

Data Science on OpenShift Lunch and Learn Series

Do Math!!!

John Archer - Chief Architect Energy

Faizal Kadher - Solution Architect



MEET THE SPEAKERS



John Archer

Chief Architect Energy

Red Hat since 2015

Previously with BEA Systems, BSI Consulting, DocuQuest Systems, Andrews & Kurth LLP, SilverStream, Petris and Oracle

OSDU Focal point, Upstream Data Management, DoD, APIs, eCommerce, IoT, data science and blockchain

Open Group OSDU and OPAF, SPE, SEG, PPDM, HJUG, HDUG, Energistics



Faizal Khader

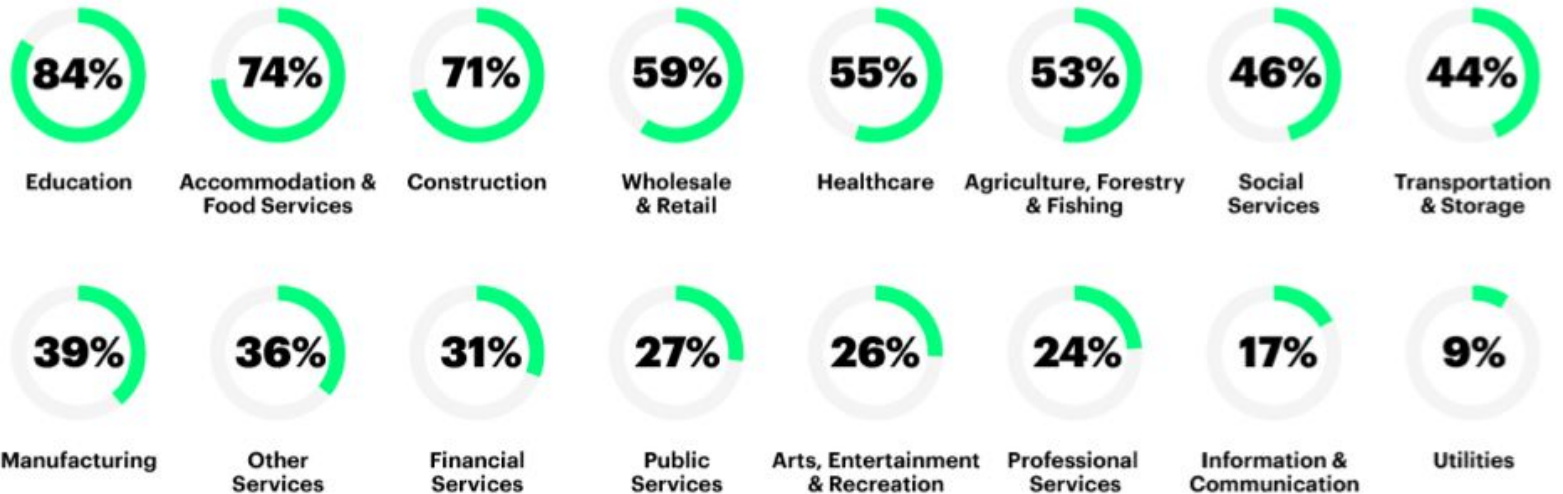
Senior Specialist Solution Architect

Covers Openshift and AppDev portfolio

Previously with TIBCO Software and EDS
Working experience with many
middleware products

Why Data Science Matters?

The impact of AI on profits by industry



Source : [Accenture Report: Artificial Intelligence Has Potential to Increase Corporate Profitability in 16 Industries by an Average of 38 Percent by 2035](#)

87

Why do 87% of data science projects never make it into production?

Customize and control Google Chrome

VB Staff July 19, 2019 4:10 AM AI

f t in



TLDR Version

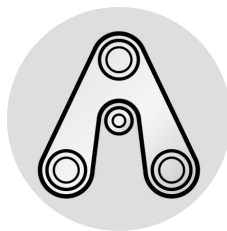
- Culture and Silos
- Lack of Collaboration
- Poor Internal Communication
- Data Access and Quality
- DevOps/MLOps Maturity

HOW RED HAT SEES AI



AI as a Workload

Represents a workload requirement for our **platforms** across the hybrid cloud.



AI Accelerates Open Source Development

Applicable to Red Hat's existing core business in order to increase **Open Source** development and production **efficiency**.



AI Enhances RH Products & Services

Valuable to our customers as specific services and product capabilities, providing an **Intelligent Platform** experience.

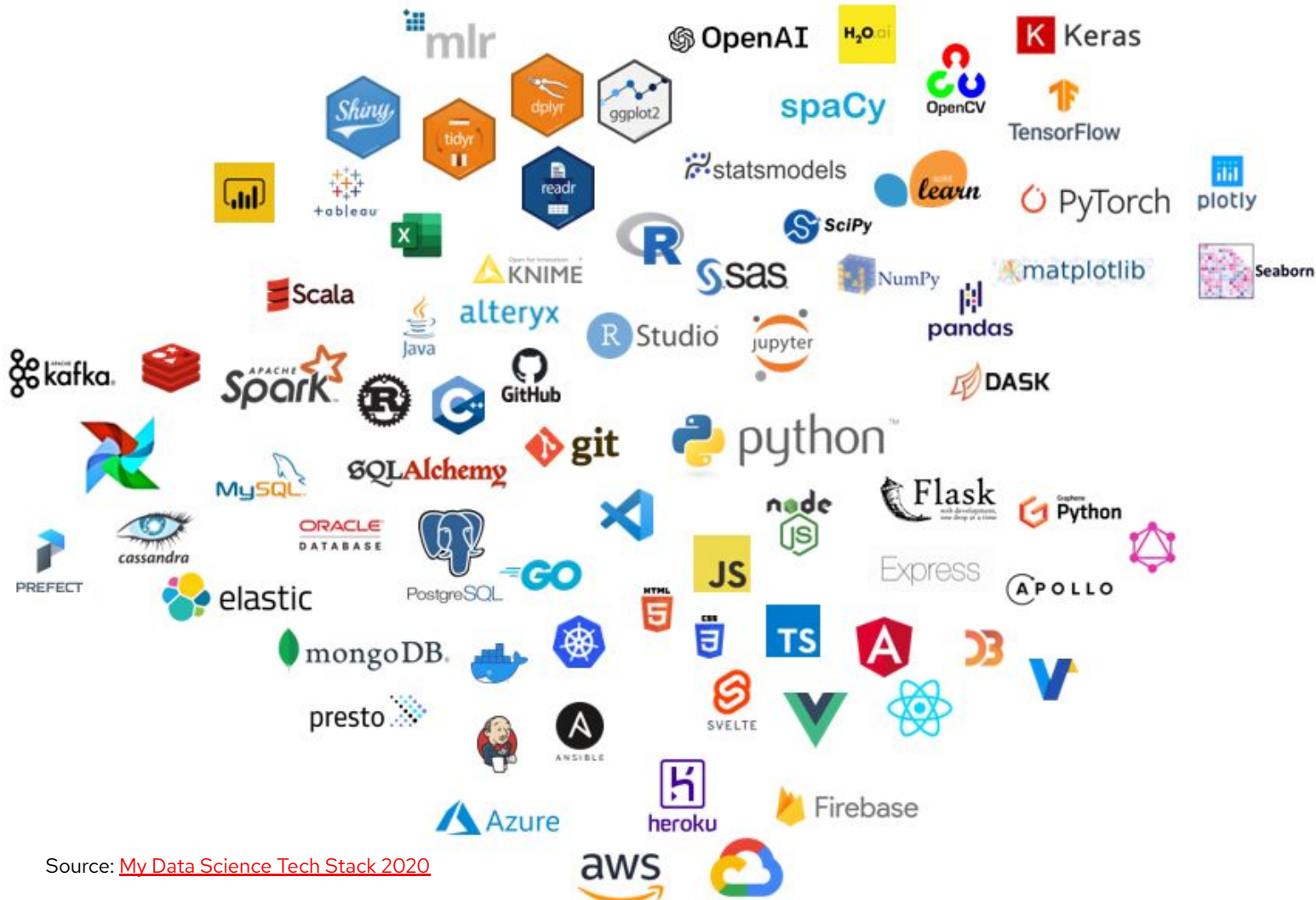


AI Enables Customers

to build **Intelligent Apps** using Red Hat products as well as our broader partner ecosystem.

010110
101010

Data as the Foundation





Source: [My Data Science Tech Stack 2020](#)

 TensorFlow




 python™



 pandas  plotly
 statsmodels



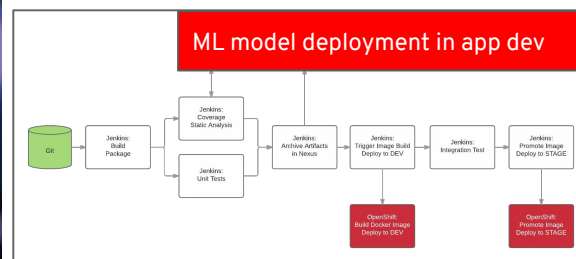
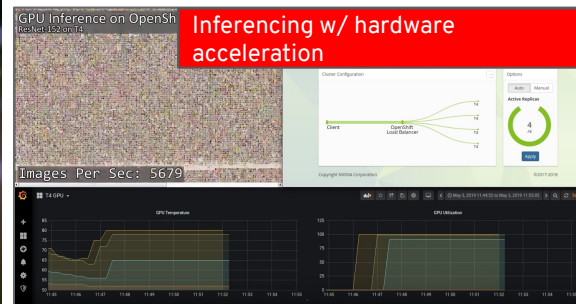
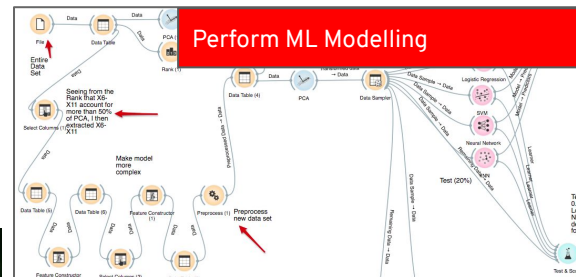
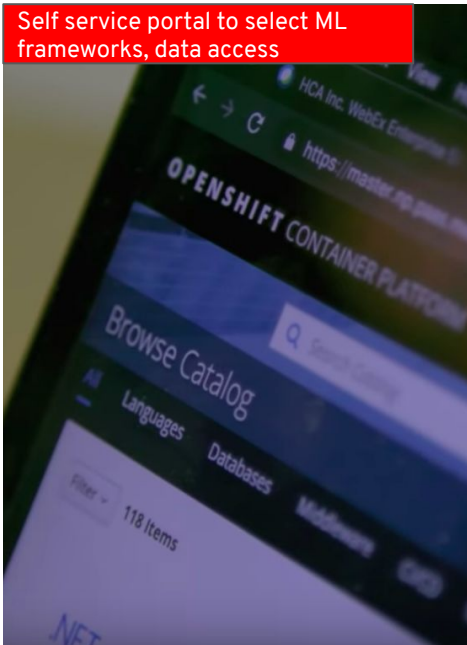
 DASK  PREFECT
 PostgreSQL

 mongoDB.

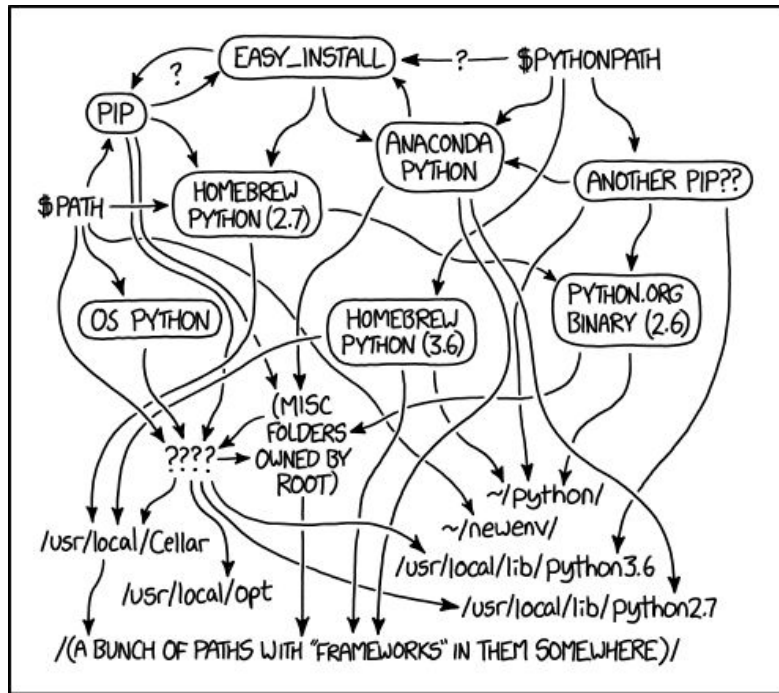


What does a Data Scientist Want?

As a Data Scientist, I want a “self-service cloud like” experience for my Machine Learning projects, where I can access a rich set of modelling frameworks, data, and computational resources, share and collaborate with colleagues, and deliver my work into production with speed, agility and repeatability to drive business value!



Data Science Team Pressures



EXPLOSIVE GROWTH

in data analytics teams and analytic tools

MULTIPLE TEAMS COMPETING

for use of the same storage and computing resources

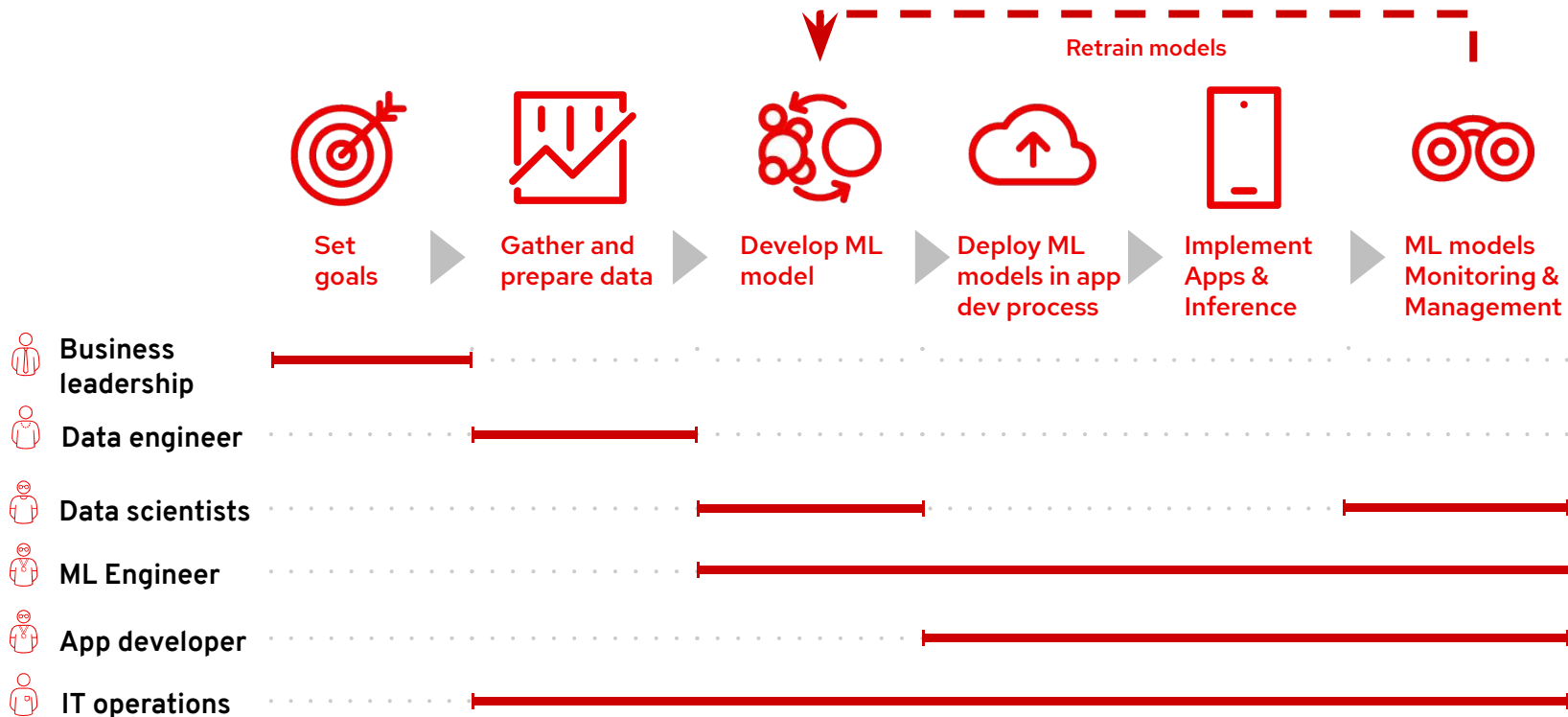
CONGESTION

in busy analytic clusters causing frustration and missed SLAs

EMERGING MLOPS

Data Scientist Developers vs Full Stack Developer agility and enablement gaps

AI/ML lifecycle and key personas



Challenges facing key personas



Gather and prepare data



Develop ML model



Deploy models in app dev process



Implement apps



ML models monitoring & management



Data engineer

Data acquisition

Duplication of data

Data management architecture



Data scientists

Data access/hygiene
Collaborating with data engineers and devs

Selecting & deploying the right tools

Dependency on IT ops
Poor performing hardware

Deployment in dev processes

Training, testing, and retraining models



App developer

Clashes between models and app code
Building APIs

Disjointed ML workflow

Deployment in dev processes

Coordinating retraining and redeployment



IT operations

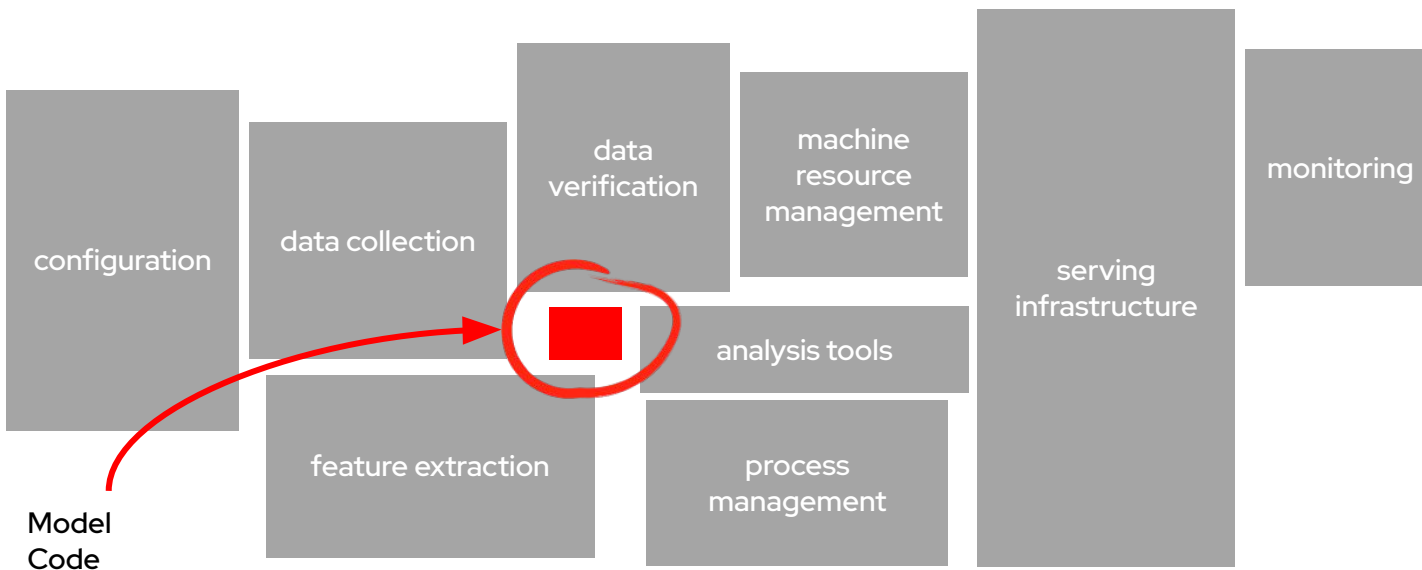
Managing security and performance

Support multiple, incompatible tools

Slow & legacy IT infrastructure and outdated processes

Intelligent applications are distributed systems

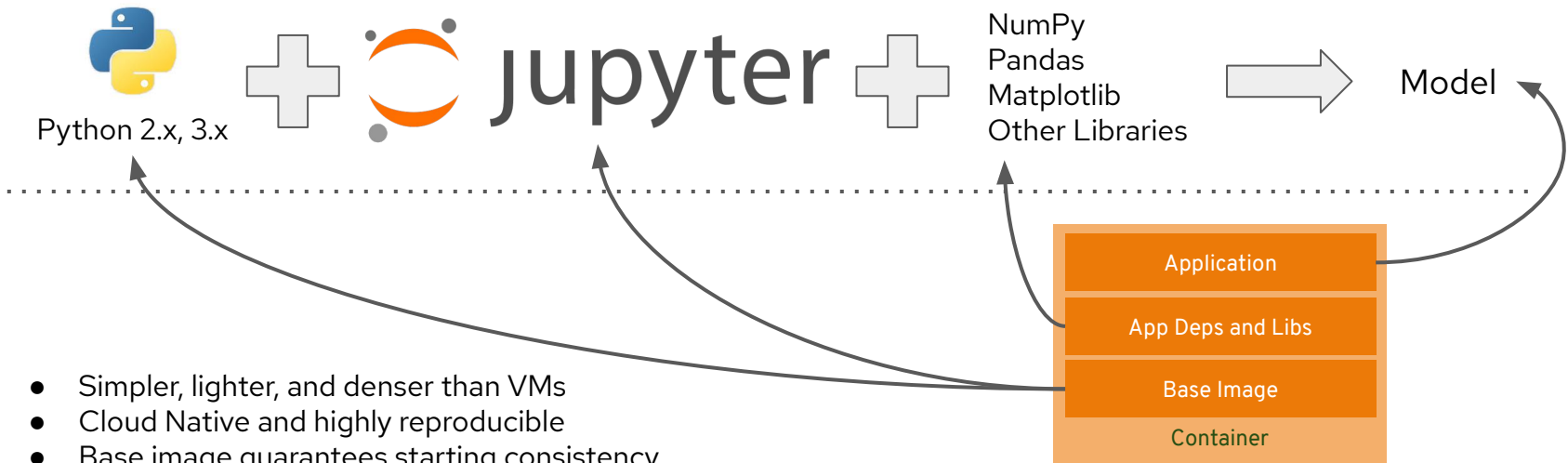
The model is a tiny part, but let's focus on it



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

Containers for data analysis

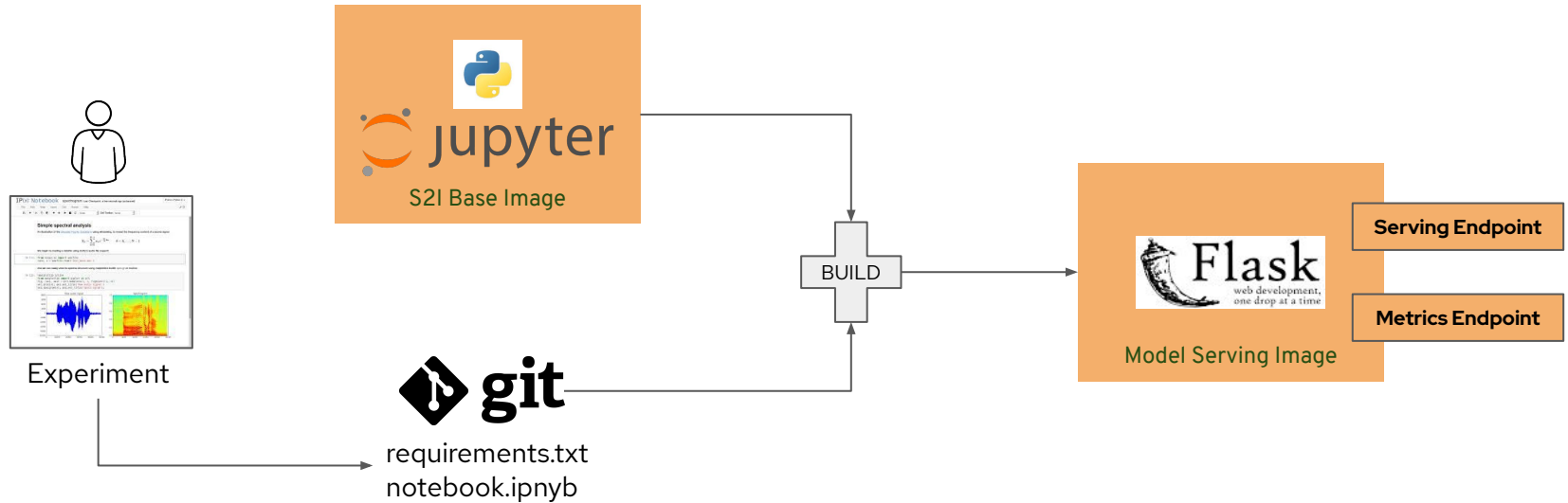
Reproducible and shareable environments



- Simpler, lighter, and denser than VMs
- Cloud Native and highly reproducible
- Base image guarantees starting consistency
- Package apps with all dependencies
- Reusable, portable, shareable as a unit

Source-to-Image for models

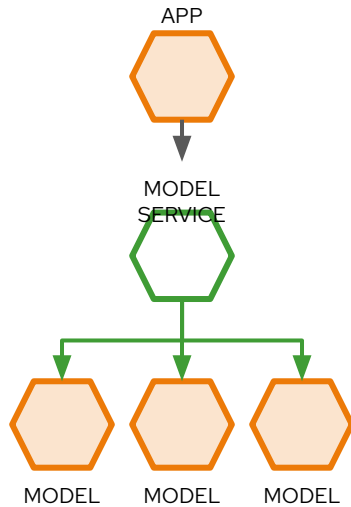
Automating build, deployment, APIs and instrumentation



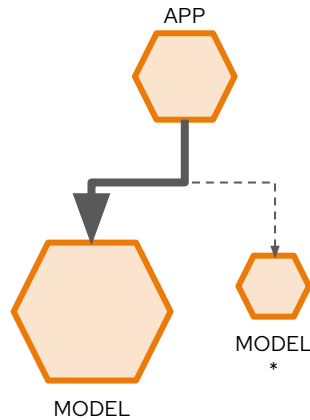
Models as stateless microservices

The power of Kubernetes

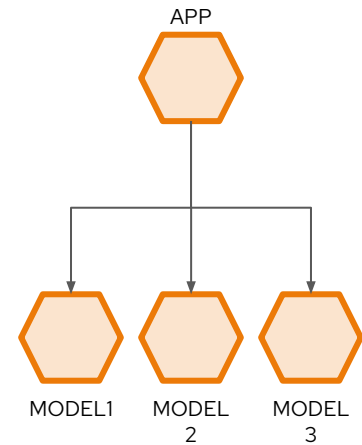
SCALE
HORIZONTALLY



PHASED
ROLLOUTS

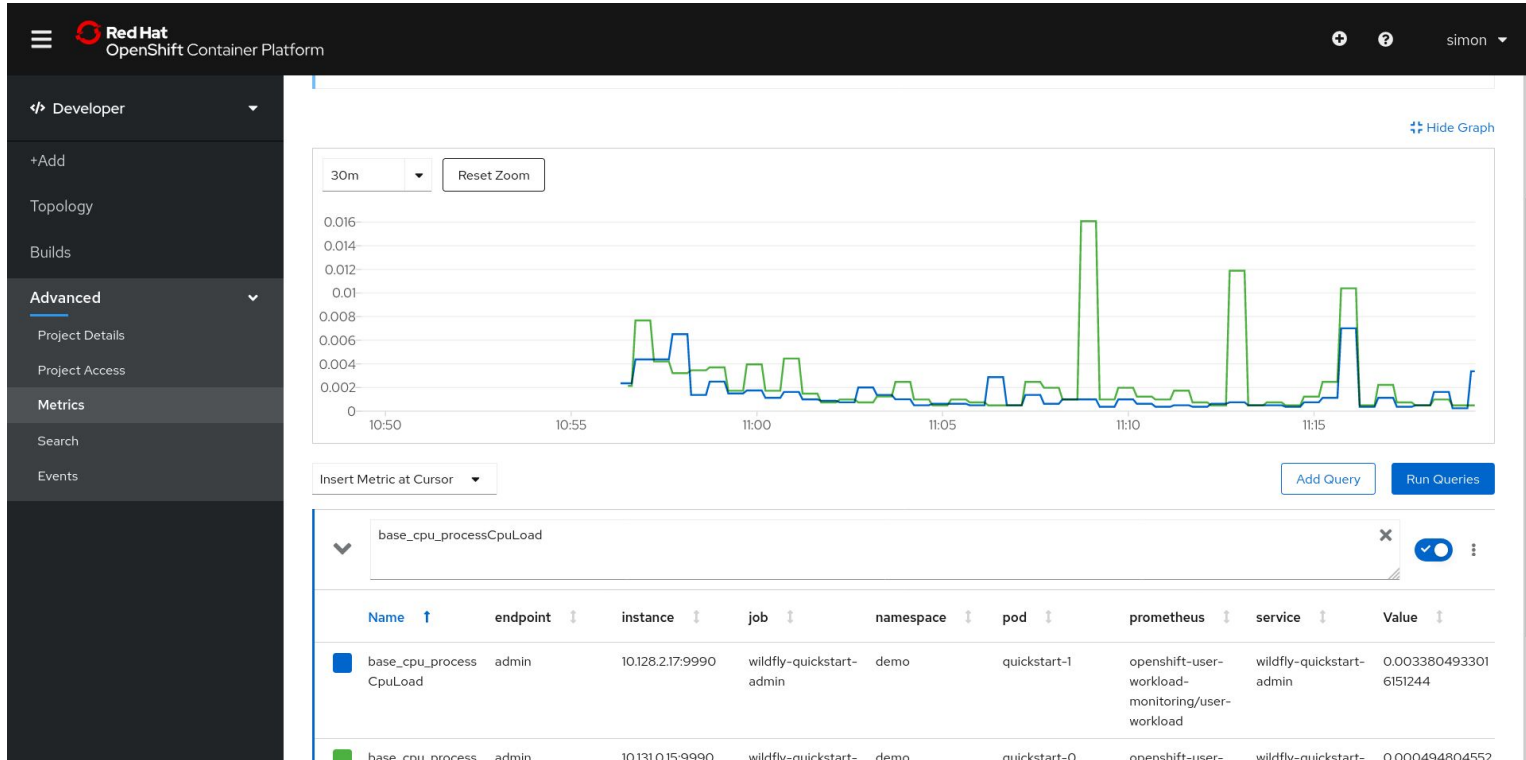


MULTIPLE
TRIALS

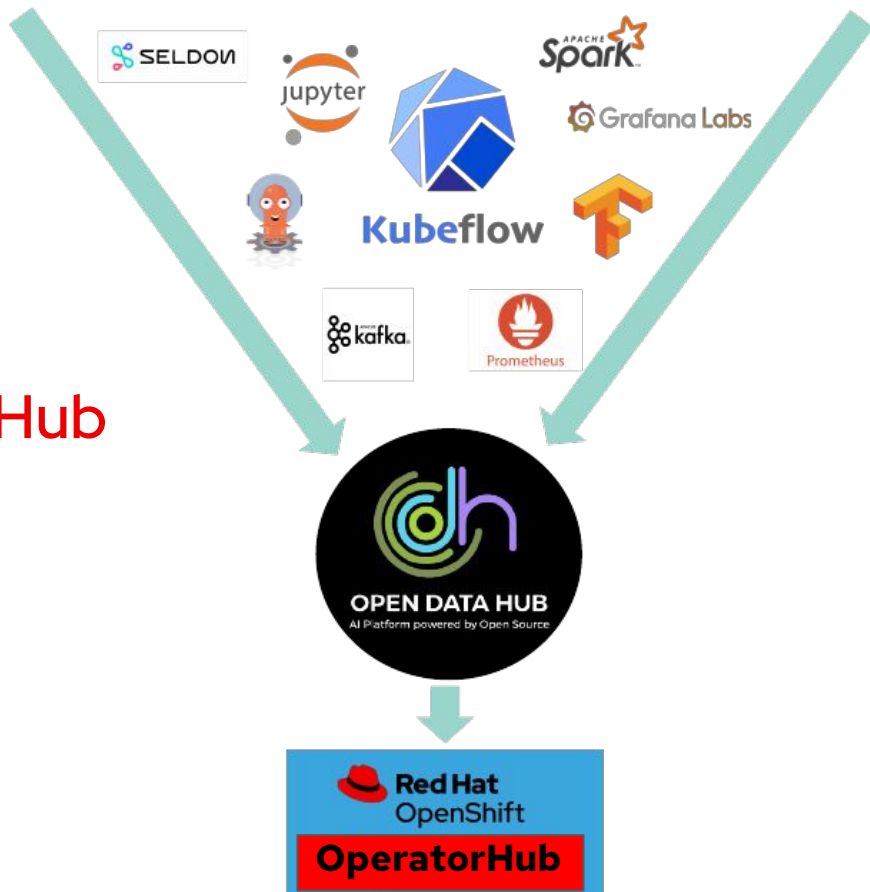


Monitoring performance and drift

Keeping track of models over time



Open Data Hub



Goals

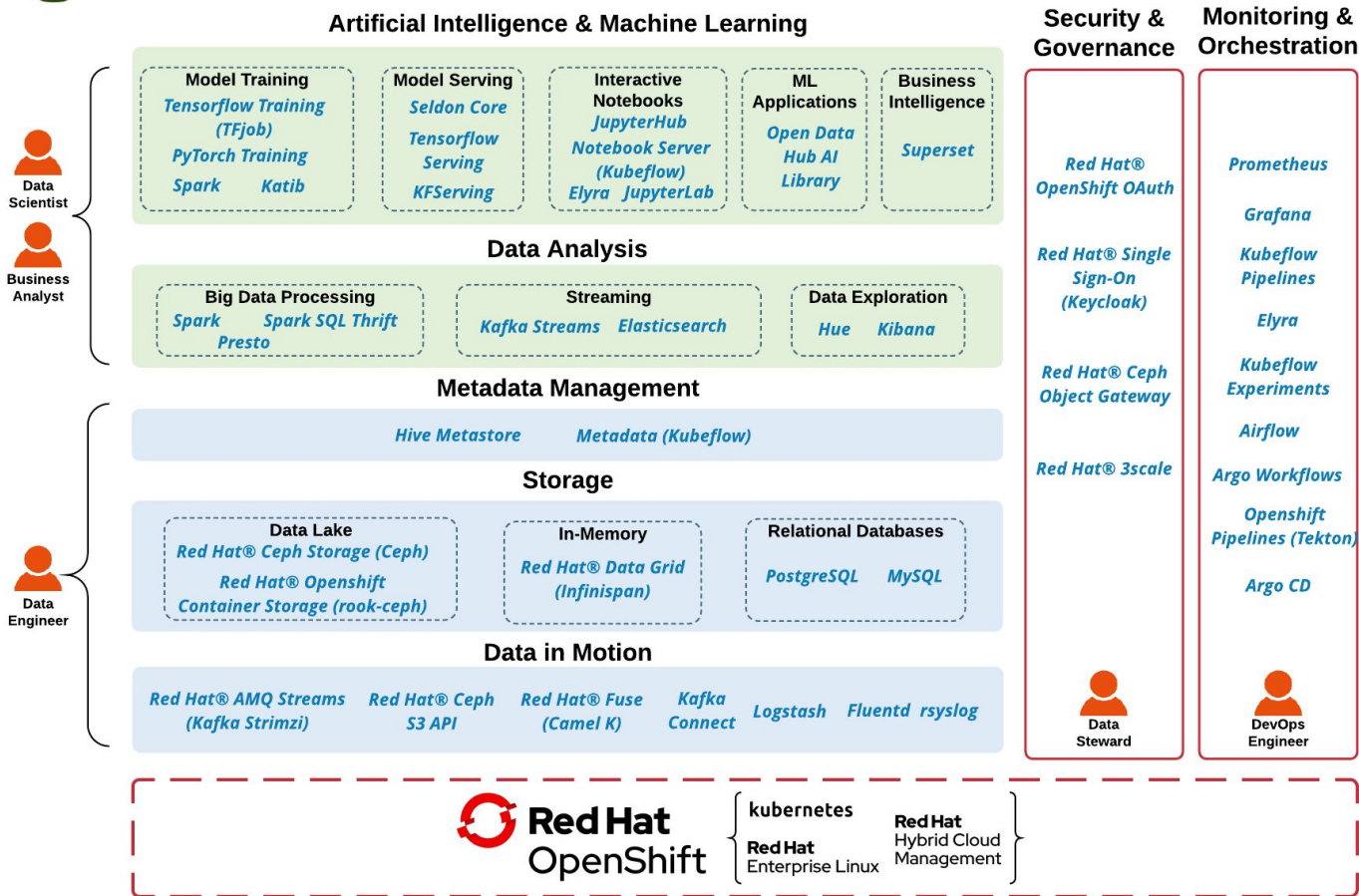
- Provide an end-to-end AI/ML platform on OpenShift
- One stop easy operator deployment for the platform on OCP
- Provide Tools for each stage in the AI/ML platform and for all AI/ML user personas optimized for OpenShift
- Provide monitoring tools for model and services used by DevOps
- Provide development tools for Data Scientists
- Provide ETL tools used by Data Engineers
- AI/ML pipelines and long processing tasks.



OPEN DATA HUB

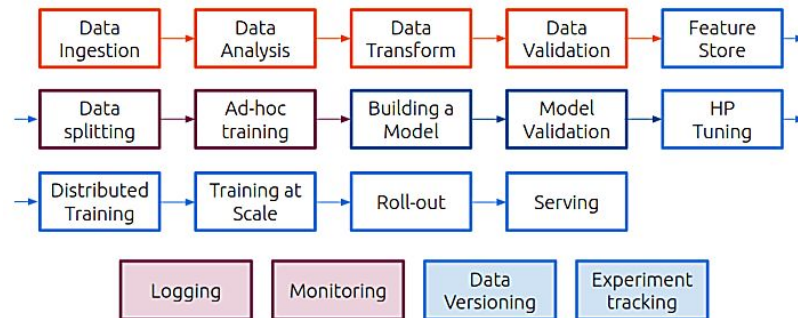
AI Platform powered by Open Source

Reference Architecture for AI on OpenShift



Kubeflow Overview

- ▶ **ML - Many steps**
 - **Maintain**
 - **Share**
 - **Collaborate**
- ▶ **Multiple environments**
- ▶ **Container's can isolate workloads**
- ▶ **Less learning for DS**



- ▶ **Kubeflow**
 - **Codified representation of ML - Pipelines**
 - **Reproducible**
 - **Composable**
 - **Scalable**
 - **Shareable**

```
mnist-pipeline-4ff6-771465083
Artifacts  Input/Output  Volumes  Manifest  Logs
-----
26 dropout (Dropout) (None, 612) 0
27 dens_1 (Dense) (None, 10) 5130
28 .....
30 Total params: 407,050
31 Trainable params: 407,050
32 Non-trainable params: 0
33
34 None
35 Train on 60000 samples
36 Epoch 1/10 ..... ETA: 6:37 - loss: 2.4625 - accuracy: 0.4938 99%
37 32/60000 .....
38 Epoch 2/10 .....
39 32/60000 ..... ETA: 3s - loss: 0.3463 - accuracy: 0.8750 1054/6
40 .....
41 Epoch 3/10 ..... ETA: 3s - loss: 0.4677 - accuracy: 0.8438 1084/6
42 .....
43 Epoch 4/10 ..... ETA: 3s - loss: 0.3979 - accuracy: 0.8125 1085/6
44 .....
45 Epoch 5/10 ..... ETA: 3s - loss: 0.1486 - accuracy: 0.9688 1084/6
46 .....
47 Epoch 6/10 ..... ETA: 3s - loss: 0.3367 - accuracy: 0.8438 1084/6
48 .....
49 Epoch 7/10 ..... ETA: 3s - loss: 0.1289 - accuracy: 1.0000 1084/6
50 .....
51 Epoch 8/10 ..... ETA: 3s - loss: 0.4910 - accuracy: 0.8125 1120/6
52 .....
53 Epoch 9/10 ..... ETA: 3s - loss: 0.4991 - accuracy: 0.8438 1084/6
54 .....
55 Epoch 10/10 ..... ETA: 3s - loss: 0.4288 - accuracy: 0.8125 1084/6
56 .....
57 10000/10000 - 0s - loss: 0.3245 - accuracy: 0.8864
58 Test accuracy: 0.8864
59
60
```

Elyra Overview

- Notebook Pipelines visual editor
- Ability to run notebooks as batch jobs
- Hybrid runtime support (based on Jupyter Enterprise Gateway)
- Python script execution capabilities within the editor
- Notebook versioning based on Git integration
- Reusable configuration for runtimes

The image displays two screenshots of the Elyra interface. The top screenshot shows the 'Launcher' window with a pipeline editor. The pipeline consists of two nodes: 'generate-contrib...' and 'overview', connected by a flow arrow. The bottom screenshot shows the 'generate-stats.ipynb' notebook editor. The code in the notebook includes:

```
[0]: !pip install PyGithub pandas > /dev/null 2>&1

[1]: import os
import datetime
import pandas as pd
from github import Github

[2]: github = Github(os.environ['GITHUB_TOKEN'])

Jupyter Community Stats

[3]: repositories = {}
repositories['jupyter'] = ['notebook', 'jupyter_client', 'nb2kg', 'enterprise_gateway', 'kernel_gateway']
repositories['jupyterlab'] = ['jupyterlab']
repositories['jupyterhub'] = ['jupyterhub']
repositories['ipython'] = ['ipython']
repositories['irkernel'] = ['irkernel']
repositories['apache'] = ['incubator-toree']

community_stats = {}
index = 0
for org, repos in repositories.items():
    github_org = github.get_organization(org)
    for repo in repos:
        github_repo = github_org.get_repo(repo)
        contributors = github_repo.get_contributors().totalCount
        community_stats[index] = ('org':org, 'repo':repo, 'full_name':github_repo.full_name, 'stars':github_repo.stargazers_count, 'index = index')

community_stats_df = pd.DataFrame.from_dict(community_stats, orient='index')

[4]: community_stats_df.to_csv('community_stats.csv', index=False)
```

The bottom right corner of the interface features the Red Hat logo.

OpenDataHub Demo

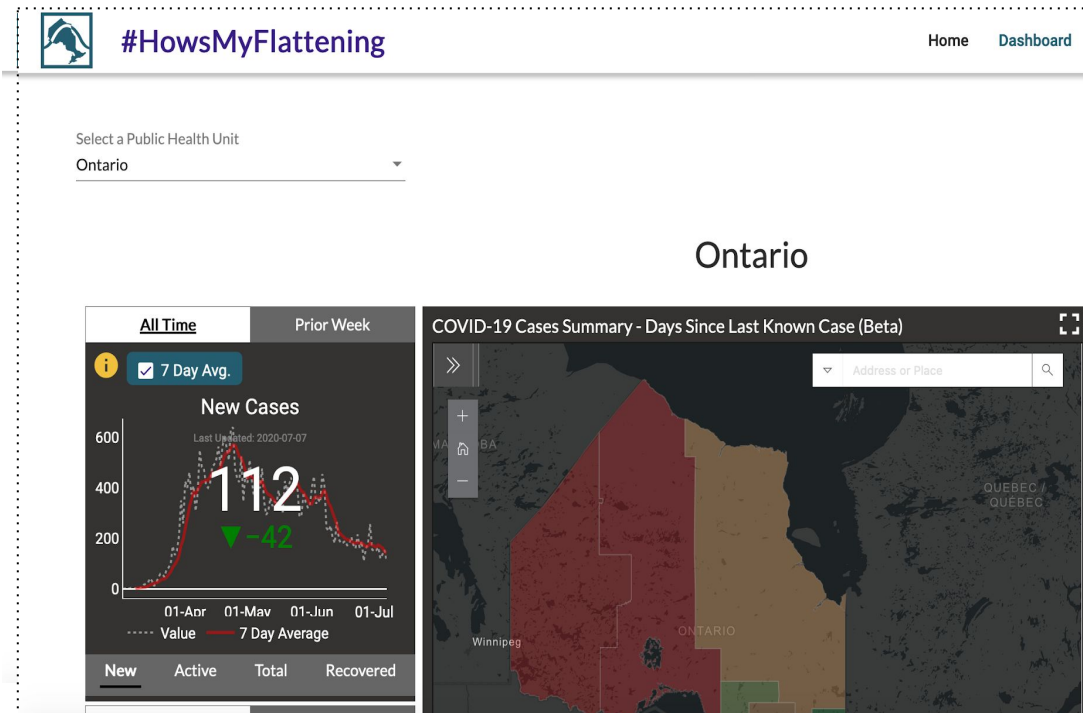
Use Cases

What is #HowsMyFlattening?

howsmyflattening.ca



- A community powered, volunteer group that analyzes, visualizes and shares COVID-19 related data and interventions; (c. March 14/2020)
- More than 275 clinicians, health informaticists, data scientists and developers are working on this project
- Leveraged by citizens, other clinicians, public health officials, politicians, military command
- Now also focused on return to work metrics and economic challenges for the province of Ontario

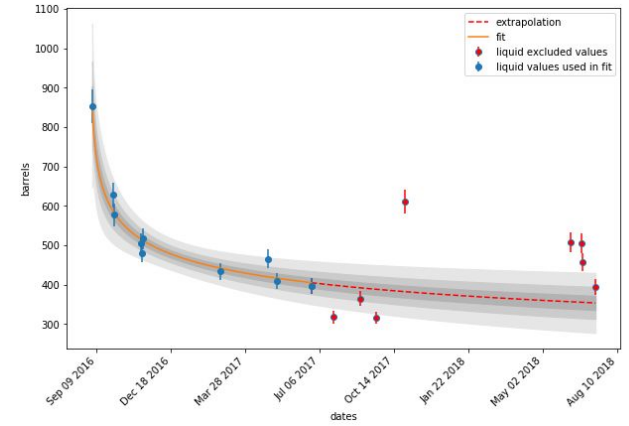
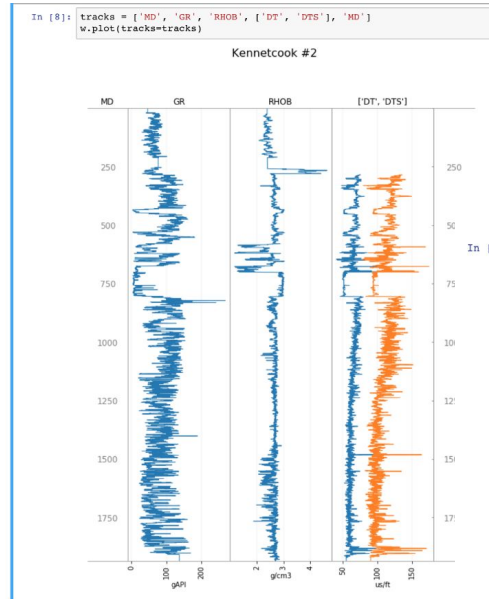


Oil and Gas Data Science

Unique performance computing requirements for Artificial Intelligence, Machine Learning, Neural Networks and GPUs

Multiple Data Science images:

- TensorFlow
- Pyro/PyTorch
- Scikit-learn
- CNN/GANs
- Keras
- Seldon
- RAPIDS.AI
- Apache Arrow
- Welly (LAS)
- Equinor segyio (SEGY)
- OSDU APIs



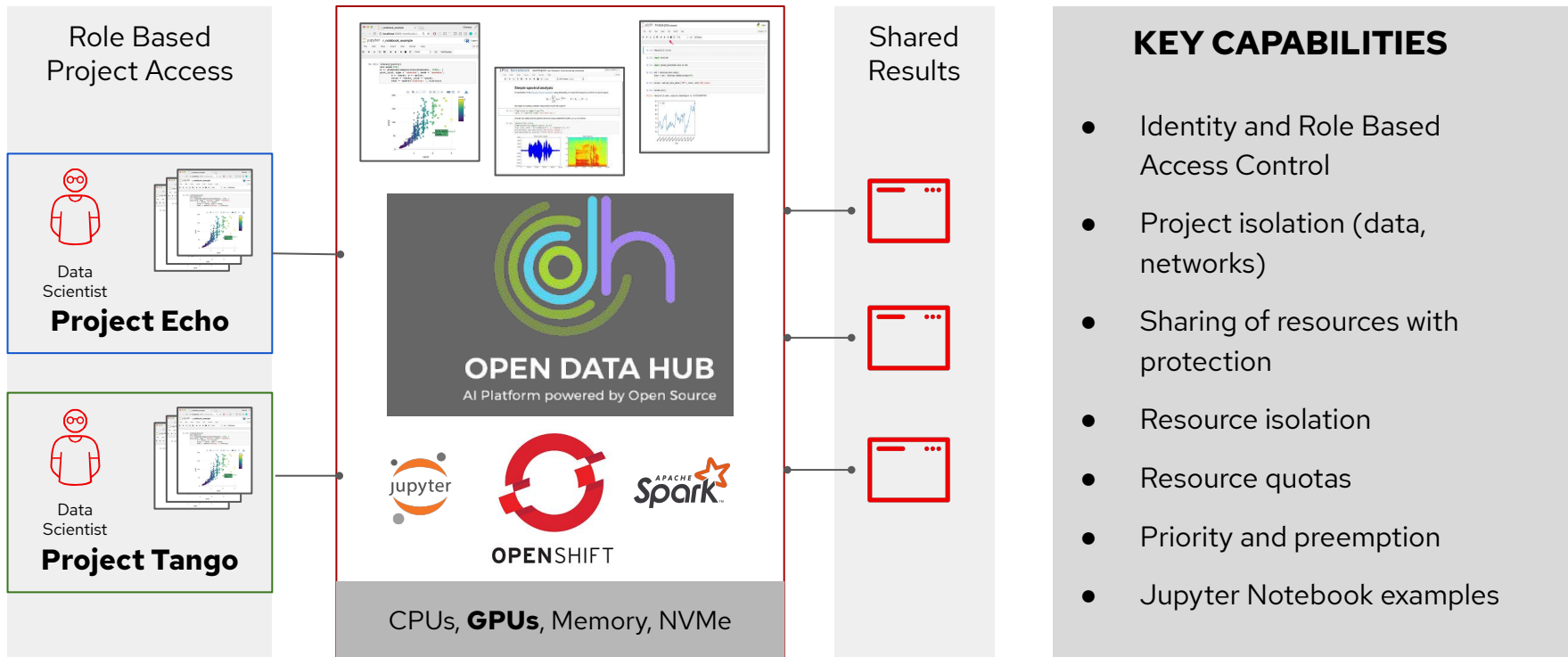
Plot final resampled file

```
In [16]: plot_seggy(destination)
```

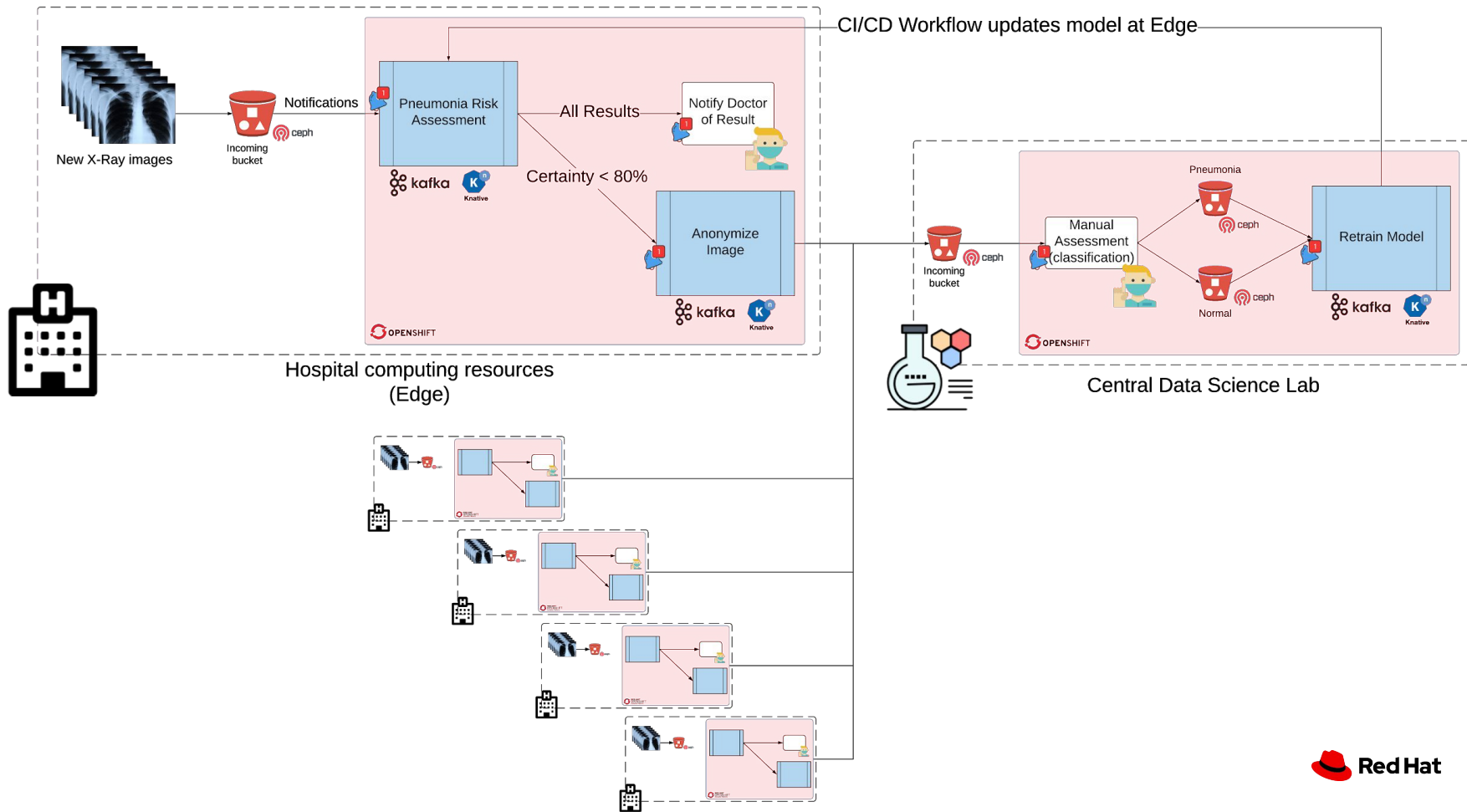
HMF Data Science Solution Pattern

Shared, Secured, Automated Research Environment

howsmyleftening.ca



Medical images analysis has to scale, let's automate it!



Medical images analysis has to scale, let's automate it!

The dashboard displays a pipeline progress overview with the following stages:

- Images uploaded:** 275
- Images processed:** 251
- Images anonymized:** 11

The pipeline flow is: Images uploaded (via S3 and Kafka) → Pneumonia Risk Assessment (via Kafka) → Image Anonymization (via Kafka) → Images anonymized (via S3).

Monitoring and deployment details:

- CPU and RAM Usage:** Shows CPU usage (0.00 to 1.00) and RAM usage (0 GB to 14 GB) over time.
- Risk assessment container:** Shows a gauge for 12 containers.
- Risk distribution:** A chart showing the distribution of risk levels: Normal, Pneumonia, and Unsure.
- Deployments:**

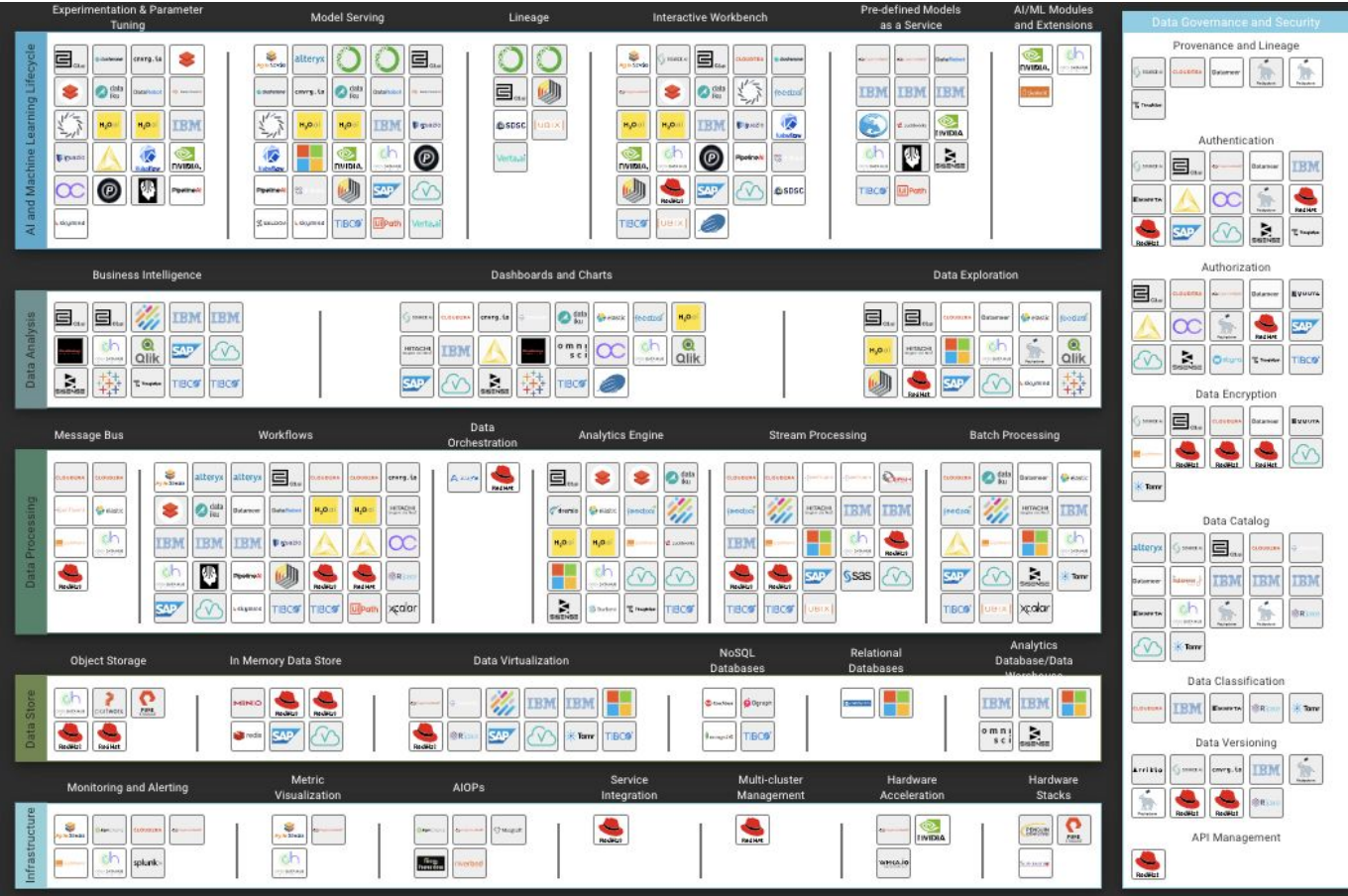
Time	Deployment	Replicas
2020-09-17 15:17:47	risk-assessment-awsb9-deployment	2
2020-09-17 15:17:47	risk-assessment-vq62d-deployment	2
- Images processed by model version:** A chart showing the number of images processed by different model versions (v1, v2) over time.

Image lists and thumbnails:

- Last 10 uploaded images:**

Time	Value
2020-09-17 15:17:47	demo_Jason Callaway_9128_1955-12-26_2018-09-18.jpeg
2020-09-17 15:17:46	demo_Robert Lopez_7898_1960-06-23_2016-03-21.jpeg
2020-09-17 15:17:45	demo_Shonta Turner_1010_1958-11-17_2017-12-08.jpeg
2020-09-17 15:17:44	demo_Armando Campbell_9901_1964-07-28_2018-04-20.jpeg
2020-09-17 15:17:43	demo_Jody Lee_7545_1957-04-15_2016-12-24.jpeg
2020-09-17 15:17:42	demo_Shirley Black_4750_1961-11-16_2015-12-03.jpeg
- Last processed image:** demo_Robert Lopez_7898_1960-06-23_2016-03-21-processed.jpeg
- Last anonymized image:** demo_XXXXXXXXX_b7c5bb0c_XXXXXX_2015-05-13.jpeg

Open Data Hub Landscape - <http://opendatahub.io/landscape>



Strategic partnerships within AI/ML ecosystem

CONFIDENTIAL designator

AI/ML Lifecycle



Data Governance & Security



Data Processing



Data Analytics



Databases



AI Ops



Infrastructure Partners



Hardware Acceleration



Databases and data analytics on Red Hat OpenShift & ecosystem



Ecosystem for databases and data analytics tools



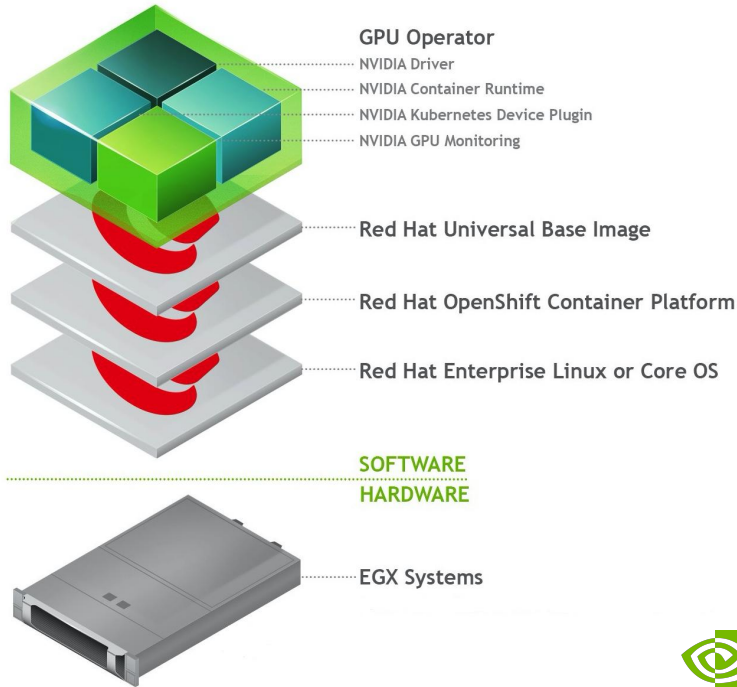
Containers and Kubernetes platform



Infrastructure



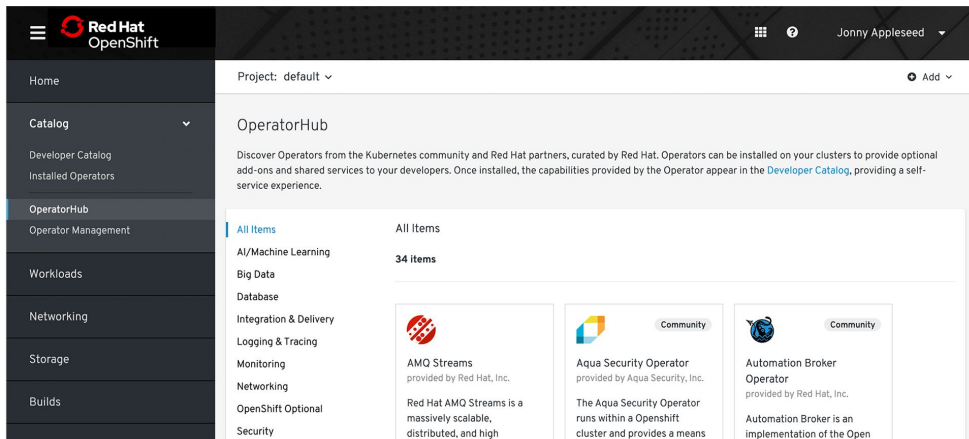
Accelerated AI EAP - Partnership with NVIDIA



Accelerates delivery of AI-powered intelligent applications across the datacenter, edge, and public clouds

- NVIDIA GPUs + NGC Containers
- Red Hat® OpenShift® An industry-leading enterprise container application and Kubernetes platform
- Starter guides for OEMs & Cloud Providers
 - HPE Accelerated AI Reference Architecture





Operatorhub

- Makes an ISV application or platform behave as if it was a cloud service
- Basic install or full auto pilot - the creator (ISV) can chose the levels and functionality
- Standards based - not just for OpenShift - but can take advantage of OpenShift features

Install

OPERATOR VERSION

0.0.1

CAPABILITY LEVEL

- Basic Install
- Seamless Upgrades
- Full Lifecycle
- Deep Insights
- Auto Pilot

PROVIDER

Redis Labs, Inc

LINKS

[Documentation](#)

REPOSITORY

N/A

CONTAINER IMAGE

[redislabs/operator-internal:562_ee119ea](#)

CREATED AT

2019-02-24T11:35:59.000Z

MAINTAINERS

Redis Labs, Inc
support@redislabs.com

CATEGORIES

Database

The Red Hat Marketplace difference



Any cloud. On-prem. Anywhere OpenShift runs.

Build once, deploy to any environment with software that allows for workload portability across clouds.



Certified enterprise software

All software is certified for Red Hat OpenShift and uses Kubernetes operators for built-in management logic.



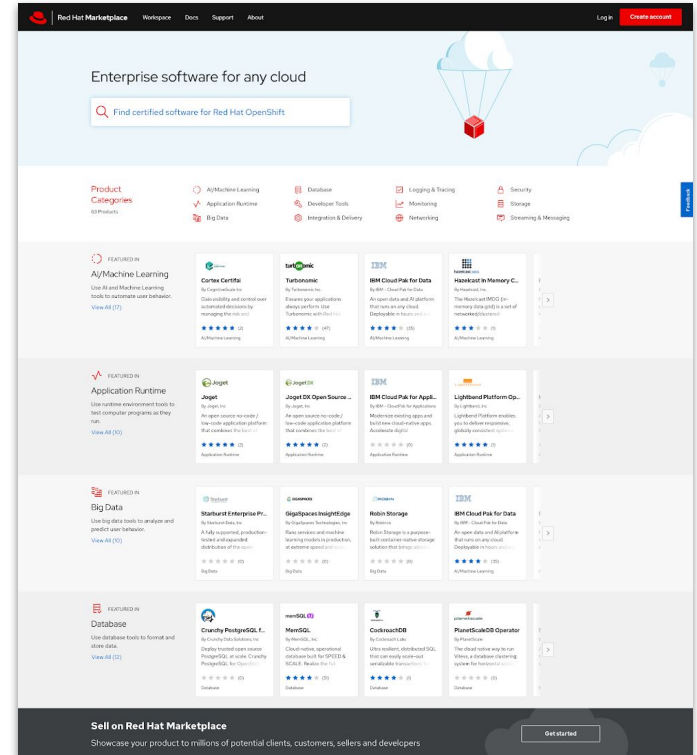
Continuous support

Central support portal offers a single point of entry for any software purchased in the marketplace.



Consolidated usage tracking & spend optimization

Pool spending across clouds. Monitor license usage, expiration, and renewals on a single dashboard.



AI/ML discovery session

A one day, no-cost planning session



GOAL:

To understand the customer's business drivers and technical use cases to propose an Open AI/ML architecture.



DETAIL:

Discussion guided by Red Hat AI/ML Consulting Architect

Attendees from Data Science team, Business decision makers, Engineering, Operations, and Application Development

Deep dive into customer use cases



RED HAT PROVIDES:

Our vision for organization-wide AI/ML adoption

Tailored proposal to solve customer use cases

High-level adoption roadmap

Red Hat AI/ML service offerings

Who attends the AI/ML discovery session?

Customer

- Data scientists
- Data engineers
- IT operations
- Intelligent App developers
- Business sponsors

Red Hat

- AI Center of Excellence (leads)
- Red Hat Services
- Open Innovation Labs
- Solutions Architect (optional)
- Account Team (optional)

AI/ML Architecture Review

Vendor lock-in and high costs of scaling AI/ML applications



Open and flexible architecture for AI/ML applications

Inputs

- Customer AI/ML use cases
- AI/ML platform requirements

Scope

- Red Hat OpenShift
- 1 week

Activities

- Day 1: Customer use cases deep dive
- Day 2: AI/ML architecture review workshop
- Day 5: Live demo in customer lab
- Mentoring and roadmap creation

Key contacts

- Neeraj Kuppam - Services Portfolio Team
- William Benton - AI Center of Excellence

Outcomes

- AI/ML platform strategy
- AI/ML architecture and roadmap
- Live demo in customer lab



Red Hat Services Solution: Open AI/ML Platform

Why

- Accelerate your AI/ML project using the Open Data Hub architecture
- Create a pilot platform for AI/ML use cases
- Bring data scientists, app developers and operations teams together
- Experience DevOps for machine learning best practices
- Build an AI powered application prototype

What

- AI/ML Consulting + Open Innovation Labs Residency
- Infra: Open Data Hub + OpenShift + Open Innovation Labs tools





NEXT STEPS

Contact Us!

archer@redhat.com

Technical overview session

Lets meet with you to discuss how Red Hat is accelerating AI/ML workflows & delivery of intelligent apps.

Visit our websites

Learn more about our AI/ML capabilities, and see success stories from existing customers.

openshift.com/ai-ml, opendatahub.io

Watch our videos

Visit our [YouTube channel](#) to discover a wide range of videos answering all your AI/ML questions.




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)

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

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

 twitter.com/RedHat

