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?

ML services, load balanced and scaled
ML microservices scheduled and orchestrated on shared resources
ML deployed across clouds, data center, and edge
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

- Business Objectives
- Data
- Data Acquisition & Preparation
- ML Modelling (Selection, Training, Testing)
- ML Model Deployment in App. Dev. Process
- Intelligent applications to achieve business outcomes

Roles:
- Data Engineer
- Data Scientists
- App Developer
- IT Operations
- Business Leadership
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](https://opendatahub.io/docs/architecture.html)
Upstream/Community Projects in the AI/ML Space

- **NVIDIA NGC**
  - GPU optimized and curated

- **Kubeflow**
  - ML toolkit and lifecycle

- **Seldon**
  - Kube

- **Others**

- **OPEN DATA HUB**
  - ML-as-a-service platform based on
    - OpenShift, Ceph, Kafka, JupyterHub,
    - Spark,

- **OperatorHub.io**
  - Home for k8s community to share operators for various apps/tools
Open Data Hub v0.4 Operator
Available Now at OpenDataHub.io

- Prometheus
  - Monitoring and alerting toolkit
  - Used to diagnose problems

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

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

- Spark
  - Unified analytics engine
  - Large-scale data access

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

- Ceph
  - Distributed Object Store
  - S3 Interface

- Kafka
  - Distributed event streaming
  - Pub/Sub Messaging

- Strimzi
  - Container-native workflow engine
  - Declaratively deploy ML pipelines and models
Open Data Hub Blueprint

Artificial Intelligence & Machine Learning
- Model Lifecycle
  - Kubeflow
  - Seldon
  - mlflow
- ML Applications
  - Open Data Hub
  - AI Library
- Interactive Notebooks
  - JupyterHub
  - Hue
- Business Intelligence
  - Superset

Data Analysis
- Big Data Processing
  - Spark
  - Spark SQL
  - Thrift
- Streaming
  - Kafka Streams
  - Elasticsearch
- Data Exploration
  - Hue
  - Kibana

Metadata Management
- Hive Metastore

Storage
- Data Lake
  - Red Hat® Ceph Storage
- In-Memory
  - Red Hat® Data Grid
  - (Infinispan)
- Relational Databases
  - PostgreSQL
  - MySQL

Data in Motion
- Red Hat® AMQ Streams
  - (Kafka Strimzi)
- Red Hat® Ceph
  - S3 API
- Kafka Connect
- Logstash
- Fluentd
- rsyslog

Security & Governance
- Red Hat® OpenShift OAuth
- Red Hat® Single Sign-On
  - (Keycloak)
- Red Hat® Ceph Object Gateway
- Red Hat® 3scale

Monitoring & Orchestration
- Prometheus
- Grafana
- Kubeflow Pipelines
- Argo Workflows
- Jenkins CI/CD

Data Engineer

Business Analyst

Data Scientist
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