MLOps with OpenShift AI

A technical overview

Cansu Kavili Örnek
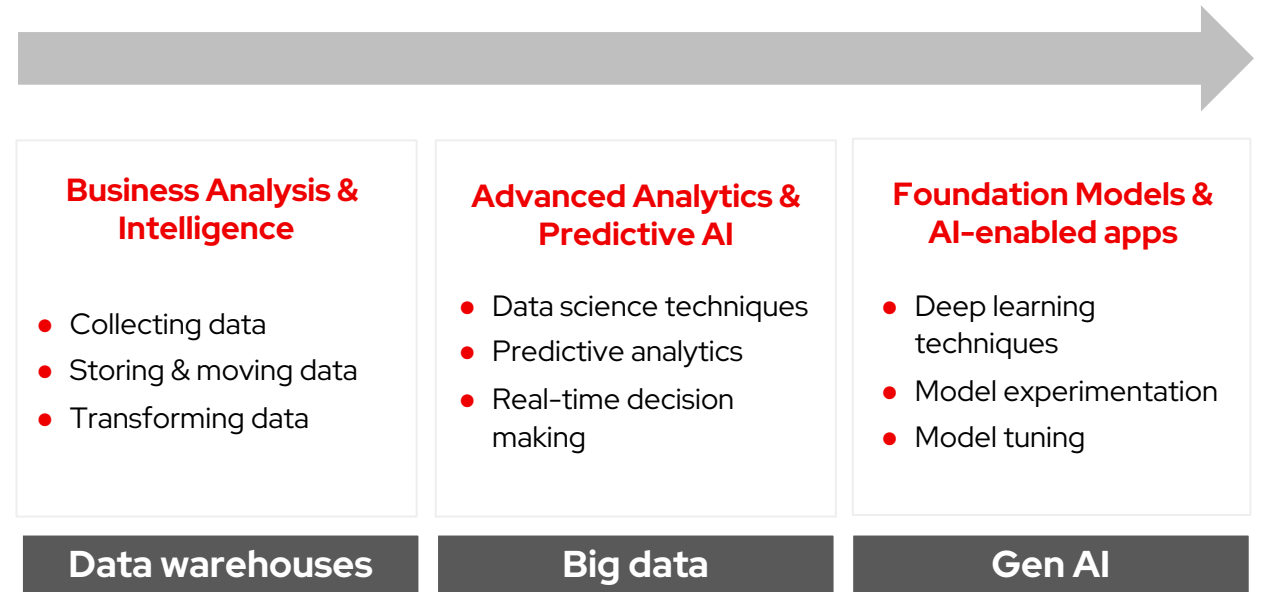
Principal AI Consultant

cansu@redhat.com

Erkan Ercan

Principal Solution Architect

eercan@redhat.com

AI has undergone significant evolution

The evolution of **Data-driven** AI: from Business Intelligence to Generative AI

▸ Predictive AI runs businesses today

▸ Foundation models provide a shortcut for realizing the value of AI

### Business Analysis & Intelligence

- Collecting data
- Storing & moving data
- Transforming data

Data warehouses

Advanced Analytics & Predictive AI

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

Big data

Foundation Models & AI-enabled apps

- Deep learning techniques
- Model experimentation
- Model tuning

Gen AI

Red Hat

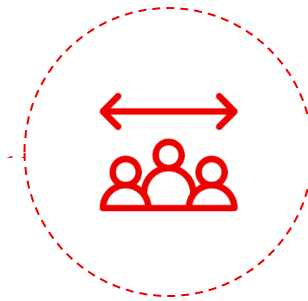# Personas

Operationalizing AI projects is a team sport

Line of Business

‣ Time to business value
‣ Reliability of predictions
‣ Responsive to evolving business
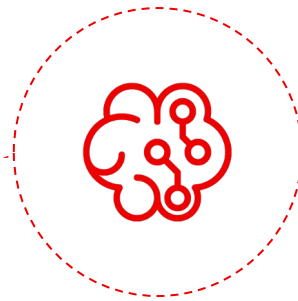
Data Scientist, Data Engineer, App Developer

‣ Self-service access to tools, data & infrastructure
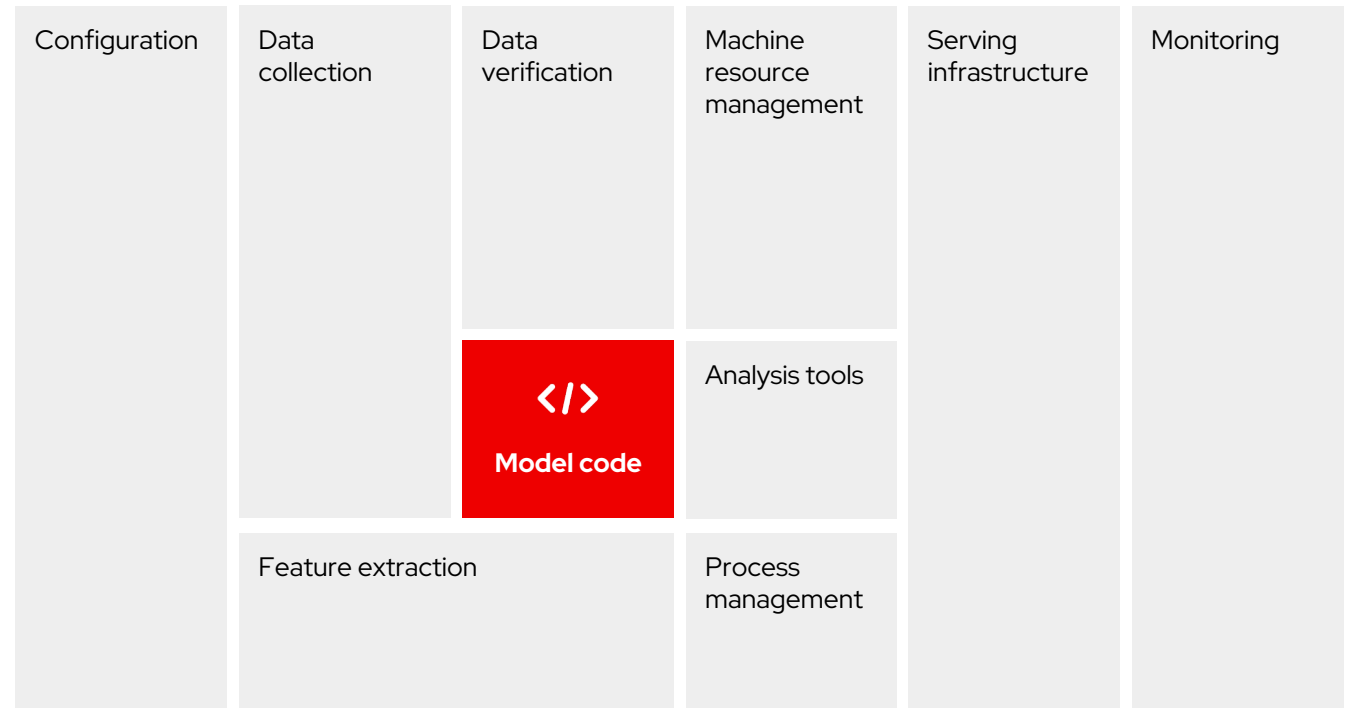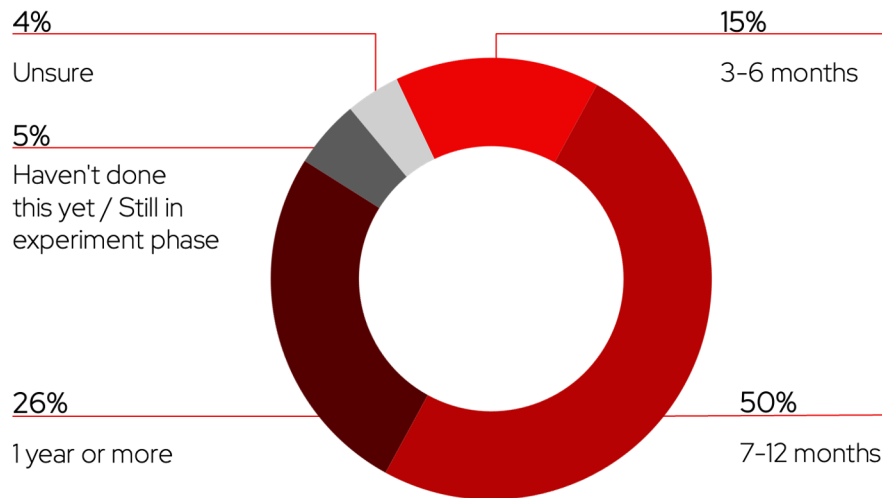‣ Collaborative environments for fast AI-enabled apps creation

IT Operations

‣ High availability
‣ Security
‣ Ease of management
‣ Scalability
‣ Enhance existing vs "rip and replace"

Operationalizing AI is one of the biggest challenges

What is the average timeline from **idea to operationalizing** the model?

4%
Unsure

5%
Haven't done this yet / Still in experiment phase

26%
1 year or more

15%
3-6 months

50%
7-12 months

Configuration	Data collection	Data verification	Machine resource management	Serving infrastructure	Monitoring
		Model code	Analysis tools		
	Feature extraction	Process management			

(Adapted from Sculley et al., "Hidden Technical Debt in Machine Learning Systems." NIPS 2015)
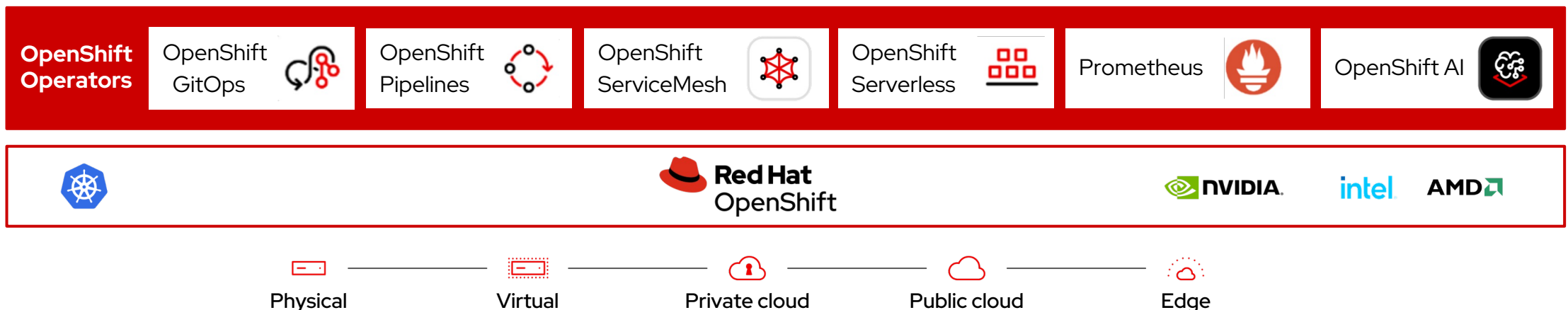
Red Hat

MLOps

The **practices, culture,** and **tools** that aim to **reliably** and **efficiently** **deploy** and **maintain** **AI/ML models** in production.

Red Hat

Red Hat OpenShift

A platform for continuous development, integration, and deployment for AI/ML models with GPU support

▶ **One platform for all.** Collaborative environments for dev, data engineers, data scientists and DevOps.

▶ **Extend capabilities with operators.** Operators allow bringing AI/ML capabilities to OpenShift.

▶ **Hybrid-cloud support.** On-premise support for model development, delivery and deployment.

▶ **Enhanced security.** Expand DevSecOps practices to protect the AI/Ml lifecycle.

OpenShift Operators	OpenShift GitOps	OpenShift Pipelines	OpenShift ServiceMesh	OpenShift Serverless	Prometheus	OpenShift AI

Red Hat OpenShift

NVIDIA   intel   AMD

Physical — Virtual — Private cloud — Public cloud — Edge

Red Hat

Red Hat OpenShift AI

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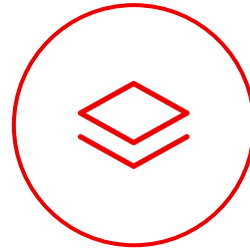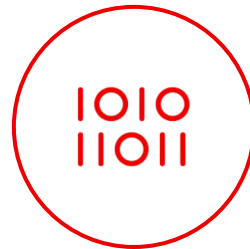

Built on top of OpenShift®
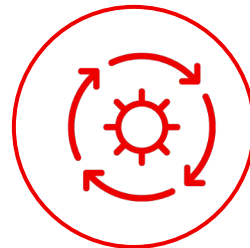Deliver consistency, **cloud-to-edge** production deployment and monitoring capabilities

Designed for machine learning
Scale to meet the workload demands of foundation models

Empowered data science
Provide a **unified platform** for **data scientists** and **intelligent application developers**

DevOps applied to ML
Set up rigorous **pipelines** and **workflows** to take you from **development** to **production**.

Red Hat OpenShift

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

Red Hat Open Source AI platforms

InstructLab

STEP 1

Learn and experiment via limited desktop-scale training method (qlora) on small datasets.
Podman Desktop integration.

Laptop / desktop

Red Hat Enterprise Linux AI

STEP 2

Production-grade model training using full synthetic data generation, teacher and critic models. Tooling focused on scriptable primitives.
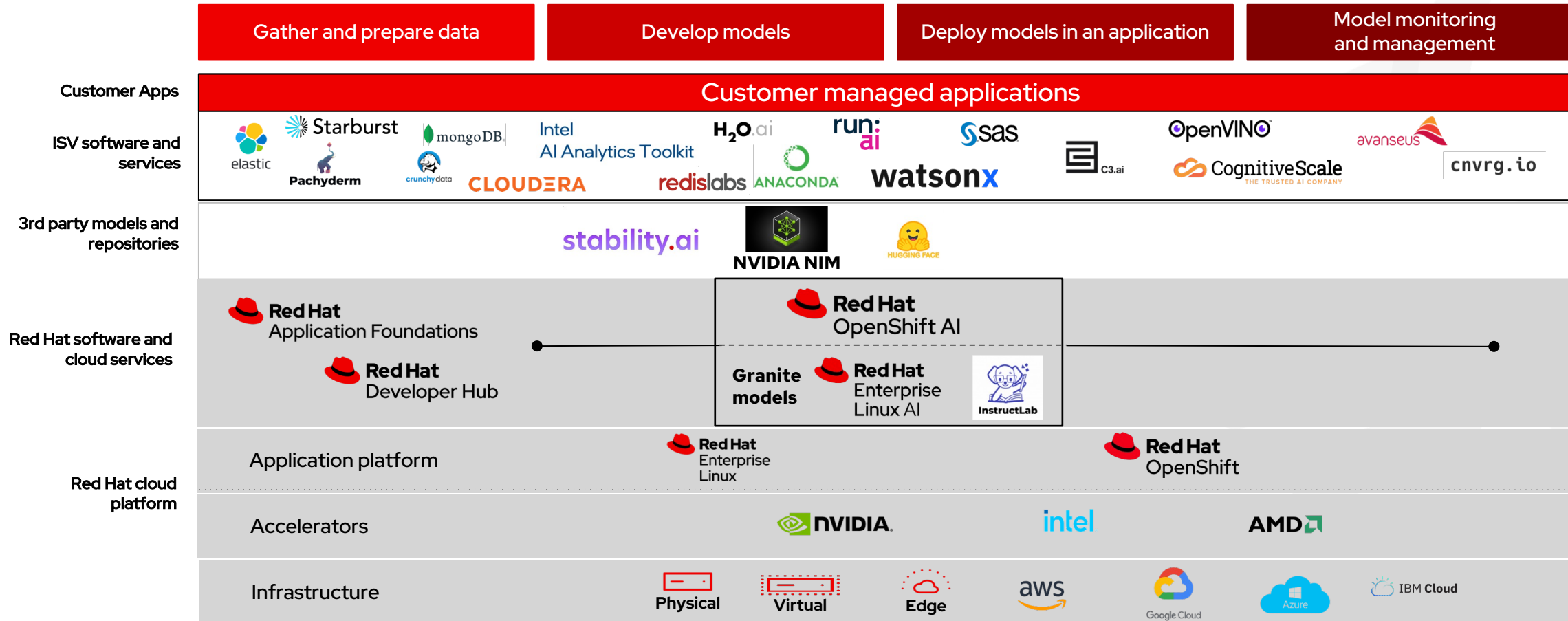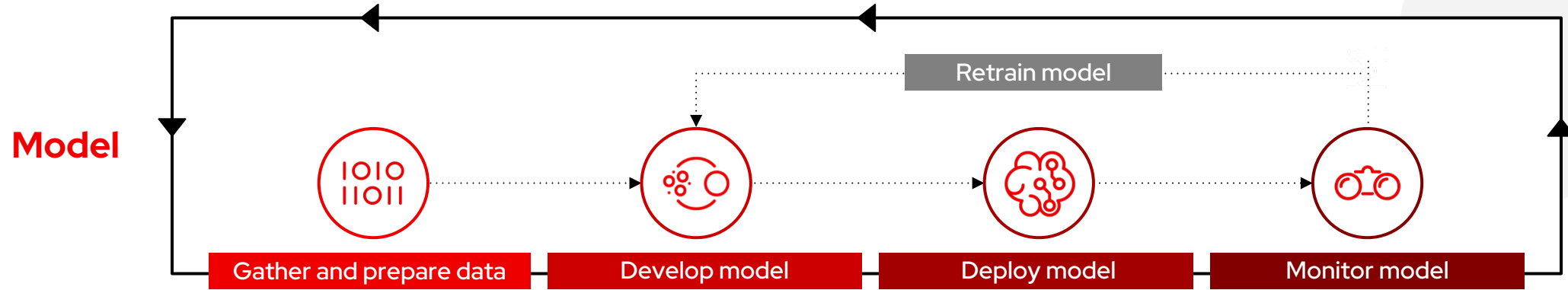
Server / VM

Red Hat OpenShift AI

STEP 3

Production-grade model training as in RHEL AI, using full power of Kubernetes scaling, automation, and MLOps services.
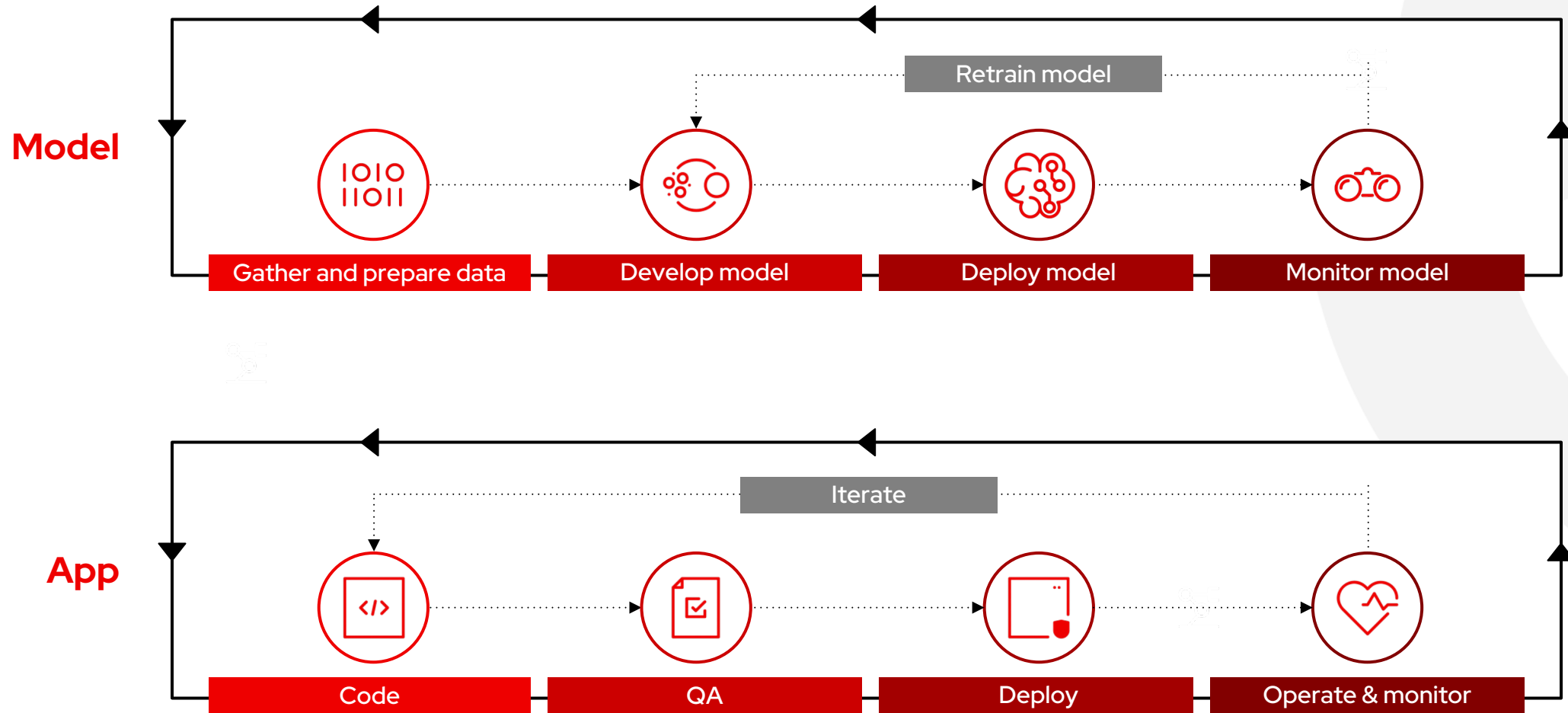
Cluster

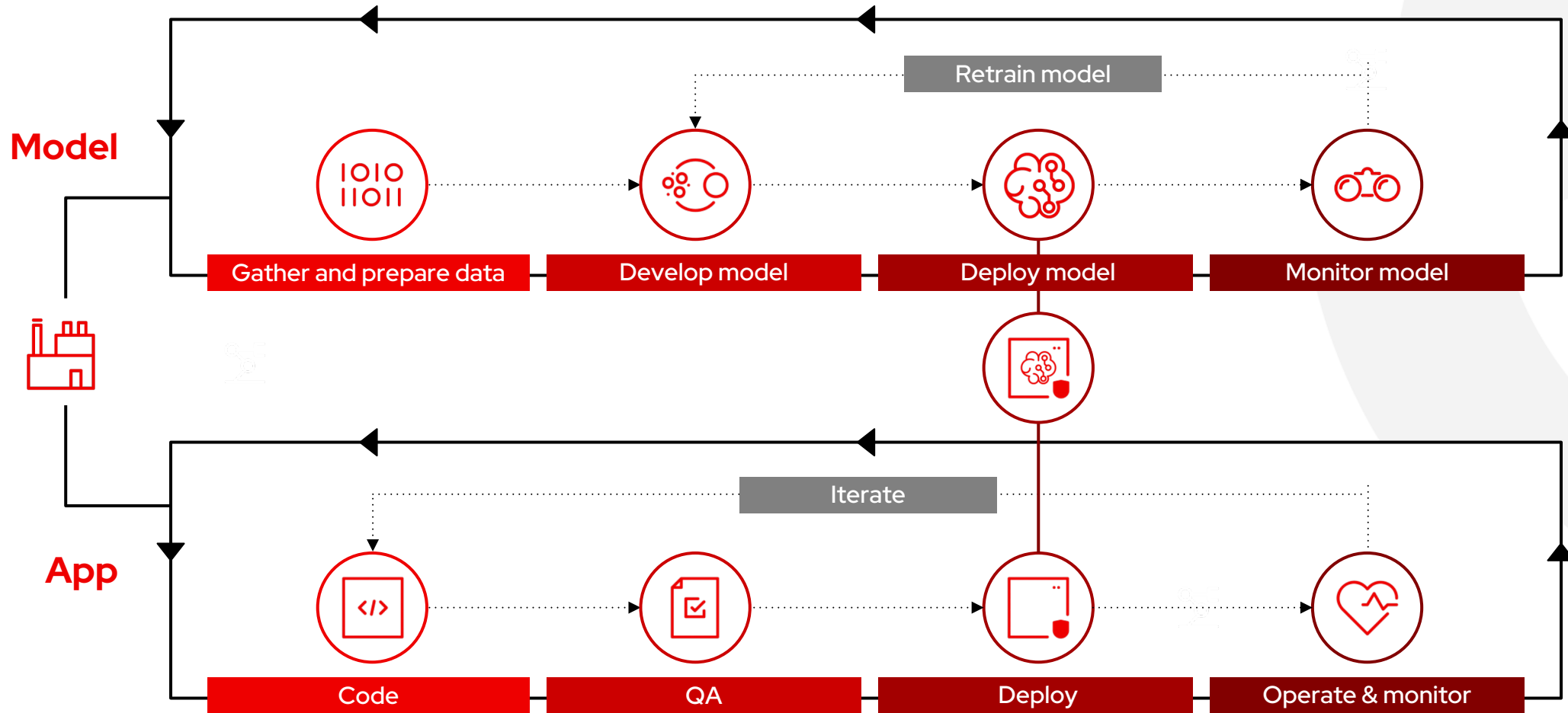Red Hat

Detailed look at OpenShift AI's partner ecosystem

Gather and prepare data	Develop models	Deploy models in an application	Model monitoring and management

Customer Apps — Customer managed applications

ISV software and services
elastic · Starburst · Pachyderm · mongoDB · crunchy data · CLOUDERA · Intel AI Analytics Toolkit · H₂O.ai · redislabs · ANACONDA · run:ai · watsonx · SAS · C3.ai · OpenVINO · CognitiveScale THE TRUSTED AI COMPANY · avanseus · cnvrg.io

3rd party models and repositories
stability.ai · NVIDIA NIM · HUGGING FACE

Red Hat software and cloud services
Red Hat Application Foundations · Red Hat Developer Hub · Red Hat OpenShift AI · Granite models · Red Hat Enterprise Linux AI · InstructLab

Red Hat cloud platform

Application platform
Red Hat Enterprise Linux · Red Hat OpenShift

Accelerators
NVIDIA · intel · AMD

Infrastructure
Physical · Virtual · Edge · aws · Google Cloud · Azure · IBM Cloud

11

Lifecycle for operationalizing models

Model

Retrain model

Gather and prepare data | Develop model | Deploy model | Monitor model

We've seen this before..

Model

Retrain model

| Gather and prepare data | Develop model | Deploy model | Monitor model |

App

Iterate

| Code | QA | Deploy | Operate & monitor |

Red Hat

Your success depends on your ability to operationalize models as well

Model

Retrain model

| Gather and prepare data | Develop model | Deploy model | Monitor model |

App

Iterate

| Code | QA | Deploy | Operate & monitor |

Red Hat

What's in the box?

- Multi-user workbench
- Data science and machine learning workflows

CodeFlare

- Scale distributed computing for AI
- Automatically adjust the underlying workers based on demands

- Streamline the entire ML lifecycle
- Accelerate model development & deployment

- Design pipeline with drag and drop ease
- Kubeflow integration

KServe & ModelMesh

- Supports multiple ML frameworks
- Deploy and scale AI models quickly and efficiently

Trusty AI

- AI Explainability Toolkit
- Aims to mitigate AI bias, enhancing trust and fairness in AI systems

Plus everything we know from OpenShift

- GitOps Controller
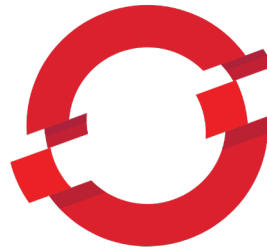- Declarative app management

- Cloud Native CI/CD pipeline
- Scalable pipeline orchestration

- Real time metrics monitoring
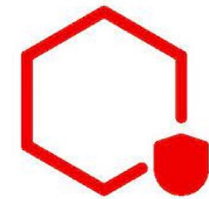- Query-based alerting

- OCI compliant image management

Autoscaling

- Demand base machine autoscaling
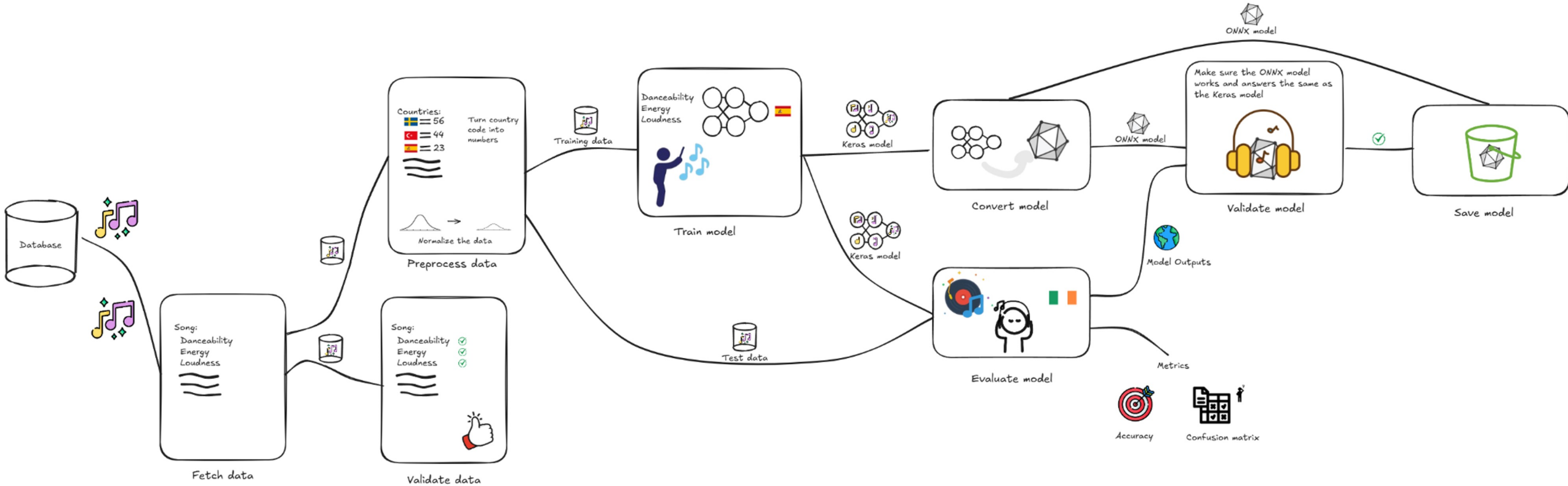- Resource optimization

Multi-tenancy

- Isolated environment
- Enhanced security boundaries

Live Demo

Demo workflow – update this with tech logos

Incremental Big Picture

Red Hat Summit

Connect

Thank you

in linkedin.com/company/red-hat

f facebook.com/redhatinc

▶ youtube.com/user/RedHatVideos

🐦 twitter.com/RedHat

Red Hat