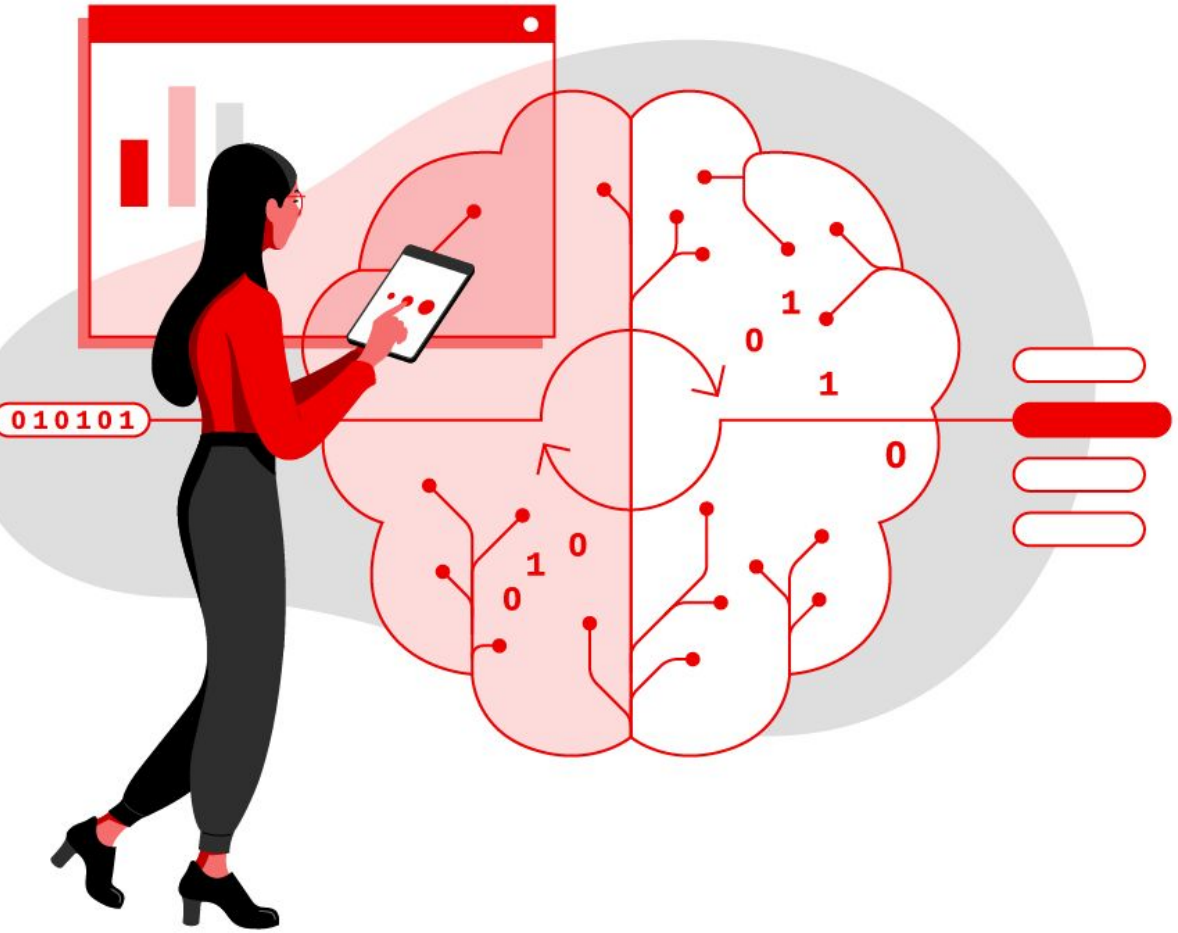
Overview of Red Hat OpenShift AI

MLOps platform for artificial intelligence/ machine learning (AI/ML) use cases

Steven Huels

Sr Director, AI Business Unit

Red Hat



Red Hat OpenShift AI

An AI-focused platform that provides tools across the full lifecycle of AI/ML experiments and models.

Red Hat

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
- Partner with model builders to offer models with OpenShift AI
- Enable out-of-the-box “bring your own model” use cases
- OpenShift AI is a foundation layer for IBM watsonx.ai and Ansible Lightspeed with IBM Watson Code Assistant
- Infuse generative AI capabilities across the Red Hat portfolio as we did with Ansible Lightspeed

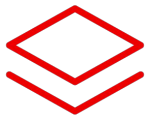
Red Hat

Our AI/ML strategy

AI workload support

Support **AI workload requirements** on Red Hat platforms

e.g., hardware acceleration, GPU Operator



Platform for AI-enabled apps

Provide a consistent, hybrid cloud **application platform for customers** to build, train, and deploy AI-enabled applications

e.g., Red Hat OpenShift AI

AI-enabled platforms

Use **AI models, tools, and services to accelerate adoption** of existing Red Hat products and services

e.g., Red Hat Ansible Lightspeed, Red Hat Developer Hub

Red Hat

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

Red Hat OpenShift AI

Hybrid MLOps platform

Collaborate within a common platform to bring IT, data science, and app dev teams together

Available as

- **managed cloud service**
- **traditional software product on-site or in the cloud!**

Model development

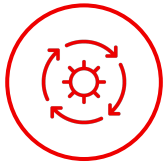
Conduct exploratory data science in JupyterLab with access to core AI / ML libraries and frameworks including TensorFlow and PyTorch using our notebook images or your own.

Model serving & monitoring

Deploy models across any cloud, fully managed, and self-managed OpenShift footprint and centrally monitor their performance.

Lifecycle Management

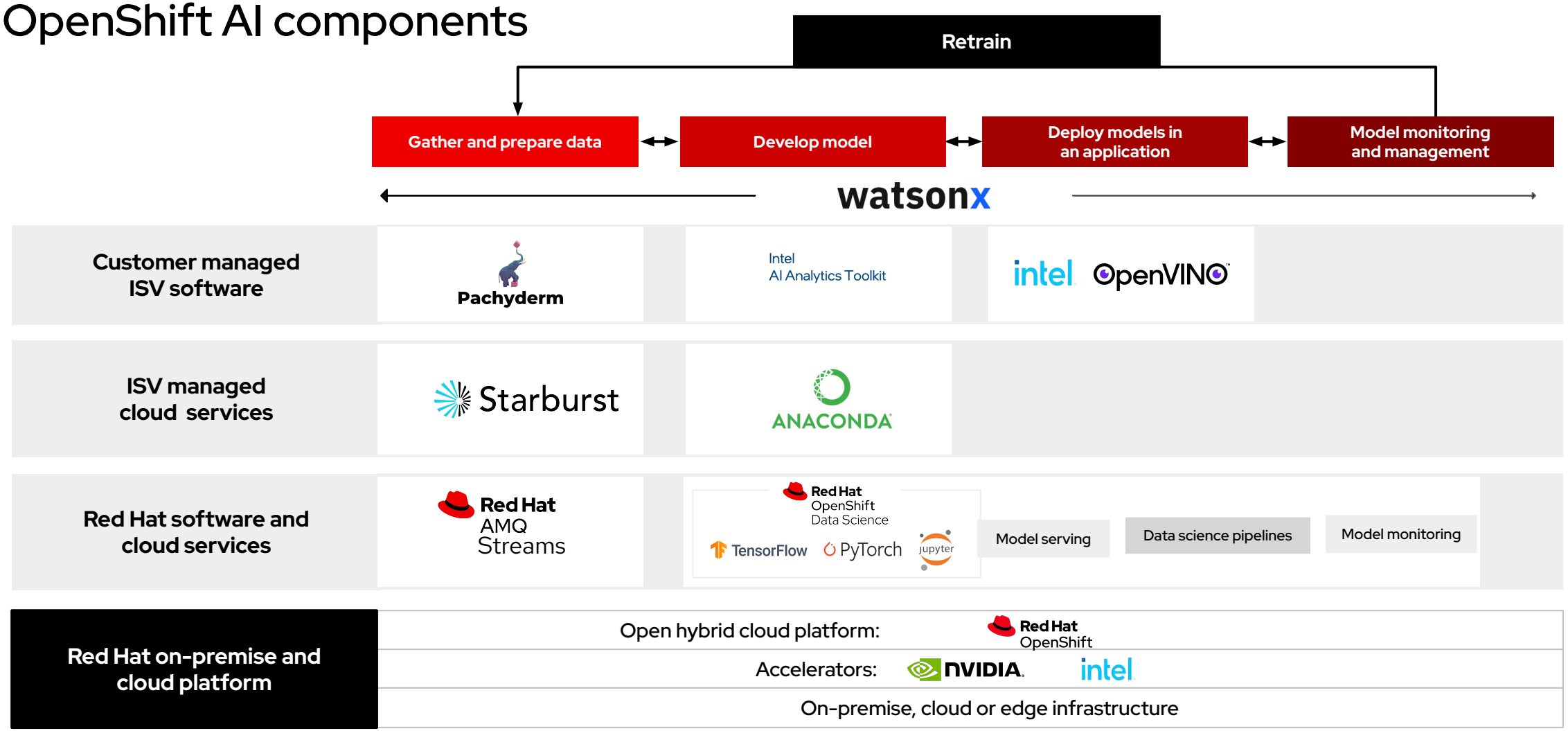
Create repeatable data science pipelines for model training and validation and integrate them with devops pipelines for delivery of models across your enterprise.



Increased capabilities / collaboration

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

OpenShift AI components



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 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

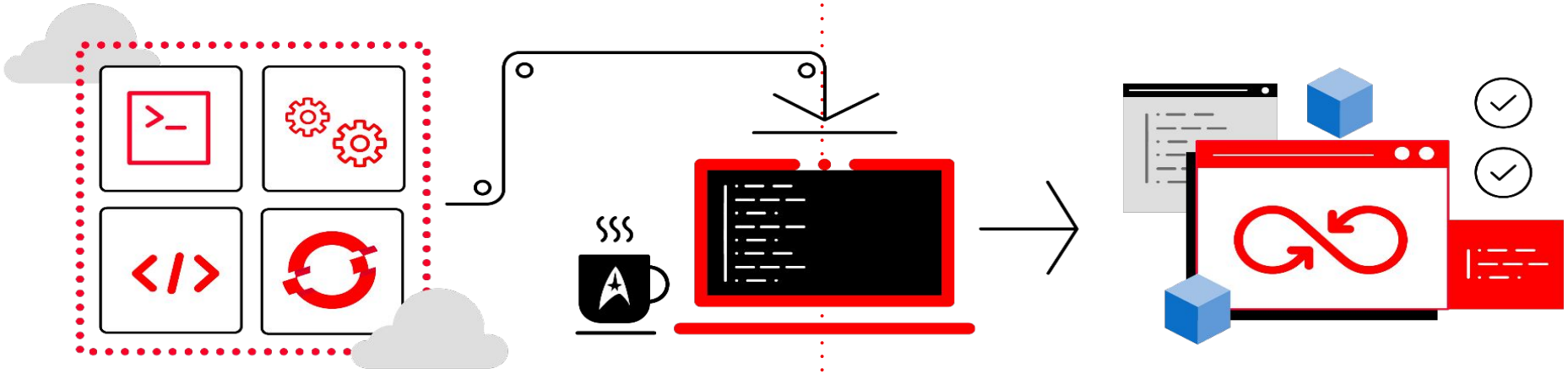
BOSTON UNIVERSITY

- **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 Consulting Services for your AI/ML Journey

Red Hat OpenShift AI Platform Engagement

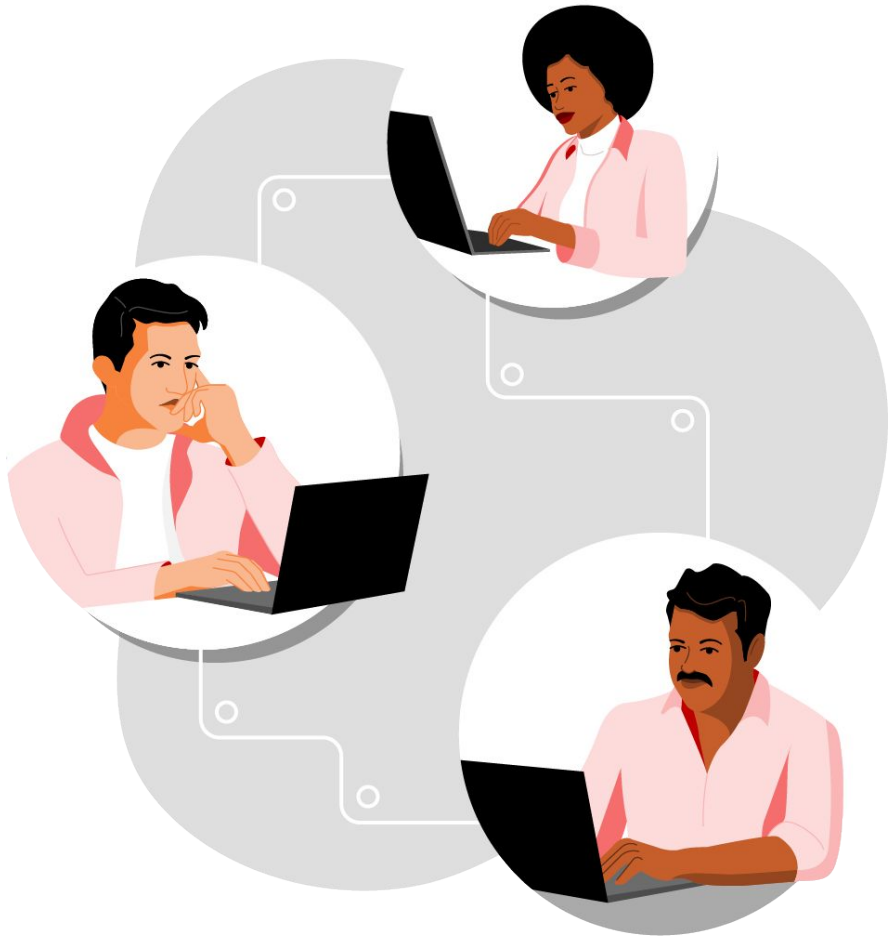
MLOps Foundation



Red Hat

Functionality Details

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

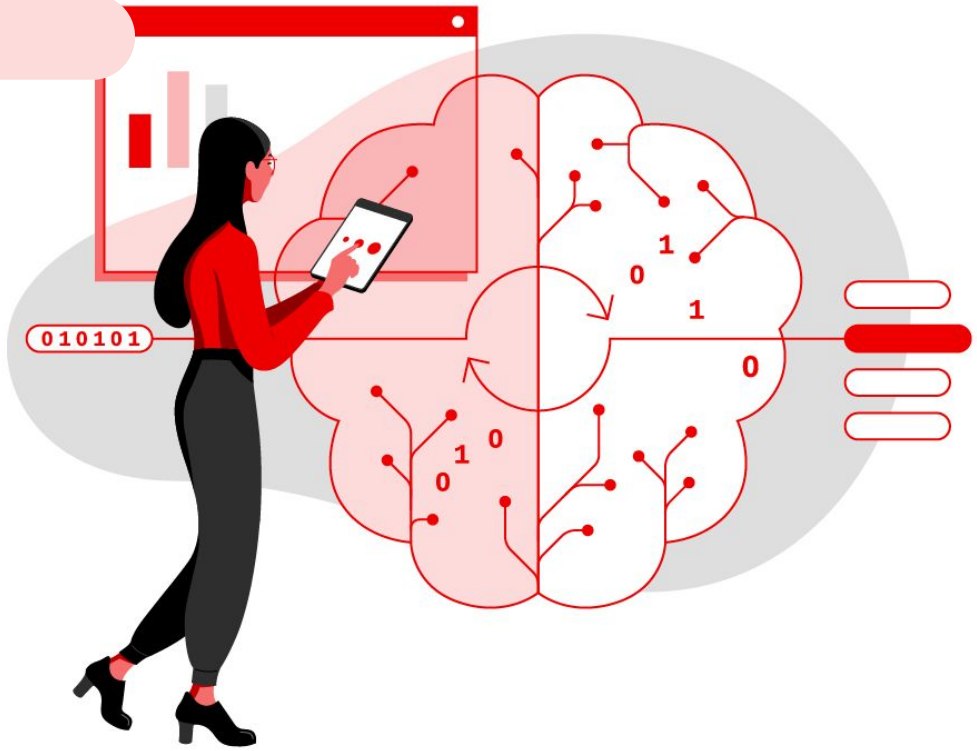
Automate setup and deployment
so you can get up and running with minimal effort

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

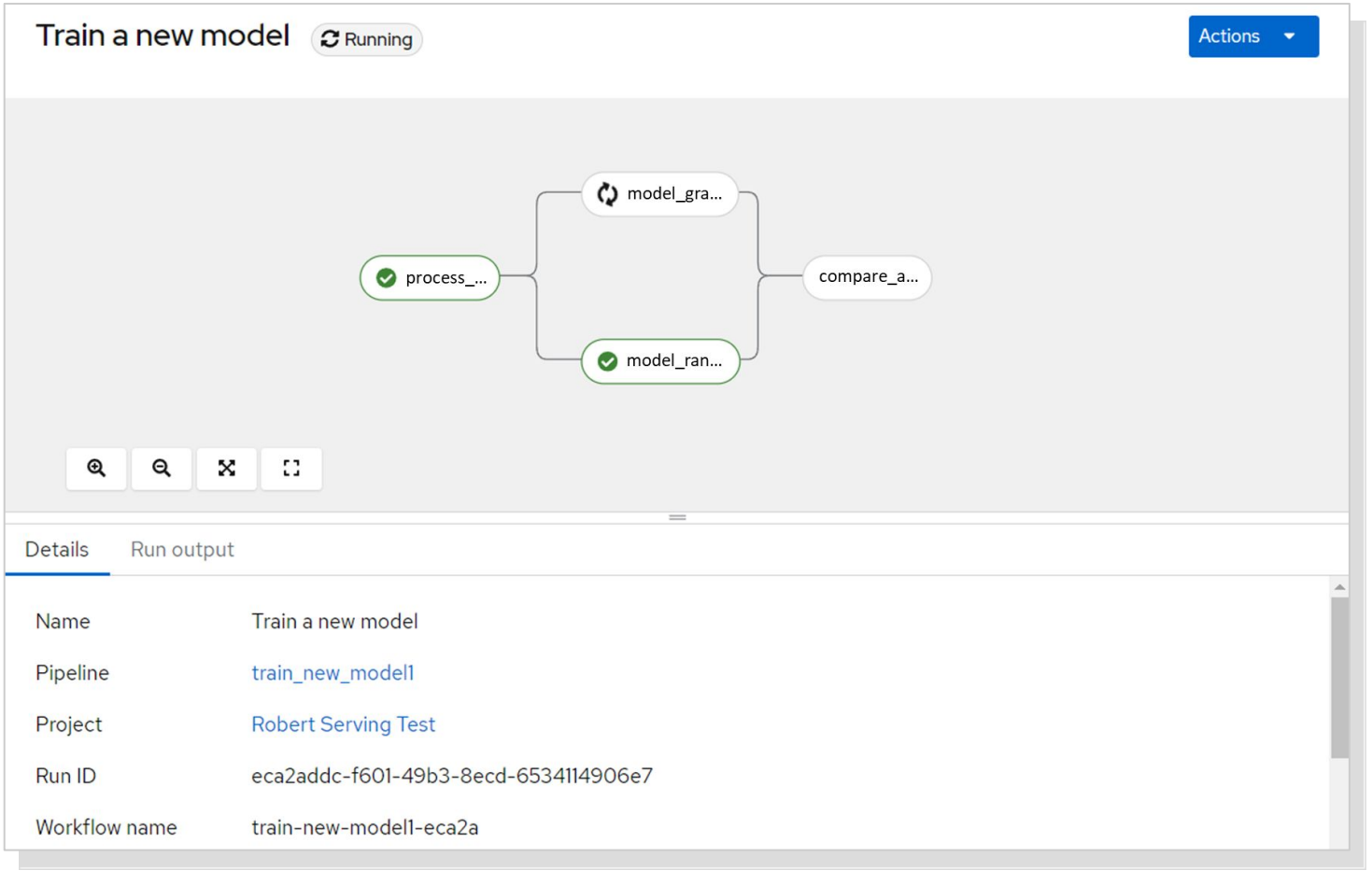
Make model serving more flexible



- **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

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

Device edge

Device
or sensor

Far edge

Red Hat OpenShift AI
Model serving

Near edge

Red Hat OpenShift AI
Model monitoring
Model registry

Enterprise

Red Hat OpenShift AI
Pipelines

Red Hat OpenShift AI
Model training

Edge

Unreliable connection

Last-mile network

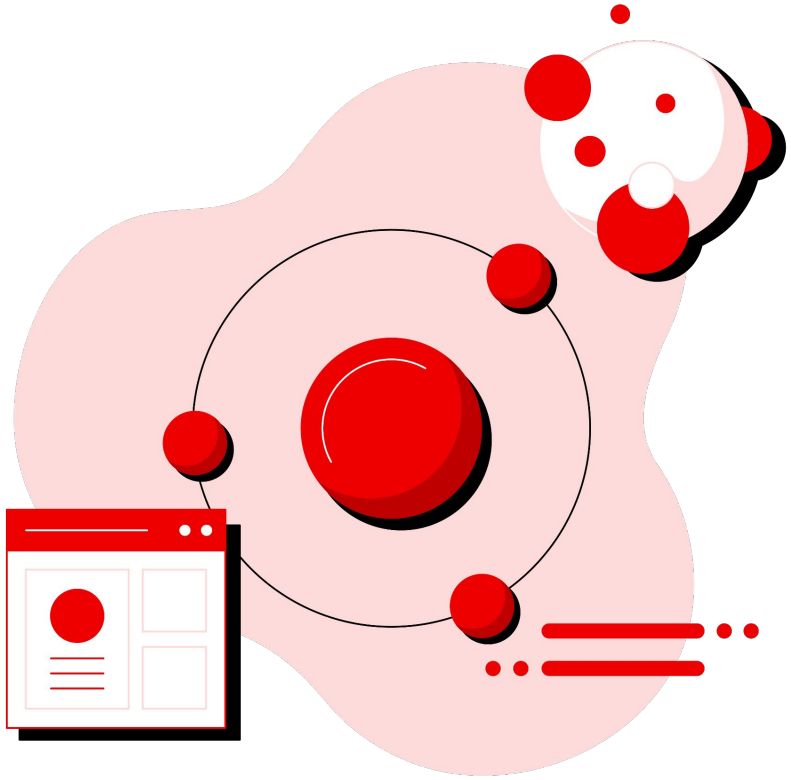
Core

Sensor data, telemetry, events, operational data, general information, etc.

Code, configuration, master data, machine learning models, control, commands, etc.

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

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