OPEN DATA HUB

AI Platform powered by Open Source

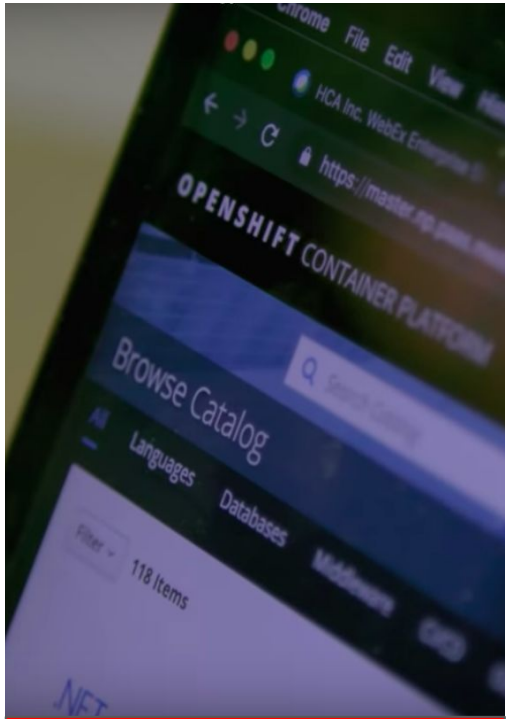
Juana Nakfour
Senior Software Engineer
Open Data Hub

Agenda

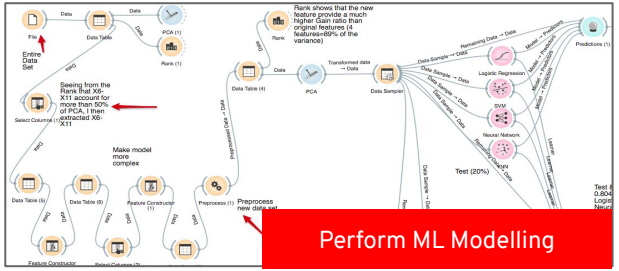
- Open Data Hub (ODH) Introduction
  - ODH **bridging the AI/ML gap** for OpenShift and Red Hat
  - ODH **Architecture**
  - ODH **Components**
  - **Kubeflow**
- Customer Success Stories
  - **Internal** ODH
  - **External** Customers and Mass Open Cloud (**MOC**)
- Open Data Hub Community
  - **Contact** and Engagement

BUILDING A PLATFORM FOR DATA SCIENCE

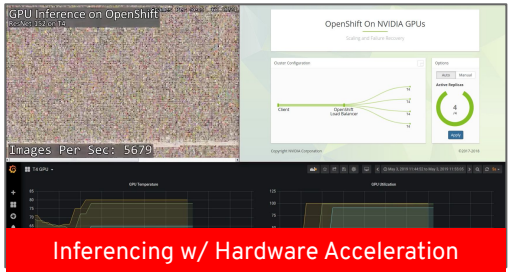
As a Data Scientist, I want a "self-service cloud like" experience for my Machine Learning projects, where I can access a rich set of modelling frameworks, data, and computational resources, share and collaborate with colleagues, and deliver my work into production with speed, agility and repeatability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



Self Service Portal to Select ML Frameworks, and Data Sources



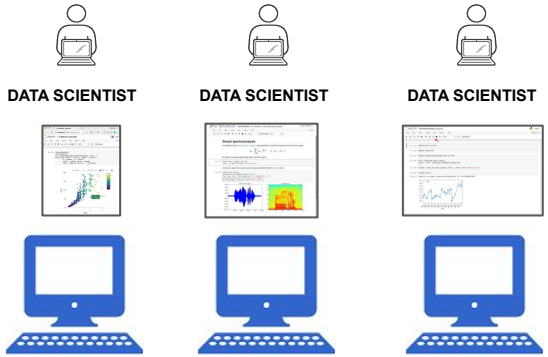
Perform ML Modelling



Inferencing w/ Hardware Acceleration

Model Deployment in App Dev

CHALLENGES WITH STATUS QUO



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

Developing the Customer's AI Platform

***Lock in** and Lose Control* with Amazon, Google or Microsoft Machine Learning

VS

An Open Source AI platform that is built on **Hybrid Cloud**

Comprehensive set of **ISVs** in AI and Data space that are motivated

Integrated and automated that reduces risk

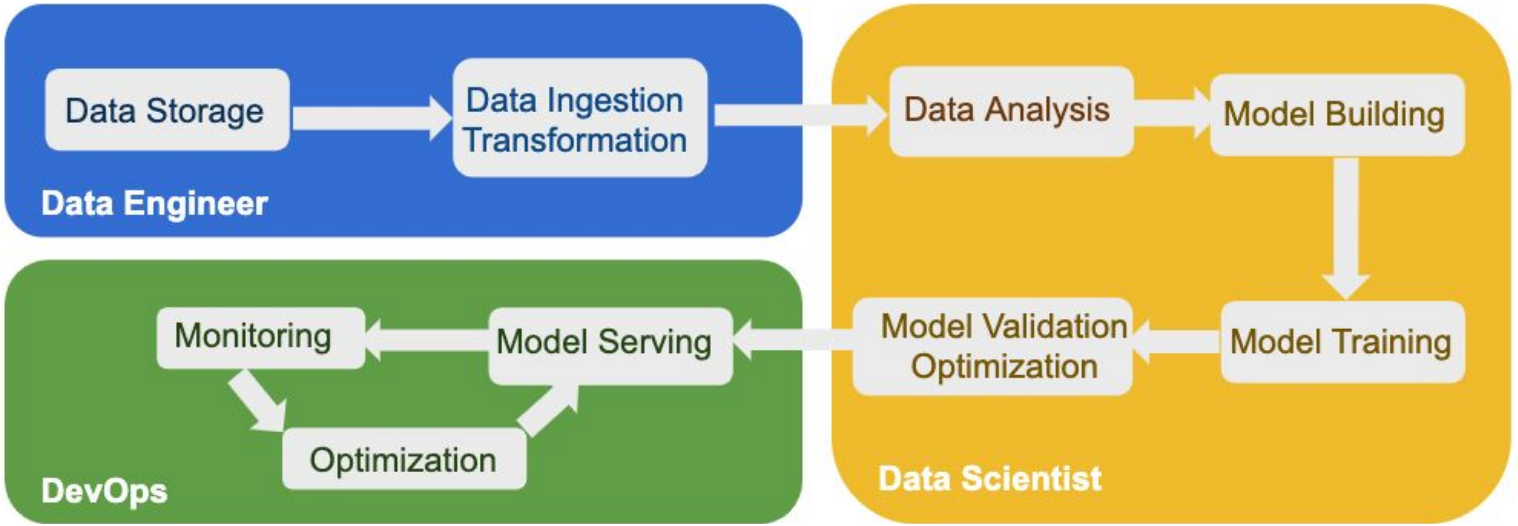
Mitigate vendor, technology and ecosystem **lock in** and control

Foundation for OPEN DATA HUB

The Open Data Hub Project

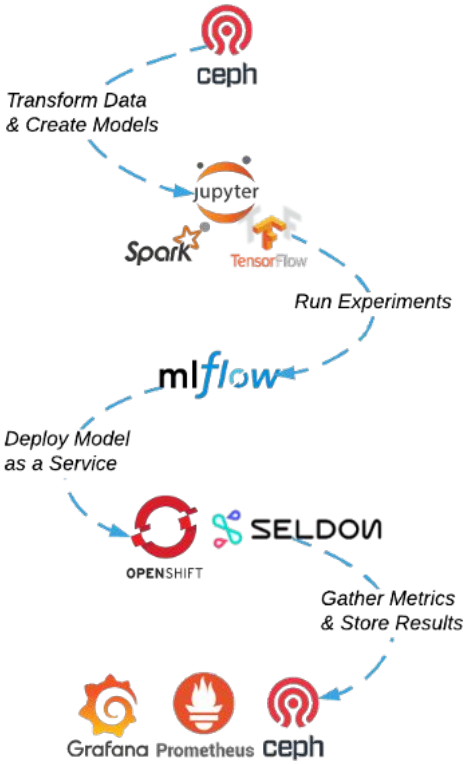
Collaborate on a Data & AI platform for the Hybrid Cloud - https://opendatahub.io/

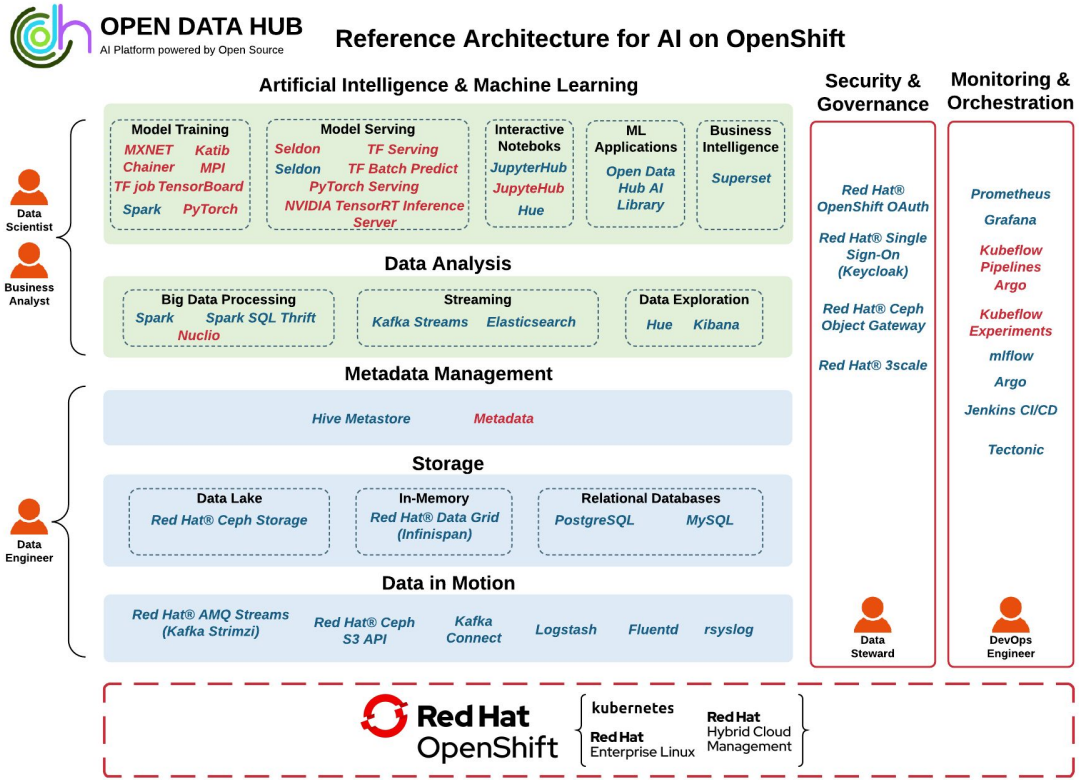
- **Meta-Project** to integrate Open Source projects into a practical service oriented solution.
- Red Hat's **internal** Data Science and AI platform.
- ODH **Documentation:** https://opendatahub.io/docs.html
- AI/ML playlist on Openshift Commons **youtube** channel: https://www.youtube.com/playlist?list=PLaR6Rq6Z4IqcG2znnCIv-xbj93Q_wcY8L
- Bi-weekly open **community meetings:** https://gitlab.com/opendatahub/opendatahub-community



OPEN DATA HUB - FUTURE

Fraud Detection:
https://youtu.be/662FccIWeOE





Open Data Hub 0.5

Available Now in Openshift 4.x Catalog

- Monitoring and alerting toolkit
- Records numeric time series data
- Used to diagnose problems

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

- Deploying machine learning models on Kubernetes
- Expose models via REST and gRPC
- Full model lifecycle management

- Unified analytics engine
- Large-scale data
- Runs on Kubernetes

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

- Distributed Object Store
- S3 Interface

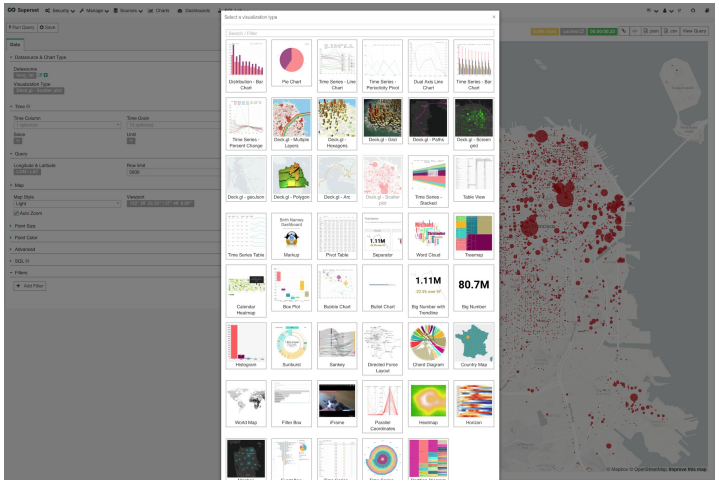
- Distributed event streaming
- Pub/Sub Messaging

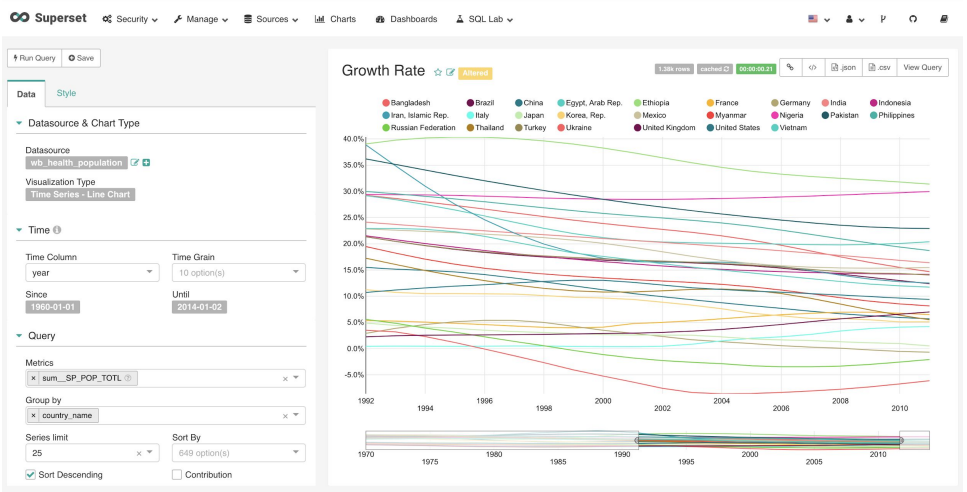
- Container Native Pipelines
- DAG workflows

Open Data Hub 0.5.0

Apache Superset

- A web portal tool for **data exploration** and **visualization**
- Connect to **SQL** Database, support for most SQL-speaking databases
- Provides many beautiful **visualizations** to showcase data
- SQL tool for data queries





Open Data Hub 0.5.0

Data Catalog

- **Data exploration** toolset based on Spark SQL Thrift Server, Hive Metastore and Cloudera Hue
- Manages Hive tables and **query data** directly from Data Lake
- Catalyst (part of Spark SQL) as the **SQL engine**
- **Business Intelligence tools** can connect to Data Catalog through an ODBC/JDBC connector

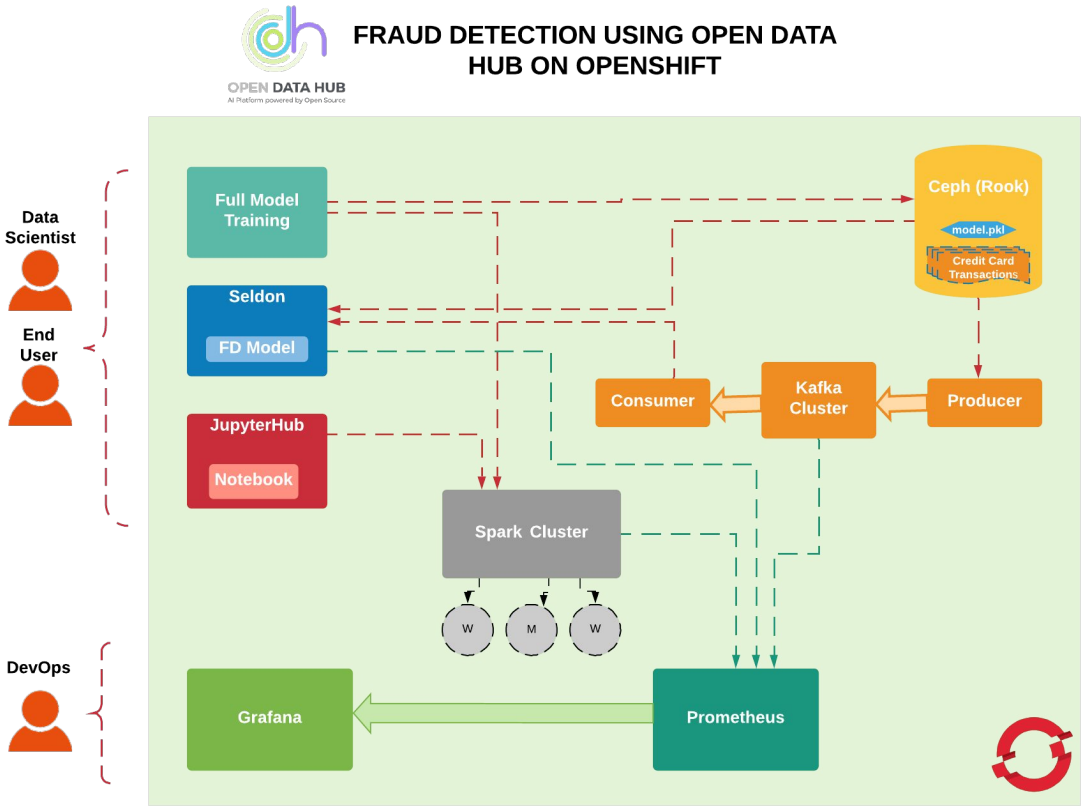
Open Data Hub Kubeflow on Openshift

- Dedicated to making deployments of machine learning (ML) workflows on Kubernetes **simple, portable and scalable -** https://www.kubeflow.org
- Core **components**: Notebook ,TensorFlow training/Serving, Pytorch Training Serving, Seldon, Katib, Pipelines(Argo)
- ODH **github** fork**:** https://github.com/opendatahub-io
- ODH **Goals:**
  - Kubeflow 0.7 o**n OCP 4.x**
  - Contribute **upstream** to Kubeflow
  - **Integrate** with ODH

Kubeflow

Fraud Detection

- **Detect fraud** credit card transactions
- Dataset used from **Kaggle**
- Provide **end-to-end AI/ML** platform
- Create an AI/ML **model** that can predict fraud transactions
- **Serve** model and collect model **metrics**
- Provide **monitoring** tools for model and services used by DevOps
- Provide **development** tools for Data Scientists
- Provide ETL tools used by **Data Engineers**
- Demo: https://youtu.be/IcQ2bhsw_kQ



Supporting the Open Data Hub

- Open Data Hub is a reference architecture and community operator, not a product
- No official full-stack support planned
- Support can be extended with professional service agreements

What is supported by Red Hat?

RED HAT AMQ

Red Hat Data Grid

Red Hat Process Automation Manager

Red Hat Single Sign-On

Red Hat Enterprise Linux

AI/ML Partner Enablement

Mission: *With partners, make best-in-class AI/ML hybrid/multi-cloud ready for mainstream and enterprise markets*

- **Primary Focus on OpenShift** - kubernetes-powered multi/hybrid-cloud AI/ML platform
- **Hardware Ecosystem for Performance** - GPUs
- **SW Ecosystem Enablement**
  - Projects w/ Vendor Support - Certified vendors w/ commercial support
  - Balance Platforms & Components - Provide choice
- **Cloud-Like Experience -** Operators, operatorhub.io, embedded operatorhub

Via:

- **AI Reference Architecture** - Open Data Hub - RH Ref Arch for AI for upstream/products/partners
- **AI Use Case Definitions** - define core use cases - RH + Partners Enable these use cases

Customer Success Stories

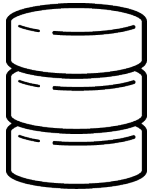
ODH Red Hat Internal Clusters

Stage

Shared OpenShift cluster managed by PSI team.

Production

Dedicated OpenShift cluster managed by PSI team.



Ceph S3

Shared Ceph cluster managed by PSI team, migrating to a dedicated cluster.

ODH Red Hat Internal Customers

These are some significant users of the Internal Data Hub:

PnT DevOps

Applications in the product release pipeline store their runtime logs in our system. These groups are also engaged for anomaly detection (Factory 2.0)

Telemeter

Operational metrics from OpenShift clusters. AIOps is engaged here.

Grokket

Automated data science on insights data

Customer Insights

Storage of customer data like SOSReports, customer feedback, etc.

Customers

External Customers in the Gas and Oil (Exxon Mobil) Industry have successfully deployed Open Data Hub supporting multiple jupyterhub notebook kernels and GPU since March 2019.

Currently running proof of concepts with Financial Industry clients.

Mass Open Cloud (MOC)

Led by Boston University, the MOC is a collaborative effort among BU, Harvard, UMass Amherst, MIT, and Northeastern University, as well as the Massachusetts Green High-Performance Computing Center (MGHPCC) and Oak Ridge National Laboratory (ORNL).

It is supported by a broad alliance of industry partners, including Red Hat.

Open Data Hub Community

Contact and Engagement

- Open Data Hub **site**: opendatahub.io
- ODH **Gitlab**: https://gitlab.com/opendatahub
- ODH-Kubeflow **Github**: https://github.com/opendatahub-io
- Community **Mailing lists**: announcements@lists.opendatahub.io, contributors@lists.opendatahub.io
- Join our open **community meetings**: https://gitlab.com/opendatahub/opendatahub-community
- AI/ML playlist on Openshift Commons **youtube** channel: https://www.youtube.com/playlist?list=PLaR6Rq6Z4IqcG2znnCIv-xbj93Q_wcY8L