# Red Hat OpenShift AI Workshop

For Developers & Data Scientists

Eoin Crosbie
Technical Partner Enablement Manager



Red Hat

# Background

# AI is becoming a part of our everyday lives

Chat GPT

**Ansible Lightspeed**

with IBM **Watson** Code Assistant

Bard

Bing

GitHub Copilot

DALL·E 2

# AI/ML is changing the World


The disappearing computer -- and a world where you can take AI everywhere

360,055 views | Imran Chaudhri • TED2023

Share Add Like (10K) Read transcript

In this exclusive preview of groundbreaking, unreleased technology, former Apple designer and Humane cofounder Imran Chaudhri envisions a future where AI enables our devices to "disappear." He gives a sneak peek of his company's new product -- shown for the first time ever on the TED stage -- and explains how it could change the way we interact with tech and the world around us. Witness a stunning vision of the next leap in device design.

# AI has undergone significant evolution in the last decade

## The evolution of AI: from BI to Chat GPT

### 2000s

**Business Analysis & Intelligence**

- Collecting data
- Storing & moving data
- Transforming data

### 2010s

**Advanced Analytics & Traditional AI**

- Data science techniques
- Predictive analytics
- Real-time decision making

### 2020s

**Foundation Models & AI-enabled apps**

- Deep learning techniques
- Model experimentation
- Model tuning

- Traditional AI/ML still runs businesses today
- Traditional AI prepares organizations to develop strong AI strategies
- Foundation models provide a shortcut for realizing the value of AI
- ChatGPT only solves for content generation use cases only

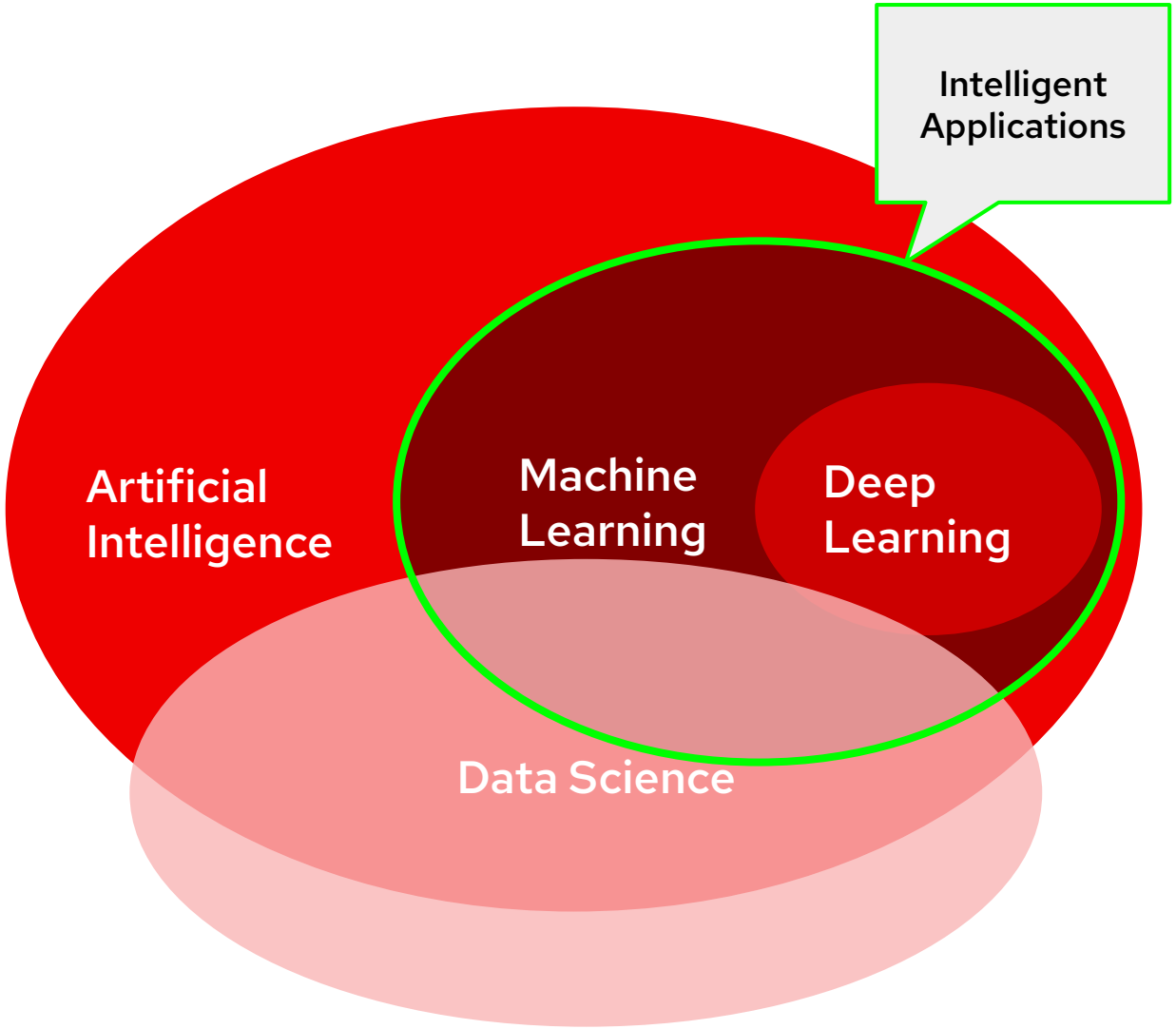
# The Terminology

**Artificial Intelligence (AI)**
Machines imitating intelligent human behavior. The term AI is primarily used by the business community.

**Machine Learning (ML)**
Subset of AI, gives computers the ability to learn without being explicitly programmed. The term ML is primarily used by the technical community.

**Deep Learning (DL)**
Subset of ML, uses layers to progressively extract higher level features from the raw input. Applications include computer vision, image recognition, etc.

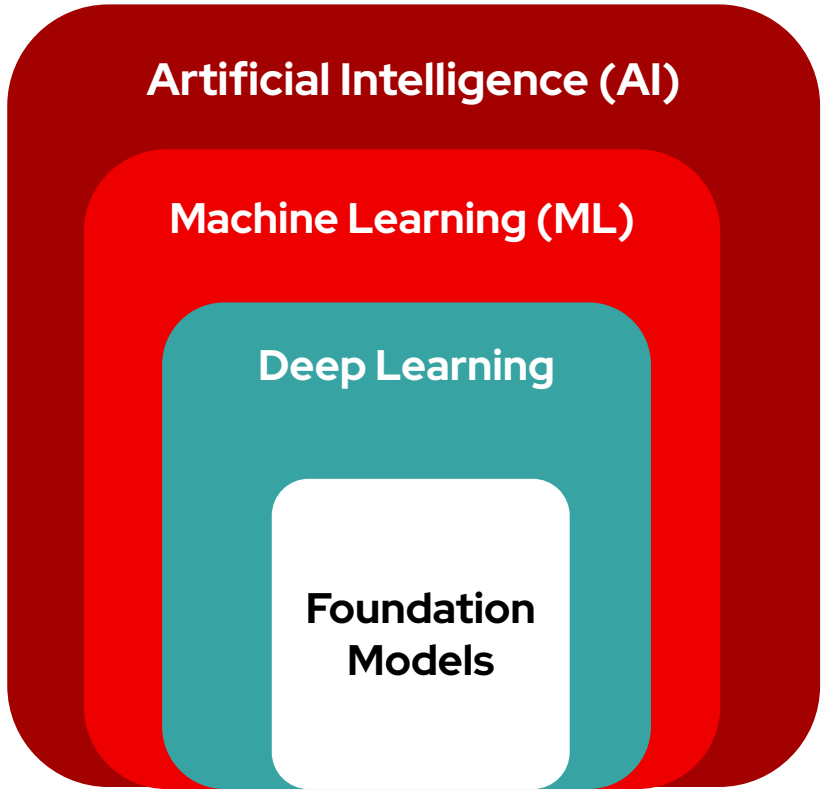
**Data Science**
The discipline of deriving value out of data using analysis for insights, statistics for causality, or machine learning for predictions. The term data science is most often associated with the data scientist persona.




Source:
https://www.redhat.com/en/blog/what-machine-learning
https://towardsdatascience.com/what-is-data-science-8c8fbaef1d37

# Where do Foundation Models fit within AI?

Foundation models serve as the basis for a wide range of downstream AI apps like Chat GPT




**Artificial Intelligence (AI)** - is the field that combines computer science and robust datasets to enable problem-solving.

**Machine Learning (ML)** - is a subset of AI that solves business problems by training statistical models to extract knowledge and patterns from data.

**Deep learning techniques** - subfield of ML that uses large data sets for training models.

**Foundation models (FM) -** large, pre-trained models trained on extensive and diverse datasets to learn general features and patterns, making them versatile for various tasks.

# Generative AI applications are powered by foundation models

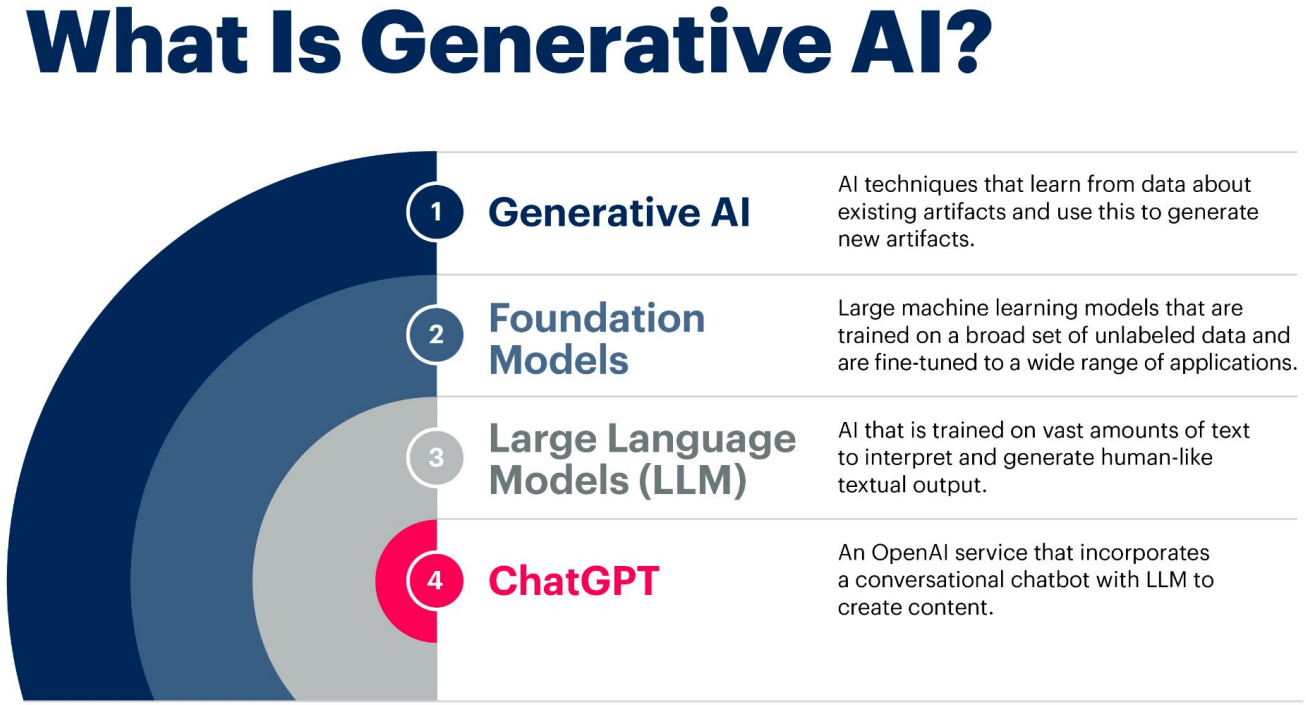
## Foundation models allow developing specialized AI-enabled applications

**Benefits of foundation models:**

- Time to value
- Accuracy
- Accessibility
- Versatility

**Most common Gen AI applications:**

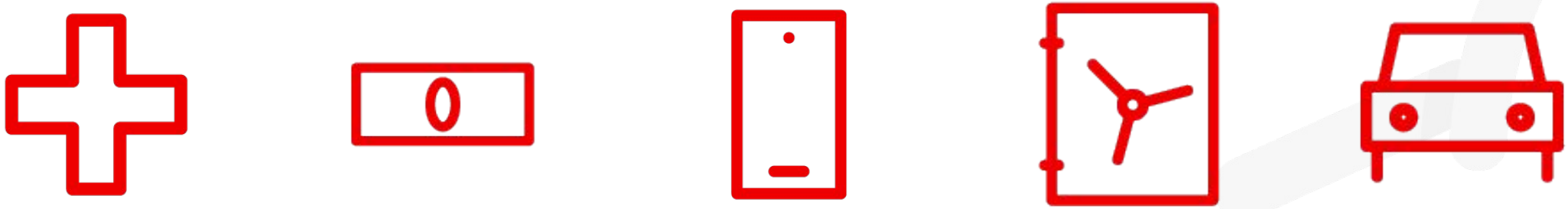
- Text summarization
- Text, code and image generation
- Sentiment analysis
- Classification
- Conversational Q&A




Source: Gartner


Gartner

# Every business has a use for AI/ML



## Healthcare

- Increased clinical efficiency
- Faster/better diagnosis
- Improved outcomes

## Financial services

- More personalized services
- Improved risk analysis
- Reduced fraud
- Better predictions

## Telcos

- Better customer insights/experiences
- Optimized network performance & operations
- Improved threat detection

## Insurance

- Automated claims processing and handling
- Usage-based insurance services

## Automotive

- Autonomous driving
- Predictive maintenance
- Improved supply chains

# Companies say all phases of AI/ML are demanding on infrastructure

## Data preparation and management

30%

Of companies said this was the most infrastructure-intensive

## AI/ML model training

40%

Of companies said this was the most infrastructure-intensive

## Inferencing

30%

Of companies said this was the most infrastructure-intensive

# Enterprises are investing in platforms for AI/ML

Abundance of computing power, data, and the availability of open source ML frameworks are helping make AI/ML real

**$13T**

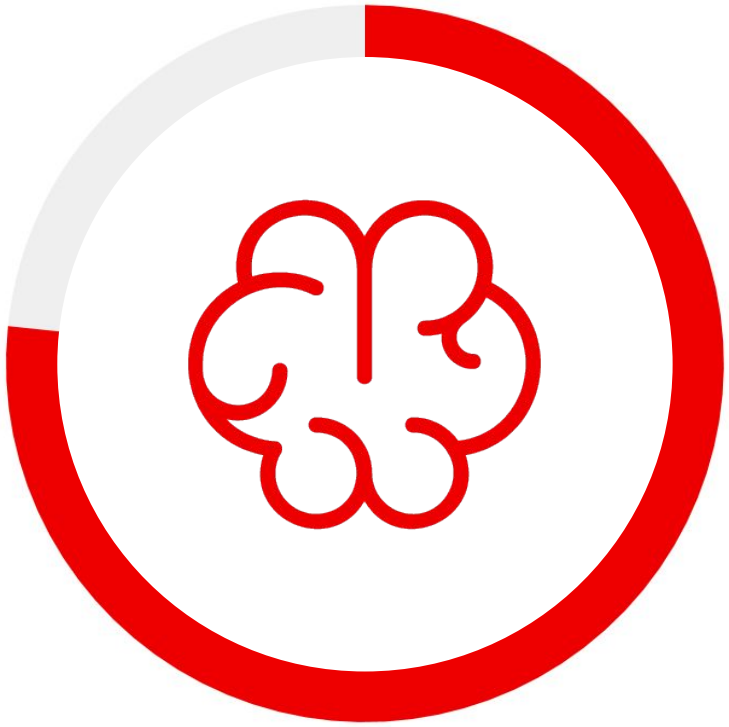
AI has the potential to deliver additional global economic activity of around $13 trillion by 2030.[2]

**51%**

of enterprises indicate that their current AI infrastructure will not be able to meet future demands.[1]



Source:
1. 451 Research, part of S&P Global Market Intelligence - Voice of the Enterprise: AI & Machine Learning, Infrastructure 2020
2. Notes from the AI frontier: Modeling the impact of AI on the world economy, McKinsey Global Institute, September 2018



# 69%

of enterprises use a mix of open source and cloud-based software to power AI initiatives.



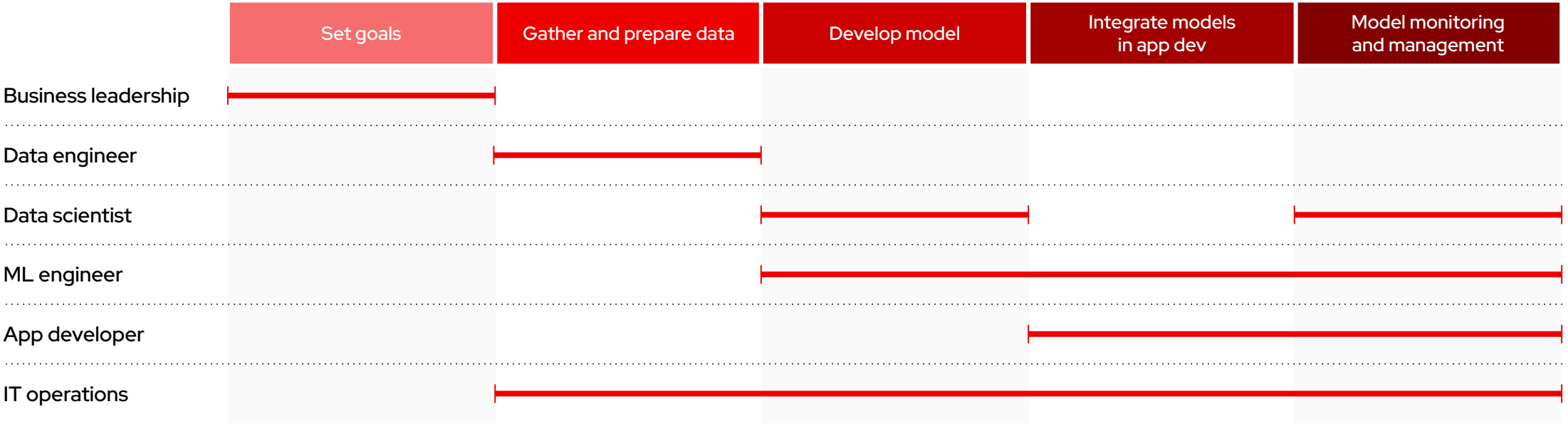
PULSE

Source:
Data collected from Dec. 16, 2020 - Jan. 22, 2021 | Respondents: 100 enterprise IT and data leaders

# Red Hat and AI

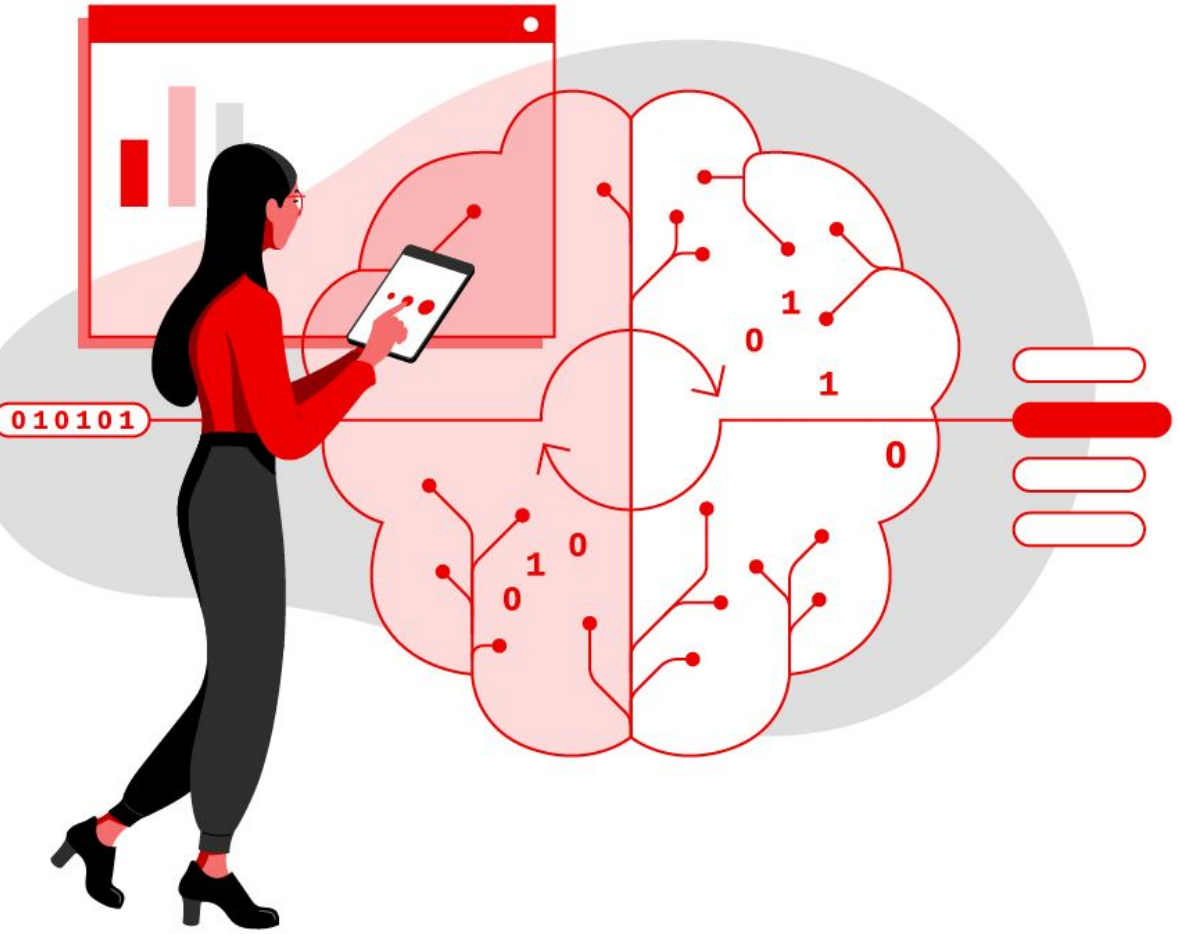
# Operationalizing AI/ML is not trivial

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






Red Hat




# Red Hat OpenShift AI

An AI-focused platform that provides tools to train, tune, serve, monitor and manage AI/ML experiments and models.

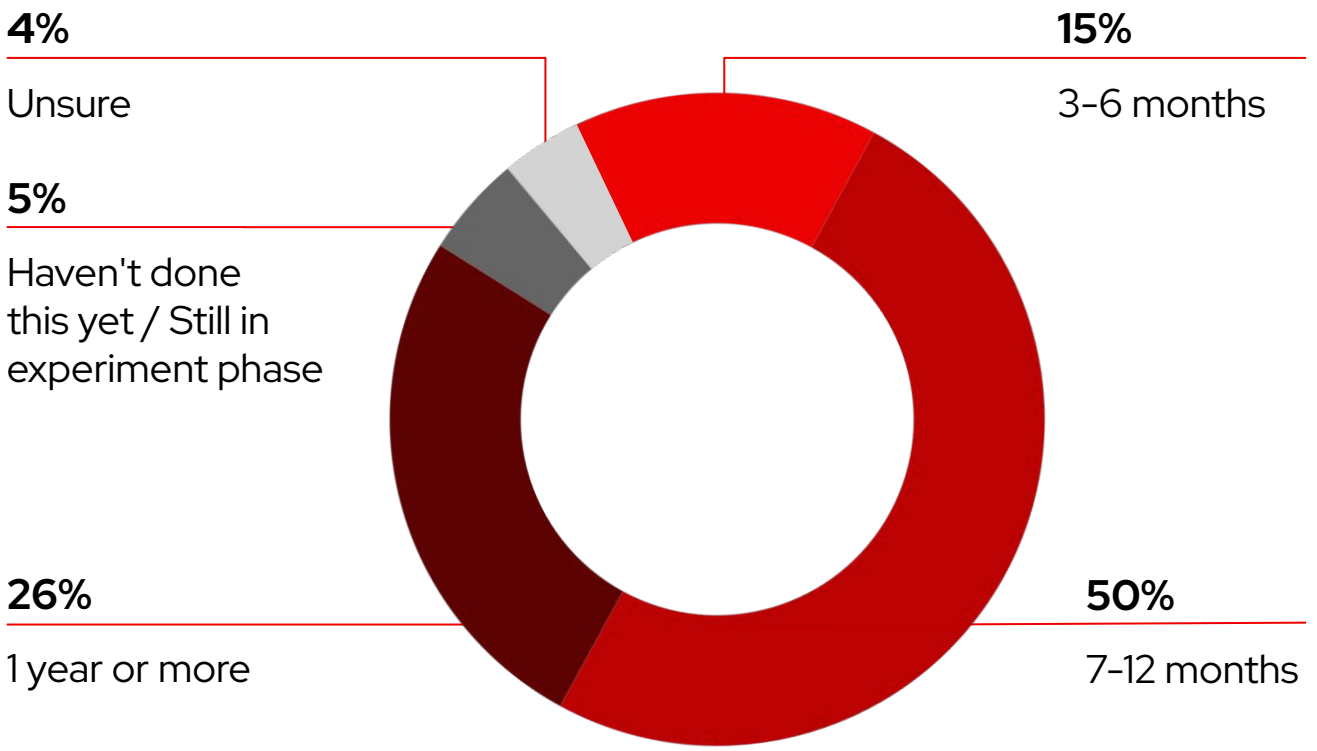


Red Hat

# Operationalizing AI is still a challenging process

## What is the average AI/ML timeline from idea to operationalizing the model?

Half of respondents (50%) say their average AI/ML timeline from idea to operationalizing the model is 7-12 months.






Source: Gartner Peer Insights, Open Source AI for Enterprise survey, 2023

# Red Hat strategy around generative AI and foundation models

- Developing the infrastructure stack for distributed workloads, scheduling for building, prompt-tuning, fine-tuning and serving foundation models
  - OpenShift AI is a foundation layer for IBM watsonx.ai and Ansible Lightspeed with IBM Watson Code Assistant
- Provide Granite models as part of RHEL AI
- Enable out-of-the-box “bring your own model” use cases to OpenShift AI
- Red Hat will infuse Generative AI capabilities into more of its portfolio.
  - OpenShift Lightspeed preview and RHEL Lightspeed vision announced at Red Hat Summit

# Red Hat's AI portfolio

**Trust** | **Choice** | **Consistency**

## Models

**RHEL AI**

Base Model | Alignment Tuning | Methodology & Tools | Platform Optimization & Acceleration

## AI platform

**OpenShift AI**

Development | Serving | Monitoring & Lifecycle | Resource Management

## AI enabled portfolio

**Lightspeed portfolio**

Usability & Adoption | Guidance | Virtual Assistant | Code Generation

## AI enabled apps

**App & Developer services**

Model Evaluation and Testing | App Connectivity | Secure Supply Chain

## Open Hybrid Cloud Platforms

**Red Hat Enterprise Linux | Red Hat OpenShift | Red Hat Ansible Platform**

Acceleration | Performance | Scale | Automation | Observability | Security

## Partner Ecosystem

Hardware | Software | Accelerators | Models | Delivery



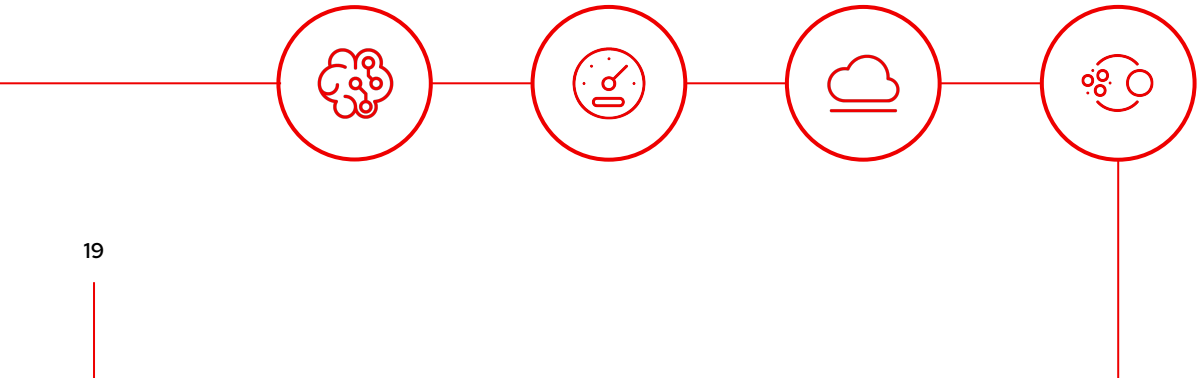
Red Hat

# AI for the open hybrid cloud

## Enterprise-grade open source hybrid AI and MLOps platform

**Red Hat OpenShift AI**

Develop, train, serve, monitor, and manage the life cycle of AI/ML models from experiments to production.



- Provide a unified platform for data scientists and intelligent application developers
- Scale to meet the workload demands of foundation models: data volume, training time, model size, acceleration, and scalability
- Deliver consistency, cloud-to-edge production deployment and monitoring capabilities
- Underlying platform for training, serving, and tuning foundation models in Red Hat Ansible Lightspeed with IBM Watson Code Assistant

# Red Hat's AI/ML engineering is 100% open source




Red Hat

# Red Hat OpenShift AI

## Integrated MLOps platform

Create and deliver GenAI and predictive models at scale across hybrid cloud environments.

Available as

- Fully managed cloud service
- Traditional software product on-site or in the cloud!

### Model development

Provides flexibility and composability by supporting multiple AI/ML libraries, frameworks, and runtimes.

### Model serving and monitoring

Deploy models across any OpenShift footprint and centrally monitor their performance.

### Lifecycle management

Expands DevOps practices to MLOps to manage the entire AI/ML lifecycle.

### Increased capabilities / collaboration

Create projects and share them across teams. Combine Red Hat components, open source software, and ISV certified software.

# ... and integrating our partner ecosystem


| Gather and prepare data | Develop model | Integrate models in app dev | Model monitoring and management |
|---|---|---|---|

**Customer managed ISV software**

Customer managed applications

elastic, Starburst, Pachyderm, mongoDB, crunchydata, Intel AI Analytics Toolkit, CLOUDERA, H2O.ai, redislabs, SAS, watsonx, C3.ai, OpenVINO, CognitiveScale THE TRUSTED AI COMPANY, avanseus, cnvrg.io

**ISV managed cloud services**

Starburst Galaxy, ANACONDA

**Red Hat software and cloud services**

Red Hat AMQ — Red Hat OpenShift AI

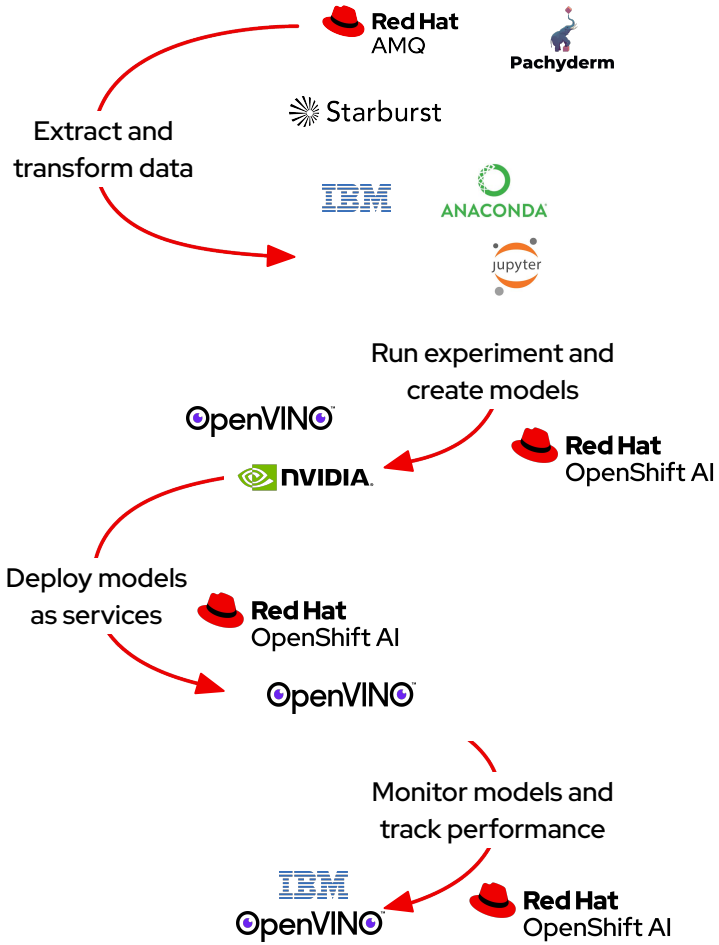
**Red Hat managed cloud platform**

| | | | |
|---|---|---|---|
| Open hybrid cloud platform with self service capabilities | Red Hat OpenShift | Red Hat OpenShift Dedicated | Red Hat OpenShift Service on Amazon Web Services |
| Accelerators | NVIDIA | intel | |
| Cloud infrastructure | aws | Google Cloud | Azure |



Red Hat

# ... and the model operationalization life cycle




**Starburst Enterprise**
Unlock the value of your data by making it fast and easy to access data across hybrid cloud.

**Pachyderm**
Brings data versioning and governance to your most precious asset

**Anaconda Professional**
Curated access to an extensive set of data science packages to be used in your Jupyter projects.

**IBM Watsonx.ai**
Build, run, and manage generative AI models at scale.

**NVIDIA**
GPU-enabled hardware makes it easier for customers to stand up resource-intensive environments and accelerate their data science experiments.

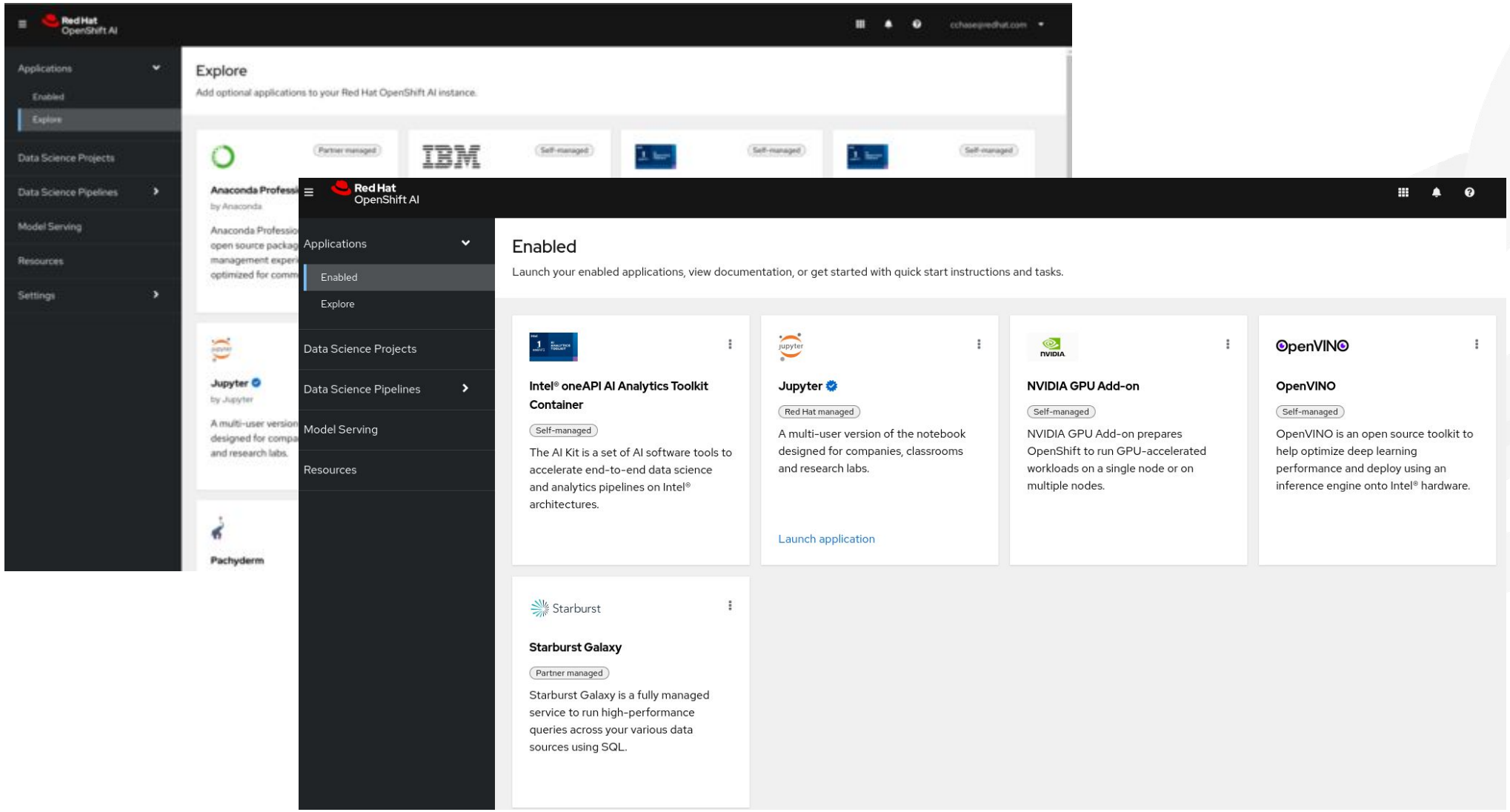
**Intel OpenVINO Notebook Images**
Toolkit of pre-trained models optimized for intel processors and GPUs

**Intel OpenVINO Model Server**
Scalable, High-Performance serving engine

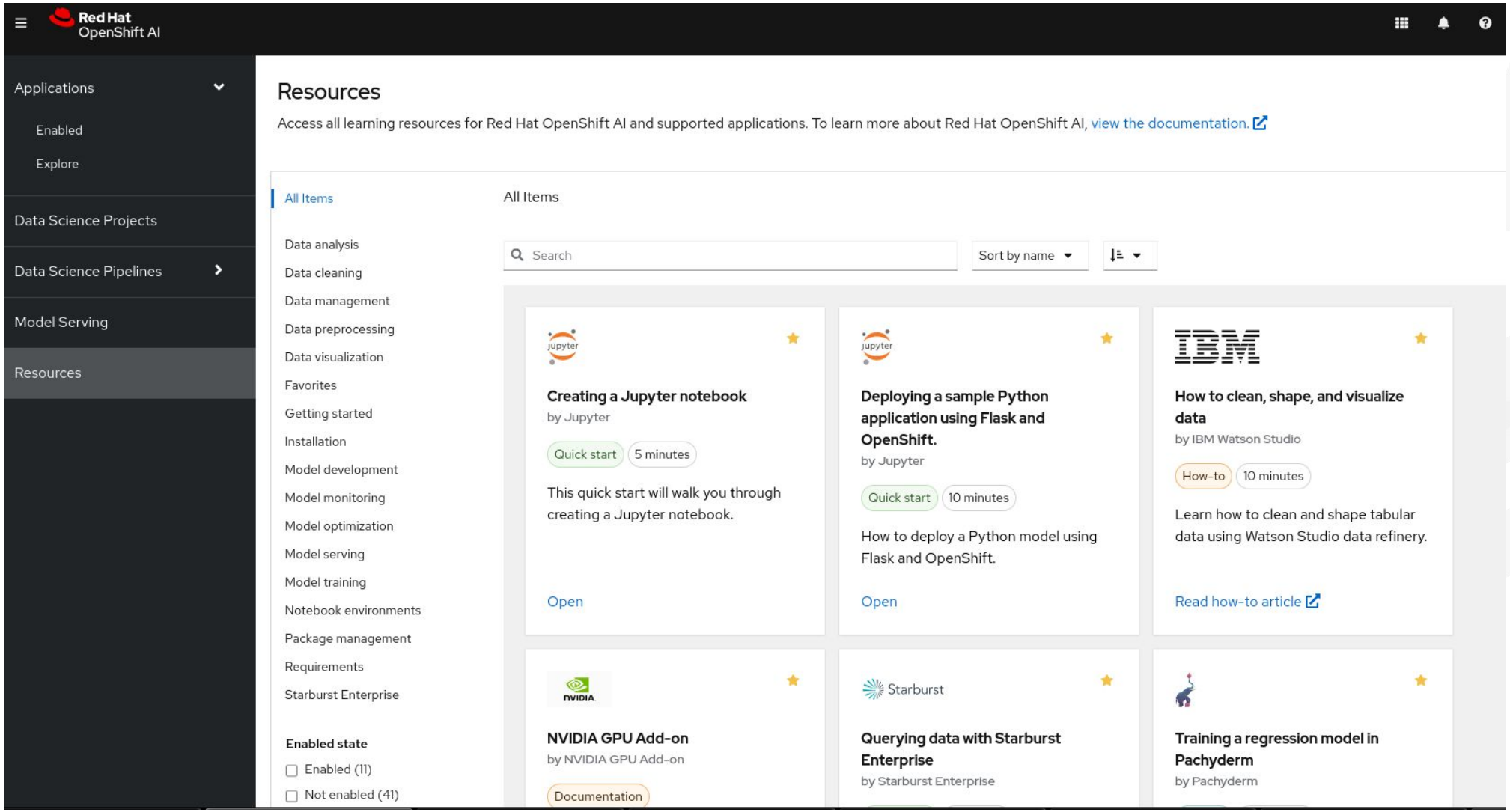


Red Hat

# Dashboard user interface



# Dashboard resources






Red Hat

# What differentiates us

## Hybrid cloud

Deploy models in containerized format for intelligent apps on-premise or in cloud

## Easy to manage

Simple configurations on a secure and proven platform, that you can scale up or down with low effort

## Collaborate

Collaborate on a common, extensible platform to bring IT, data science and application development teams together

## Open Source

Red Hat tracks changes and fixes to open source AI/ML tooling and enables customer access to upstream innovation

## U.S. Department of Veterans Affairs

*Suicide has no single cause, and no single strategy can end this complex problem. That's why Mission Daybreak is fostering solutions across a broad spectrum of focus areas.*

*A diversity of solutions will only be possible if a diversity of solvers answer the call to collaborate and share their expertise.*

# Red Hat, Team Guidehouse named winner in Mission Daybreak challenge to reduce Veteran suicides

### Challenge

Develop new data-driven means of identifying Veterans at risk for suicide.

### Solution

Red Hat teamed with global consulting services provider Guidehouse and Philip Held, Ph.D. of Rush University Medical Center, to develop a new data-driven means of identifying Veterans at risk for suicide running on Red Hat OpenShift, leveraging Red Hat OpenShift API Management and Red Hat OpenShift AI.

### Results

- Named a winner in the Mission Daybreak challenge, Phase 2, of the U.S. Department of Veterans Affairs' (VA) Mission Daybreak Grand Challenge in support of cutting-edge suicide prevention solutions
- Moved forward with a solution for the VA's efforts to reduce Veteran suicides
- Showcased the repeatability and scalability of open source-enabled solutions



Red Hat

Source: Red Hat blog. Red Hat, Team Guidehouse named winner in Mission Daybreak challenge to reduce Veteran suicides, Mar. 2023.

BOSTON UNIVERSITY

- **Implemented interactive lecture and lab environment** for computer scientists and engineers based on Red Hat OpenShift AI
- **Currently over 300 users** including over 100 concurrent
- **Integrates with the Boston University online textbook material,** also authored using the Red Hat OpenShift AI
- **Fast time to solution:** cloud services environment enabled BU to configure and deploy in December for classes that started in January
- **Lowers cost:** auto-scales based on demand; enables bursty interactive use cases at optimized cost



Red Hat

# 2024 AI Awards for OpenShift AI





2024 Finalist - Artificial Intelligence Excellence Awards

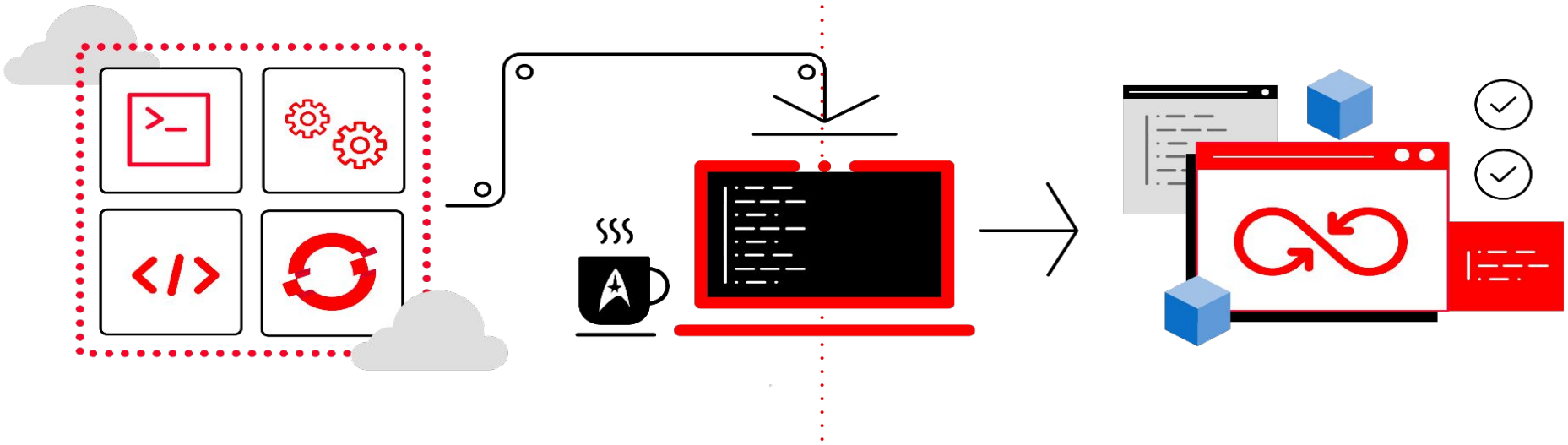


Red Hat

# Red Hat Consulting Services for your AI/ML Journey

Red Hat OpenShift AI Platform Engagement

MLOps Foundation



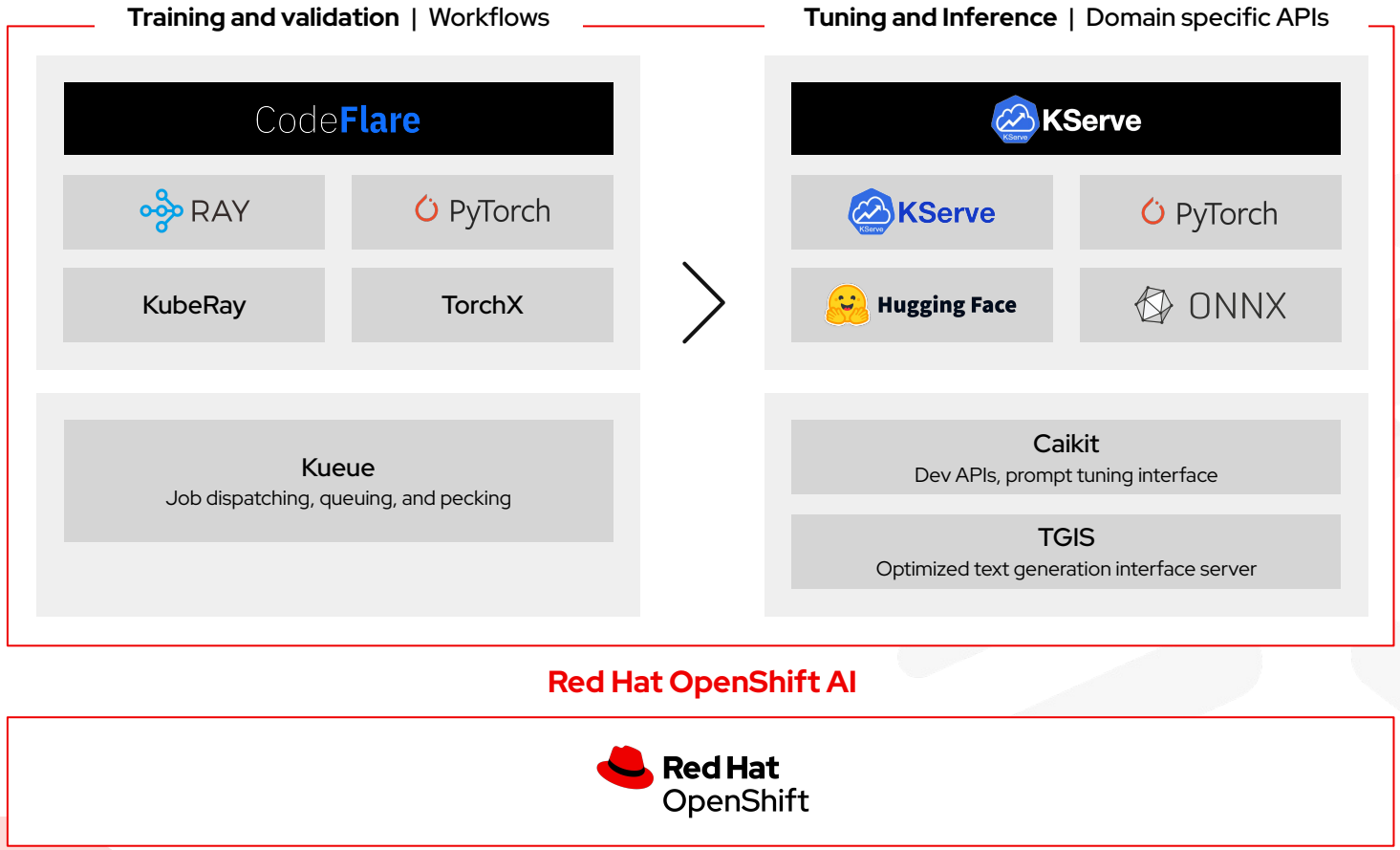


Red Hat

# RHOAI Functionality

# An open source platform for foundation models

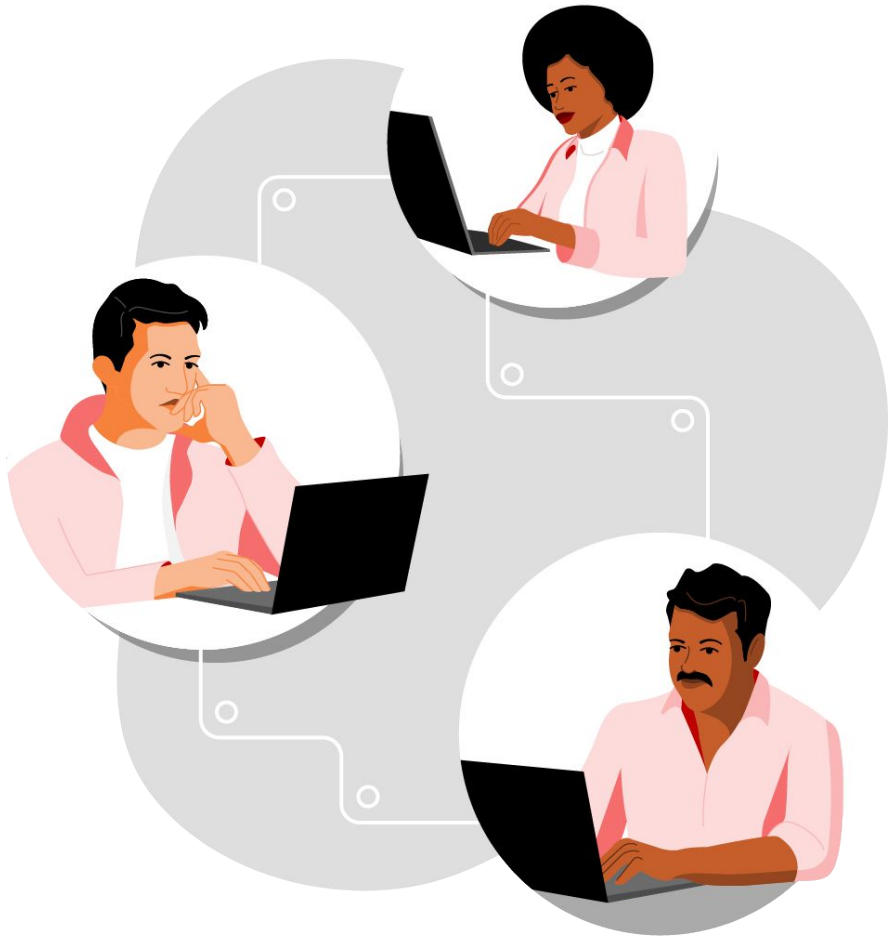
## Train or fine tune conversational and generative AI






Red Hat

# Model training highlights



**Support a variety of use cases**
including generative AI by accelerating and managing model training and tuning workloads

**Improve performance and scalability**
with distributed training

**Initiate and manage batch training**
in single- or multi-cluster environments with an easy-to-use interface

**Meet scale and performance needs**
by selecting from a range of accelerators

**Automate foundation model pipelines**

# Distribute workloads to enhance efficiency

## Focus on modeling, not infrastructure

by dynamically allocating computing power

## Prioritize and distribute job execution

using advanced queuing for tasks like large-scale data analyses

## Automate setup and deployment

so you can get up and running with minimal effort

## Manage resources and submit jobs

using a Python-friendly SDK, which is a natural fit for data scientists

## Streamline data science workflows

with seamless integration into the OpenShift AI ecosystem

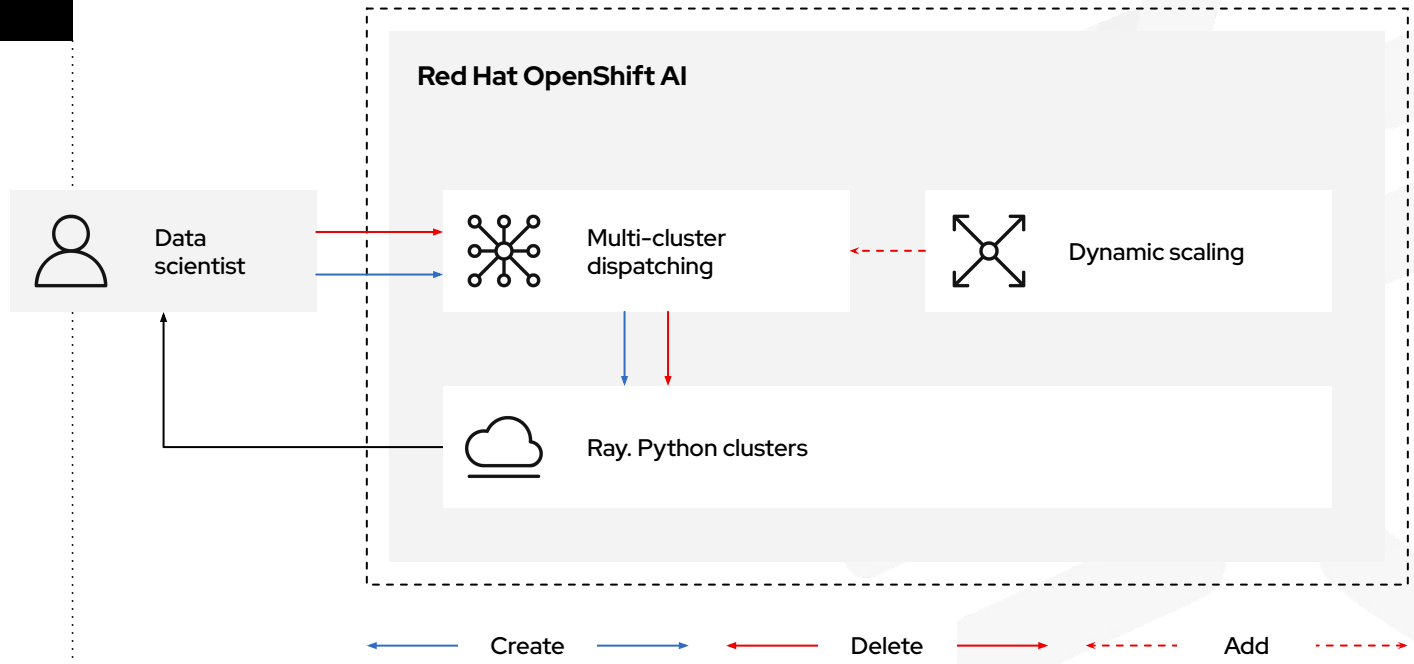


Red Hat

# Configure distributed workload clusters more easily

## Process

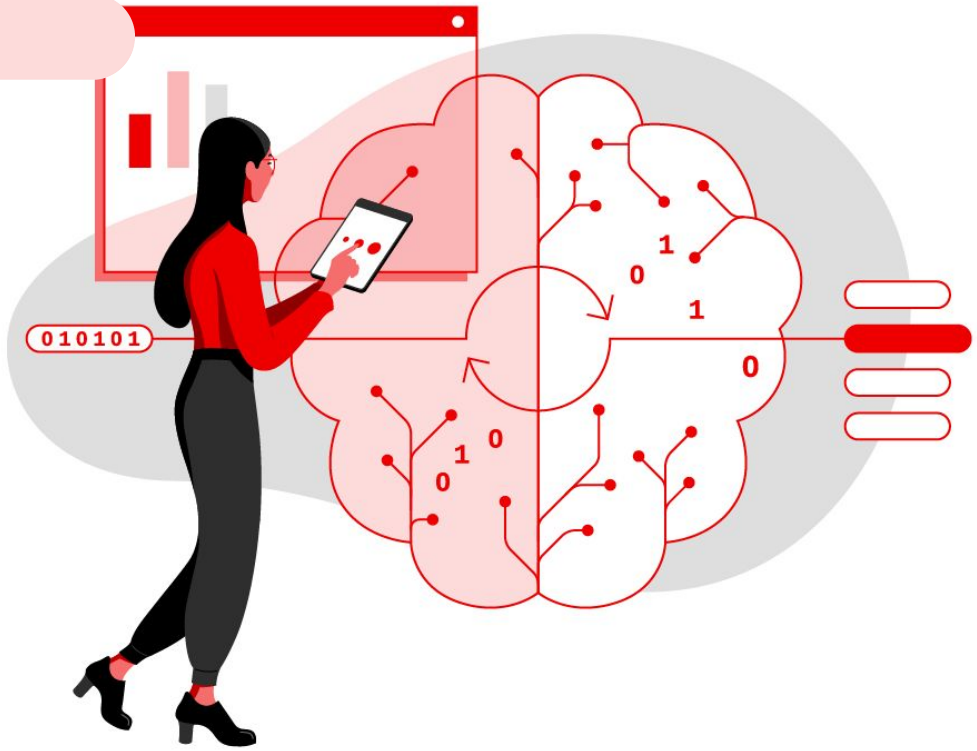
1. **Send request** for creation of cluster
2. **Evaluate clusters** for aggregated resources and dispatch when available
3. **Watch pending clusters** to provide aggregated resources and guarantee workload execution
4. **Develop and submit jobs** to retrieve statuses and logs from the clusters
5. **Send request** for deletion of clusters






Red Hat

# Make model serving more flexible and transparent




- **Use model-serving user interface (UI)**
  integrated within product dashboard and projects workspace
- **Serve open source models**
  from providers like Hugging Face
- **Support a variety of model frameworks**
  including TensorFlow, PyTorch, and ONNX
- **Choose inference servers**
  either out-of-the-box options optimized for foundation models or your own custom inference server
- **Scale cluster resources**
  up or down as your workload requires

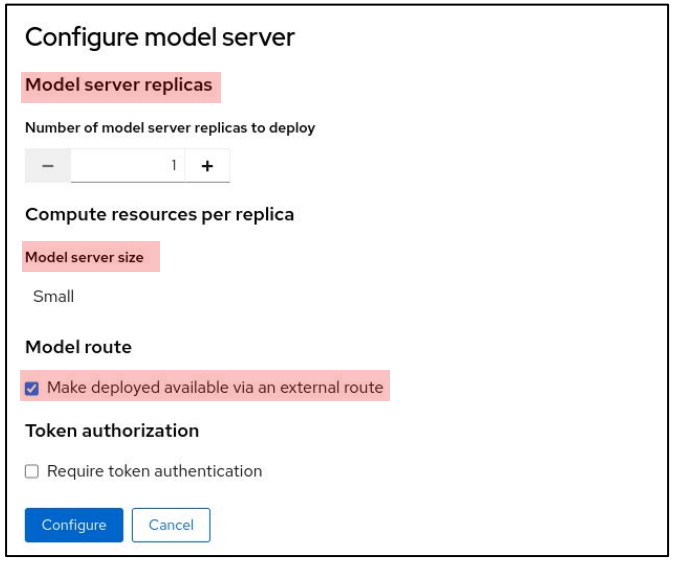


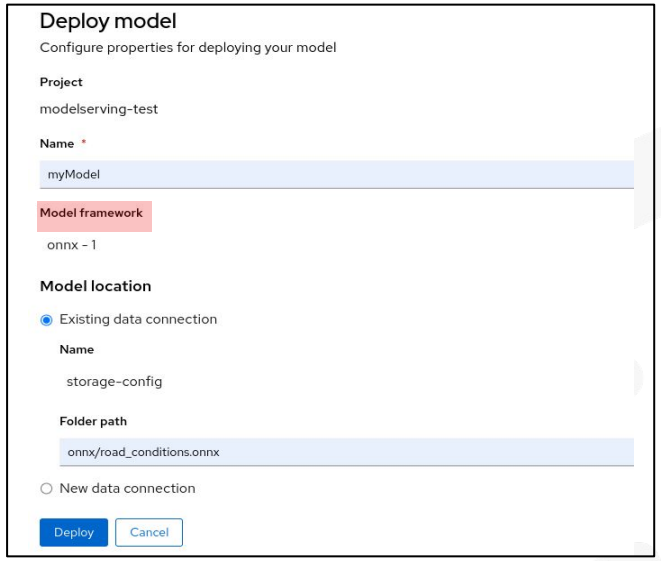
Red Hat

# Serve, scale, and monitor your models

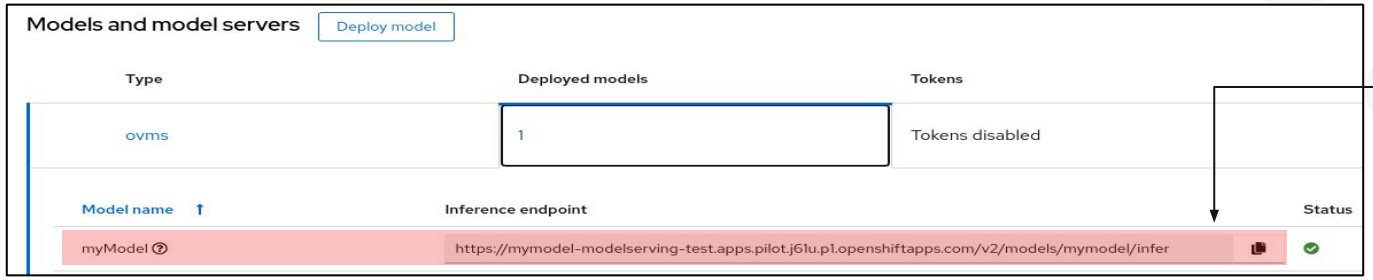
Select the required resources and scale model serving as needed

Make your model public and secure




Select your model framework




View your deployed model fleet endpoints



Red Hat

# Model performance metrics

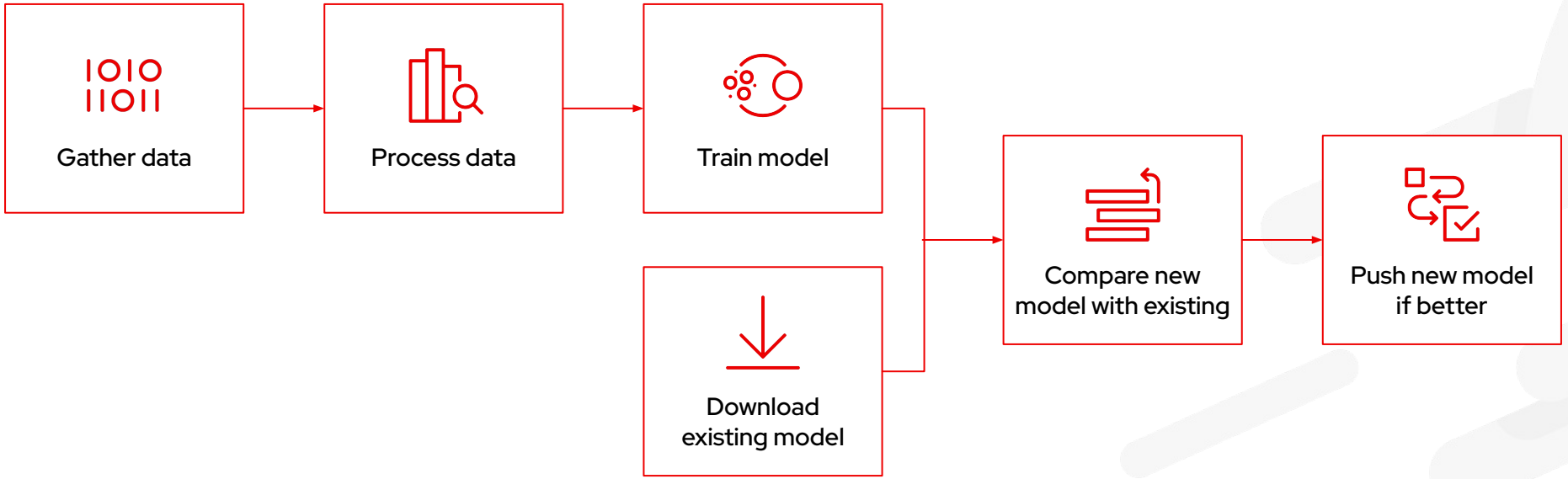
Access a range of model performance metrics to build your own visualizations or integrate data with other observability services


- Out-of-the-box visualizations for performance and operations metrics



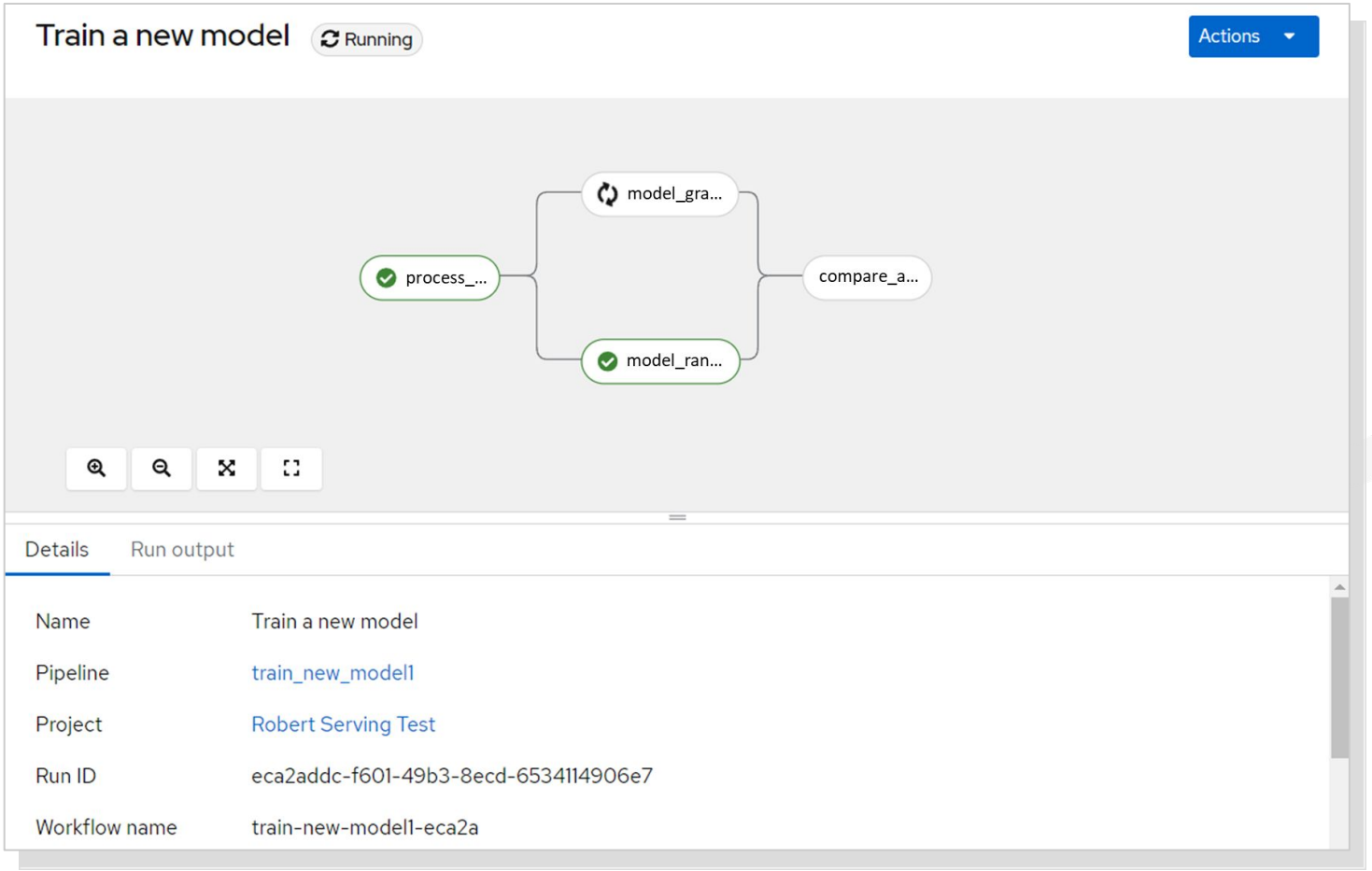
Red Hat

# Data science pipelines component




- Continuously deliver and test models in production
- Schedule, track, and manage pipeline runs
- Easily build pipelines using graphical front end
- Orchestrate data science tasks into pipelines
- Chain together processes like data prep, build models, and serve models

# Red Hat OpenShift data science pipelines user interface



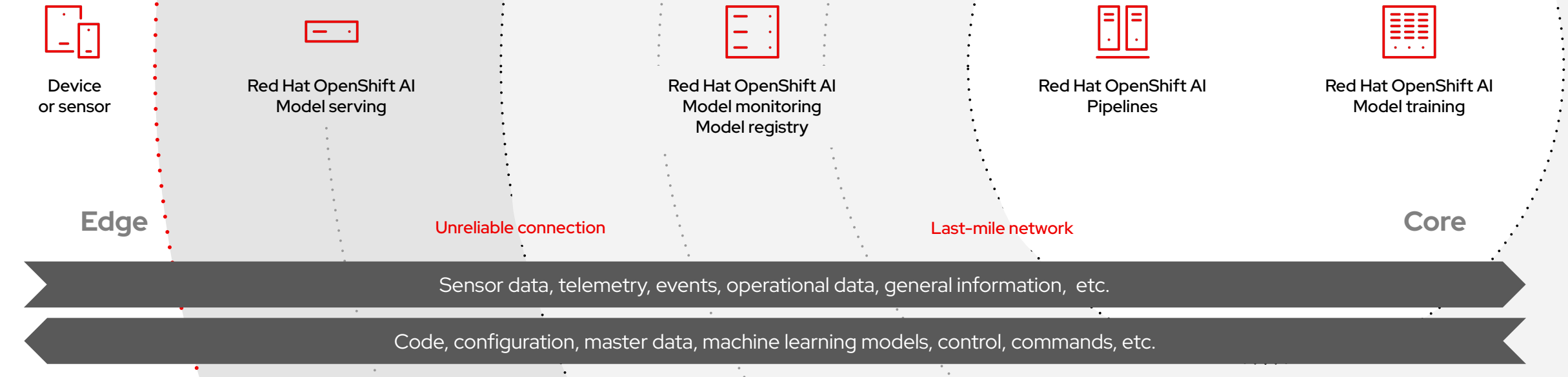

The OpenShift AI user interface enables you to track and manage pipelines and pipeline runs.



Red Hat

Edge AI

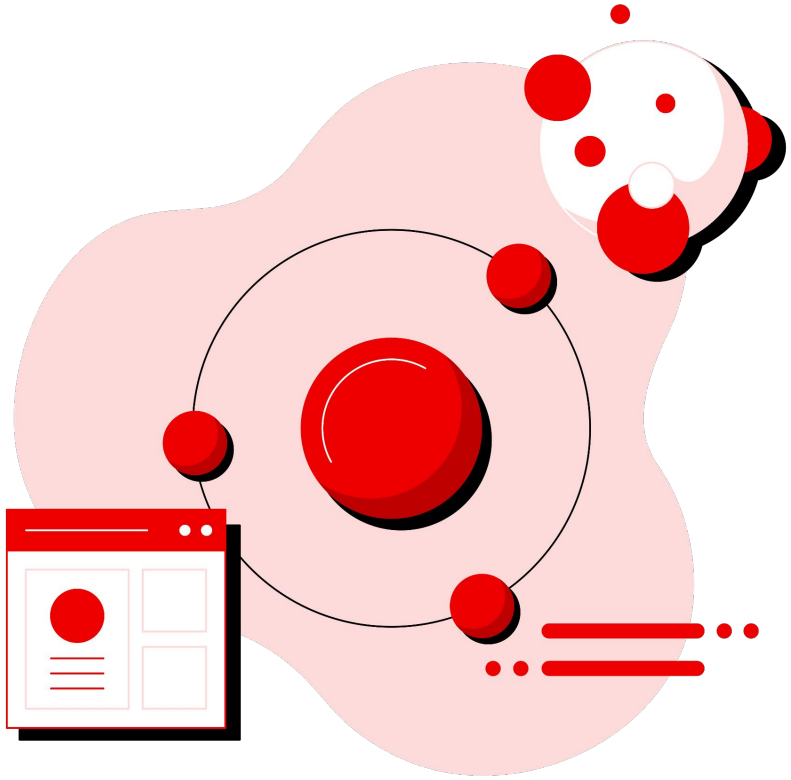
# Flexibility at the edge






Red Hat

# Red Hat OpenShift AI at the edge



## Consistently deploy and manage intelligent applications

- **Deploy** centrally to the near edge using GitOps approach
- **Monitor** operations using centralized Grafana dashboard
- **Provide** data scientists with actionable insights
- **Automate** deployment throughout stages with repeatable MLOps pipelines

# Timeline

## 1H '24

### Next

- **MLOps**
  - Enhance OOTB model monitoring - perf & ops metrics
  - Enhance support for LLM serving
  - Model deployments to near-edge locations (tech preview)
- **Model development**
  - Distributed workloads GA
  - VS Code OOTB support (Tech Preview)
  - Update OOTB workbench images
  - Data Science Projects UX enhancements
- **Platform / integrations**
  - Redesigned home/starting page
  - Vector DB partner solution

## 2H'24

### Future

- **MLOps**
  - Data Science Pipelines v2 incl. experiment tracking
  - Model registry (GA)
  - Enhance OOTB model monitoring
  - Model deployments to edge locations (GA)
- **Model development**
  - Local IDE plugins (eg. VS Code)
  - Feature store (Tech Preview)
  - Enhanced foundation model tuning capabilities
- **Platform/integrations**
  - Support AMD GPUs
  - Run:ai integration
  - Expand admin UI config capabilities
  - Fractional GPUs for training & inference
  - Continue to expand accelerator support

# Red Hat OpenShift AI

Learn more ▸

Try it ▸



Red Hat

# Lab Time!
# Password: rhoai

# Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.


linkedin.com/company/red-hat


youtube.com/user/RedHatVideos


facebook.com/redhatinc


twitter.com/RedHat



Red Hat

# Red Hat Enterprise Linux AI

## Foundation Model Platform

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

### 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.

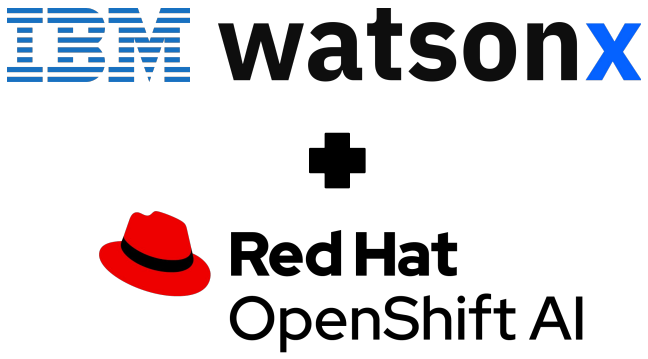
### Optimized bootable model runtime instances

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.

# Integration with IBM watsonx.ai




Red Hat OpenShift AI and IBM watsonx provide an

**AI-focused portfolio for the open hybrid cloud**

that

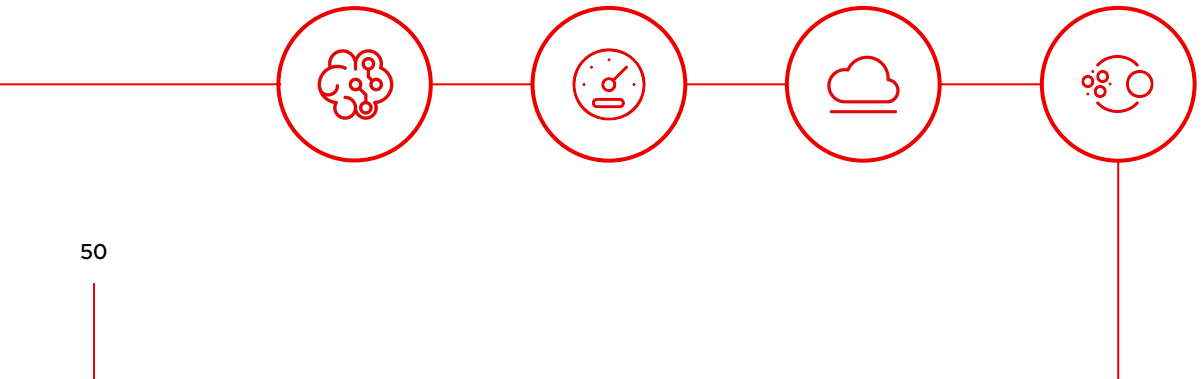
- accelerates the adoption of generative AI
- simplifies the process of managing the AI lifecycle
- reduces the complexities of incorporating AI into the business

# AI for the open hybrid cloud

## Enterprise-grade open source hybrid AI and MLOps platform

**Red Hat OpenShift AI**

Train, serve, monitor, and manage the life cycle of AI/ML models and applications, from experiments to production.



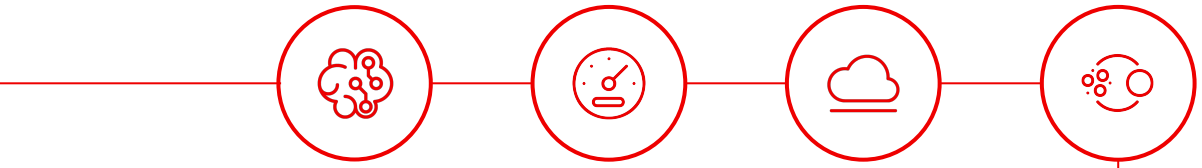
- Provide a unified platform for data scientists and intelligent application developers
- Scale to meet the workload demands of foundation models: data volume, training time, model size, acceleration, and scalability
- Deliver consistency, cloud-to-edge production deployment and monitoring capabilities

# Extend OpenShift AI with watsonx.ai

## Innovative toolset studio to train, validate, tune and deploy Gen AI

**watsonx.ai**

Accelerate generative AI adoption with curated IBM and open source foundation models



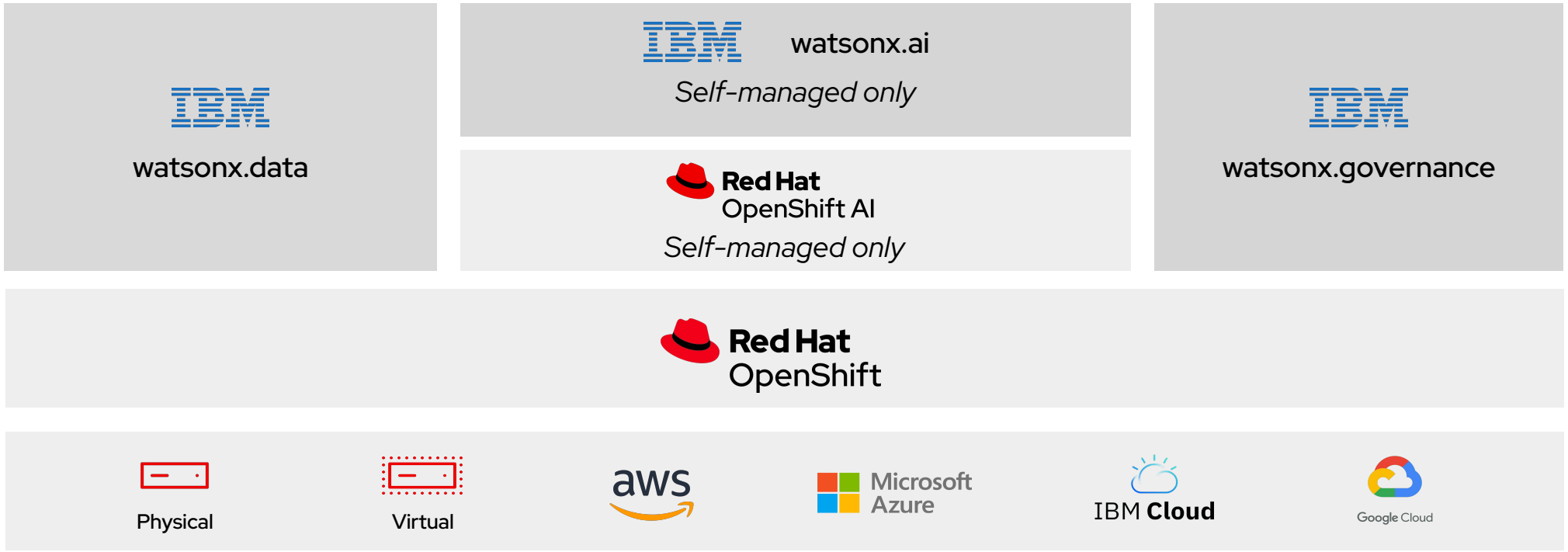
- **Reduce model discovery time and risk**: Quickly begin with confidence using IBM's foundation or curated open-source models.
- **Improve tuning inputs and results:** Prompt Lab provides different AI-builders an intuitive experience for building effective prompts for use in tuning foundation models.
- **Validate tuned models and iterate quickly:** Quickly identify areas for improvement and iterate quickly to achieve better results before going into production.



Red Hat

# IBM watsonx and Red Hat OpenShift AI

## High-performing, cloud-native AI open source stack runs on Red Hat OpenShift AI

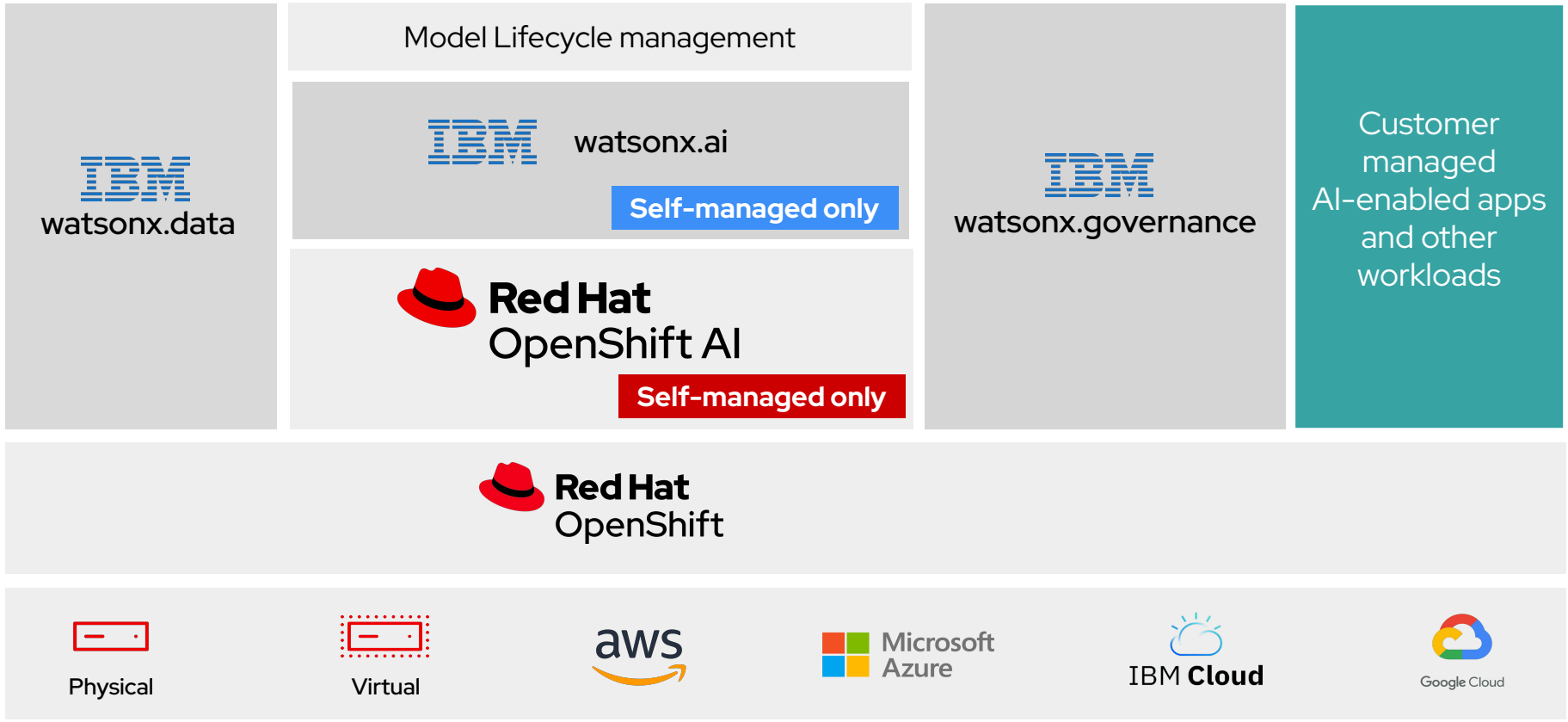





Red Hat

# Red Hat OpenShift AI and IBM watsonx

## High-performing, cloud-native AI open source stack runs on Red Hat OpenShift AI






Red Hat

# What is IBM watsonx?

AI and data platform designed to help scale and accelerate the impact of AI with trusted data across the business


IBM watsonx provides access to curated open-source foundation models, visual tools for machine learning, data processing and storage solutions, and governance toolkit for the AI lifecycle.

# IBM watsonx

AI and data platform that brings together three core components to enhance AI capabilities and data management

### watsonx.data

A fit-for-purpose data repository built on an open data lakehouse architecture to efficiently store, organize, and access data.

### watsonx.ai

A studio toolset to train, validate, tune and deploy generative AI, foundation models and machine learning capabilities. Includes the foundation models.

### watsonx.governance

A toolkit to accelerate AI workflows that are built with responsibility, transparency and explainability. It helps manage the risk and support regulatory compliance.

The portfolio's three main components run on Red Hat OpenShift

# Combine the power of watsonx.ai with Red Hat OpenShift AI

## Get access

Access underlying APIs powering interface and tuning (calkit and tgis)

## Extend the platform

Deploy custom AI software stacks to Red Hat Openshift

# Self-managed vs. managed cloud service

# Comparison of Red Hat OpenShift AI offerings

| | Red Hat OpenShift AI cloud service | Red Hat OpenShift AI self managed |
|---|---|---|
| **Initial Red Hat OpenShift platform deployment** | Customer responsibility | Customer responsibility |
| **Initial Red Hat OpenShift AI deployment** | Customer responsibility via Add-on | Customer responsibility |
| **Red Hat OpenShift and Red Hat OpenShift AI maintenance** | **Red Hat (included with service)** | **Customer responsibility** |
| **Red Hat OpenShift AI Features** | No limitations | No limitations |
| **Integration with Red Hat and certified ISV products and services** | No limitations | No limitations |
| **Customer support** | **Premium support** | **Standard or premium support options** |
| **Uptime SLA** | **Red Hat OpenShift (99.95%)<br>Red Hat OpenShift AI (99.95%)** | **Customer responsibility** |
| **Pricing** | **Yearly and consumption-based options** | **Yearly core-based and bare metal options** |
| **Platforms** | **Amazon Web Services, Google Cloud Platform** | **Any platform where OpenShift can be deployed (eg. Amazon Web Services, Microsoft Azure, Google, on-prem bare metal, on-prem virtual)** |
| **Deploy custom apps** | Customer responsibility | Customer responsibility |



Red Hat

# Appendix: Trials

# Getting started with Red Hat OpenShift AI

## Free

### Sandbox environments

- Developer portal
  red.ht/rhods_sandbox
- Support for OpenShift AI only (Limited access to ISV software)
- Single-user environment
- Small cluster size only
- 30-day maximum
- No cost for infrastructure, OpenShift, OpenShift AI

## $

### Trial environments

- 60 day trial (self-managed w/ OpenShift Container Platform included as option)
  red.ht/rhods_60_trial
- Support for OpenShift AI with ISV software trials available from partners
- Multi-user environment
- Full range of cluster sizes supported
- If interested in managed cloud service option, check out Level Up ROSA program

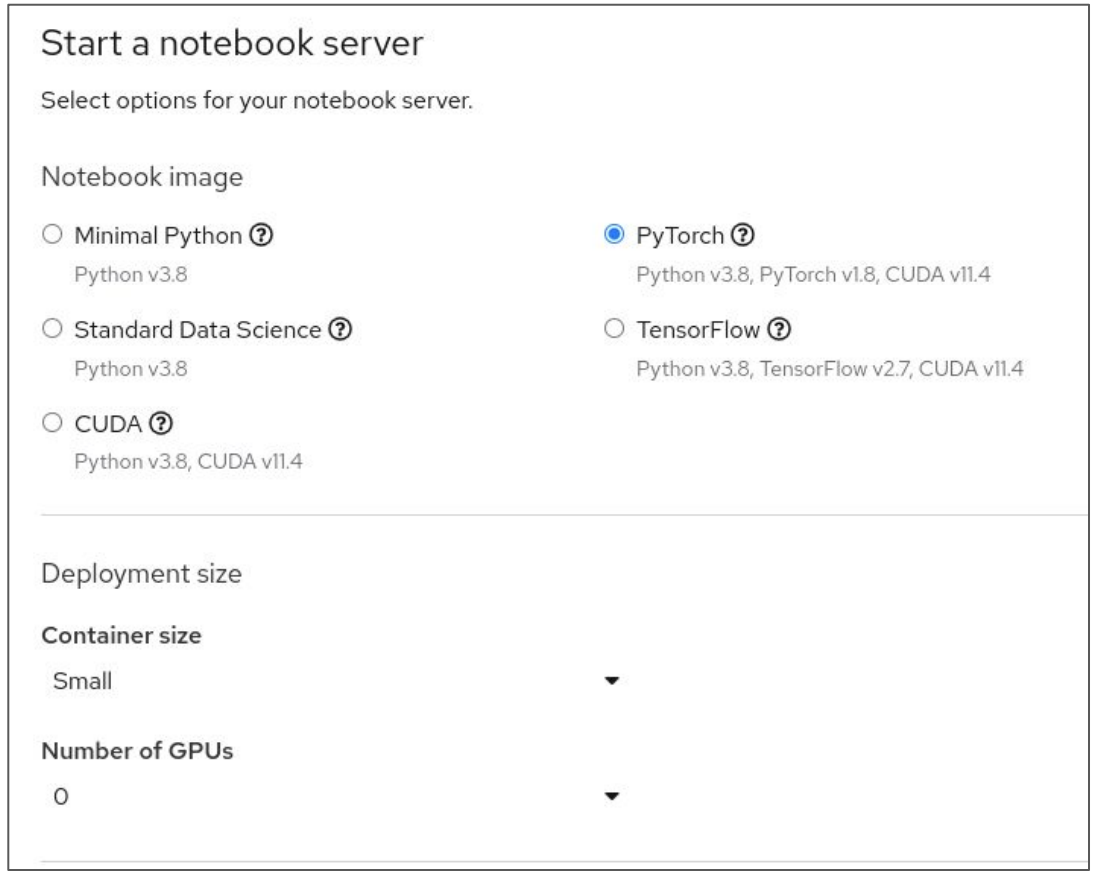
## $$

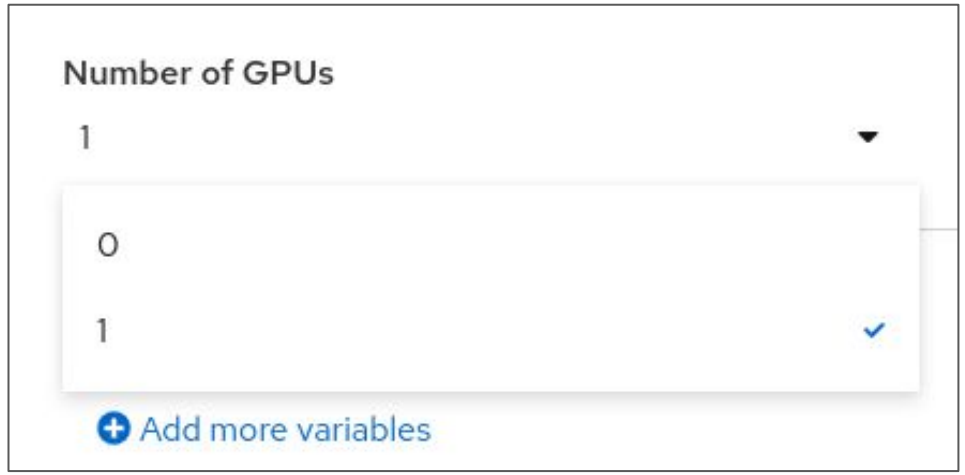
### Paid environments

- Support for OpenShift AI with ISV software available from partners
- Multi-user environment sized based on customer needs
- Annual and On-demand Hourly Costs per vCPU (managed cloud) and Yearly (managed and self managed)

# Appendix: GPU enablement

# Leveraging GPUs in Red Hat OpenShift AI







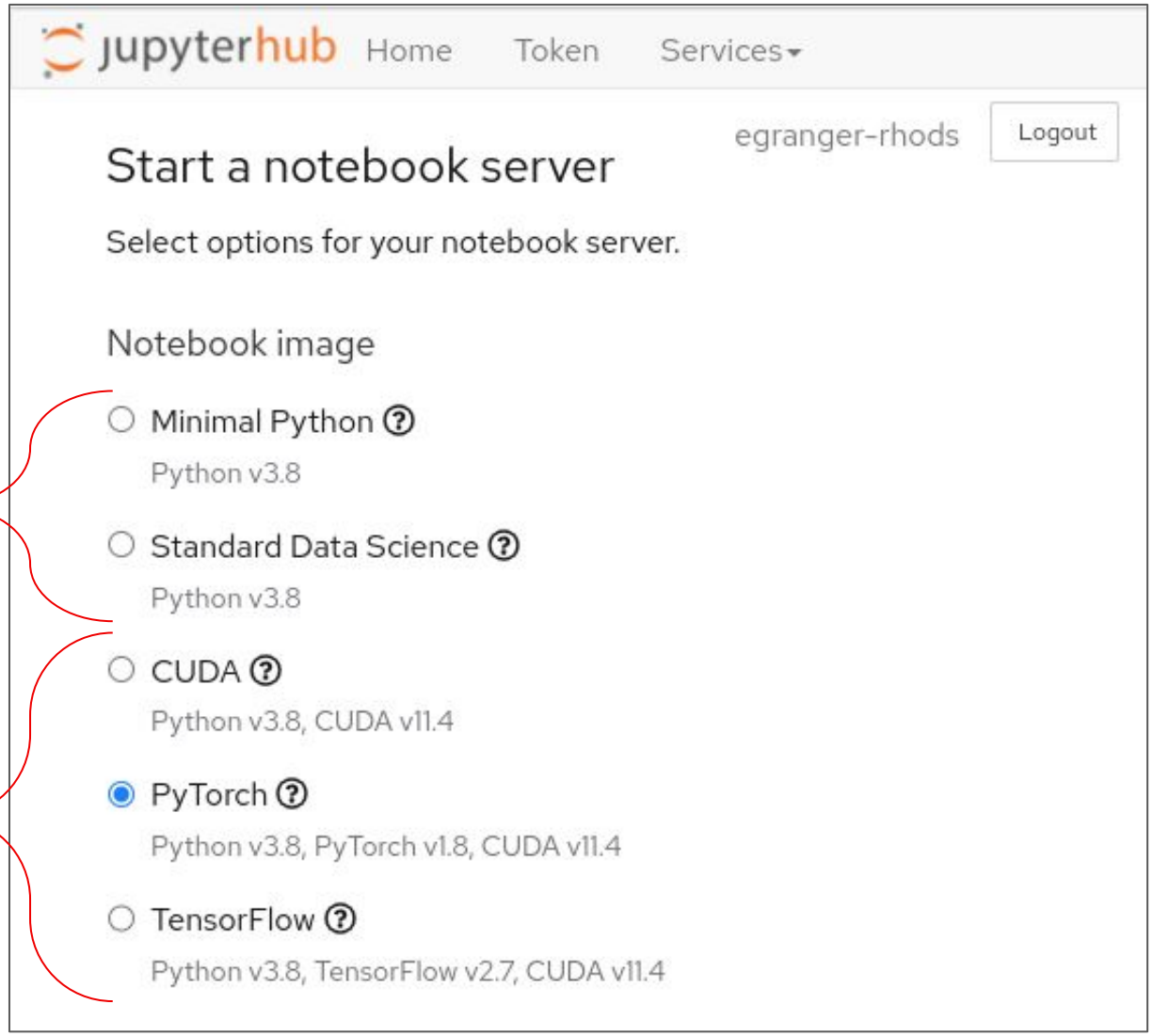
Red Hat

# Leveraging GPUs in OpenShift AI

Notebook Images are built to support and leverage GPUs

CPU only

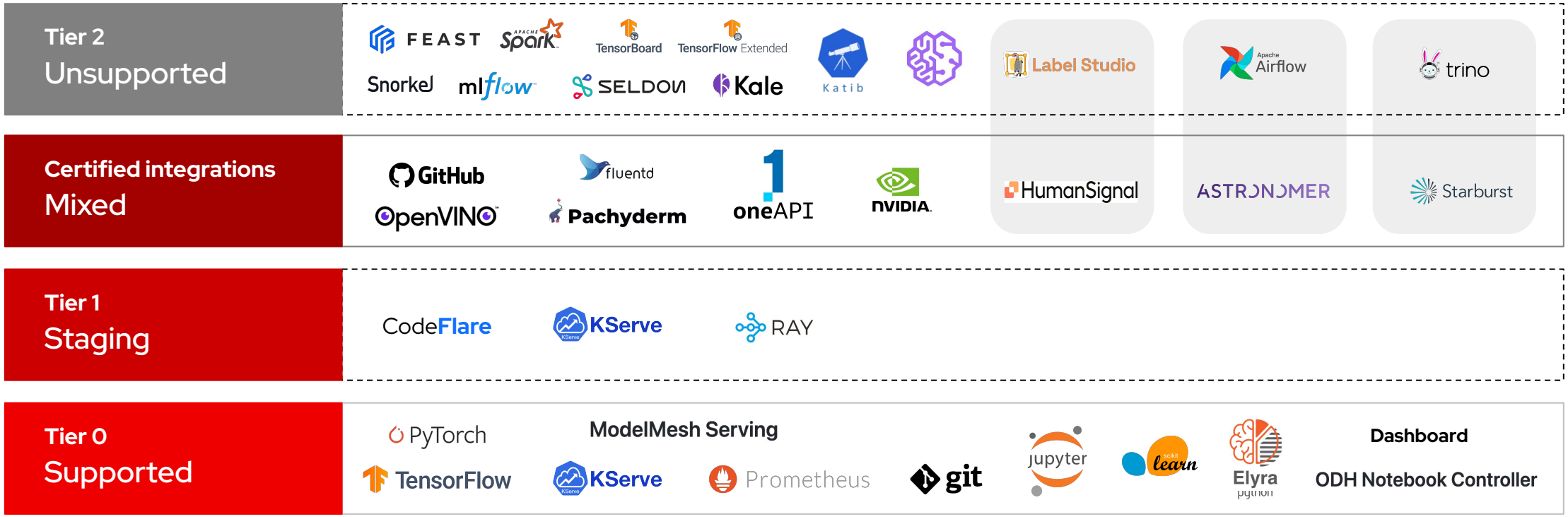
CPU & GPU



# Appendix: Open Data Hub ecosystem support levels

# Open Data Hub to OpenShift AI technology

## Ecosystem support levels






Red Hat