YapıKredi Teknoloji

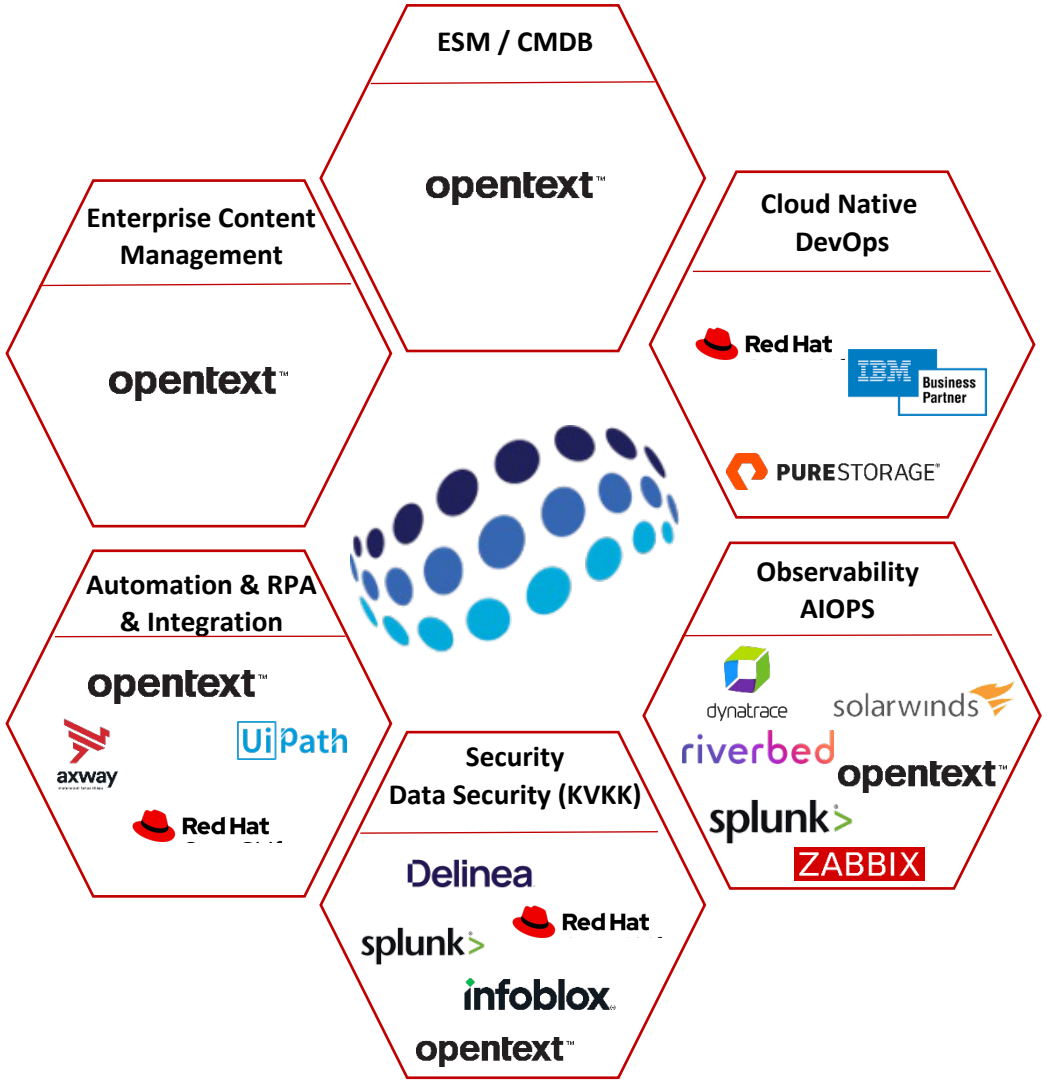
MLOps: Journey Walkthrough with Red Hat OpenShift AI

November’25

YapıKredi

Konsalt

Domains, Partnerships & References (2011-2025)



A “niche” system integrator

40+ Alanında Uzman Danışman

Yurt İçi & Yurt Dışı Proje & Destek / Danışmanlık

E-posta : info@konsalt.com.tr
sales@konsalt.com.tr
Web. : www.konsalt.com.tr

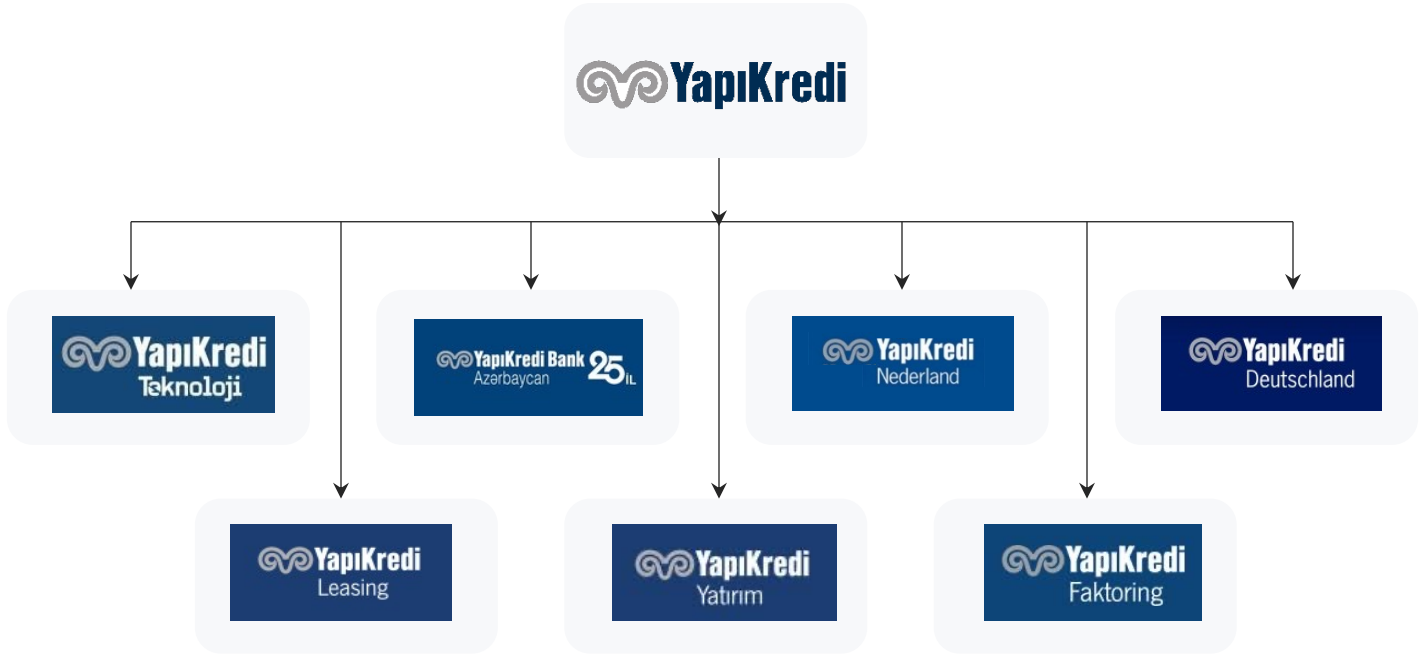
Konsalt

YapıKredi Teknoloji

Yapı Kredi

Yapı Kredi Bank was established in 1944 as **Turkey's first retail focused private bank** with a nationwide presence.

In 2006 two of the most strongest financial organization, Yapı Kredi Bank A.Ş. and Koçbank A.S., merged as one organization under the strong leadership of Koç Financial Services. Yapi Kredi Bank has always played an important role in the development of the domestic economy and has set standards in the Turkish banking sector with many innovative products and services.



The starting point:

“Why we needed a change?”

The Business Hurdles

Shared Resources

Generic user accounts leading to a high-friction environment.

Parallel Project Gridlock

User jobs directly interfering with each other’s performance and stability.

High Operational Overhead

Significant time loast in identifying and resolving environment issues instead of building models.

Delayed Time-to-Market

Critical ML models were not delivered on time.

DevOps Perspective

Manual Reengineering

APIs needed to be written specifically for the models.

Manual Handoff

The model is thrown over the wall, without registry.

CI for Code, not for Models

CI pipeline tests the API, not the model logic, quality, or accuracy.

Chaotic Versioning

No naming conventions nor semantic versioning.

Infrastructure Challenges

Inefficient use of compute and GPU resources.

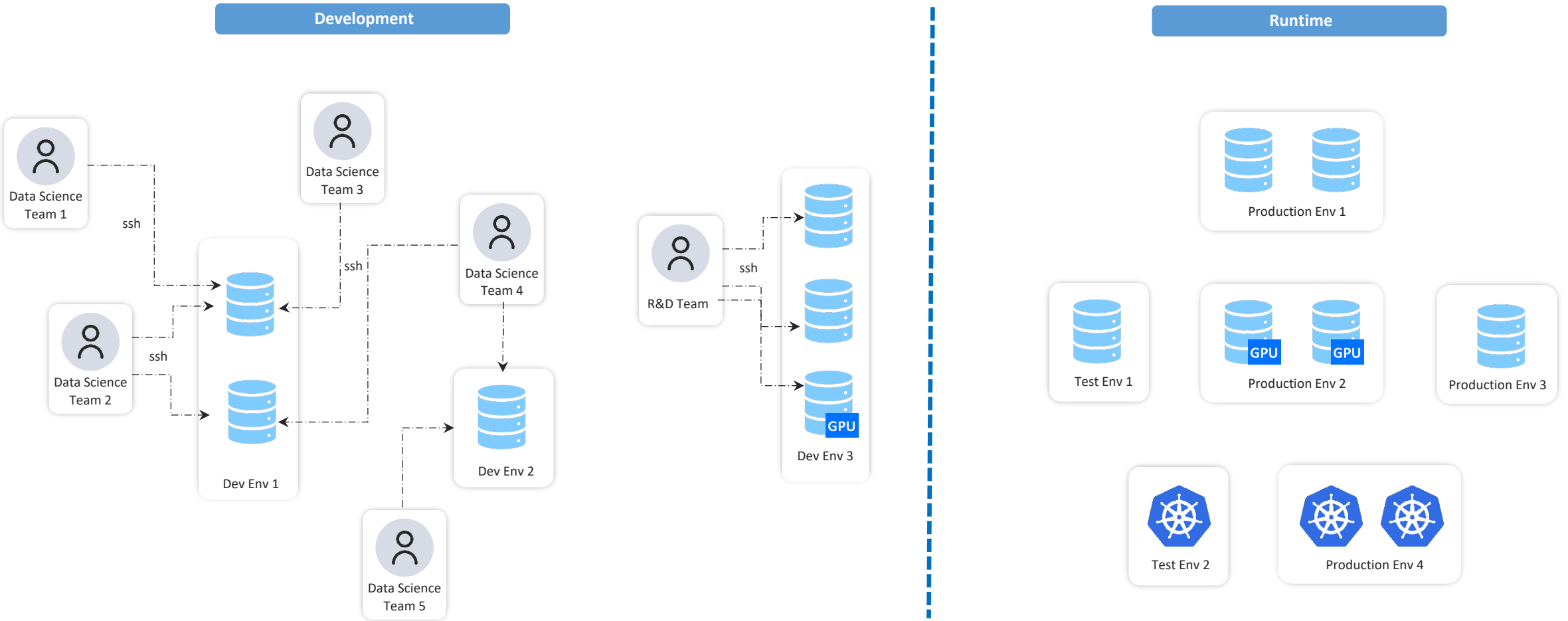
Distributed, hard-to-scale physical/virtual servers.

Complex and incossistent Access control.

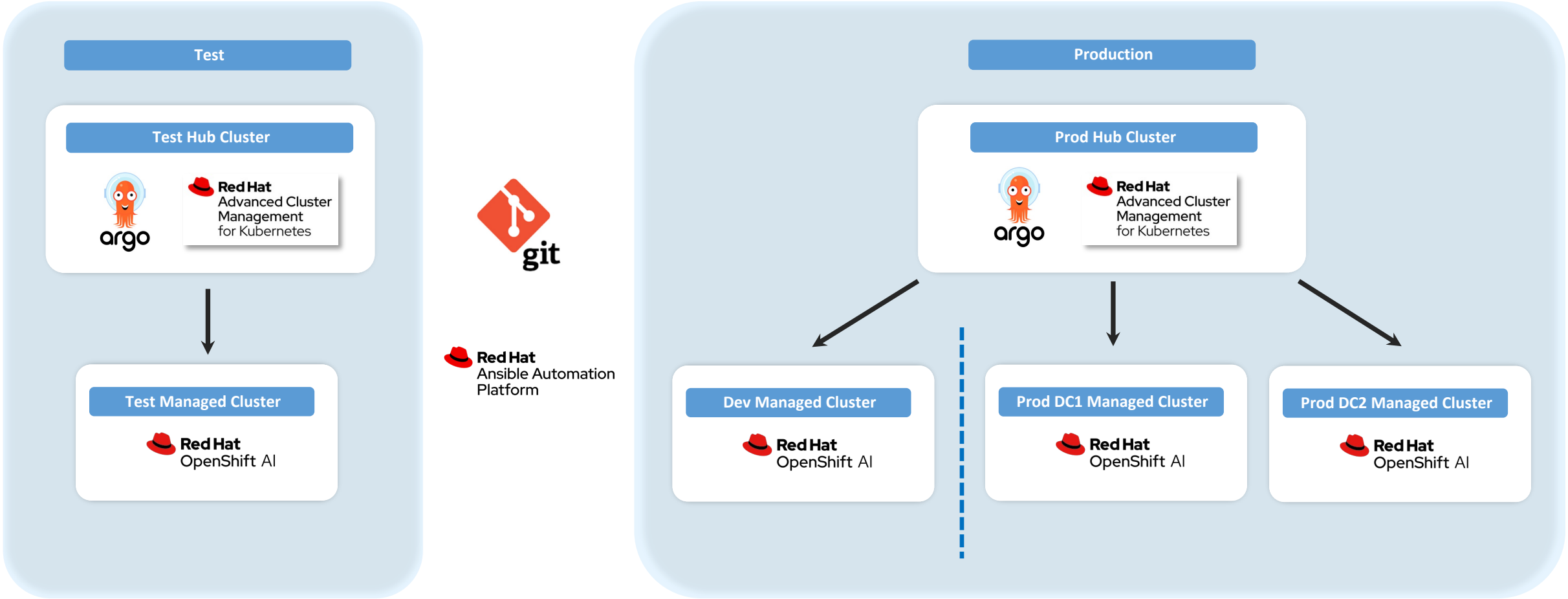
Heavy operational overhead for patching and maintenance.

Unclear ownership of the ML compute environment.

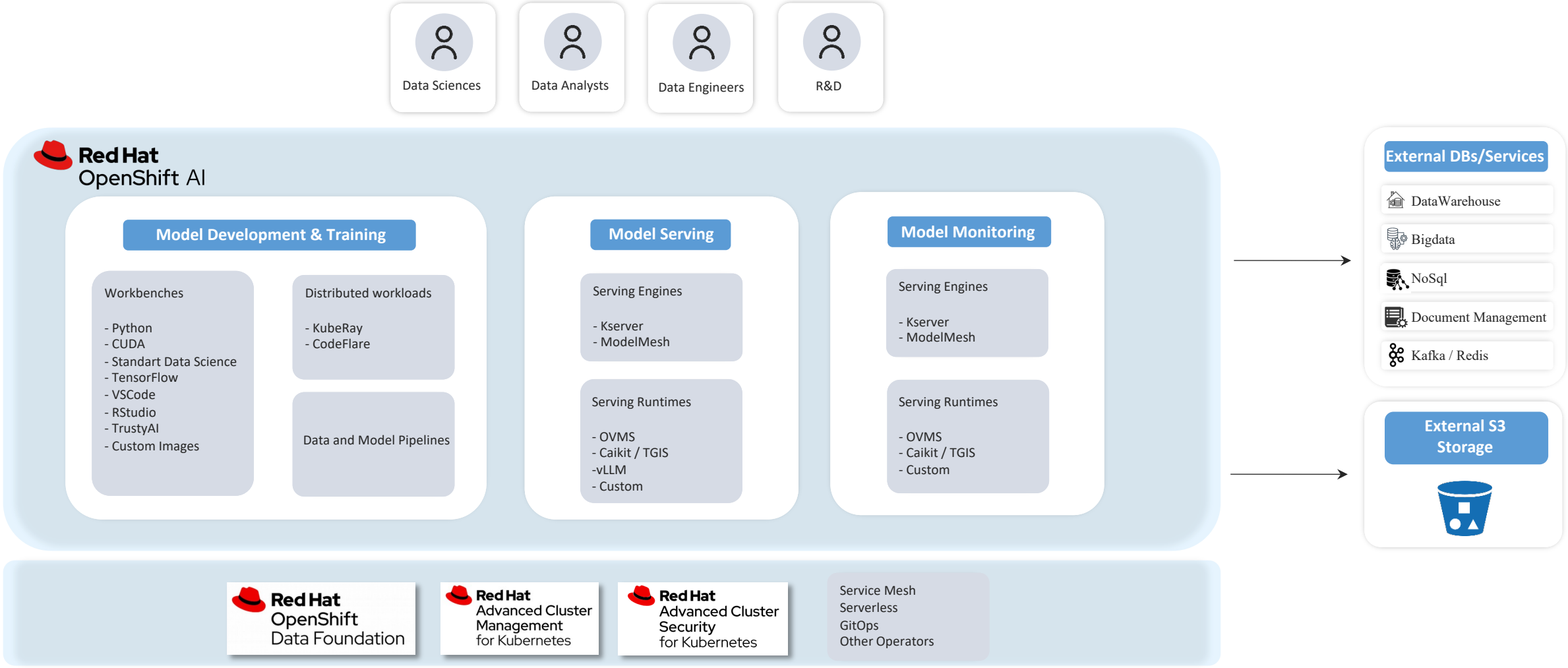
AI Platform: Then



AI Platform: Now / High Level



AI Platform: Now / Development



Onboarding a Team


```yaml
apiVersion: v1
kind: List
items:
  -kind: Namespace
  apiVersion: v1
  metadata:
    name: dsp-TEAMNAME
    labels:
      type: dsp
      modelmesh-enabled: 'false'
      opendatahub.io/dashboard: 'true'
      managedBy: sdlc
      size: medium
      sizeStorage: xsmall
      gpu: enable
      rbacEdit: cloud-ocp-TEAMNAME
      egress: enable
      networkPolicy: default
    annotations:
      egressips: |
        - EGRESSIP1
        - EGRESSIP2
```

Key Concept: One Manifest to Rule Them All

Goal

Streamline team onboarding to Openshift AI using a single, declarative approach

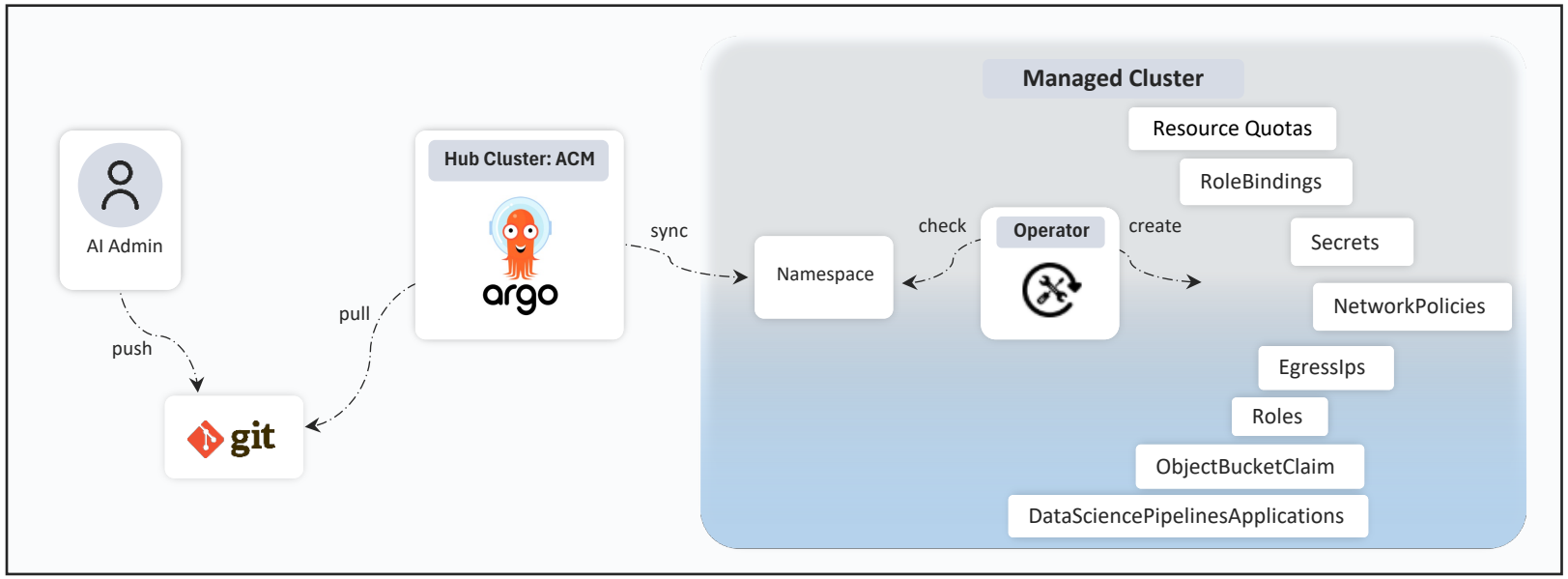
Method

A single Namespace Kubernetes manifest is applied

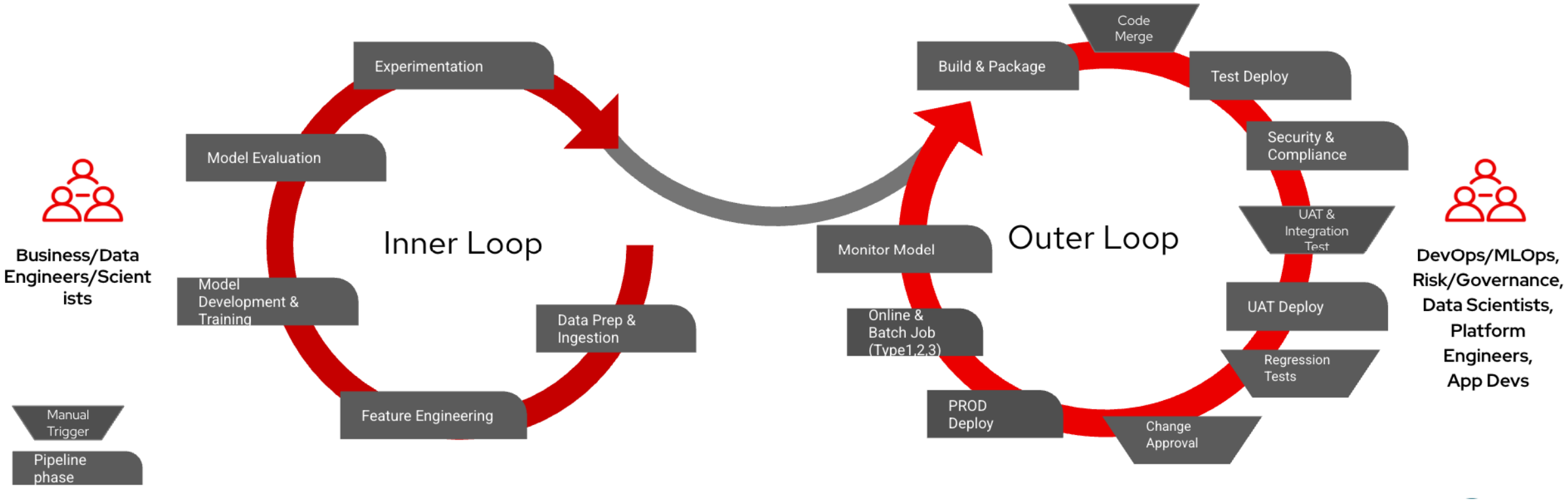
Automation Engine

The Namespace Operator handles the heavy lifting, driven purely by using config templates.

Manifest Breakdown: Labels & Annotations



YKT Inner and Outer Loops



The Foundation: Secure Workbenches

Build and Supply: Custom Workbench Preparation

The target : Building Air-gapped MLOps Pipelines on OpenShift AI.

Secure Workbenches: The Data Scientist’s ultimate launchpad.

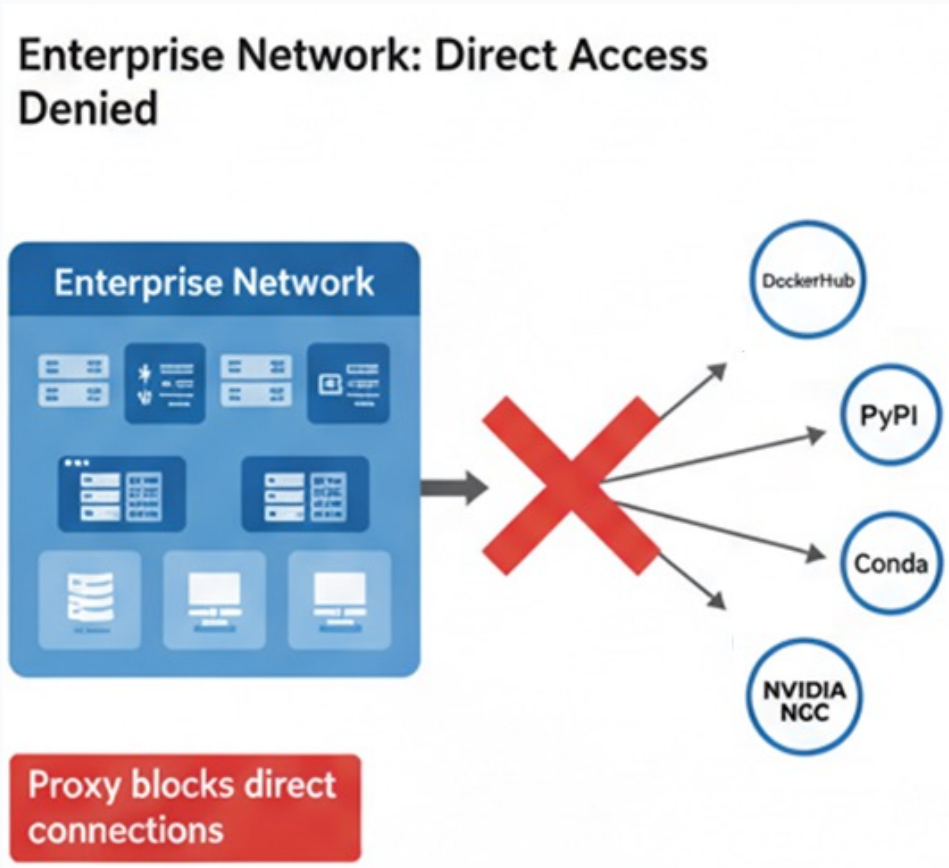
The Challenge: The Private Registry Constraint

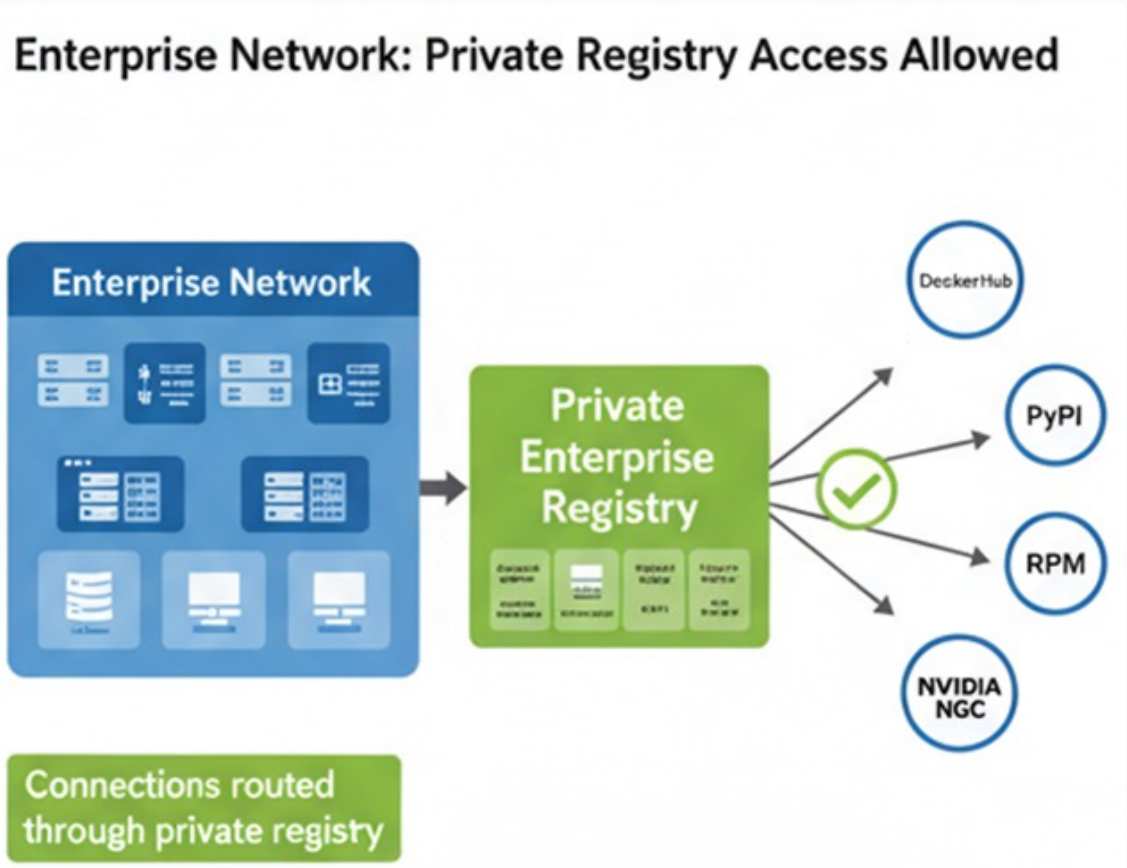
The Enterprise Reality: No public pulls are allowed.

Problem: Projects require specific tools, but enterprise security mandates using private registries.

Solution: A centralized Base Image Factory that provides pre-vetted, security-scanned, and *custom-built* **workbench** and **serving runtime** images.

The Challenge: The Private Registry Constraint





From Inner to Outer Loop

The MLOps Glue: CI/CD and GitOps

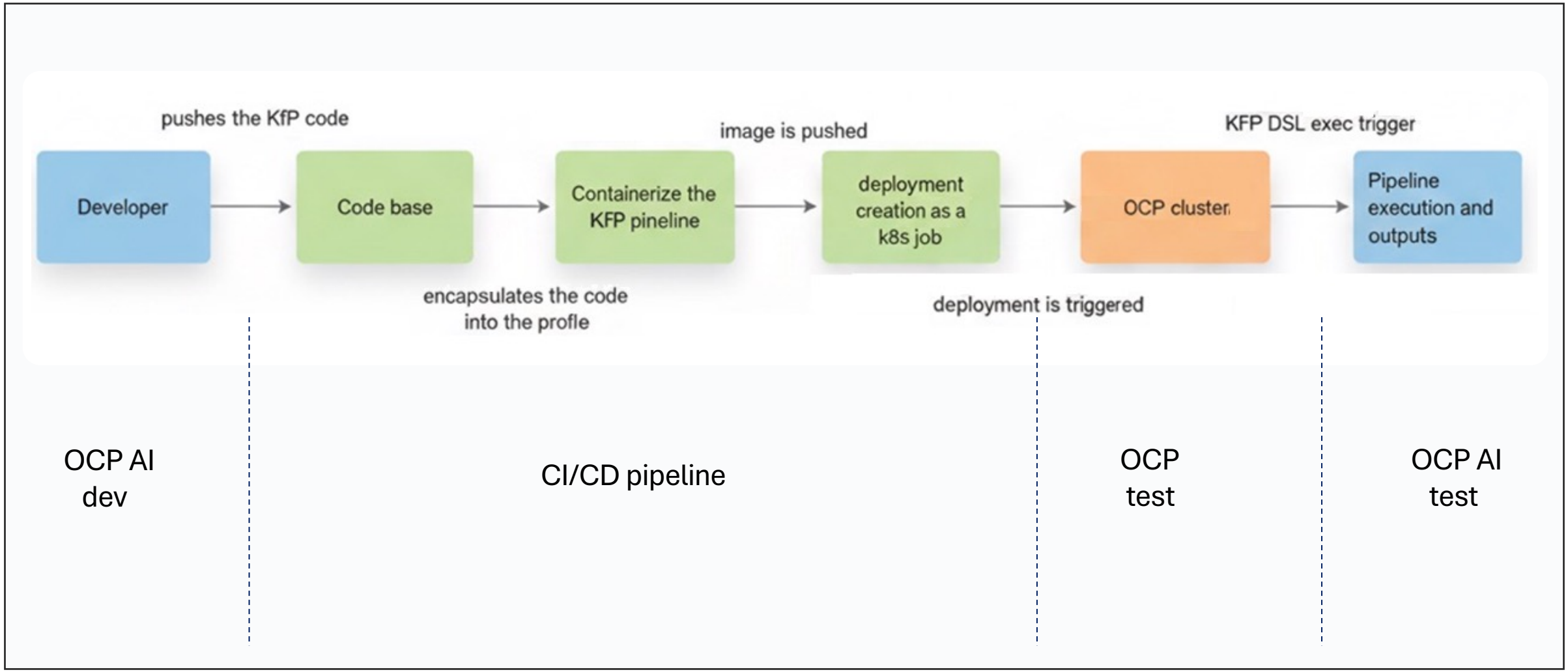
CI of Image Factory : A pipeline with security scans, tag, and push all custom images to the private registry.

CD via GitOps : Whenever a new *Workbench* or *Serving Runtime* image produced or updated, the platform is populated with them in GitOps manner.

CI/CD Pipeline of Kubeflow Pipelines

Pipeline in the Runtime: Passes through SDLC processes, deployed and executed back in OCP AI

CI/CD Pipeline of Kubeflow Pipelines



To Prod via MLOps: Custom Serving Runtimes

Not Only Development: Serving the Models via Custom Runtimes

OCP AI as Runtime Env: KFP pipelines, LLM, and tree-based models are served simply, efficiently

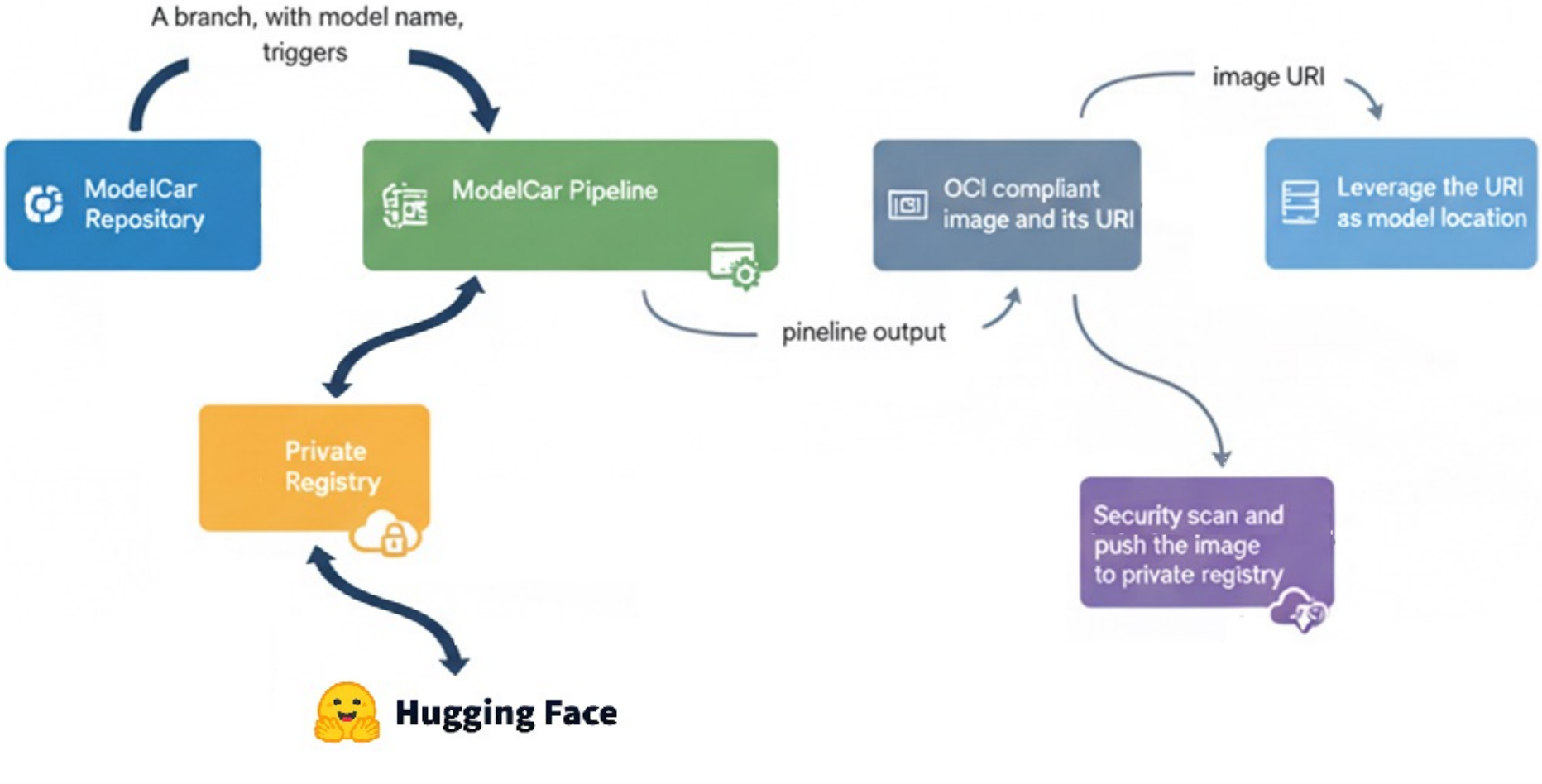
Tree Model Serving: Specialized Speed with FIL Backend

The Need: Standard LLM or generic frameworks are inefficient for classical ML models like XGBoost.
The Solution: Creation of a dedicated custom serving runtime based on Triton that includes the FIL BE.
Result: Dramatically reduced latency and higher throughput.

Single Arfifact of Large Model: The ModelCar Approach

Model as Image: Model, now, is an image artifact, consumable via Serving Runtime

Single Artifact of a Model: The ModelCar Approach



The New Era: A unified analytics & AI Platform

Beyond Data Science:
A Bank-wide service

18
Teams

+200
Users

The Business Impact:
Speed Efficient & Elasticity

-%50
Env. troubleshooting

-%75
Onboarding

A Centralized Hub:

The single, governed source for all analytical workloads.

True End-to-End MLOps:

Covering the full lifecycle from data prep to production monitoring.

The New Frontier:

Our platform is now the home for Generative AI and Agent/Chatbot initiatives.

YapıKredi Teknoloji

Thank You

YapıKredi

Konsalt