Pioneering Product Development & Technology Consulting for a Tech-Powered Future.

TEB Arf

A Rising Future

Revolutionizing AI Infrastructure:

How TEB and OBSS Optimized AI Ops With OpenShift AI

Yusuf Tok

Chief Technology Officer

OBSS

Aykut Koltarla

Head of AI and Data Science

TEB Arf

OpenShift AI Features

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 AI - KEY FEATURES

Model Development

Self-service workbenches AI/ML tooling, libraries, frameworks, etc.

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, and tuning

Responsible AI

Monitoring capabilities for finding and fixing biases before, after and during production

RED HAT'S AI/ML ENGINEERING IS 100% OPEN SOURCE

Upstream Projects

RAY

CodeFlare

TrustyAI

vLLM

TensorFlow

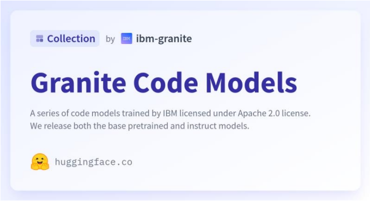
PyTorch

ModelMesh Serving

jupyter

KServe

Kubeflow



Collection by ibm-granite

Granite Code Models

A series of code models trained by IBM licensed under Apache 2.0 license.
We release both the base pretrained and instruct models.

huggingface.co

InstructLab

→

Community Projects

OPEN DATA HUB

AI Platform powered by Open Source

podman desktop

Podman AI Lab

→

Product

Red Hat OpenShift AI

Red Hat Enterprise Linux AI

TEB Requirements and Pain Points

TEB Possible Pain Points

- Operational Complexity
- Infrastructure Management
- Data Security and Compliance
- Talent and Expertise
- Model Lifecycle Management
- Performance and Optimization
- Tooling and Technology Stack

OPERATIONAL COMPLEXITY

Model Deployment

Challenges in moving models from development to production reliably and efficiently.

Continuous Monitoring

Maintaining robust monitoring and alerting systems to track model performance and detect issues early.

Model Retraining and Updating

Automating the retraining process and ensuring that models are consistently updated with the latest data.

Pipeline Management

Managing complex ML pipelines involving multiple steps from data ingestion to deployment and monitoring.

INFRASTRUCTURE MANAGEMENT

Scalability

Difficulty in scaling infrastructure to handle large-scale data processing and model training efficiently.

Resource Management

Optimizing resource allocation for high-cost GPU and compute resources while minimizing expenses.

Integration with Existing IT

Seamlessly integrating the MLOps platform with legacy systems and existing IT infrastructure.

DATA SECURITY AND COMPLIANCE

Regulatory Compliance

Ensuring adherence to financial regulations and data privacy laws (e.g., GDPR, CCPA, or specific banking regulations)

Data Security

Protecting sensitive financial and personal customer data during model development, training, and deployment.

Data Management

Ensuring that training data is clean, high-quality, and relevant for building accurate models. Managing versions of datasets to support reproducibility and traceability of model training.

TALENT AND EXPERTISE

Skill Gap

Limited internal expertise in MLOps practices and technologies, making platform development and maintenance challenging.

Training and Upskilling

The need for continuous training and upskilling of staff to effectively use and manage the MLOps platform.

Cross-Team Collaboration

Facilitating effective collaboration between data scientists, developers, and IT/operations teams.

MODEL LIFECYCLE MANAGEMENT

Model Explainability

Ensuring that models are interpretable and explainable, which is critical for gaining stakeholder trust and meeting regulatory requirements.

Bias Detection and Mitigation

Implementing tools to identify and mitigate biases in models to promote fairness and accuracy.

Version Control

Keeping track of different model versions and their associated metadata for reproducibility.

PERFORMANCE AND OPTIMIZATION

Latency and Response Time

Ensuring low-latency predictions in real-time applications.

Optimization

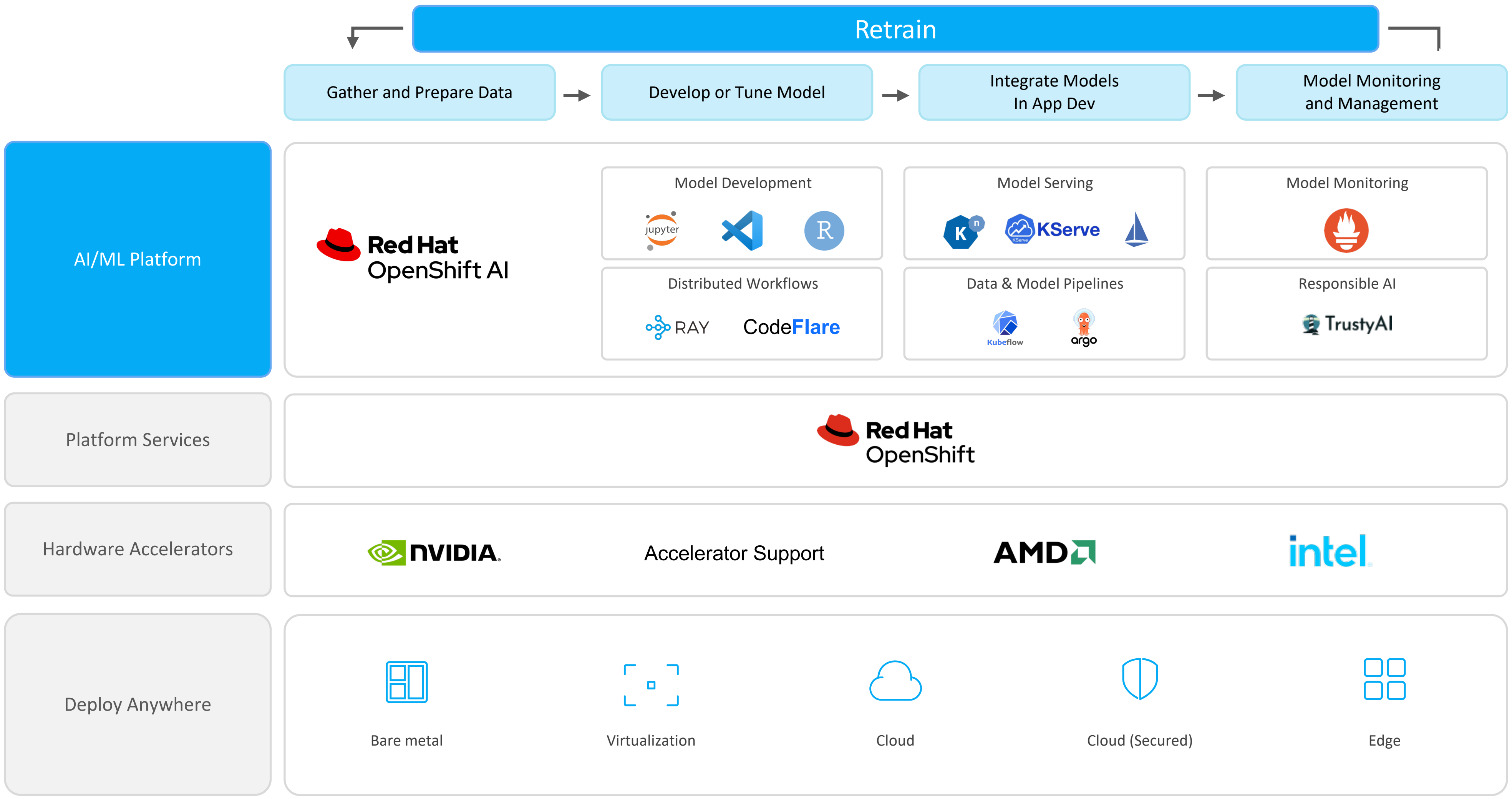
Optimizing models for performance without compromising accuracy.

Resource Utilization

Efficiently utilizing resources to avoid overuse and cost overruns.

RED HAT OPENSHIFT AI

Red Hat's AI/ML platform for predictive and Gen AI applications



Retrain

Gather and Prepare Data → Develop or Tune Model → Integrate Models In App Dev → Model Monitoring and Management

AI/ML Platform

Red Hat OpenShift AI

Model Development: jupyter, R

Model Serving: KServe

Model Monitoring

Distributed Workflows: RAY, CodeFlare

Data & Model Pipelines: Kubeflow, argo

Responsible AI: TrustyAI

Platform Services

Red Hat OpenShift

Hardware Accelerators

NVIDIA. Accelerator Support AMD intel.

Deploy Anywhere

Bare metal

Virtualization

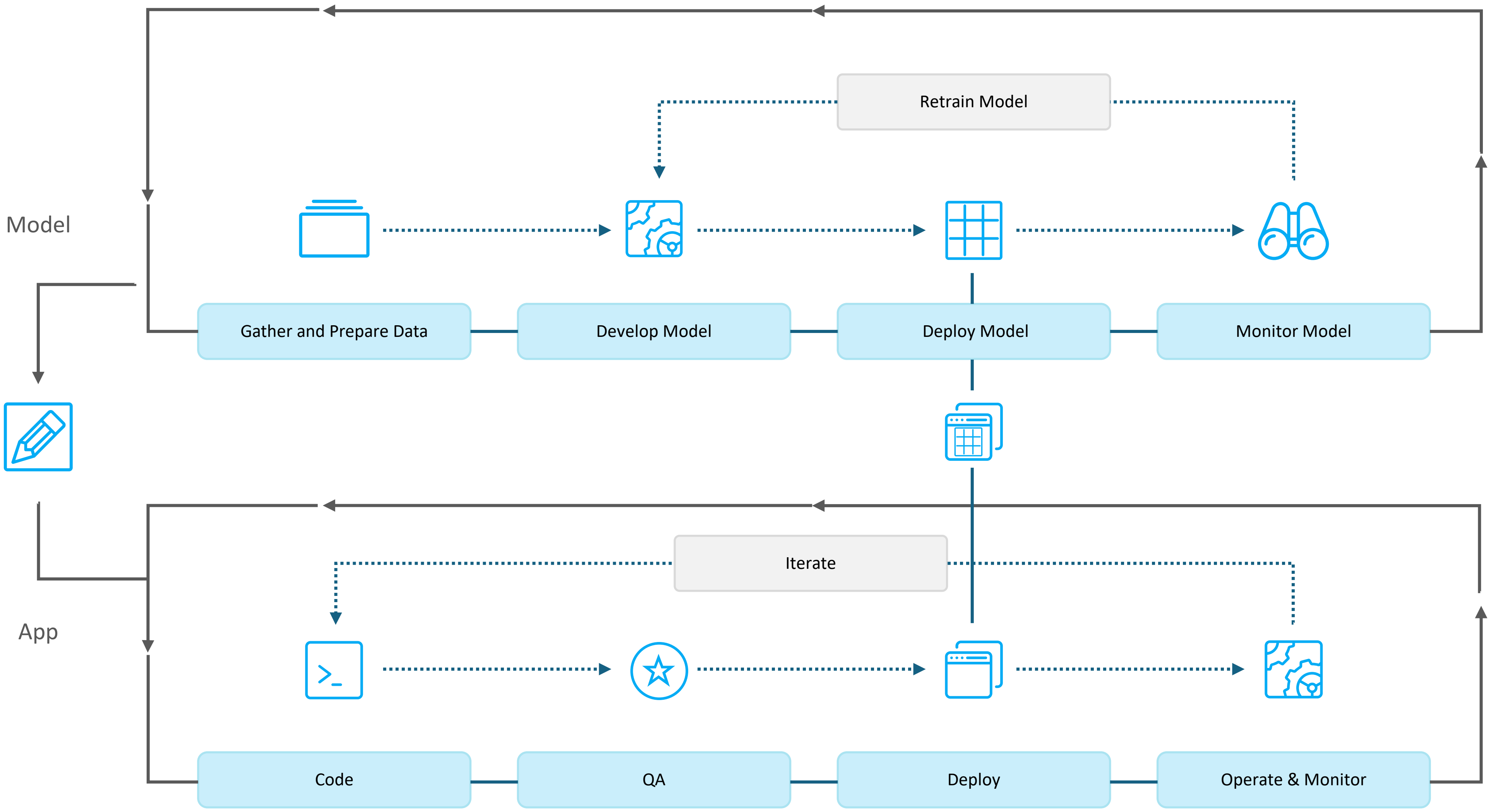
Cloud

Cloud (Secured)

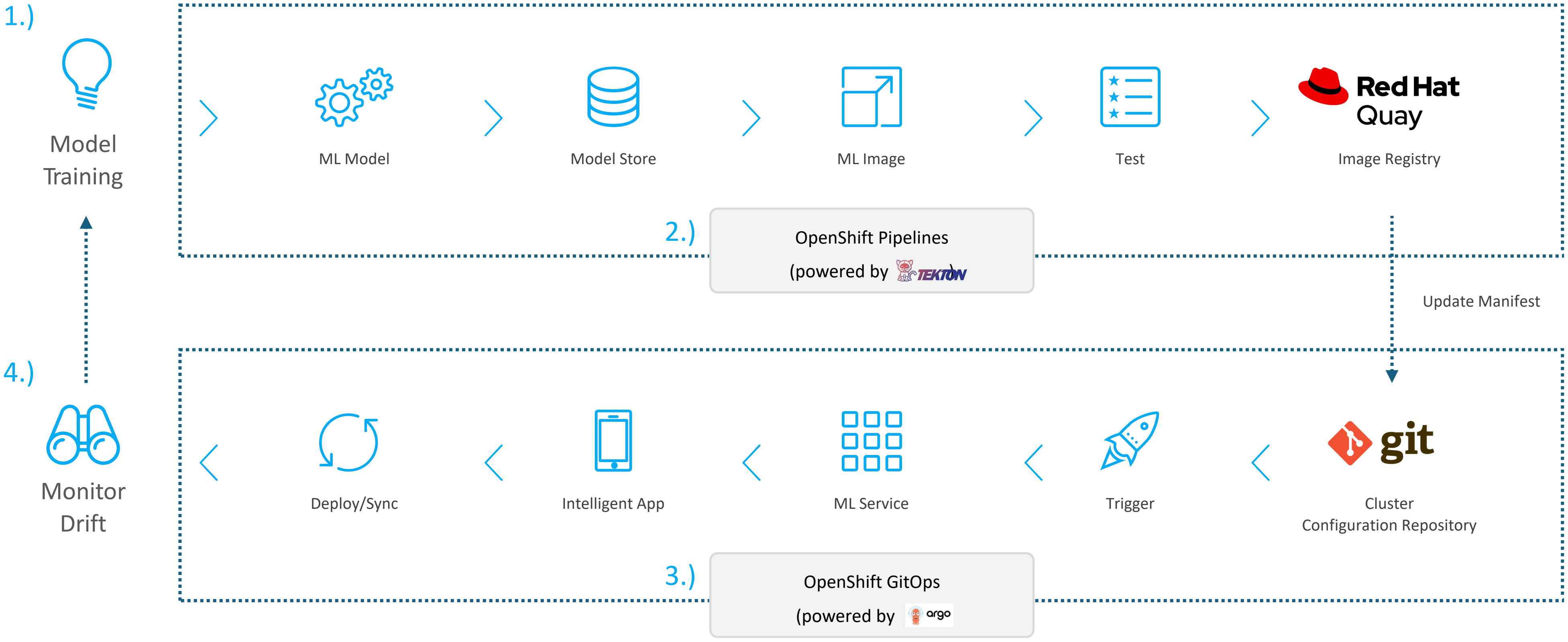
Edge

REALIZING VALUE FROM AI/ML

Lifecycle for operationalizing models



MLOPS WITH RED HAT OPENSHIFT



1.) Model Training

ML Model

Model Store

ML Image

Test

Red Hat Quay

Image Registry

2.) OpenShift Pipelines (powered by TEKTON)

Update Manifest

git

Cluster Configuration Repository

Trigger

ML Service

Intelligent App

Deploy/Sync

3.) OpenShift GitOps (powered by argo)

4.) Monitor Drift

Red Hat OpenShift

Red Hat OpenShift Data Foundation

TEB Arf

A Rising Future

Thank You!