Using AIOps to Speed up CI/CD processes

Aakanksha Duggal
Oindrilla Chatterjee

Aakanksha Duggal
Data Scientist

Oindrilla Chatterjee
Senior Data Scientist

Boston, MA, USA

Emerging Tech,
Office of the CTO

Red Hat

Index

- Open Source Operations Data

  - What is it?

  - Where is it coming from?

  - Why is it important?

- What is AIOps?

- Project AI4CI and its tools

- Demo

- Resources

Open Source operations data - what, where and why?

- Data originating from **real world production systems.**

- Data made only available by **operating softwares and apps** in production.

- Eg: CI/CD data, telemetry, logs, operational dashboards.
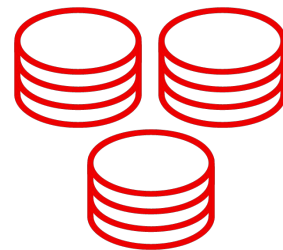
Open Source operations data - what, where and why?

Operate First

- Operate First makes **operations open source**.

- An initiative centered around learning and developing code and practices in an open **production community cloud**.

- Deploy and maintain apps in an open environment leading to **open operations data** which include logs, issues, metrics.

Kubernetes

- Kubernetes testing infrastructure open sources CI/CD data originating from components such as Prow, Testgrid, Github

www.operate-first.cloud/

Open Source operations data - what, where and why?

- Open operations mean open sourcing **SRE best practices** leading to **better collaboration** between developers, operators, data scientists and better software.

- Open production **operations datasets are rare** and very valuable for AI communities.

- Open operations datasets can help **enable AI tools** to assist with cloud operations.

Red Hat

AI + Ops

Using AI tools to support Operations

AIOps + CI

Supporting CI/CD by using AI capabilities

AI4CI

Artificial Intelligence for Continuous Integration

An collection of open source data science tools to collect and analyze CI data built using open operations data.

Red Hat

AI4CI: Open Source AIOps toolkit

Problem

- Need for **AIOps** - Automated monitoring, analysis, alerting with Ops (CI/CD, development processes)

- **Open Source data** originating from real world production systems is a rarity for public datasets.

- Lack of AI driven metrics for open source community health.

Opportunity

- **Open operations data** made available by running open source software and applications in production.

- Data includes CI/CD data, code, telemetry, logs, operational dashboards.

- Eg: Kubernetes testing infrastructure, Fedora make their testing data available open source.

Solution

- Collection of intelligent and open source **data science tools** to collect and analyze the CI/CD data.

- **AIOps models** like Github time-to-merge service, optimal stopping time prediction, build log classifier

- KPI and Metric dashboards

- Goal is to foster an open source AIOps community with open ops data, AI tools and services.
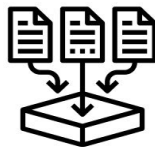
Red Hat

AI4CI supports CI/CD and software dev processes

What is AI4CI?

Collection of **Open Source AIOps tools** including scripts, notebooks, pipelines, dashboards and data sources.

Data collection
Collection of open operations data from Kubernetes testing platforms eg: Testgrid, Github, and Prow.

Metrics
Collects metrics and **KPIs** and visualization dashboards.

ML Services
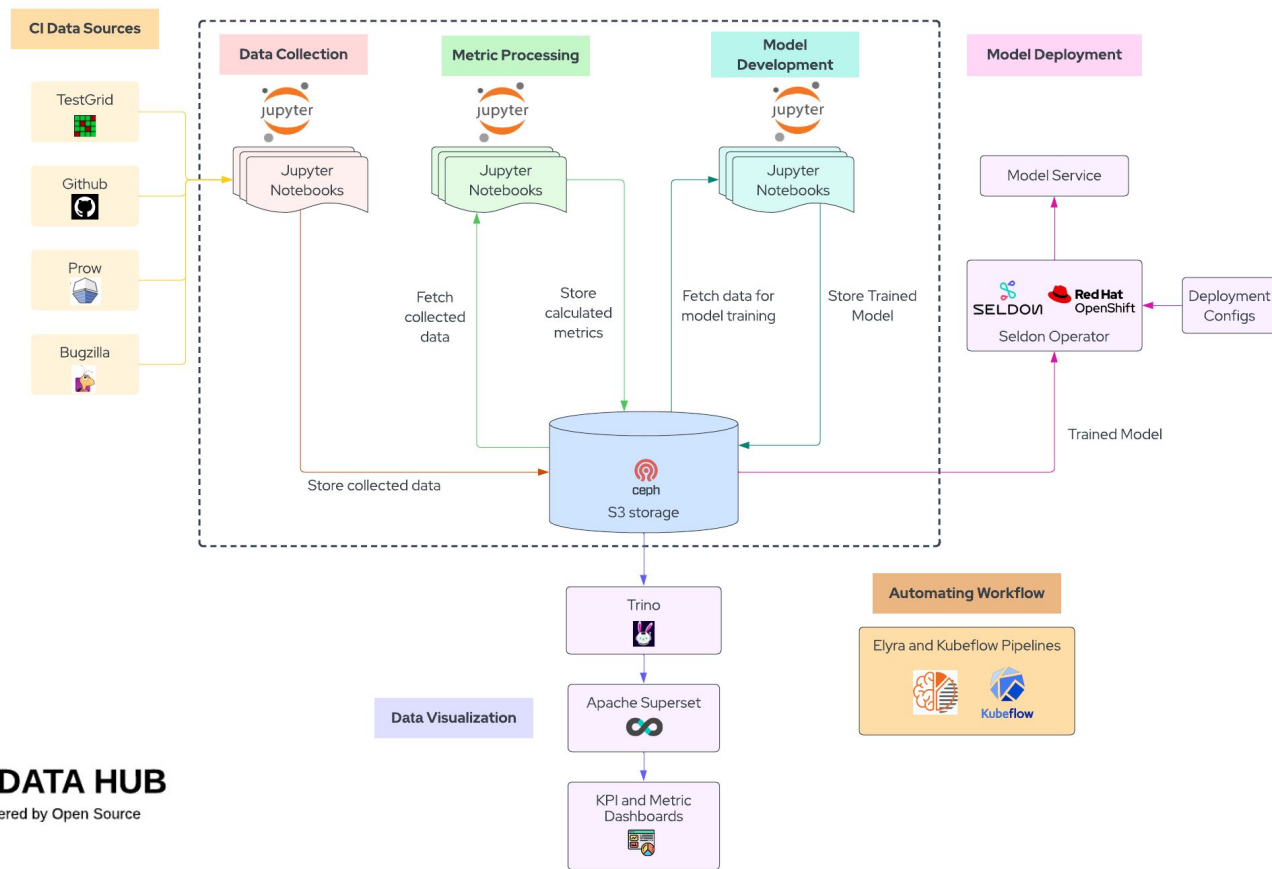**ML services** which can support CI/CD processes.

Open source AIOps template
Resource for open source AIOps communities (notebooks, scripts, automated ML pipelines, dashboards, services tools)
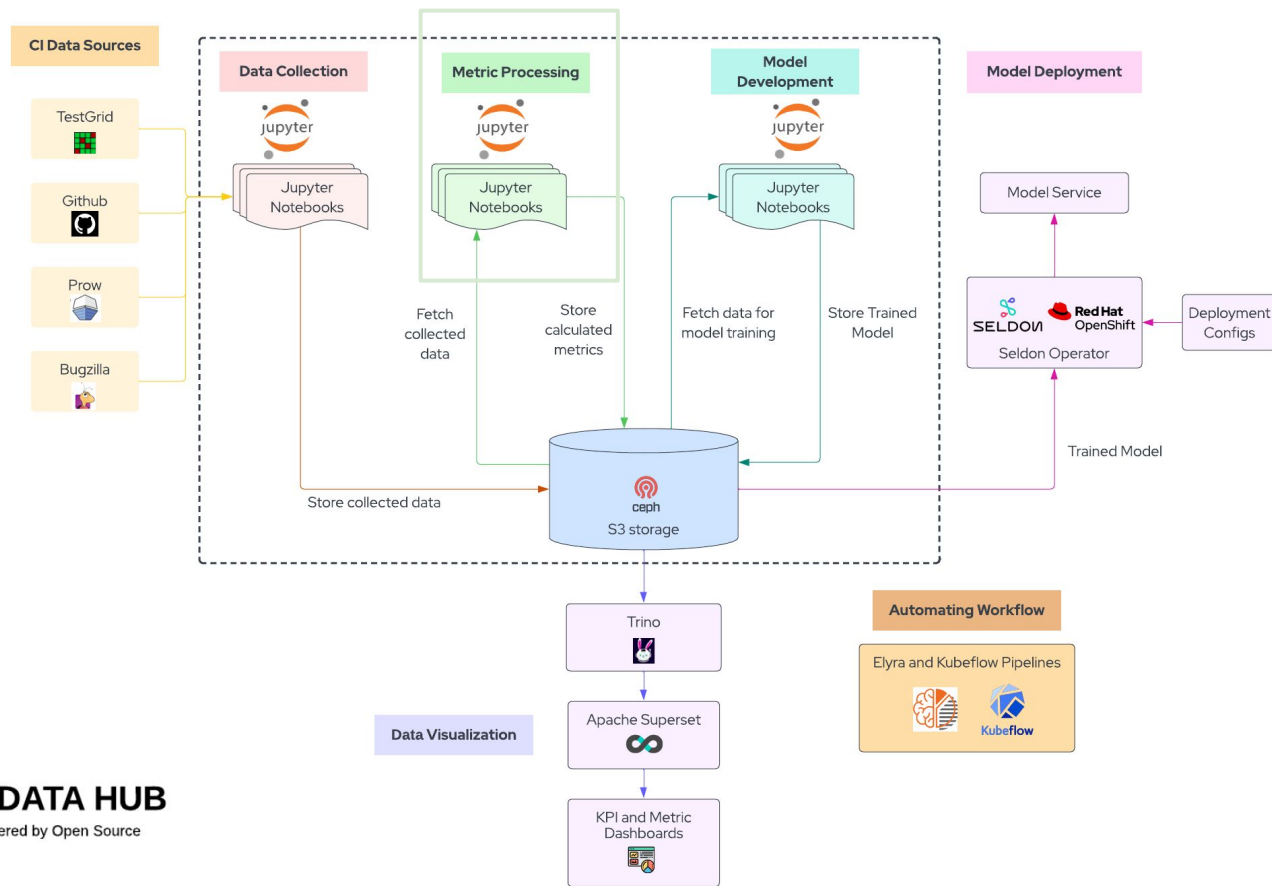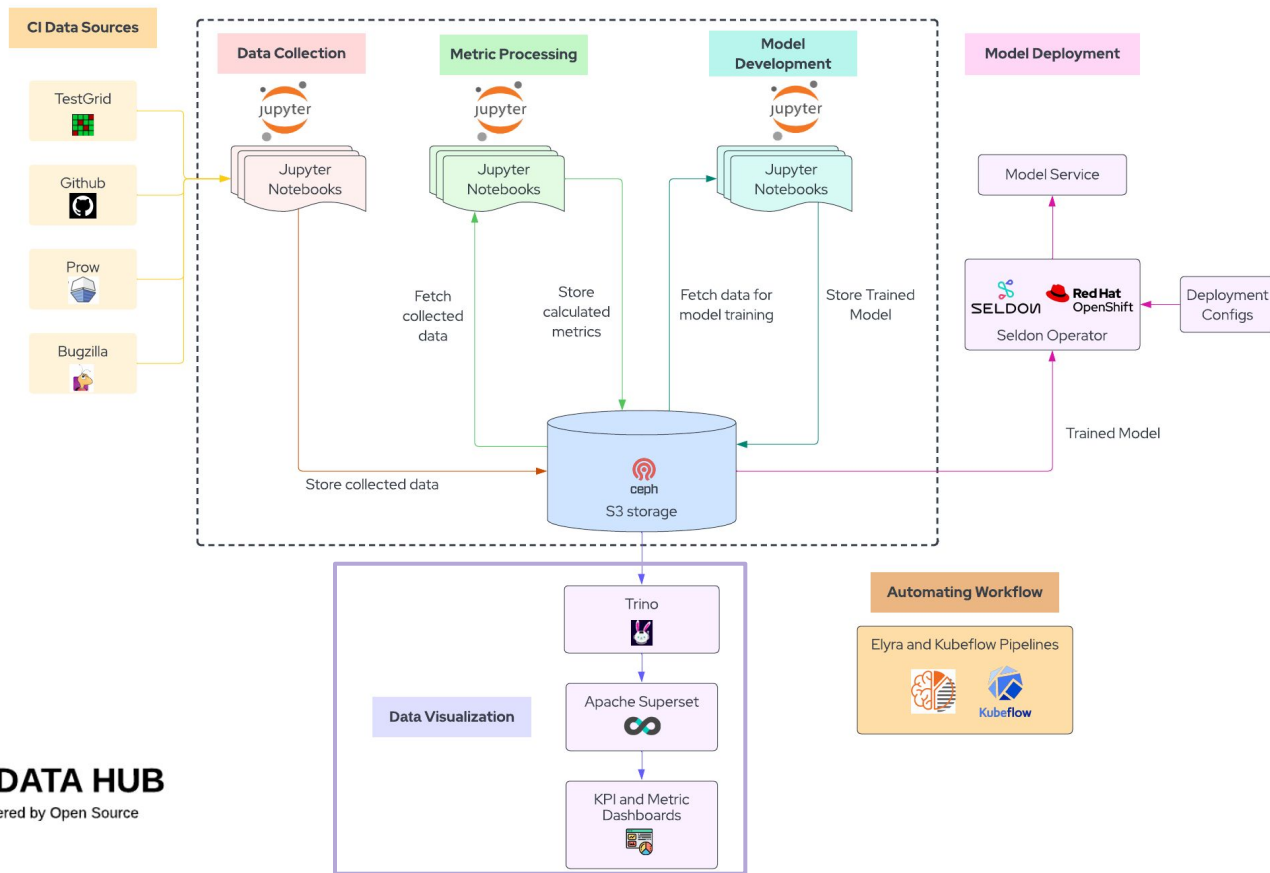
Red Hat

Architecture

Architecture

Architecture

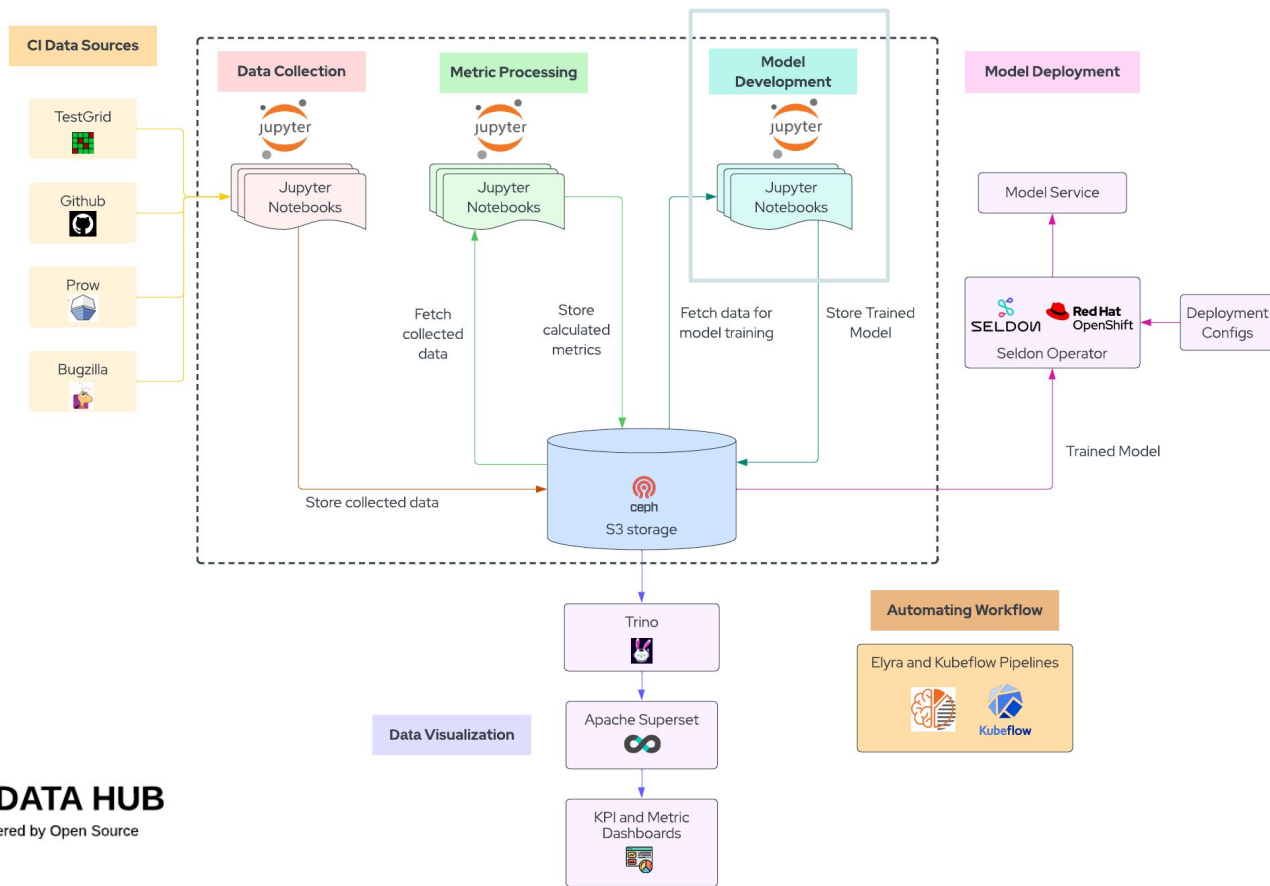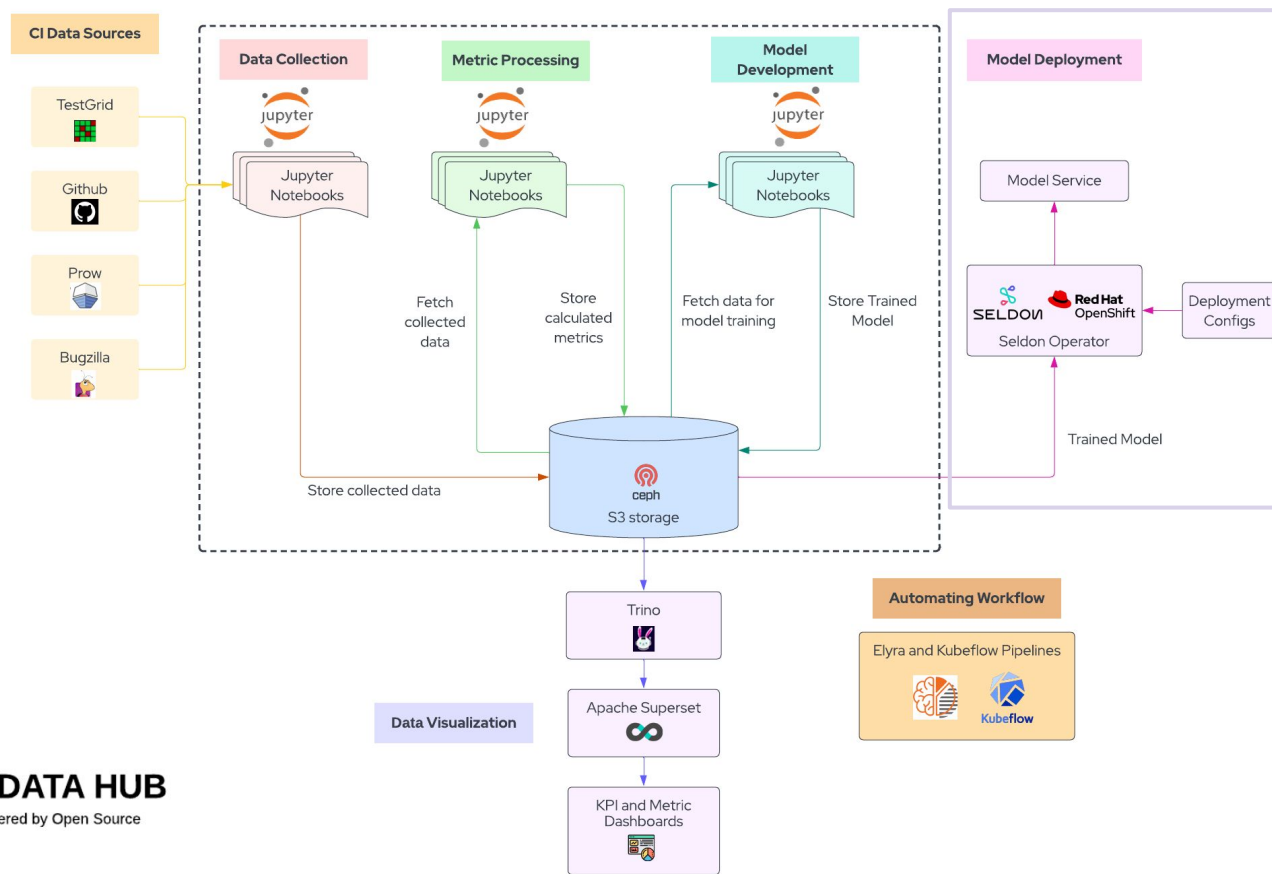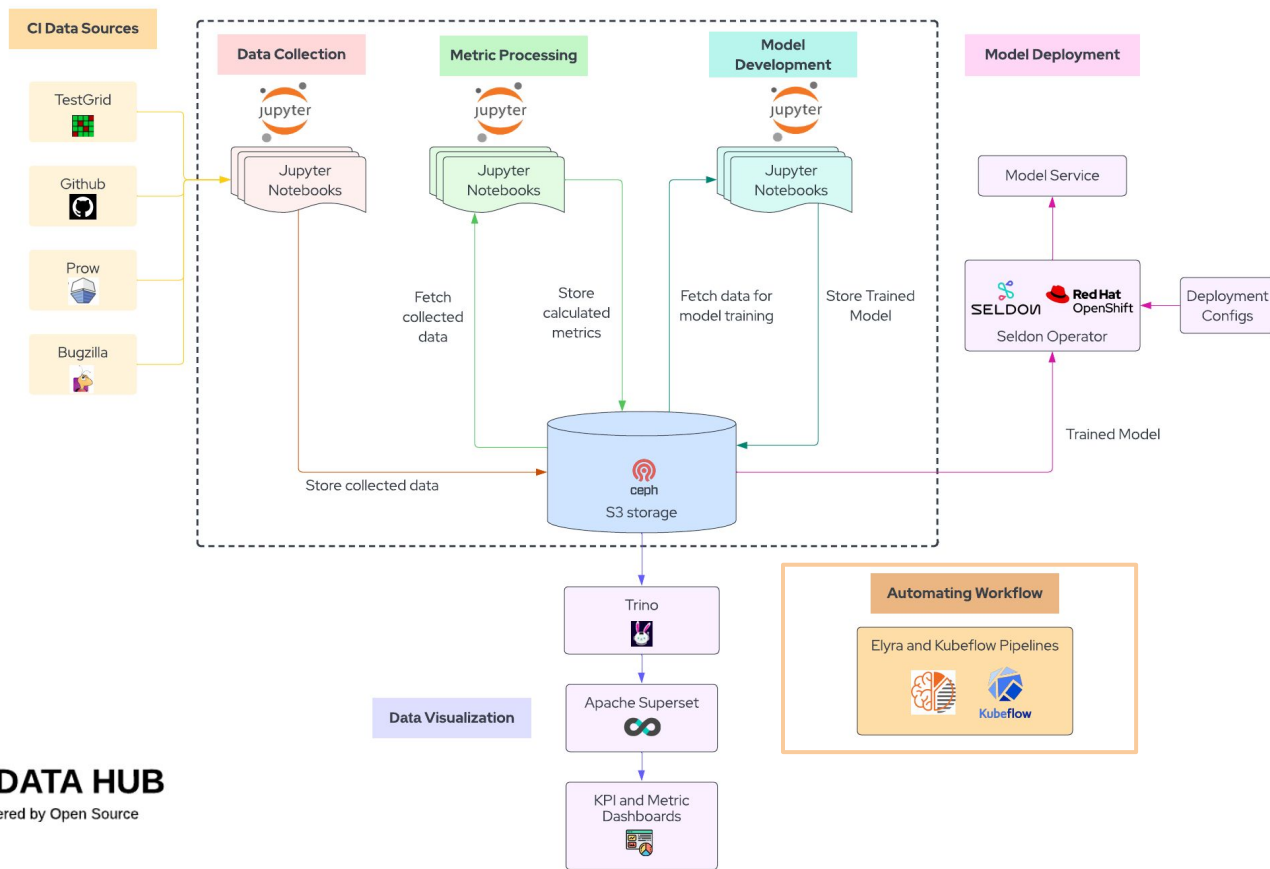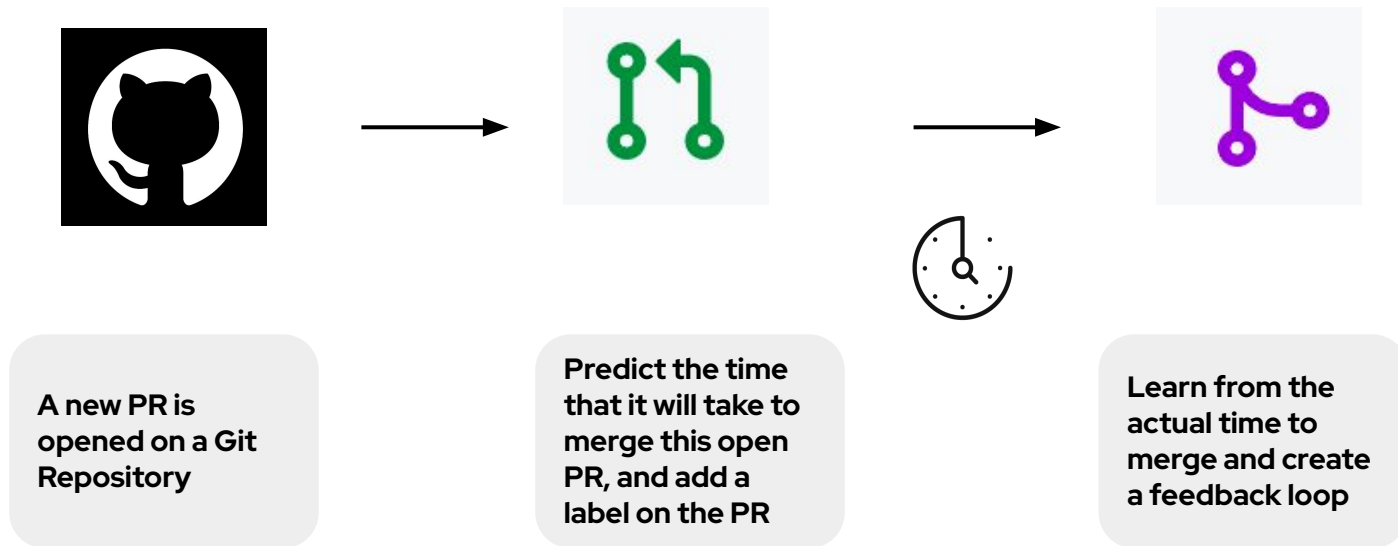# Architecture

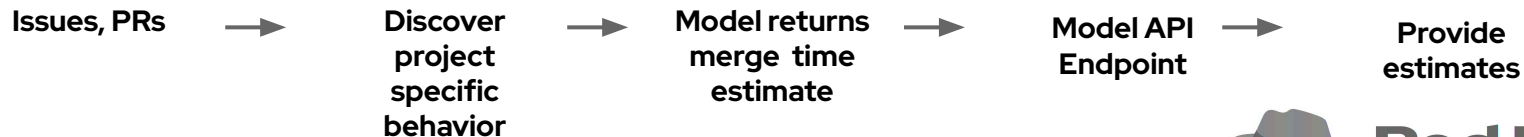# Architecture
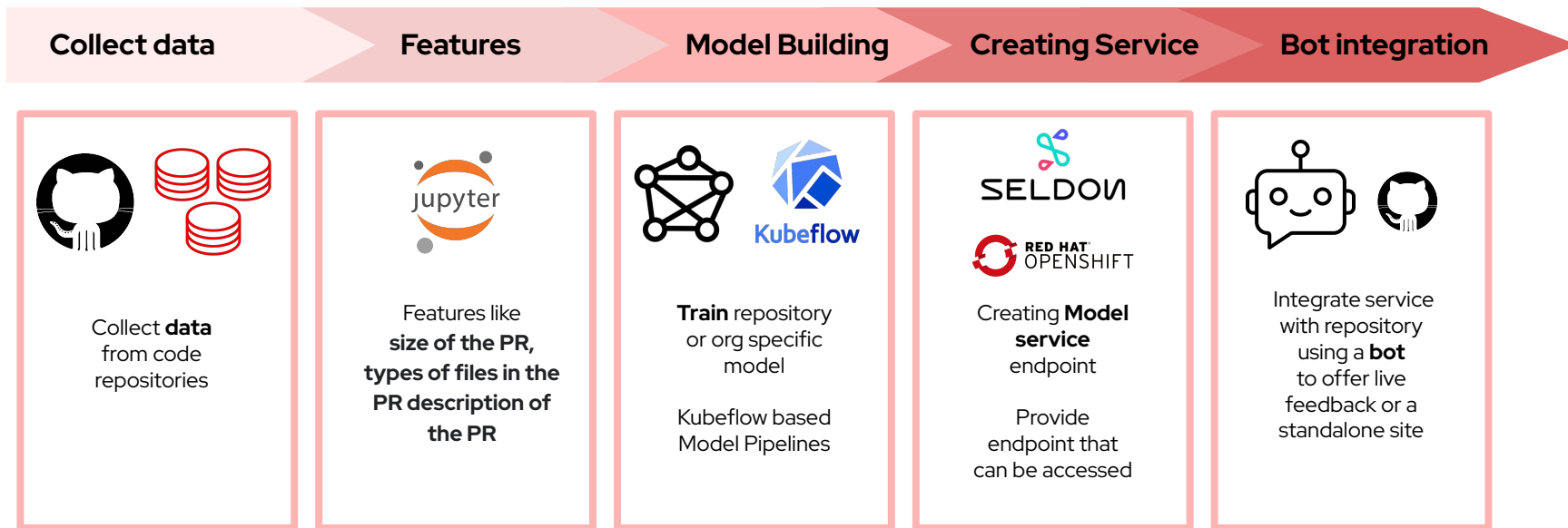

CI Data Sources

TestGrid

Github

Prow

Bugzilla

Data Collection

Jupyter

Jupyter Notebooks

Metric Processing

Jupyter

Jupyter Notebooks

Model Development

Jupyter

Jupyter Notebooks

Model Deployment

Fetch collected data

Store calculated metrics

Fetch data for model training

Store Trained Model

Store collected data

ceph
S3 storage

Model Service

SELDON   Red Hat OpenShift

Seldon Operator

Deployment Configs

Trained Model

Trino

Data Visualization

Apache Superset

KPI and Metric Dashboards

Automating Workflow

Elyra and Kubeflow Pipelines

Kubeflow

OPEN DATA HUB
AI Platform powered by Open Source

Red Hat

Architecture

Architecture

Architecture

ML Service: Time to Merge Prediction

A new PR is opened on a Git Repository

Predict the time that it will take to merge this open PR, and add a label on the PR

Learn from the actual time to merge and create a feedback loop

Red Hat

Current workflow: Github time to merge prediction service

Collect data	Features	Model Building	Creating Service	Bot integration

Collect **data** from code repositories

Features like **size of the PR, types of files in the PR description of the PR**

Train repository or org specific model

Kubeflow based Model Pipelines

Creating **Model service** endpoint

Provide endpoint that can be accessed

Integrate service with repository using a **bot** to offer live feedback or a standalone site

Issues, PRs → **Discover project specific behavior** → **Model returns merge time estimate** → Model API Endpoint → **Provide estimates**

ML Service: Github Time to Merge Model

COLLECT DATA

openshift / **origin**

srcopsmetrics 2.11.1

```
pip install srcopsmetrics
```

Red Hat

ML Service: Github Time to Merge Model

GITHUB PR DATA → **FEATURE ENGINEERING**

🖥 openshift / **origin**

size of the PR,

types of files in the PR

description of the PR

Feature importance scores for Mutual Information Statistic

ML Service: Github Time to Merge Model

GITHUB PR DATA

openshift / **origin**

FEATURE ENGINEERING

size of the PR,

types of files in the PR

description of the PR

MODEL TRAINING

Naive Bayes

XGBoost

Random Forest

SVM

ML Service: Github Time to Merge Model

| GITHUB PR DATA | FEATURE ENGINEERING | MODEL TRAINING | MODEL DEPLOYMENT |

openshift / **origin**

size of the PR,

types of files in the PR

description of the PR

Naive Bayes

XGBoost

Random Forest

SVM

SELDON

RED HAT OPENSHIFT
Container Platform

Red Hat

ML Service: Github Time to Merge Model
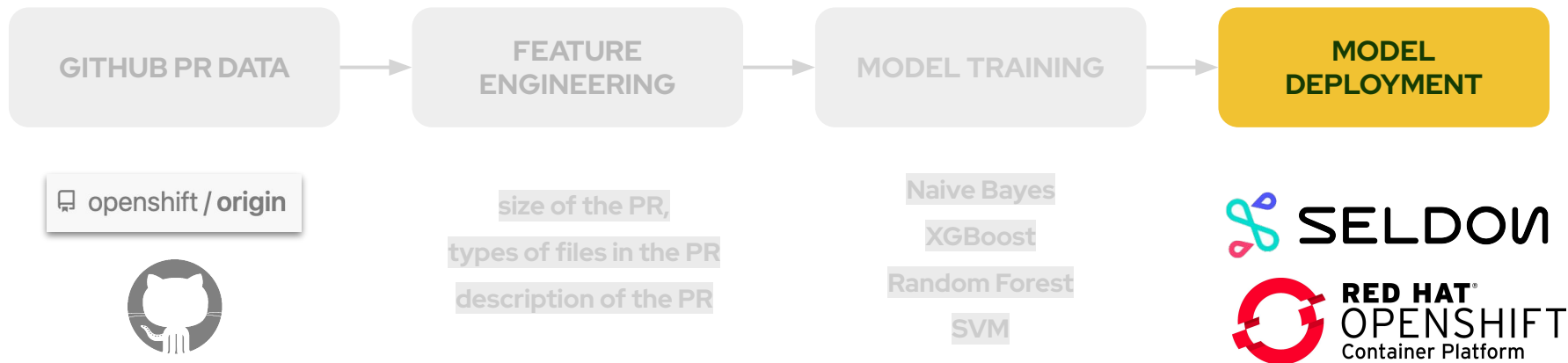
ML Service: Optimal Stopping Point Model

Sometimes tests/builds take longer than expected to run

Find an Optimal Stopping Point after which the test will fail.
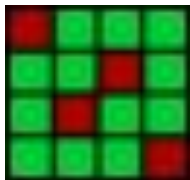
We can better allocate and save resources.
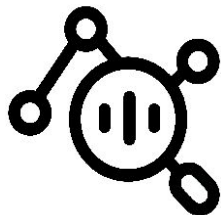
Red Hat

ML Service: Optimal Stopping Point Model
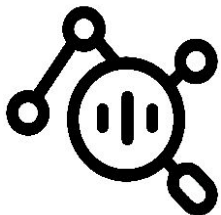
Data Collection

ML Service: Optimal Stopping Point Model
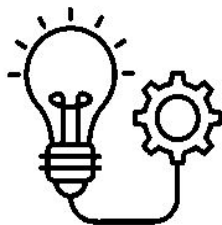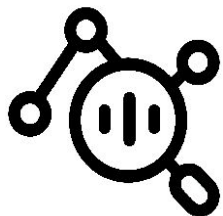

Data Collection

Feature Engineering

Find the **distribution type for passing and failing tests**. Probability density plots are used to find the probabilities of test duration
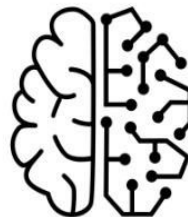
Red Hat

ML Service: Optimal Stopping Point Model

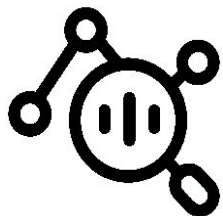**Data Collection**

Feature Engineering
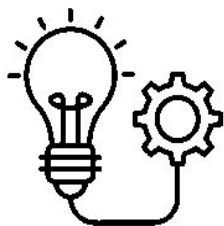
Model Training

Predict optimal stopping point by finding the point where:
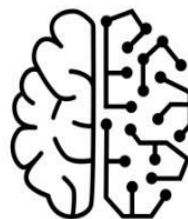**probability of failure > probability of passing**
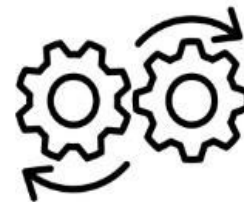
ML Service: Optimal Stopping Point Model

Data Collection **Feature Engineering** **Model Training** **Model Deployment**

Demo time!

Engage

Get Started

https://github.com/aicoe-aiops/ocp-ci-analysis/blob/master/docs/get-started.md

Open Data Sources	Notebooks	Dashboards	Model Endpoints	Automated Workflows	Video Playlist

https://tinyurl.com/aiforci

Red Hat

Operate First : operate-first.cloud/

- Get Started with the Operate First Cloud and Services :
  https://www.operate-first.cloud/getting-started
- Join the Operate First Data Science Community –
  https://www.operate-first.cloud/data-science/operate-first-data-science
  -community/docs/meetup-landing-page.md
- Video Playlist – https://www.youtube.com/c/OperateFirst

OPERATE FIRST>>

🐦 **@OperateFirst**

Red Hat