Red Hat Summit

Connect

Navigating AI/ML Adoption with OpenShift AI

From Trial to Experimentation to Adoption and Scale

Adnan Drina

Specialist Solution Architect

AppDev & AI

Navigating AI/ML Adoption

From safe **experimentation** to collaborative **development** to scaled AI **integration**

Stage 1: Trial

How do I *evaluate* models and pick the best one for my use case?

Stage 2: Experiment

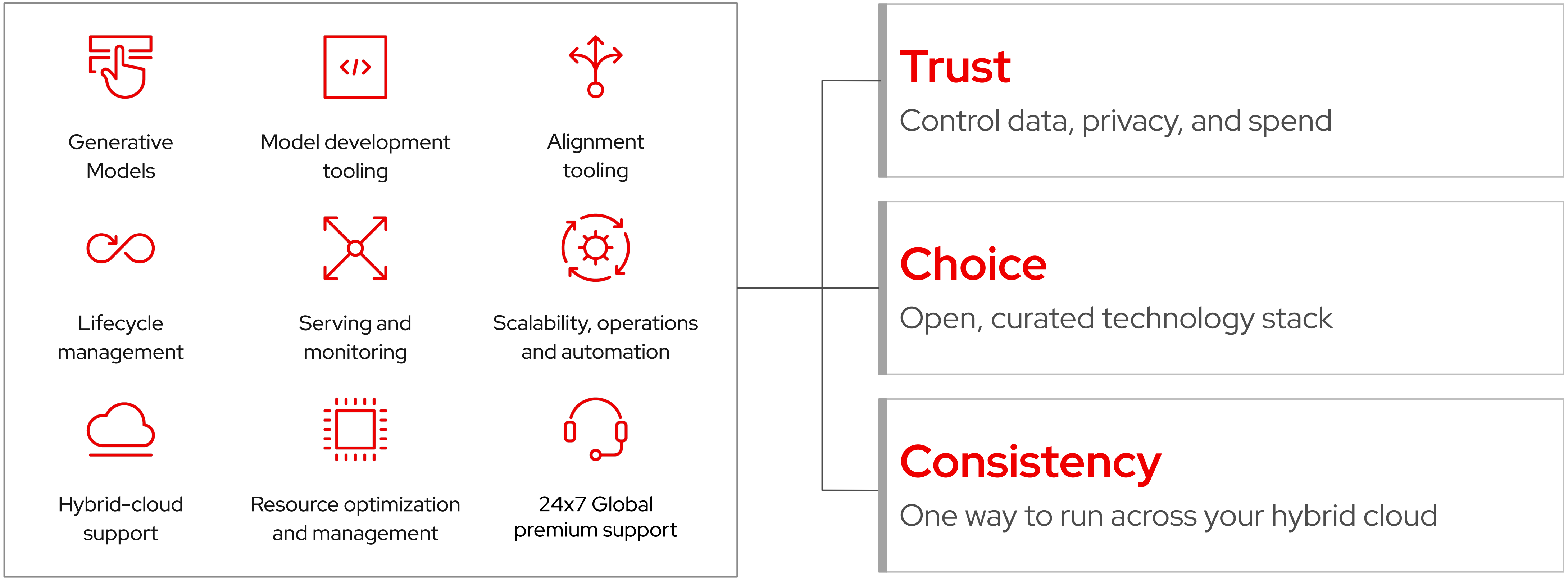
How do I *build and integrate* an AI application?

Stage 3: Adoption & Scale

How do I *deploy and operate* my AI application?

Building AI applications requires more than just models

Red Hat offers generative AI and MLOps capabilities for building flexible, trusted AI solutions at scale



Generative Models

Model development tooling

Alignment tooling

Lifecycle management

Serving and monitoring

Scalability, operations and automation

Hybrid-cloud support

Resource optimization and management

24x7 Global premium support

Trust

Control data, privacy, and spend

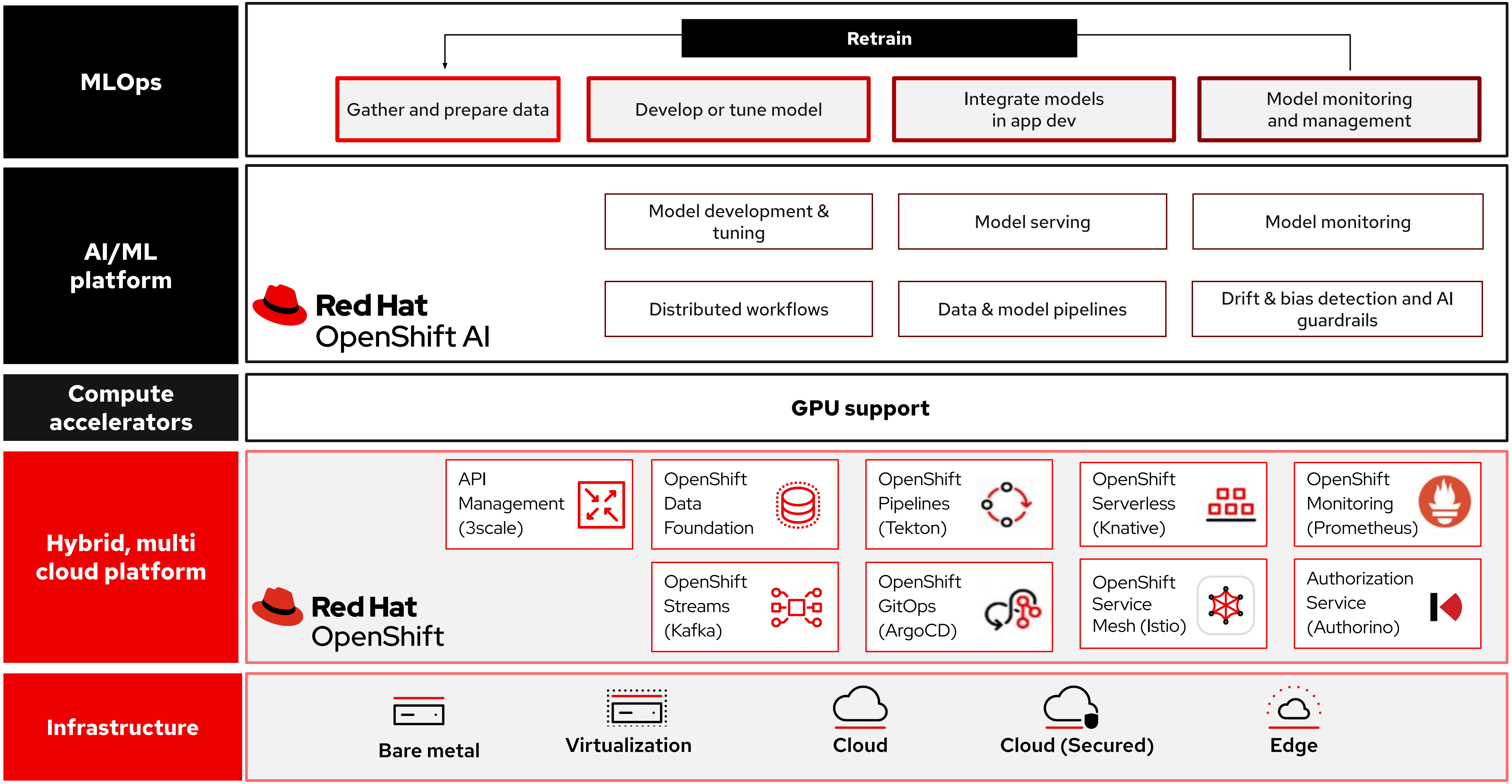
Choice

Open, curated technology stack

Consistency

One way to run across your hybrid cloud

Single platform to run *any* model, on *any* accelerator, on *any* cloud



MLOps

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 & tuning

Model serving

Model monitoring

Distributed workflows

Data & model pipelines

Drift & bias detection and AI guardrails

Compute accelerators

GPU support

Hybrid, multi cloud platform

Red Hat OpenShift

API Management (3scale)

OpenShift Data Foundation

OpenShift Pipelines (Tekton)

OpenShift Serverless (Knative)

OpenShift Monitoring (Prometheus)

OpenShift Streams (Kafka)

OpenShift GitOps (ArgoCD)

OpenShift Service Mesh (Istio)

Authorization Service (Authorino)

Infrastructure

Bare metal

Virtualization

Cloud

Cloud (Secured)

Edge

AI/ML Adoption Is a Collaborative Journey

Every member of your team plays a critical role—from early exploration to scaled integration

Business leadership

Define AI vision and success criteria

Data engineer

Prepare and pipeline data for modeling

Data scientists

Train, fine-tune, and evaluate models

ML engineer

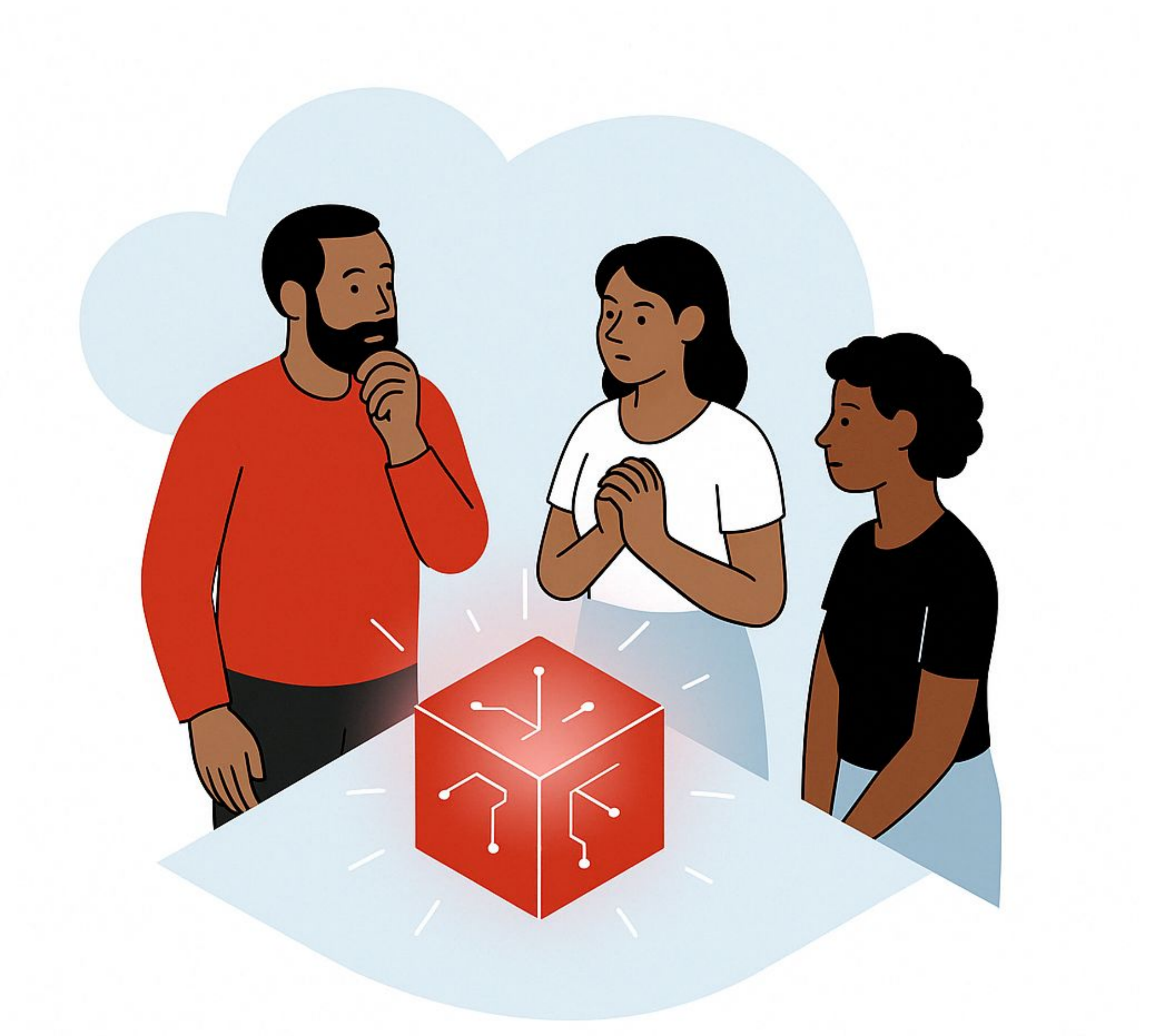
Operationalize models with MLOps

App developer

Integrate models into business applications

IT operations

Manage platform and ensure secure, scalable access



Stage: Trial
Evaluate Models

Strategic Goal

Establish a low-risk entry point for AI adoption. *Focus on enabling initial infrastructure and secure access to pre-trained models* to support evaluation and early alignment.

Why This Stage Matters

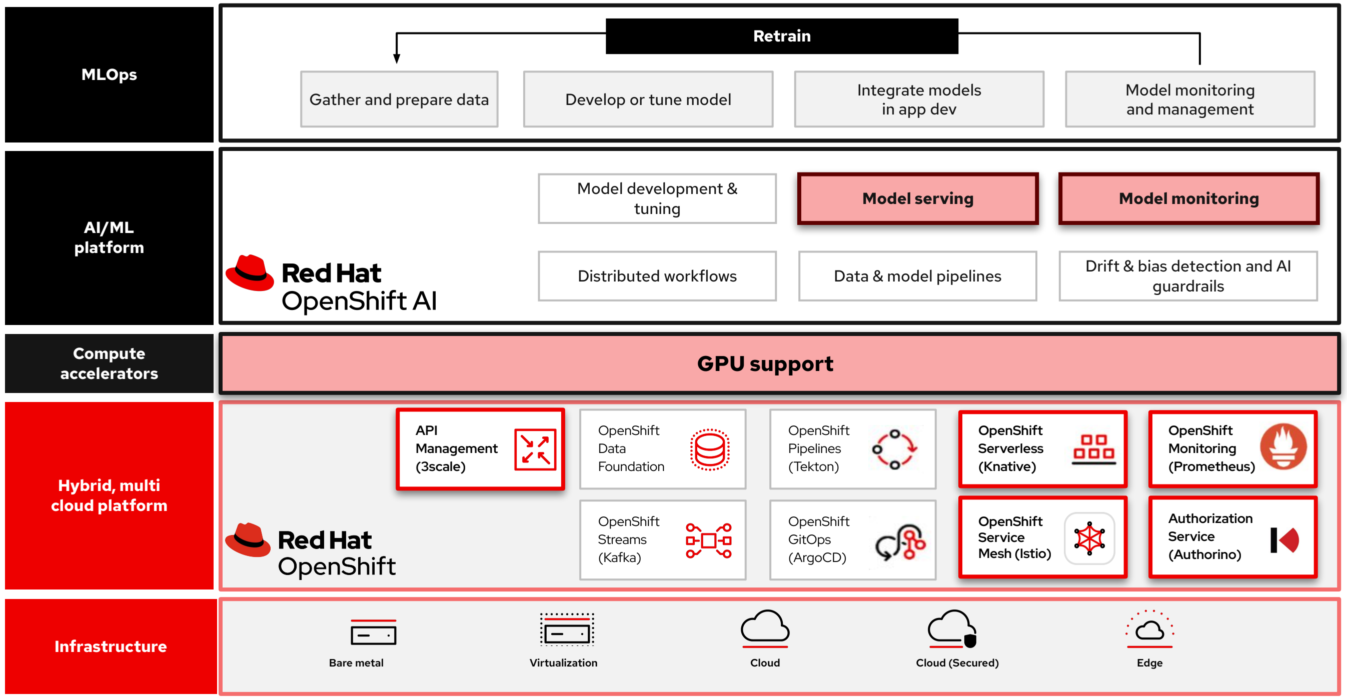
- *Establishes clarity*
- *Aligns teams*
- *Lays governance*

Enabling Foundational AI Capabilities

Focus on establishing secure, observable, and scalable **infrastructure** to support early experimentation

Platform Objectives:

- Provide a secure and scalable AI infrastructure
- Support model inference
- Expose models through secure APIs
- Provide a governed model catalog



MLOps

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 & tuning

Model serving

Model monitoring

Distributed workflows

Data & model pipelines

Drift & bias detection and AI guardrails

Compute accelerators

GPU support

Hybrid, multi cloud platform

Red Hat OpenShift

API Management (3scale)

OpenShift Data Foundation

OpenShift Pipelines (Tekton)

OpenShift Serverless (Knative)

OpenShift Monitoring (Prometheus)

OpenShift Streams (Kafka)

OpenShift GitOps (ArgoCD)

OpenShift Service Mesh (Istio)

Authorization Service (Authorino)

Infrastructure

Bare metal

Virtualization

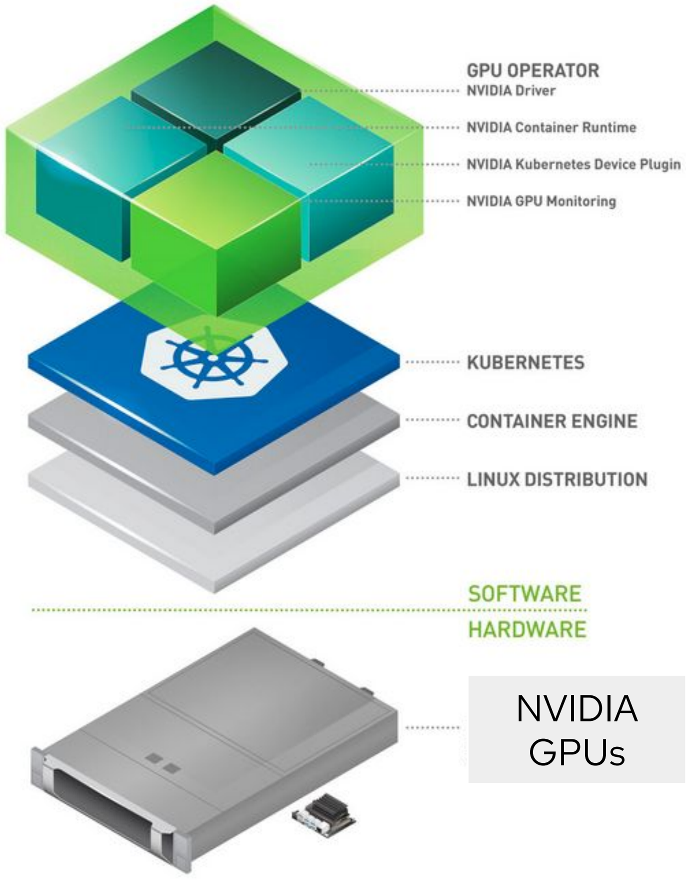
Cloud

Cloud (Secured)

Edge

NVIDIA GPU Operator on OpenShift

Automated GPU provisioning and monitoring for scalable AI workloads



1. Build NVIDIA GPU driver for RHCOS

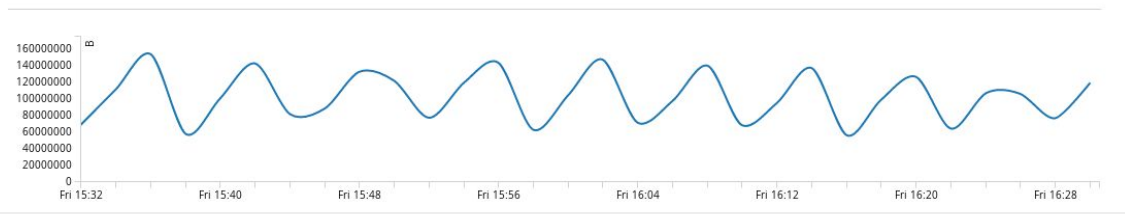
2. Expose NVIDIA GPUs as K8s extended resources

```
nvidia.com/gpu:2
```

3. Advertise NVIDIA GPU features with Node labels

```
nvidia.com/gpu.count=2
nvidia.com/gpu.memory=20096
nvidia.com/gpu.family=ampere
nvidia.com/gpu.product=A100-PCIE-40GB-MIG-...
...
```

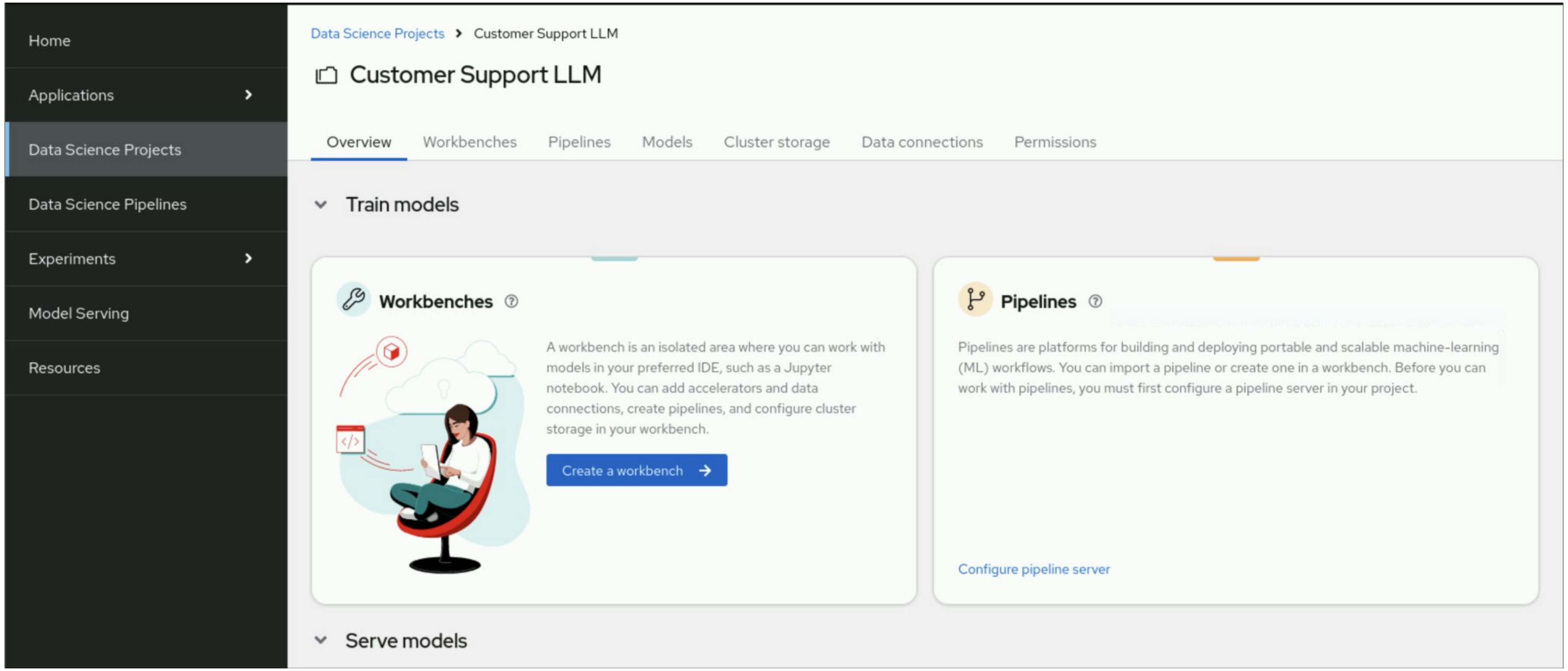
4. Expose NVIDIA GPU metrics to OpenShift



5. Handle operator upgrades

OpenShift AI Data Science Projects

Unified, secure workspaces for building, managing, and collaborating on AI projects



A workbench is an isolated area where you can work with models in your preferred IDE, such as a Jupyter notebook. You can add accelerators and data connections, create pipelines, and configure cluster storage in your workbench.

Pipelines are platforms for building and deploying portable and scalable machine-learning (ML) workflows. You can import a pipeline or create one in a workbench. Before you can work with pipelines, you must first configure a pipeline server in your project.

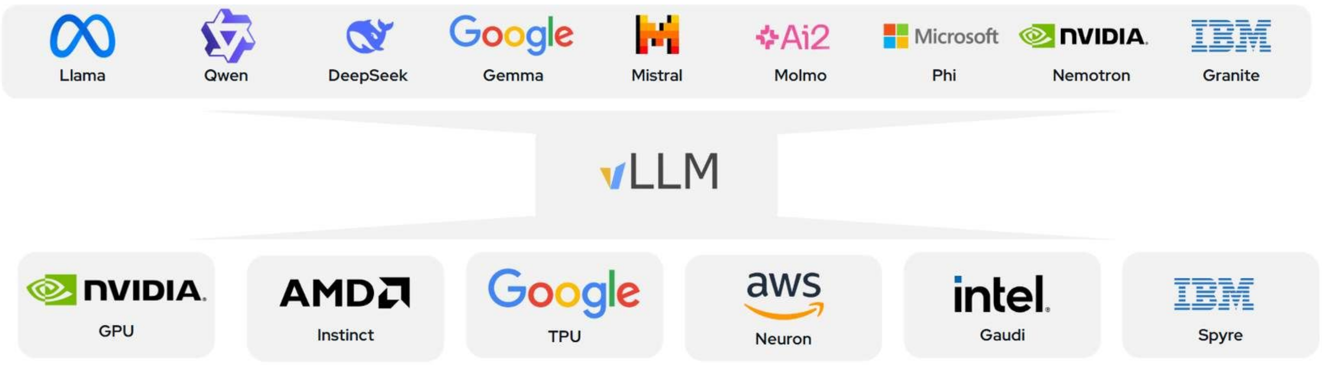
Data Science Projects enable teams to *organize, collaborate, and manage end-to-end AI projects*—offering workbenches, data connections, storage, pipelines, and model serving in a unified, secure environment.

Key Features:

- Scoped project workspaces: Kubernetes namespaces encapsulating notebooks, compute, storage, and pipelines
- Integrated development workbenches: Launch Jupyter, VS Code, or RStudio environments directly within projects
- Managed data & storage: Attach cluster storage and external data connections (e.g. S3) via UI

vLLM Inference Server

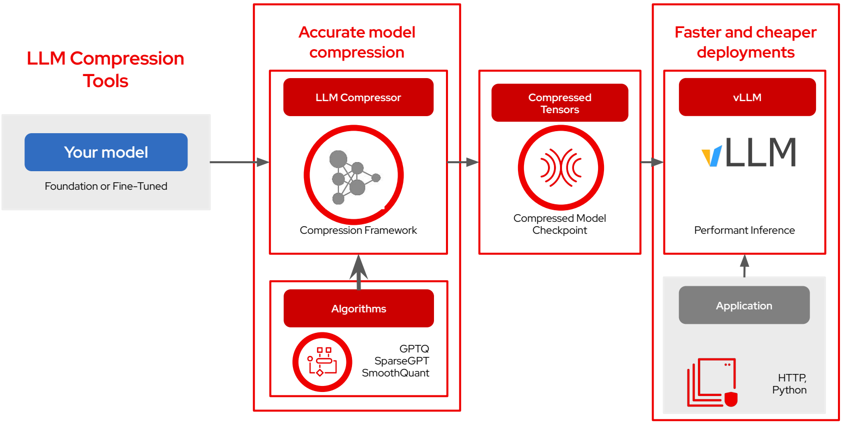
Fast, memory-efficient, production-grade LLM serving for GenAI workloads



Llama
Qwen
DeepSeek
Gemma
Mistral
Molmo
Phi
Nemotron
Granite

vLLM

GPU
Instinct
TPU
Neuron
Gaudi
Spyre



LLM Compression Tools

Your model
Foundation or Fine-Tuned

Accurate model compression

LLM Compressor
Compression Framework

Algorithms
GPTQ
SparseGPT
SmoothQuant

Compressed Tensors
Compressed Model Checkpoint

Faster and cheaper deployments

vLLM
vLLM
Performant Inference

Application
HTTP, Python

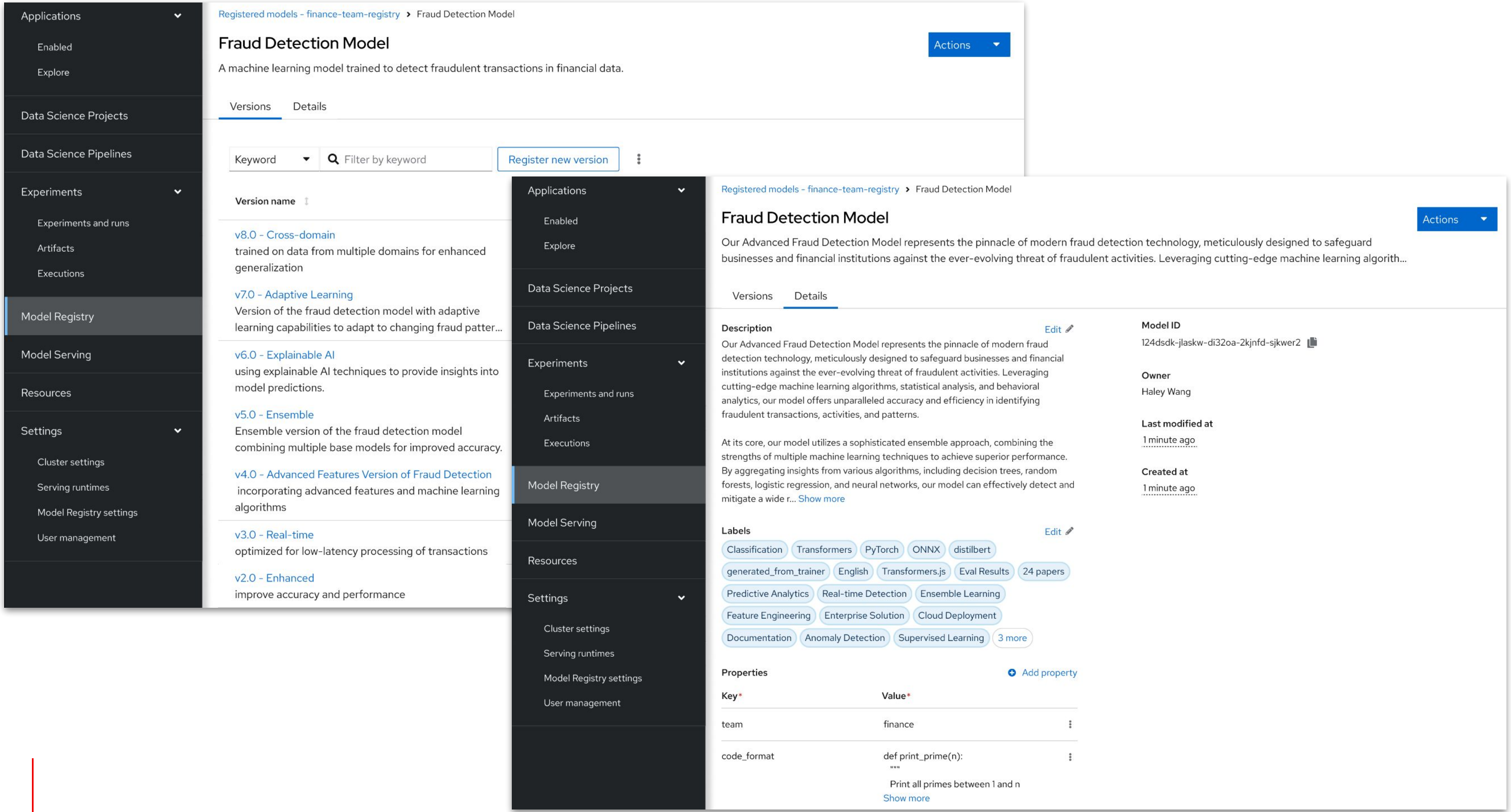
vLLM Inference Server enables efficient, production-grade LLM serving while *significantly reducing infrastructure and compute costs*—making it ideal for real-time, scalable AI applications across enterprises and research environments.

Key Features:

- **Faster response time:** Higher throughput handles more requests in less time for faster responses.
- **Efficient memory management:** Organized memory enables larger models a on existing hardware.
- **Reduced hardware costs:** Efficient resource utilization reduces required GPUs and overall infrastructure costs.
- **Designed for security and scale:** Self-hosting with vLLM strengthens data privacy, control, and scalable growth.

OpenShift AI Model Registry

Establishing model governance from data to deployment



Registered models - finance-team-registry › Fraud Detection Model

Fraud Detection Model

A machine learning model trained to detect fraudulent transactions in financial data.

Version name

v8.0 - Cross-domain
trained on data from multiple domains for enhanced generalization

v7.0 - Adaptive Learning
Version of the fraud detection model with adaptive learning capabilities to adapt to changing fraud patter...

v6.0 - Explainable AI
using explainable AI techniques to provide insights into model predictions.

v5.0 - Ensemble
Ensemble version of the fraud detection model combining multiple base models for improved accuracy.

v4.0 - Advanced Features Version of Fraud Detection
incorporating advanced features and machine learning algorithms

v3.0 - Real-time
optimized for low-latency processing of transactions

v2.0 - Enhanced
improve accuracy and performance

Registered models - finance-team-registry › Fraud Detection Model

Fraud Detection Model

Our Advanced Fraud Detection Model represents the pinnacle of modern fraud detection technology, meticulously designed to safeguard businesses and financial institutions against the ever-evolving threat of fraudulent activities. Leveraging cutting-edge machine learning algorith...

Description Edit

Our Advanced Fraud Detection Model represents the pinnacle of modern fraud detection technology, meticulously designed to safeguard businesses and financial institutions against the ever-evolving threat of fraudulent activities. Leveraging cutting-edge machine learning algorithms, statistical analysis, and behavioral analytics, our model offers unparalleled accuracy and efficiency in identifying fraudulent transactions, activities, and patterns.

At its core, our model utilizes a sophisticated ensemble approach, combining the strengths of multiple machine learning techniques to achieve superior performance. By aggregating insights from various algorithms, including decision trees, random forests, logistic regression, and neural networks, our model can effectively detect and mitigate a wide r... Show more

Labels Edit

Classification · Transformers · PyTorch · ONNX · distilbert · generated_from_trainer · English · Transformers.js · Eval Results · 24 papers · Predictive Analytics · Real-time Detection · Ensemble Learning · Feature Engineering · Enterprise Solution · Cloud Deployment · Documentation · Anomaly Detection · Supervised Learning · 3 more

Properties Add property

Key*	Value*
team	finance
code_format	def print_prime(n): """ Print all primes between 1 and n Show more

Model ID
124dsdk-jlaskw-di32oa-2kjnfd-sjkwer2

Owner
Haley Wang

Last modified at
1 minute ago

Created at
1 minute ago

A centralized AI/ML model registry *that tracks models from registration through deployment*, enabling collaboration, version control, governance, and streamlined MLOps workflows.

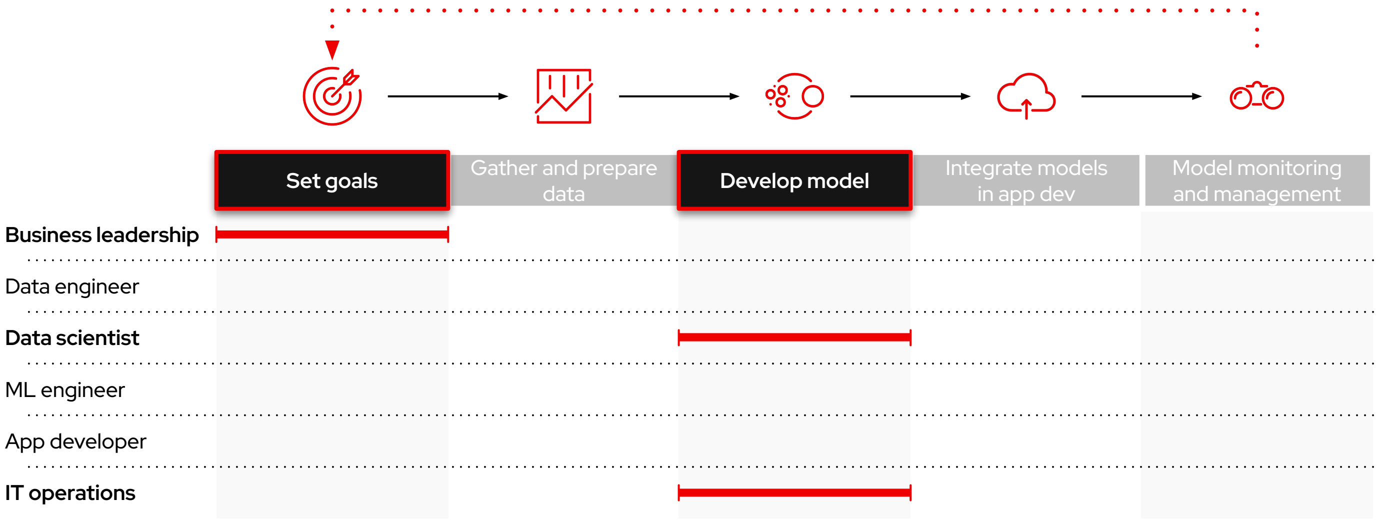
Key Features:

- Central hub: stores models, parameters, metrics, and deployment events
- Model versioning: register models and versions with object storage
- Collaboration & governance: role-based access control

DEMO

Early Role Alignment for AI Trial

Introducing core teams to AI through goal-setting and hands-on evaluation of pre-trained models



	Set goals	Gather and prepare data	Develop model	Integrate models in app dev	Model monitoring and management
Business leadership	▬				
Data engineer					
Data scientist			▬		
ML engineer					
App developer					
IT operations			▬		

- **Business leadership** sets initial goals and success criteria
- **Data scientists** evaluate pre-trained models to test feasibility and value.
- **IT operations** enables secure access and controlled infrastructure

Laying the Groundwork for Enterprise AI

Establishing an AI/ML foundation with secure infrastructure, and early alignment across teams

Stage: Trial - Outcomes:

- **Platform Enablement**

  ✅ *Foundational infrastructure is operational* (GPU provisioning, platform metrics).
  ✅ *Secure access to models and usage controls is enabled via curated catalog and secure APIs.*
  ✅ *As-a-Service capabilities activated* (GPU, MaaS) to accelerate experimentation.

- **Team Enablement**

  ✅ *Cross-functional alignment* between platform teams, data scientists, and business stakeholders.
  ✅ *AI roles and responsibilities clarified* across cross-functional teams.
  ✅ *Teams gain first hands-on AI experience* in a controlled environment.

- **Business Readiness**

  ✅ *Early success criteria* and AI use cases align with business goals.
  ✅ *Clarity on AI value* through trial workloads using pre-trained models.
  ✅ *A governance foundation* is established to prevent “shadow AI” and support safe experimentation.

Stage: Experiment *Fine-tune Models*

Strategic Goal

Enable experimentation with real enterprise data. *Focus on building collaborative environments, customizing models, and defining repeatable workflows* to support scalable and secure AI development.

Why This Stage Matters

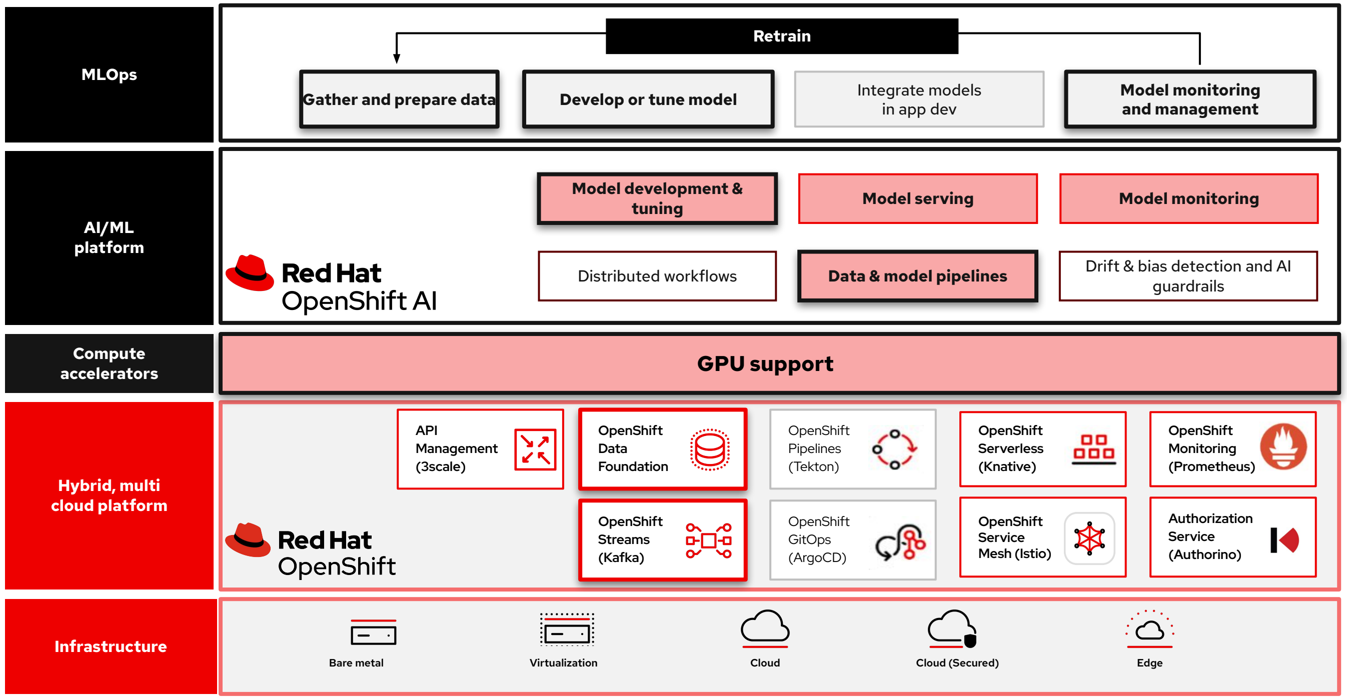
- *Introduces enterprise data*
- *Promotes collaboration*
- *Enables model customization*

Enabling Collaborative AI Experimentation

Focus on establishing secure, observable, and scalable infrastructure to support early experimentation

Platform Objectives:

- Ingest, store, and prepare enterprise data
- Enable collaborative development environments
- Support model training for differentiated AI



MLOps

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 & tuning

Model serving

Model monitoring

Distributed workflows

Data & model pipelines

Drift & bias detection and AI guardrails

Compute accelerators

GPU support

Hybrid, multi cloud platform

Red Hat OpenShift

API Management (3scale)

OpenShift Data Foundation

OpenShift Pipelines (Tekton)

OpenShift Serverless (Knative)

OpenShift Monitoring (Prometheus)

OpenShift Streams (Kafka)

OpenShift GitOps (ArgoCD)

OpenShift Service Mesh (Istio)

Authorization Service (Authorino)

Infrastructure

Bare metal

Virtualization

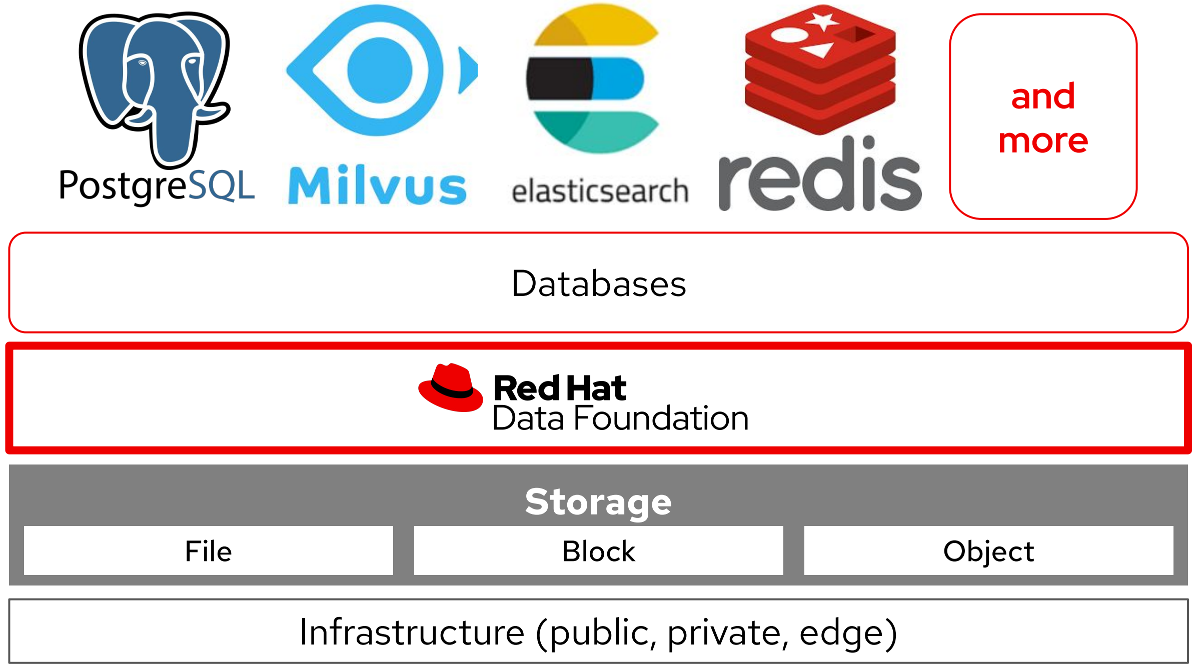
Cloud

Cloud (Secured)

Edge

OpenShift Data Foundation and OpenShift Operators

Scalable, resilient storage for AI/ML workloads with integrated vector database support



and more

Databases

Red Hat Data Foundation

Storage

File

Block

Object

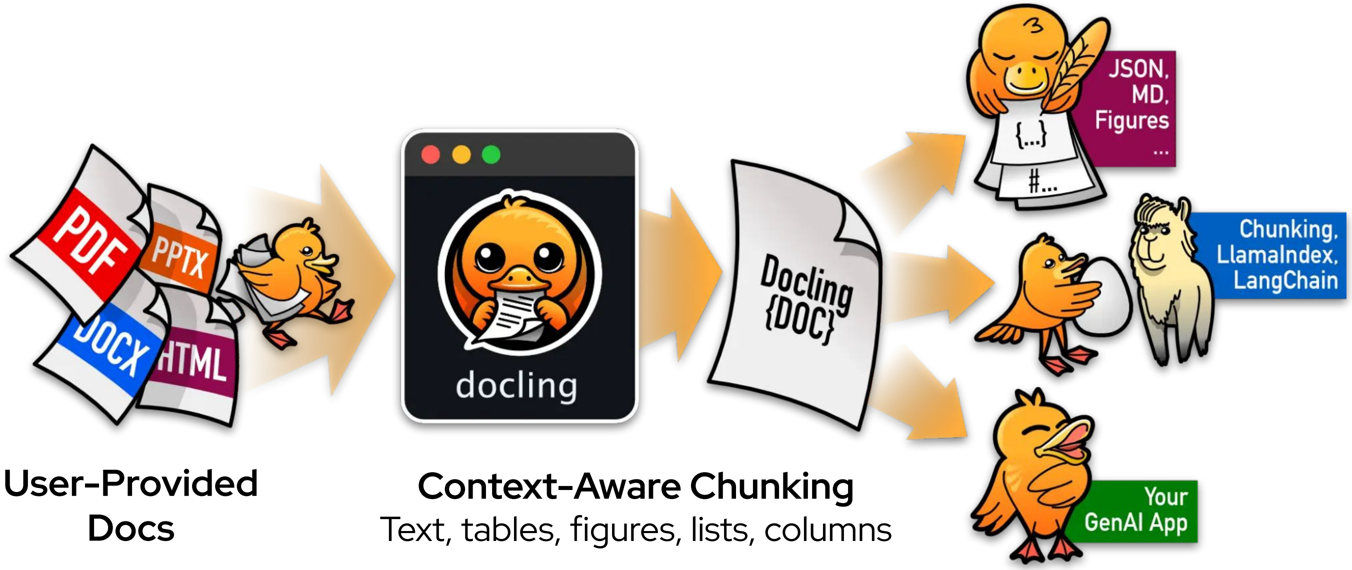
Infrastructure (public, private, edge)

OpenShift Data Foundation provides the *scalable, resilient storage backbone for AI/ML workloads*, supporting object, block, and file storage. Combined with OpenShift Operators, it *enables seamless deployment and lifecycle management of vector databases* like:

- PostgreSQL: Stores structured data, model metadata, and vectors via `pgvector`.
- Milvus (Vector DB): High-performance embedding storage and vector search in RAG pipelines.
- Elasticsearch: Hybrid text and vector search for RAG and large-scale indexing.
- Redis: Low-latency feature serving, caching, and vector search for real-time inference workloads.

Docling

Automated document transformation for GenAI and retrieval-augmented generation (RAG)



User-Provided Docs

Context-Aware Chunking
Text, tables, figures, lists, columns

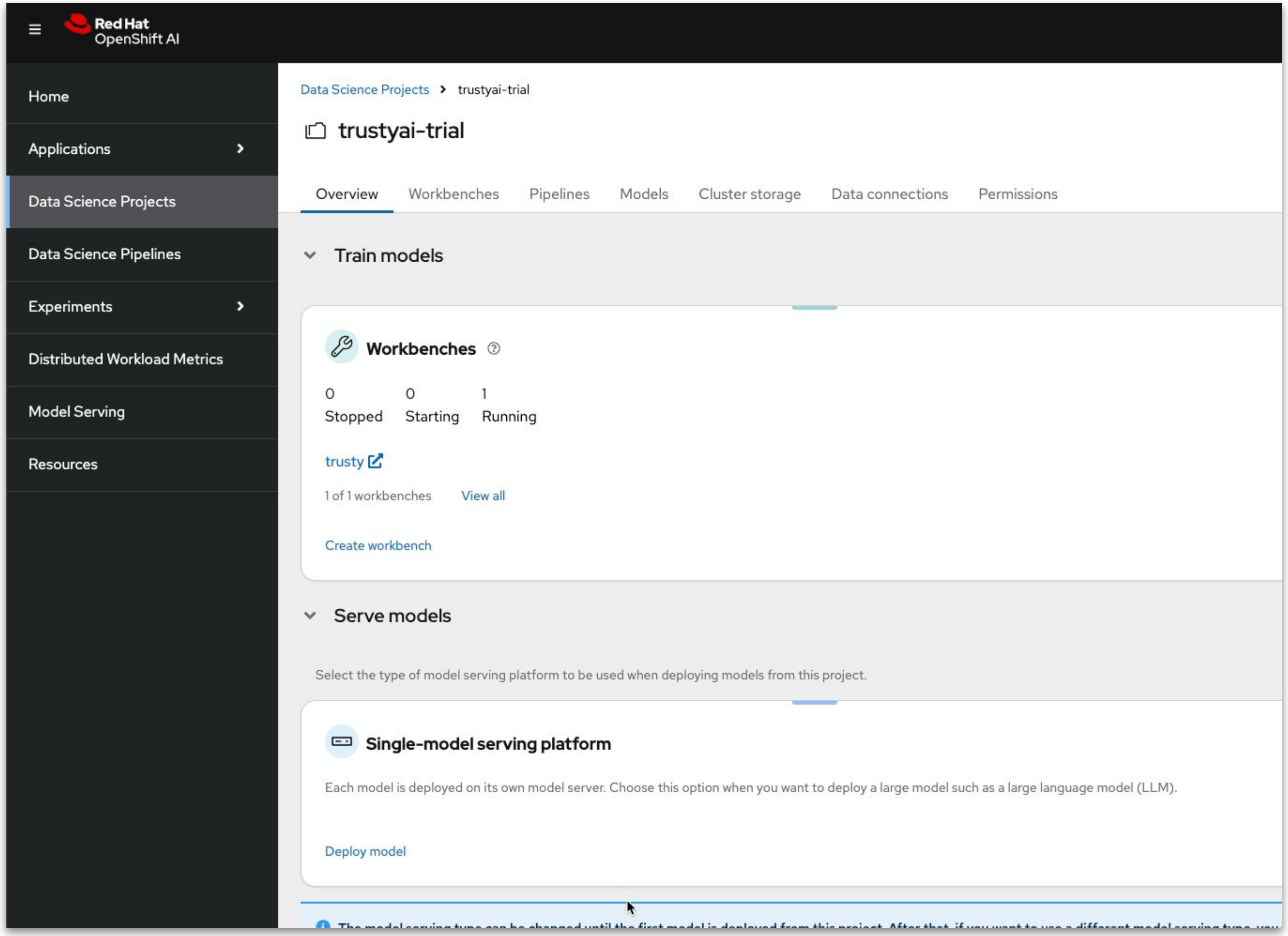
Docling simplifies document processing by *converting complex files into structured, AI-ready formats*—making it easier to use documents and audio in search, RAG, and analytics workflows.

Key Features:

- Multi-format parsing: supports PDF, DOCX, PPTX, HTML —even scanned or layout-complex formats
- Advanced layout & table extraction: leverages DocLayNet and TableFormer models for accurate structure, formulas, code, and tables
- Built-in Automatic Speech Recognition: transcribes audio (WAV/MP3) into structured text

JupyterLab on OpenShift AI

Web-based notebooks for scalable, containerized AI development



Select the type of model serving platform to be used when deploying models from this project.

Each model is deployed on its own model server. Choose this option when you want to deploy a large model such as a large language model (LLM).

The model serving type can be changed until the first model is deployed from this project. After that, if you want to use a different model serving type, you

JupyterLab delivers a web-based notebook IDE running on Kubernetes—*empowering teams to develop, train, and deploy AI workflows* with cluster-grade compute, storage, and data integrations.

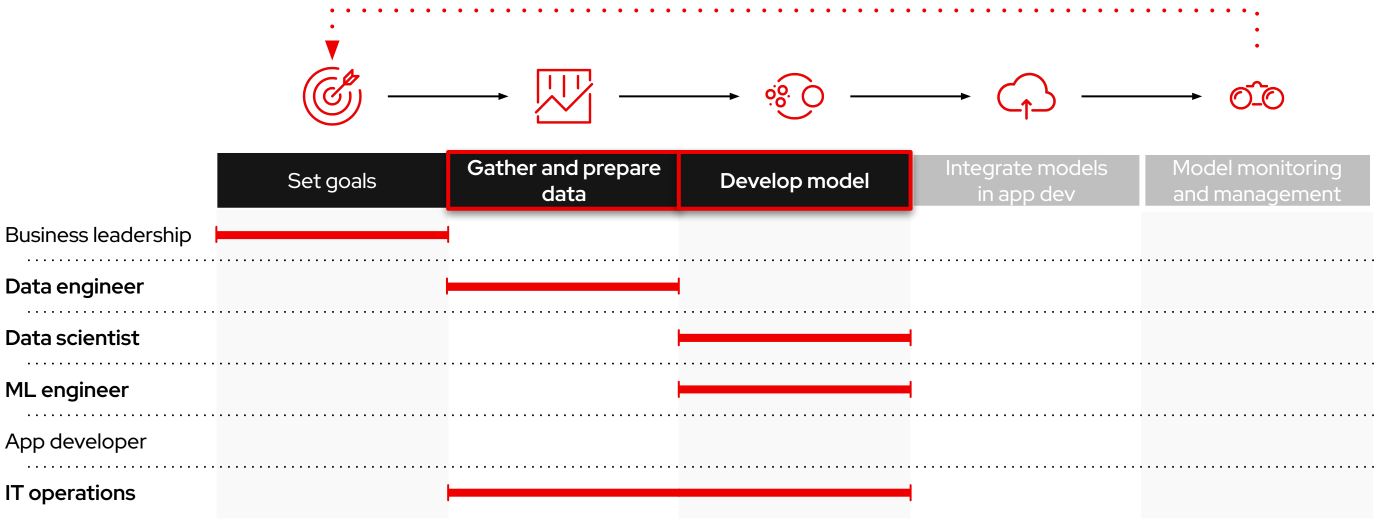
Key Features:

- Managed notebook server: Launch JupyterLab containerised environments for secure and scalable development.
- Integrated IDE options: Choose from JupyterLab, VS Code, or RStudio—all containerized within projects.
- Persistent storage & data access: Attach cluster or S3-compatible external data sources seamlessly.
- Cluster-powered compute: Run code on powerful CPUs/GPUs in the cluster.

DEMO

Role Expansion for AI Experimentation

Enabling hands-on collaboration with enterprise data, pipelines, and model customization



	Set goals	Gather and prepare data	Develop model	Integrate models in app dev	Model monitoring and management
Business leadership	━				
Data engineer		━			
Data scientist			━		
ML engineer			━		
App developer					
IT operations		━	━		

- **Data engineers** build pipelines and prepare data for modeling.
- **Data scientists** customize and evaluate models with enterprise context data.
- **ML engineers** fine-tune models and manage experimentation workflows.
- **IT operations** maintains environment and storage access

Accelerating Collaborative AI Development

Establishing scalable, secure, and reusable foundations for experimentation and model customization

Stage: Experiment – Outcomes:

- **Platform Enablement**

  ✅ *Enterprise data pipelines operational* (ingestion, storage, reuse).
  ✅ *Training-ready environments available* to support model fine-tuning.
  ✅ *As-a-Service capabilities enabled* (Data as a Service, Data Science as a Service) to streamline experimentation.

- **Team Enablement**

  ✅ *Cross-functional collaboration extends* across data engineers, scientists, and platform teams.
  ✅ *Shared environments in place* for collaborative development.
  ✅ *Automated workflows in use* for reproducibility and governance.

- **Business Readiness**

  ✅ *Real enterprise data in use* to test business-aligned use cases.
  ✅ *Differentiated AI development enabled* via model fine-tuning and customization.
  ✅ *Compliance controls introduced* to support scale and secure use.

Stage 3: Adopt
Scale & Integrate

Strategic Goal

Operationalize AI at scale across the enterprise. *Focus on automating model delivery, embedding AI into business applications, and ensuring trust, governance, and observability* in production.

Why This Stage Matters

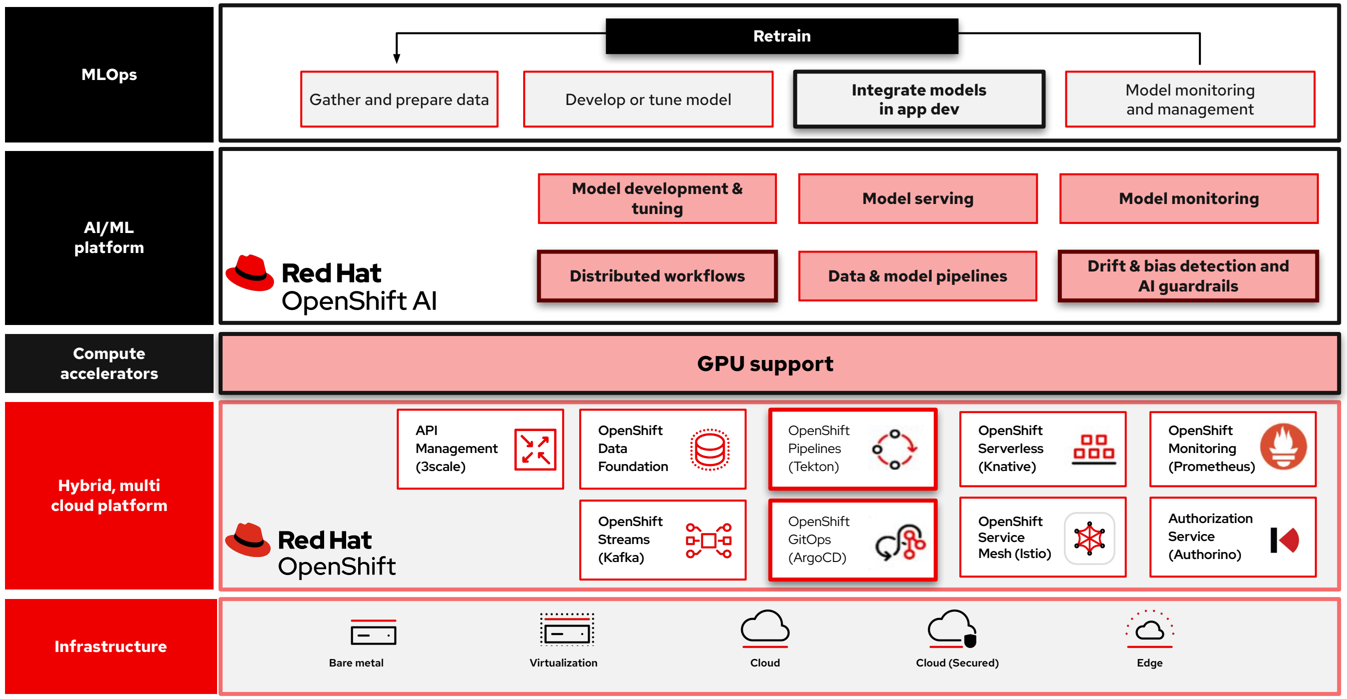
- *Integrates AI*
- *Establishes trust*
- *Sustainable AI operations*

Enabling Scalable and Trusted AI Integration

Focus on automating delivery, ensuring trust, and embedding AI into enterprise applications

Platform Objectives:

- Embed trust, observability, and accountability
- Integrate AI models into enterprise applications
- Automate model delivery workflows



MLOps

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 & tuning

Model serving

Model monitoring

Distributed workflows

Data & model pipelines

Drift & bias detection and AI guardrails

Compute accelerators

GPU support

Hybrid, multi cloud platform

Red Hat OpenShift

API Management (3scale)

OpenShift Data Foundation

OpenShift Pipelines (Tekton)

OpenShift Serverless (Knative)

OpenShift Monitoring (Prometheus)

OpenShift Streams (Kafka)

OpenShift GitOps (ArgoCD)

OpenShift Service Mesh (Istio)

Authorization Service (Authorino)

Infrastructure

Bare metal

Virtualization

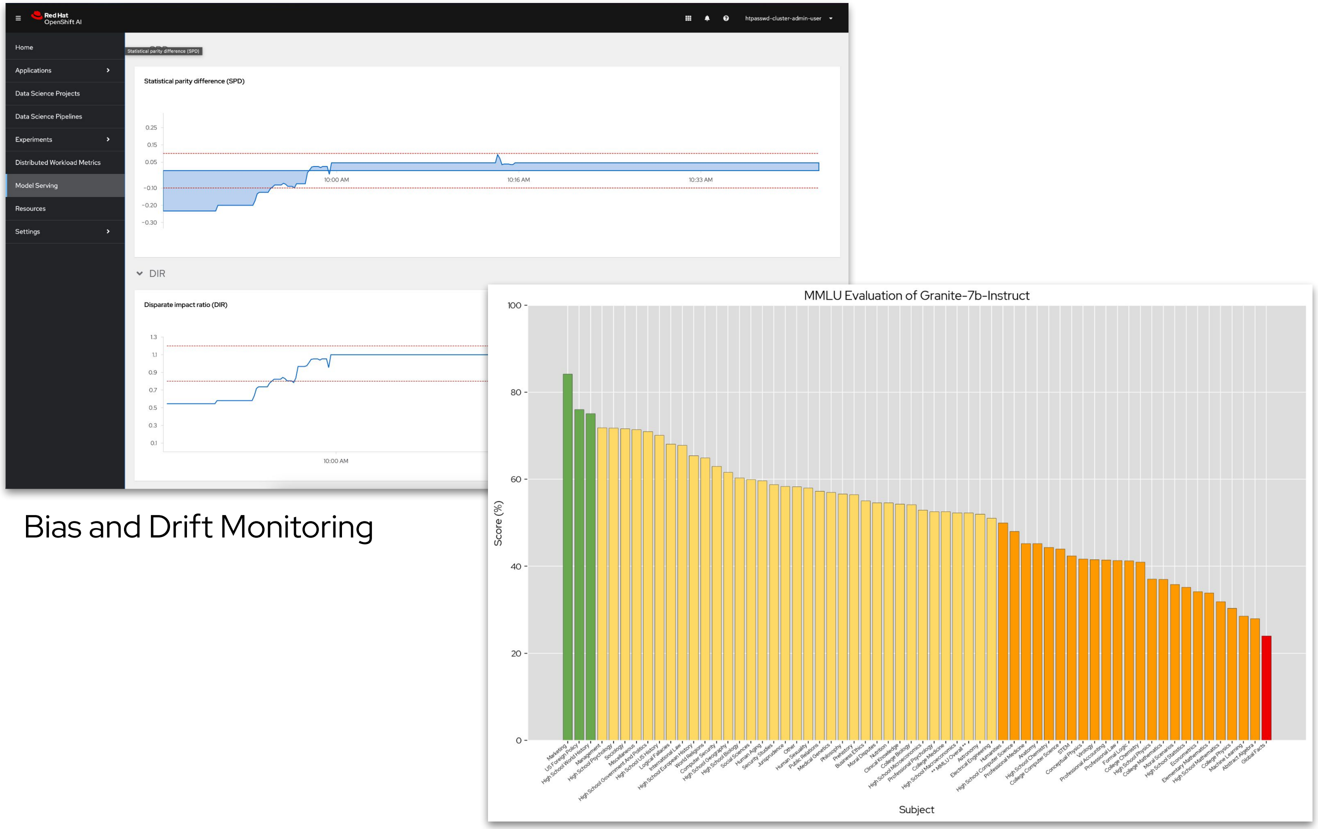
Cloud

Cloud (Secured)

Edge

TrustyAI on OpenShift AI

Automate monitoring, fairness checks, and safety enforcement for responsible AI



Bias and Drift Monitoring

Language Model Evaluation

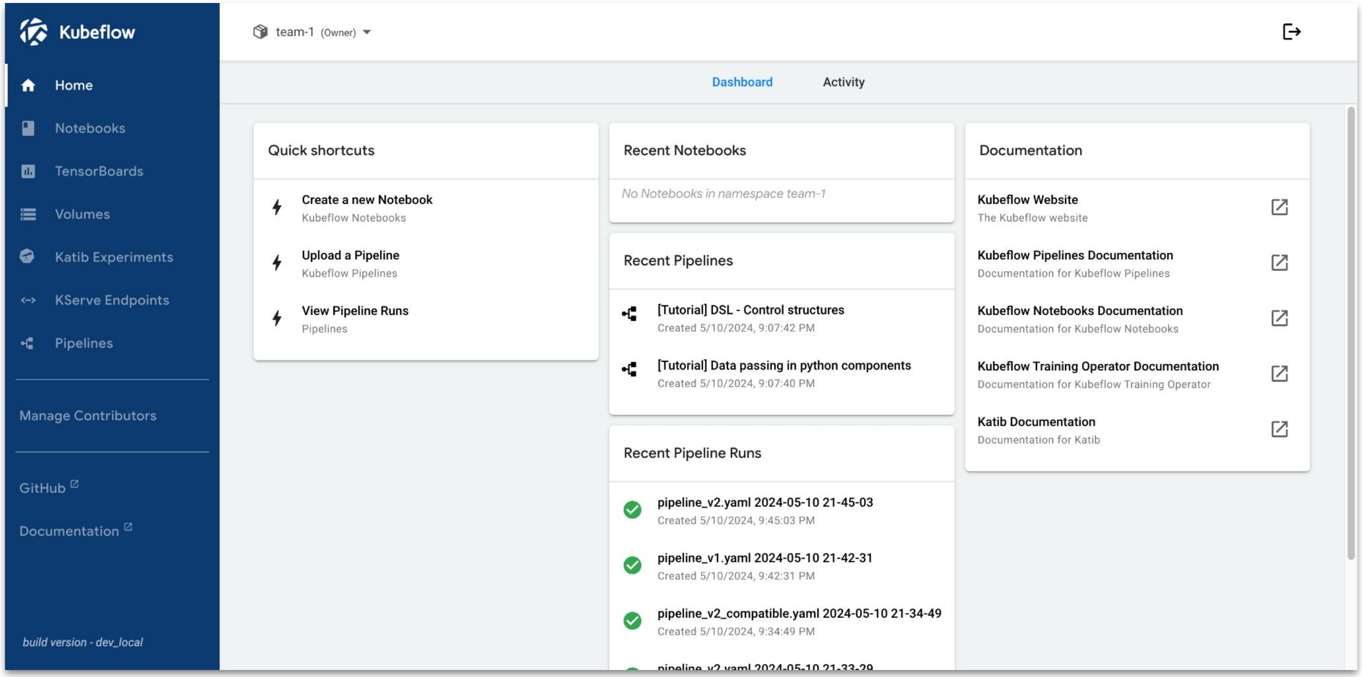
TrustyAI a comprehensive toolkit for responsible AI offering explainability, fairness, drift monitoring, and guardrails. Integrated via an Operator, it *empowers teams to deploy transparent, auditable models with automated monitoring and safety enforcement* as part of their MLOps pipelines.

Key Features:

- Explainability built-in: Offers local and global model explainers (e.g., LIME, SHAP, counterfactuals) to interpret predictions.
- Fairness & drift detection: Continuously monitors bias and data drift metrics in production.
- AI guardrails: Deploys modular detectors (e.g., profanity, prompt injection) to enforce safety and policy compliance on LLM inputs/outputs.

Kubeflow on OpenShift AI

Run scalable, reproducible ML workflows across hybrid environments



Quick shortcuts

Create a new Notebook
Kubeflow Notebooks

Upload a Pipeline
Kubeflow Pipelines

View Pipeline Runs
Pipelines

Recent Notebooks

No Notebooks in namespace team-1

Recent Pipelines

[Tutorial] DSL - Control structures
Created 5/10/2024, 9:07:42 PM

[Tutorial] Data passing in python components
Created 5/10/2024, 9:07:40 PM

Recent Pipeline Runs

pipeline_v2.yaml 2024-05-10 21-45-03
Created 5/10/2024, 9:45:03 PM

pipeline_v1.yaml 2024-05-10 21-42-31
Created 5/10/2024, 9:42:31 PM

pipeline_v2_compatible.yaml 2024-05-10 21-34-49
Created 5/10/2024, 9:34:49 PM

pipeline_v2.yaml 2024-05-10 21-33-29

Documentation

Kubeflow Website
The Kubeflow website

Kubeflow Pipelines Documentation
Documentation for Kubeflow Pipelines

Kubeflow Notebooks Documentation
Documentation for Kubeflow Notebooks

Kubeflow Training Operator Documentation
Documentation for Kubeflow Training Operator

Katib Documentation
Documentation for Katib

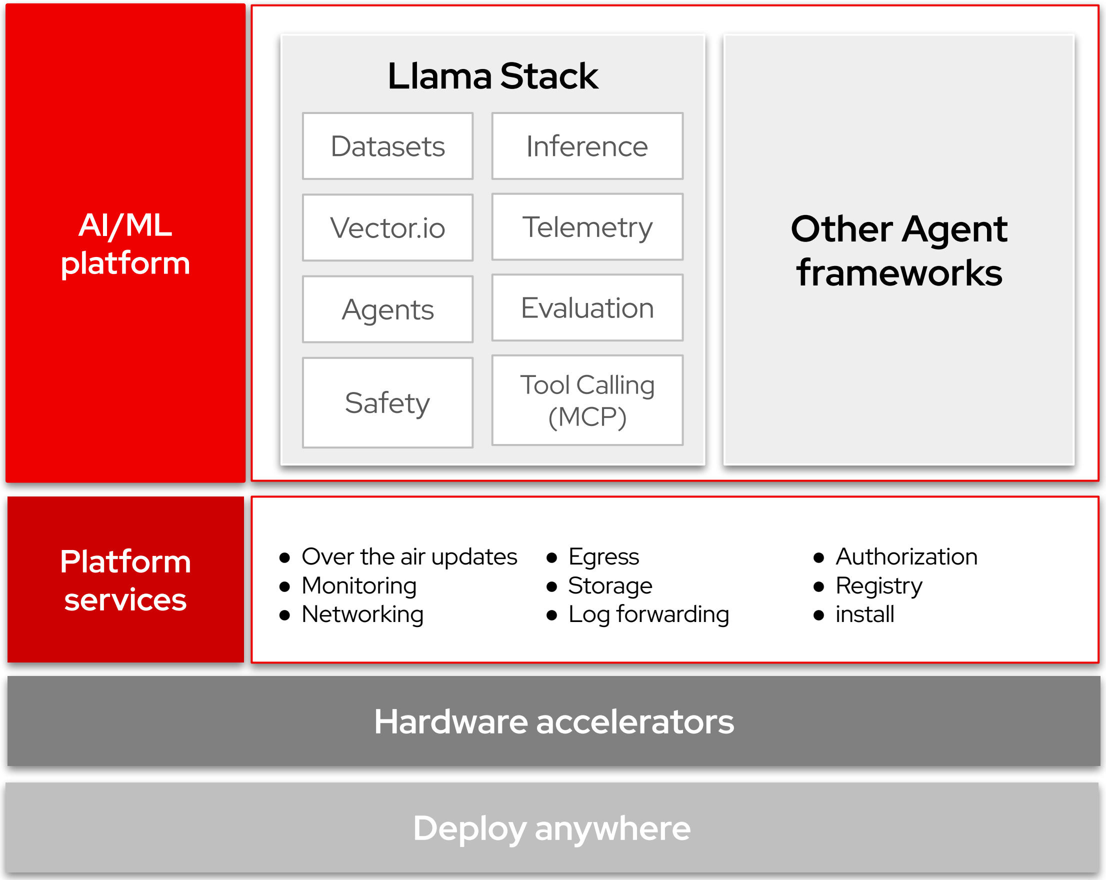
Kubeflow a *Kubernetes-native MLOps platform that supports the full AI lifecycle*—from model development to distributed training and serving. It offers multi-tenancy, GPU autoscaling –*empowering AI/ML teams to build scalable, composable, and cloud-portable workflows*.

Key Features:

- Modular pipeline: Includes Kubeflow Pipelines for building, scheduling, and managing reproducible ML workflows.
- Distributed training operators: Scale training jobs with built-in support for TensorFlow (TFJob), PyTorch (PyTorchJob), and MPI-based workloads.
- Hyperparameter tuning (Katib): Automate model optimization using scalable search algorithms across distributed training jobs.

Llama Stack on OpenShift AI

Standardize and accelerate AI agent development with modular components



Llama Stack a modular, *agent-capable AI framework* that combines LLM inference, RAG, and tool integration. It *simplifies building intelligent agents by offering standardized APIs* and seamless orchestration—*accelerating AI application development with agentic workflows* and operational tooling.

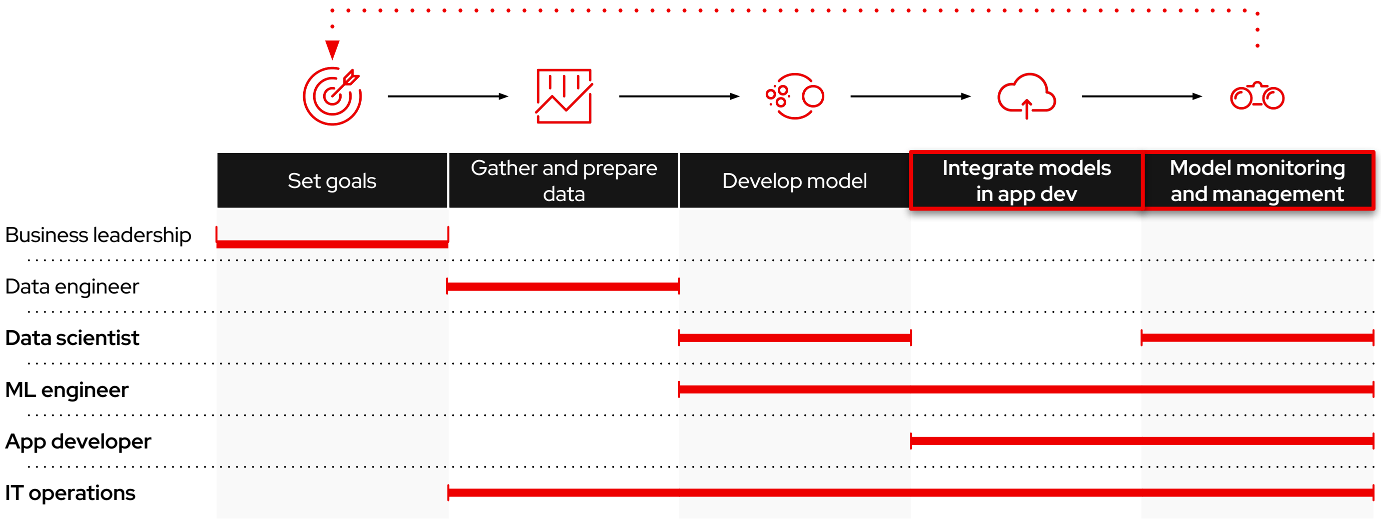
Key Features:

- Unified API: Standard interface for inference, embeddings, RAG, and tool execution.
- MCP-based tool integration: Use the Model Context Protocol to discover and invoke external tools via APIs
- Built-in RAG support: Natively ingests documents and retrieves context using vector-DB backends
- Interactive agents: Build multi-step agents that retrieve, reason, and act with real-time context.

DEMO

Full Role Integration for Scaled AI Delivery

Operationalizing models across apps with CI/CD, monitoring, and business impact in mind



	Set goals	Gather and prepare data	Develop model	Integrate models in app dev	Model monitoring and management
Business leadership	🟥				
Data engineer		🟥			
Data scientist			🟥		🟥
ML engineer			🟥	🟥	🟥
App developer				🟥	🟥
IT operations		🟥	🟥	🟥	🟥

- **Data scientists** validate fairness, explainability, and post-deployment performance.
- **ML engineers** manage MLOps pipelines for model lifecycle, from deployment to rollback.
- **App developers** integrate AI models into production applications using APIs, service mesh, or custom logic.
- **IT operations** deliver secure, compliant, and stable infrastructure for model deployment.

Operationalizing AI Across the Enterprise

Scaling delivery, embedding AI into business systems, and reinforcing trust in production

Stage: Adopt - Outcomes:

- **Platform Enablement**

  ✅ *Production platform fully operational* with automated governance controls.
  ✅ *MLOps as a Service enabled* for model lifecycle management, promotion, and rollback.
  ✅ *AI as a Service enabled* for scalable access and reuse.

- **Team Enablement**

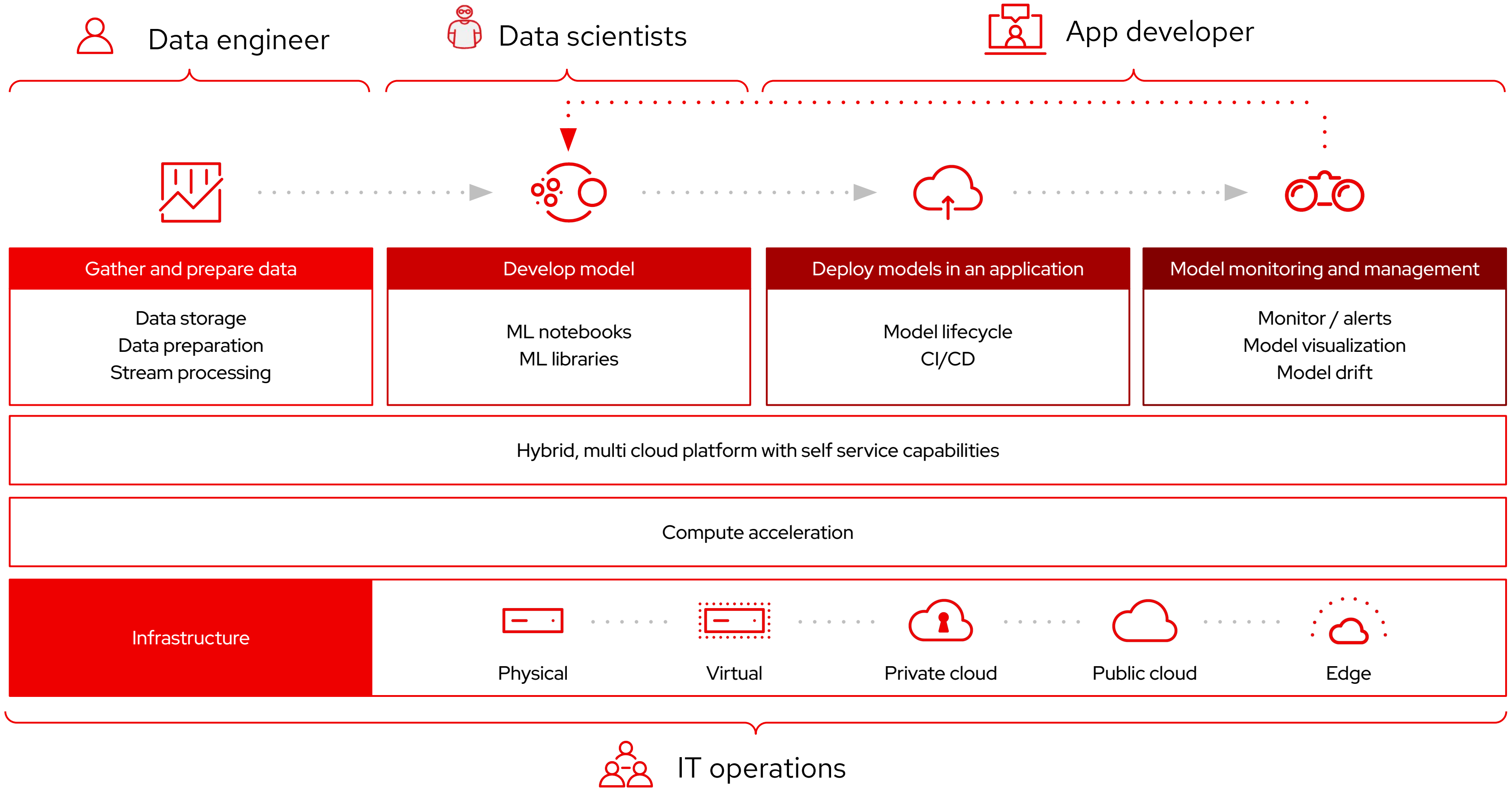
  ✅ *End-to-end collaboration established* from model training to monitoring.
  ✅ *AI knowledge extended* across AI/ML engineers, App developers, and IT operations.
  ✅ *Governed delivery in place* with clear ownership and measurable outcomes.

- **Business Readiness**

  ✅ *AI integrated into business applications* through secure APIs and workflows.
  ✅ *Trust and compliance reinforced* with drift detection, explainability, and fairness checks.
  ✅ *Scalability foundation established* for enterprise-wide AI adoption.

Navigating AI/ML Adoption: What We've Achieved

A milestone-driven journey from safe experimentation to scalable, responsible AI adoption



- Stage 1: Trial

  A foundational platform

- Stage 2: Experiment

  An AI development environment

- Stage 3: Adopt & Scale

  AI models are integrated

Red Hat Summit

Connect

Thank you

Adnan Drina

Specialist Solution Architect
AppDev & AI