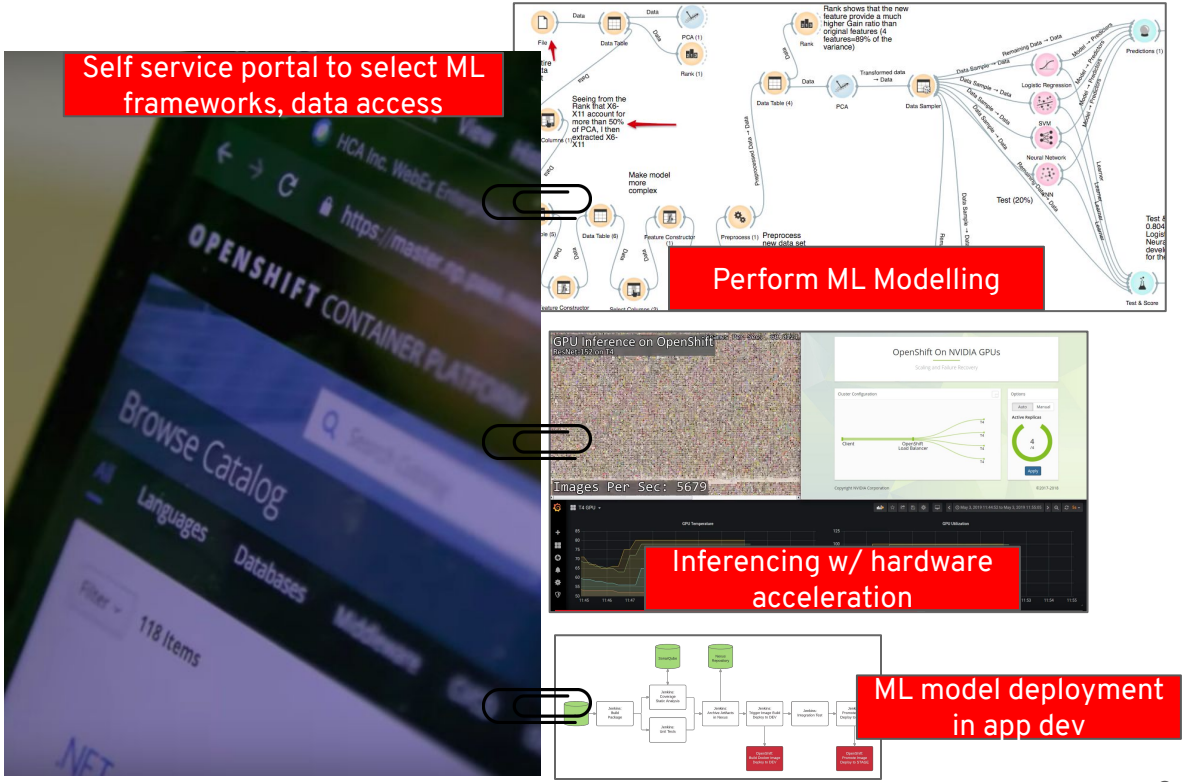
Open Data Hub - A Deep Dive

OpenShift Commons Gathering San Francisco

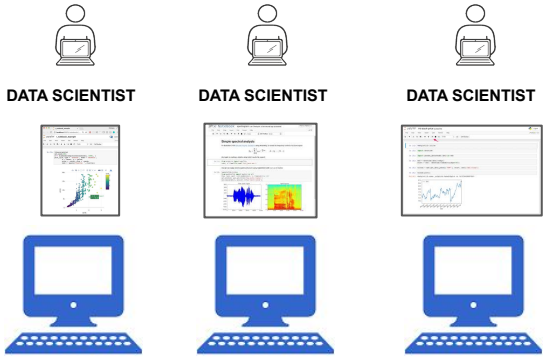
Sherard Griffin
Senior Manager, Red Hat
AI Center of Excellence

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!

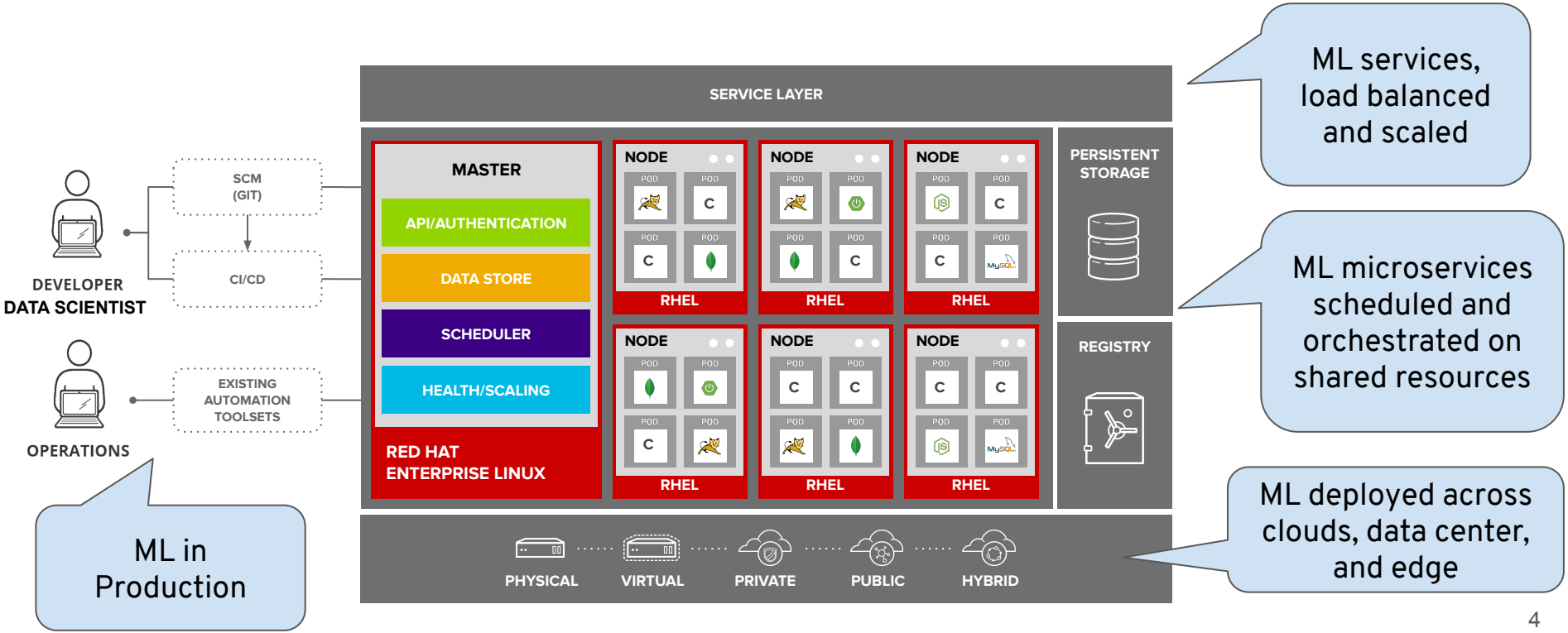


Current Situation and Challenges



- ➔ Team(s) of Data Scientists and Developers
- ➔ Sharing and collaboration, if any, is difficult, manual, error prone and take time
- ➔ Access to limited non-shared resources means modeling takes long times or can't achieve desired accuracy
- ➔ Delivering into models into production is a challenge

Why OpenShift And Cloud Platforms for ML Workloads?



AI/ML on OpenShift Momentum is Strong

Connected Drive &
Autonomous Driving

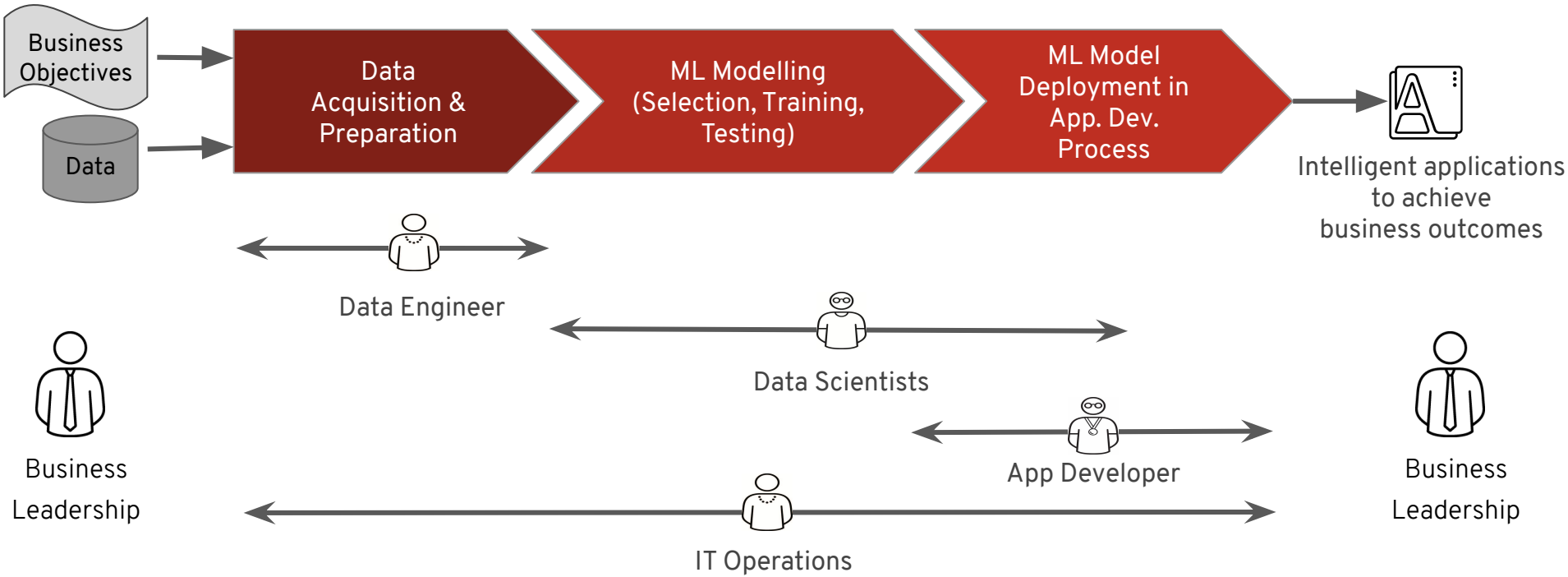
Data driven diagnosis
of “Sepsis”

Digital banking with
personalized services

Autonomous Driving

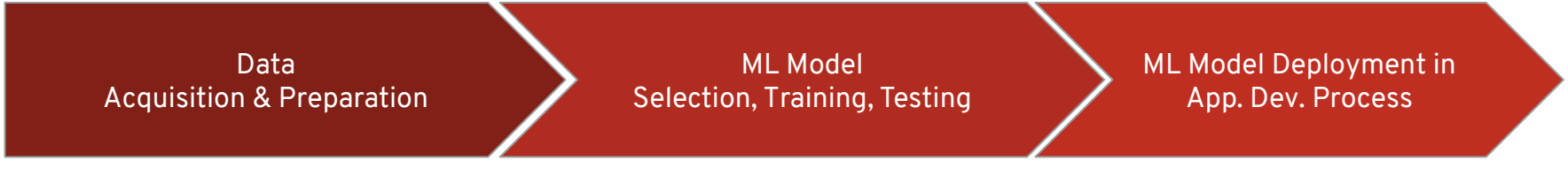
Oil & Gas Exploration

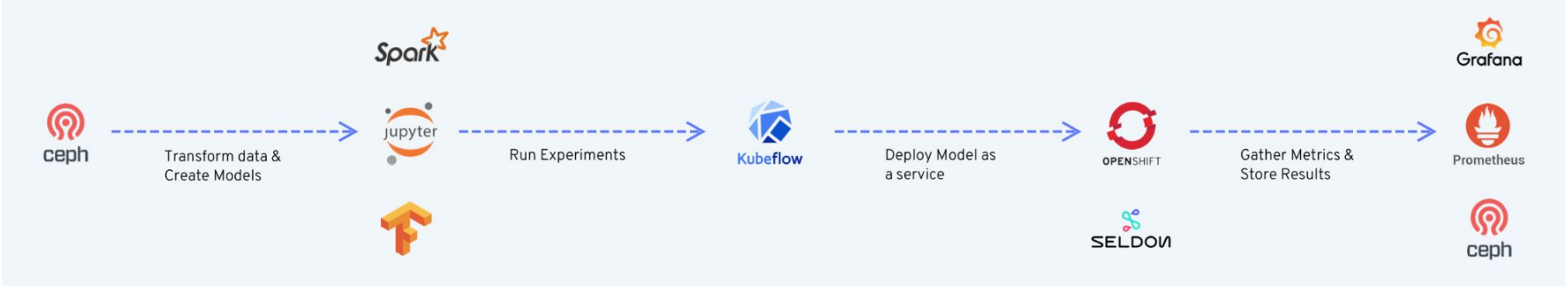
Machine Learning Pipeline



Open Data Hub Project - OpenDataHub.io

- Community meta-operator that integrates best open source AI/ML/Data projects
- Blueprint architecture for end to end AI/ML on OpenShift
- Used as Red Hat's internal data science and AI platform
- Open Data Hub Architecture: https://opendatahub.io/docs/architecture.html





Upstream/Community Projects in the AI/ML Space

NVIDIA NGC
GPU optimized
and curated

Kubeflow

SELDON

ML toolkit and lifecycle
Kube

Tensorflow PYTORCH

Spark jupyter

Others

OPEN DATA HUB

ML-as-a-service platform based on OpenShift, Ceph, Kafka, JupyterHub, Spark,

The registry for
Kubernetes Operators
OperatorHub.io

Home for k8s community to share operators for various apps/tools

Open Data Hub v0.4 Operator

Available Now at OpenDataHub.io

- Monitoring and alerting toolkit
- Used to diagnose problems

- Analytics platform for all metrics
- Query, visualize and alert on metrics

- Deploying machine learning models as micro-services
- Full model lifecycle management

- Unified analytics engine
- Large-scale data access

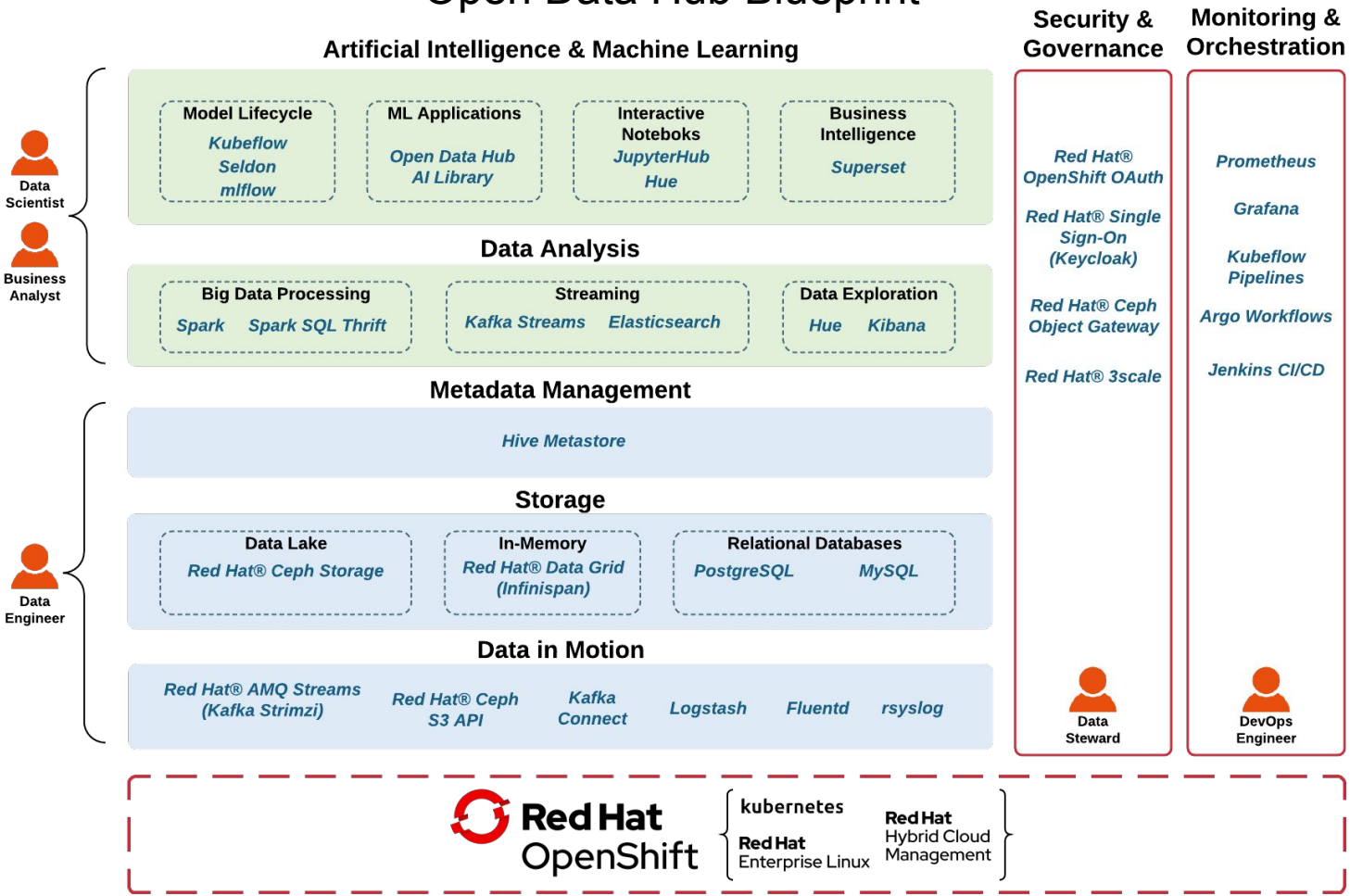
- Multi-user Jupyter
- Used for data science and research

- Distributed Object Store
- S3 Interface

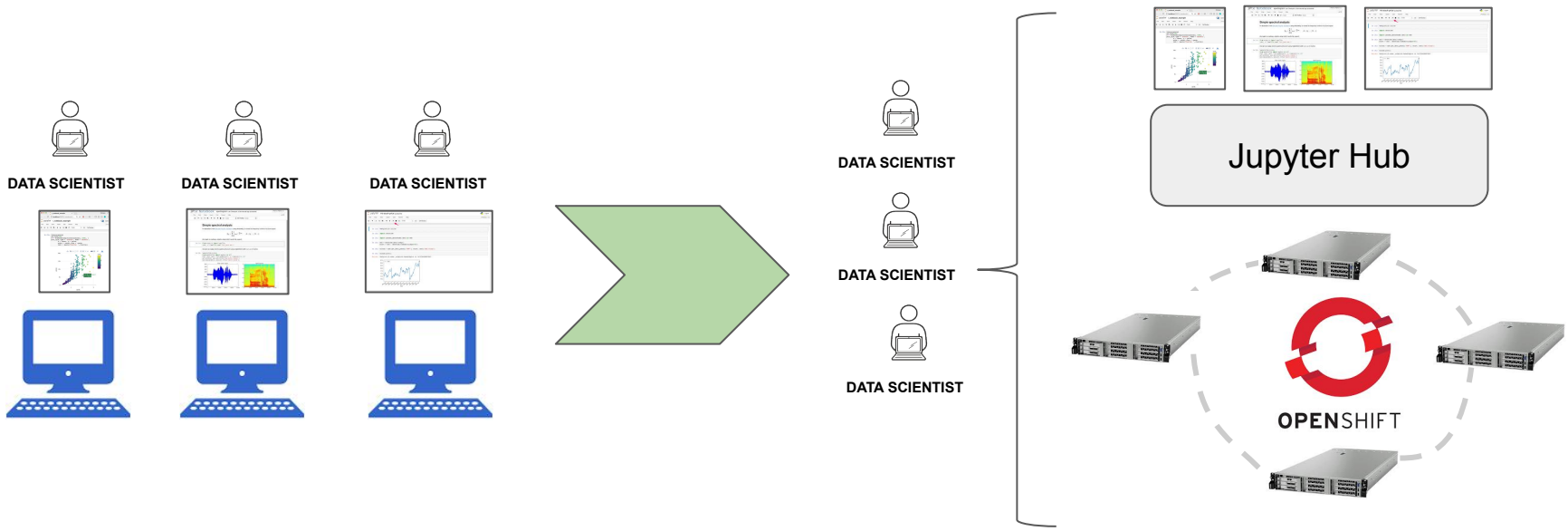
- Distributed event streaming
- Pub/Sub Messaging

- Container-native workflow engine
- Declaratively deploy ML pipelines and models

Open Data Hub Blueprint



Pooling and Sharing Resources with JupyterHub on OpenShift



Demo

https://console-openshift-console.apps.cluster-raleigh-8a2a.raleigh-8a2a.open.redhat.com/operatorhub/ns/opendatahub-user1

Thank you

OpenDataHub.io