- A good way to do this is to have an external resource which the projects have access to.
 - Examples:
 - A git repository with shared code
 - An object storage with shared artifacts
 - A structured database with shared data

Data Science Projects



Data Science Projects

Workbenches

Cluster Storage

Data connections

Pipelines

Models and model servers

Data Science Projects

Red Hat OpenShift AI

Applications

Enabled

Explore

Data Science Projects

Data Science Pipelines

Model Serving

Resources

Settings

Data Science Projects

View your existing projects or create new projects.

Data science projects

Name Find by name

Launch Jupyter

Create data science project

Project

Name **↑**

Resource names and types are used to find your resources in OpenShift.

Resource name project-1 **Resource type** Project

Workbenches

Name **Status**

project-1 panbalag@redhat.com

project-2 panbalag@redhat.com

panbalag@redhat.com

Data science projects are 'Projects' in OpenShift identified by the label under 'Resource name'

Data Science Projects

Red Hat OpenShift AI

Applications

Enabled

Explore

Data Science Projects

Data Science Pipelines

Model Serving

Resources

Data Science Projects

View your existing projects or create new projects.

Data science projects

Name Find by name

Project

Name ↑

project-1 ⓘ panbalag@redhat.com

Filter

Name Display name Status

PR project-1 project-1 Active

PR project-2 project-2 Active

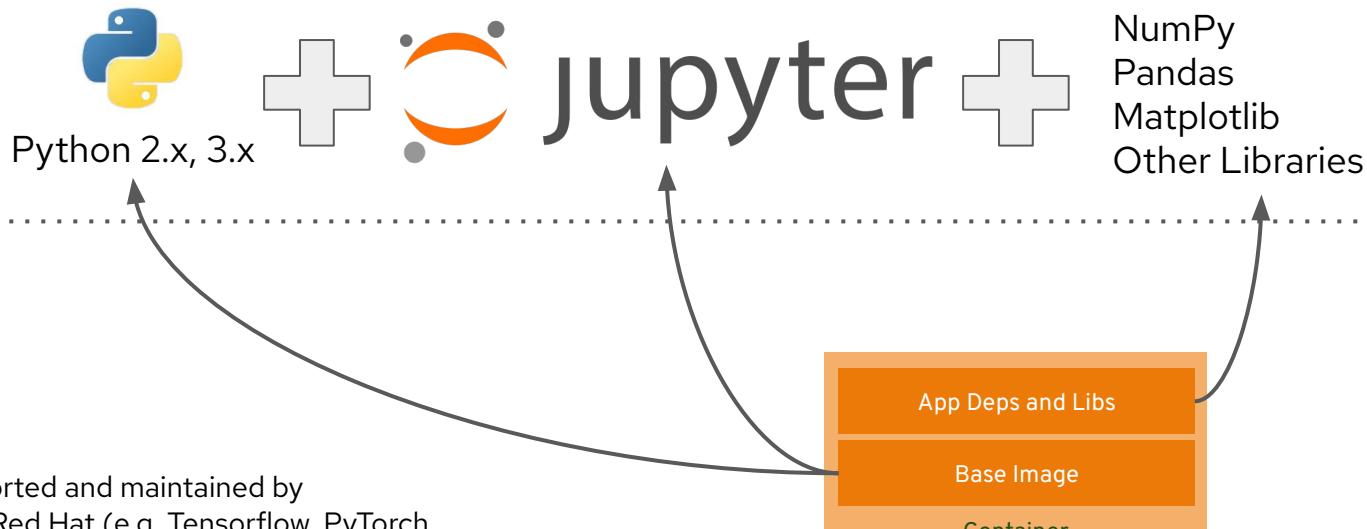
Red Hat OpenShift Dedicated

Projects

Customizing Workbenches

Base Notebook Images

Reproducible and shareable environments for building, training and serving



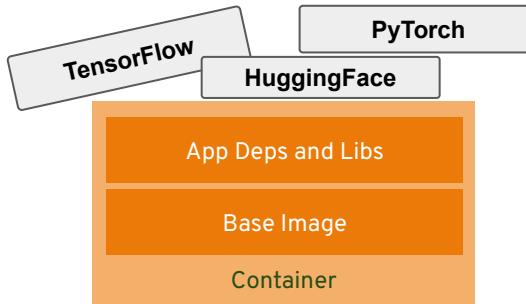
Supported and maintained by

- Red Hat (e.g. Tensorflow, PyTorch, CUDA)
- partner (Anaconda, Intel)
- you (custom notebooks)

Customizing Workbenches

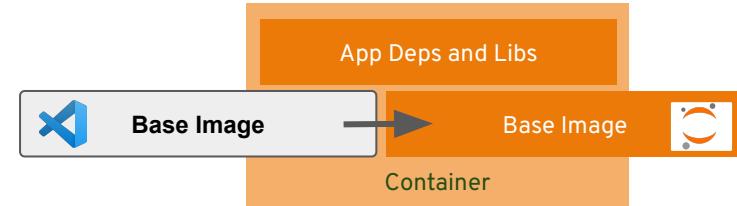
Customizing the workbench

Adding packages on top of a good image



Just remember that they are removed
when restarting the workbench*

Creating your own custom image with all
dependencies you need

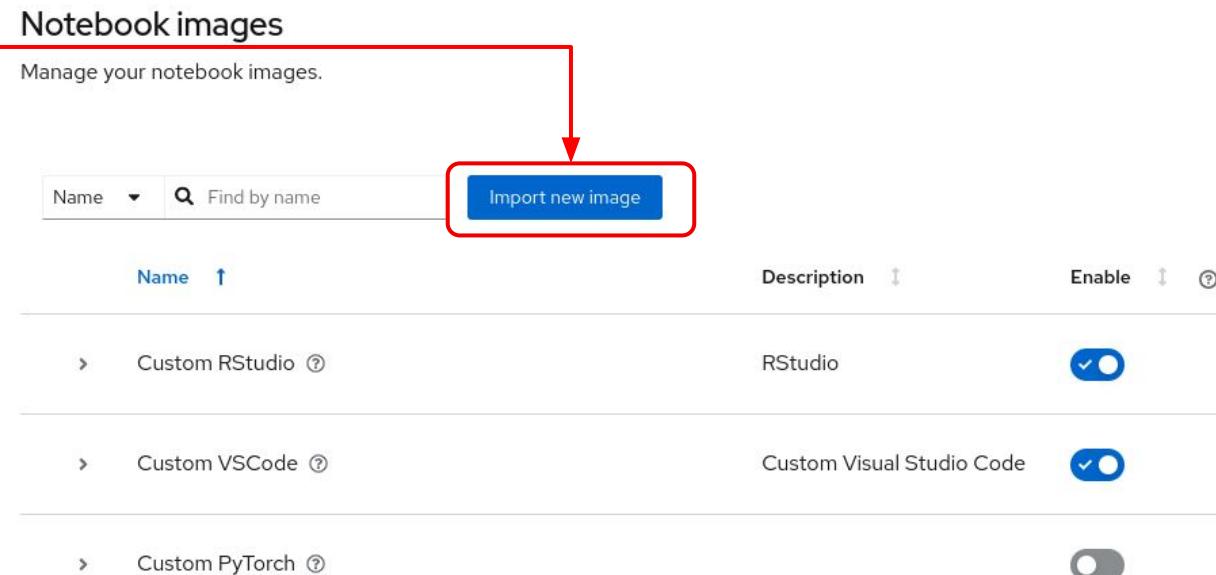
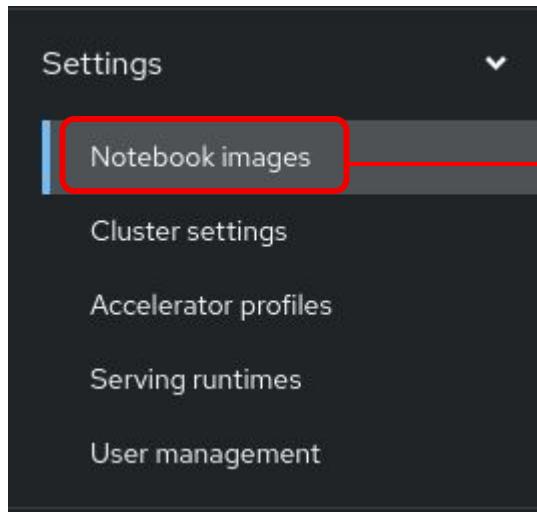


You can now version and maintain it
according to your preferences

* This is on purpose so that you can un-mess-up your environment easily if you get into dependency issues.

Custom Notebook Image

Import new image



The image shows the 'Notebook images' management interface with the following details:

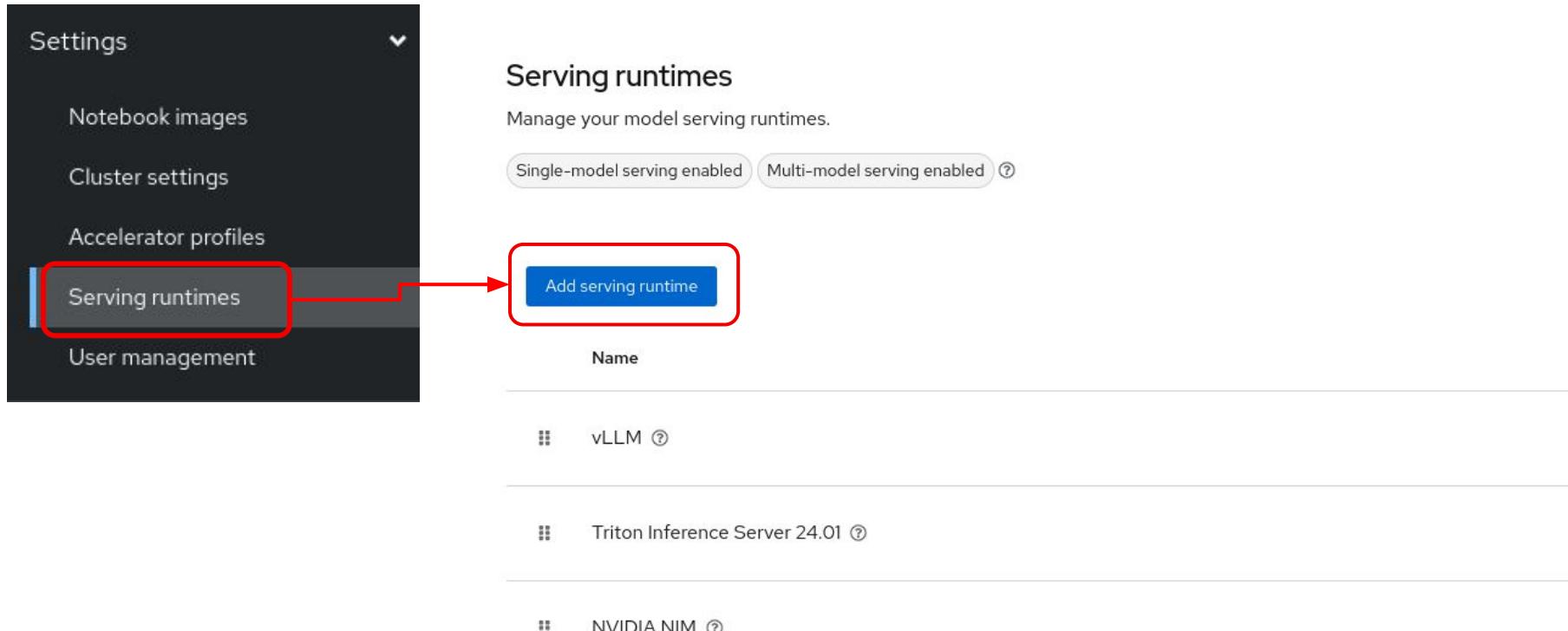
Notebook images
Manage your notebook images.

Name	Description	Enable
Custom RStudio	RStudio	<input checked="" type="checkbox"/>
Custom VSCode	Custom Visual Studio Code	<input checked="" type="checkbox"/>
Custom PyTorch		<input type="checkbox"/>

Import new image

Custom Serving Runtime

Add serving runtime



The image shows a user interface for managing model serving runtimes. On the left, a sidebar menu under 'Settings' includes 'Notebook images', 'Cluster settings', 'Accelerator profiles', 'Serving runtimes' (which is highlighted with a red box and has a red arrow pointing to the 'Add serving runtime' button), and 'User management'. The main content area is titled 'Serving runtimes' and describes managing model serving runtimes. It shows two runtimes listed: 'vLLM' and 'Triton Inference Server 24.01'. A blue 'Add serving runtime' button is prominently displayed. The 'vLLM' entry includes a 'Name' field, a three-dot icon, and a question mark icon. The 'Triton Inference Server 24.01' entry includes a three-dot icon and a question mark icon. The 'NVIDIA NIM' entry is partially visible at the bottom.

Settings

Notebook images

Cluster settings

Accelerator profiles

Serving runtimes

User management

Serving runtimes

Manage your model serving runtimes.

Single-model serving enabled

Multi-model serving enabled

Add serving runtime

Name

vLLM

Triton Inference Server 24.01

NVIDIA NIM

Customize RHOAI Cluster

Enable or disable components

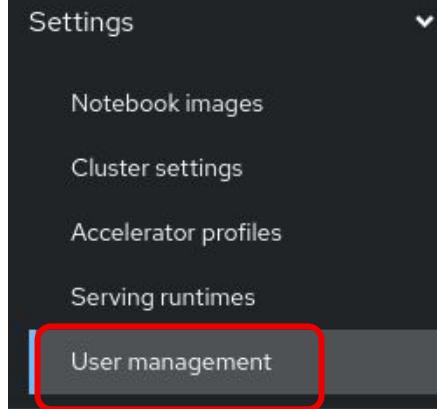
You can enable or disable RHOAI components inside of your DataScienceCluster yaml.

These are the components you can enable/disable:

- CodeFlare (for distributed training)
- Dashboard
- Data Science Pipelines
- Kserve (the component for single-model serving)
- Modelmesh serving (the component for multi-model serving)
- Ray (for distributed training)
- TrustyAI
- Workbenches

```
101  spec:
102    components:
103      codeflare:
104        devFlags: {}
105        managementState: Removed
106      dashboard:
107        devFlags: {}
108        managementState: Managed
109      datasciencelines:
110        devFlags: {}
111        managementState: Managed
112      kserve:
113        devFlags: {}
114        managementState: Managed
115        serving:
116          ingressGateway:
117            certificate:
118              secretName: knative-serving-cert
119              type: SelfSigned
120            managementState: Managed
121            name: knative-serving
122          modelmeshserving:
123            devFlags: {}
124            managementState: Managed
125          ray:
126            devFlags: {}
127            managementState: Removed
128          trustyai:
129            devFlags: {}
130            managementState: Removed
131          workbenches:
132            devFlags: {}
133            managementState: Managed
```

User Management



- Settings
- Notebook images
- Cluster settings
- Accelerator profiles
- Serving runtimes
- User management

User management

Define OpenShift group membership for Data Science administrators and users.

Data Science administrator groups

Select the OpenShift groups that contain all Data Science administrators.

cluster-admins ✕ dedicated-admins ✕ rhods-admins ✕

[View, edit, or create groups in OpenShift under User Management](#)

ⓘ All cluster admins are automatically assigned as Data Science administrators.

Data Science user groups

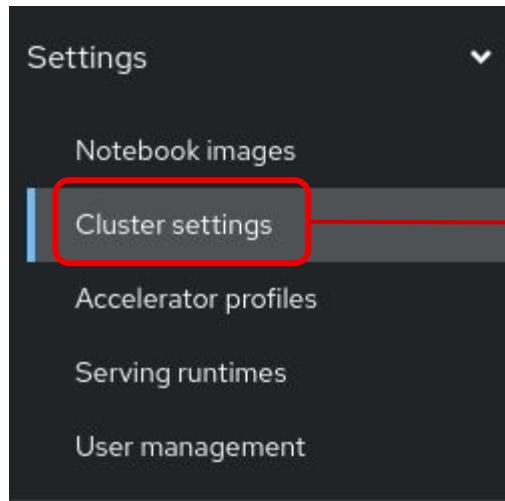
Select the OpenShift groups that contain all Data Science users.

system:authenticated ✕

[View, edit, or create groups in OpenShift under User Management](#)

[Save changes](#)

Cluster Settings



1. Model serving platforms
2. PVC size
3. Stop idle notebooks
4. Usage data collection
5. Notebook pod tolerations

Accelerator Profile

- Applications ▼
 - Enabled
 - Explore
- Data Science Projects
- Data Science Pipelines ▼
 - Pipelines
 - Runs
- Model Serving
- Resources
- Settings ▼
 - Notebook images
 - Cluster settings
 - Accelerator profiles
 - Serving runtimes
 - User management

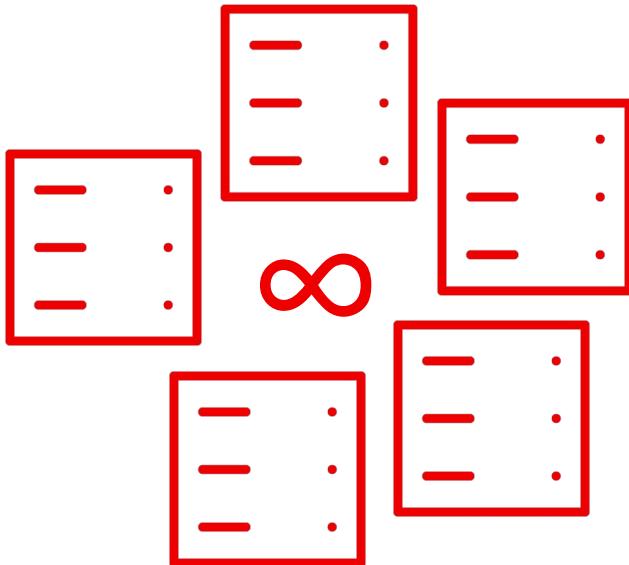
Accelerator profiles

Manage accelerator profile settings for users in your organization

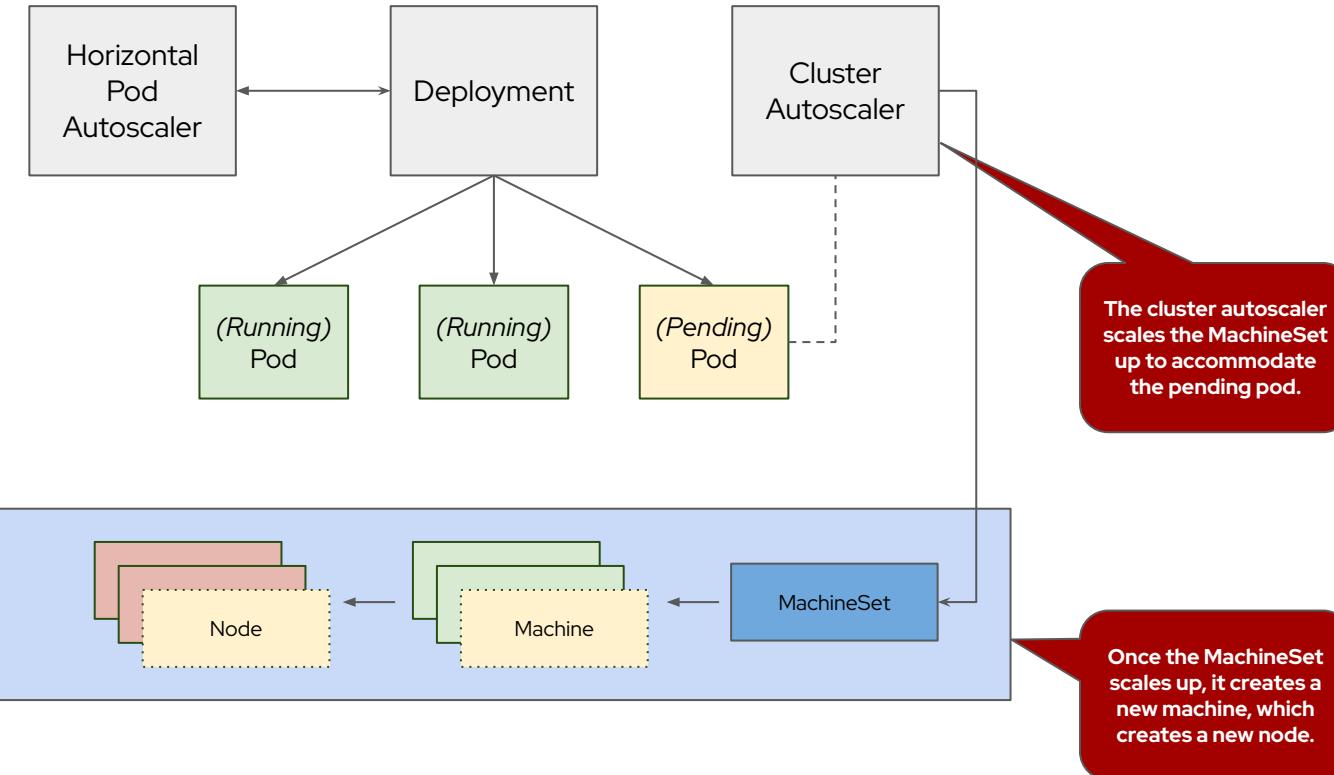
Name	Identifier	Enable
fractional small	nvidia.com/gpu-frac	<input type="checkbox"/>
1/7th of a real GPU		
Habana HPU - 1st Gen Gaudi	habana.ai/gaudi	<input type="checkbox"/>
This Accelerator Profile is for 1st Gen Gaudi Devices		
Large GPU Card	nvidia.com/gpu	<input type="checkbox"/>
NVIDIA GPU - use sparingly	nvidia.com/gpu	<input checked="" type="checkbox"/>
We have very few GPUs in this cluster. Although you can use them fo...		
tinyGPU	nvidia.com/gpu	<input type="checkbox"/>

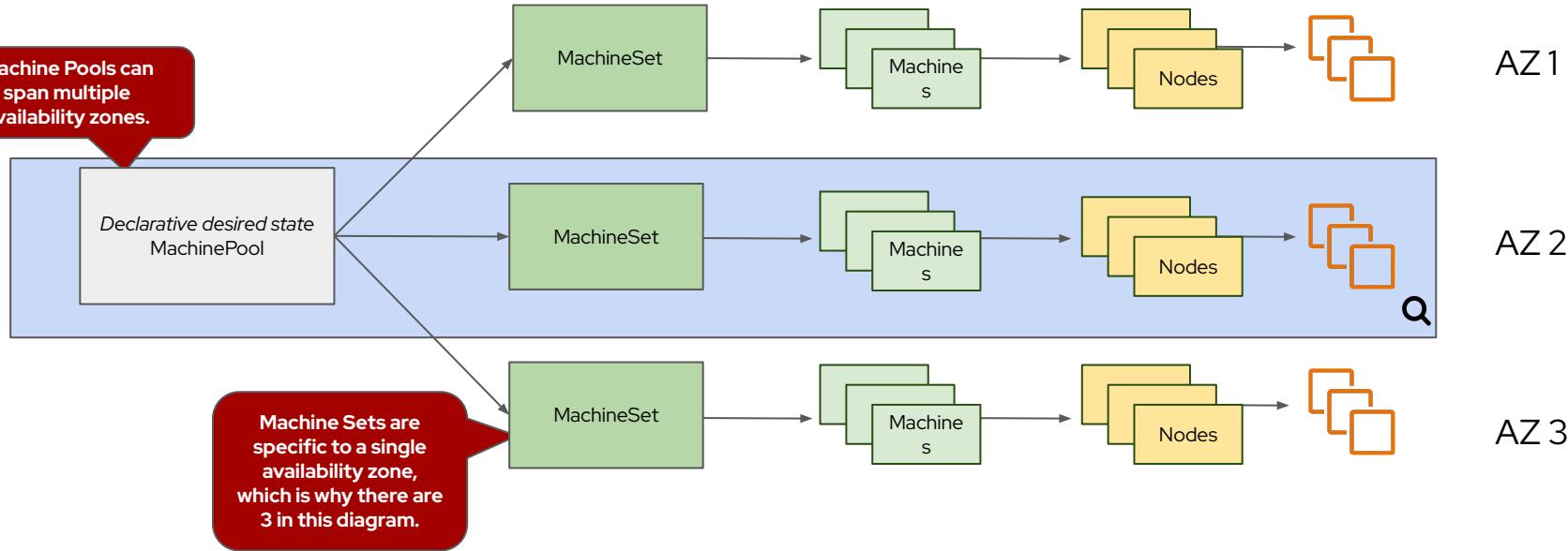
Cluster Autoscaling

Automatically responding to cluster demand.



- ▶ MachinePools can be scaled to meet applications demands.
- ▶ Cluster AutoScaler will provision additional worker nodes when pods can not be scheduled due to resource constraints.
- ▶ Cluster AutoScaler will not scale beyond predefined limits.





Monitoring

☰ Red Hat OpenShift AI Admin

Applications >

Data Science Projects

Data Science Pipelines >

Distributed Workload Metrics

Model Serving

Resources

Settings >

Distributed Workload Metrics

Monitor the metrics of your active resources.

Project Insurance Claims

Project metrics Distributed workload status

Refresh interval 5 minutes

Resource alerts Not configured Configure resource alert rules

Requested resources

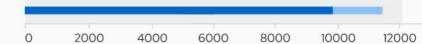
CPU Cores



0 200 400 600 800 1000 1200

- Requested by Insurance Claims: 968
- Requested by all projects: 1000
- Total shared quota: 1200

Memory GiB



0 2000 4000 6000 8000 10000 12000

- Requested by Insurance Claims: 9950
- Requested by all projects: 11060
- Total shared quota: 12000

Accelerator GPUs



0 20 40 60 80 100 120

- Requested by Insurance Claims: 90
- Requested by all projects: 110
- Total shared quota: 120

Accelerator memory GiB



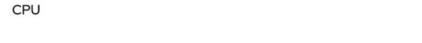
0 2000 4000 6000 8000 10000 12000

- Requested by Insurance Claims: 7461
- Requested by all projects: 9999
- Total shared quota: 12000

Resource usage over time

Workloads None selected Time range 6 hours

CPU



1000

Memory



12000

Automatic vs Manual

Change update approval strategy

What strategy is used for approving updates?

Automatic (default)

New updates will be installed as soon as they become available.

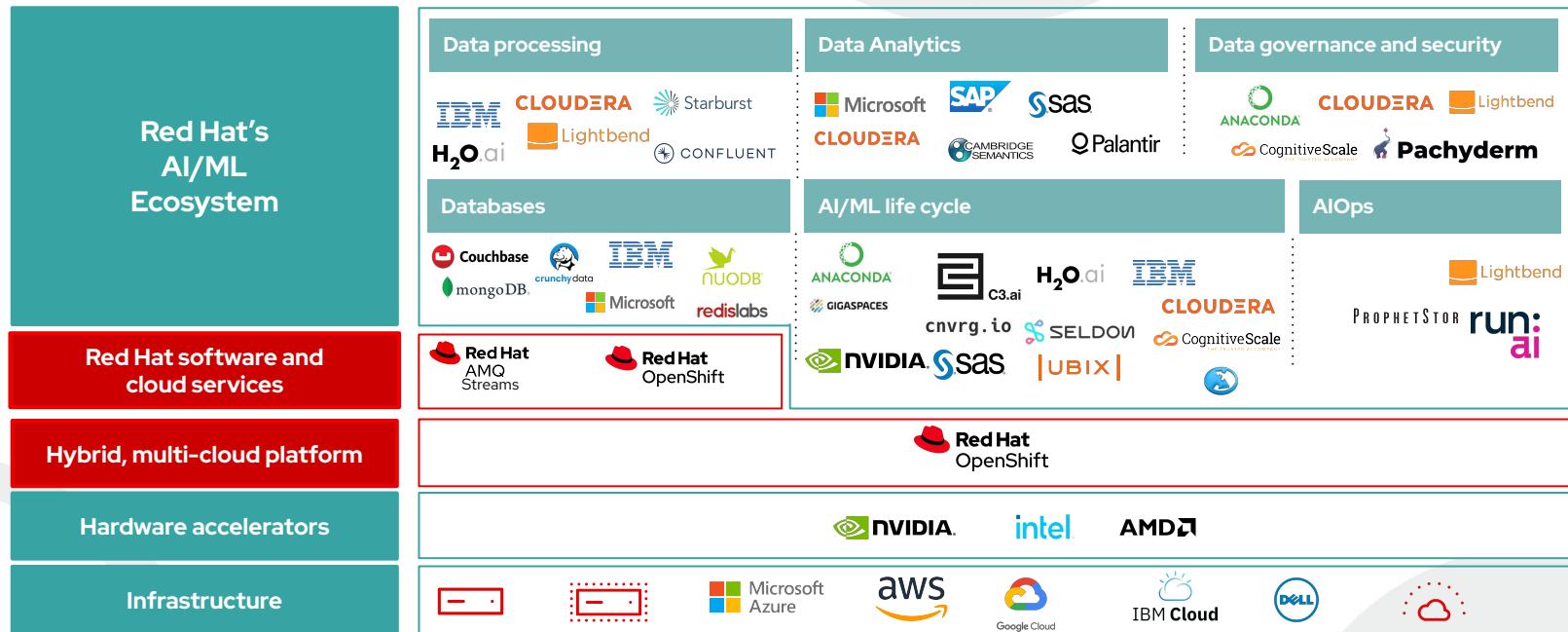
Manual

New updates need to be manually approved before installation begins.

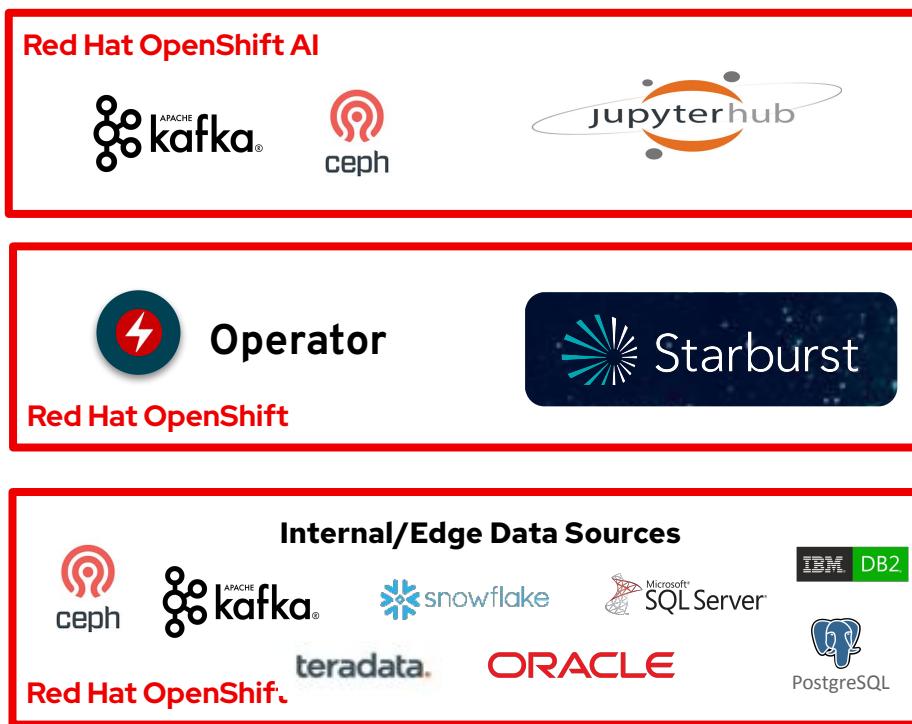
Cancel

Save

Strategic partnerships + Red Hat AI/ML offerings



Data Acquisition and Preparation



Operators on ROSA for AWS Services



Welcome to OperatorHub.io

OperatorHub.io is a new home for the Kubernetes community to share Operators. Find an existing Operator or list your own today.

CATEGORIES

- AI/Machine Learning
- Application Runtime
- Big Data
- Cloud Provider
- Database
- Developer Tools
- Drivers and plugins
- Integration & Delivery
- Logging & Tracing
- Modernization & Migration
- Monitoring
- Networking
- OpenShift Optional
- Security
- Storage
- Streaming & Messaging

VIEW ■■■ SORT A-Z

Provider	Operator Name	Description
Aerospike	Aerospike Operator	Provides a Kubernetes operator for managing Aerospike clusters.
alauda	alauda Operator	Provides a Kubernetes operator for managing alauda clusters.
Alibaba Cloud	Alibaba Cloud Operator	Provides a Kubernetes operator for managing Alibaba Cloud resources.
Altinity	Altinity Operator	Provides a Kubernetes operator for managing Altinity clusters.
Basic Install	Basic Install Operator	Provides a Kubernetes operator for basic installations.
NetApp	Astra Trident Operator	Provides a Kubernetes operator for managing Astra Trident installations.
NetApp	AWS Auth Operator	Provides a Kubernetes operator for managing AWS auth-cm ConfigMap management.
Amazon	AWS Controllers for Kubernetes - Amazon ACM	Provides a Kubernetes operator for managing AWS Controllers for Kubernetes - Amazon ACM.
Amazon	AWS Controllers for Kubernetes - Amazon ACM PCA	Provides a Kubernetes operator for managing AWS Controllers for Kubernetes - Amazon ACM PCA.
Amazon	AWS Controllers for Kubernetes - Amazon API Gateway v2	Provides a Kubernetes operator for managing AWS Controllers for Kubernetes - Amazon API Gateway v2.
Amazon	AWS Controllers for Kubernetes - Amazon Application Auto Scaling	Provides a Kubernetes operator for managing AWS Controllers for Kubernetes - Amazon Application Auto Scaling.
Amazon	AWS Controllers for Kubernetes - Amazon CloudFront	Provides a Kubernetes operator for managing AWS Controllers for Kubernetes - Amazon CloudFront.
Amazon	AWS Controllers for Kubernetes - Amazon CloudTrail	Provides a Kubernetes operator for managing AWS Controllers for Kubernetes - Amazon CloudTrail.
Amazon	AWS Controllers for Kubernetes - Amazon CloudWatch	Provides a Kubernetes operator for managing AWS Controllers for Kubernetes - Amazon CloudWatch.
Amazon	AWS Controllers for Kubernetes - Amazon CloudWatch Logs	Provides a Kubernetes operator for managing AWS Controllers for Kubernetes - Amazon CloudWatch Logs.
Amazon	AWS Controllers for Kubernetes - Amazon DocumentDB	Provides a Kubernetes operator for managing AWS Controllers for Kubernetes - Amazon DocumentDB.
Amazon	AWS Controllers for Kubernetes - Amazon DynamoDB	Provides a Kubernetes operator for managing AWS Controllers for Kubernetes - Amazon DynamoDB.
Amazon	AWS Controllers for Kubernetes - Amazon EC2	Provides a Kubernetes operator for managing AWS Controllers for Kubernetes - Amazon EC2.
Amazon	AWS Controllers for Kubernetes - Amazon ECR	Provides a Kubernetes operator for managing AWS Controllers for Kubernetes - Amazon ECR.
Amazon	AWS Controllers for Kubernetes - Amazon ECS	Provides a Kubernetes operator for managing AWS Controllers for Kubernetes - Amazon ECS.

Strategic partnerships within AI/ML ecosystem

AI/ML life cycle



Data processing



Databases



Infrastructure partners



Data governance and security



Data analytics



AI Ops



Hardware acceleration





Training and Certification

Close skills gaps and hone your teams' Red Hat product expertise; flexible learning options to meet your business needs

coming soon

Introduction to Red Hat OpenShift AI (AI262)	Red Hat OpenShift AI Administration (AI263)	Creating Machine Learning Models with Red Hat OpenShift AI (AI264)	Deploying Machine Learning Models with Red Hat OpenShift AI (AI265)	Automation using Data Science Pipelines (AI266)
Learn about the general architecture and main features of RHOAI and create some basic projects.	Get hands-on administration experience with RHOAI including install, upgrades, management and configuration.	Learn basic Machine Learning concepts and create and train Machine Learning models.	Deploy and serve Machine Learning models on RHOAI and troubleshoot deployed models.	Learn about and get hands-on experience with KubeFlow and Elyra pipelines.
Developing and Deploying AI/ML Applications on Red Hat OpenShift AI (AI267): 4 day instructor-led virtual training				
Red Hat Certified Specialist: OpenShift AI (EX267): 3 hour exam				

Conclusion



Thank you!

Yury Titov

ytitov@redhat.com



linkedin.com/company/red-hat



youtube.com/user/RedHatVideos



facebook.com/redhatinc



twitter.com/RedHat