Accelerate Data Access for Kubernetes/Openshift Workloads with **Datashim**

Yiannis Gkoufas, IBM Research Europe, Ireland

LF AI & DATA

Background: Kubernetes and Data

- **Kubernetes < 1.6**
  Mostly used for stateless applications. Users could only leverage local storage

- **Kubernetes 1.6**
  Introduction of Dynamic Storage Provisioning and Storage Classes

- **Kubernetes 1.13**
  Container Storage Interface becomes GA

Motivation for our work

- **Non-power kubernetes users** face complexities when accessing remote data

- Becomes increasingly common to store/retrieve data from remote stores in **Machine Learning and AI workloads**

- **Kubernetes administrators** have tools to limit/track resource usage (CPUs, RAM) but not for data access

Introduction

LF AI & DATA
INCUBATION PROJECT

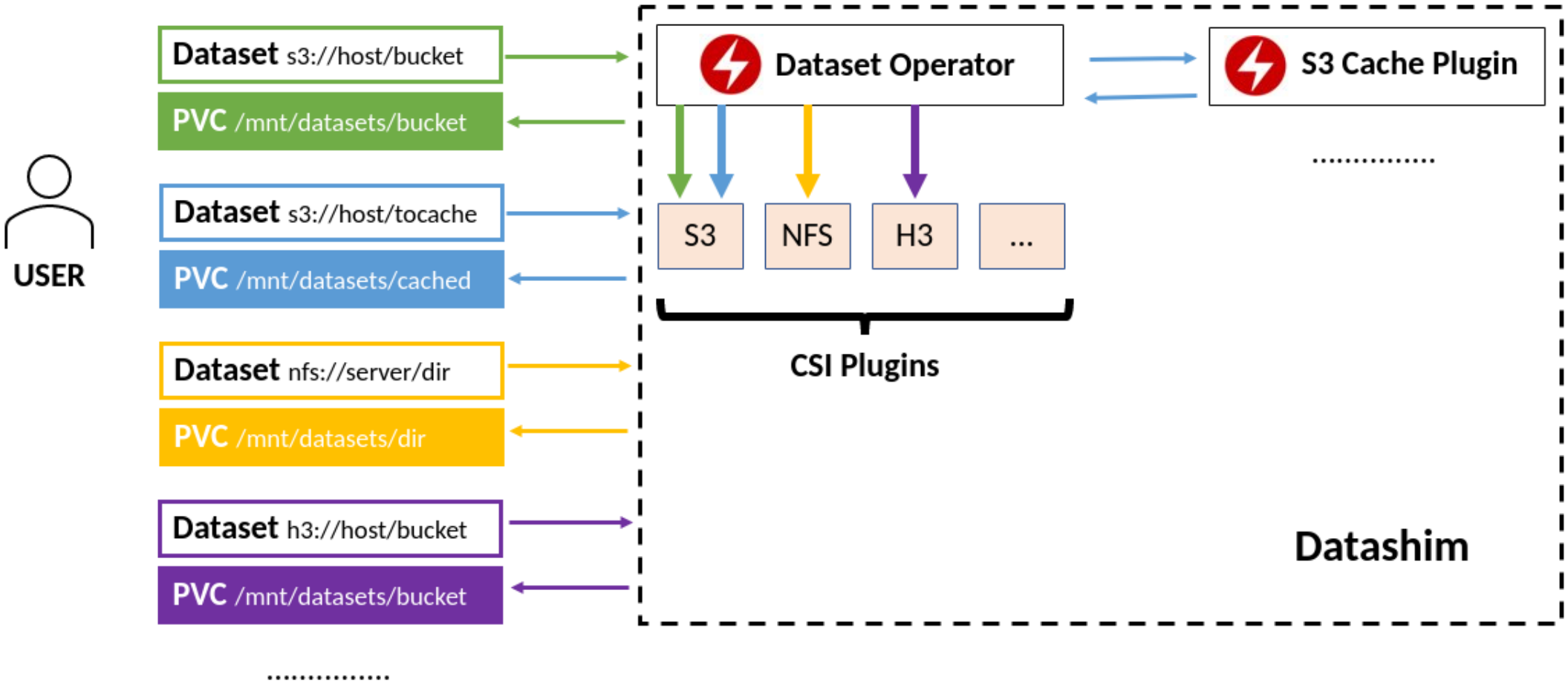
- **Kubernetes Framework** to assist users to access various types of **Datasources** in a consistent and performant manner

Goals

- Improve **User Experience** and **Performance** of K8s/Openshift workloads

- Be **highly extensible** to support all major types of Data sources

Project Overview



Dataset CRD Specification

```yaml
apiVersion: com.ie.ibm.hpsys/v1alpha1
kind: Dataset
metadata:
  name: test
spec:
  local:
    type: "COS"
    accessKeyID: "testKeyId"
    secretAccessKey: "testKey"
    endpoint: "https://s3.eu.cloud-object-storage.appdomain.cloud"
    bucket: "test-yiannis"
    region: "" #it can be empty
```

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: nginx
  labels:
    dataset.0.id: "test"
    dataset.0.useas: "mount"
spec:
  containers:
    - name: nginx
      image: nginx
```

Benefits

- **Data scientists/engineers:** Focus on workload/experiments development and not on configuring/tuning data access
- **Storage Providers:** Increase adoption since the framework is extensible without hindering the User Experience
- **Data-oriented Frameworks:** Can build capabilities (caching, scheduling) on top of Datashim using a declarative way to access/manage data sources

1000 Genomes Project

- Most detailed catalogue of **human genetic variation**
- **2500 individuals** from **26 different populations**
- **12 Terabytes** of Data hosted in an Amazon S3 Bucket

Collaboration with European Bioinformatics Institute

- Connected in **Kubeflow** community call
- Adopted **Datashim** on their pipeline which allowed them to minimize effort for accessing the **1000 Genomes** bucket
- Work to be presented in **VHPC 2021 : 16th Workshop on Virtualization in High-Performance Cloud Computing**

Github Metrics

- **Released:** September 2019
- **Donated to LF AI & Data:** January 2021
- **103 Stars, 28 Forks** (as of Jun 2020)
- **56 Closed Pull Requests, 27 Closed Issues**

Thank you!

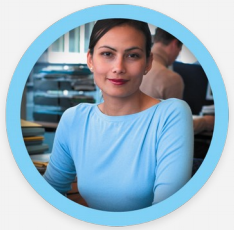

https://datashim.io/


https://github.com/datashim-io/datashim


Yiannis.gkoufas@gmail.com

APPENDIX

Scenario



Data Scientist

Search

Catalog

Publish datasets

Data Lake

Manage

Data Provider

Pain points
- *Locate datasets*
- *Manage credentials*
- *Configure Jobs*
- *Reduce wait*

Pain points
- *Configure access*
- *Enforce governance*

Launch Job

Access data

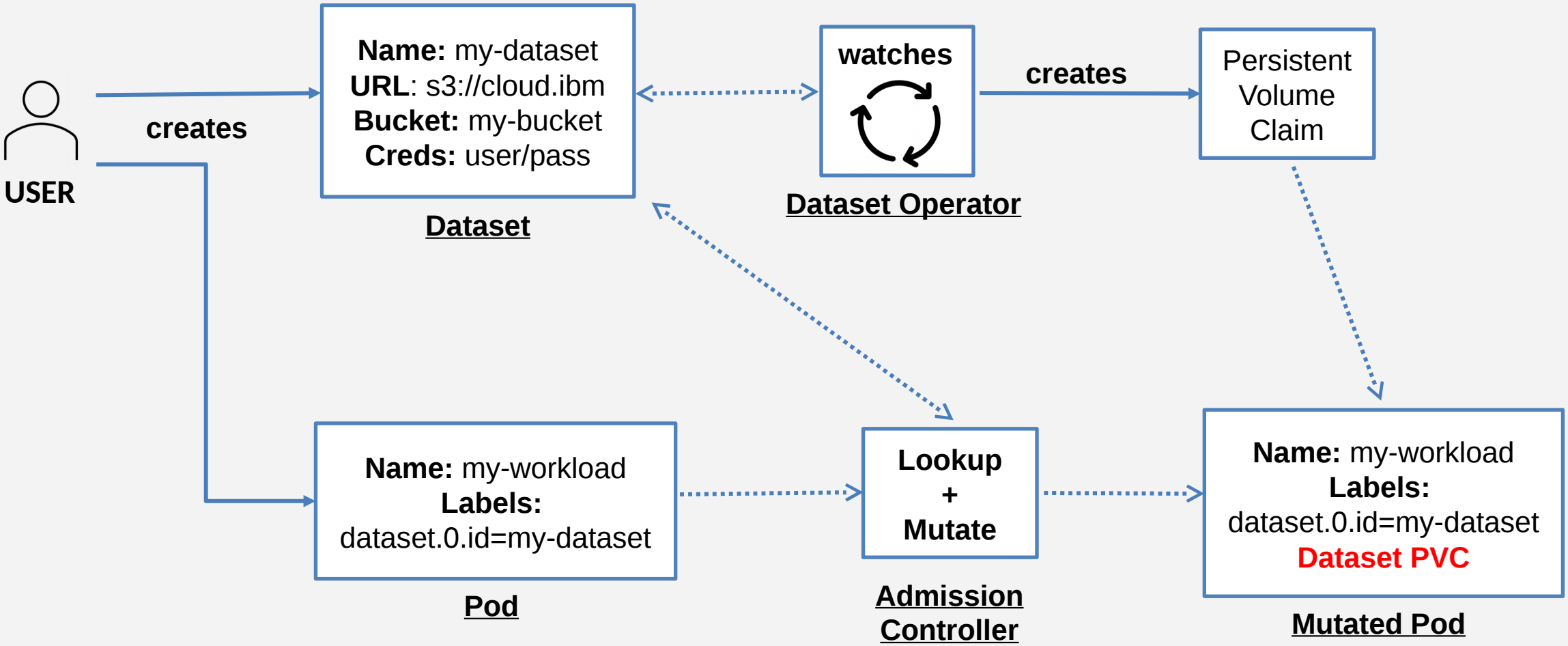
kubernetes OPENSHIFT

Produce

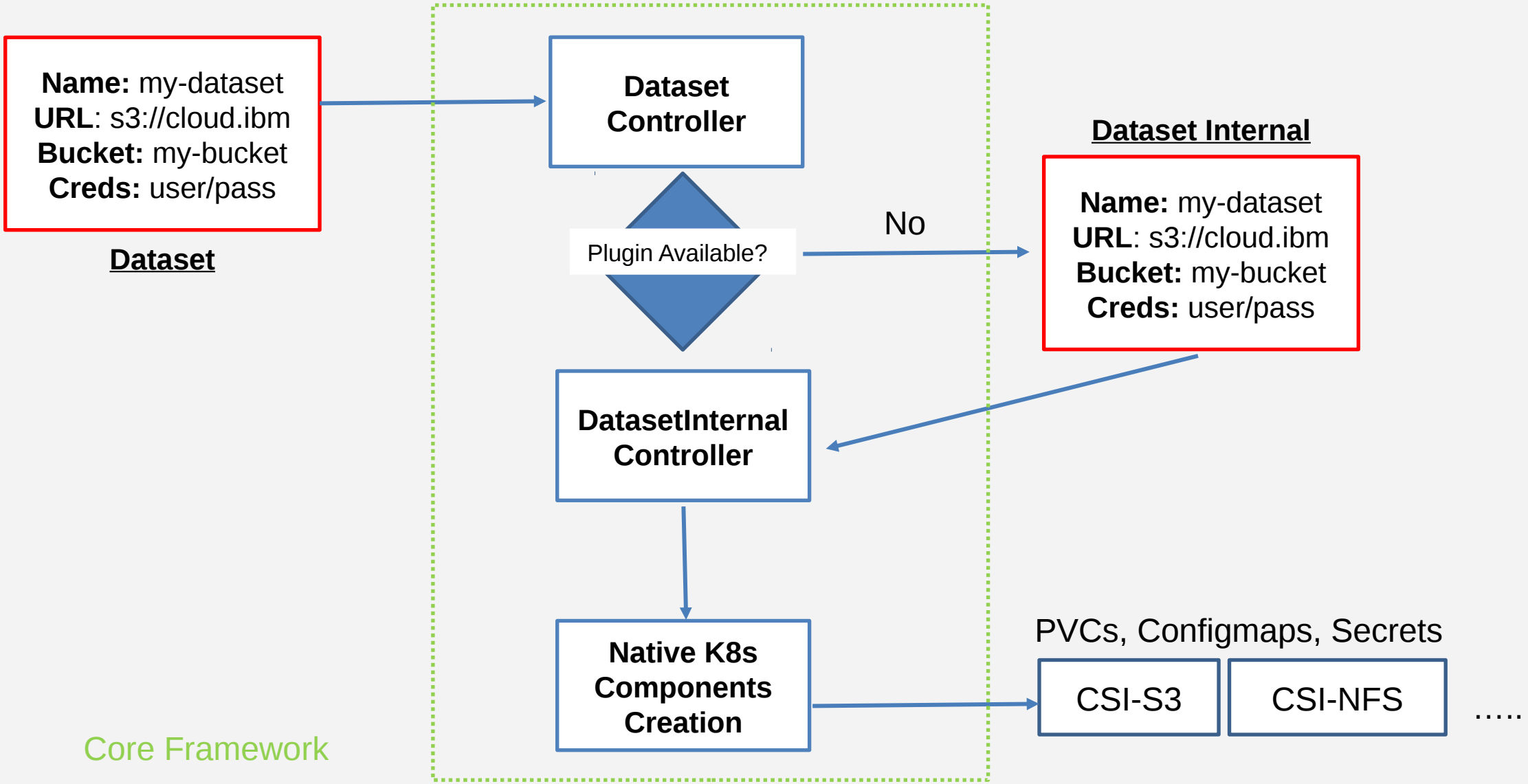
Insights

Cloud resources

Components


USER
creates
Name: my-dataset
URL: s3://cloud.ibm
Bucket: my-bucket
Creds: user/pass
Dataset
watches
Dataset Operator
creates
Persistent Volume Claim
Name: my-workload
Labels:
dataset.0.id=my-dataset
Pod
Lookup + Mutate
Admission Controller
Name: my-workload
Labels:
dataset.0.id=my-dataset
Dataset PVC
Mutated Pod

Transparent Caching



Transparent Caching