

Introduction to, using and use cases of KubeFlow

Jonathan Gershater & Boris Lublinsky

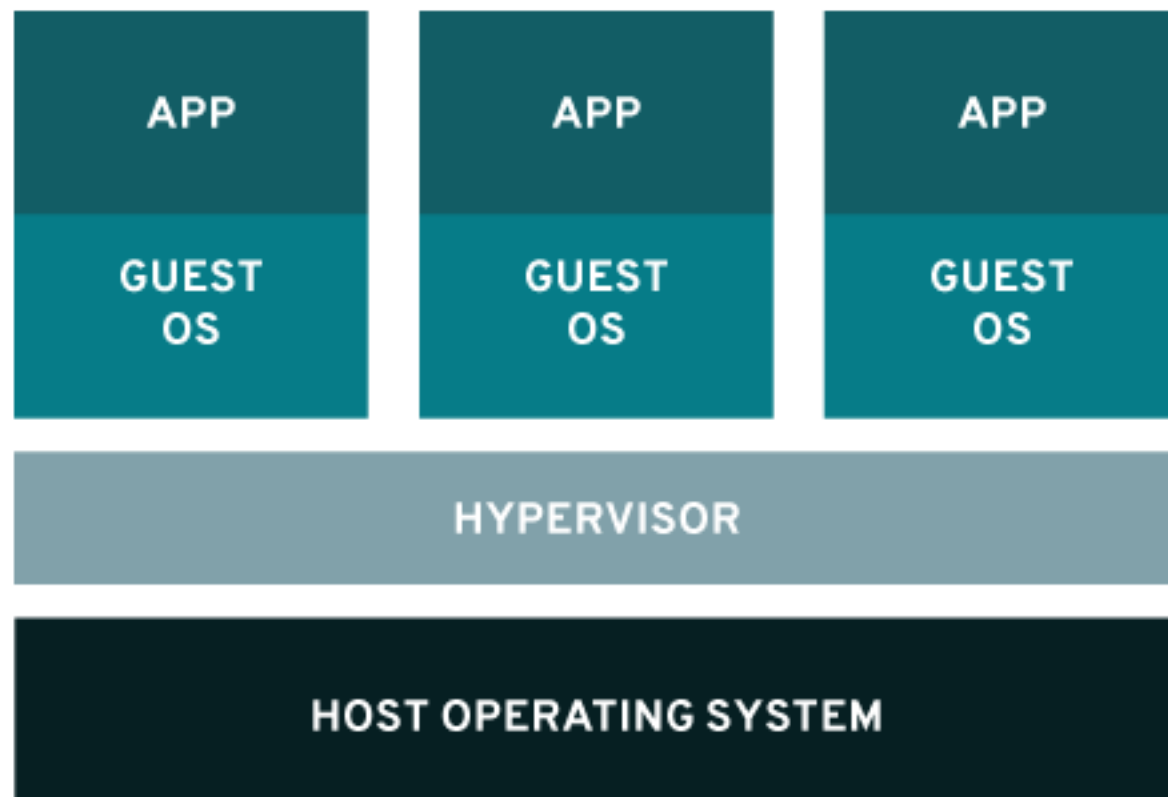


Agenda

- What is Kubernetes
- What is Kubeflow
- Workflow components
- End-to-end example

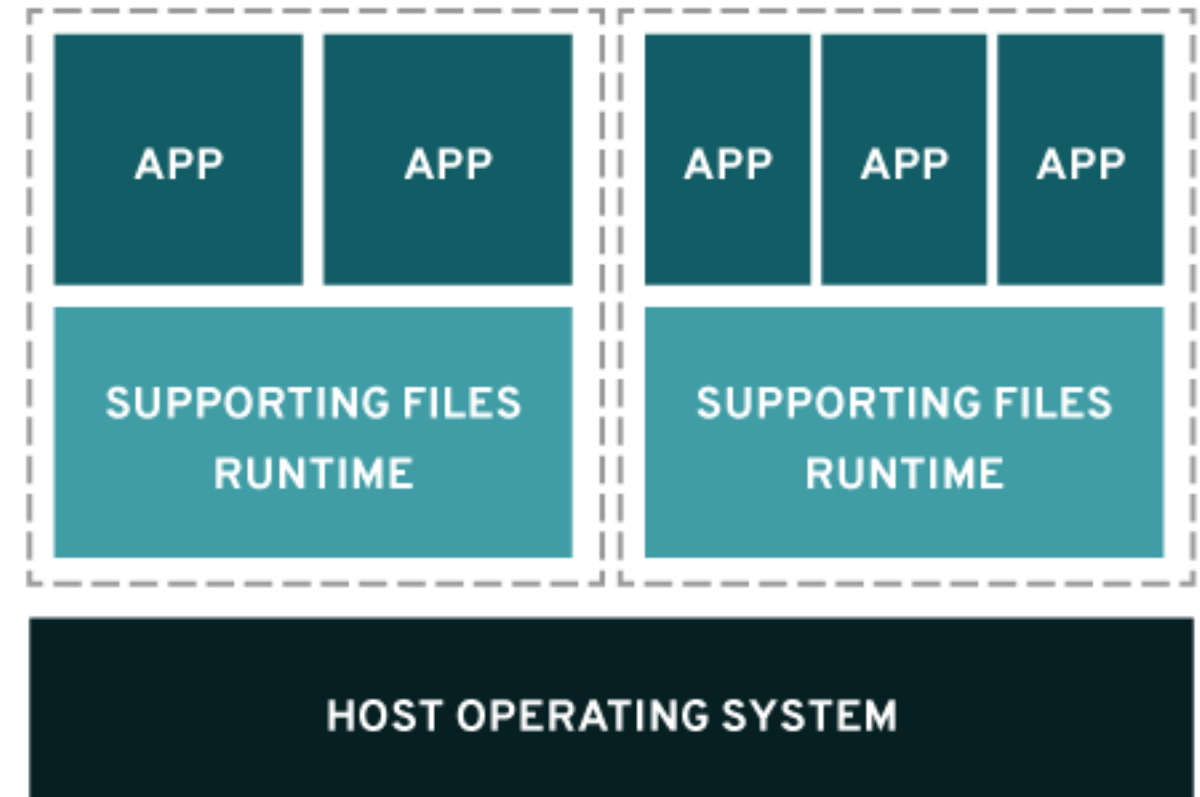
What are containers?

VIRTUALIZATION

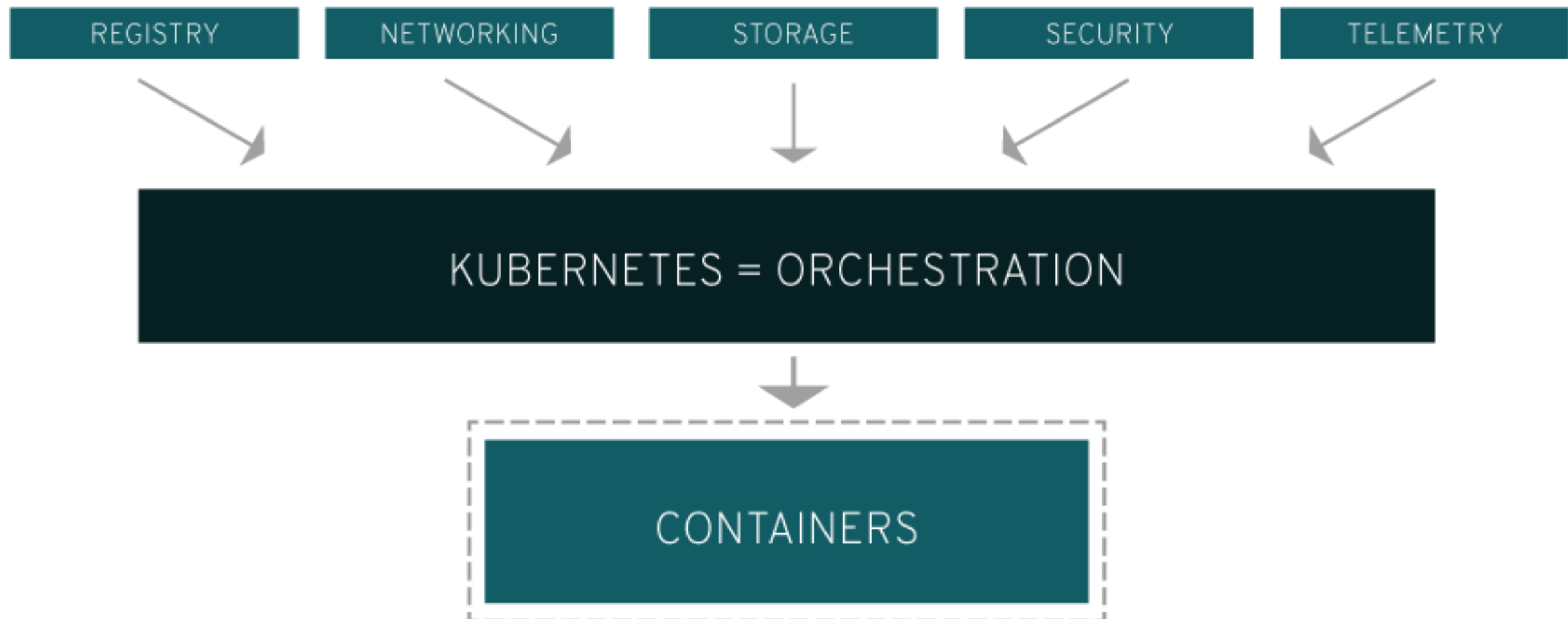


VS.

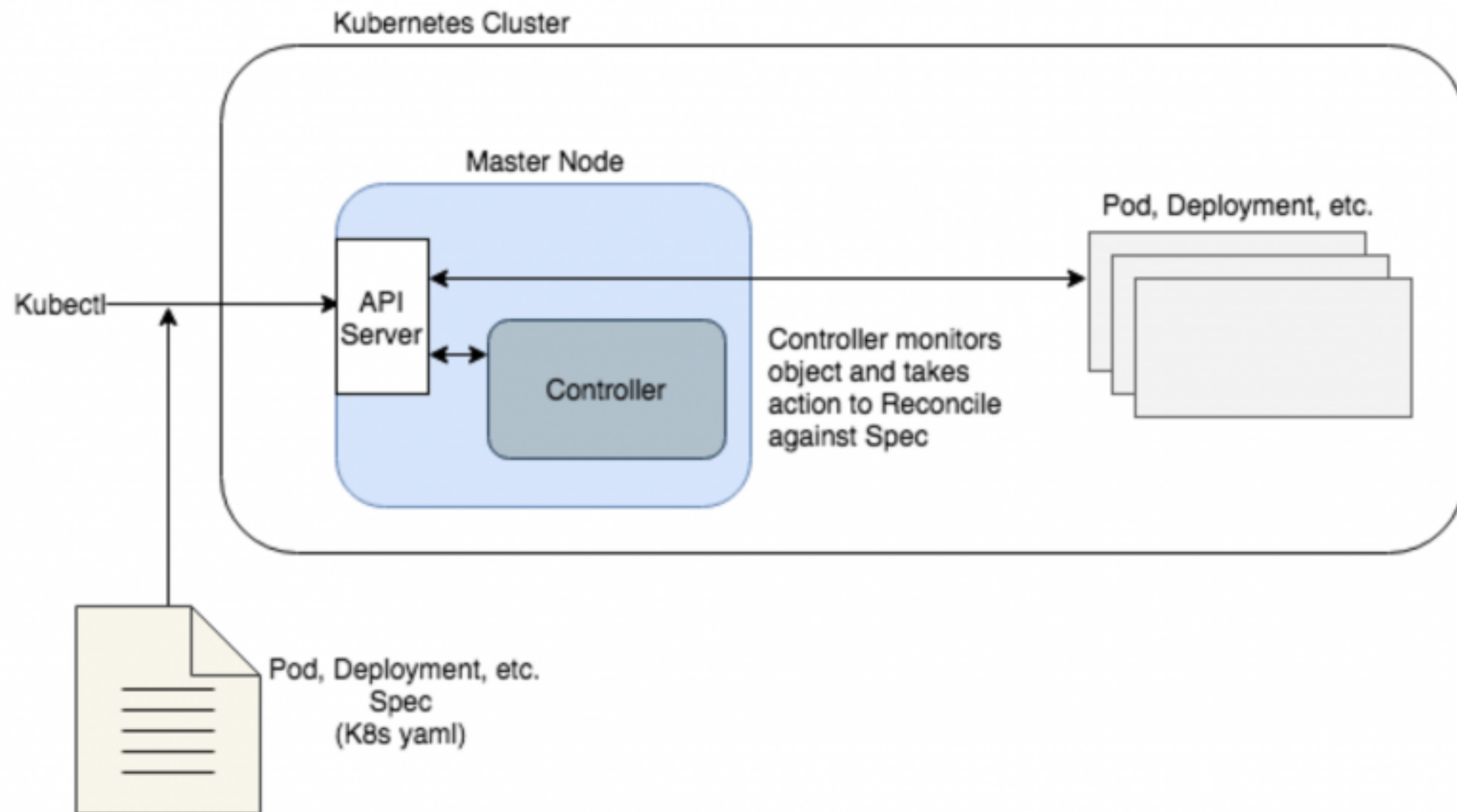
CONTAINERS



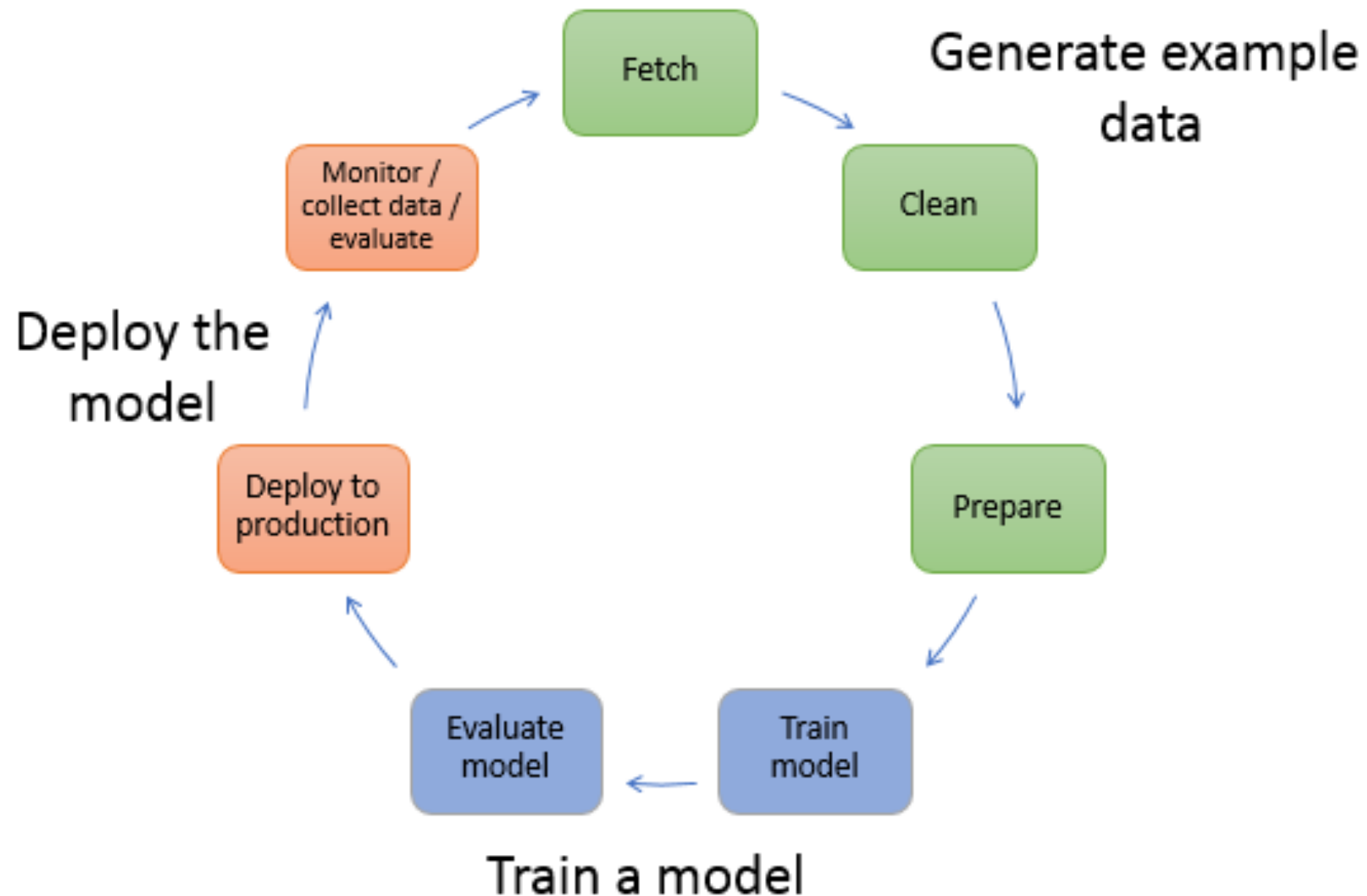
What is kubernetes



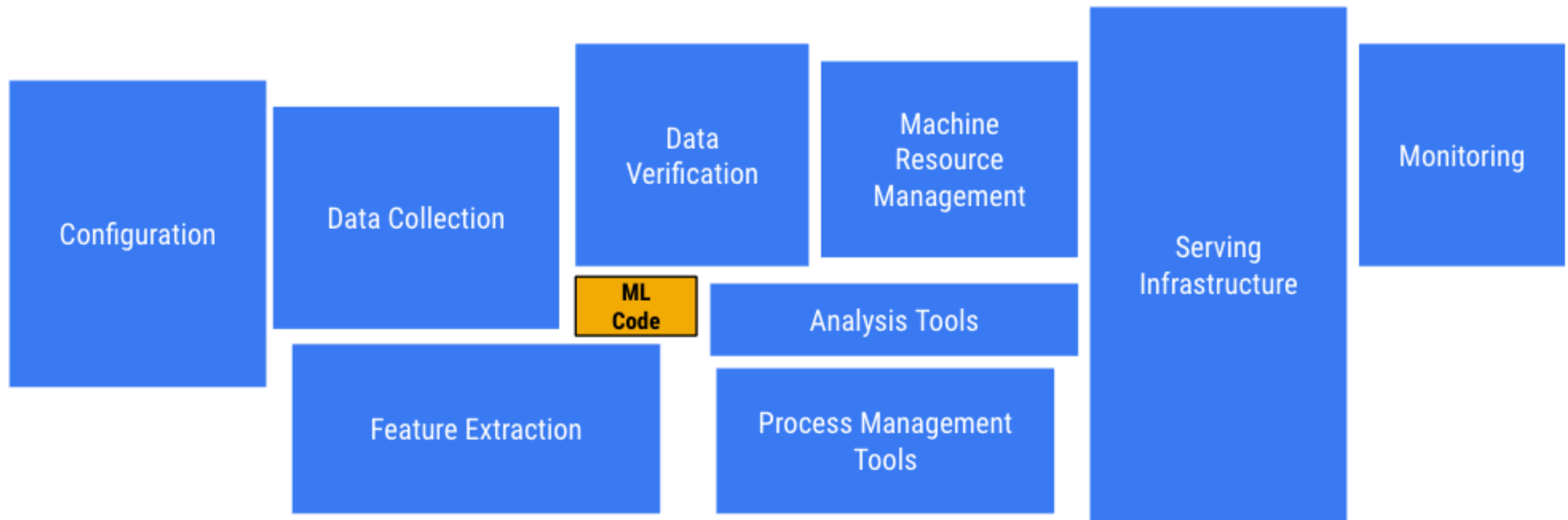
Kubernetes desired state management



What is machine learning?



Challenging of machine learning in production



Source <https://medium.com/kubeflow/why-kubeflow-in-your-infrastructure-56b8fabf1f3e>

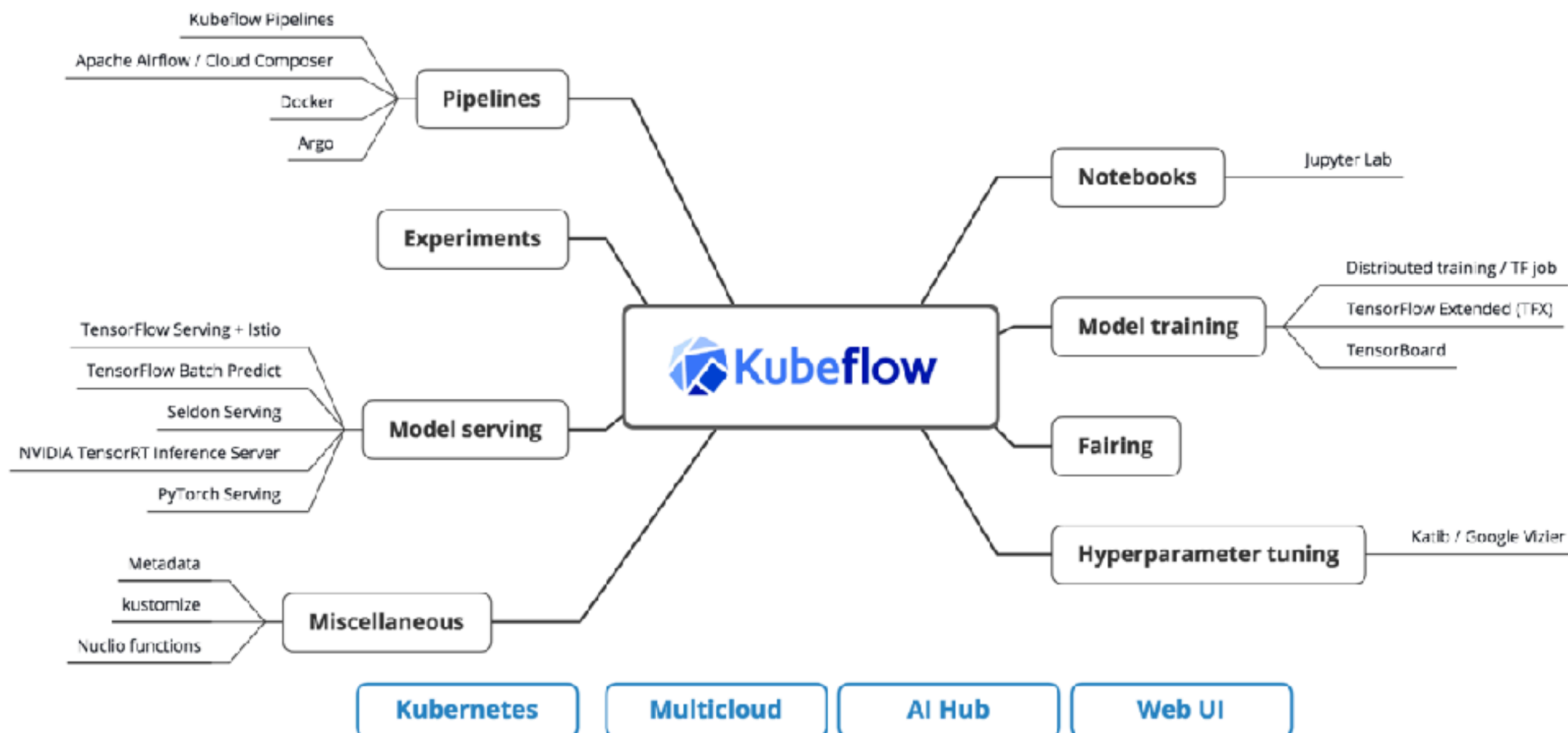
What is kubeflow?

Make deployments of machine learning (ML) workflows on Kubernetes simple, portable and scalable.

Provide a straightforward way to deploy open-source systems for ML to diverse infrastructures.

Anywhere you are running Kubernetes, you should be able to run Kubeflow.

Kubeflow components



Version 1.1 20190807 @MichalBrys

Source <https://medium.com/@michal.brys/kubeflow-a-machine-learning-toolkit-for-kubernetes-d8686f6c91b6>

Scalability

- Machine learning (deep learning) today is a result of the larger scale and capacity available in the cloud.
- And various machine types and hardware-specific accelerators (e.g., graphics processing units/Tensor processing units),
- And data locality for improved performance.
- Scale teams through collaboration and simplify the running of a large number of experiments.

What's new in 0.6?

- Secure architecture for multi-user support by leveraging a new integration with Istio
- Extensions to the Kubeflow Pipelines' Domain Specific Language (DSL) by adding new primitives that enable data and execution context versioning in Kubeflow Pipelines
- New metadata component, along with a metadata API and initial corresponding clients. These allow users to track their artifacts and execution contexts through an end-to-end ML workflow.
- Several documentation updates and valuable new operational and configuration capabilities, most notably the introduction of Kustomize as a complete replacement of ksonnet.

Meet Kubeflow

The screenshot shows the Kubeflow dashboard in a web browser. The browser's address bar displays the URL `istio-ingress-istio-system.apps.streampipe.streampipe.lightbend.com/?ns=istio-system`. The dashboard has a dark blue sidebar on the left with the Kubeflow logo and navigation links: Home, Pipelines, Notebook Servers, Katib, Artifact Store, GitHub, and Documentation. The main content area has a top bar with 'istio-system' and a user profile icon. Below this, there are tabs for 'Dashboard' (selected) and 'Activity'. The dashboard is divided into several sections: 'Quick shortcuts' with links to upload a pipeline, view pipeline runs, create a new notebook server, view Katib studies, and view metadata artifacts; 'Recent Notebooks' showing 'No Notebooks in namespace istio-system'; 'Recent Pipelines' listing five sample pipelines with their creation times; 'Recent Pipeline Runs' showing 'None Found'; and 'Documentation' with links to 'Getting Started with Kubeflow', 'MiniKF', 'Microk8s for Kubeflow', 'Minikube for Kubeflow', 'Kubeflow on GCP', 'Kubeflow on AWS', and 'Requirements for Kubeflow'. The footer of the sidebar contains 'Privacy • Usage Reporting' and 'build version v1beta1'.

istio-system

Dashboard Activity

Quick shortcuts

- ⚡ Upload a pipeline
Pipelines
- ⚡ View all pipeline runs
Pipelines
- ⚡ Create a new Notebook server
Notebook Servers
- ⚡ View Katib Studies
Katib
- ⚡ View Metadata Artifacts
Artifact Store

Recent Notebooks

No Notebooks in namespace istio-system

Recent Pipelines

- 🔧 [Sample] Basic - Exit Handler
Created 8/8/2019, 12:15:23 PM
- 🔧 [Sample] Basic - Conditional execution
Created 8/8/2019, 12:15:21 PM
- 🔧 [Sample] Basic - Parallel execution
Created 8/8/2019, 12:15:20 PM
- 🔧 [Sample] Basic - Sequential execution
Created 8/8/2019, 12:15:19 PM
- 🔧 [Sample] ML - TFX - Taxi Tip Prediction Model T...
Created 8/8/2019, 12:15:18 PM

Recent Pipeline Runs

None Found

Documentation

- Getting Started with Kubeflow**
Get your machine-learning workflow up and running on Kubeflow
- MiniKF**
A fast and easy way to deploy Kubeflow locally
- Microk8s for Kubeflow**
Quickly get Kubeflow running locally on native hypervisors
- Minikube for Kubeflow**
Quickly get Kubeflow running locally
- Kubeflow on GCP**
Running Kubeflow on Kubernetes Engine and Google Cloud Platform
- Kubeflow on AWS**
Running Kubeflow on Elastic Container Service and Amazon Web Services
- Requirements for Kubeflow**
Get more detailed information about using Kubeflow and its components

Privacy • Usage Reporting
build version v1beta1

Cluster view

Red Hat

OpenShift Container Platform

kubeadmin

Home

Catalog

Workloads

Pods

Deployments

Deployment Configs

Stateful Sets

Secrets

Config Maps

Cron Jobs

Jobs

Daemon Sets

Replica Sets

Replication Controllers

Horizontal Pod Autoscalers

Networking

Storage

Builds

Monitoring

Compute

Administration

You are logged in as a temporary administrative user. Update the [cluster OAuth configuration](#) to allow others to log in.

Project: kubeflow

Resources

Dashboard

Actions

Health

Kubernetes API

UP

All good

OpenShift Console

UP

All good

Alerts Firing

0

Alerts

Crashlooping Pods

0

Pods

Resource Usage

CPU Usage

Memory Usage

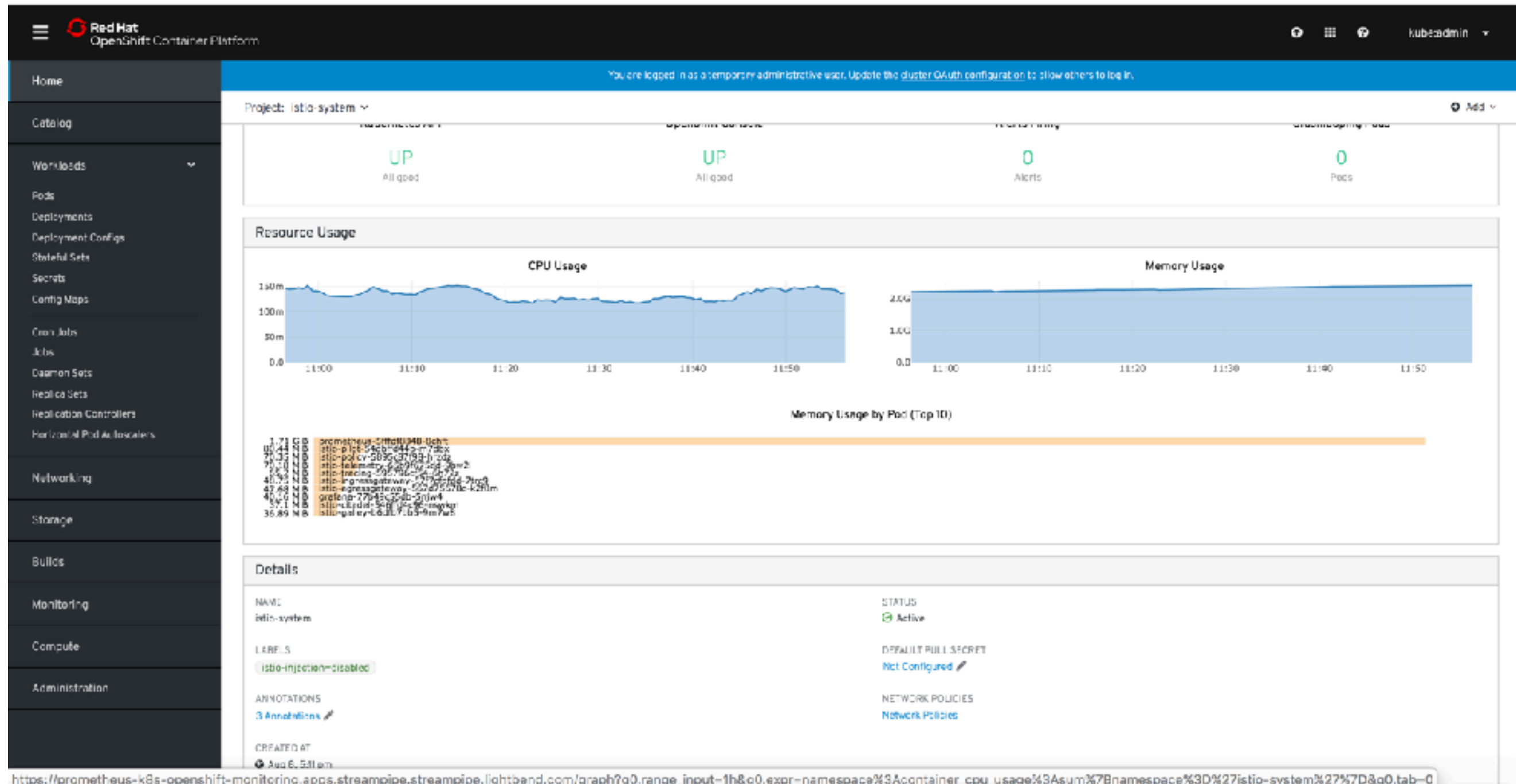
Memory Usage by Pod (Top 10)

555 MB	mycibf0588959-scrt
312 MB	katib-cb-659e189d-mbrat
309 MB	metacat-43-2f-nc05b1d-wjmw2
272 MB	minio-05c65c144-br879
195 MB	ml-pipeline-cb-70529d-m2dc
139 MB	tensorboard-71d3cd75c-6zdc
120 MB	katib-controller-0750c744-cp7w2
88 MB	katib-experiment-runner-7f6b8f1f7-zmug
61 MB	centraldashboard-55b4b975d-2loeb
70 MB	ml-pipeline-ui-68c889ce48-ctam

Details

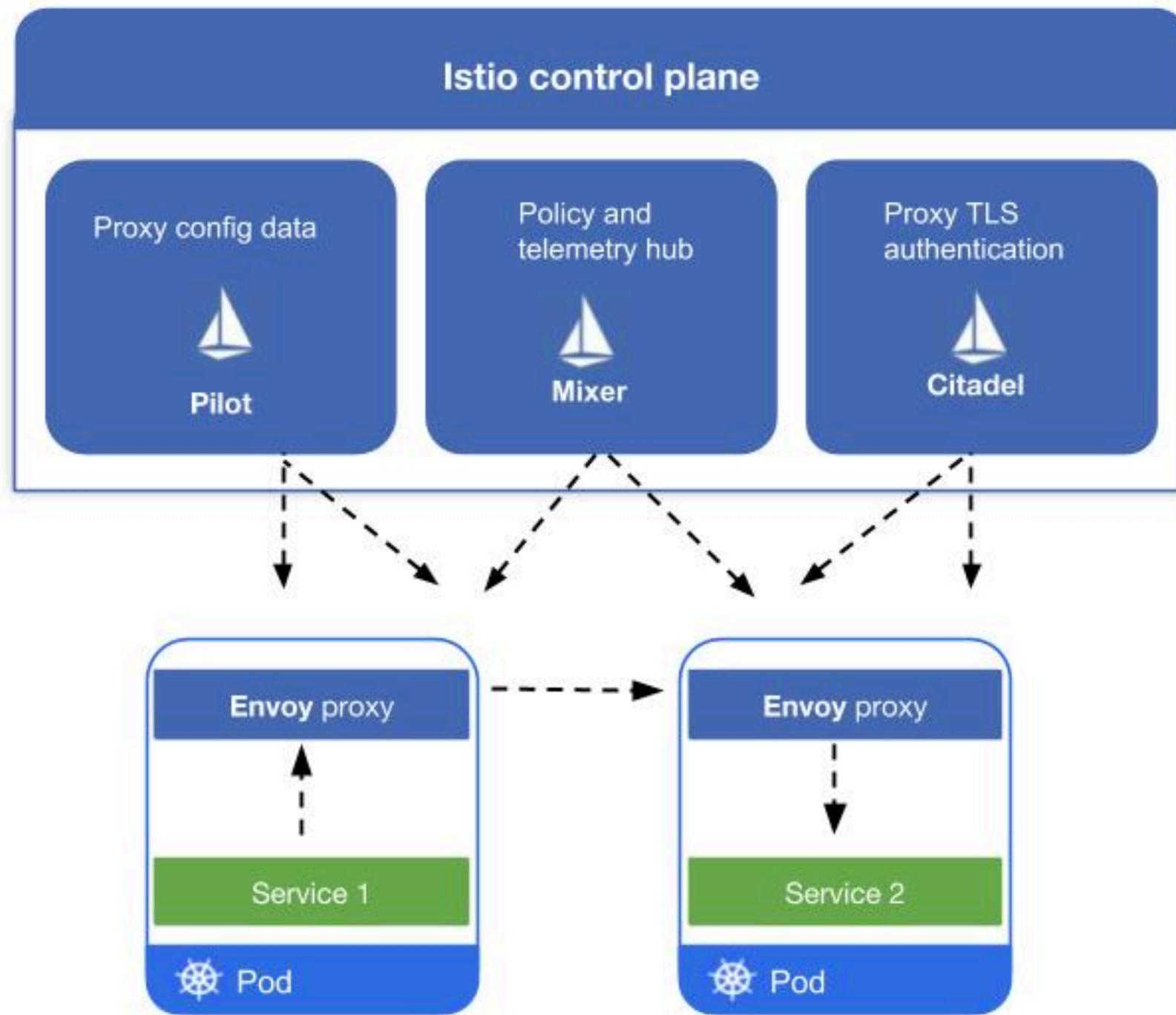
NAME	STATUS
kubeflow	Active
LABELS	DEFAULT PULL SECRET
No labels	Not Configured
ANNOTATIONS	NETWORK POLICIES

Cluster view



https://prometheus-k8s-openshift-monitoring.apps.streampipe.streampipe.lightband.com/graph?g0.range_input=1h&g0.expr=namespace%3Acontainer_cpu_usage%3Asum%7Bnamespace%3D%27istio-system%27%7D&g0.tab=0

Istio

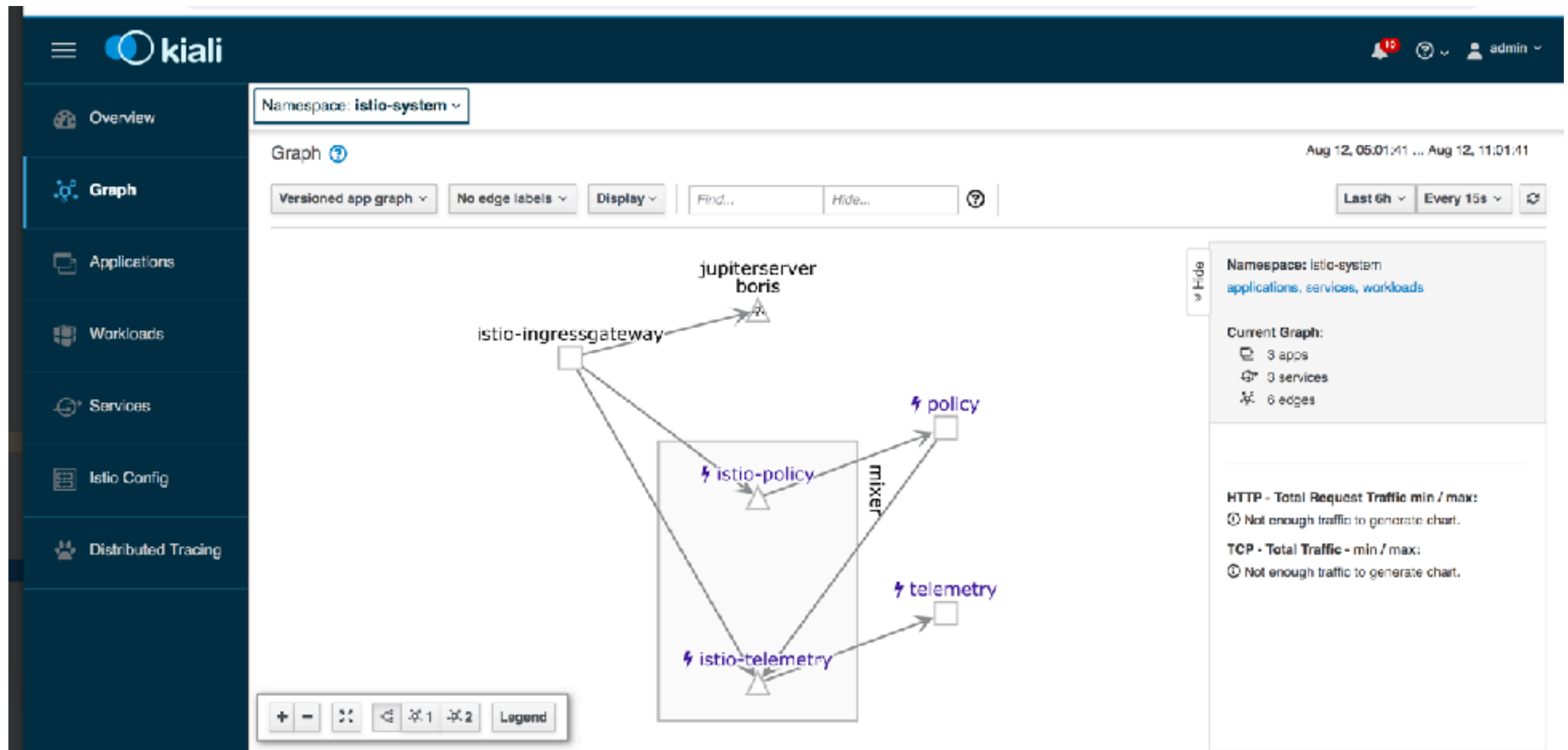


Istio in KubeFlow

The screenshot displays the Kiali dashboard interface. On the left is a dark blue sidebar with navigation links: Overview, Graph, Applications, Workloads, Services, Istio Config, and Distributed Tracing. The main content area is titled 'Namespaces' and features a grid of 12 namespace health cards. Each card shows the namespace name, the number of applications, a health status (green checkmark for healthy, N/A for not applicable), and four small icons at the bottom representing different Istio components. The top of the dashboard includes a header with the Kiali logo, a user profile 'admin', and filters for 'Name', 'Filter by Name', 'Show health for Apps', and refresh intervals 'Last 1m' and 'Every 15s'.

Namespace	Applications	Health
boris	5	Healthy
default	0	N/A
istio-system	12	Healthy
kubeflow	20	Healthy
openshift	0	N/A
openshift-apiserver	1	Healthy
openshift-apiserver-operator	1	Healthy
openshift-authentication	1	Healthy
openshift-authentication-operator	1	Healthy
openshift-cloud-credential-operator	0	N/A
openshift-cluster-machine-approver	1	Healthy
openshift-cluster-node-tuning-operator	0	N/A

Istio in KubeFlow



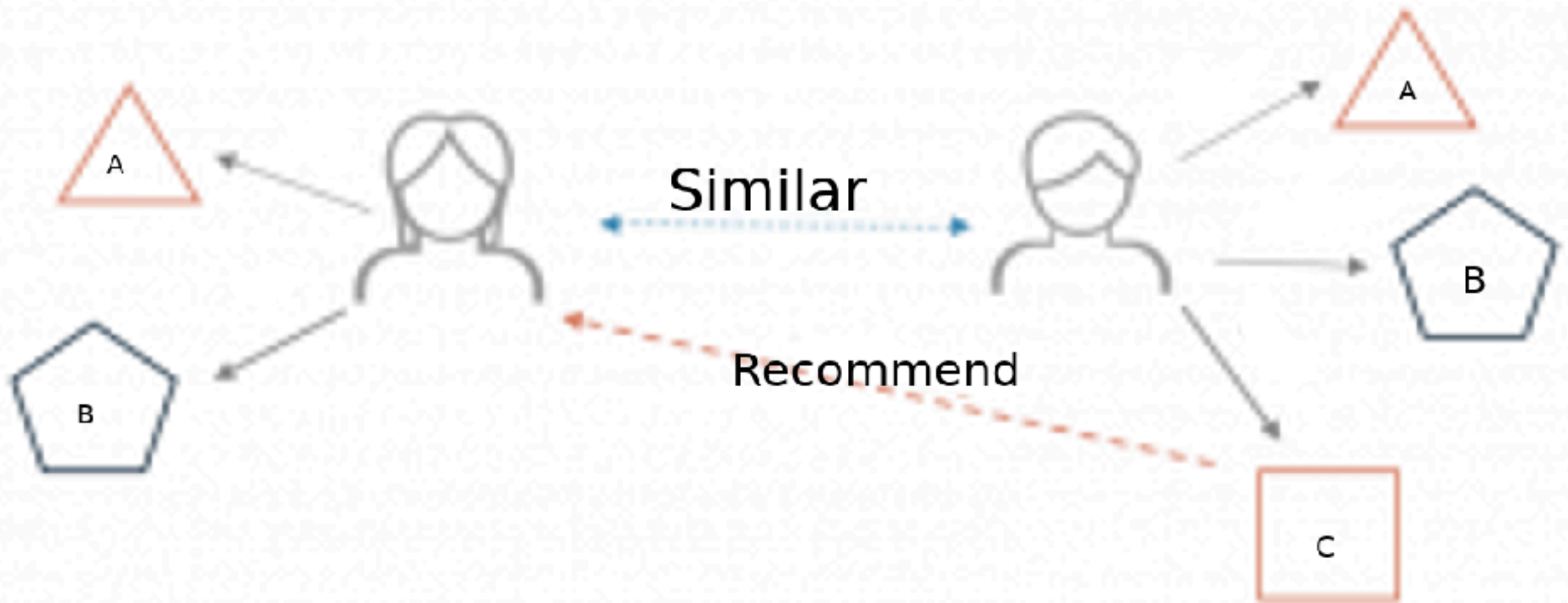
Motivating example

For our example we will use implementation of a product recommender.



Source <https://towardsdatascience.com/recommender-system-a1e4595fc0f0>

Collaborative filtering



Source <https://towardsdatascience.com/how-to-build-a-simple-recommender-system-in-python-375093c3fb7d>

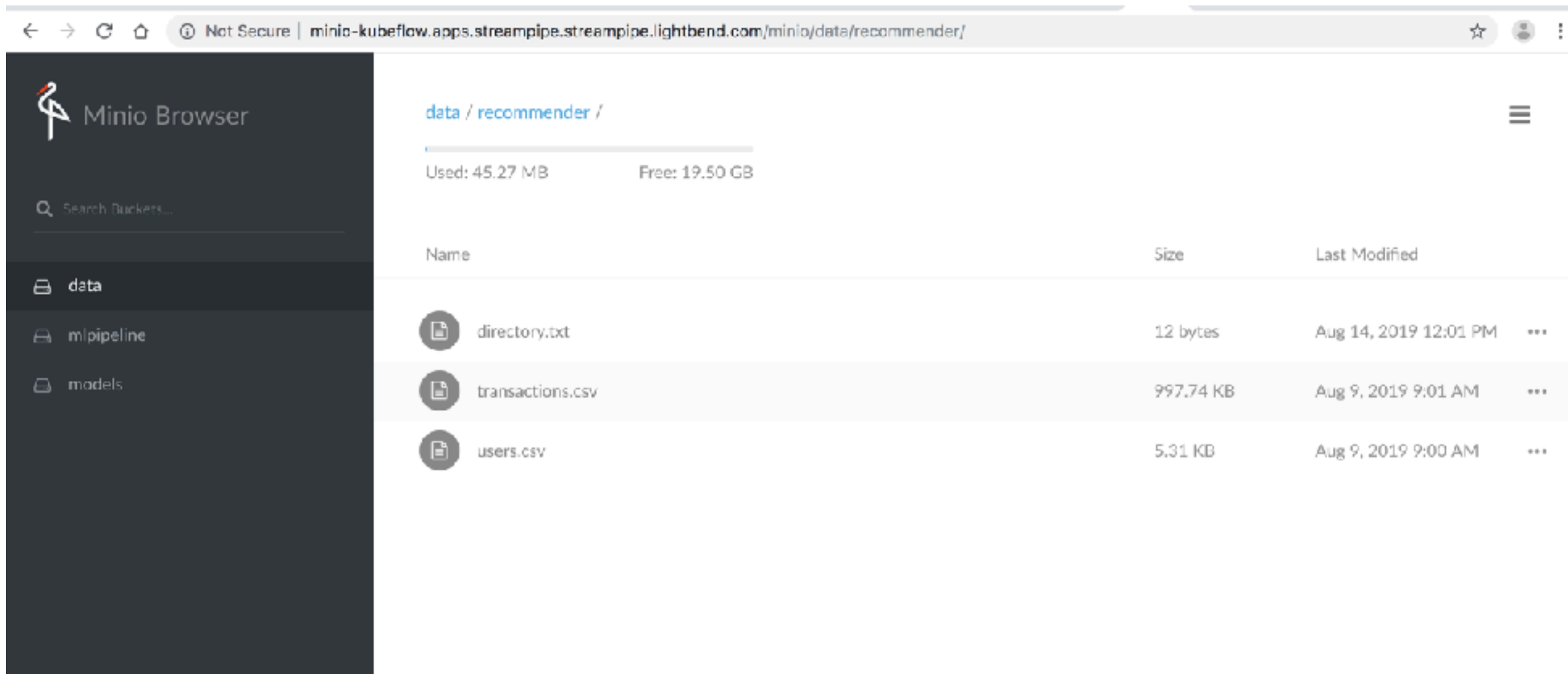
Rating Matrix

Items

		1	2	3	4	...	n
Users	1	4	3			5	
	2	5		4		4	
	3	4		5	3	4	
	4		3				5
	...		4				4
	m			2	4		5

Using Minio as a centralized storage

Minio, a high-performance distributed object storage server, designed for large-scale private cloud infrastructure.

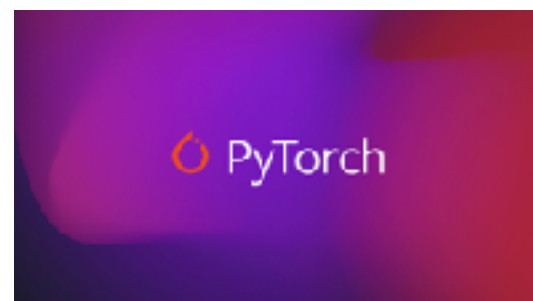


The screenshot displays the Minio Browser web interface. The left sidebar shows a navigation menu with 'data', 'mlpipeline', and 'models' folders. The main content area shows the 'data / recommender /' directory. A progress bar indicates 'Used: 45.27 MB' and 'Free: 19.50 GB'. Below this, a table lists the files in the directory:

Name	Size	Last Modified
directory.txt	12 bytes	Aug 14, 2019 12:01 PM
transactions.csv	997.74 KB	Aug 9, 2019 9:01 AM
users.csv	5.31 KB	Aug 9, 2019 9:00 AM

Kubeflow options for machine learning and model serving




Learning




Serving



Using Jupiter for creating implementation

 Kubeflow  boris 

 **Name**


Specify the name of the Notebook Server and the Namespace it will belong to.

Name

jupyterboris

Namespace

boris


 **Image**

A starter Jupyter Docker Image with a baseline deployment and typical ML packages.

☐ Custom Image

Image

gcr.io/kubeflow-images-public/tensorflow-1.13.1-notebook-cpu:v0.5.0

 **CPU / RAM**

Specify the total amount of CPU and RAM reserved by your Notebook Server. For CPU-intensive workloads, you can choose more than 1 CPU (e.g. 1.5).



CPU

2.0

Memory




20.0Gi

Using Jupiter for creating implementation

 Kubeflow  boris

Notebook Servers

[+ NEW SERVER](#)

Status	Name	Age	Image	CPU	Memory	Volumes	
	jupiterserver	15 mins ago	tensorflow-1.13.1-notebook-cpu:v0.5.0	2.0	10.0Gi		 CONNECT 

 jupyter


Quit

Files **Running** Clusters

Select items to perform actions on them.

Upload New 

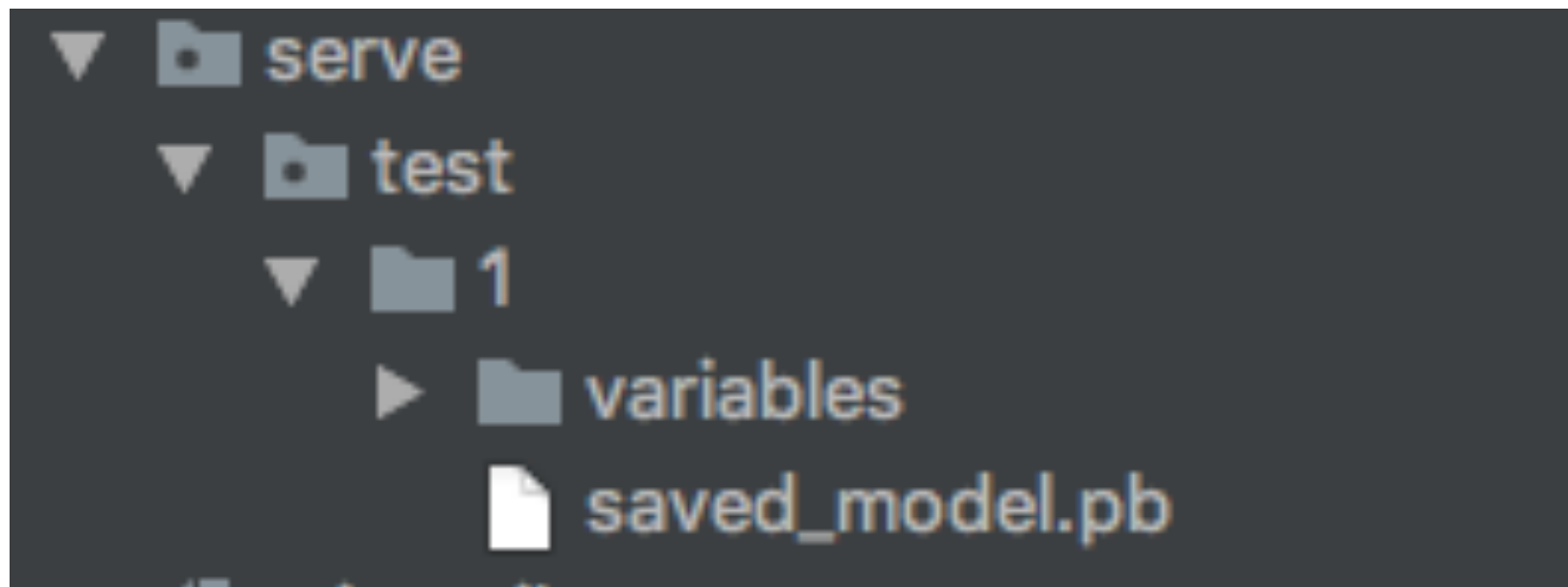
☐ 0  /

Name  Last Modified File size

The notebook list is empty.

Jupyter - exporting model

Export model in a standard tensorflow [saved model format](#), which defines both the layout of the content of the saved model on disk and the information stored in the model. The layout on disk looks as follows:



Converting implementation to TFJob

A distributed TensorFlow job typically contains some of the following processes:

- Chief - responsible for orchestrating training and performing tasks like checkpointing the model.
- Ps - parameter servers; these servers provide a distributed data store for the model parameters.
- Worker - who does the actual work of training the model. In some cases, worker 0 might also act as the chief.
- Evaluator - can be used to compute evaluation metrics as the model is trained.

Converting implementation to TFJob

- Export notebook as Python code.
- Build docker image

```
FROM tensorflow/tensorflow:1.12.0-devel-py3  
RUN pip3 install --upgrade pip  
RUN pip3 install pandas --upgrade  
RUN pip3 install keras --upgrade  
RUN pip3 install minio --upgrade  
RUN mkdir -p /opt/kubeflow  
COPY Recommender_Kubeflow.py /opt/kubeflow/  
ENTRYPOINT ["python3", "/opt/kubeflow/  
Recommender_Kubeflow.py"]
```

Converting implementation to TFJob

```
apiVersion: "kubeflow.org/v1"
kind: "TFJob"
metadata:
  name: "recommederjob"
  namespace: boris
spec:
  cleanPodPolicy: None
  tfReplicaSpecs:
    Worker:
      replicas: 1
      restartPolicy: Never
      template:
        spec:
          containers:
            - name: tensorflow
              image: lightbend/ml-tf-recommender:0.0.1
              imagePullPolicy: "Always"
              env:
                - name: "MINIO_URL"
                  value: "minio-service.kubeflow.svc.cluster.local:9000"
                - name: "MINIO_KEY"
                  valueFrom: { secretKeyRef: { name: "minioaccess", key:
"AWS_ACCESS_KEY_ID" } }
                - name: "MINIO_SECRET"
```

Running TF Job

kubeflow/tf-operator

boris/recommenderjob

Namespace

All namespaces

 **recommenderjob**
worker: 1

Name: recommenderjob


Namespace: boris

Created on: 2019-08-10T17:41:07Z

Status: TFJobRunning

Replicas: 1

Image: lightbend/ml-tf-recommender:0.0.1

Name	Status	Logs
recommenderjob-worker-0	Running	

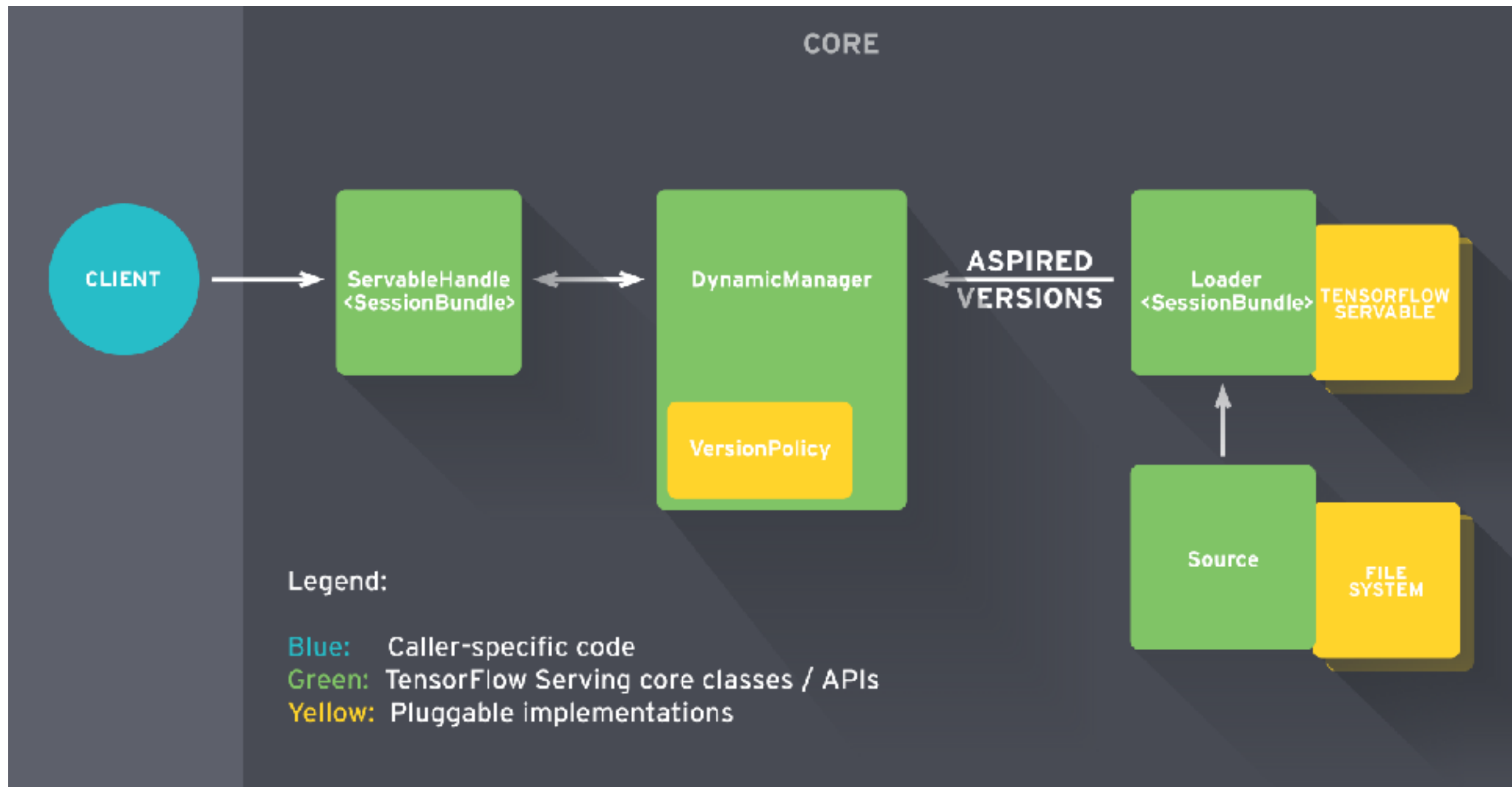
Running TF Job

Logs

Layer (type)	Output Shape	Param #	Connected to
=====			
user (InputLayer)	(None, 1)	0	
product (InputLayer)	(None, 1)	0	
embedding_1 (Embedding)	(None, 1, 50)	1430300	user[0][0]
embedding_2 (Embedding)	(None, 1, 50)	15000	product[0][0]
reshape_1 (Reshape)	(None, 50)	0	embedding_1[0][0]
reshape_2 (Reshape)	(None, 50)	0	embedding_2[0][0]
concatenate_1 (Concatenate)	(None, 100)	0	reshape_1[0][0]

CLOSE

TF - serving



Deploying TF-serving

Create and deploy secret to access Minio

```
apiVersion: v1
kind: Secret
metadata:
  name: minioaccess
data:
  AWS_ACCESS_KEY_ID: xxxxxxxxxxxx
  AWS_SECRET_ACCESS_KEY: xxxxxxxxxxxxxxxxxxxxxxxxxxxx
```

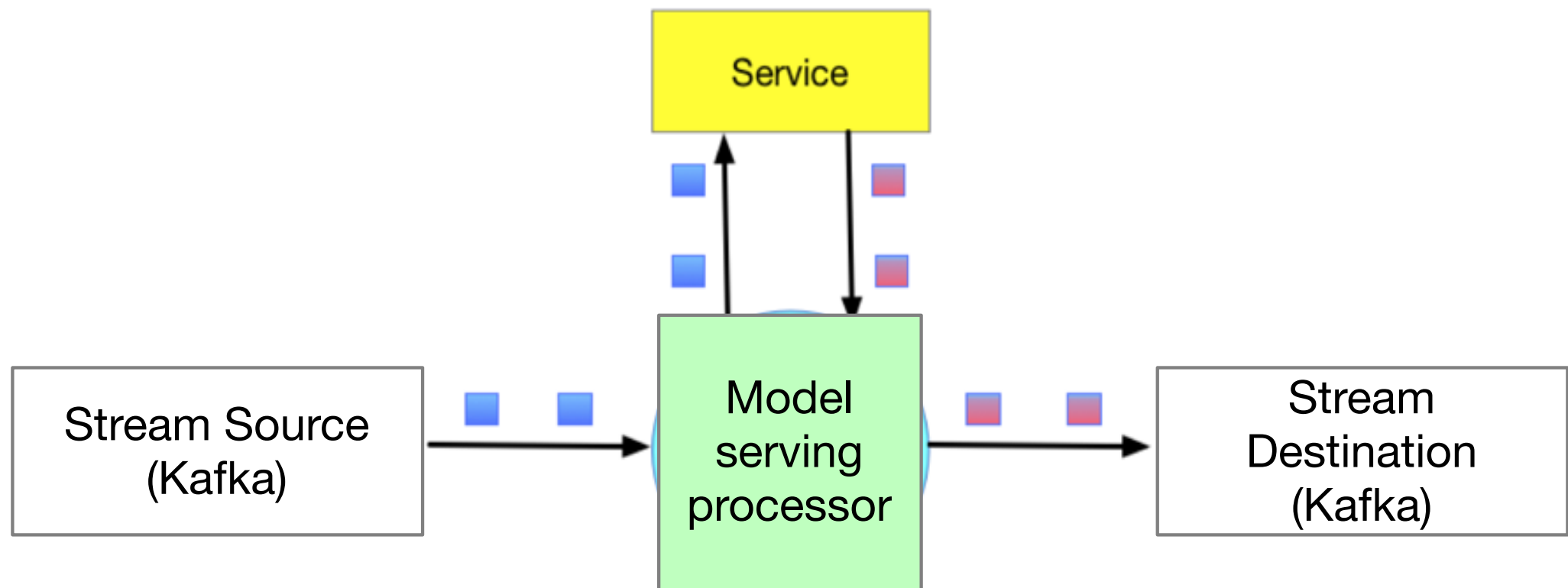

Deploying TF-serving

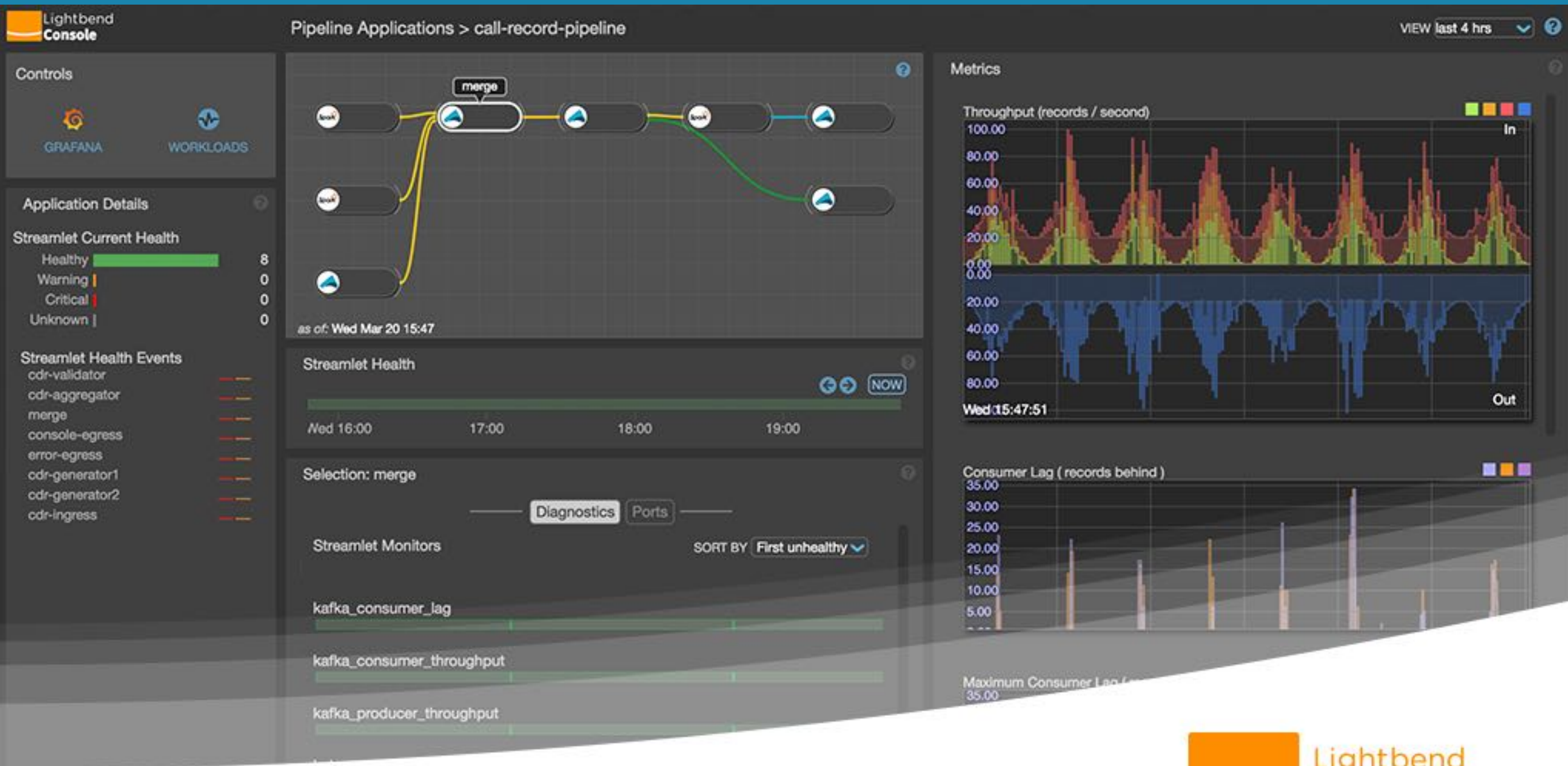
TF-serving is not yet ported into Kubeflow 0.6. As a result I used custom Helm chart

Validating TF-Serving

```
curl -X POST http://recommender-service-
kubeflow.lightshift.lightbend.com/v1/models/recommender/
versions/1:predict -d
'{"signature_name":"serving_default","inputs": {"products":
[[1],[2]], "users" : [[25], [3]]}}'
{
  "outputs": {
    "model-version": [
      "1"
    ],
    "recommendations": [
      [
        0.185666
      ],
      [
        0.0364329
      ]
    ]
  }
}
```

Using TF Serving in streaming applications





What we're up to at Lightbend...
lightbend.com/lightbend-pipelines-demo



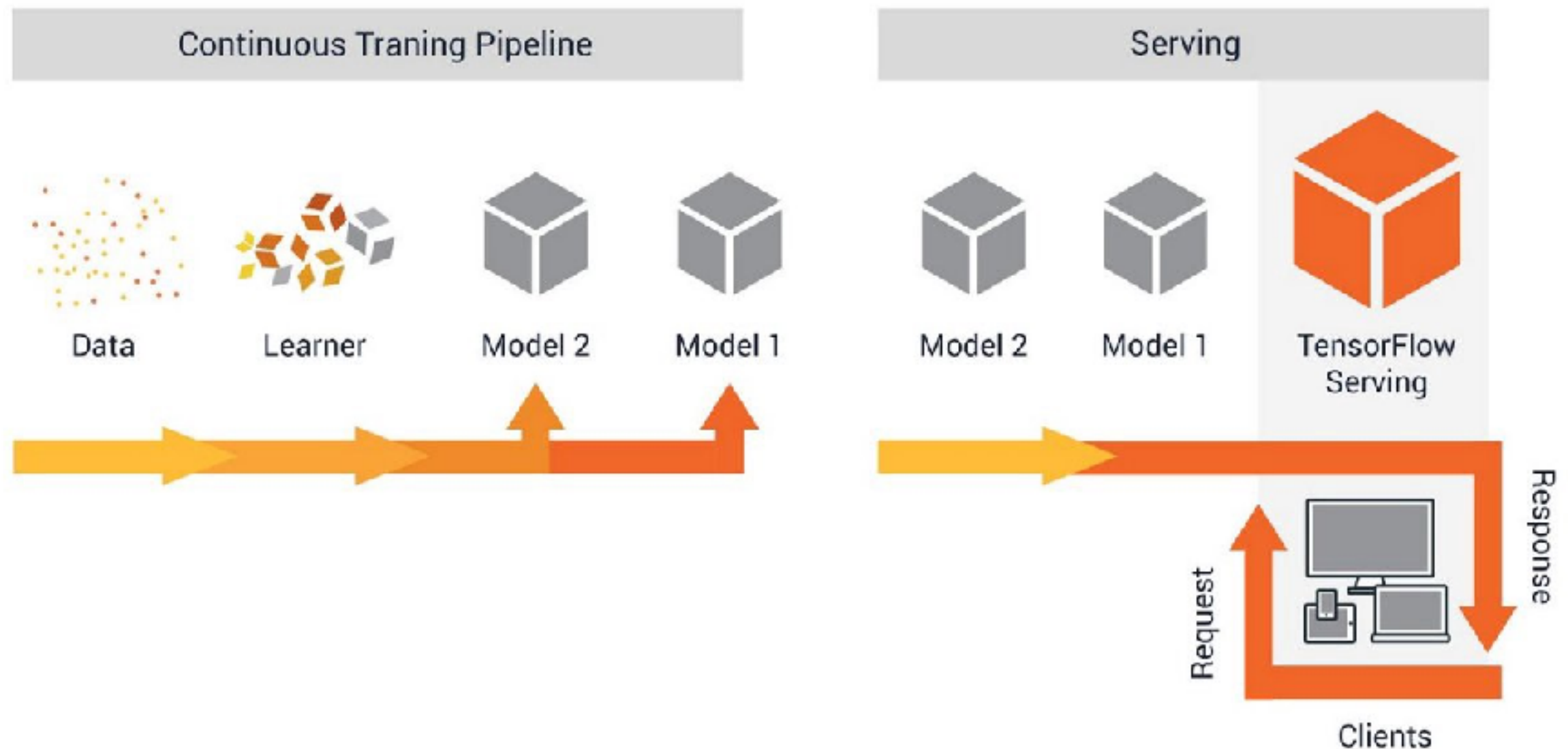
Concept drift

In most challenging data analysis applications, data evolve over time and must be analyzed in near real time. Patterns and relations in such data often evolve over time, thus, models built for analyzing such data quickly become obsolete over time. In machine learning and data mining this phenomenon is referred to as concept drift.

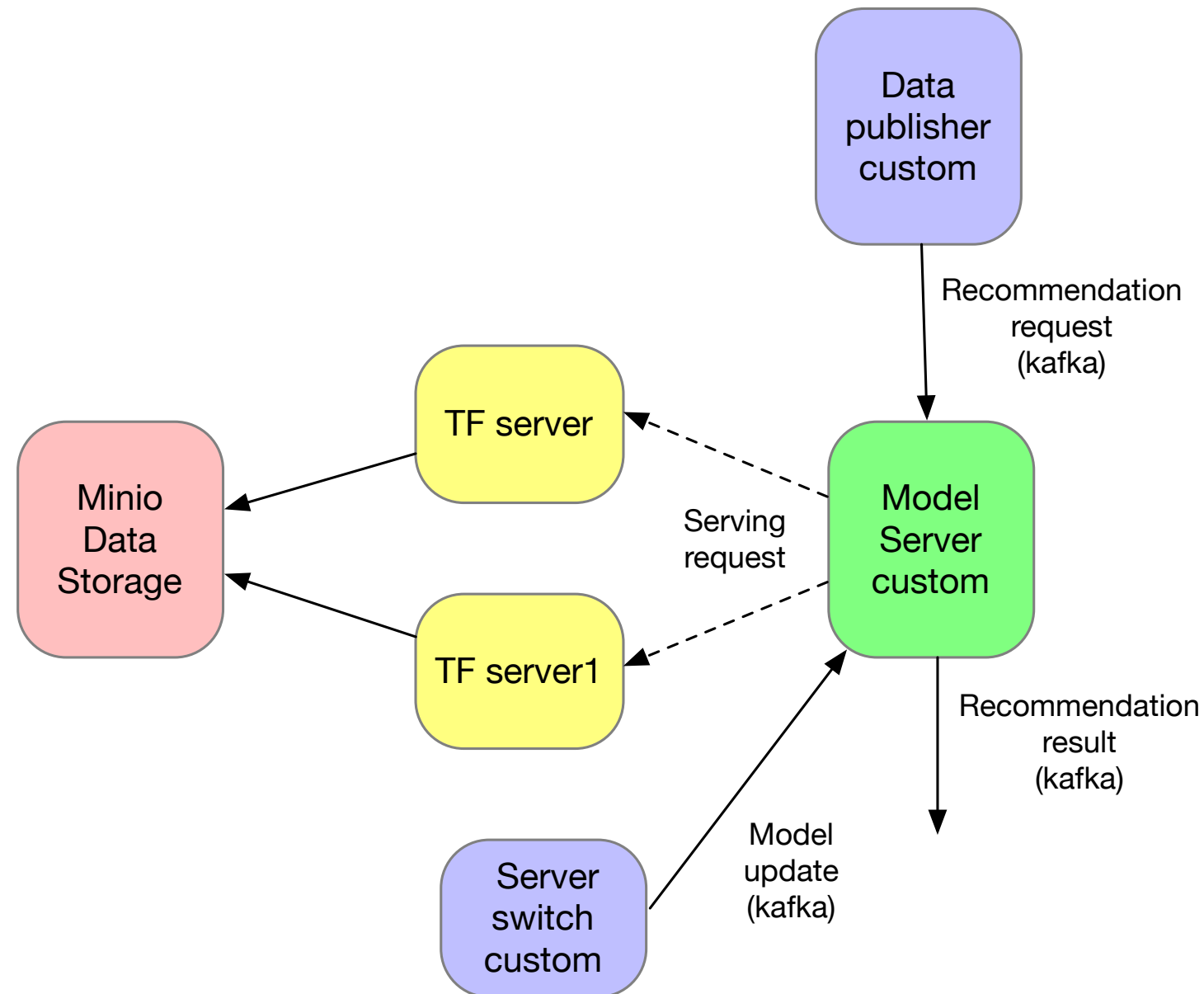
Source https://www.win.tue.nl/~mpechen/publications/pubs/CD_applications15.pdf

Continuous model updates

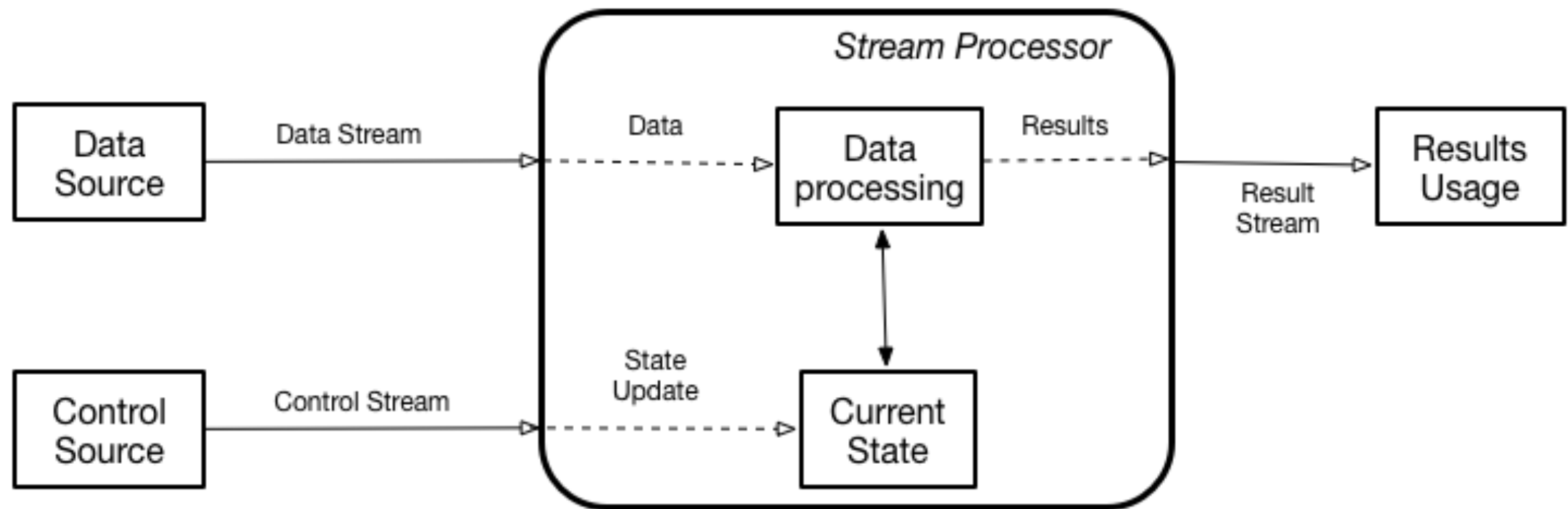
Serve models in production with TensorFlow Serving



Continuous model updates implementation



TF Serving in streaming applications



Additional custom components

- Model serving data provider - generator to produce recommendation requests
- Data updater for model serving - component responsible for publishing of new data for machine learning
- Model updater for model serving - component responsible for updating Tensorflow server information for model serving

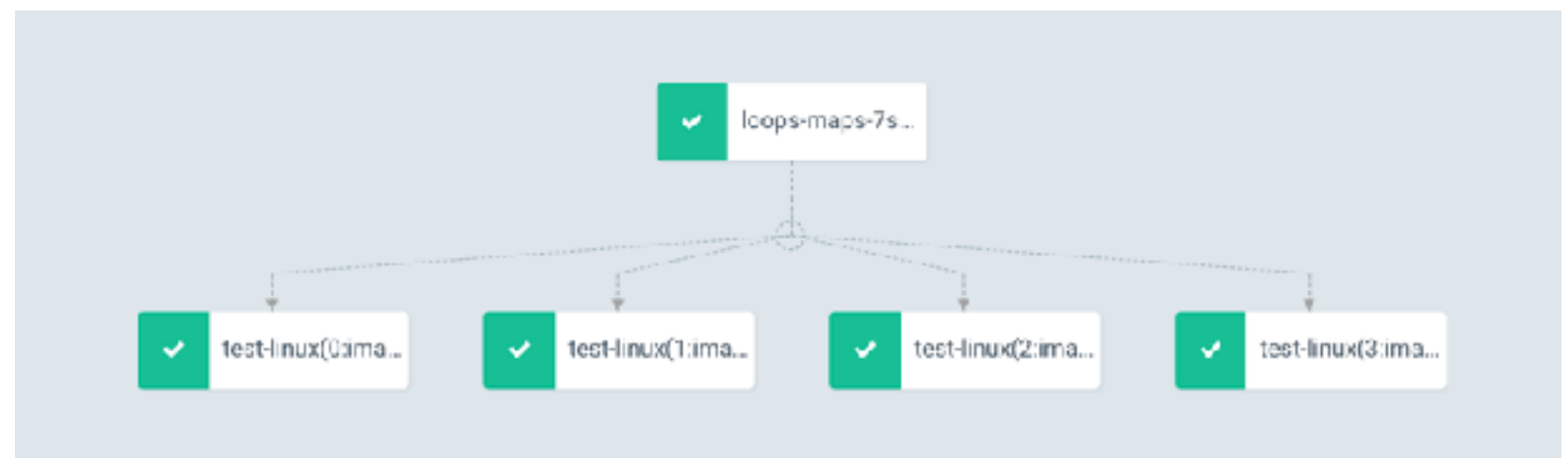
Kubeflow pipelines

Kubeflow Pipelines is a platform for building and deploying portable, scalable machine learning (ML) workflows based on Docker containers. The Kubeflow Pipelines platform consists of:

- A user interface (UI) for managing and tracking experiments, jobs, and runs.
- An engine for scheduling multi-step ML workflows
- An SDK for defining and manipulating pipelines and components.
- Notebooks for interacting with the system using the SDK.

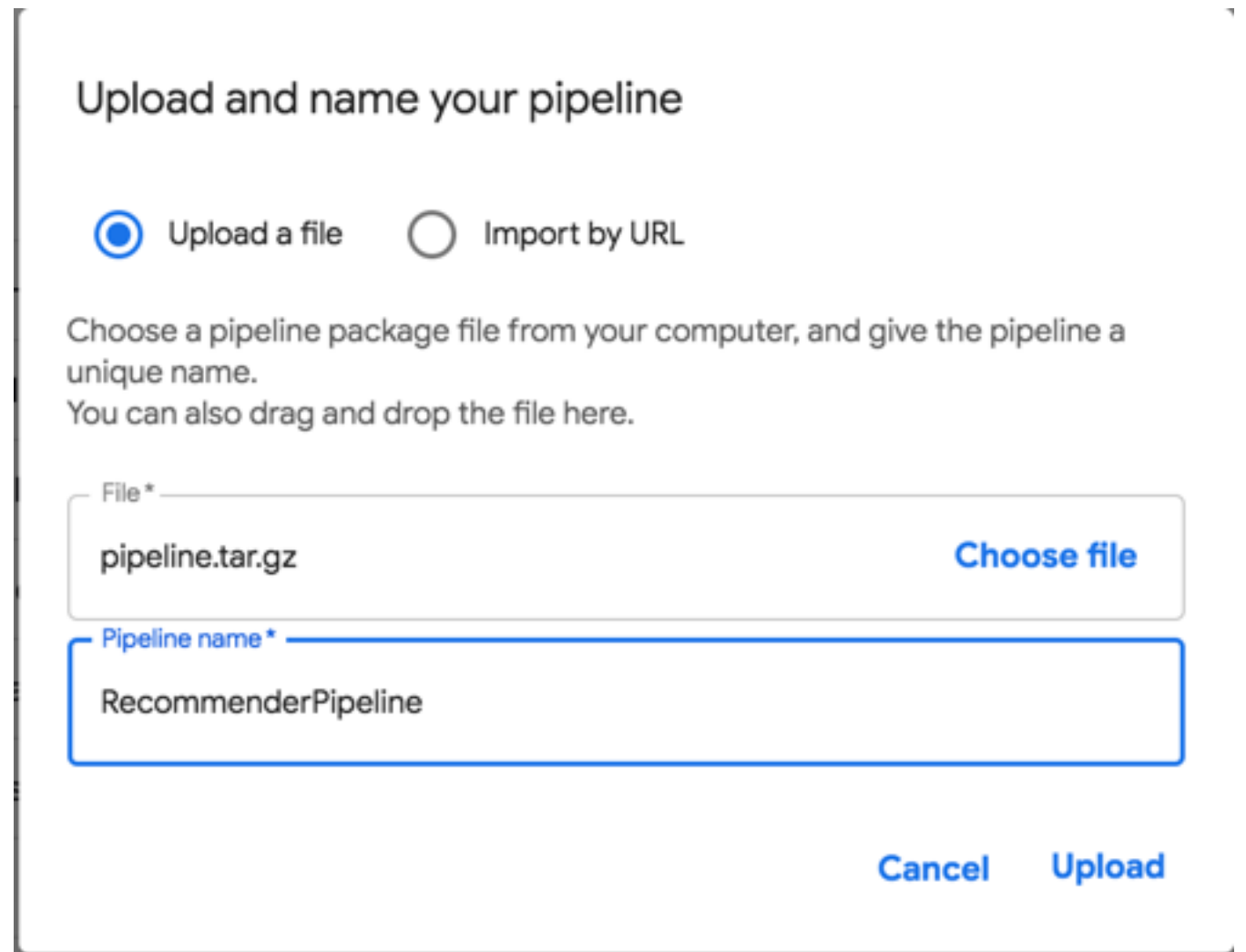
.

Argo workflow



Implementing Kubeflow pipelines

- Based on the Python library
- Fully supported in Jupiter
- Jupiter implementation can be exported to Python
- Python code creates pipeline.tar.gz file, which now can be uploaded to the Kubeflow pipelines UI



Upload and name your pipeline

☒ Upload a file ☐ Import by URL



Choose a pipeline package file from your computer, and give the pipeline a unique name.
You can also drag and drop the file here.


File*
pipeline.tar.gz [Choose file](#)


Pipeline name*
RecommenderPipeline


[Cancel](#) [Upload](#)

Pipeline execution




 Pipelines

 Experiments

 Archive

Experiments > mdupe22


←  pipeline1

[Clone run](#) [Terminate](#) [Archive](#)


Graph

Run output


Config

updatedata 

↓

trainmodel 

↓

publishmodel 

① Runtime execution graph. Only steps that are currently

×

recommender-model-update-wh7m6-4175455579

Artifacts

Input/Output

Volumes


Manifest

Logs

```
1014 2019-08-14 17:03:00.871156: I tensorflow/core/platform/s3/aws_logging.cc:54] Connection has been released
1015 2019-08-14 17:03:00.917338: I tensorflow/core/platform/s3/aws_logging.cc:54] Found secret key
1016 2019-08-14 17:03:00.917444: I tensorflow/core/platform/s3/aws_logging.cc:54] Connection has been released
1017 2019-08-14 17:03:00.921115: I tensorflow/core/platform/s3/aws_logging.cc:54] Found secret key
1018 2019-08-14 17:03:00.921245: I tensorflow/core/platform/s3/aws_logging.cc:54] Connection has been released
1019 2019-08-14 17:03:00.924738: I tensorflow/core/platform/s3/aws_logging.cc:54] Deleting file: /tmp/s3_files
1020 2019-08-14 17:03:00.924861: I tensorflow/core/platform/s3/aws_logging.cc:54] Found secret key
1021 2019-08-14 17:03:00.924971: I tensorflow/core/platform/s3/aws_logging.cc:54] Connection has been released
1022 2019-08-14 17:03:00.925818: I tensorflow/core/platform/s3/aws_logging.cc:54] Found secret key
1023 2019-08-14 17:03:00.925952: I tensorflow/core/platform/s3/aws_logging.cc:54] Connection has been released
1024 2019-08-14 17:03:01.011936: I tensorflow/core/platform/s3/aws_logging.cc:54] Found secret key
1025 2019-08-14 17:03:01.012172: I tensorflow/core/platform/s3/aws_logging.cc:54] Connection has been released
1026 2019-08-14 17:03:01.015311: I tensorflow/core/platform/s3/aws_logging.cc:54] Found secret key
1027 2019-08-14 17:03:01.015418: I tensorflow/core/platform/s3/aws_logging.cc:54] Connection has been released
1028 2019-08-14 17:03:01.033165: I tensorflow/core/platform/s3/aws_logging.cc:54] Found secret key
1029 2019-08-14 17:03:01.033787: I tensorflow/core/platform/s3/aws_logging.cc:54] Connection has been released
1030 2019-08-14 17:03:01.038648: I tensorflow/core/platform/s3/aws_logging.cc:54] Deleting file: /tmp/s3_files
1031 Done training!
1032 input 0 user:0
1033 input 1 product:0
1034 input [<tf.Tensor 'user:0' shape=(?, 1) dtype=float32>, <tf.Tensor 'product:0' shape=(?, 1) dtype=float32>]
1035 output 0 strided_slice:0
1036 output 1 strided_slice_1:0
1037 output Tensor("dense_2/BiasAdd:0", shape=(?, 1), dtype=float32)
1038 Exporting trained model to s3://models/recommender1/1/
```

Build commit: 812ca7f

Execution of pipelines on schedule

 Kubeflow

This run will be associated with the following experiment:

Experiment *

mdupdate2

Choose

Run Type

☐ One-off ☒ Recurring

Run trigger

Choose a method by which new runs will be triggered

Trigger type *

Periodic

Maximum concurrent runs *

1

☐ Has start date

☐ Has end date

Run every

30

Minutes

Run parameters

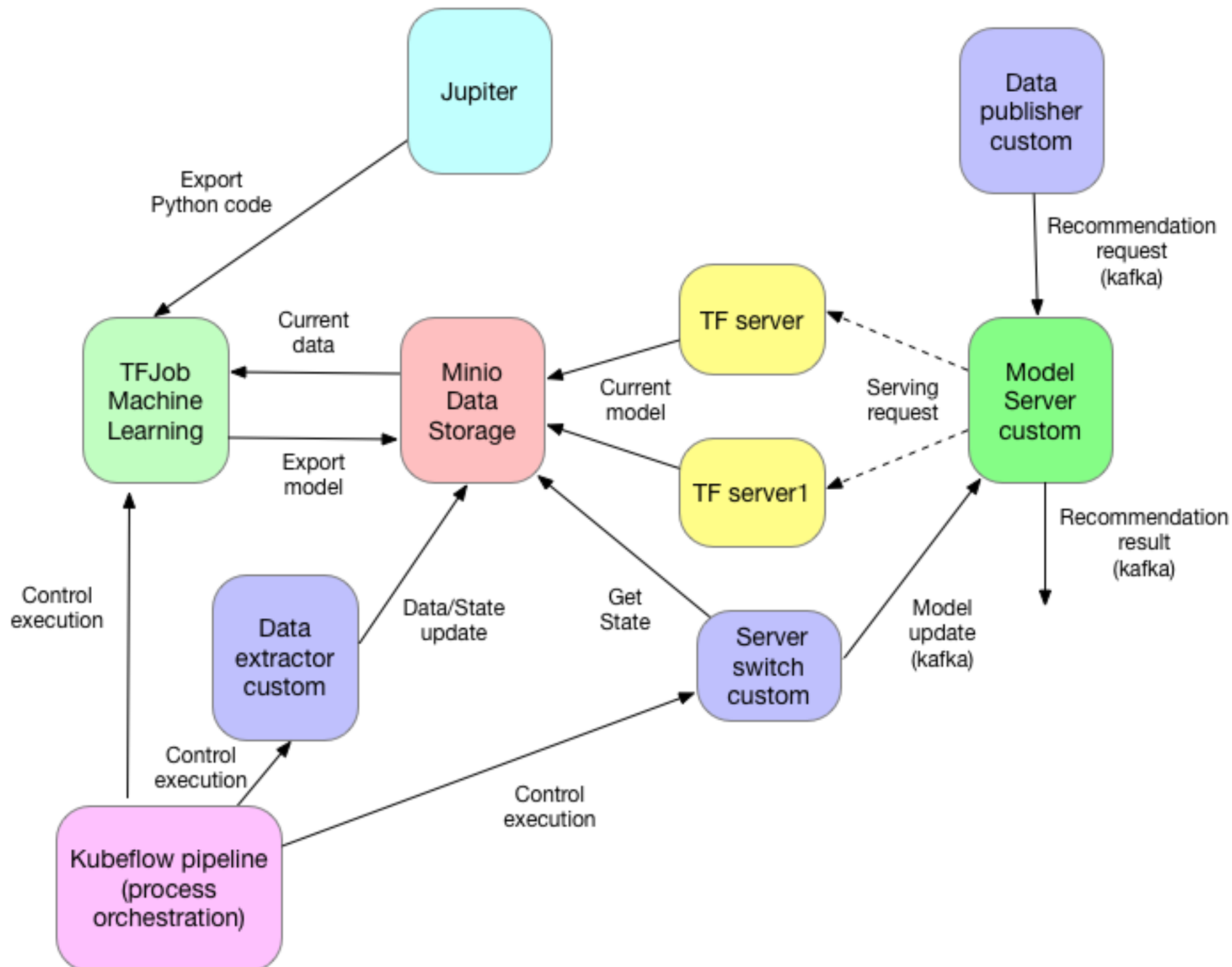
This pipeline has no parameters

Start

Cancel

Build commit: B12ca7f

Bringing it all together



Try this yourself

- Recommender implementation is heavily based on this [blog post](#)
★ <https://medium.com/datadriveninvestor/how-to-build-a-recommendation-system-for-purchase-data-step-by-step-d6d7a78800b6>
- We used purchases data provided [here](#) to test our implementation.
★ https://github.com/moorissa/medium/blob/master/items-recommender/data/trx_data.csv
- The full code, including Kubeflow setup information, is available [here](#)
★ <https://github.com/lightbend/kubeflow-recommender>

Thank you

Any questions?