Red Hat Summit

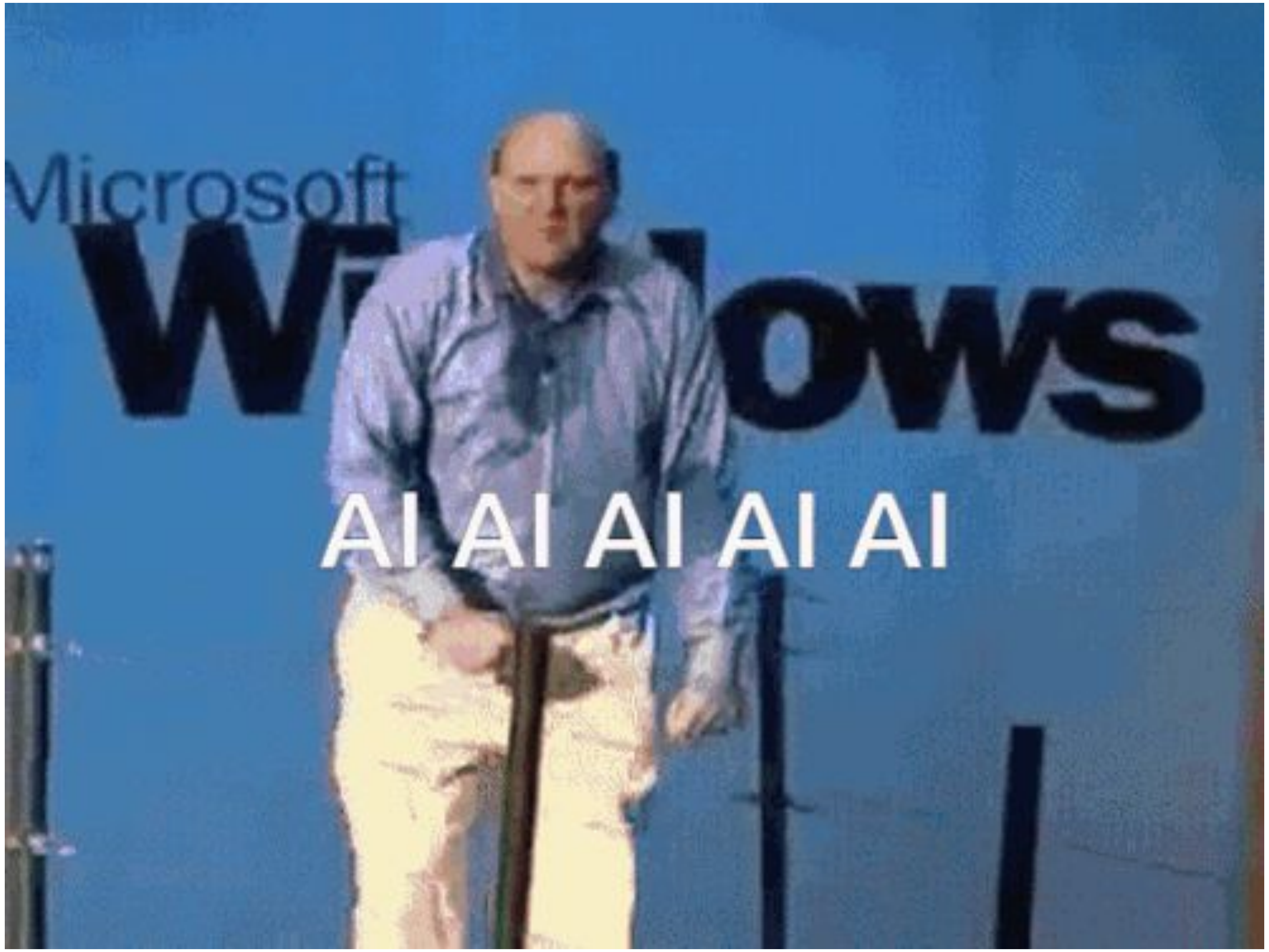
Connect

From Model to Market: Scaling AI with Red Hat OpenShift

Industry Insights

Johana Limka
Senior Solution Architect

Juo Auskalnis
Senior Solution Architect


Microsoft
AI AI AI AI AI

Johana Limka

Senior Solution Architect
Red Hat

Juo Auskalnis

Senior Solution Architect
Red Hat

"For the first time in decades, software itself is changing

from something we write, to something we *speak to*"

Goals:

- Simplified, prescriptive guidance.
- Faster time to value.
- Operational stability at scale.

~95% fail to deliver value

Experimental / POC Phase

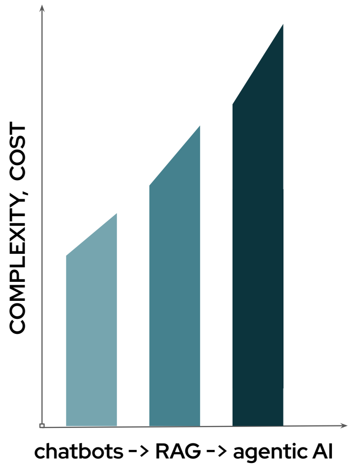
- Isolated Use Cases
- Off-the-shelf Models
- Emerging Tech
- Low Risk

AI Sprawl

High number of Easily accessible models and frameworks.

Shadow AI

- Immediate productivity gains
- Easy-to-use consumer tools



Inflection point

Scale Phase

- Strategic Focus
- Bring-Your-Own-Model
- IT Service Delivery
- High Risk

Approved Models

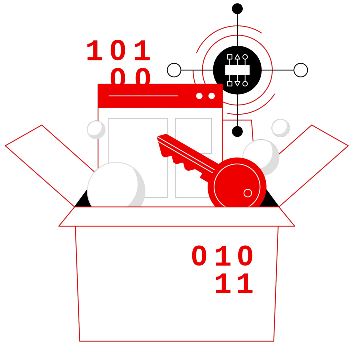
- Business Requirements
- Security
- Data Sovereignty and Privacy
- Cost at Scale
- Operational stability at Scale

"The hardest problem in enterprise AI isn't building a model—it's scaling it with control, compliance, and predictable cost. [...] The journey from a clever model to an ROI-driving production service is where most financial institutions hit a wall."

Red Hat Customer, FSI

People

- Making AI part of everyone’s job
- **Re-imagining roles** and preparing the **workforce** for AI
- **"People revert to old ways under stress."**



Business

- **Efficient** (cost-saving) vs **effective** (maximizing output)
- AI to **re-imagine** processes
- **"What are your most important strategic objectives?"**
- **Create value? Start with the customer!**
- **Cost and Funding**

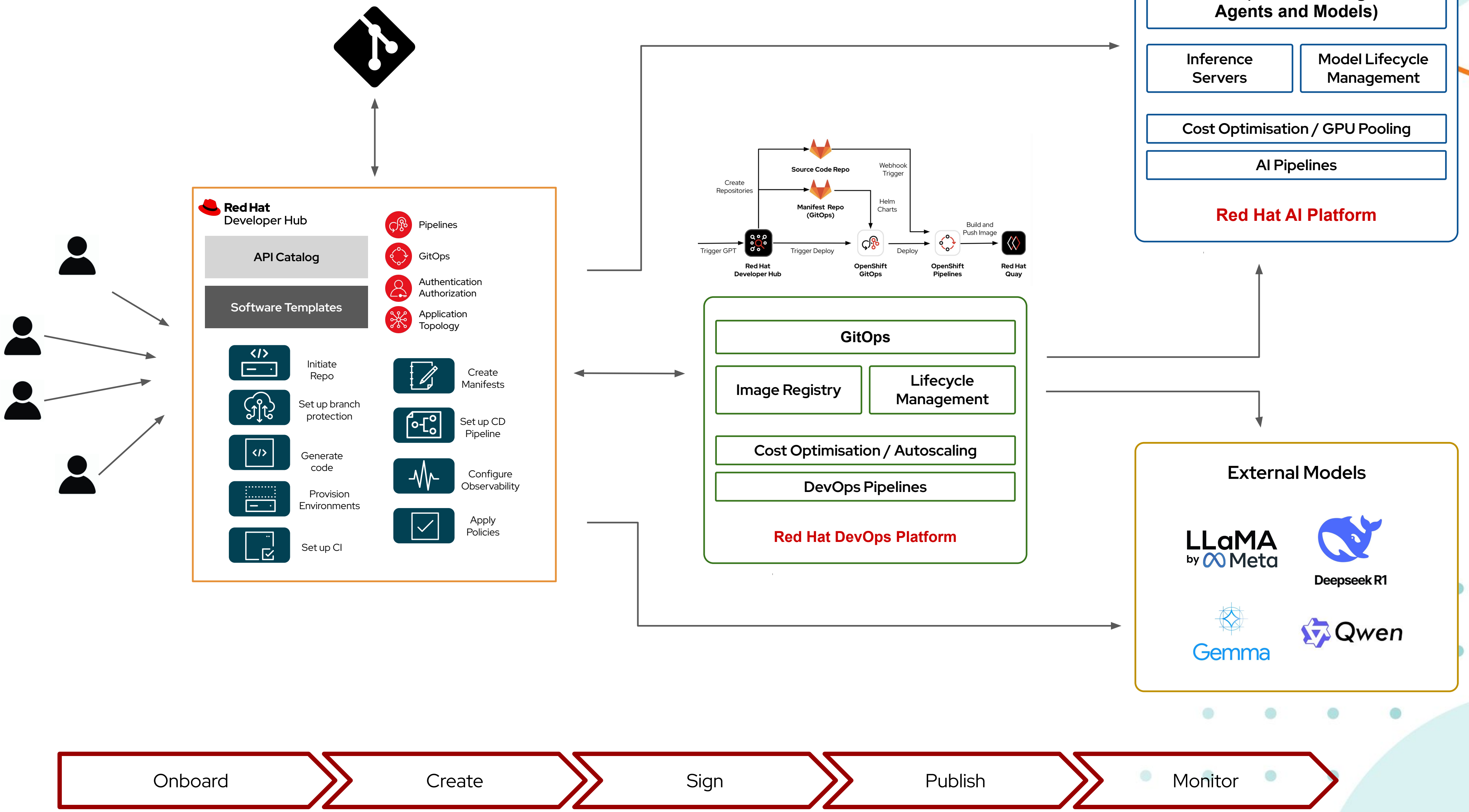
Technology

- **Experimentation vs. Scaling**
- **Small vs (LLM)**
- **Hybrid consistency**
- **Inference and accelerators to reduce cost.**
- **Digital Sovereignty**

The Red Hat Platform: Your Path to Scale

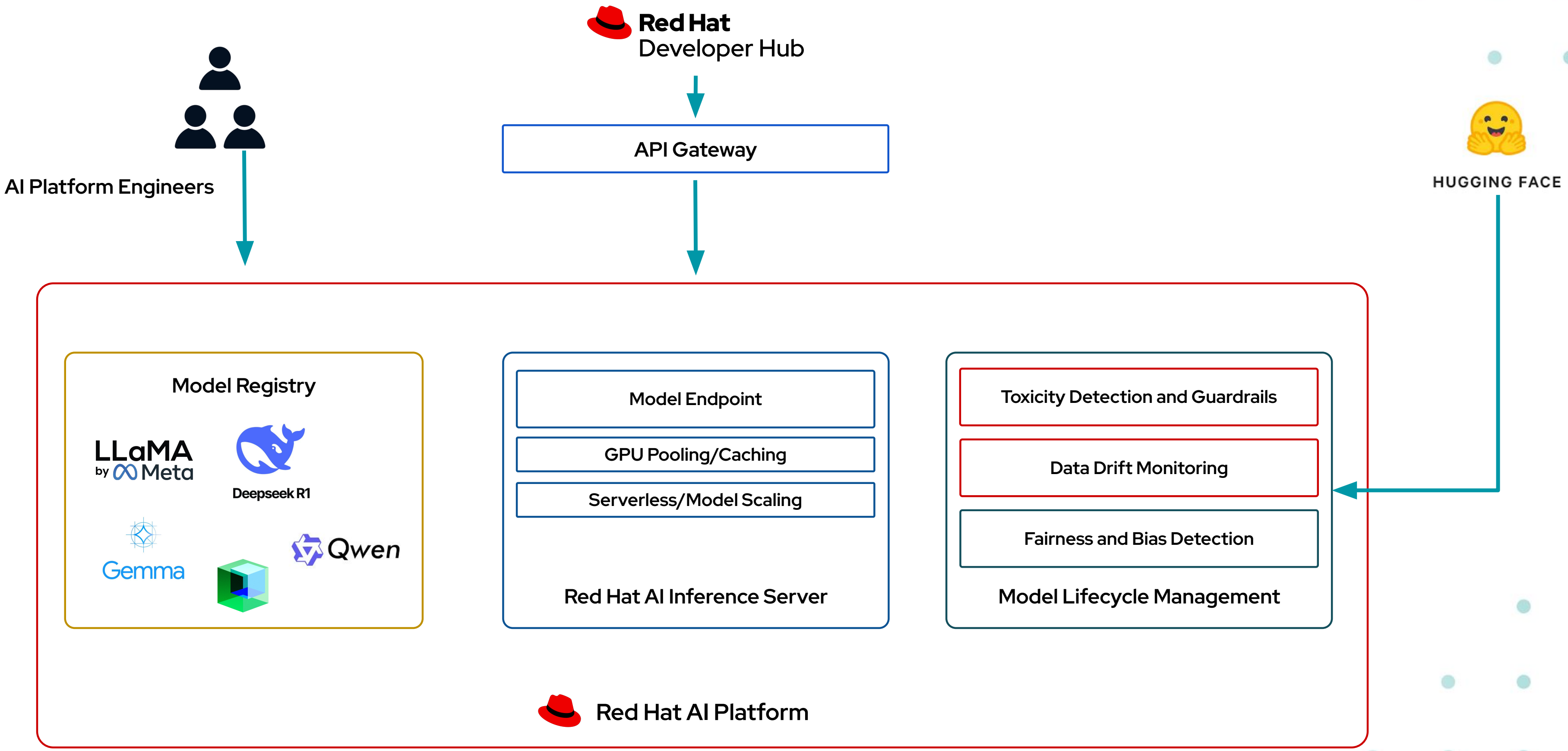
Red Hat Summit Connect London 2025

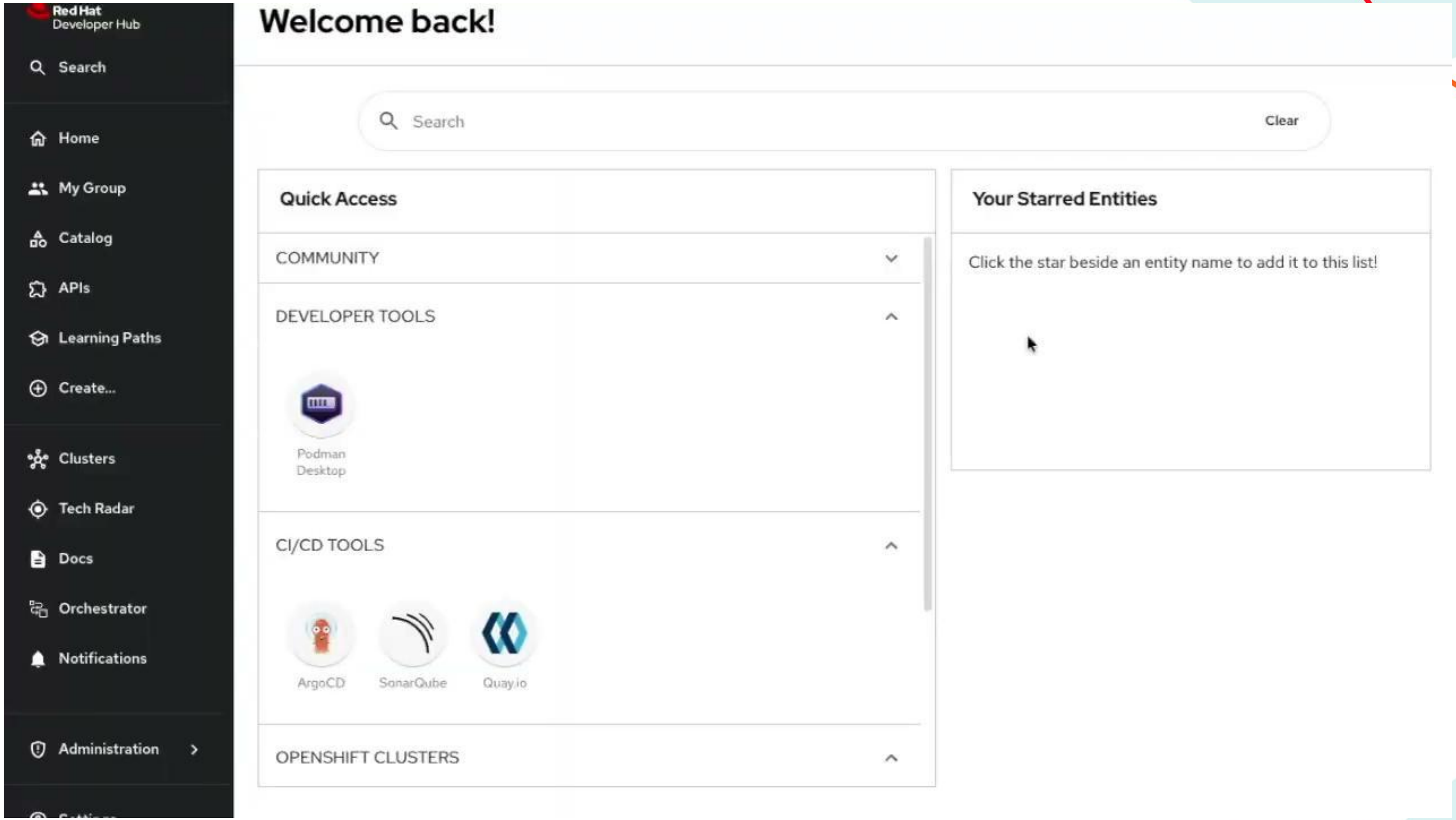
Demo Diagram



Red Hat Summit Connect London 2025

Demo Diagram




Red Hat
Developer Hub
Search
Home
My Group
Catalog
APIs
Learning Paths
Create...
Clusters
Tech Radar
Docs
Orchestrator
Notifications
Administration
Welcome back!
Search
Clear
Quick Access
COMMUNITY
DEVELOPER TOOLS
Podman Desktop
CI/CD TOOLS
ArgoCD
SonarQube
Quay.io
OPENSHIFT CLUSTERS
Your Starred Entities
Click the star beside an entity name to add it to this list!

Call to action

- Don’t let AI live in a lab.
- Move it from *prototype to production,* from *model to market.*
- Start by asking the hard questions – what business problem are we solving, what value are we creating, and how do we scale it responsibly?

At Red Hat, our mission is to help you take AI out of the proof-of-concept stage and into the real world, safely, efficiently, and with confidence.

So let’s build that future together – open, hybrid, and ready for what’s next.

Red Hat Summit

Connect

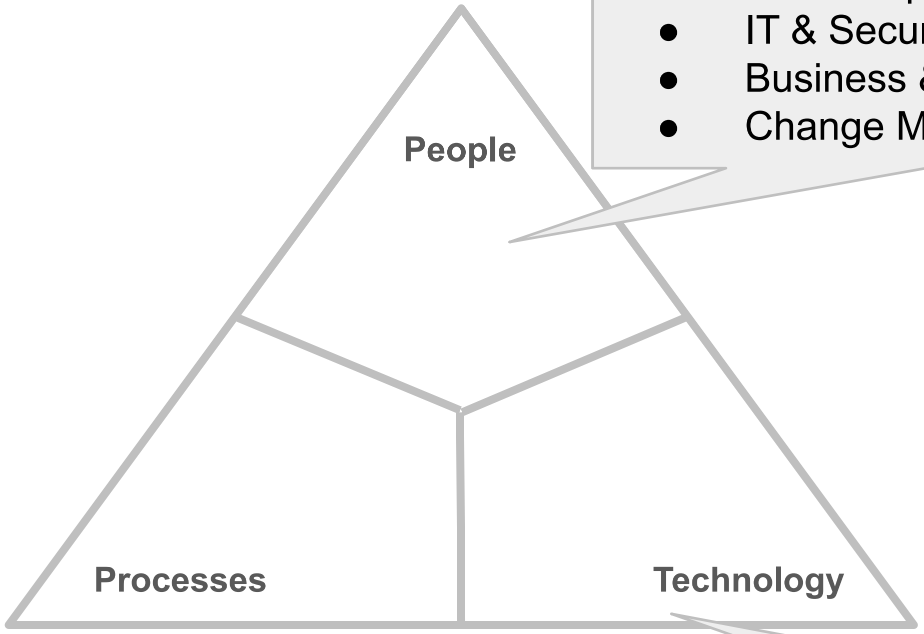
Thank you

linkedin.com/company/red-hat

facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat



Cultural

- AI/ML Experts
- IT & Security Teams
- Business & Policy Stakeholders
- Change Management Leadership

Technical

- **AI/ML Platform**
- API Layer
- Data infrastructure
- **Security & Access Controls**
- **Monitoring & Observability**
- **Automation & CI/CD**

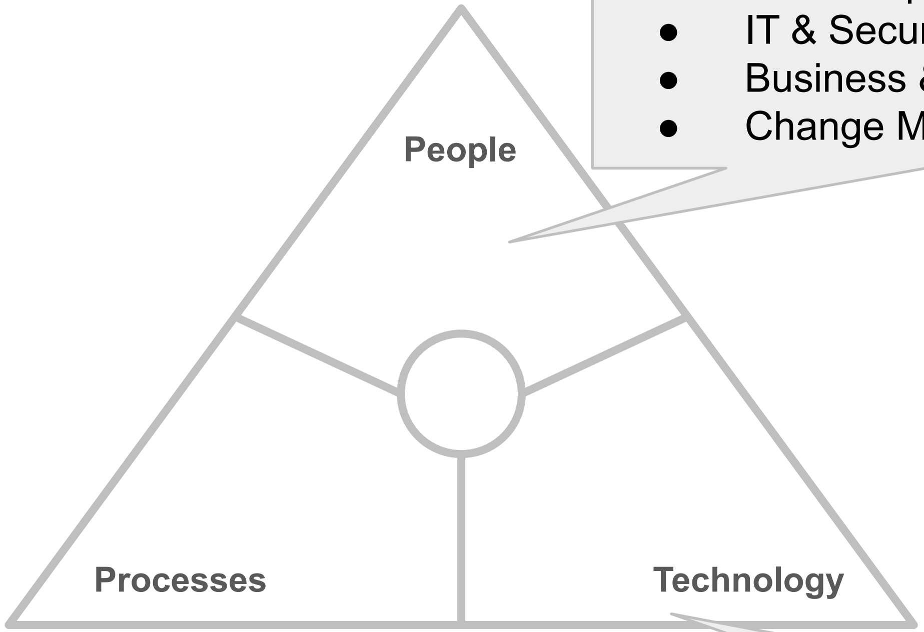
Business

- **AI to support business strategy**
-

DWP Digital

People

- AI/ML Experts
- IT & Security Teams
- Business & Policy Stakeholders
- Change Management Leadership



Processes

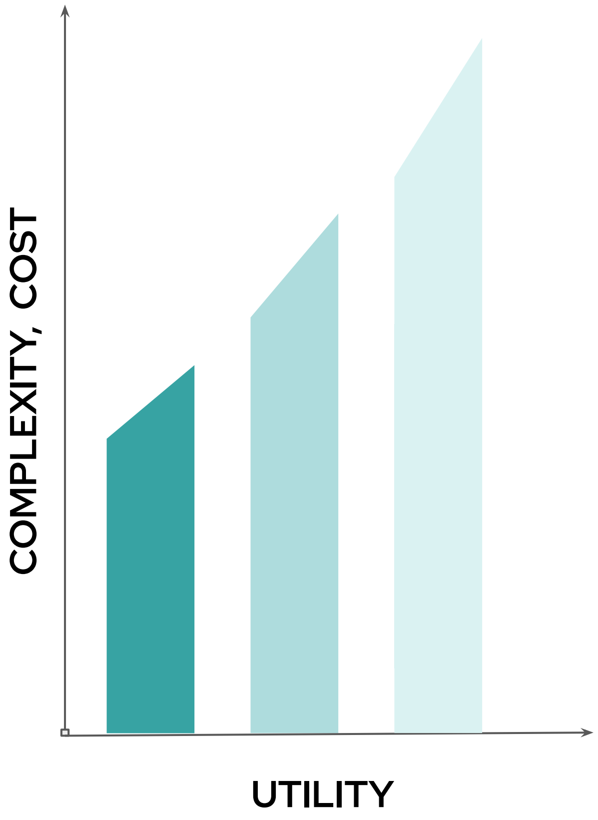
- **Model onboarding & Validation**
- Data Governance
- Security & Compliance
- **Lifecycle Management**
- Incident Response

Technologies

- **AI/ML Platform**
- API Layer
- Data infrastructure
- **Security & Access Controls**
- **Monitoring & Observability**
- **Automation & CI/CD**

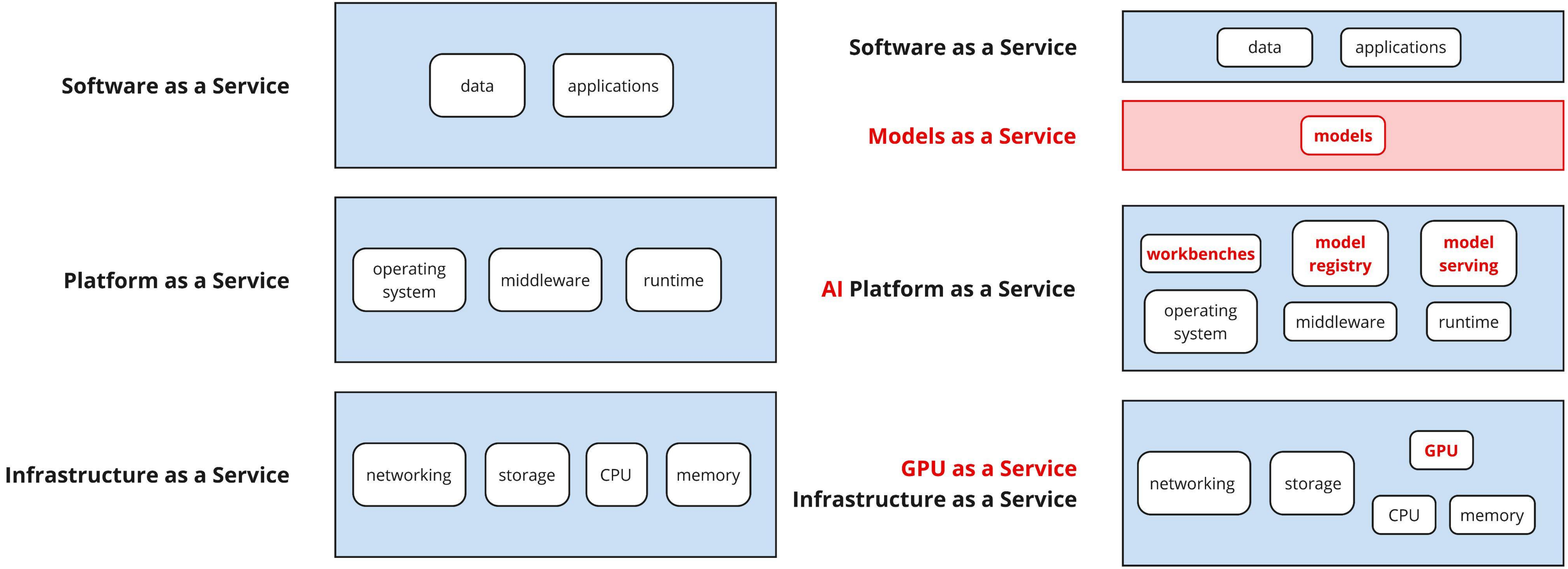
The Hidden Cost of Generative AI

More Advanced Techniques Incur Higher Costs



- **Basic chatbots** that provide direct LLM querying are relatively cheap to operate but offer limited utility.
- **Retrieval augmented generation (RAG)** allows tapping into proprietary knowledge. The cost per query multiplies.
- **Agentic AI** provides arbitrary integration and query iteration capabilities. One user query may generate many high-cost LLM queries.

Service Layers



Scaling with Red Hat OpenShift

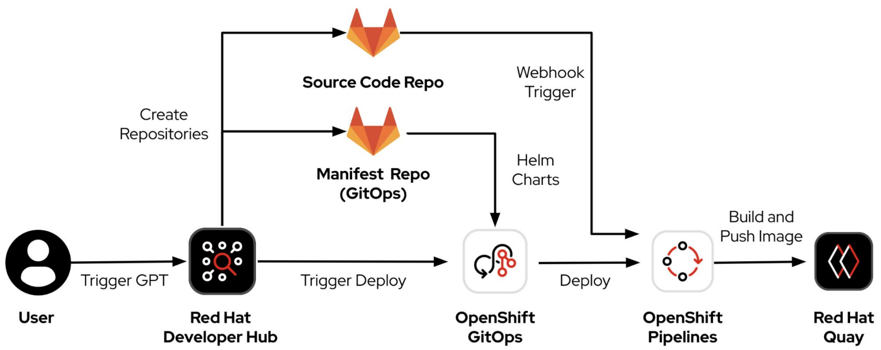
Why OpenShift for MaaS?

- Consistent platform: app + model workloads on same infra
- Operator-based automation (Ansible + OpenShift AI)
- Cloud flexibility: ROSA, ARO, on-prem OCP
- Ecosystem integrations: Bedrock BYOM, Hugging Face, RAG frameworks

Model-as-a-Service Framework

What MaaS provides

- Unified **model store + inference runtime + lifecycle governance**
- On-demand inference: scalable, autoscaling (vLLM-d, KServe)
- Multi-tenant, multi-model → single platform view
- Governance: SBOM, bias scanning, provenance
  (Visual: simple layered diagram – Data → Models → MaaS → Apps/Users)



Benefits to Business & IT

Business benefits

- Faster time-to-market for AI apps
- Reusability of models across teams
- Predictable cost and performance scaling

IT/Engineering benefits

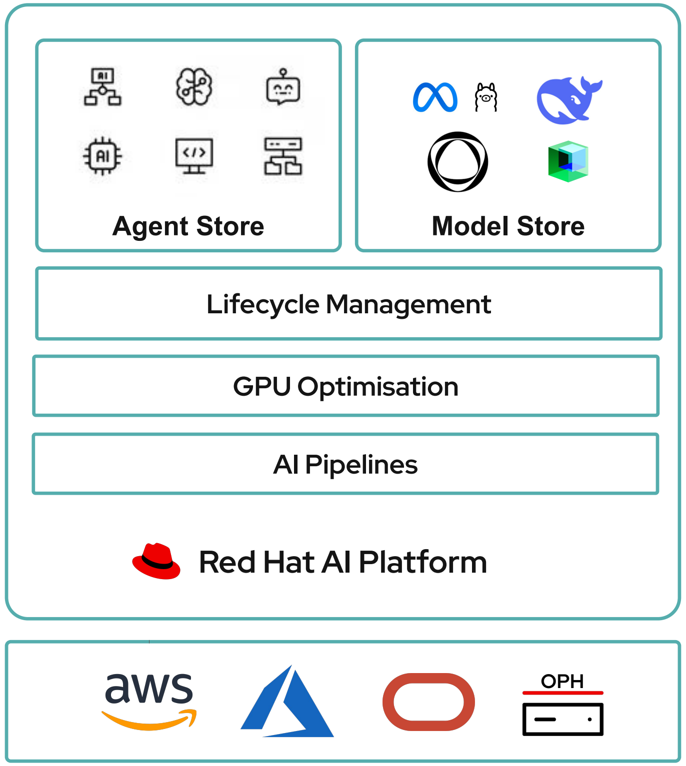
- Runs on OpenShift anywhere (cloud, on-prem, hybrid)
- GPU efficiency: partitioning, autoscaling
- Integrated CI/CD for models, observability, security

Call to Action

From model to market, at enterprise scale

- Start small: one model, one app
- Scale fast: 15+ models, 60k users, hybrid deployment
- Partner with Red Hat for **MaaS accelerators & governance**
- **Red Hat AI Incubator** → guided adoption program:
    - De-risks early AI projects
    - Provides engineering residencies & accelerators
    - Builds internal skills and production-ready pipelines
- **CTA**: “Let’s make your models production-ready with Red Hat OpenShift and AI Incubator support.”

Approach


Agent Store
Model Store
Lifecycle Management
GPU Optimisation
AI Pipelines
Red Hat AI Platform
aws
OPH

The Challenge

Why scaling AI is hard

- Fragmented tooling (training vs deployment vs
- High costs of experimentation vs production
- Model sprawl → security, compliance, lifecycle
- Throughput and user demand mismatch (60k u