

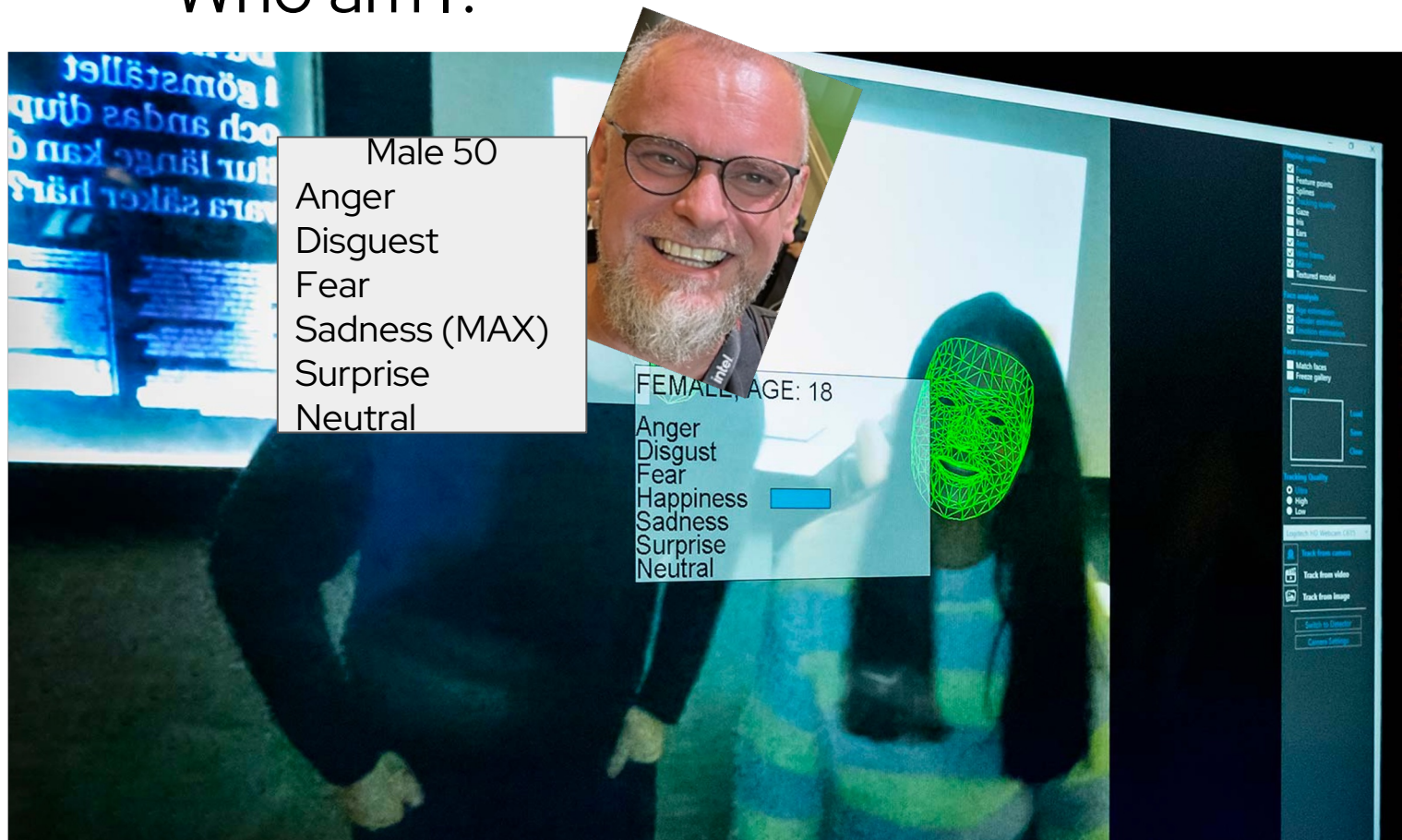


AI and Opensource -
how Red Hat and its ecosystem can
help you survive the AI hype



Who am I?

51
27 (8)
3



Navigate on the page

[Using AI to diagnose breast cancer earlier](#)

[Facts: Every image is reviewed by two radiologists](#)

[More reading](#)

Using AI to diagnose breast cancer earlier

Although mammography has dramatically reduced mortality rates for breast cancer, current technology still misses almost one third of cases. There is also a serious shortage of radiologists. Researchers believe that these problems can be solved by artificial intelligence (AI).

The chances of surviving breast cancer are significantly higher if the disease is detected at an early stage. Studies demonstrate that those who have regular mammographs have a 40% lower risk of dying from breast cancer within 10 years of a diagnosis. Unfortunately, there is a serious shortage of radiologists to examine mammographs.



Every business has a use for AI/ML



Healthcare

- Increased clinical efficiency
- Faster/better diagnosis
- Improved outcomes



Financial services

- More personalized services
- Improved risk analysis
- Reduced fraud
- Better predictions



Telcos

- Better customer insights/experiences
- Optimized network performance & operations
- Improved threat detection



Insurance

- Automated claims processing and handling
- Usage-based insurance services



Automotive

- Autonomous driving
- Predictive maintenance
- Improved supply chains

AI is becoming a part of our everyday lives



Chat GPT

Ansible Lightspeed

with IBM **Watson** Code Assistant



Bard



Bing



**GitHub
Copilot**



DALL-E 2

Chance

THIS CARD MAY BE KEPT
UNTIL NEEDED OR SOLD

GET OUT OF JAIL
FREE



©1935 Hasbro

Why is this relevant to you?

Enterprises are investing in platforms for AI/ML

The tools and technologies are here so it is happening

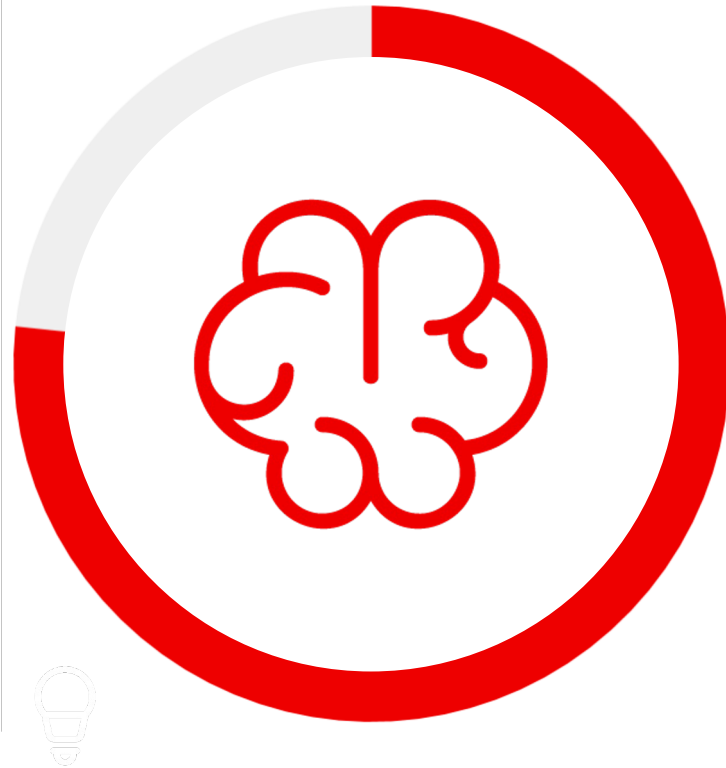
\$13T

AI has the potential to deliver additional global economic activity of around \$13 trillion by 2030.²

51%

of enterprises indicate that their current AI infrastructure will not be able to meet future demands.¹

Open source and cloud-based software to power AI initiatives



69%

of enterprises use a mix of open source and cloud-based software to power AI initiatives.

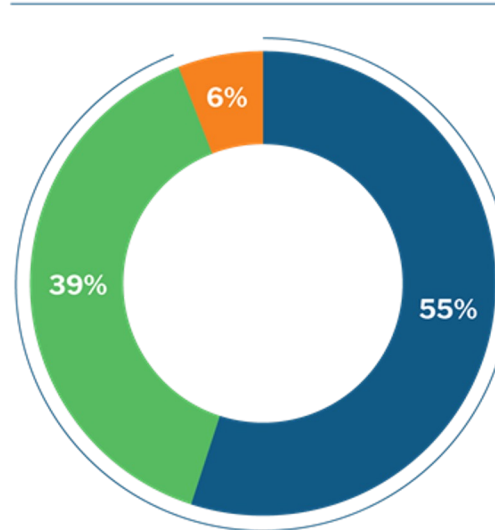
AI Adopters Are Keen on Using Containers To Improve AI Workloads

Benefits of Containers for AI adopters

94%
of AI adopters are using or plan to use containers within one year¹

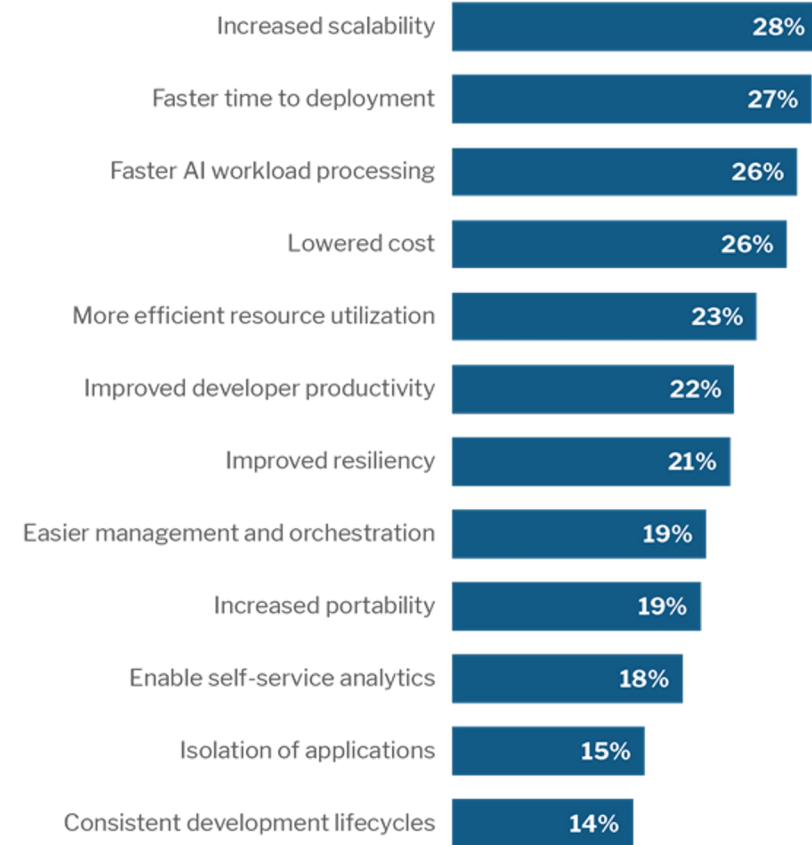
Use of Containers To Deploy Machine Learning Models in Production

94% of AI adopters are using or plan to use containers within one year¹



- Currently using
- Not currently using but plan to within the next year.
- Not currently using containerized environments and no plans in next year

Benefits of Containers



Q: Are you using a containerized environment to deploy machine learning models to production? (n=493)

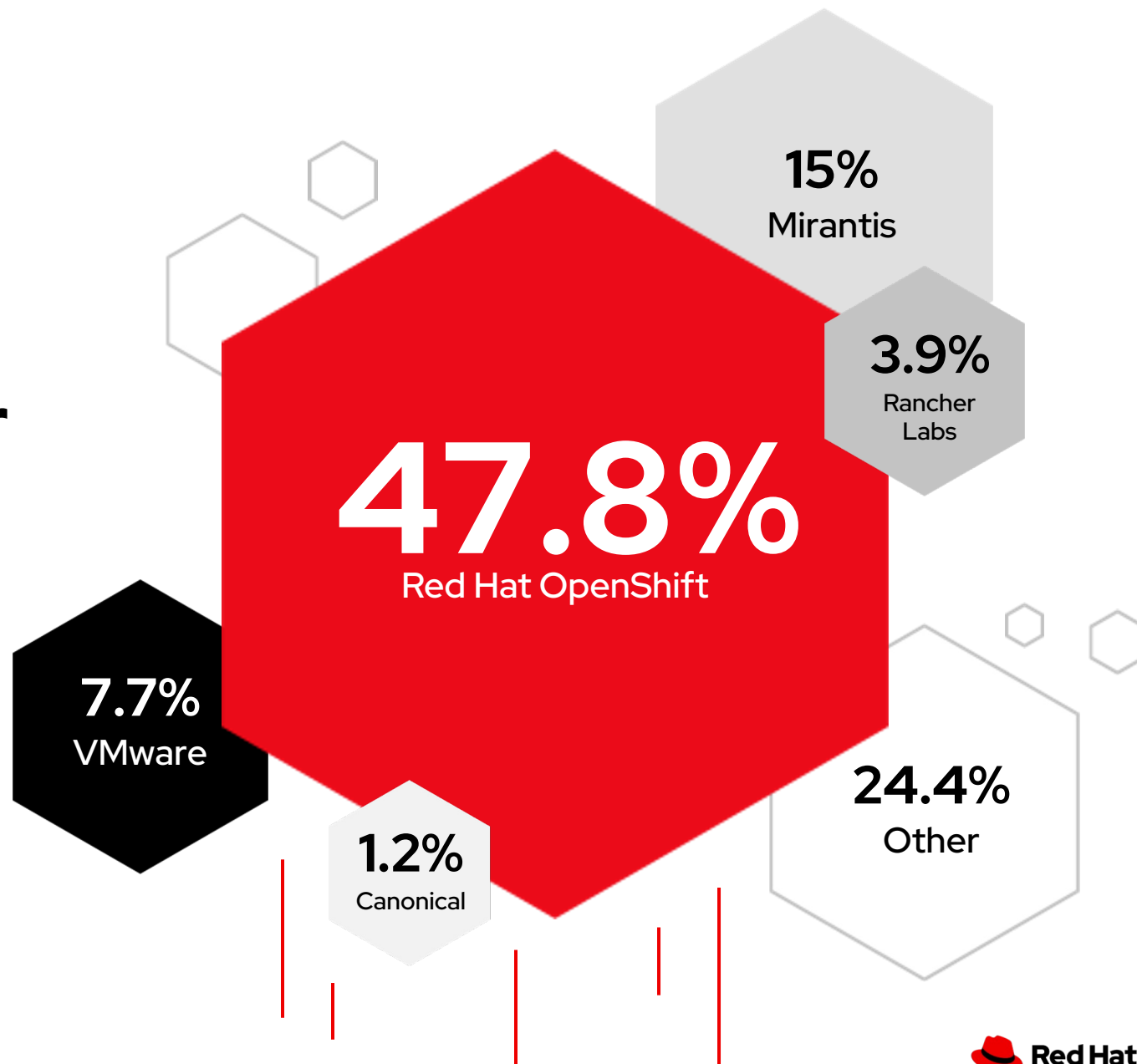
Q: What benefits [do or will] containerized environments bring to your AI initiatives? Please select up to 3. (n=638)

Base: ML is in production or proof of concept, or there are plans to use ML in next 12 months.1

Source: 451 Research, part of S&P Global Market Intelligence – Voice of the Enterprise: AI & Machine Learning, Infrastructure 2020

RED HAT OPENSIFT

Container platform market share leader



Wrap up

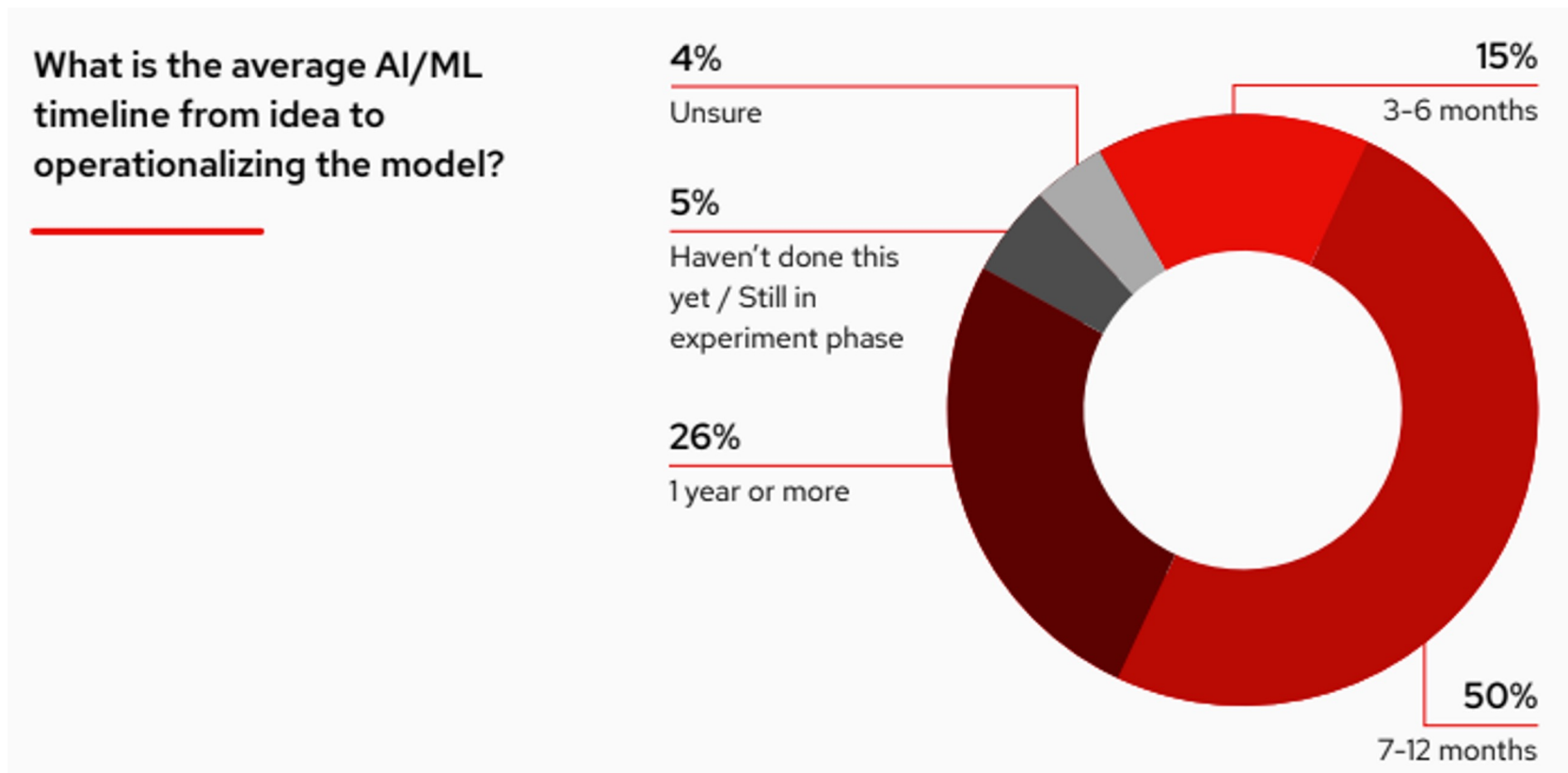
Big investments are going into AI because of clear advantages

Red Hat is a leader in the platform of choice for this development

Red Hat's value is centered in helping you "getting out of jail" by integrating whatever you need into Openshift.

Operationalizing AI is still a challenging process

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



Foundation models - to your rescue?

- **Training**

- Initial process of training foundation models

Very resource and time intensive

- **Fine tuning**

- Tailoring a base model for specific task by creating a new model based on task-specific data
- Requires significantly less data than original training process

Smart optimization on a large range of parameters to a smaller footprint of parameters

- **Prompt tuning**

- Efficient method for adapting foundation model to specific task
- Much less compute intensive than fine tuning - does not retrain model and update model weights

- **Still compute intensive**
- **Requires distributed handling of workloads**

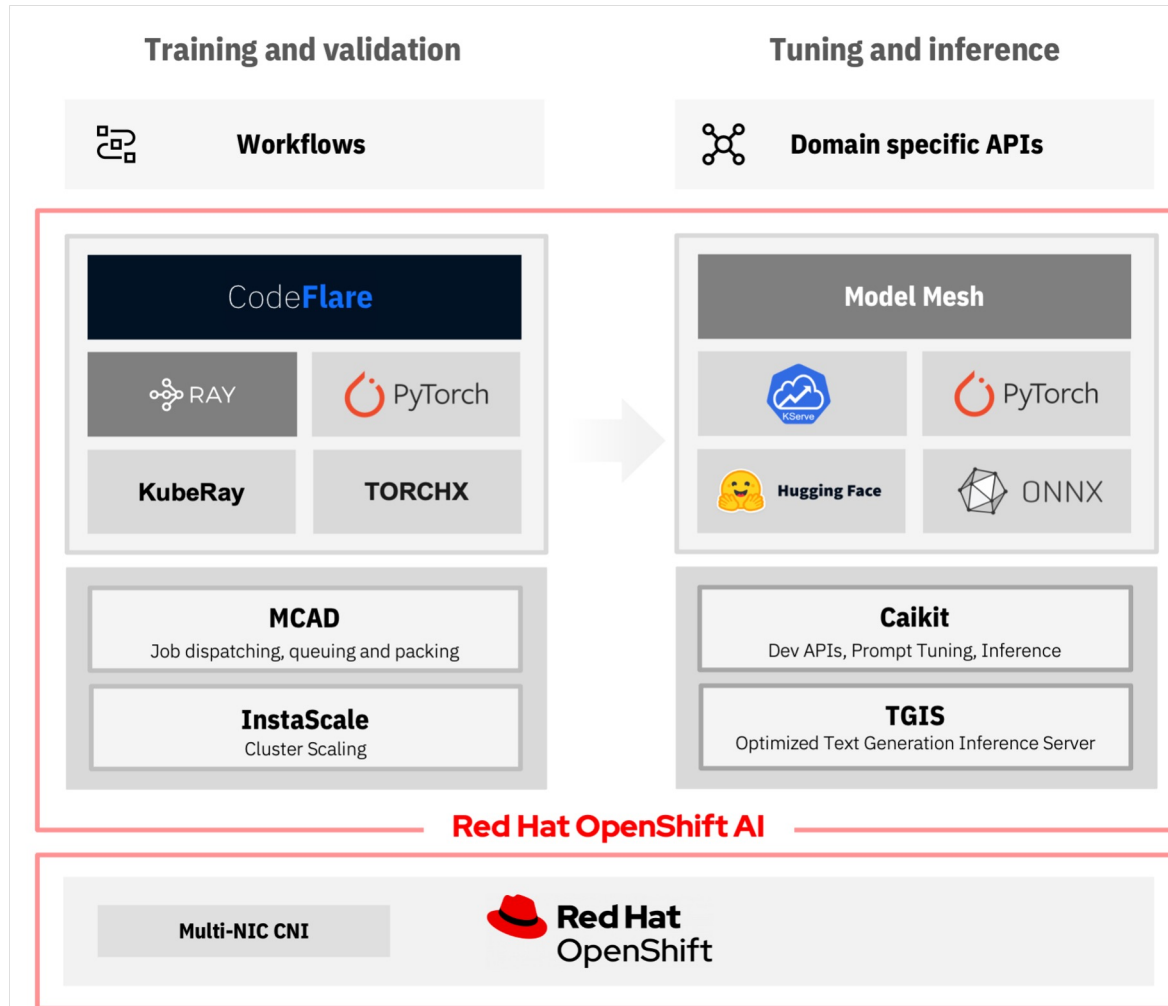
- **Serving**

- Hosting the model as an endpoint to receive input requests
- Inference - model returns result based on input data

Performance, scalability and abstracting API management.

An Open Source Platform for Foundation Models

Deploy, Train or Fine Tune Conversational and Generative AI



First "customer": Ansible Lightspeed with IBM Watson Code assistant

Ansible Lightspeed with IBM Watson Code Assistant is a generative AI service accessed via the Ansible VSCode extension, allowing users to accept and run recommended code directly in their code editing environment while creating Ansible Playbooks.

A *Tech Preview* for the service will be available for all Ansible users in late June, with a commercial offering to follow this fall.

The **IBM Watson Code Assistant** integration is infused with IBM's Ansible foundation model. This foundation model combines Ansible Galaxy data and Red Hat subject matter expertise to deliver highly relevant code automation recommendations that adhere to Ansible best practices.

IBM Watson Code Assistant is built on the **Red Hat OpenShift AI** platform.

Ansible Lightspeed
with IBM Watson Code Assistant



Watson Code Assistant

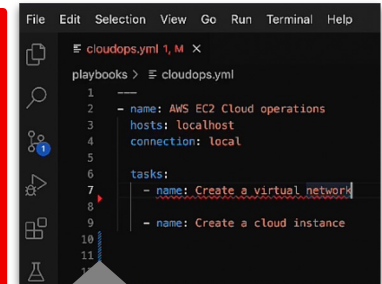
WatsonX



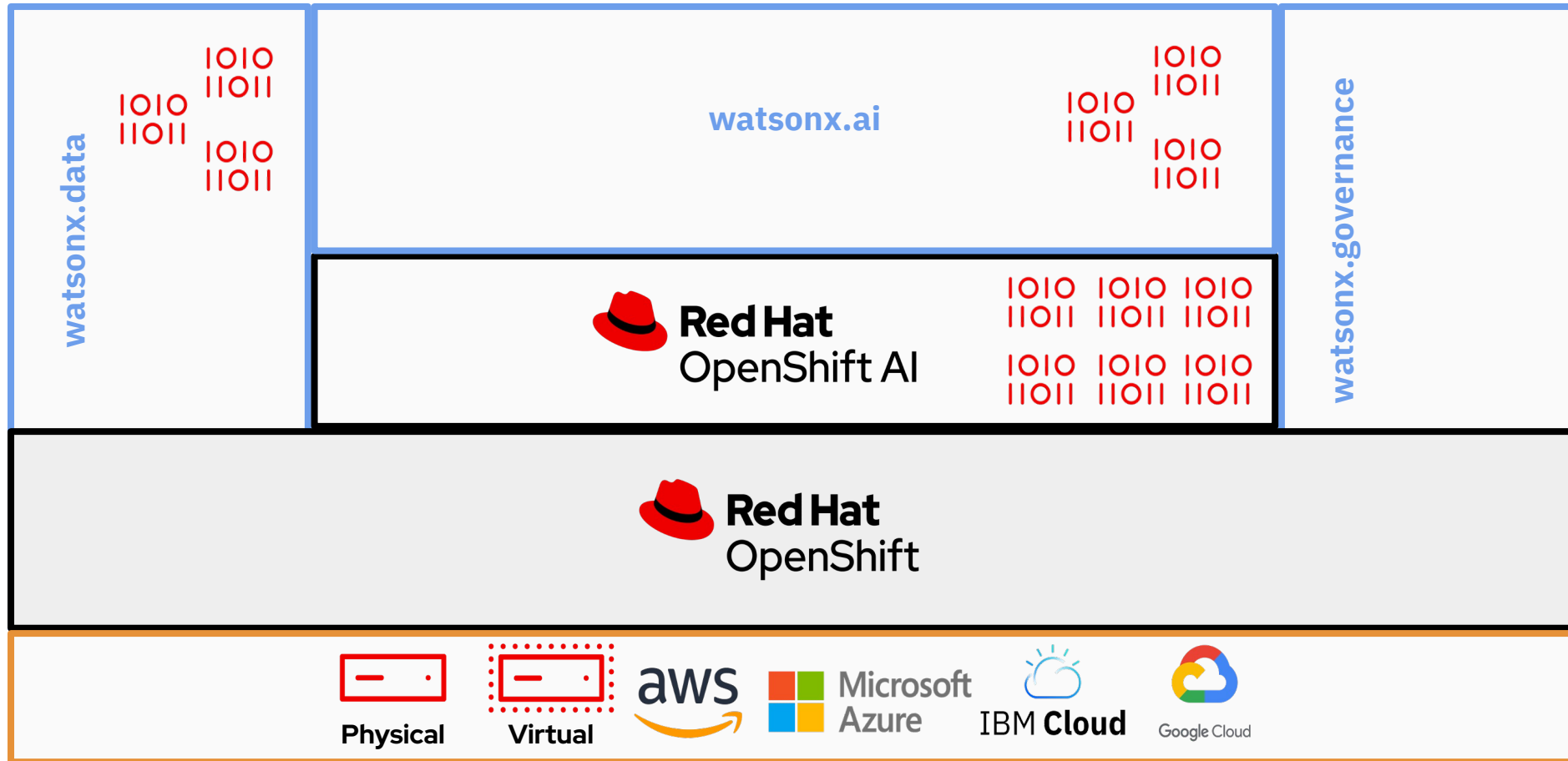
 **Red Hat**
OpenShift AI



 **Red Hat**
OpenShift



```
File Edit Selection View Go Run Terminal Help
cloudops.yml 1, M X
playbooks > cloudops.yml
1
2 - name: AWS EC2 Cloud operations
3   hosts: localhost
4   connection: local
5
6
7 tasks:
8   - name: Create a virtual network
9
10  - name: Create a cloud instance
11
```



OpenShift AI + Watsonx.ai

- ▶ Extend to include data processing, storage and governance along with visual foundation model tuning in an integrated offering with [Watsonx.ai](https://www.ibm.com/watsonx)
- ▶ **Accelerate Generative AI adoption**
 - Using **IBM's suite of curated foundation models** (through IBM's partnership with Hugging Face), **'Bring your own' foundation models** and open source foundation models.
 - Using **Prompt Lab** to customize foundation models with advanced prompt engineering capabilities.
- ▶ Advanced **MLOps capabilities** enabled visually or with code through a unified data+AI collaborative studio.
 - **AutoAI** automates end to end stages in AI/ML Lifecycle.
 - **Automated pipelines** with advanced features such as automated machine learning, model management and model monitoring pipelines.

Red Hat strategy around Generative AI and Foundation models



- ▶ **Training tools developed with IBM Research and open sourcing the infrastructure stack for distributed workloads, scheduling for building, prompt-tuning, fine-tuning and serving foundation models**
- ▶ This stack is being matured in the upstream Open Data Hub project and integrated into Red Hat OpenShift AI
- ▶ OpenShift AI is a **foundation layer for IBM watsonx.ai** and Ansible Lightspeed with IBM Watson Code Assistant
- ▶ **Easy enablement of out-of-the-box “bring your own model” use cases**
- ▶ **Red Hat has no plans to build the actual foundation models**
- ▶ Over time, Red Hat will infuse Generative AI capabilities into more of its portfolio. Ansible Lightspeed is the first example.

AI on Openshift - ecosystem

AI/ML Lifecycle



Data Governance & Security



Data Processing



Data Analytics



Databases



AI Ops



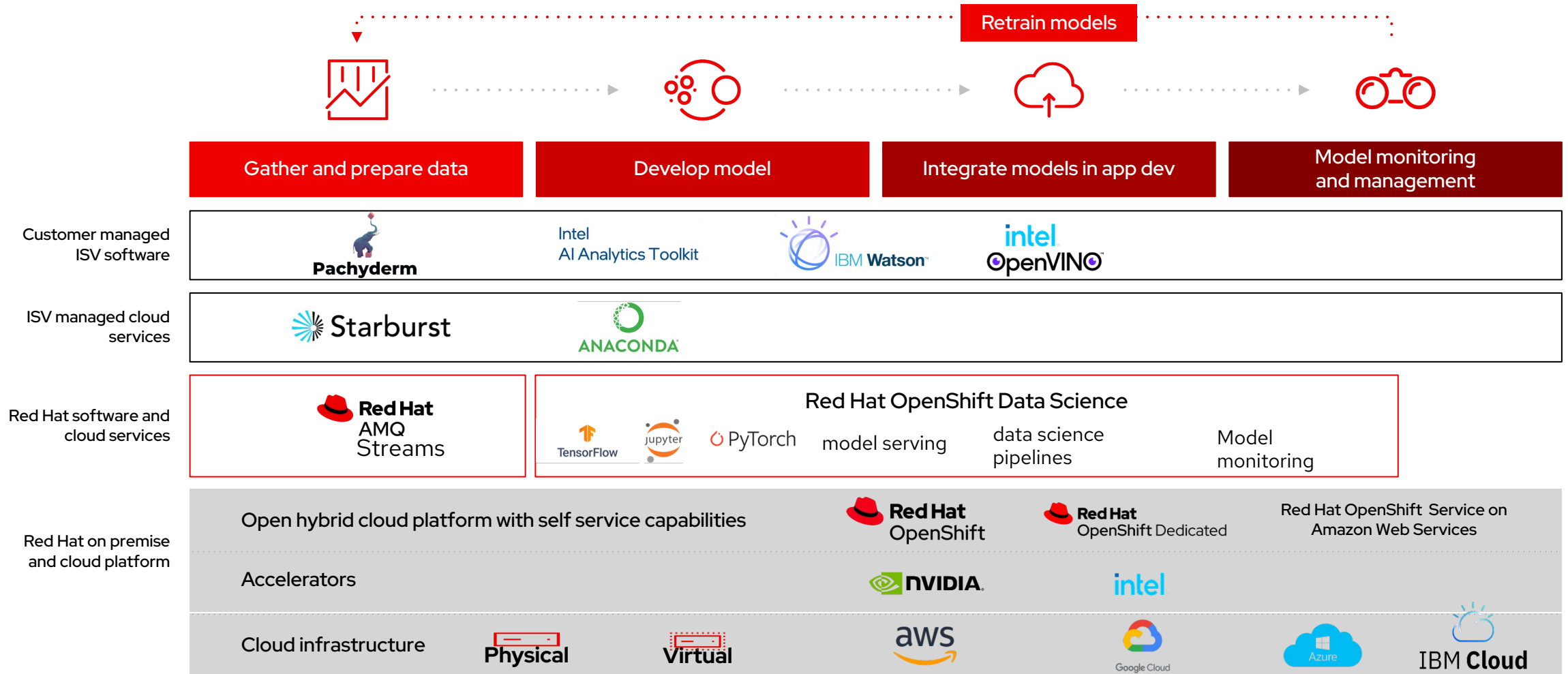
Infrastructure Partners



Hardware Acceleration



Cloud service and self-managed components



Operator Install – just a few clicks and your done!

The screenshot shows the Red Hat OpenShift OperatorHub interface. The top navigation bar includes the Red Hat OpenShift logo and a sidebar menu with categories like Administrator, Home, Operators, Installed Operators, Workloads, Networking, Storage, Builds, Observe, and Compute. The main content area is titled 'OperatorHub' and contains a search bar with 'RHODS' entered. Below the search bar, five operator cards are displayed, each with a logo, a badge (Marketplace, Certified, or Red Hat), and a brief description. The cards are: OpenVINO Toolkit Operator (Marketplace, provided by Intel), OpenVINO Toolkit Operator (Certified, provided by Intel), Prometheus Operator (Red Hat, provided by Red Hat), Red Hat OpenShift Data Science (Red Hat, provided by Red Hat), and RHODS CodeFlare Operator (Red Hat, provided by CodeFlare).

Project: All Projects ▾

OperatorHub

Discover Operators from the Kubernetes community and Red Hat partners, curated by Red Hat. You can purchase commercial software through [Red Hat Marketplace](#). You can install Operators on your clusters to provide optional add-ons and shared services to your developers. After installation, the Operator capabilities will appear in the [Developer Catalog](#) providing a self-service experience.

All Items

Search: RHODS

5 items

- OpenVINO Toolkit Operator** (Marketplace) provided by Intel
OpenVINO Toolkit Operator manages OpenVINO components in OpenShift. Currently there...
- OpenVINO Toolkit Operator** (Certified) provided by Intel
OpenVINO Toolkit Operator manages OpenVINO components in OpenShift. Currently there...
- Prometheus Operator** (Red Hat) provided by Red Hat
Manage the full lifecycle of configuring and managing Prometheus and Alertmanager...
- Red Hat OpenShift Data Science** (Red Hat) provided by Red Hat
Operator for deployment and management of Red Hat OpenShift Data Science
- RHODS CodeFlare Operator** (Red Hat) provided by CodeFlare
CodeFlare allows you to scale complex pipelines anywhere ***
This operator is not ready to be...

Very often part of the process: 1 day Workshop

Time	Session
09:00 - 09:15	Welcome & Acquaintance
09:15 - 09:45	Data Science at Infineon
09:45 - 10:15	Implementing MLOps on Open Hybrid Cloud the Red Hat Way
10:15 - 10:30	<i>Short Break</i>
10:30 - 10:45	OpenShift Data Science (Product Overview, Walkthrough)
10:45 - 11:00	Lab Walkthrough / Use Cases / Testing Access
11:00 - 12:00	Hands-On Lab - Implementing Object Detection with RHODS (Part 1)
12:00 - 13:00	<i>Lunch</i>
13:00 - 15:00	Hands-On Lab - Implementing Object Detection with RHODS (Part 2)
15:00 - 15:15	<i>Short Break</i>
15:15 - 16:00	Feedback Gathering, Roadmaps, Next Steps & Additional Slides (if time permits)

Can be done virtual as well as onsite.

Best results when we have teams of DS, APP dev, ML engineers

Chance

THIS CARD MAY BE KEPT
UNTIL NEEDED OR SOLD

GET OUT OF JAIL
FREE



©1935 Hasbro

software.

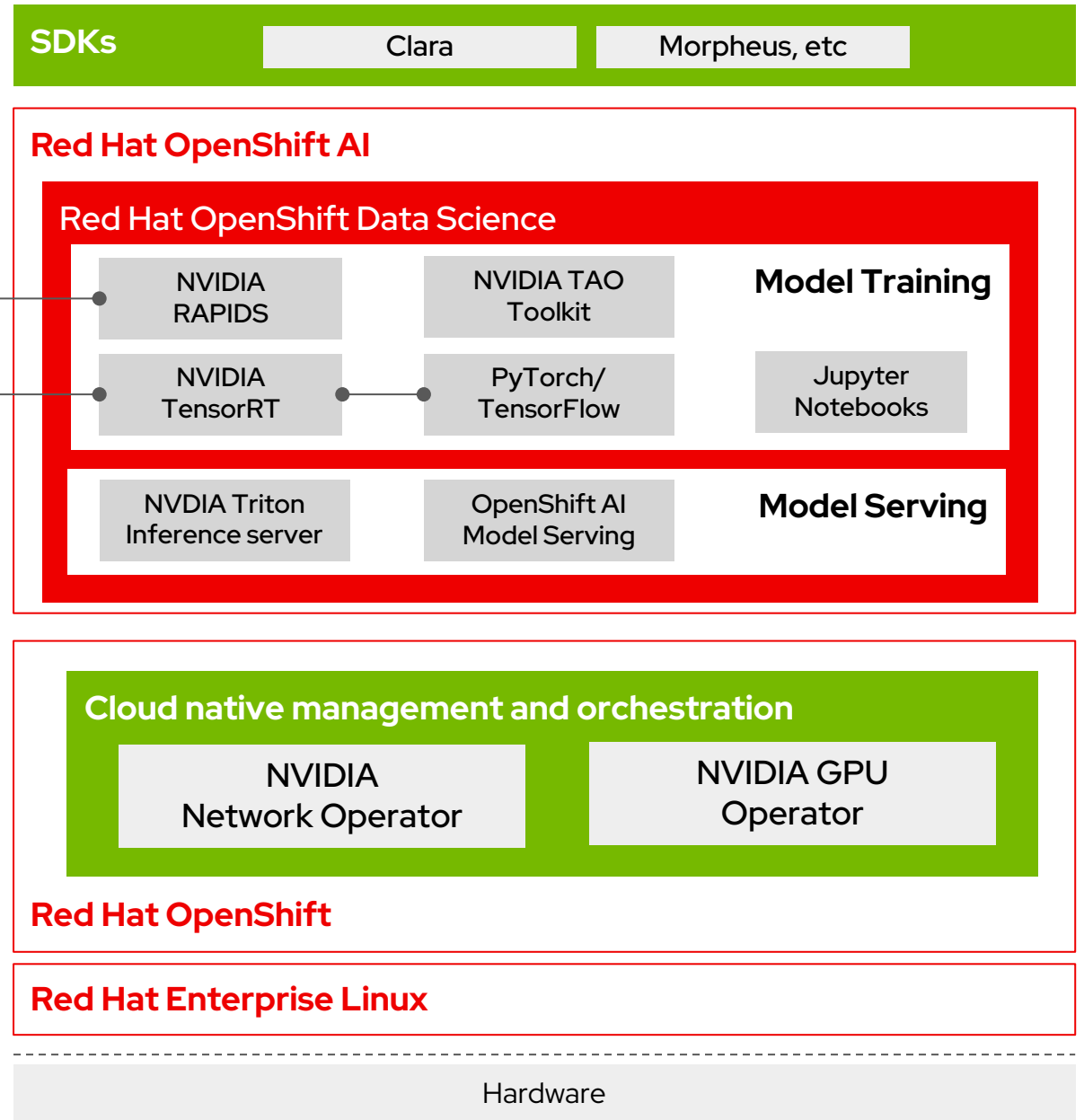
Important partner: nVidia

NVIDIA RAPIDs + OpenShift Data Science

Accelerate model training time by accessing data science libraries (numpy, pandas, scikit-learn, etc.) through Red Hat OpenShift Data Science Notebooks.

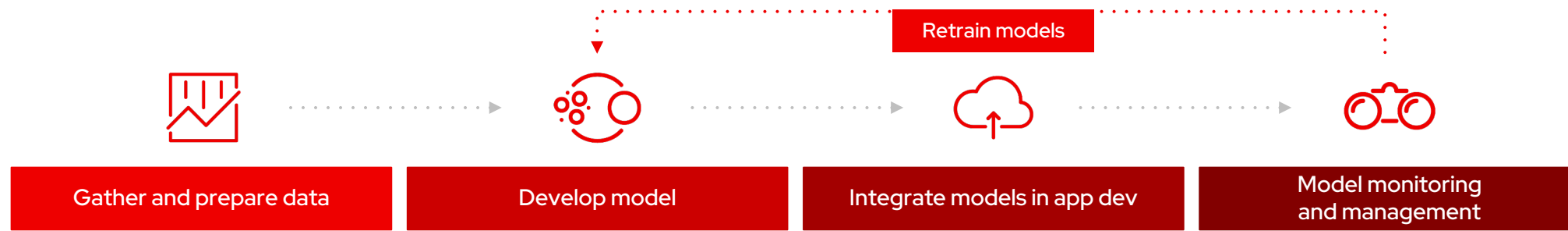
TensorFlow, PyTorch & NVIDIA TensorRT + OpenShift Data Science

Leverage GPU optimized deep learning and standard frameworks directly from Red Hat OpenShift Data Science Notebooks.



Important partner: Intel

Accelerate data science using Intel hardware



Accelerate model training Intel
AI Analytics Toolkit
Out-of-the-box speed with AI Analytics Toolkit

Intel AI Analytics Toolkit (AI Kit) Benefits with RHODS

- **Drop-in acceleration** with minimal code changes directly in notebooks
- Use low-level **optimizations** with popular Python AI frameworks
 - Tensorflow, PyTorch, NumPy & more on heterogeneous architectures
 - Speed up CPU intensive packages: Pandas, Scikit-Learn, & XGBoost
- **High Performance Intel Python distribution offers optimized and distributed compute.** Scale Pandas and Scikit-learn CPU and GPU workloads to multiple cores and nodes with minimal code changes.
- Increased model **accuracy** and **performance** using optimized algorithms within scikit-learn and XGBoost
- **Quantization** capabilities with the Intel Neural Compressor
- **Automated retraining and transfer learning**

Accelerate model inference OpenVINO™
High performance inference using Intel CPUs

Intel OpenVINO Benefits with RHODS

- **High performance model inference** from edge to cloud
 - Support for multiple Deep Learning frameworks including TensorFlow, Caffe, PyTorch, MXNet, Keras, ONNX
 - Applicable to Machine & Deep Learning tasks: computer vision, speech recognition, natural language processing, and more
- **Easy Deployment of Model Server** at Scale in Kubernetes and OpenShift
- **Support multiple storage options** (S3, Azure Blob, GSC, local)
- **Configurable Resource Restrictions and Security Context** with OpenShift resource requirements
- **Quantization**
- **Configurable Service Options** based on infrastructure requirements

References:

- [AI Analytics Toolkit](#)



Important partner: Starburst

Data Services for Modern AI/ML Use Cases

Performance

From petabytes to exabytes – query data from disparate sources using SQL – with high concurrency

Control your price/performance with the latest cost-based optimizer

Caching available for frequently accessed data

Connectivity

40+ supported enterprise connectors

High performance parallel connectors for Oracle, Teradata, Snowflake and more



Security

Kerberos, LDAP & SSO Integration

Global Security for fine-grained access control

Data Encryption/Masking

Higher security posture than vanilla K8's



Management

Configuration

Autoscaling & High Availability

Query/Cluster Monitoring

Deploy Anywhere

Multi-Cluster Management



AI on Openshift - more with the ecosystem

AI/ML Lifecycle



Data Governance & Security



Data Processing



Data Analytics



Databases



AI Ops



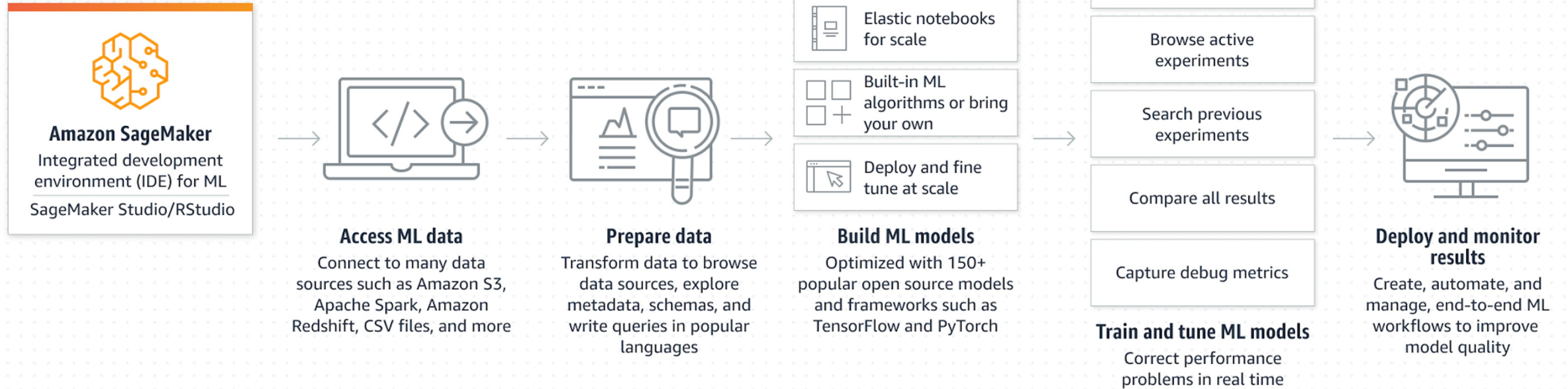
Infrastructure Partners



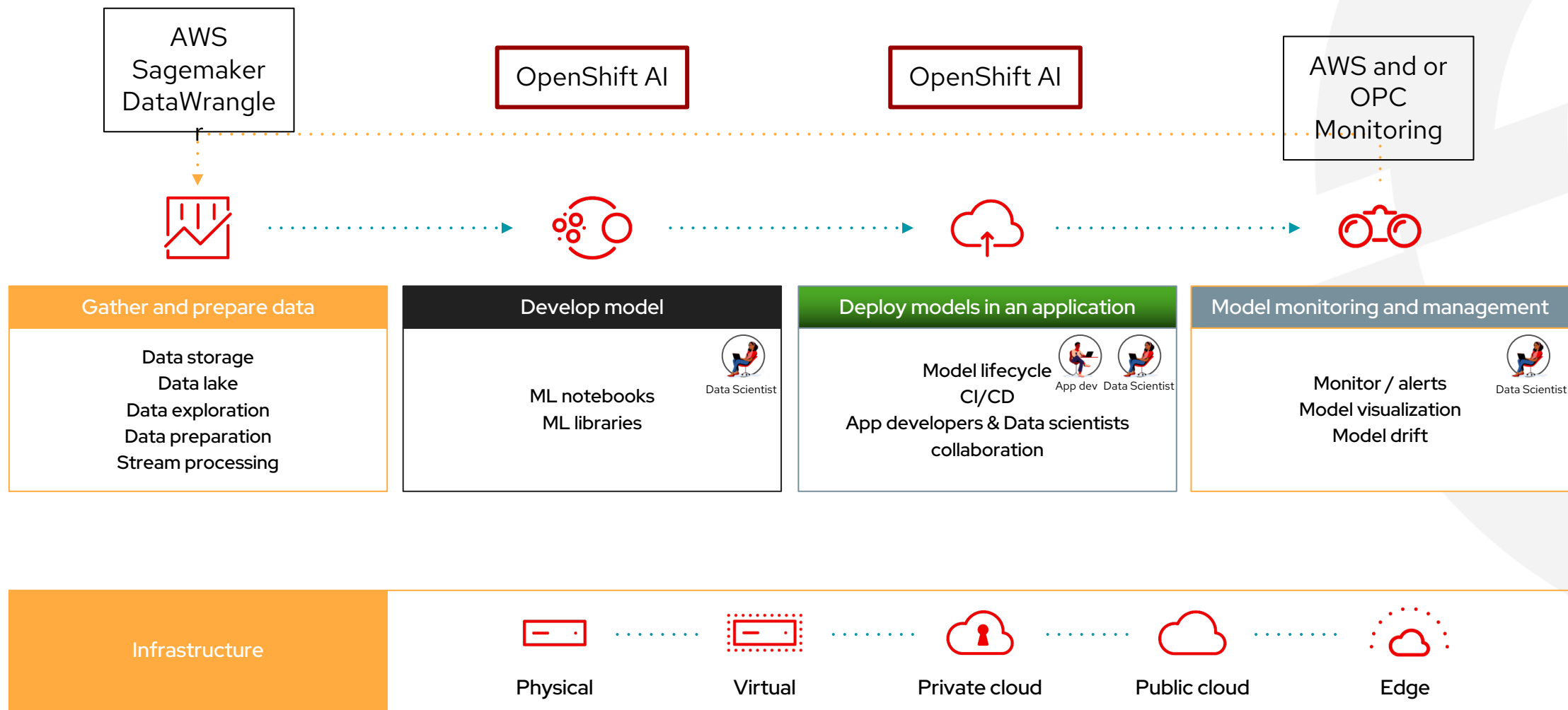
Hardware Acceleration



AWS Sagemaker process



Integrate AWS SageMaker and OpenShift AI on ROSA



Runs on anything

Our approach to AI/ML

1

AI Workload Support

Become the preferred choice for AI deployments by optimizing for the **requirements of AI workloads on Red Hat platforms** allowing users to **Deploy AI across Hybrid Cloud**.

e.g., hardware acceleration

2

Platform for AI enable apps

Accelerate time to value by providing a consistent, purpose-built **application development platform for customers** to build and deploy AI-enabled applications.

e.g., Red Hat OpenShift AI

3

AI enabled Products

Increase user productivity through **AI-enabled tools and services to drive adoption of existing Red Hat products and services**.

e.g., Ansible Lightspeed, Developer Hub

Ecosystem integration

Things you can do

1. Log into developer.redhat.com and test the OCP AI in the sandbox there
1. Invite us to do a workshop with your team - talk to me or your local contacts
1. Download and run the OCP AI Operator and try it out in your DC. If you let us know you can get 60 days of support but you can get it up in secret
1. If you want to test the managed service you can get 60 days for free on AWS (T&Cs apply)

Chance

THIS CARD MAY BE KEPT
UNTIL NEEDED OR SOLD

GET OUT OF JAIL
FREE



©1935 Hasbro

Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.

 [linkedin.com/company/red-hat](https://www.linkedin.com/company/red-hat)

 [youtube.com/user/RedHatVideos](https://www.youtube.com/user/RedHatVideos)

 [facebook.com/redhatinc](https://www.facebook.com/redhatinc)

 twitter.com/RedHat