Red Hat
Cloud Services

Harnessing Red Hat AI on ROSA (AWS)

Practical Life Cycle Tips for Intelligent Applications

Yury Titov

Sr. EMEA Black Belt for Managed Cloud Services

Red Hat

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

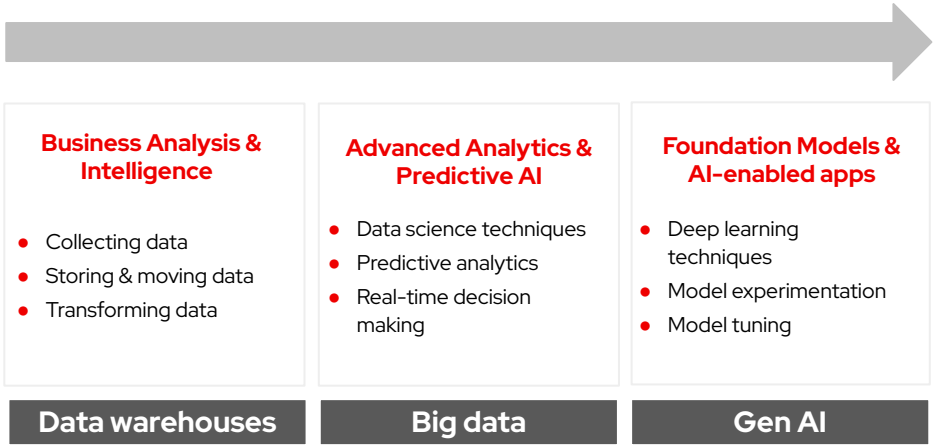
watsonx Code Assistant

IBM Granite Models

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

Operationalize AI with Red Hat OpenShift AI

What is Machine Learning?

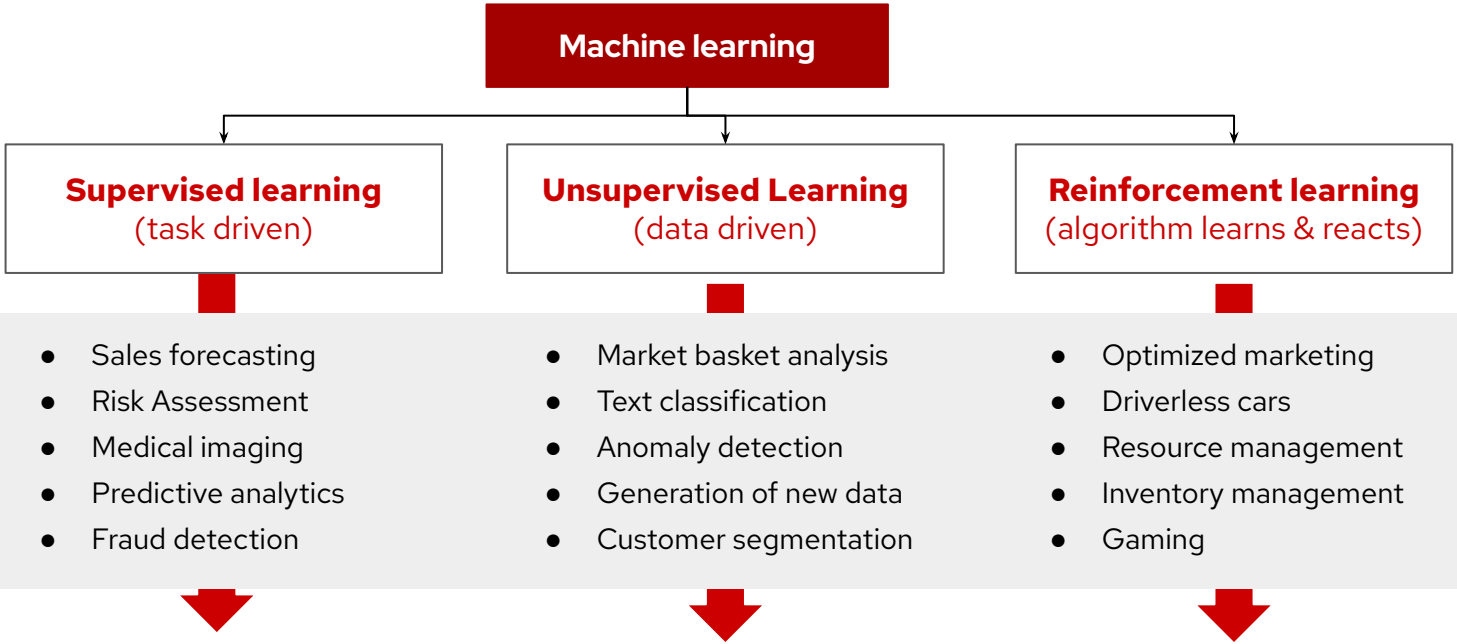
Machine learning can solve business problems

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



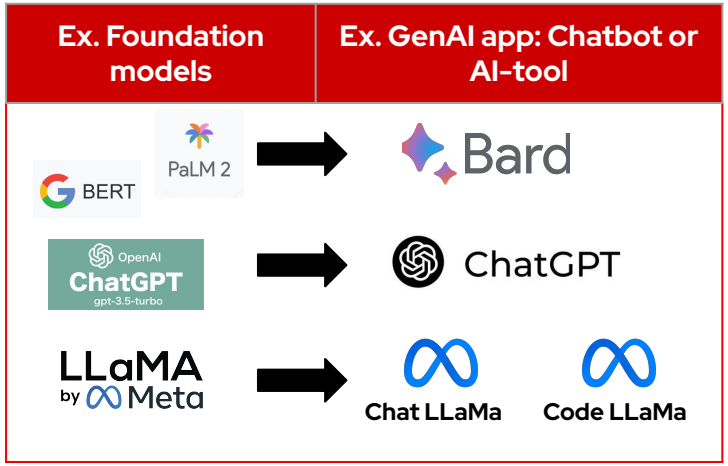
Sources: Types Of Machine Learning Techniques, 4 types of learning in machine learning explained

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

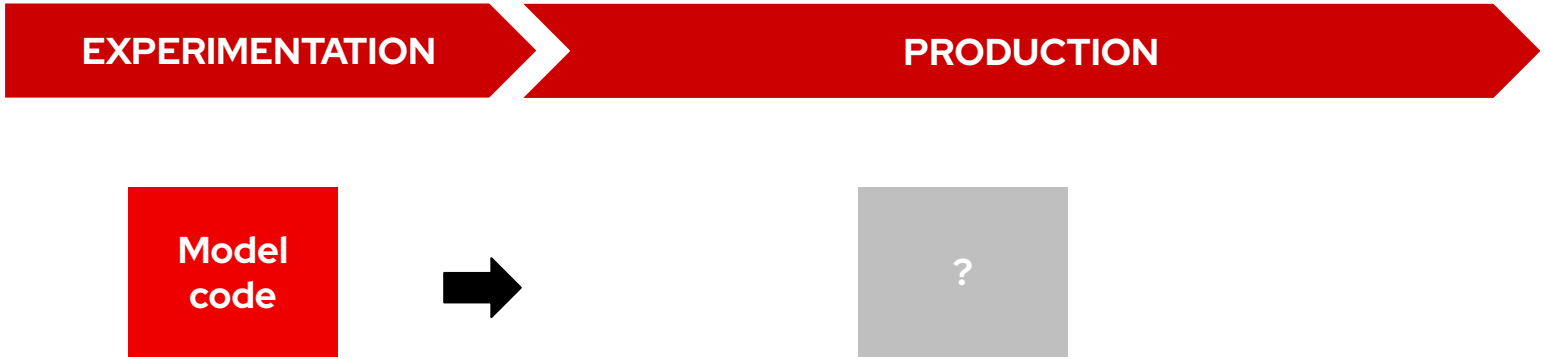
It’s not magic.
It’s math.

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

Red Hat

Real Life View of Technical Teams on AI*

*gathered from real life experience in EMEA ;)

Legacy
Monolith

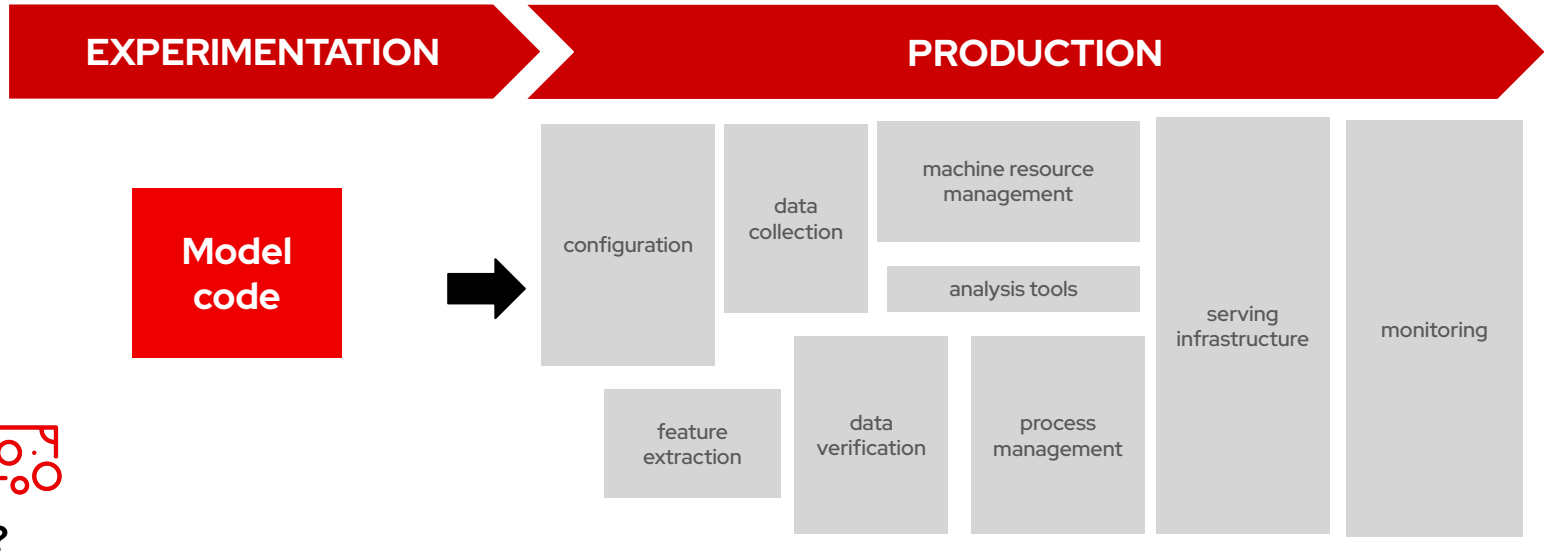
Modern
Microservices

AI

Red Hat

Poorly designed systems lead to failed ML projects

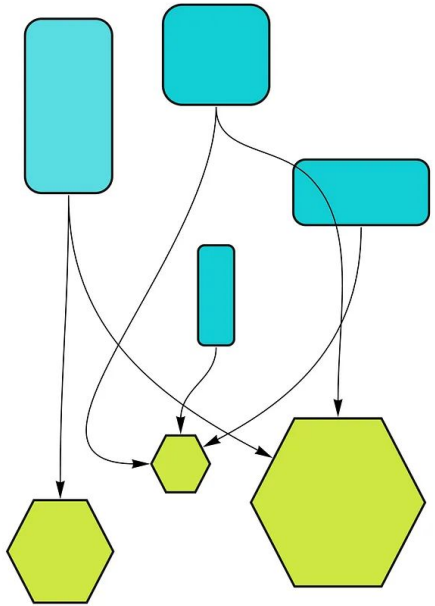
Lack of focus on end-to-end system builds technical debt



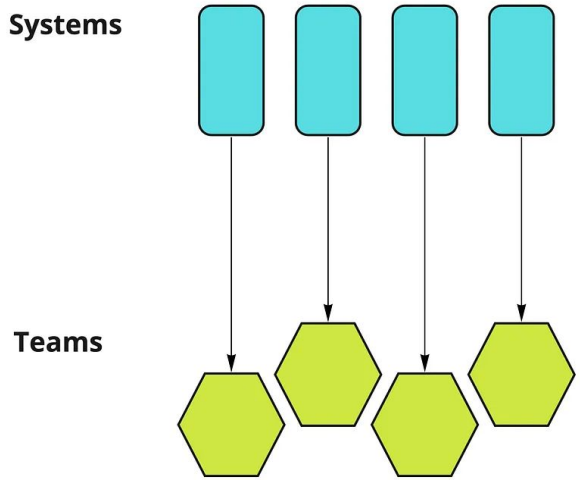
Technical debt is a barrier to production

Expected Outcome

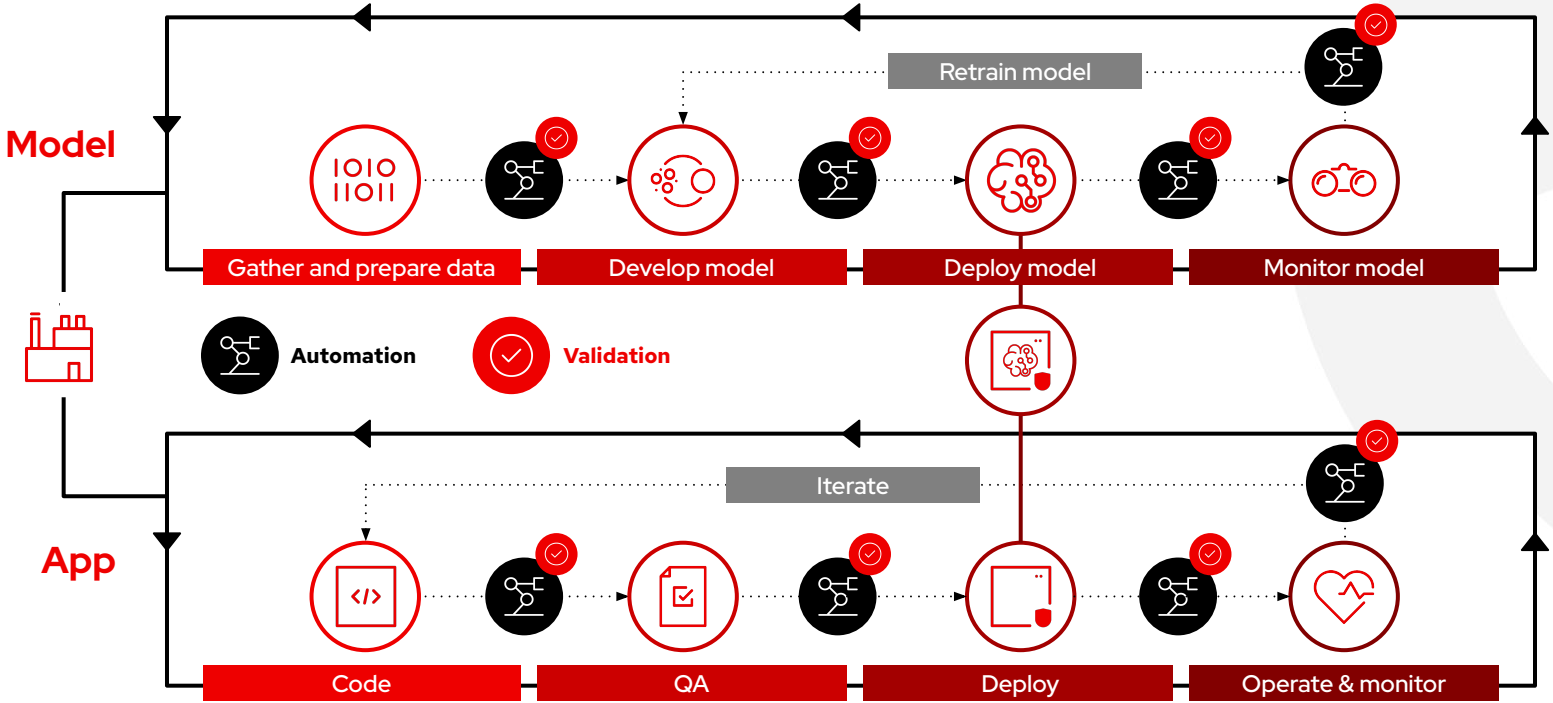
Typical mapping between systems and teams



Mapping between systems and teams in a Team-First approach



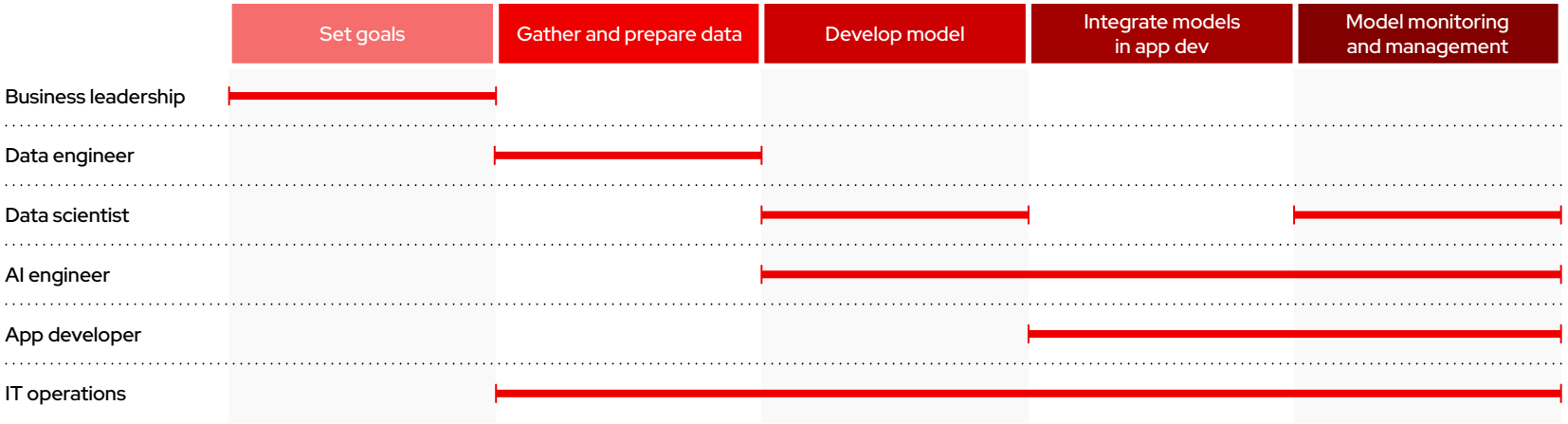
Lifecycle for operationalizing (containerised) models



Red Hat

Operationalizing AI/ML requires collaboration

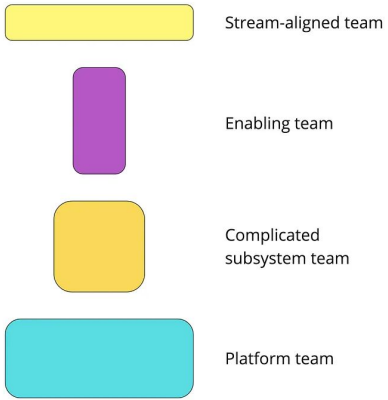
Every member of your team plays a critical role in a complex process



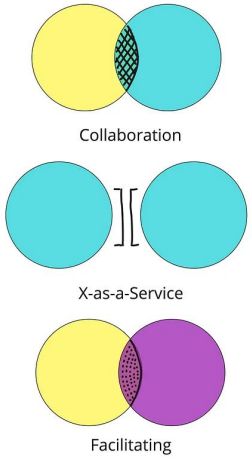
Red Hat

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

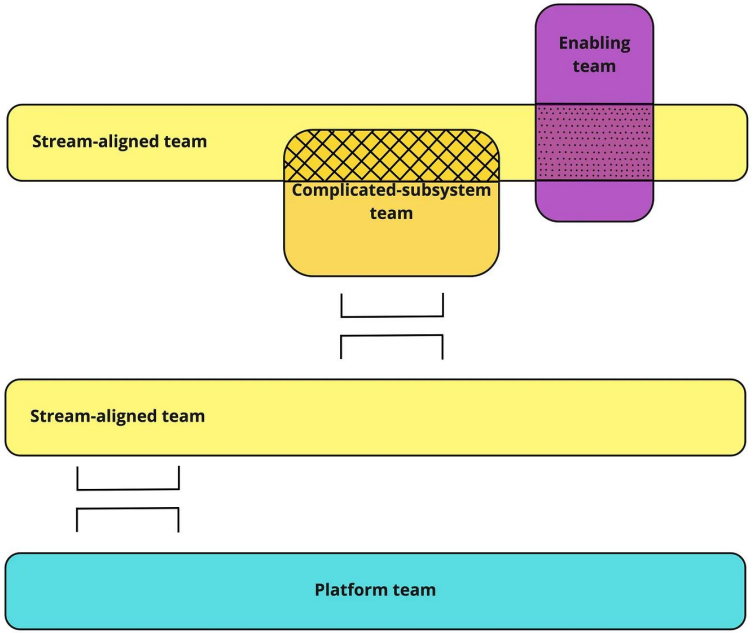
provide internal services like deployment platforms or data services

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

Red Hat’s approached are informed by, and align with, Team Topologies

Red Hat

Typical interactions between different teams



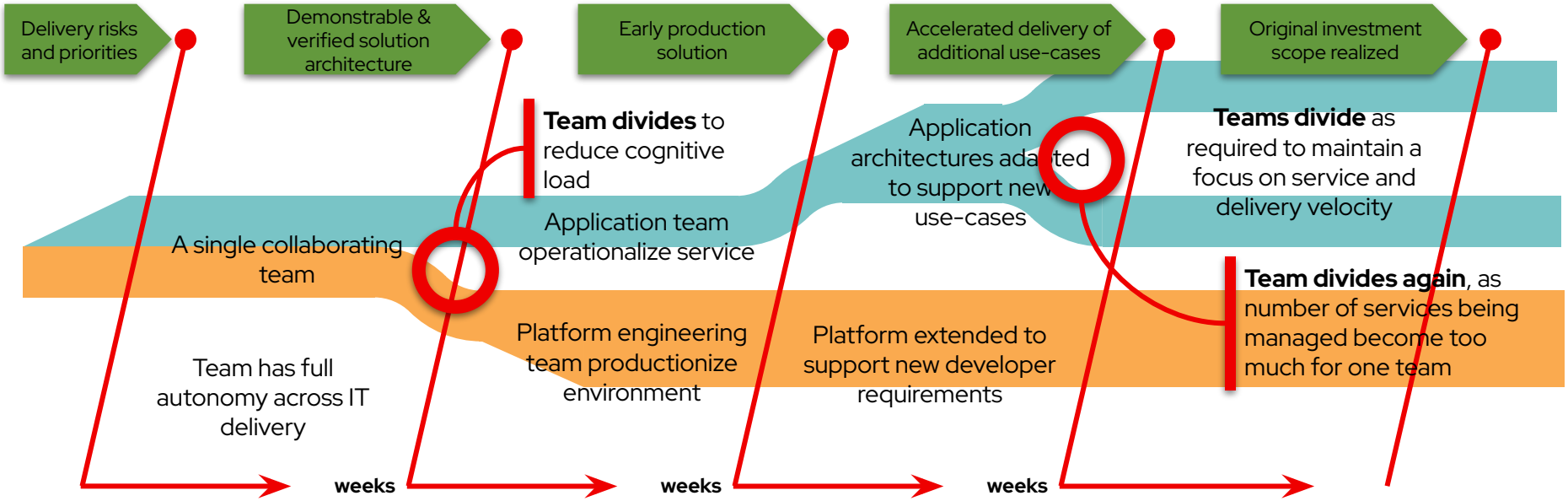
Team Topologies: Organizing Business and Technology Teams for Fast Flow, Pias & Skelton
ISBN: 9781942788812

Red Hat's approached are informed by, and align with, Team Topologies

Red Hat

Red Hat recommends an evolutionary approach to organisational change

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



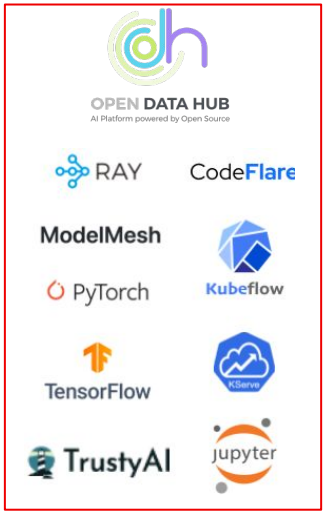
Team Topologies: Organizing Business and Technology Teams for Fast Flow, Pias & Skelton
ISBN: 9781942788812

Red Hat's approached are informed by, and align with, Team Topologies

Red Hat

Simplify AI adoption

Designed to increase AI adoption and enhance trust in AI initiatives



Red Hat OpenShift AI

Flexible

A composable platform for rapid dev and delivery of AI-enabled apps

Expand

A certified AI partner ecosystem for delivering an E2E AI/ML experience

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

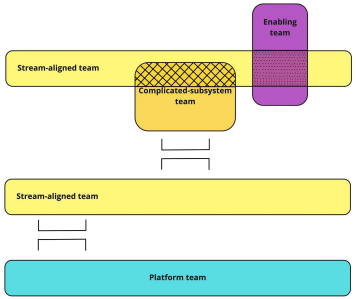
Build an AI platform for E2E AI lifecycle management

Red Hat OpenShift + **Red Hat OpenShift AI** + **Red Hat's AI Partner Ecosystem**

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

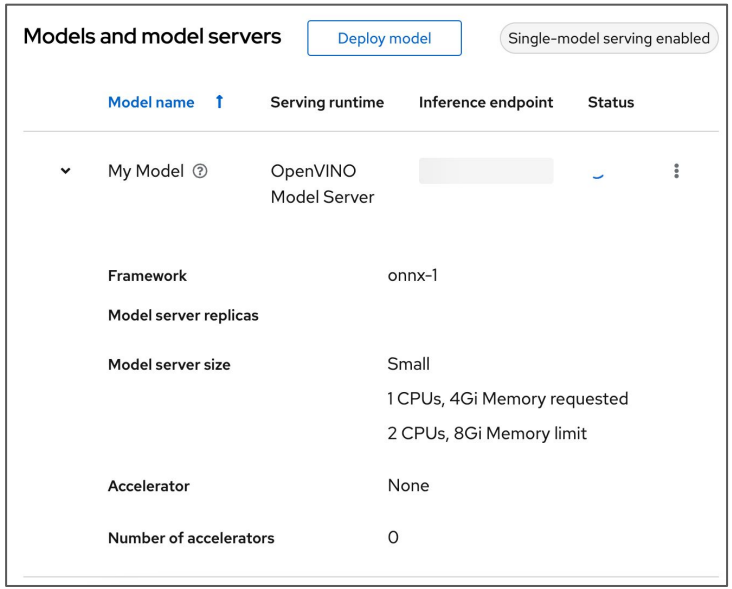
Best-of-breed AI technologies from a certified partner ecosystem to complement or extend Red Hat's AI capabilities

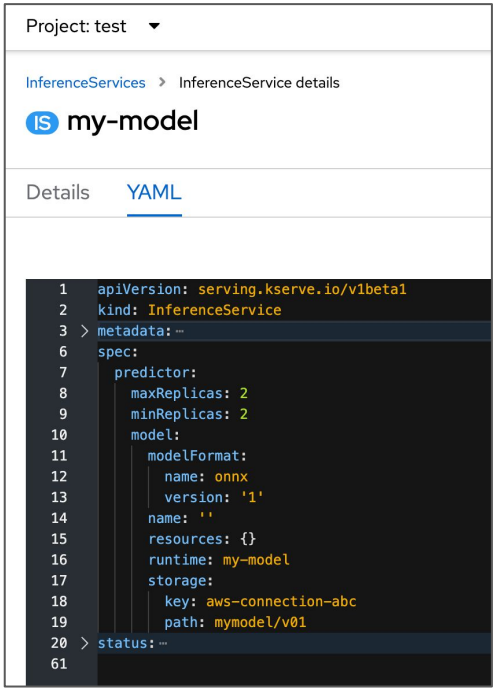


Red Hat

UI to Yaml

Everything in RHOAI has an OpenShift representation

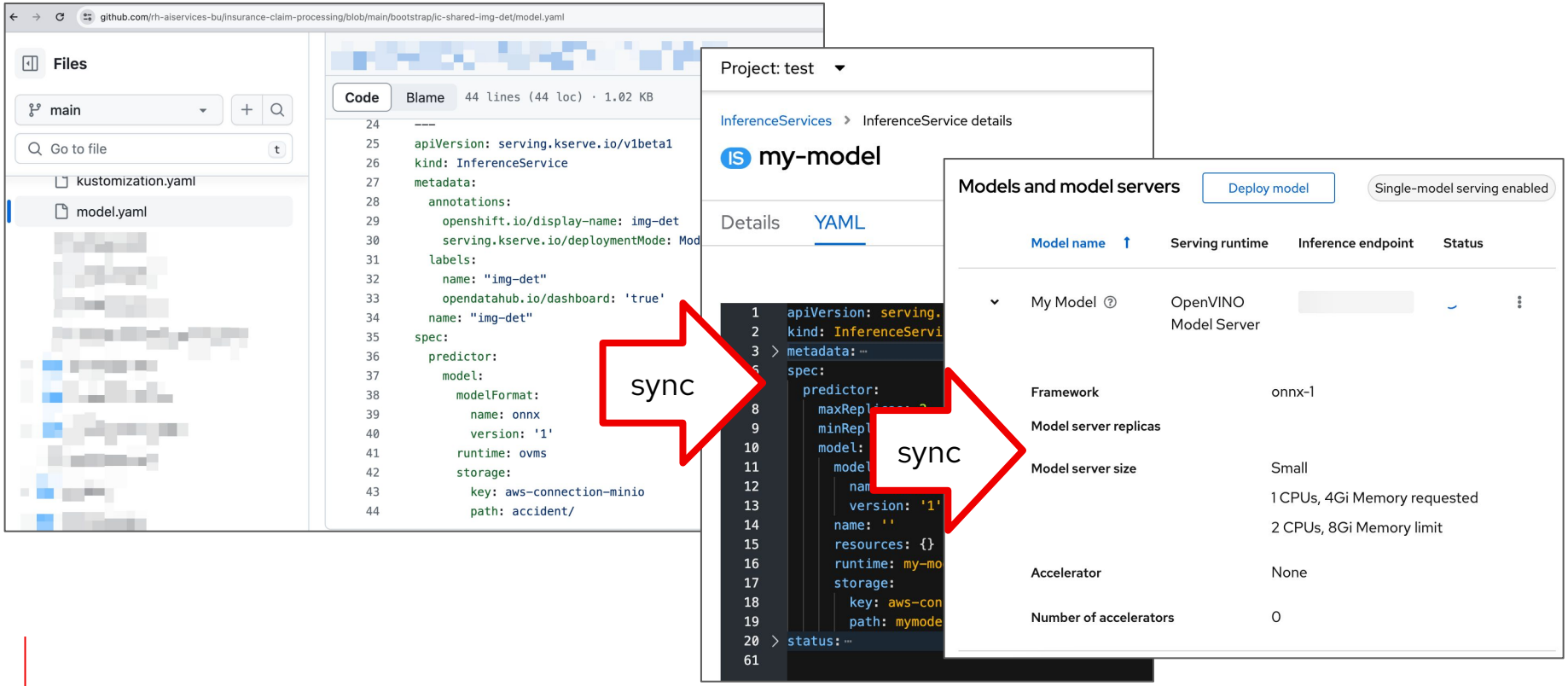




Red Hat

UI to Yaml

GitOps



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 to 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.

Red Hat

Product (and Open Source Projects) Overview

DataOps

- S3 protocol support
- Starburst integration
- Watsonx.data
- Elastic Vector Database

Model Training

- **UI**
  - JupyterLab
  - IBM watsonx.ai
- **Notebook Images**
  - Out-of-the-box
  - Custom
- **Frameworks**
  - PyTorch
  - Tensorflow
- **GPU/Accelerators**
  - NVIDIA, Intel, AMD
  - NVIDIA NIM
  - NVIDIA Rapids
  - Intel AI Analytics
- **Distributed Training**
  - CodeFlare stack
  - NVIDIA TAO Toolkit
  - Watsonx.ai Tuning Studio
- **Version control** (Git)
- **Package Management** (Anaconda)

Integrate models in app dev

- **Model Serving**
  - KServe
  - ModelMesh
  - OpenVINO Model Server
  - Custom runtimes
  - Caikit
  - TGIS
  - vLLM
- **Workflows**
  - Data Science Pipelines
  - GitOps
  - Watsonx.ai

Model monitoring and management

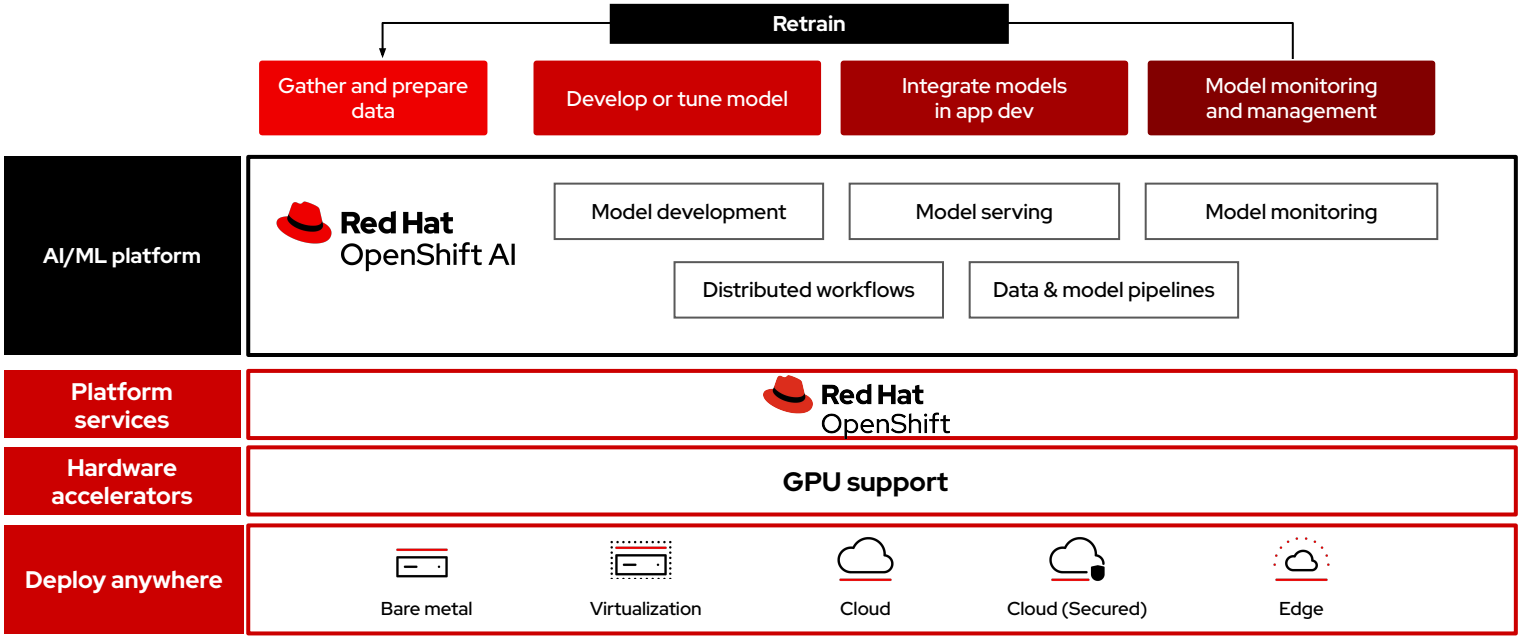
- **Monitoring**
  - Model Mesh metrics
  - Prometheus
  - Out-of-the box performance and Ops metrics
- **Governance**
  - Watsonx.governance
  - Pachyderm

Optional ISVs

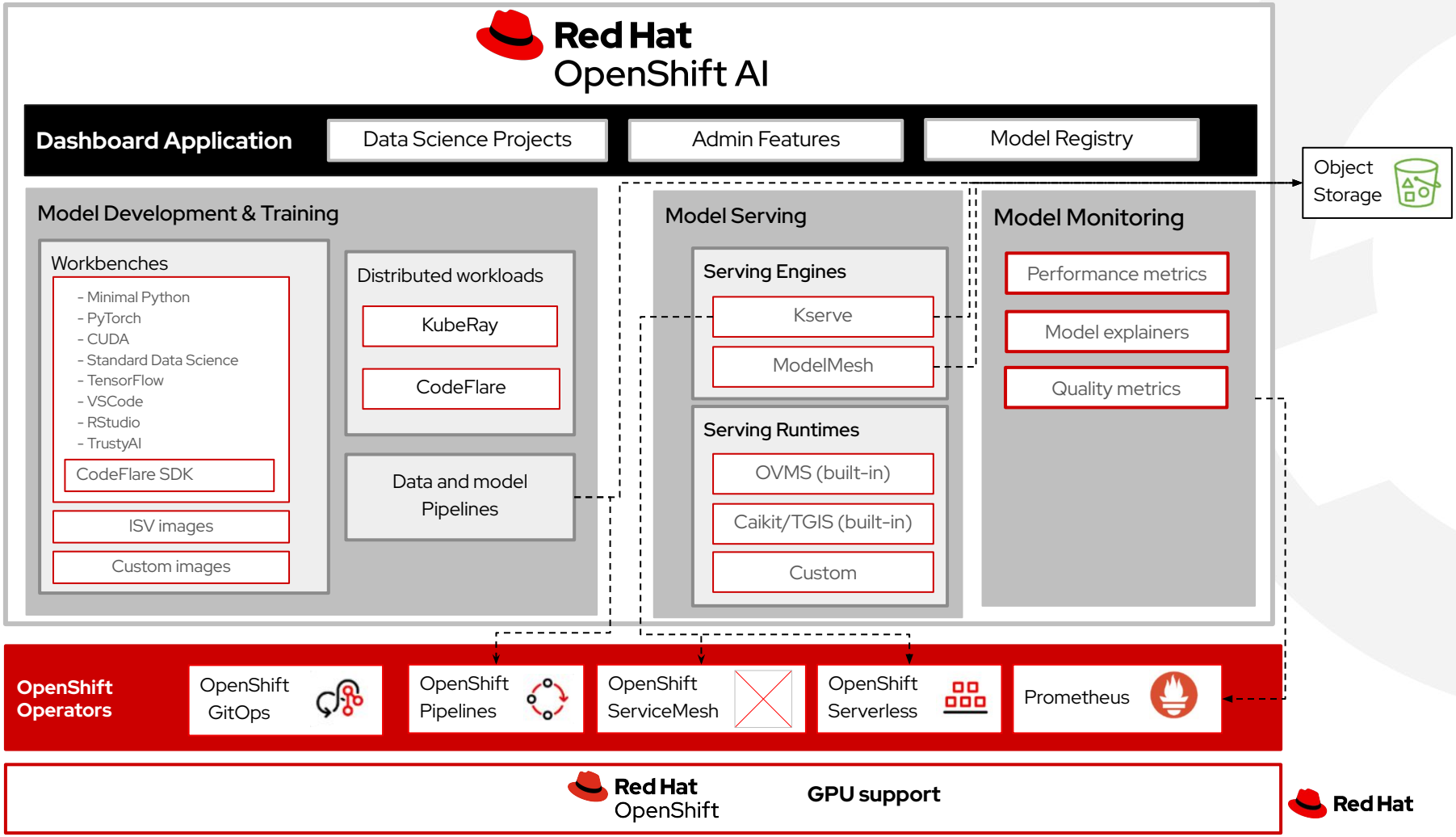
Red Hat

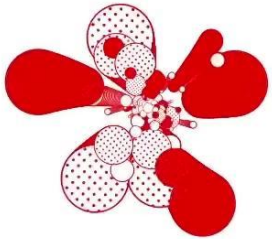
Red Hat OpenShift AI

Red Hat's AI/ML platform



Red Hat


Red Hat
OpenShift AI
Dashboard Application
Data Science Projects
Admin Features
Model Registry
Object Storage
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
Red Hat OpenShift
GPU support
Red Hat



Red Hat OpenShift cloud services

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

aws

Azure

IBM Cloud

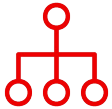
Google Cloud

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.

Red Hat

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)	

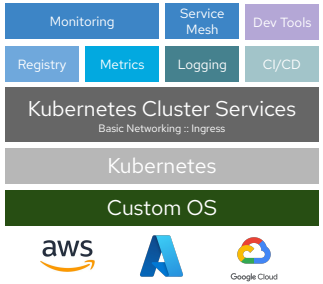
Flavors of RHOAI

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)	

Red Hat

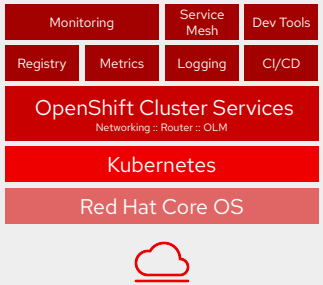
Build and run a platform *versus* using a turnkey cloud service

The Parts



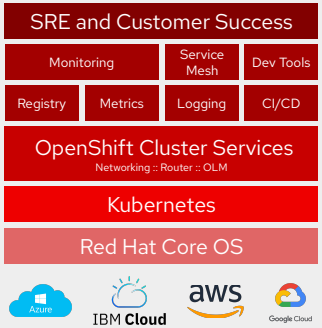
xKS + 'native' services

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

Red Hat

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

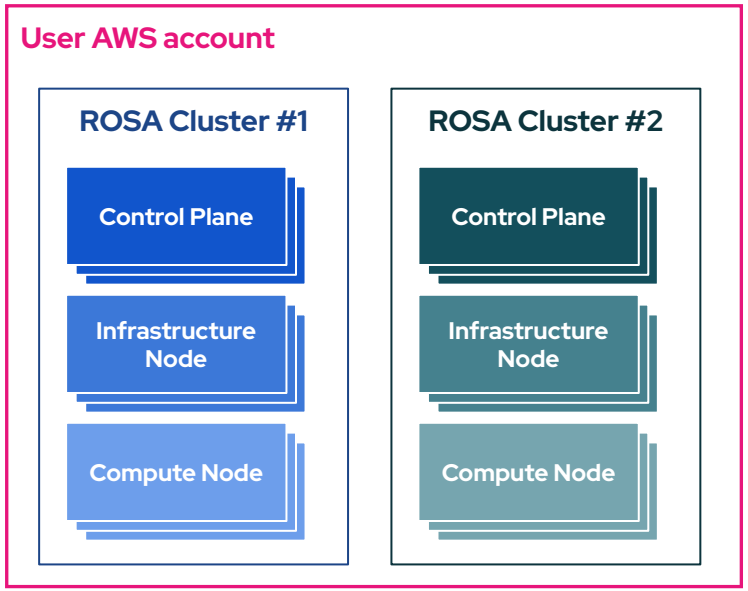
Red Hat

1 "The Total Economic Impact™ of Red Hat OpenShift Cloud Services by Forrester, "Jan. 2022.

ROSA Variants

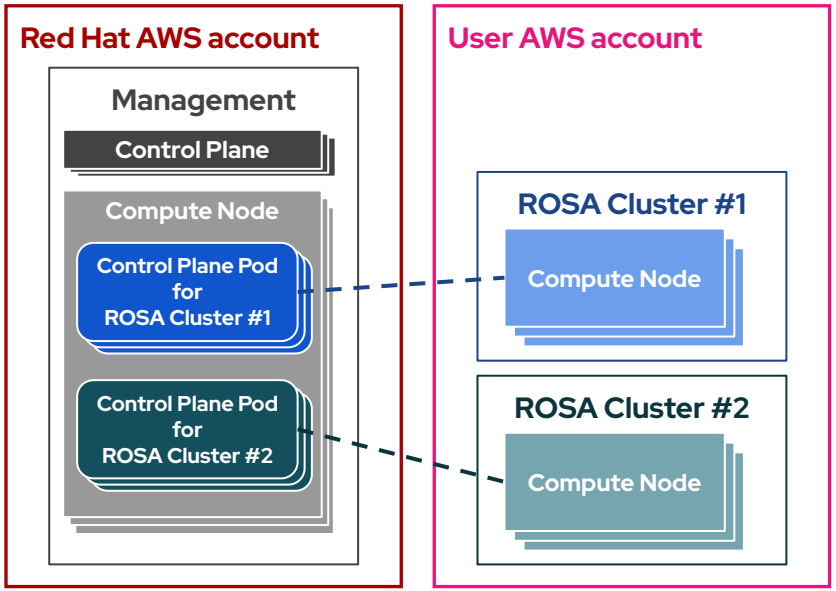
Generally Available

ROSA Classic



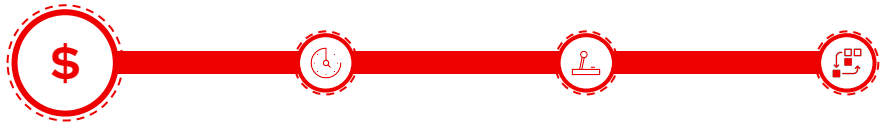
1 or 3 AZs per MachinePool

ROSA with Hosted Control Plane (HCP)



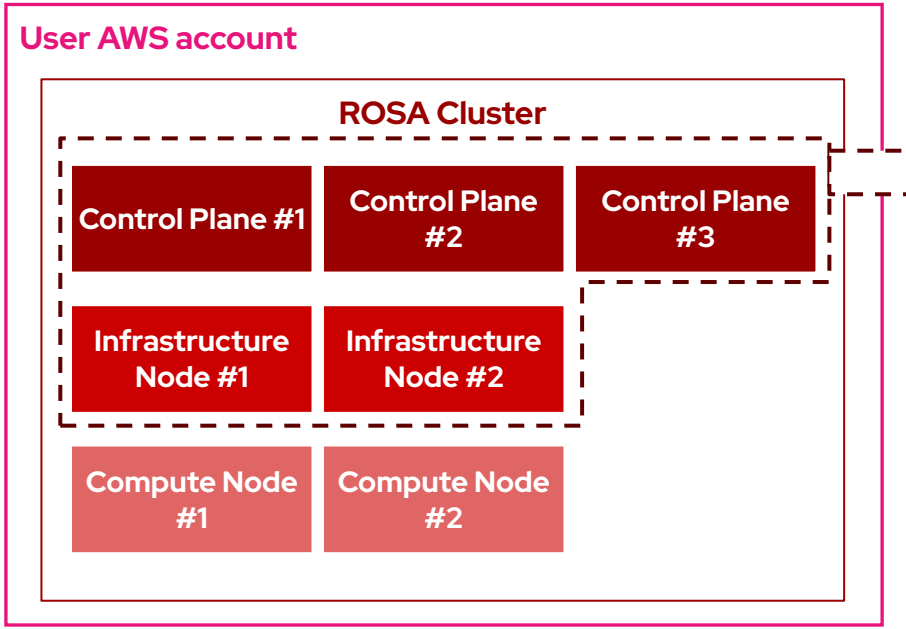
Cluster per Region (each 3 AZs)

Red Hat

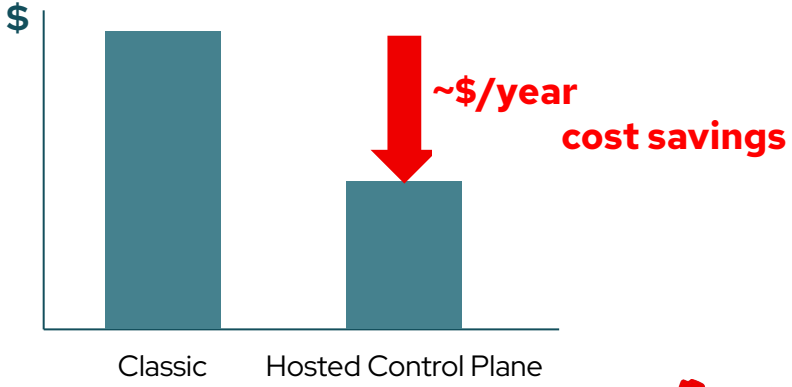


Reduce AWS infrastructure costs

Minimum configuration of ROSA Classic



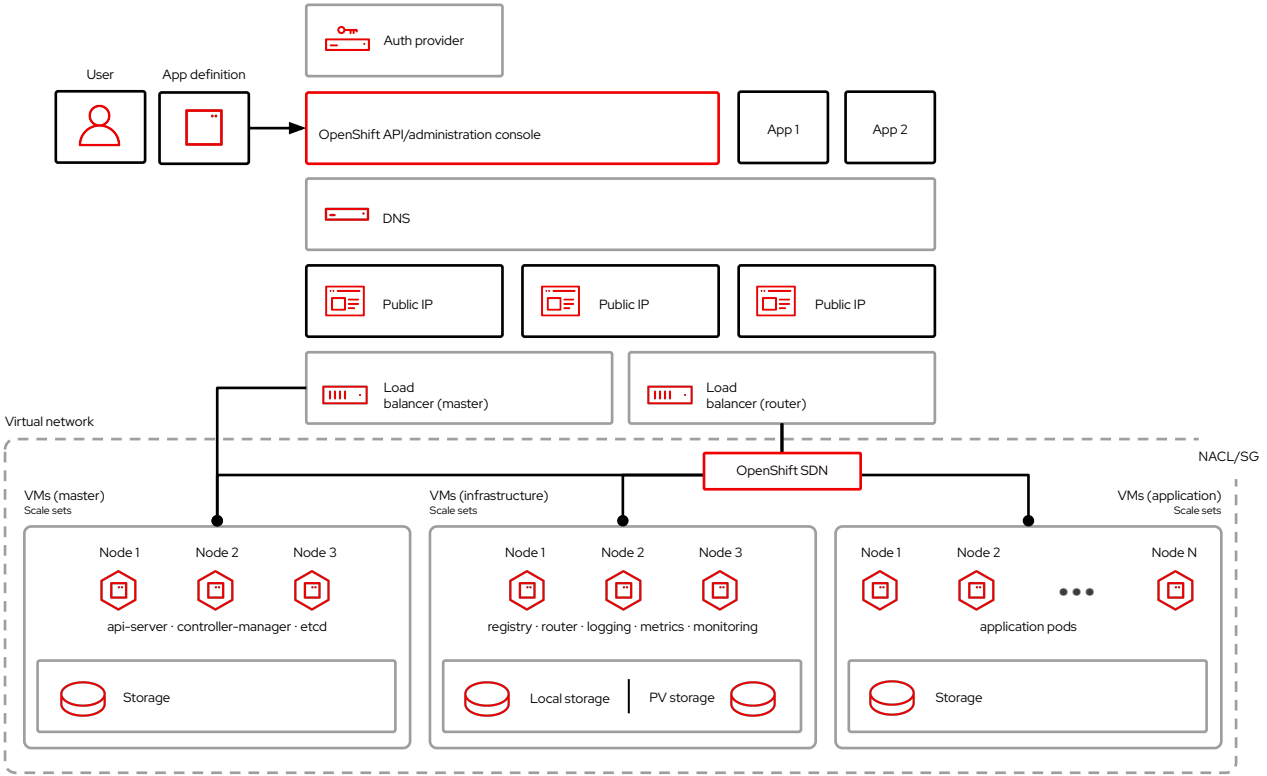
Control Plane Infrastructure Node reduces "total cost of ownership" (TCO)



Red Hat

Complexity of running your own Kubernetes Cluster

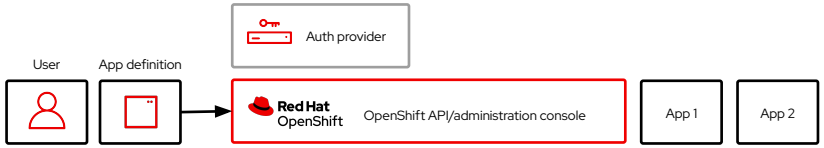
Responsibilities	
User management	Customer
Project and quota management	Customer
Application life cycle	Customer
Cluster creation	Customer
Cluster management	Customer
Monitoring and logging	Customer
Network configuration	Customer
Software and security updates	Customer
Platform support	Cloud provider & Red Hat



Simplify with fully managed clusters

Red Hat OpenShift cloud services

Responsibilities	
User management	Customer
Project and quota management	Customer
Application life cycle	Customer
Cluster creation	Shared
Cluster management	Shared
Monitoring and logging	Shared
Network configuration	Shared
Software and security updates	Cloud provider and Red Hat
Platform support	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...

Red Hat

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

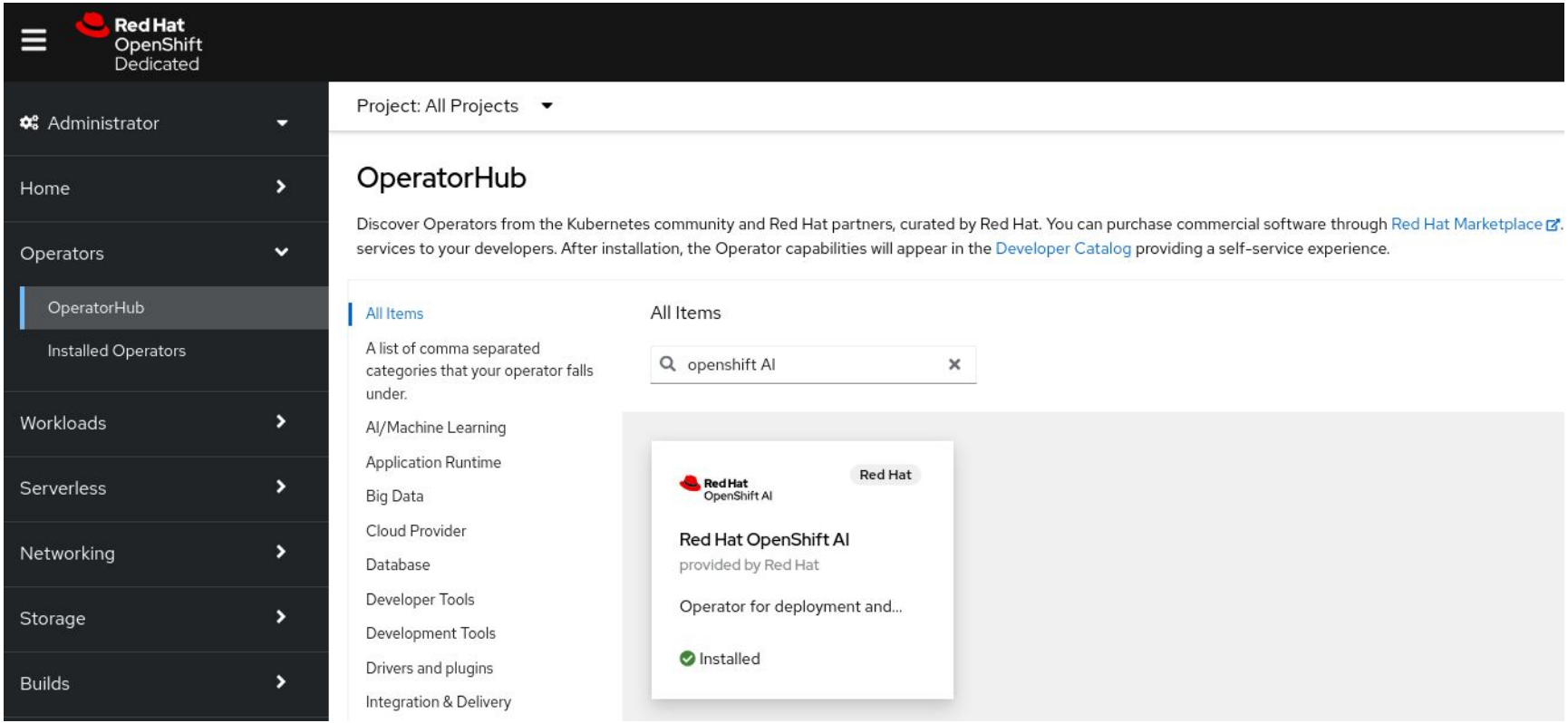
Compliance	Red Hat OpenShift Service on AWS (ROSA)	Red Hat OpenShift Service on AWS (ROSA) with hosted control planes (HCP)
HIPAA Qualified	Yes	No
ISO 27001	Yes	Yes
ISO 27017	Yes	Yes
ISO 27018	Yes	Yes
PCI DSS	Yes	Yes
SOC 1 Type 2	Yes	Yes
SOC 2 Type 2	Yes	Yes
SOC 3	Yes	Yes
FedRAMP High[1]	Yes (GovCloud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

https://docs.openshift.com/rosa/rosa_architecture/rosa_policy_service_definition/rosa-policy-process-security.html

Red Hat

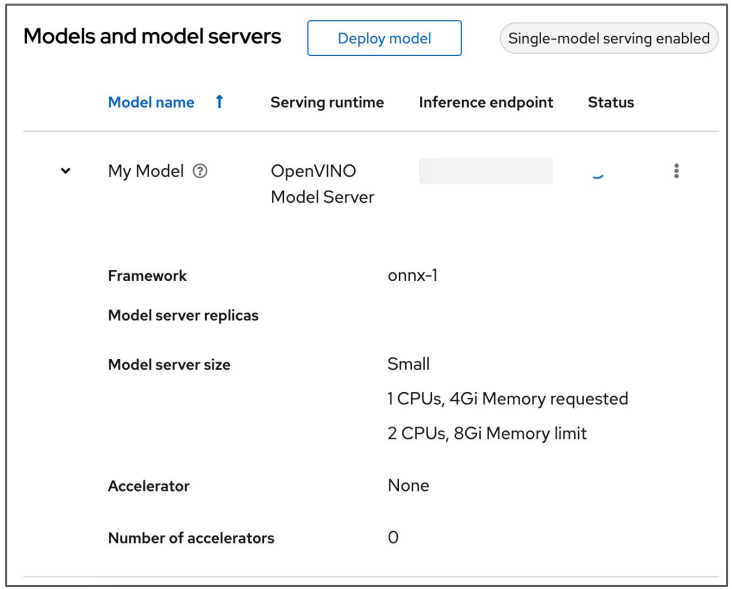
Where do we start?

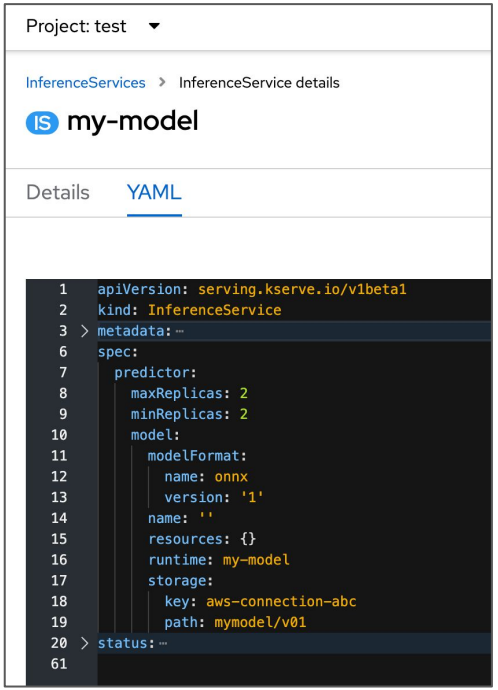
Install RHOAI



UI to Yaml

Everything in RHOAI has an OpenShift representation

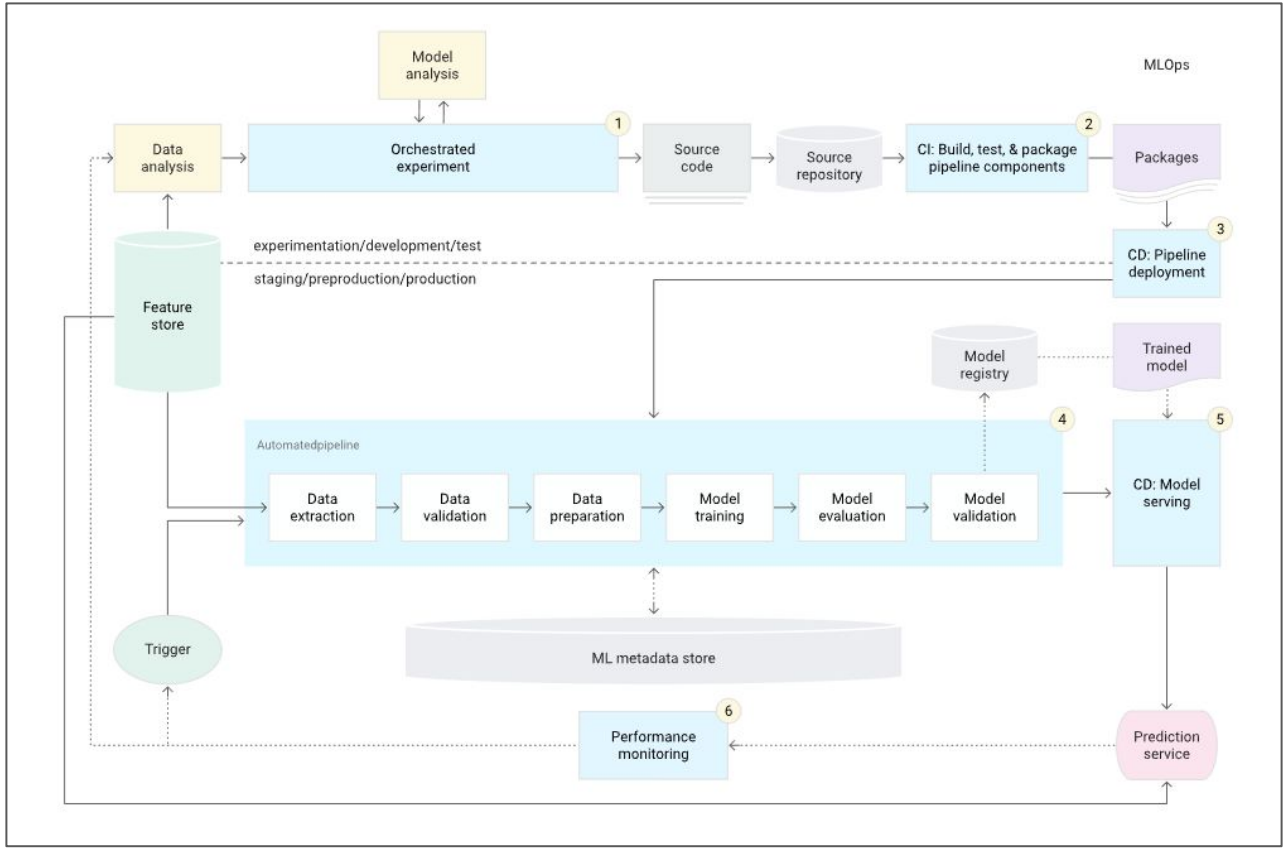




Red Hat

MLOps Automation

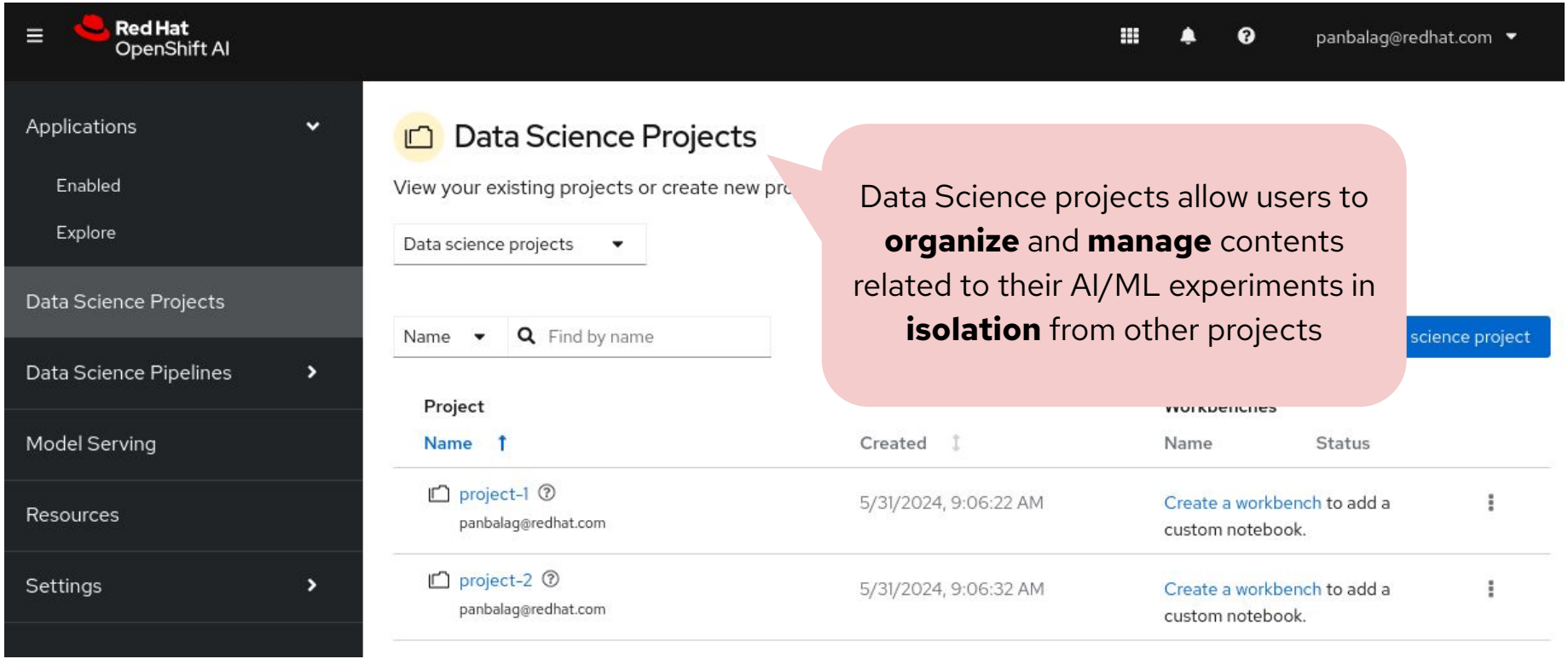
Mature MLOps Flow



https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

Red Hat OpenShift

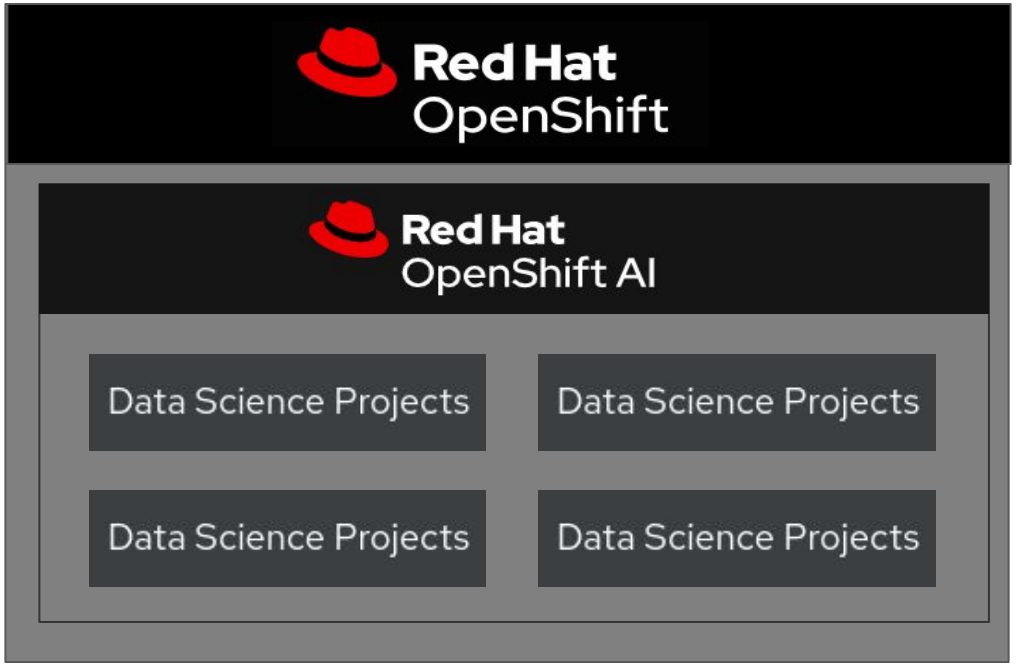
Data Science Projects



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

Red Hat

Data Science Projects



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

Data Science Projects

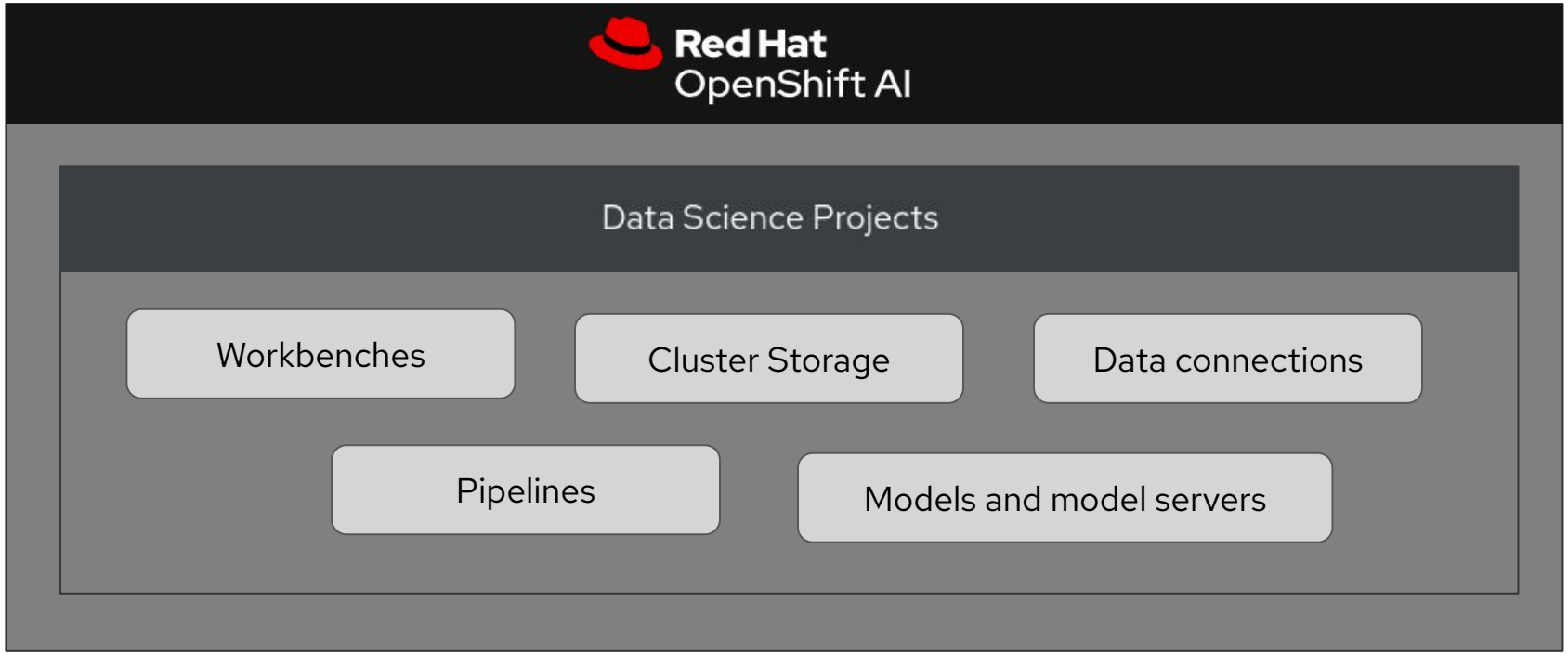
Collaborate within a project

- 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

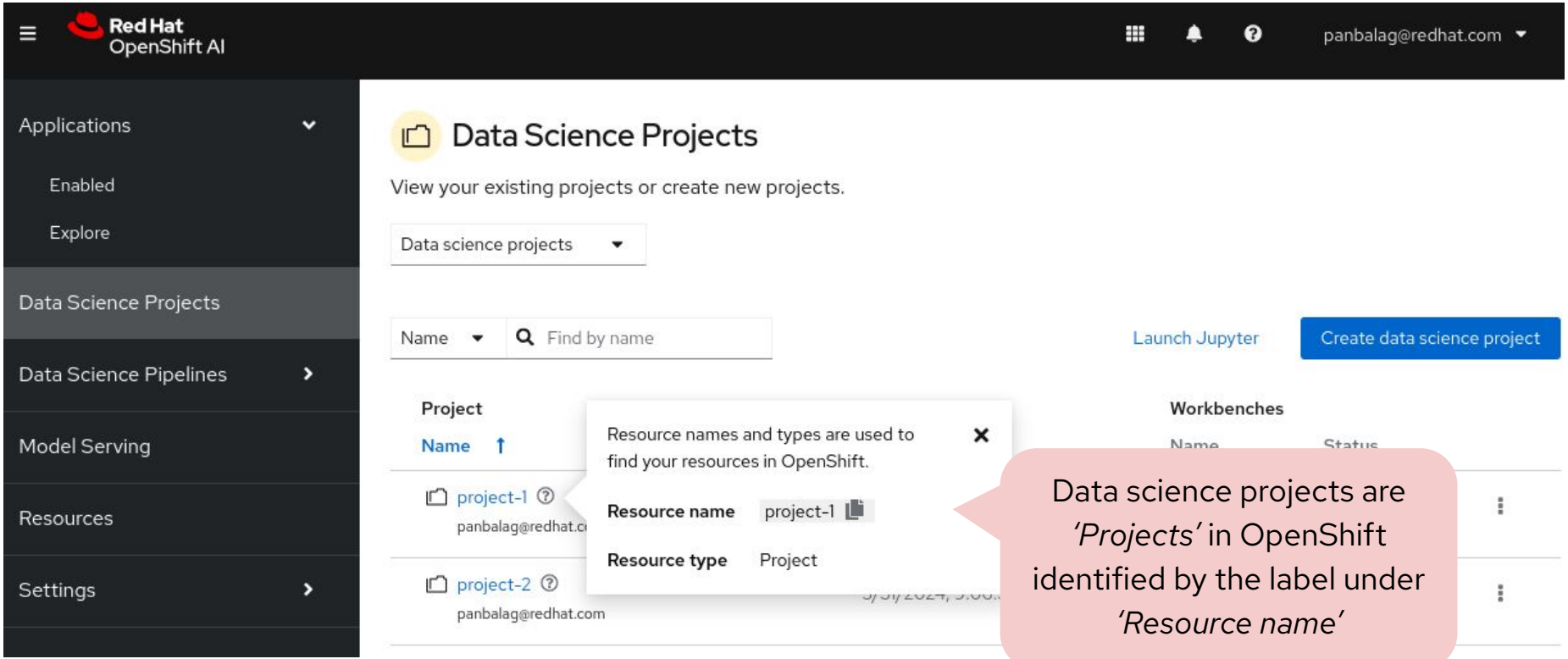
Collaborate between projects

- 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

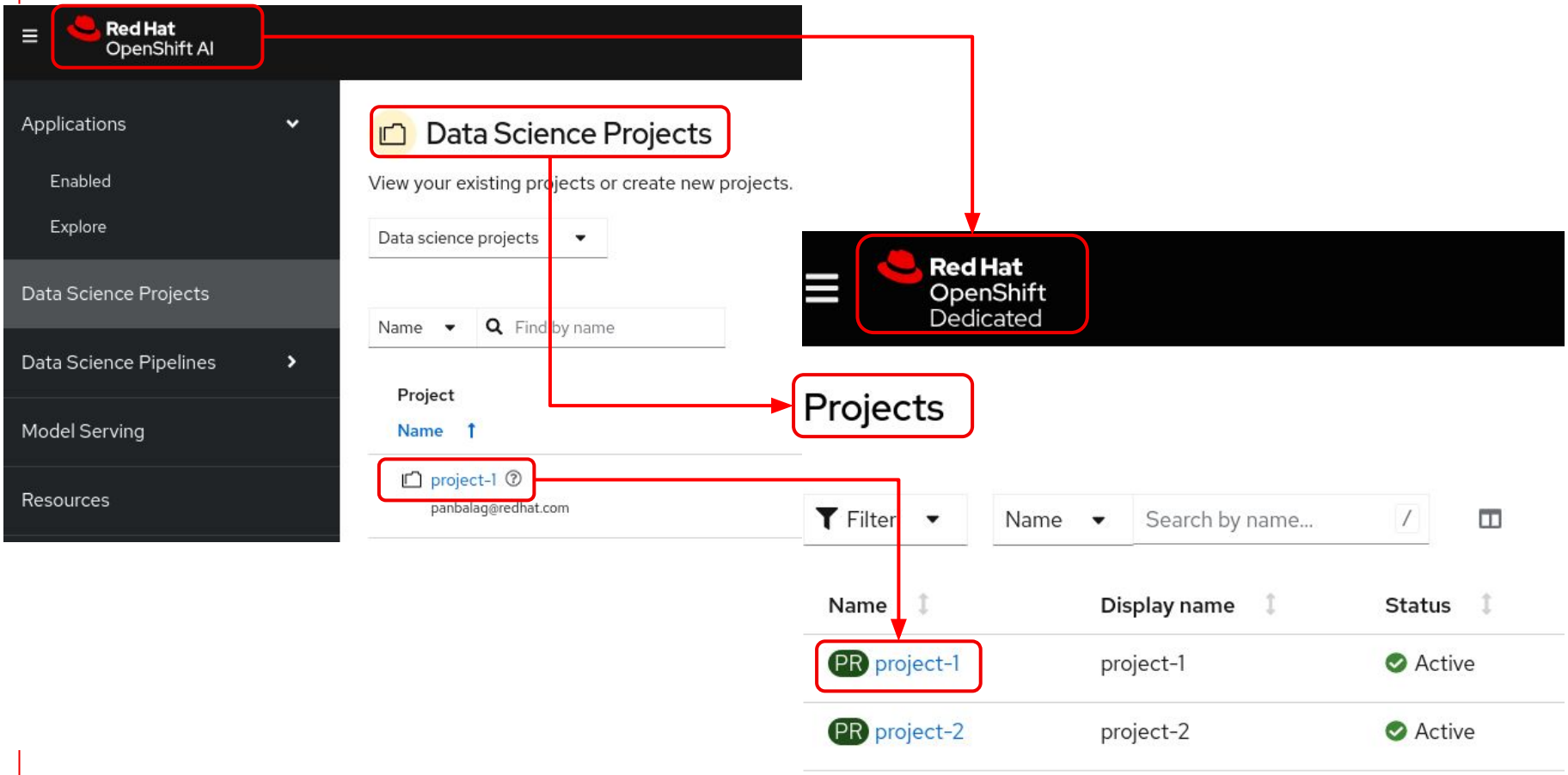


Data Science Projects



Red Hat

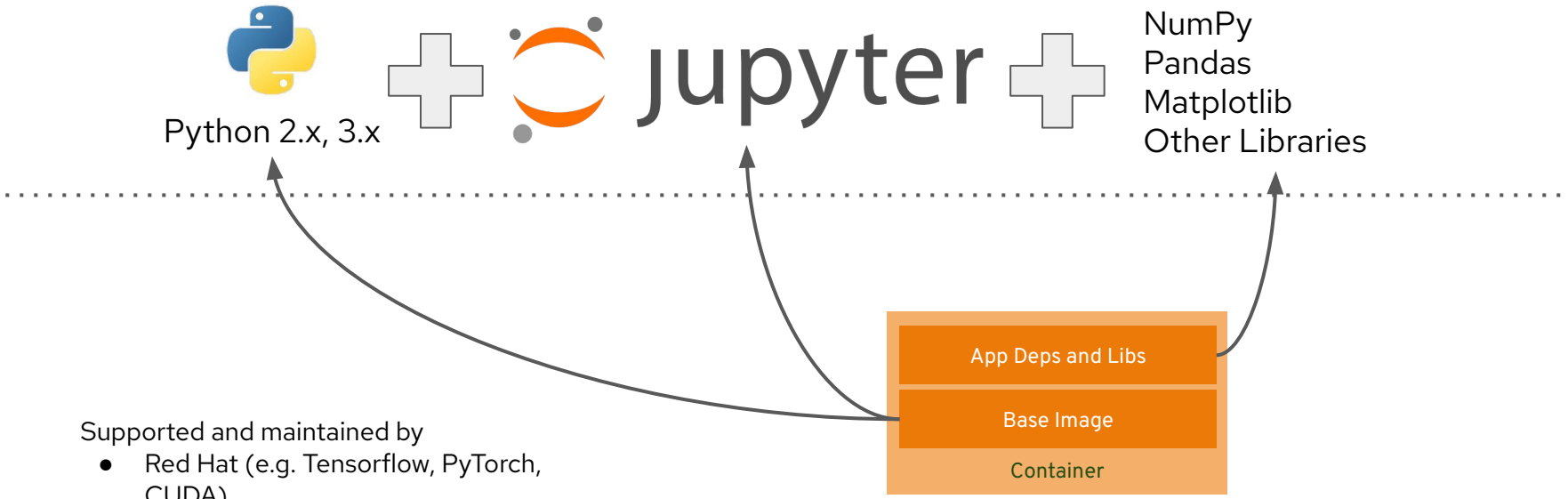
Data Science Projects



Customizing Workbenches

Base Notebook Images

Reproducible and shareable environments for building, training and serving



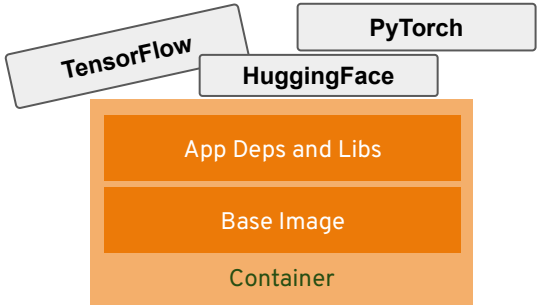
Supported and maintained by

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

Customizing Workbenches

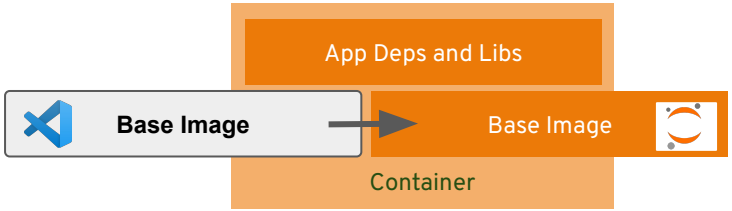
Customizing the workbench

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



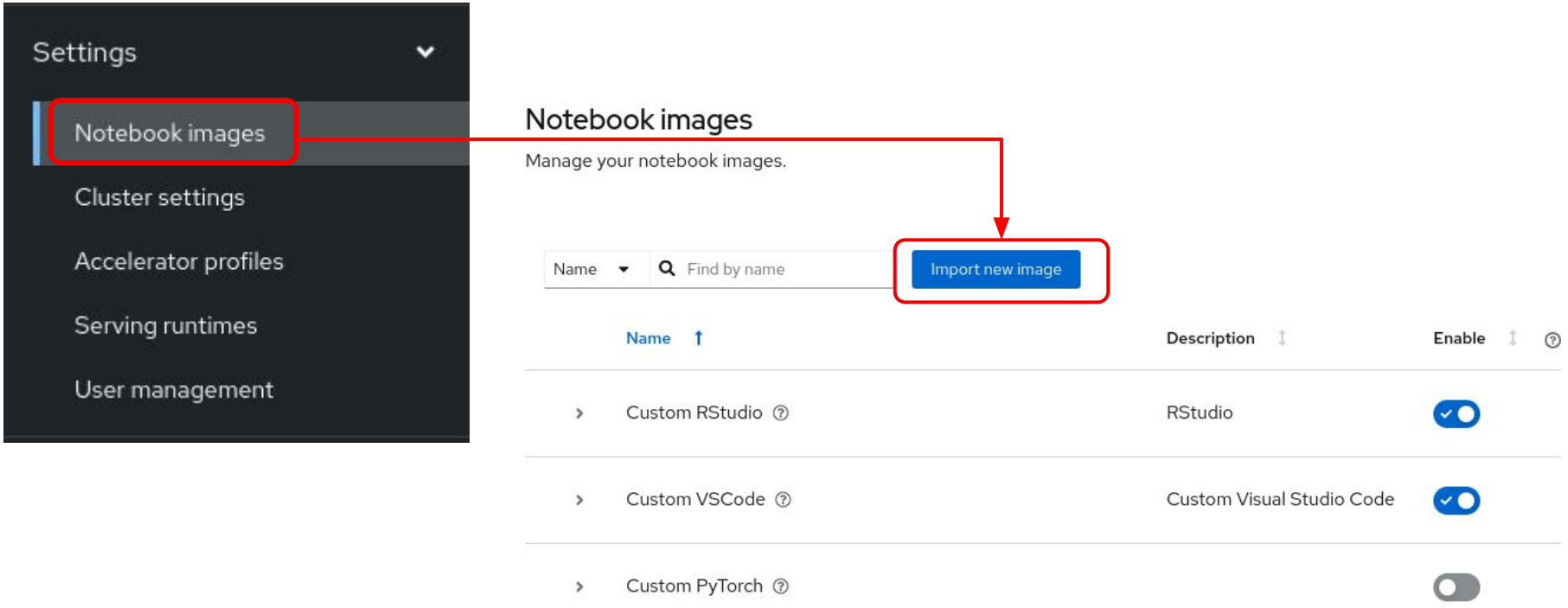
You can now version and maintain it according to your preferences

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

Red Hat

Custom Notebook Image

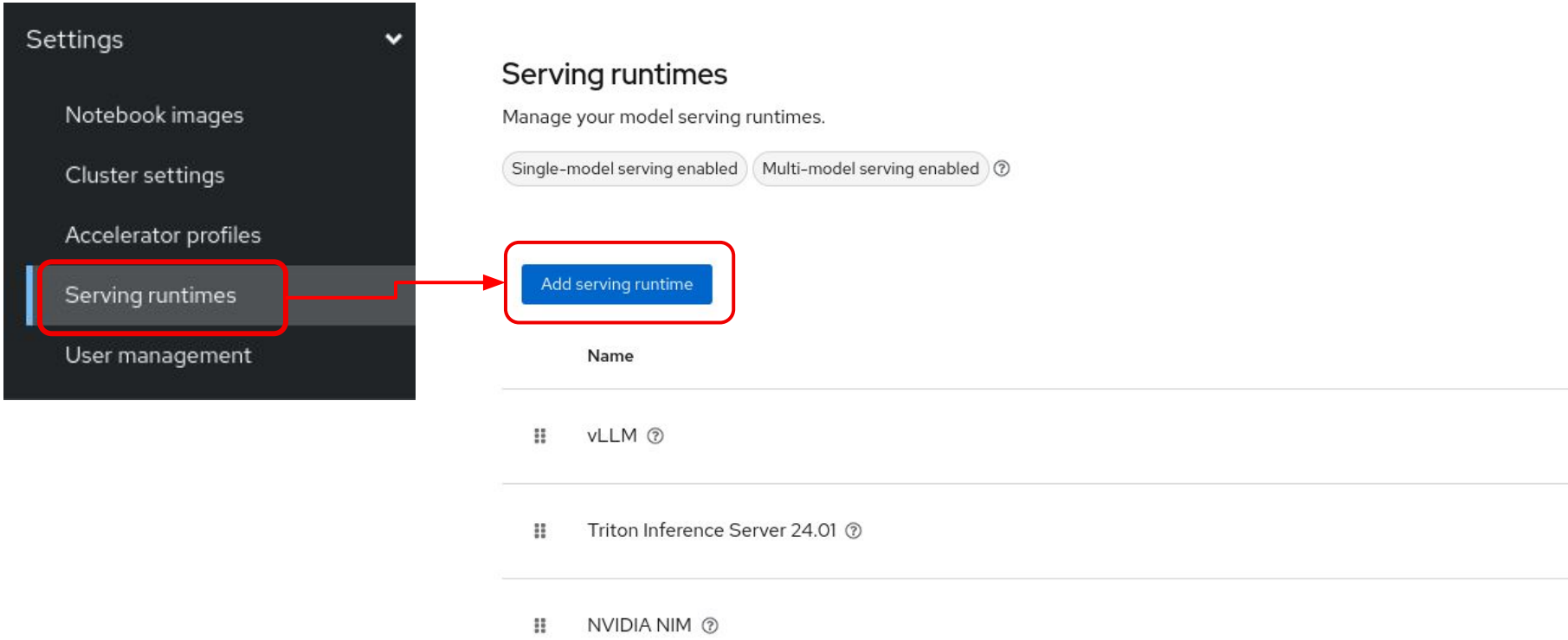
Import new image



Red Hat

Custom Serving Runtime

Add serving runtime



Red Hat

Customize RHOAI Cluster

Enable or disable components

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

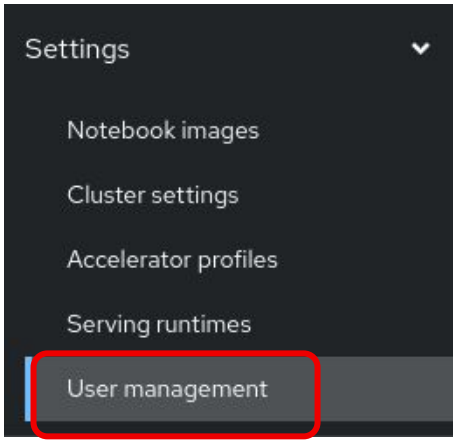
These are the components you can enable/disable:

- 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

```yaml
spec:
  components:
    codeflare:
      devFlags: {}
      managementState: Removed
    dashboard:
      devFlags: {}
      managementState: Managed
    datasciencepipelines:
      devFlags: {}
      managementState: Managed
    kserve:
      devFlags: {}
      managementState: Managed
      serving:
        ingressGateway:
          certificate:
            secretName: knative-serving-cert
            type: SelfSigned
        managementState: Managed
        name: knative-serving
    modelmeshserving:
      devFlags: {}
      managementState: Managed
    ray:
      devFlags: {}
      managementState: Removed
    trustyai:
      devFlags: {}
      managementState: Removed
    workbenches:
      devFlags: {}
      managementState: Managed
```

Red Hat

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.

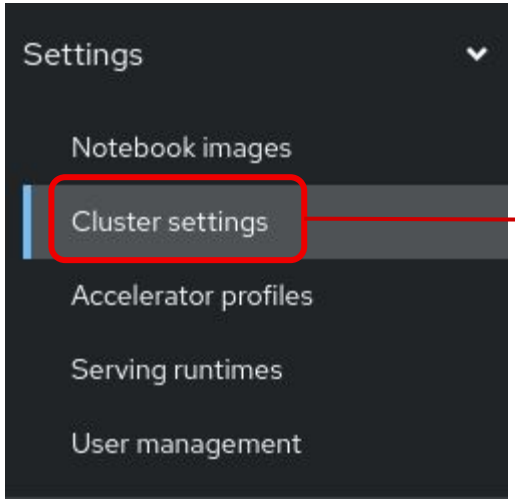
system:authenticated ✕

View, edit, or create groups in OpenShift under User Management

Save changes

Red Hat

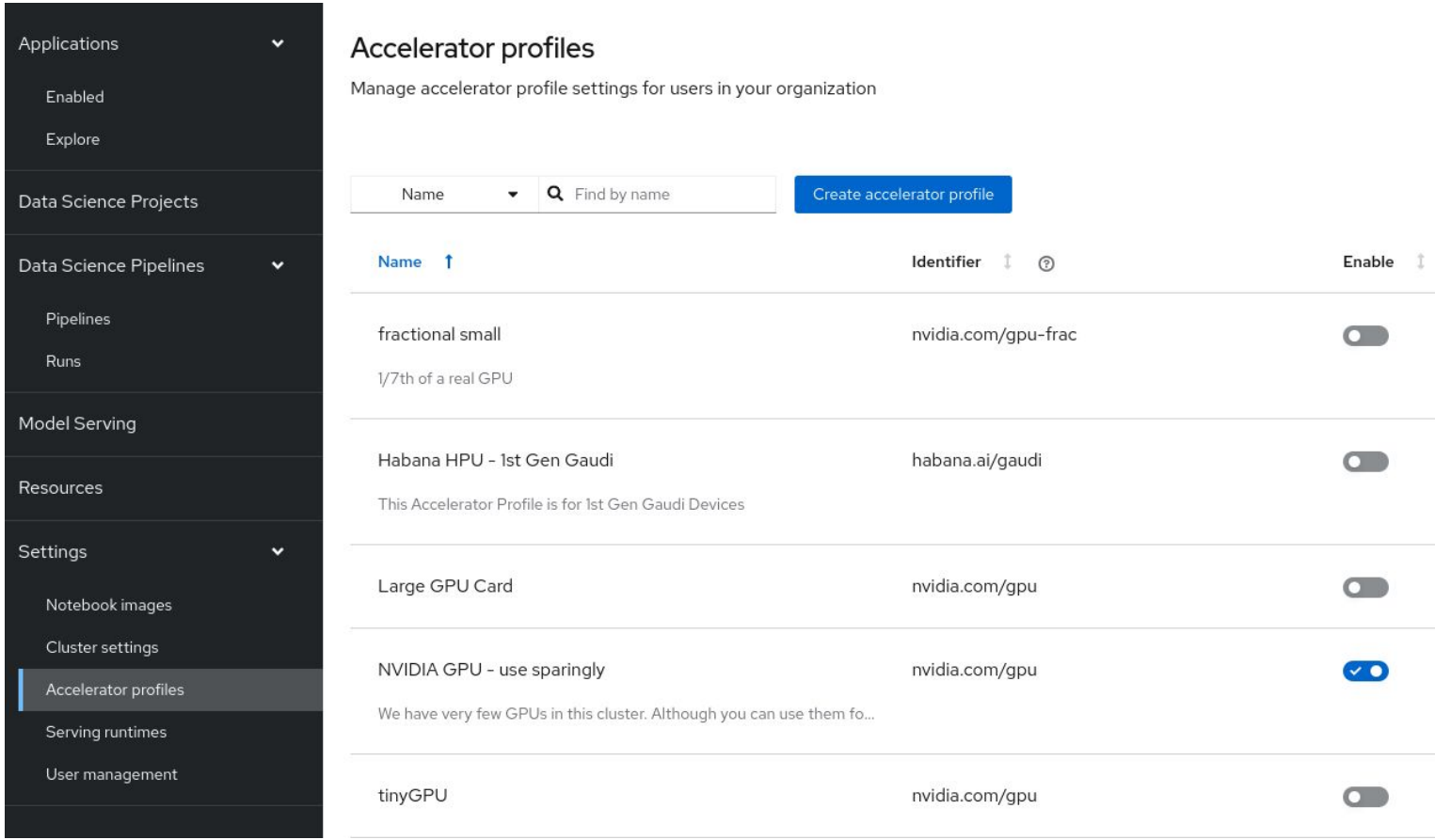
Cluster Settings



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

Red Hat

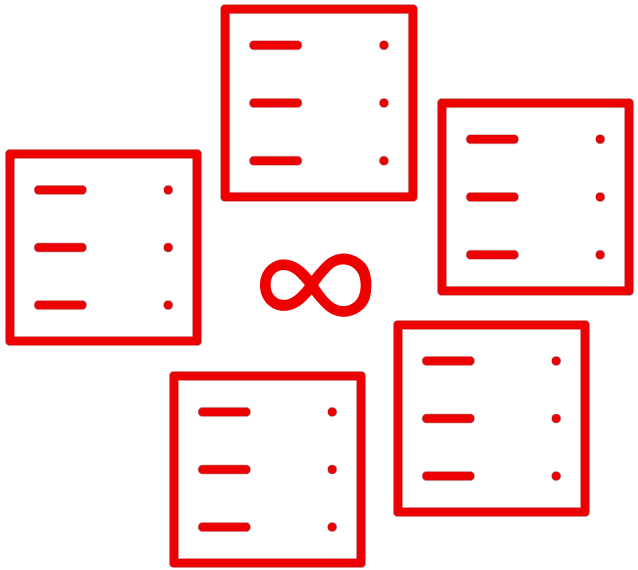
Accelerator Profile



Red Hat

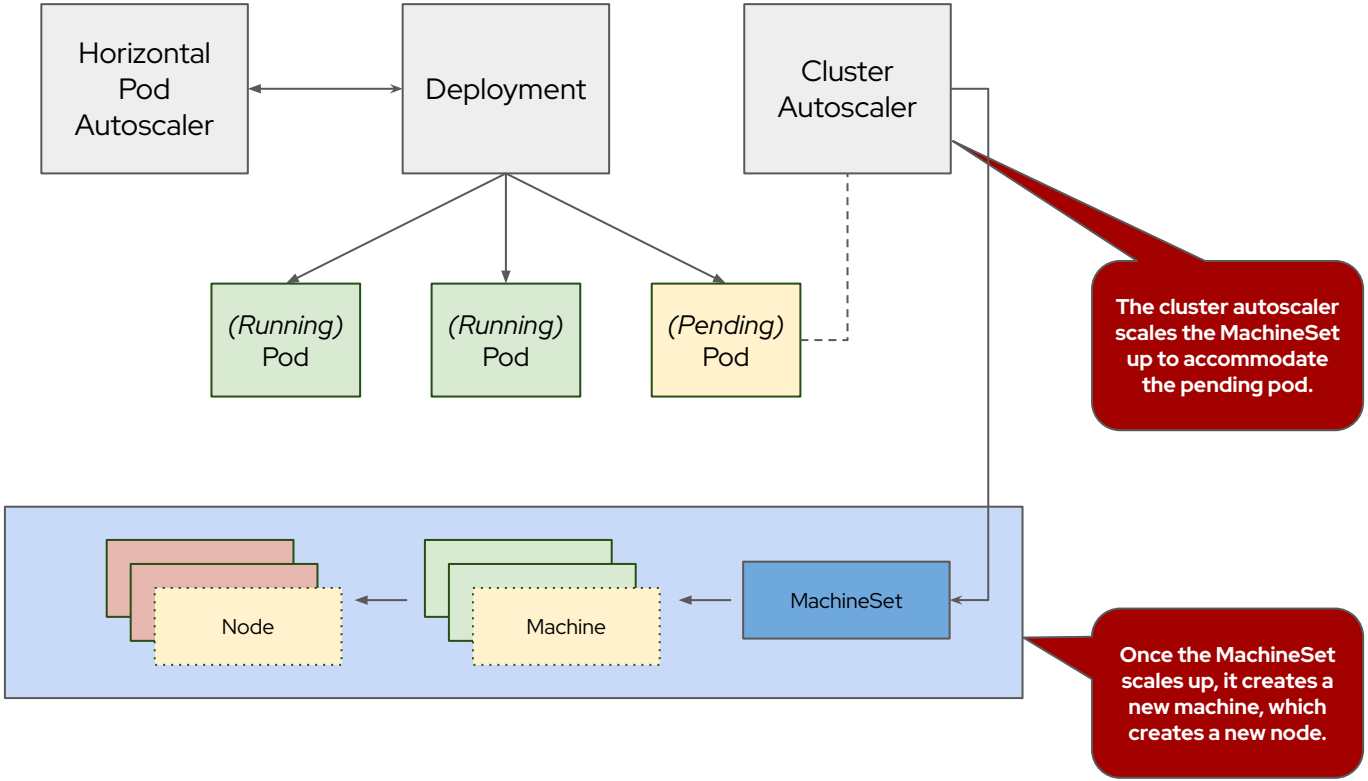
Cluster Autoscaling

Automatically responding to cluster demand.

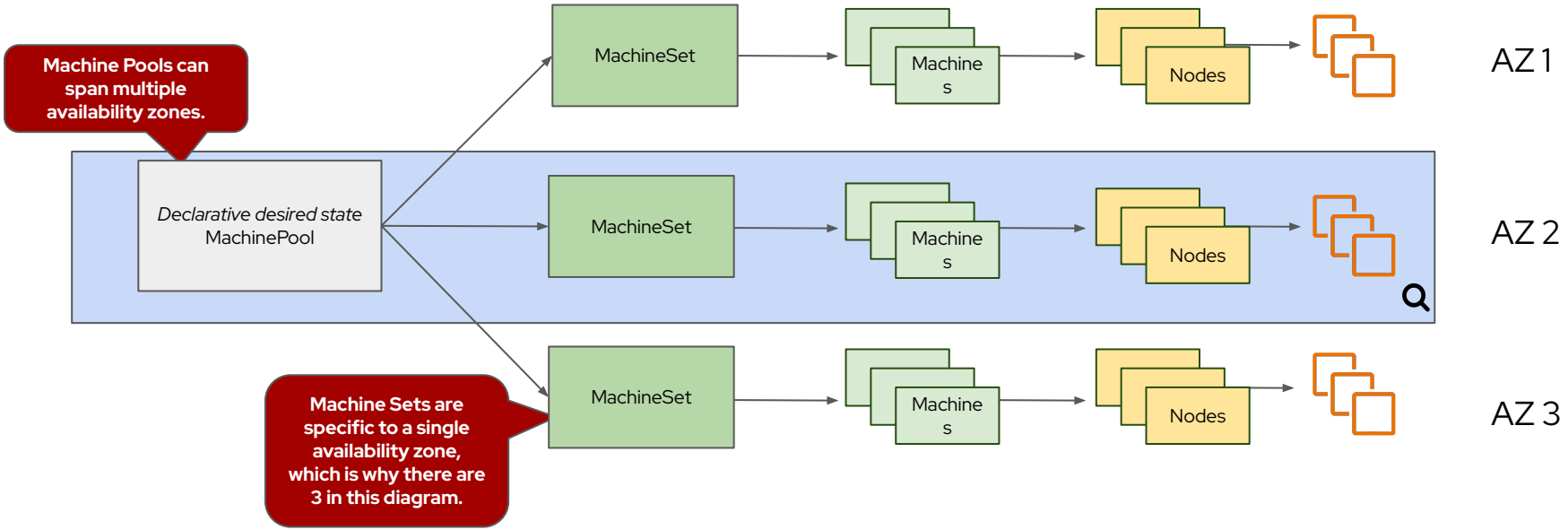


- MachinePools can be scaled to meet applications demands.
- Cluster AutoScaler will provision additional worker nodes when pods can not be scheduled due to resource constraints.
- Cluster AutoScaler will not scale beyond predefined limits.

Learn more about Cluster Autoscaling here:
https://docs.openshift.com/rosa/rosa_cluster_admin/rosa_nodes/rosa-nodes-about-autoscaling-nodes.html

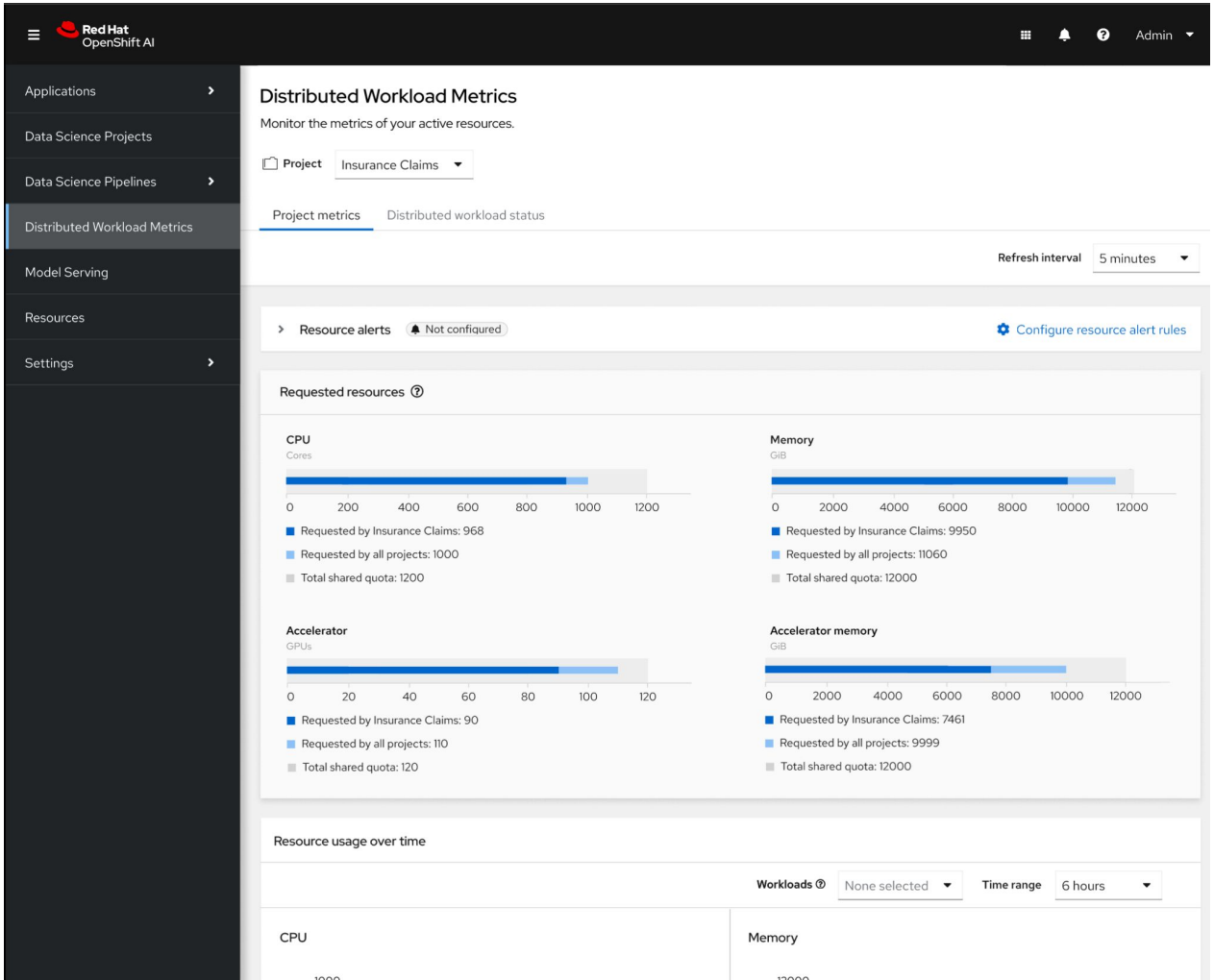

Horizontal Pod Autoscaler
Deployment
Cluster Autoscaler
(Running) Pod
(Running) Pod
(Pending) Pod
The cluster autoscaler scales the MachineSet up to accommodate the pending pod.
Node
Machine
MachineSet
Once the MachineSet scales up, it creates a new machine, which creates a new node.

Red Hat


Machine Pools can span multiple availability zones.
Declarative desired state
MachinePool
MachineSet
MachineSet
MachineSet
Machine s
Machine s
Machine s
Nodes
Nodes
Nodes
AZ 1
AZ 2
AZ 3
Machine Sets are specific to a single availability zone, which is why there are 3 in this diagram.

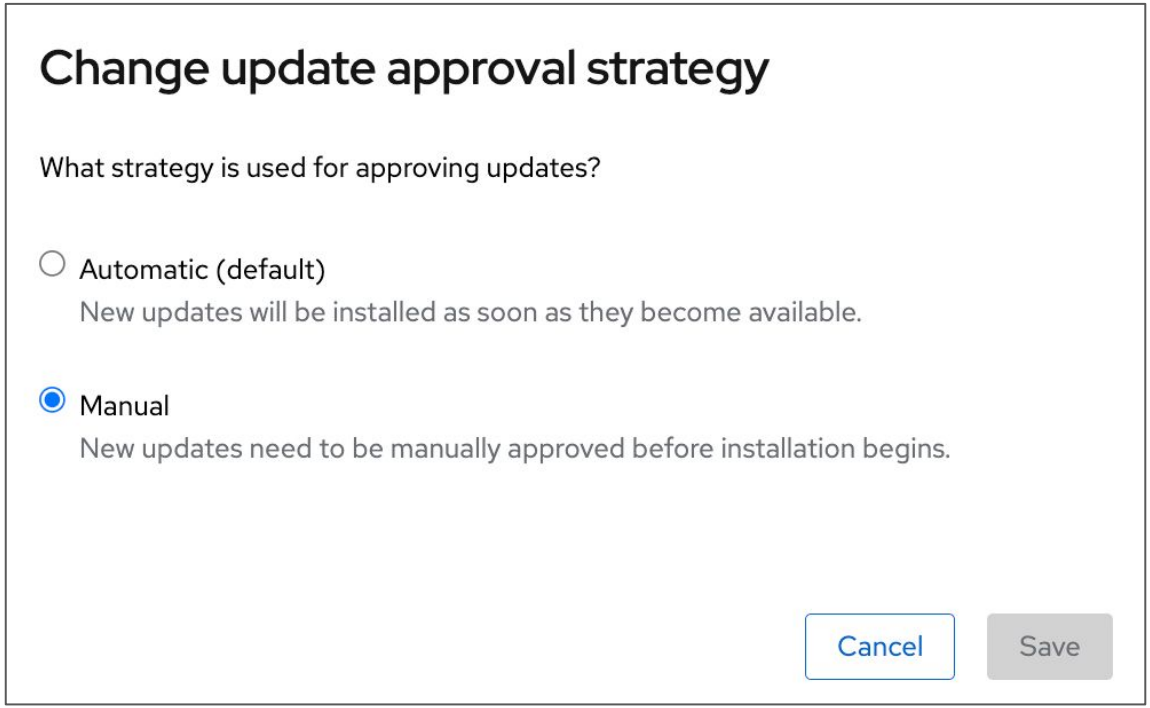
Red Hat

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
Memory

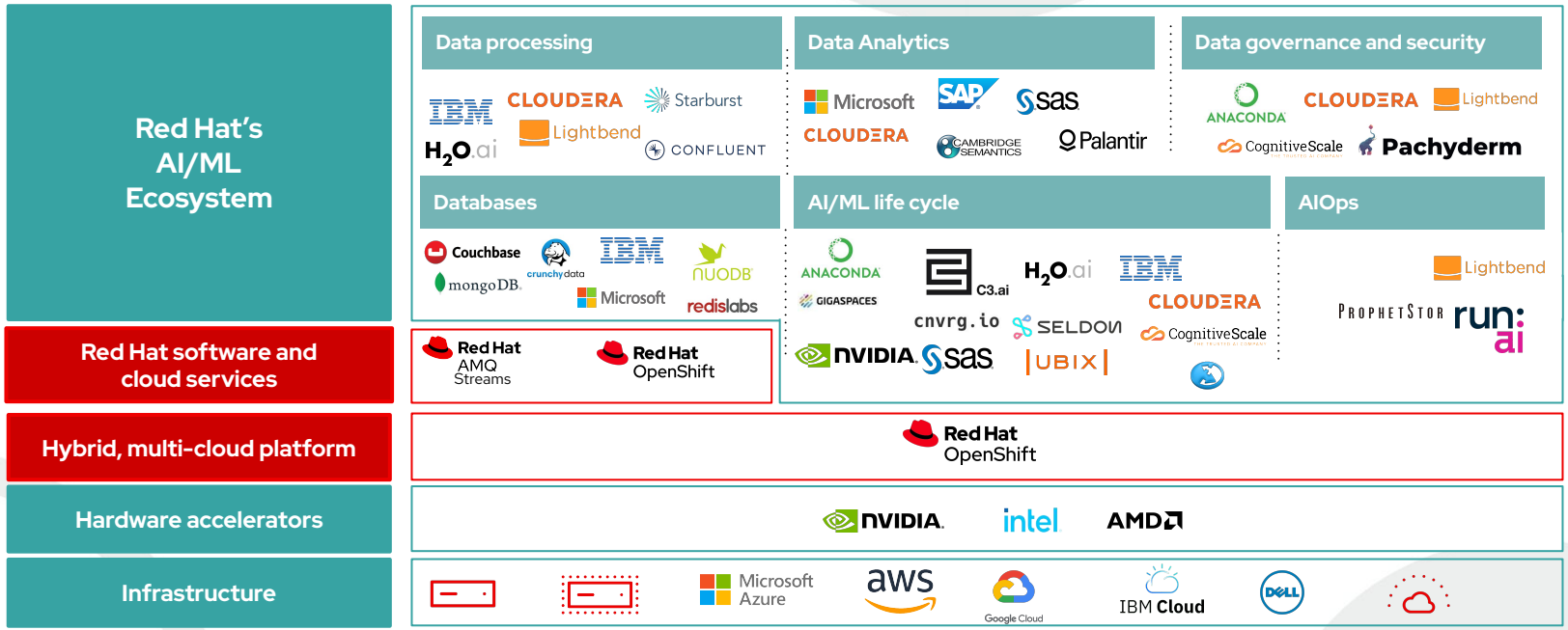
Update RHOAI

Automatic vs Manual



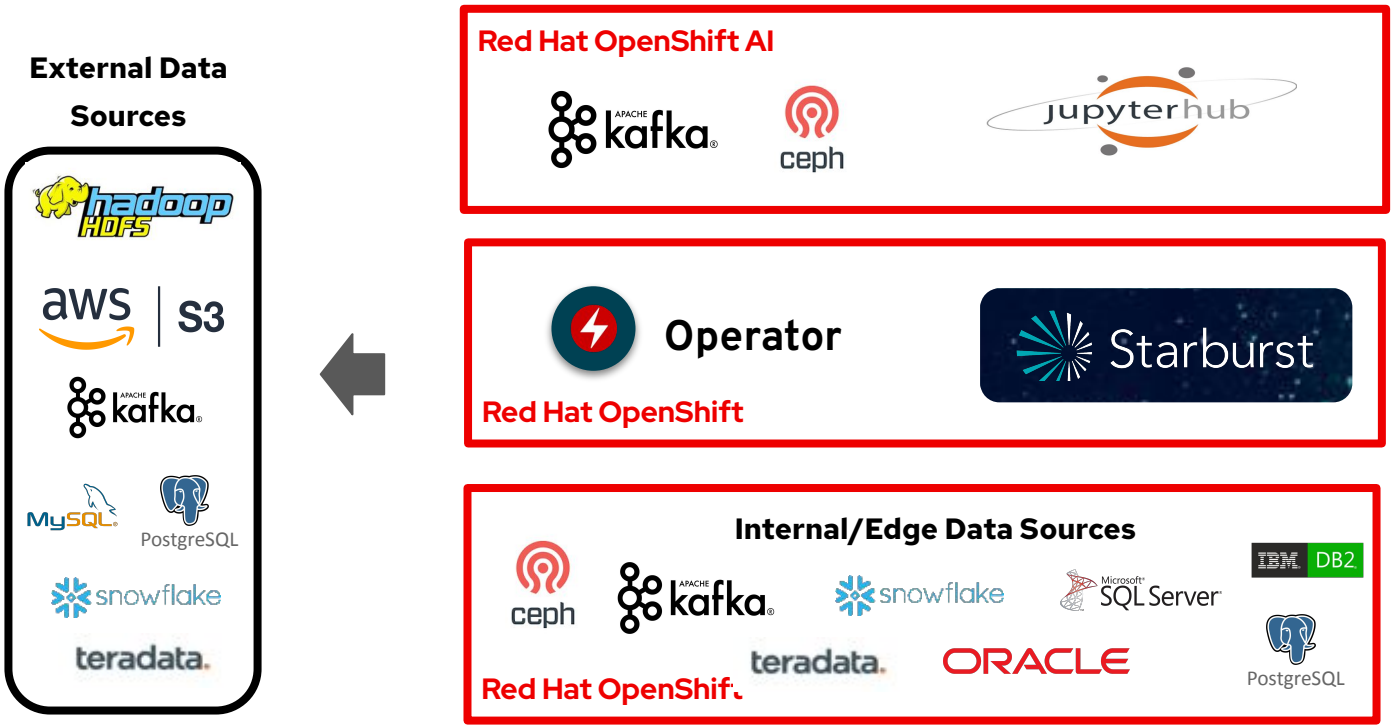
Red Hat

Strategic partnerships + Red Hat AI/ML offerings

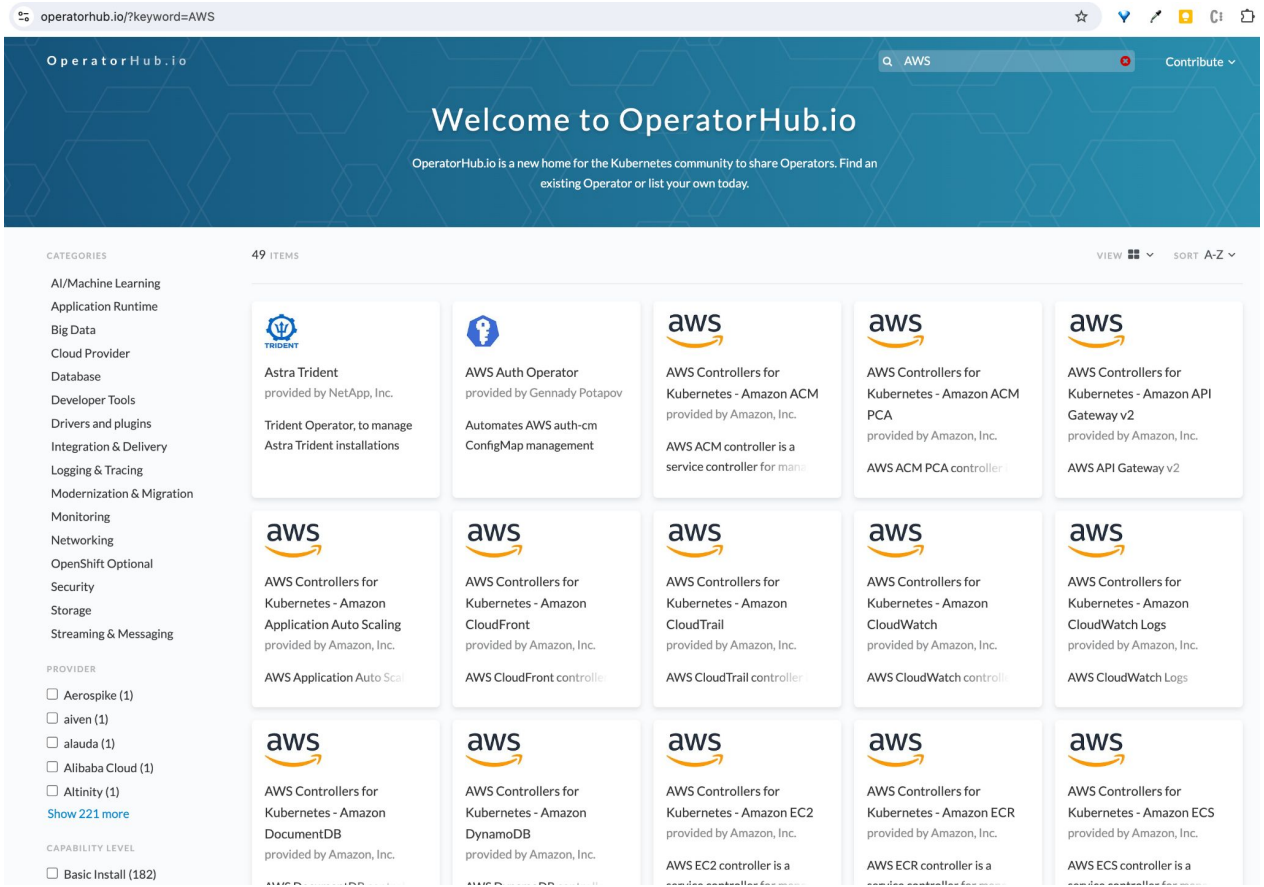


Red Hat

Data Acquisition and Preparation



Operators on ROSA for AWS Services



Strategic partnerships within AI/ML ecosystem

AI/ML life cycle

Data governance and security

Data processing

Data analytics

Databases

AI Ops

Infrastructure partners

Hardware acceleration

Red Hat Services

Training and Certification

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

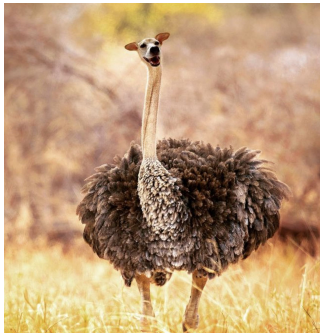
coming soon

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				

Red Hat

Conclusion

- How the AI/ML landscape is evolving: market opportunities
- 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?



Thank you!

Yury Titov
ytitov@redhat.com


linkedin.com/company/red-hat


youtube.com/user/RedHatVideos


facebook.com/redhatinc


twitter.com/RedHat

Red Hat