Red Hat Summit

Connect

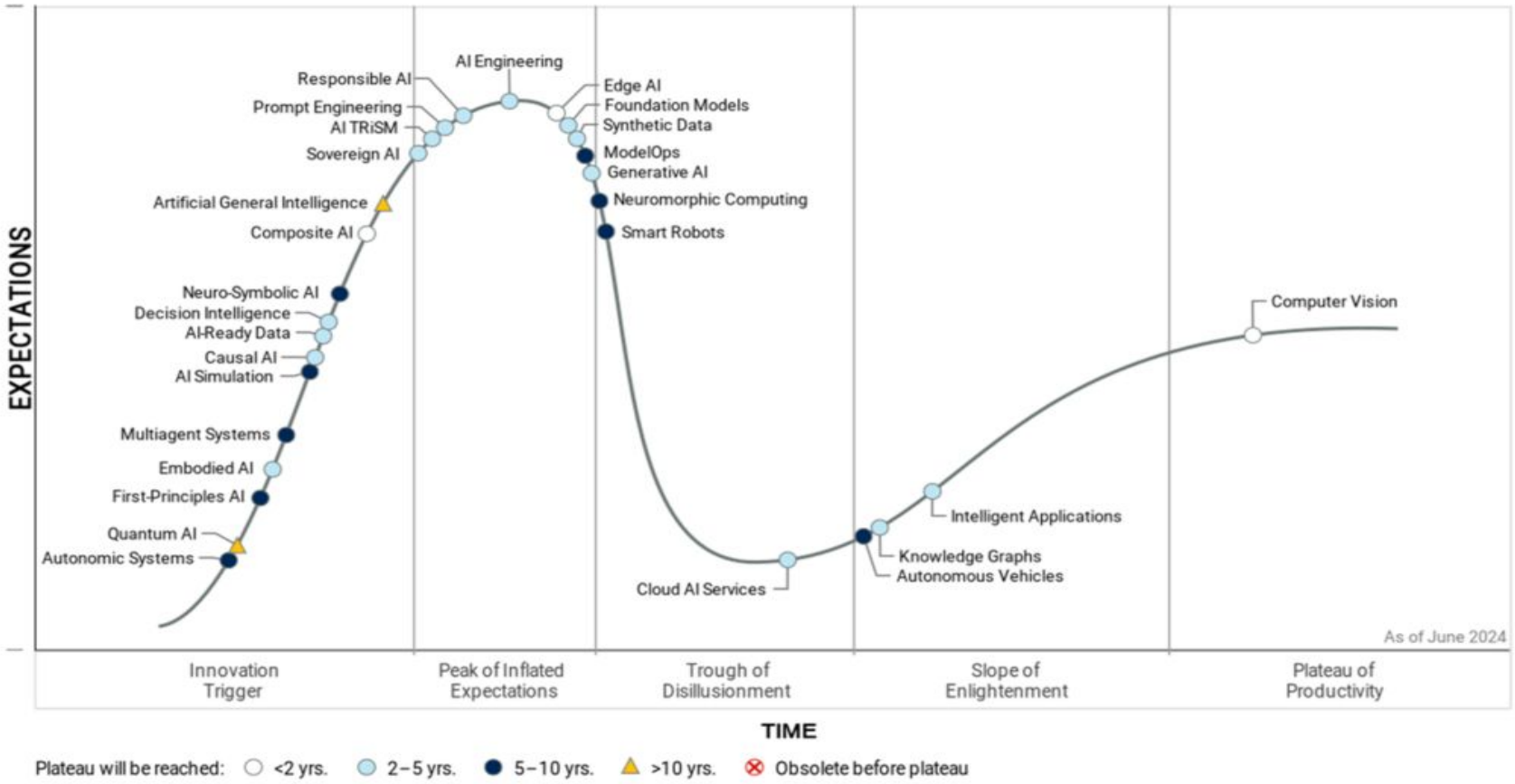
Red Hat OpenShift in an AI centric world

Andreas Bergqvist
Red Hat AI
EMEA SSP

Magnus Gadd
Director
Red Hat EMEA

Red Hat

Hype Cycle for Artificial Intelligence, 2024

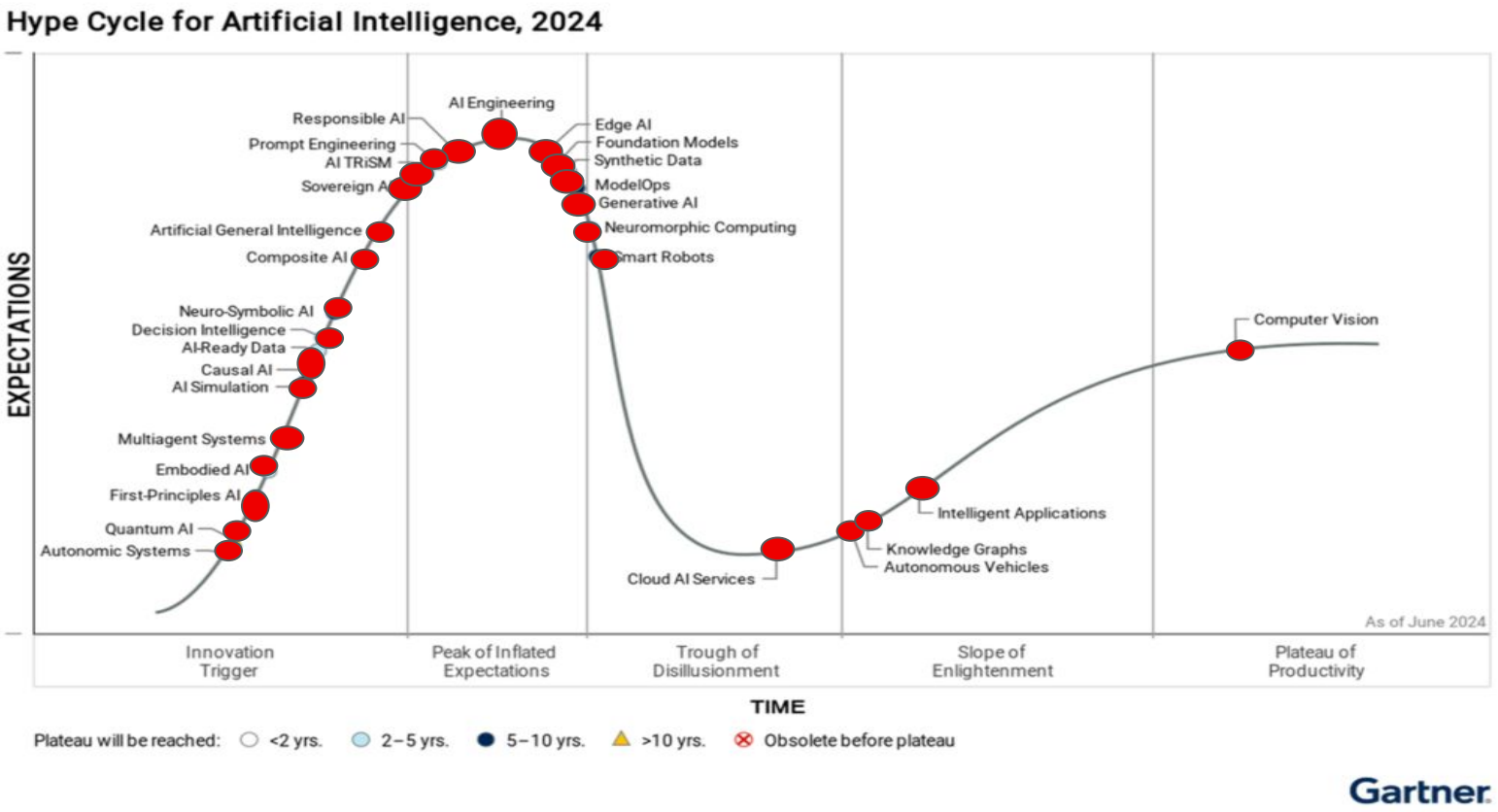


Gartner

https://www.pasqal.com/news/new-gartner-ai-hype-cycle-report/

Red Hat

Where is Red Hat in this AI centric world?



But why is then Red Hat not top of mind of world + dog when discussing AI?

Red Hat

Because Red Hat and our Eco system is a (**very significant) part** of the plumbing that is driving AI.

Unfortunately the brand of the plumbing is rarely used in commercials

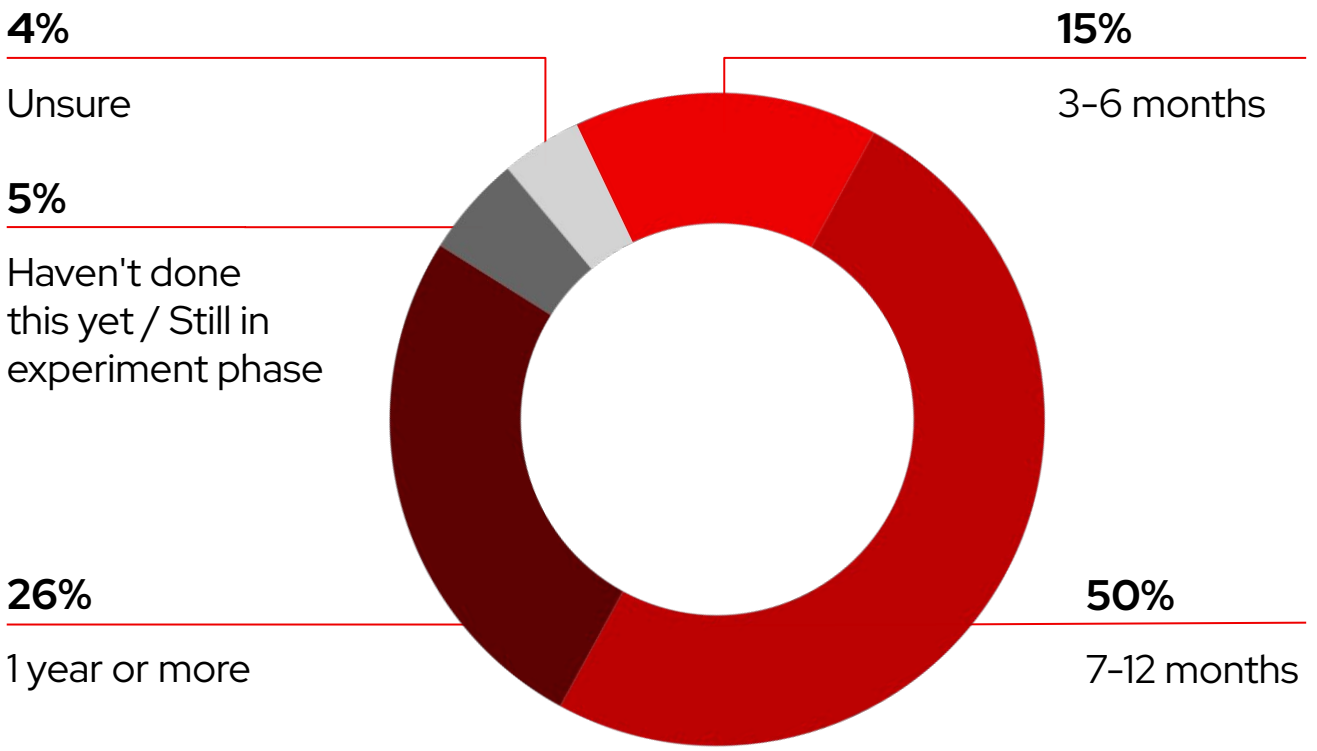
So we are already part of the solution, but there is a big need to do better.

Why?

Because all is not good in AI Land

What is the average AI/ML timeline from idea to operationalizing the model?

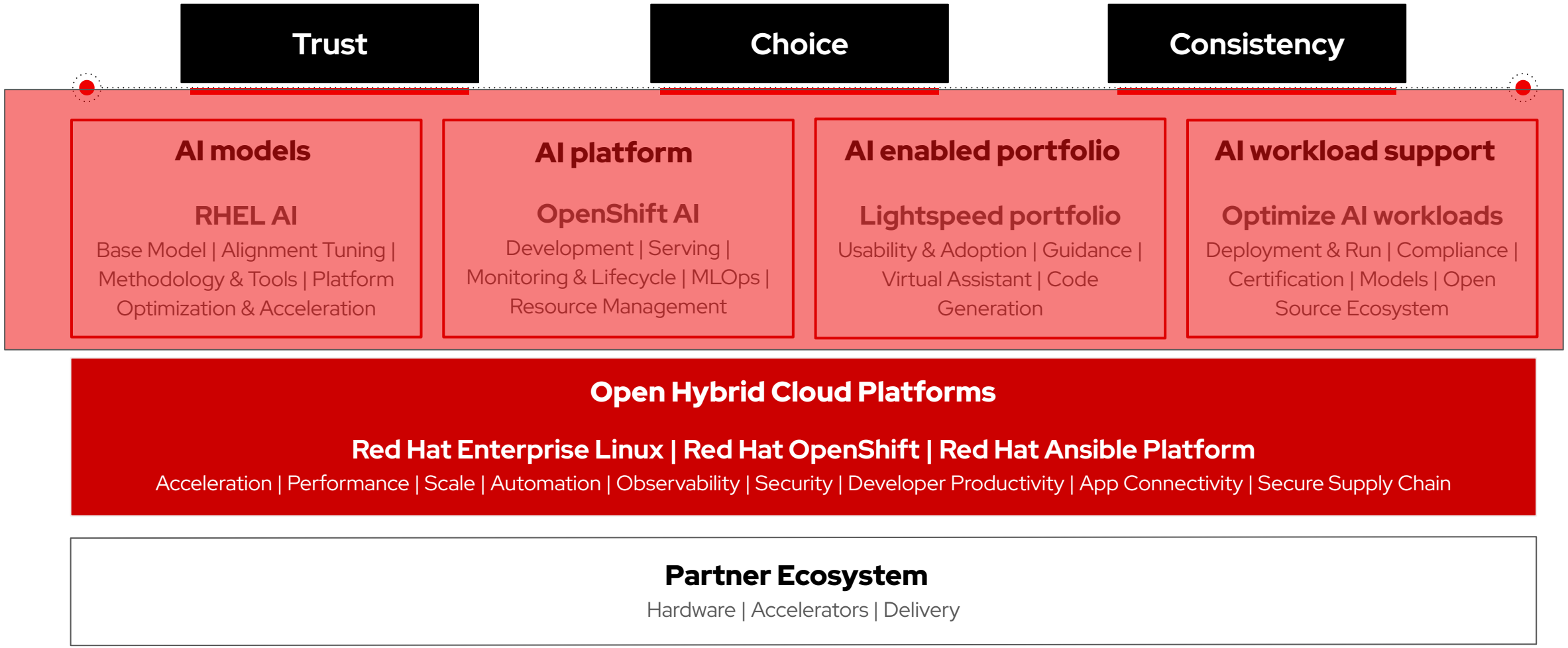
Half of respondents (50%) say their average AI/ML timeline from idea to operationalizing the model is 7-12 months.



Source: Gartner Peer Insights, Open Source AI for Enterprise survey, 2023

Red Hat

Red Hat's AI portfolio strategy aims to Make AI Great Again



Red Hat

OpenShift as the AI Orchestrator

How OpenShift facilitates the development and delivery of all stages of the AI lifecycle.

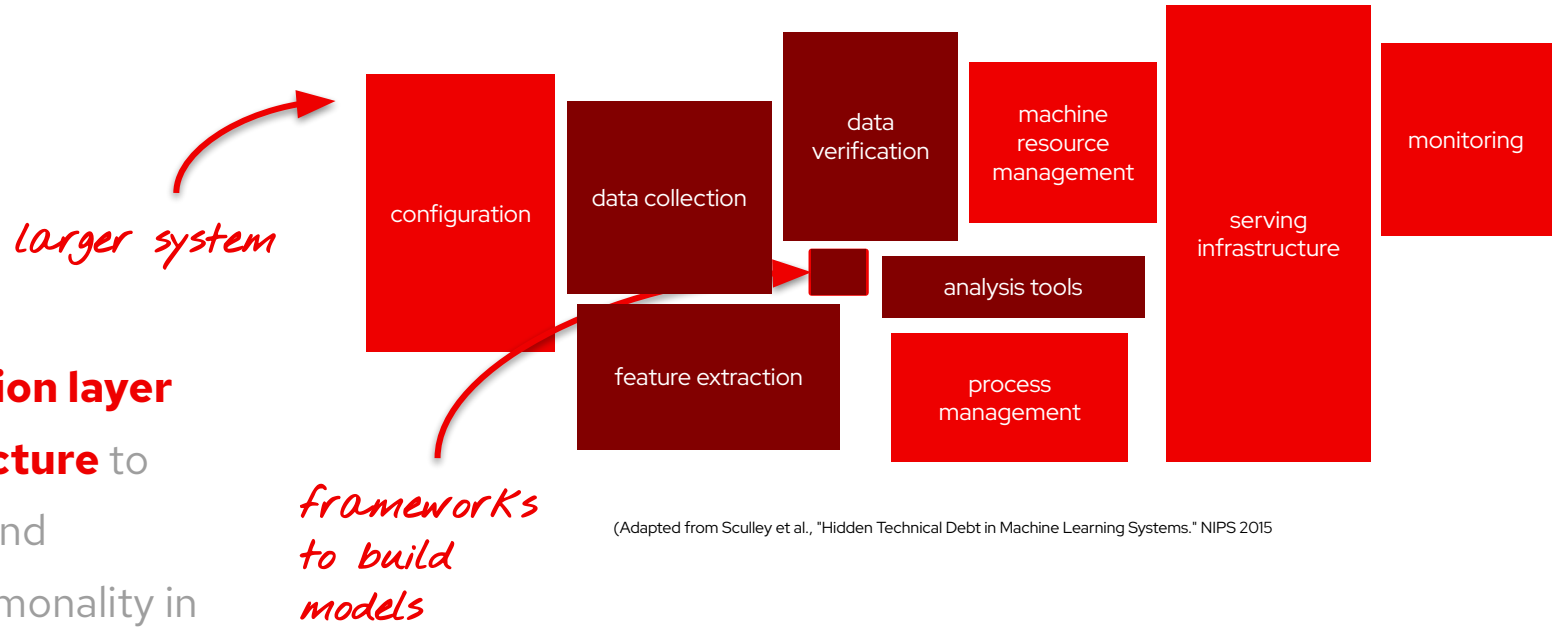
Red Hat

Data Science about More than Data Science

How to tackle the Hidden Technical Debt

"**a consistent application platform** for the management of existing, modernized, and cloud-native applications that runs on any cloud."

"**a common abstraction layer across any infrastructure** to give both developers and operations teams commonality in how applications are packaged, deployed, and managed."

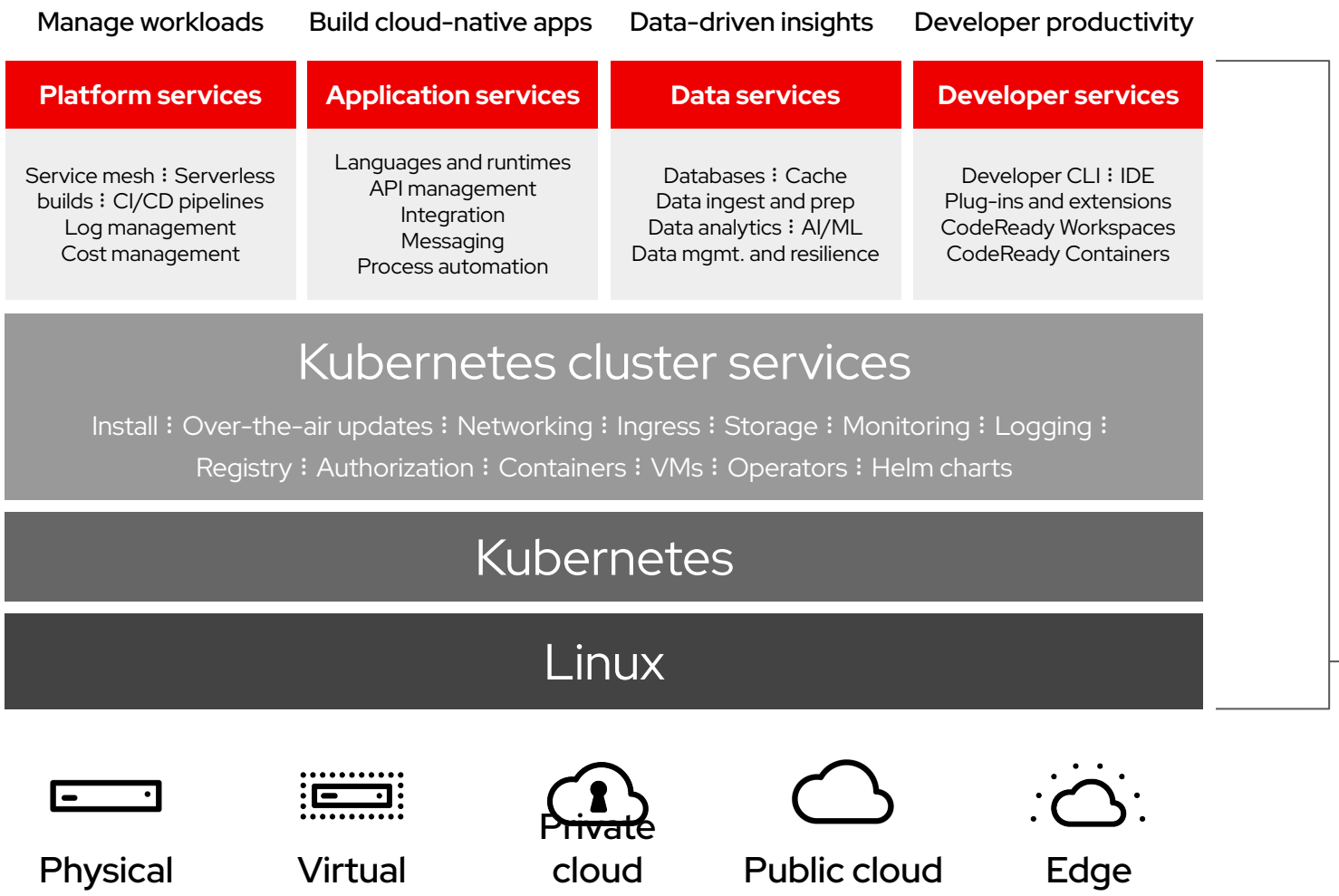


(Adapted from Sculley et al., "Hidden Technical Debt in Machine Learning Systems." NIPS 2015

Source: https://www.redhat.com/en/technologies/cloud-computing/openshift

The value of Red Hat OpenShift

A complete digital platform



Automated, full-stack installation from the container host to application services

Seamless Kubernetes deployment to any cloud or on-premises environment

Autoscaling of cloud resources

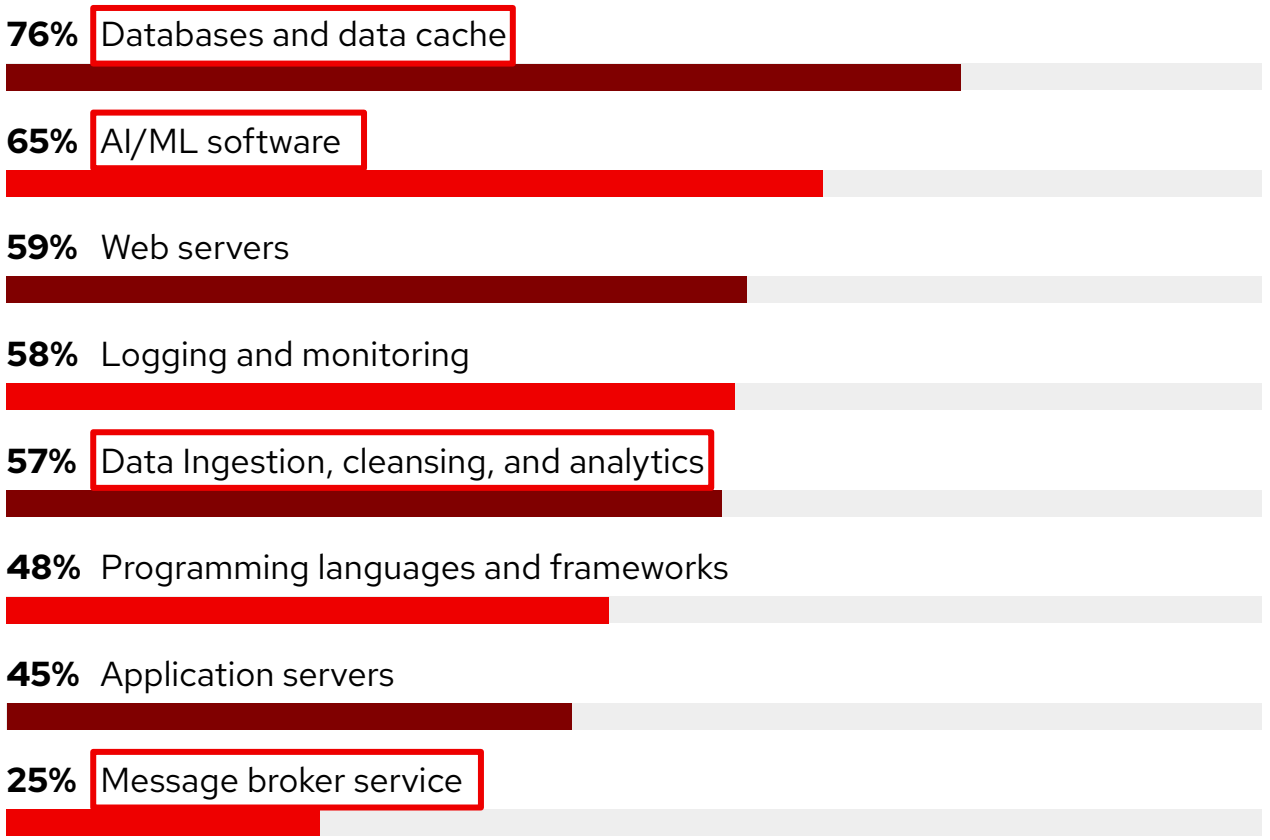
One-click updates for platform, services, and applications

Red Hat OpenShift

Red Hat

OpenShift is already heavily utilized for Data & AI-driven Digital Products

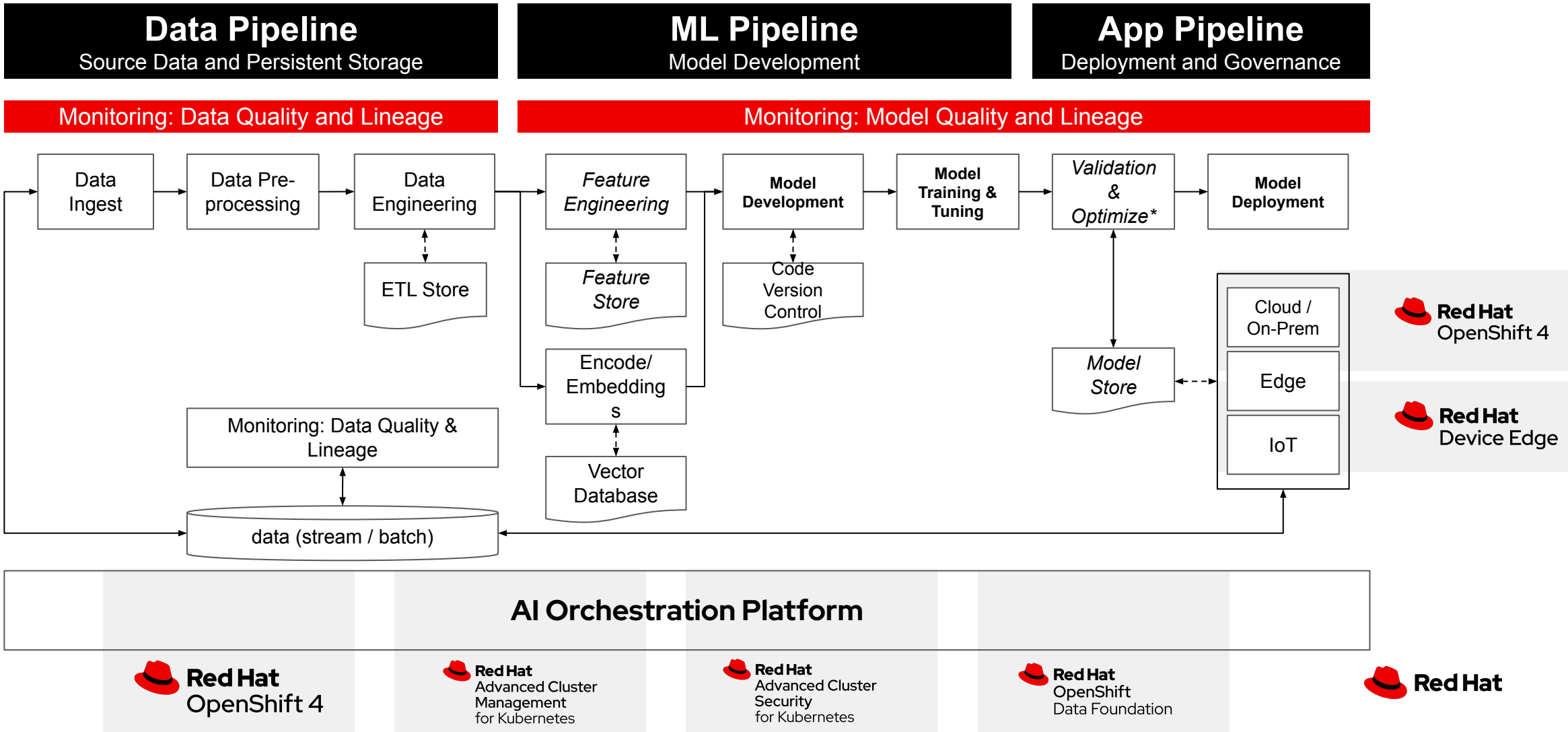
Types of workloads deployed in containers and Kubernetes environments[1]



1 Pulse, sponsored by Red Hat. "State of workloads adoption on containers and Kubernetes," November 2021.

Red Hat

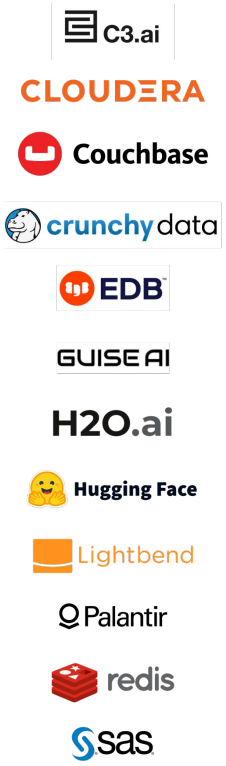
Orchestrating the AI Lifecycle in **ANY** Environment



Some members of Red Hat's AI Partner Ecosystem

Integrated ISVs

AI and general ISVs



Delivery partners

Cloud partners

Hardware

Red Hat

Red Hat's AI portfolio strategy aims to Make AI Great Again

Trust | **Choice** | **Consistency**

AI models

RHEL AI

Base Model | Alignment Tuning | Methodology & Tools | Platform Optimization & Acceleration

AI platform

OpenShift AI

Development | Serving | Monitoring & Lifecycle | MLOps | Resource Management

AI enabled portfolio

Lightspeed portfolio

Usability & Adoption | Guidance | Virtual Assistant | Code Generation

AI workload support

Optimize AI workloads

Deployment & Run | Compliance | Certification | Models | Open Source Ecosystem

Open Hybrid Cloud Platforms

Red Hat Enterprise Linux | Red Hat OpenShift | Red Hat Ansible Platform

Acceleration | Performance | Scale | Automation | Observability | Security | Developer Productivity | App Connectivity | Secure Supply Chain

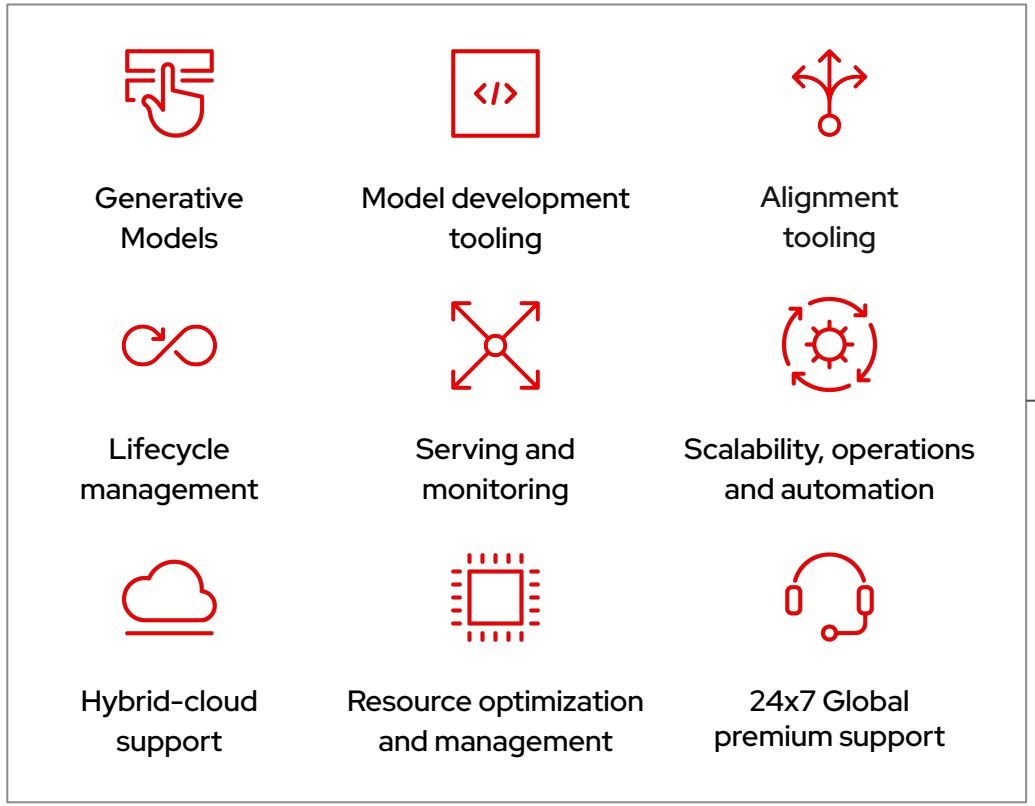
Partner Ecosystem

Hardware | Accelerators | Delivery

Red Hat

Red Hat AI platforms

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



Trust

Meet data location, privacy and security constraints while controlling costs by owning the decision of where to train, deploy and run AI models

Choice

Access to ready-to-use environments, models and AI/ML capabilities from curated open source and emerging technologies to expedite AI adoption

Consistency

Streamlines the process of managing and monitoring the lifecycle of models and AI-enabled applications at scale and across clouds

Red Hat

Red Hat AI platforms

Red Hat Enterprise Linux AI

Foundation model platform for developing, testing, and running Granite family LLMs

- Provides a simplified approach to get started with generative AI that includes open source models
- Makes AI accessible to developers and domain experts with little data science expertise
- Provides the ability to do training & inference on individual production server deployments

Red Hat OpenShift AI

Integrated MLOps platform for model lifecycle management at scale anywhere

- Provides support for both generative and predictive AI models with a BYOM approach
- Includes distributed compute, collaborative workflows, model serving and monitoring
- Offers enterprise MLOps capabilities and the ability to scale across hybrid-clouds
- Includes Red Hat Enterprise Linux AI, including the Granite family models

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

Red Hat

Red Hat OpenShift AI

Integrated AI platform

Create and deliver gen AI and predictive models at scale across hybrid cloud environments.

Available as

- Fully managed cloud service
- Traditional software product on-site or in the cloud!

Model development

Bring your own models or customize Granite models to your use case with your data. Supports integration of multiple AI/ML libraries, frameworks, and runtimes.

Model serving and monitoring

Deploy models across any OpenShift footprint and centrally monitor their performance.

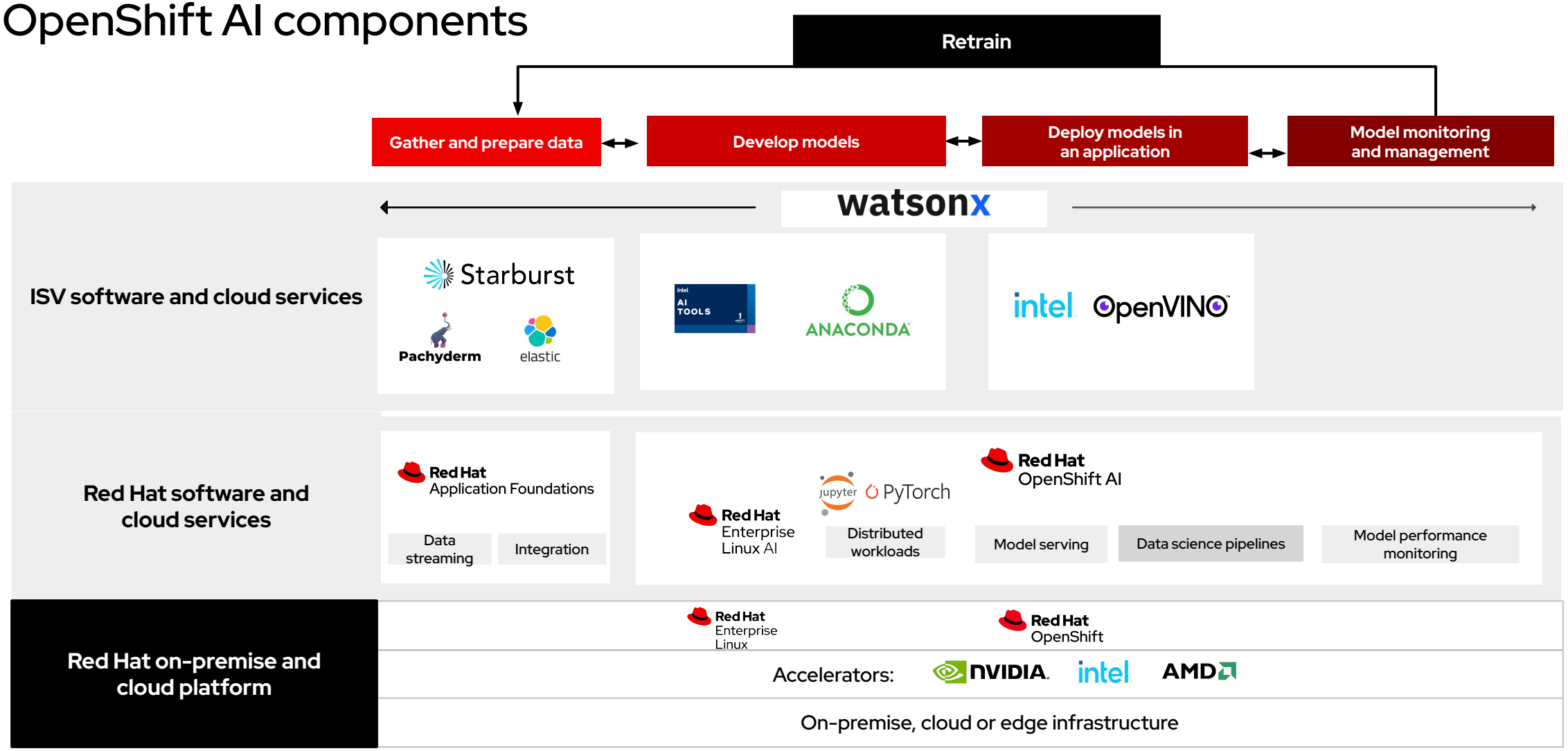
Lifecycle management

Expand DevOps practices to MLOps to manage the entire AI/ML lifecycle.

Resource optimization and management

Scale to meet workload demands of gen AI and predictive models. Share resources, projects, and models across environments.

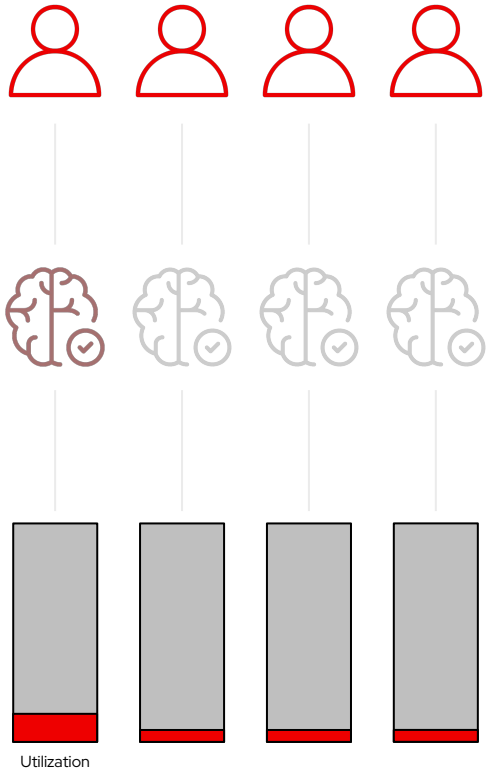
OpenShift AI components



Red Hat

Infrastructure as a Service **won't scale** beyond Data Scientists

Under-utilized GPUs, redundant workloads & wasted resources



Different users, same model, wasted GPUs

Associates are not Data Scientists

*Models as a Service **is NOT** giving an Associate their own LLM & GPUs*

Infrastructure as a Service (IaaS) offers hardware access, ie. storage or GPUs, to Data Scientists or AI Engineers

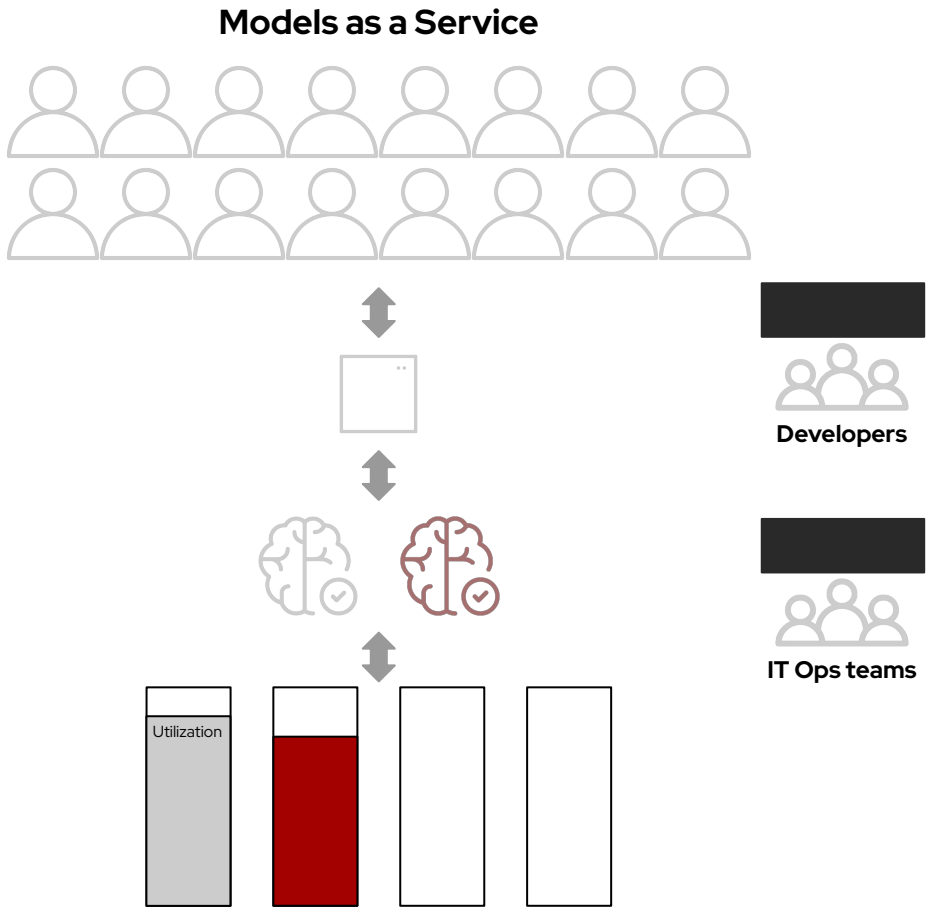
- Besides Data Scientists, few can use GPUs correctly
- GPUs are often under-utilized
- "Throwing GPUs at the problem" won't scale
- Diminishing returns on most expensive resources
- High cost and redundant workloads curbs innovation

Red Hat

Describing **Models as a Service**

Models as a Service (MaaS) solves the economies of scale problem. MaaS drives model consumption to a wider audience so all can innovate regardless of GPU experience

- IT serves and maintains models centrally
- Remove technical blockers, reduce time to market
- Right access: Devs get endpoints, users get apps
- Shared resource business model keeps costs down
- Secure hosting with private workloads

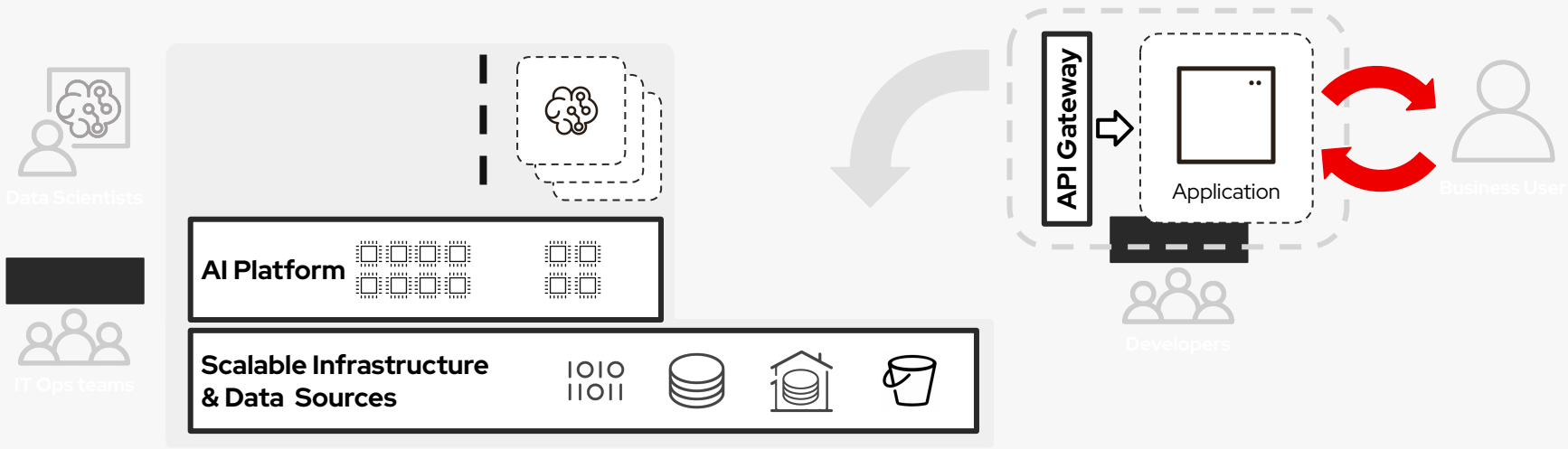


Give Associates tools they can use

They don't care about GPUs or model endpoints. Give them a centrally managed LLM service with an application interface

Red Hat

Today’s infrastructure + tomorrow’s strategy



What’s not new:

- IT remains optimal team to manage AI infrastructure
- Platform, hardware & access centrally managed
- Data Scientists use GPU service to customize models
- IT & Data Scientists monitor & evaluate performance

What’s new:

- Model serving is operationalized for wider audience
- IT adds API Gateway for production serving
- Developers build using standardized endpoints
- Associates consume Private AI Services

Red Hat

Red Hat Enterprise Linux AI

Foundation Model Platform

Seamlessly develop, test, and run Granite family large language models (LLMs) for enterprise applications.

Granite family models

Open source-licensed LLMs, distributed under the Apache-2.0 license, with complete transparency on training datasets.

InstructLab model alignment tools

Scalable, cost-effective solution for enhancing LLM capabilities and making AI model development open and accessible to all users.

Optimized bootable model runtime instances

Granite models & InstructLab tooling packaged as a bootable RHEL image, including Pytorch/runtime libraries and hardware optimization (NVIDIA, Intel and AMD).

Enterprise support, lifecycle & indemnification

Trusted enterprise platform, 24x7 production support, extended model lifecycle and model IP indemnification by Red Hat.

An open source **community** project for GenAI model development

instructlab

Type / to search

Overview | Repositories 7 | Discussions | Projects 1 | Packages | People 21

InstructLab

Unfollow

README.md

Welcome to the 🐶 InstructLab Project

InstructLab 🐶 uses a novel synthetic data-based alignment tuning method for Large Language Models (LLMs.) The "lab" in InstructLab 🐣 stands for Large-Scale Alignment for ChatBots [1].

[1] Shivchander Sudalairaj*, Abhishek Bhandwaldar*, Aldo Pareja*, Kai Xu, David D. Cox, Akash Srivastava*. "LAB: Large-Scale Alignment for ChatBots", arXiv preprint arXiv: 2403.01081, 2024. (* denotes equal contributions)

Why InstructLab

There are many projects rapidly embracing and extending permissively licensed AI models, but they are faced with three main challenges:

- Contribution to the models themselves is not possible directly. They show up as forks, which forces consumers to choose a "best-fit" model that isn't easily extensible, and the forks are expensive for model creators to maintain.
- The ability to contribute ideas is limited by a lack of AI/ML expertise. One has to learn how to fork, train, and refine models in order to see their idea move forward. This is a high barrier to entry.
- There is no direct community governance or best practice around review, curation, and distribution of forked models.

Top discussions this past month

Discussions are for sharing announcements, creating conversation in your community, answering questions, and more.

Start a new discussion

People

View all

Top languages

Python Shell TypeScript Jupyter Notebook

Report abuse

https://github.com/instructlab

IBM Research Red Hat

IBM Granite model family

Released under the Apache 2 license

Granite

IBM Granite
Language models

English Base
Granite-7B-Base

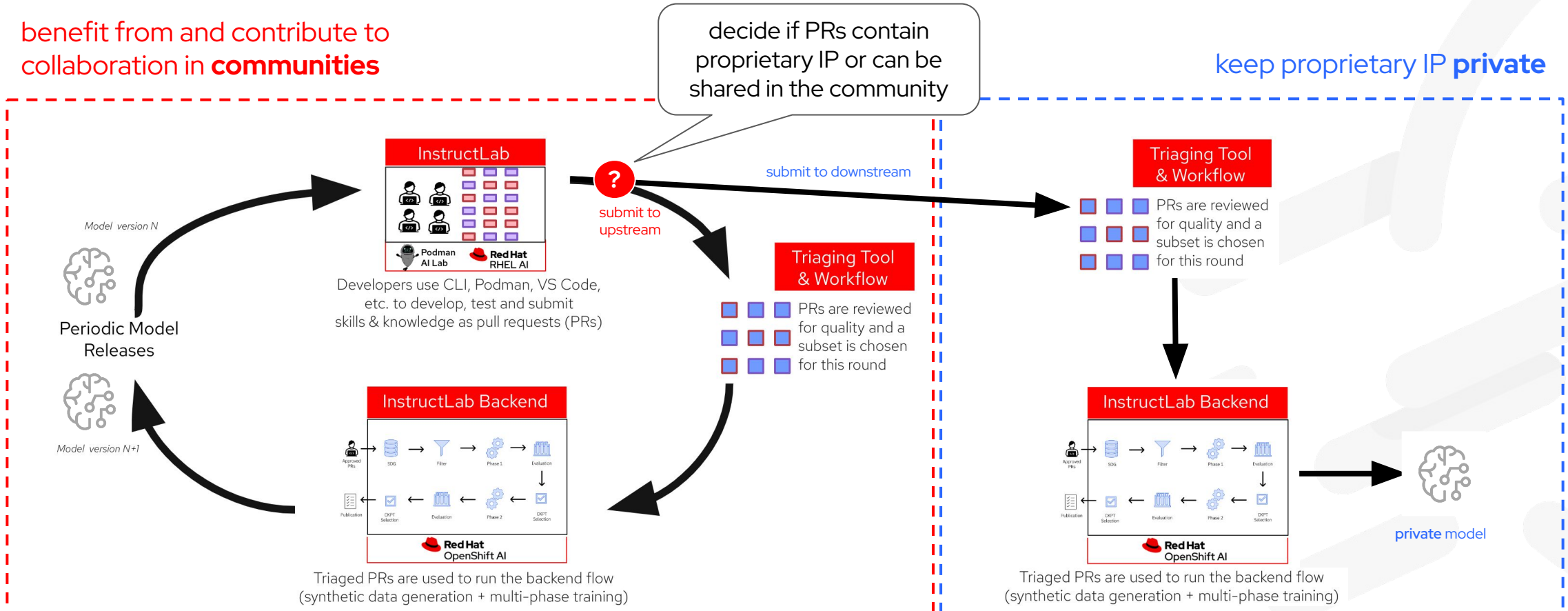
English Instruction-tuned
Granite-7B-Instruct

IBM Granite
Code models

Base
Granite-34B-Code-Base
Granite-20B-Code-Base
Granite-8B-Code-Base
Granite-3B-Code-Base

Instruction-tuned
Granite-34B-Code-Instruct
Granite-20B-Code-Instruct
Granite-8B-Code-Instruct
Granite-3B-Code-Instruct

Contributing while keeping proprietary IP private

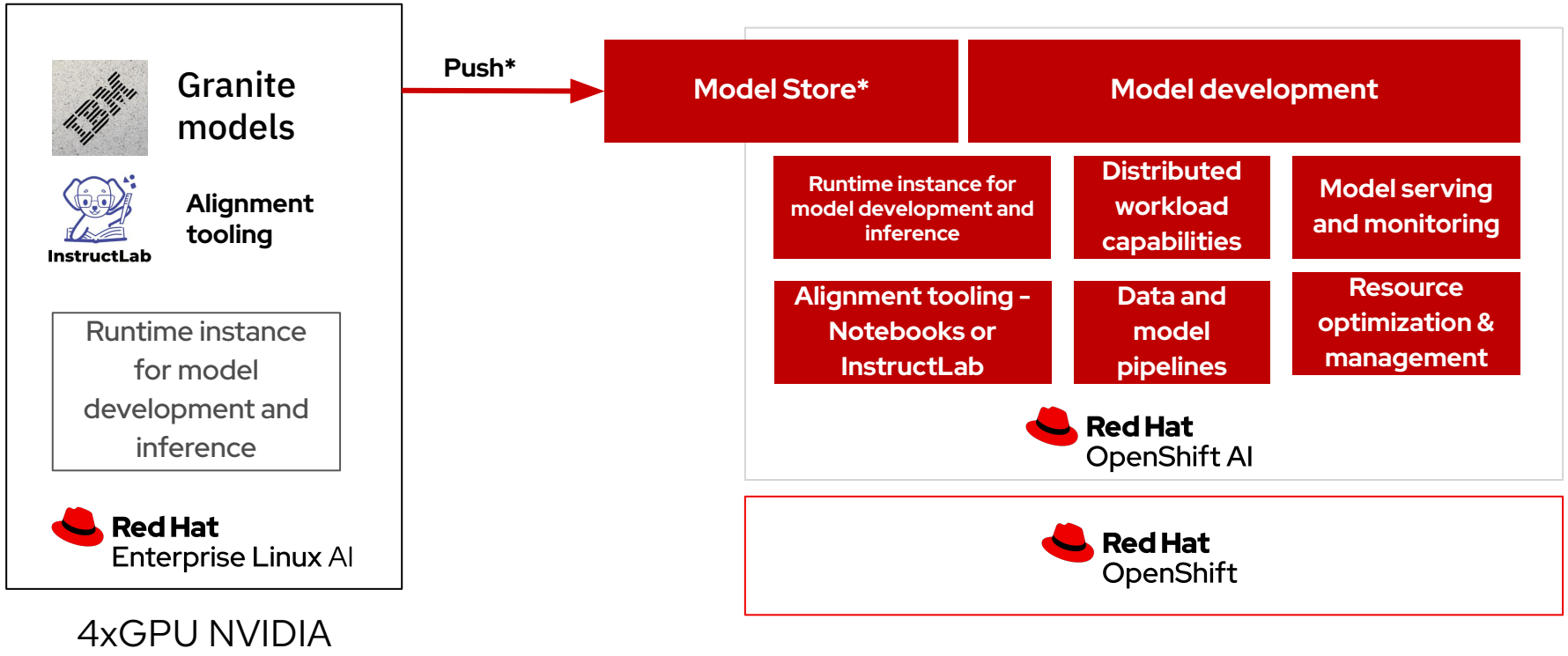


Skills and knowledge that can be shared with the community are contributed upstream. These come back for free with the next version of the model, thus reducing the resources required for in-house fine-tuning of the private model, and potentially improved by other collaborators.

Proprietary skills and knowledge, that shall not be shared, are not submitted upstream but retained in-house. These have to be re-added to each new version of the upstream base model.

Red Hat

Fitting RHEL AI + OpenShift AI together



Simple 'Build / Deploy' approach + RHEL AI = LLM private knowledge compiler.

*roadmap

Red Hat and IBM AI Portfolio

InstructLab

STEP 1

Learn & experiment via limited desktop-scale training method (qlora) on small datasets. *Future potential 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

watsonx

STEP 4

Comprehensive AI solution including AI optimized infrastructure, runtimes, middleware, data services, governance and applications.

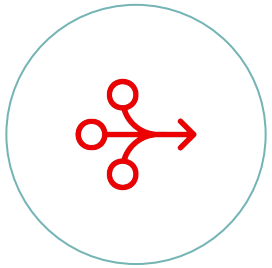
Cluster

The value of Red Hat AI

What differentiates us?

Simplify AI adoption

Promotes freedom of choice and access to latest innovation on AI/ML technologies



Drive AI/ML operational consistency

Streamline the process of moving models from experiments to production

Gain hybrid cloud flexibility

Deploy models in containerized format across on-prem, clouds and edge, including disconnected environments

Things we have discussed

- Red Hat is no stranger to AI
- Our strategy is to deliver choice - hybrid, model, tooling
- We help you use what you have and what is yours - your investment in IT and skills - and your data!

Next steps: Workshops, labs, discovery sessions, customer reference calls, training, PoC, MVP....
We are ready to help. Our partners are eager to assist.
We have the knowledge and the experience to help you wherever you are on your AI journey.

Red Hat Summit

Connect

Thank you

 linkedin.com/company/red-hat

 facebook.com/redhatinc

 youtube.com/user/RedHatVideos

 twitter.com/RedHat

Red Hat