



Accelerating Machine Learning DevOps with Kubeflow

Derek Ferguson

Head of Engineering, JP Morgan Chase Commercial Bank

Who am I?

- Head of Engineering, JP Morgan Chase Commercial Bank
- derek.ferguson@jpmorgan.com
- Blog at <http://derekmferguson.wixsite.com/ml4nonmath>
- Previously a tech evangelist in the Microsoft space
 - Editor-in-Chief of the .NET Developer's Journal
 - Author of Mobile.NET and Broadband Internet Access for Dummies
- Life-long Chicago native
 - Graduated DePaul & started commercial DSL with InterAccess

City Scholars

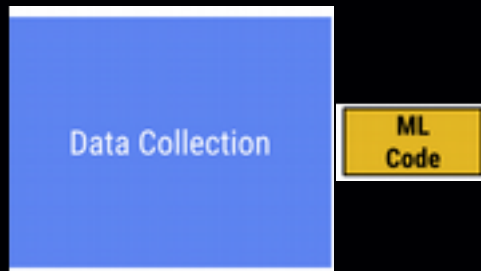
- Partnership between Chase, UIUC and City of Chicago
- Students work half-time in the Spring and full-time in Summer
- Our remits this year:
 - Build some great machine learning models for the Business
 - Improve the state of machine learning DevOps for the Industry

The Problem

The Problem



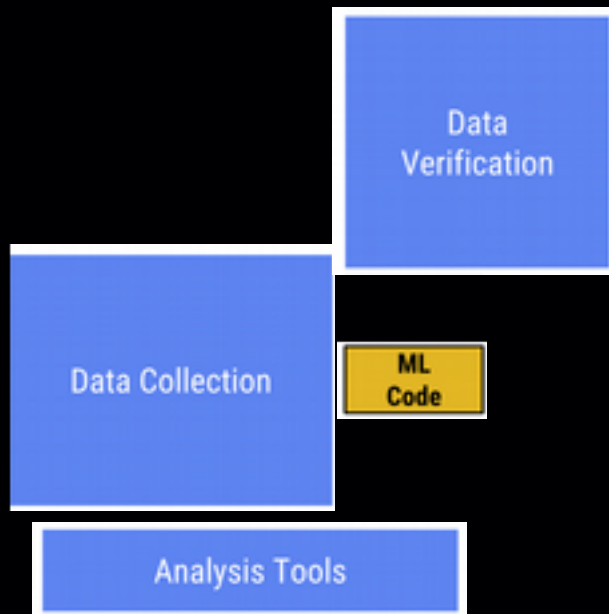
The Problem



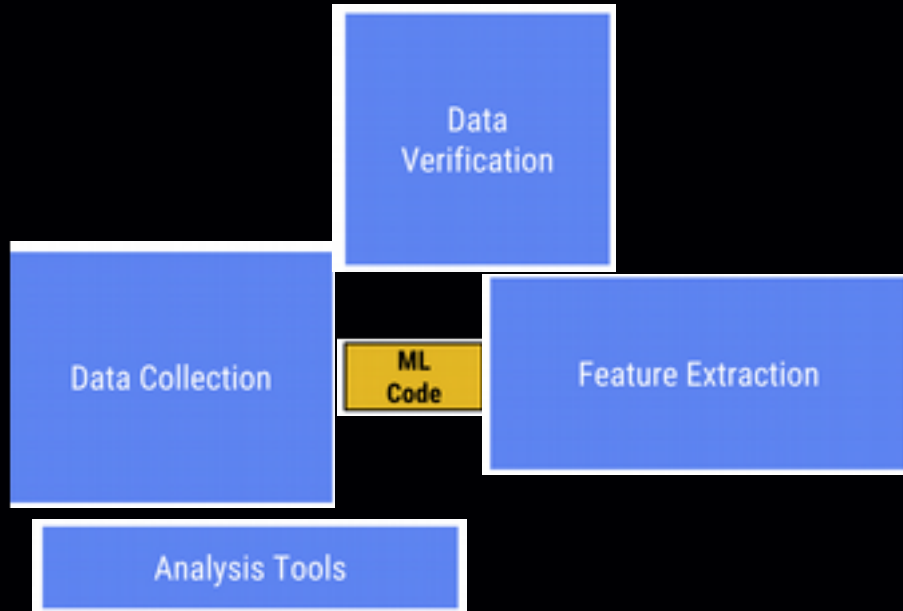
The Problem



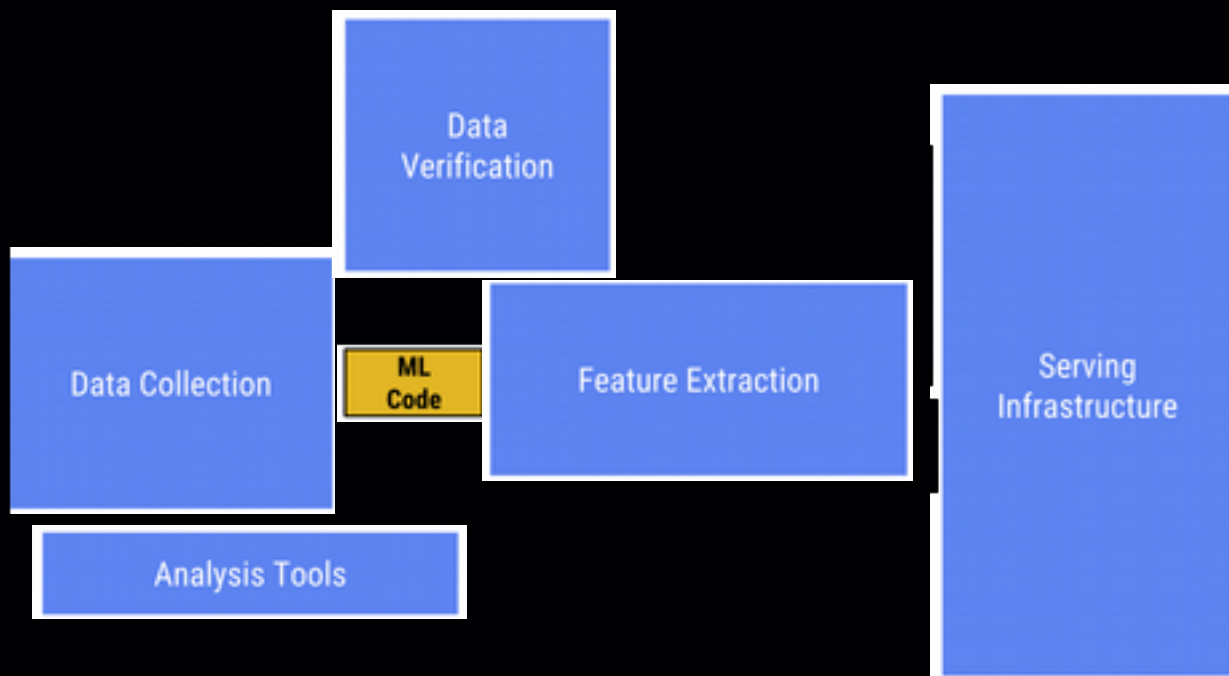
The Problem



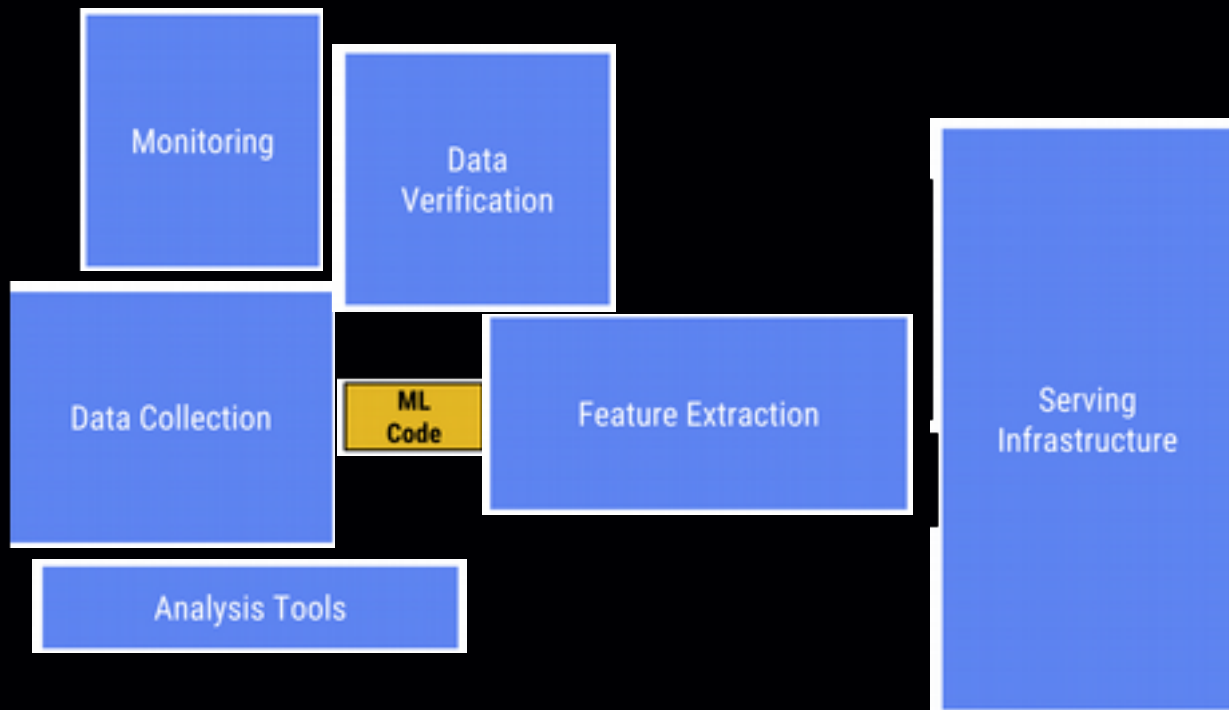
The Problem



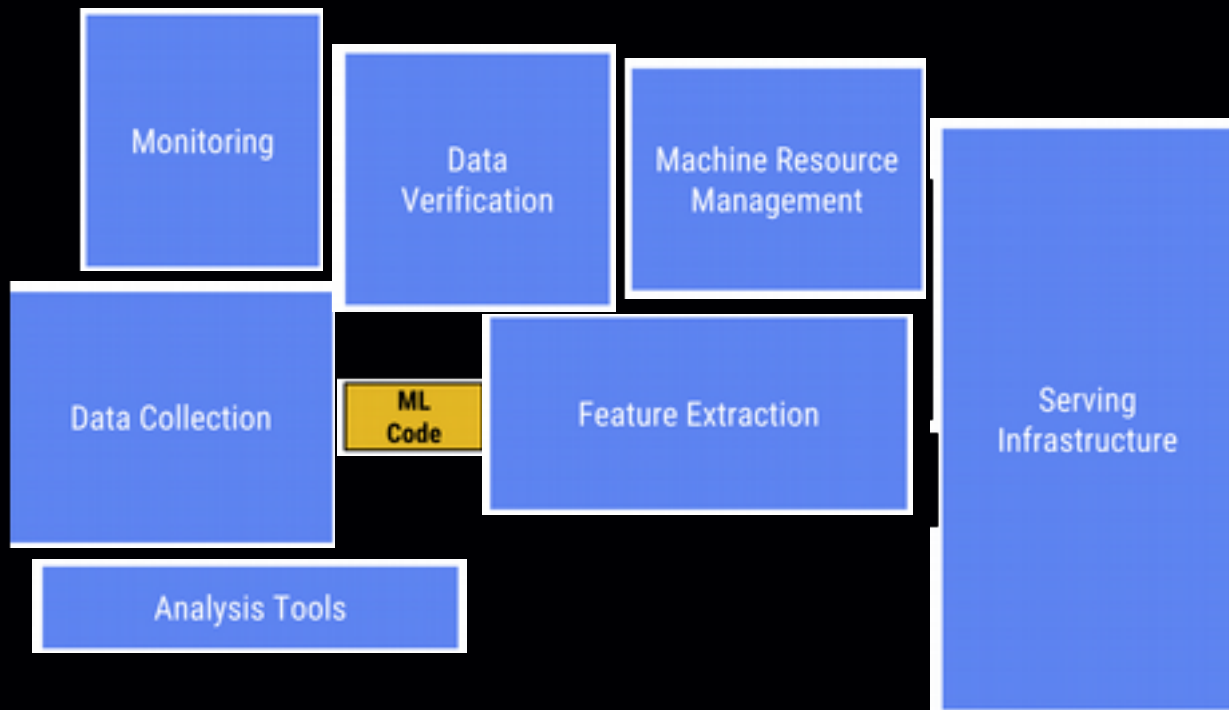
The Problem



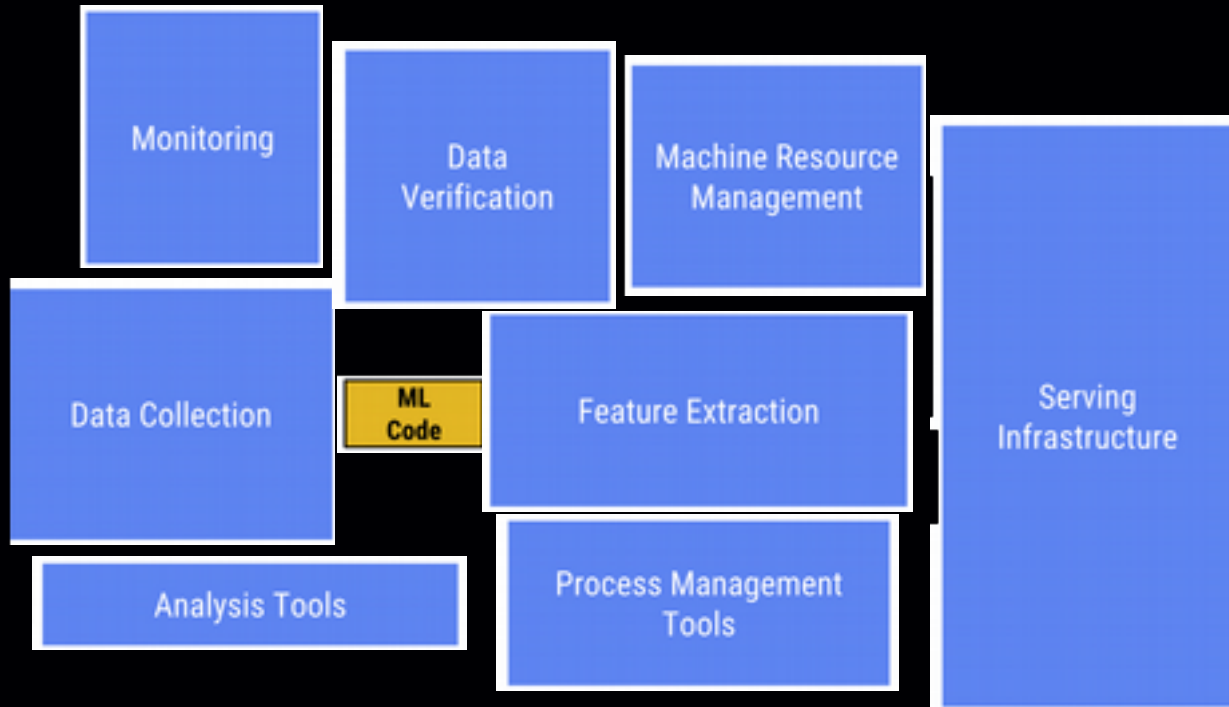
The Problem



The Problem



The Problem



Sneak Peek: the Solution

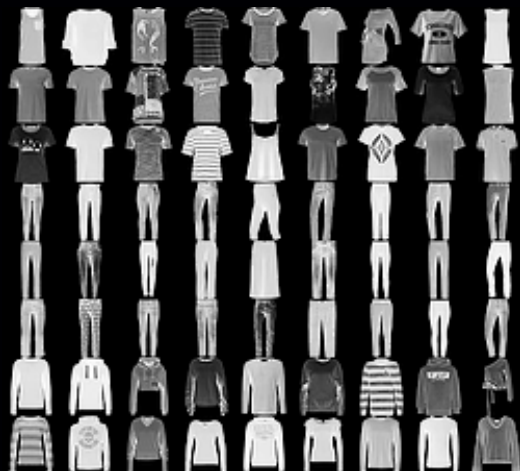


Agenda

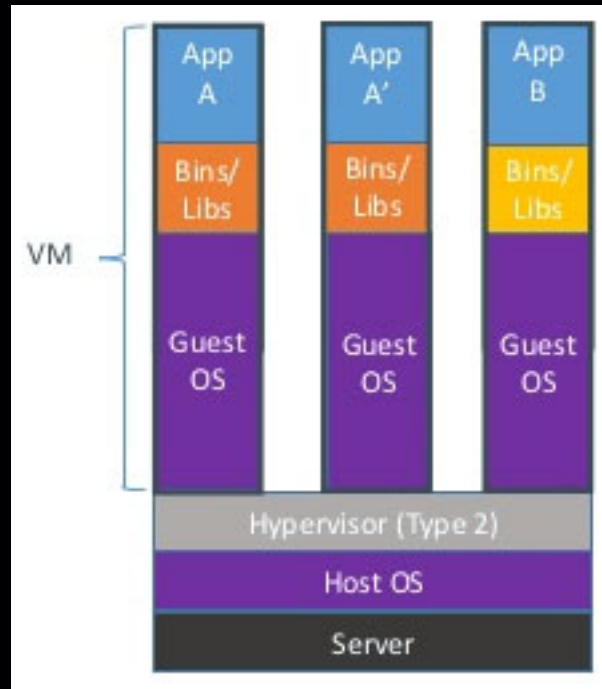
- Brief overview of the Fashion MNIST sample data
- Overview of Kubernetes
 - Starting with a brief overview of Docker
- All about Kubeflow
- Q&A

Fashion MNIST

- Sample data set with 10 different kinds of clothing
- 60k training images and labels
- 10k test images and labels
- Images start as 784-position arrays containing numbers 0 to 255



The Old Days (i.e. 5 years ago)



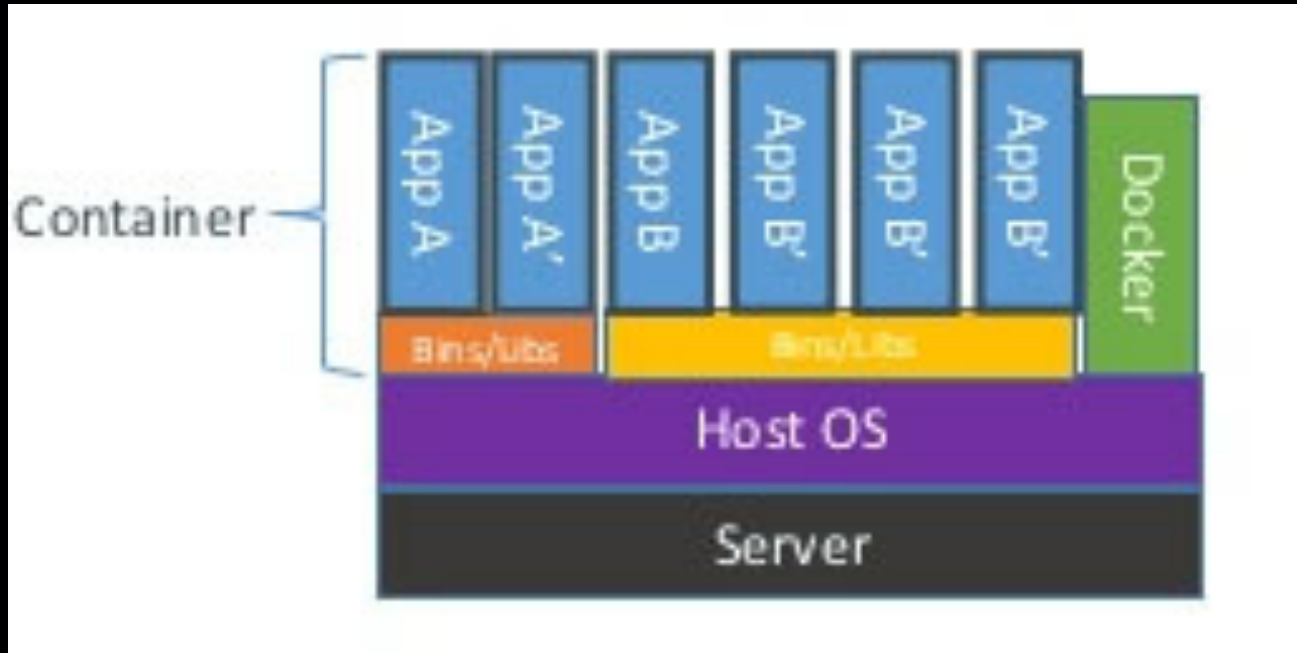
Today, IT Ops has a lot to do...



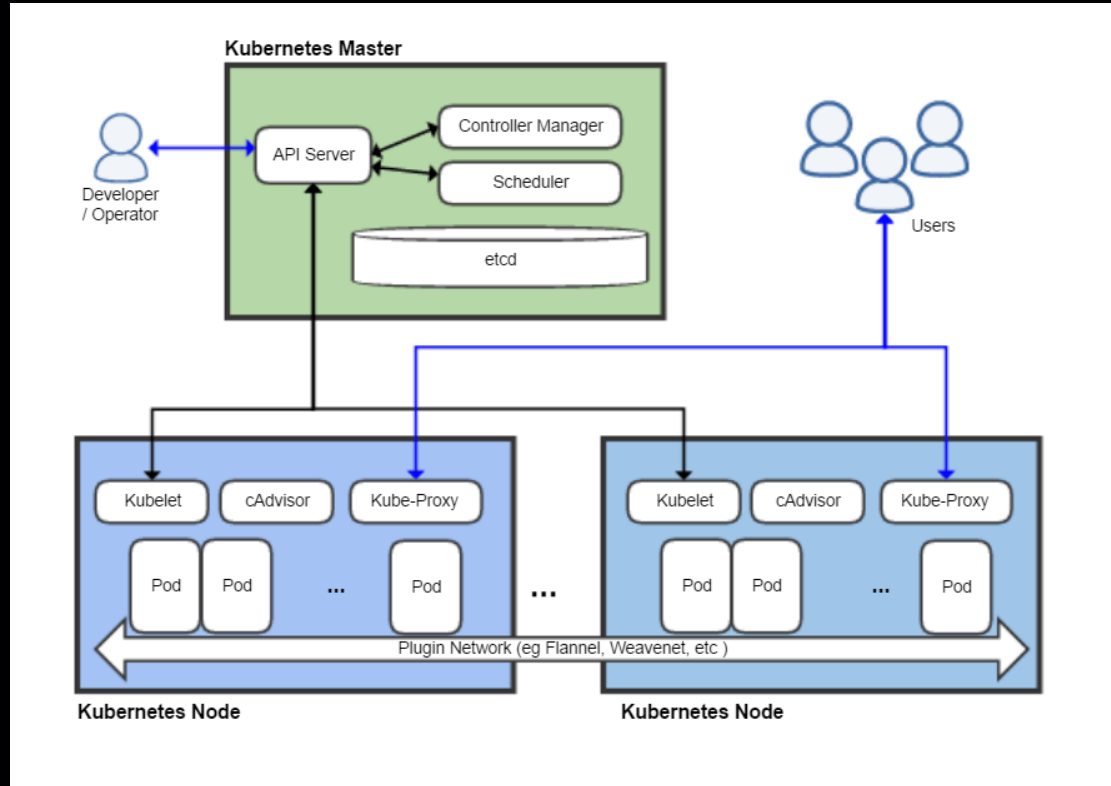
VMs take a long time to setup...



Containers cheaper and faster - but still slow

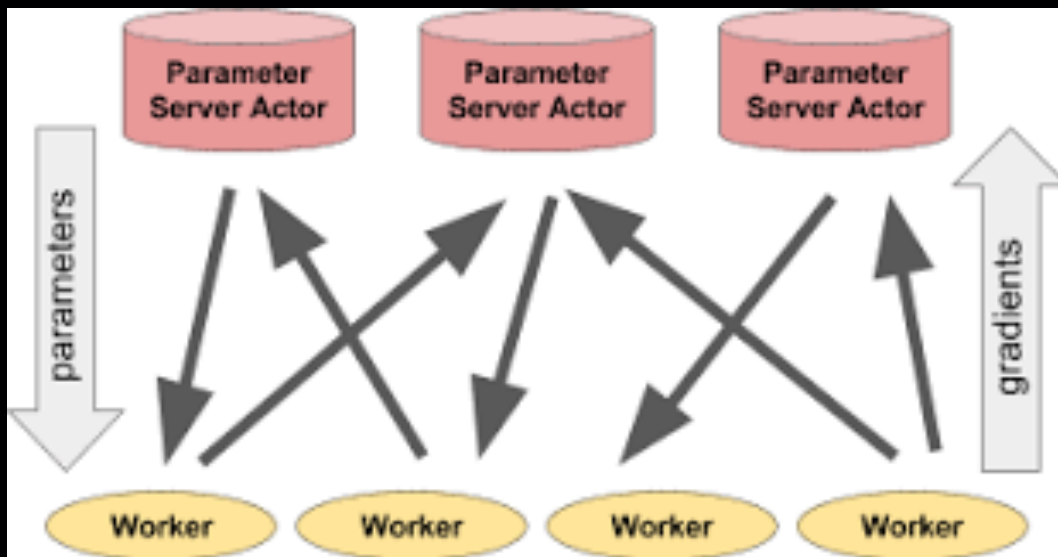


Behold, Kubernetes!



Challenge 1: Network Topology

- Framework needs to know location of Master
- Master needs to know locations of workers
- Workers need to know locations of parameter servers



Challenge 2: It doesn't turn off




Challenge 3: Everything starts open

- Jupiter notebooks open by default
- Model storage open by default
- Network calls unencrypted by default
- Network addresses and ports unauthenticated by default

What is Kubeflow?

- Kubeflow is an end-to-end lifecycle orchestration tool for machine learning
- Vision would be to let data scientists get models from initial training into Production with minimal human intervention
- Enabling technology is Kubernetes
 - There is **no** mandatory tie to Tensorflow

Secure Jupyter Notebooks



A screenshot of a web browser window showing the JupyterHub login page. The browser's address bar displays the URL `nebula.sdsc.edu:8000/hub/login`. The page features the JupyterHub logo at the top. Below the logo, there is a login form with two input fields: 'Username:' and 'Password:'. A 'Log in' button is positioned below these fields.

nebula.sdsc.edu:8000/hub/login

jupyterhub

Username:

Password:

Log in

TF Job

```
apiVersion: "kubeflow.org/v1alpha2"
kind: "TFJob"
metadata:
  name: {{workflow.parameters.job-name}}
  namespace: {{workflow.parameters.namespace}}
spec:
  tfReplicaSpecs:
    Master:
      replicas: 1
      template:
        spec:
          serviceAccountName: tf-job-operator
          containers:
            - image: {{workflow.parameters.tf-model-image}}
              name: tensorflow
              imagePullPolicy: Always
              env:
                - name: TF_MODEL_DIR
                  value: {{inputs.parameters.s3-model-url}}
                - name: TF_EXPORT_DIR
                  value: {{workflow.parameters.model-name}}
                - name: TF_TRAIN_STEPS
```

Fairing Attributes

```
In [*]: import fairing
from fairing import builders
from fairing.training import kubeflow

fairing.config.set_builder(builders.AppendBuilder(
    repository='gcr.io/mrick-gcp',
    notebook_file='/home/jovyan/work/blog-post.ipynb',
    base_image='gcr.io/kubeflow-images-public/tensorflow-1.10.1-notebook-cpu:v0.4.0'))

@kubeflow.DistributedTraining(worker_count=1, ps_count=1)
class MyModel(object):
    def train(self):
        print(get_enviro())

model = MyModel()
model.train()
```

```
Running...
Uploading gcr.io/mrick-gcp/fairing-job:cf0a8e83f0a77cc5a142d5eelf2ad4f3631d4d5d997ef6bad486f57b07a8e433
Pushed image gcr.io/mrick-gcp/fairing-job:cf0a8e83f0a77cc5a142d5eelf2ad4f3631d4d5d997ef6bad486f57b07a8e433
Training(s) launched.
Waiting for job to start...
Waiting for job to start...
```

```
b'2019-01-17 23:40:26.833343: I tensorflow/core/platform/cpu_feature_guard.cc:141] Your CPU supports instructions tha
t this TensorFlow binary was not compiled to use: AVX2 FMA'
b'fairing-job-14960ba5-worker-0'
```

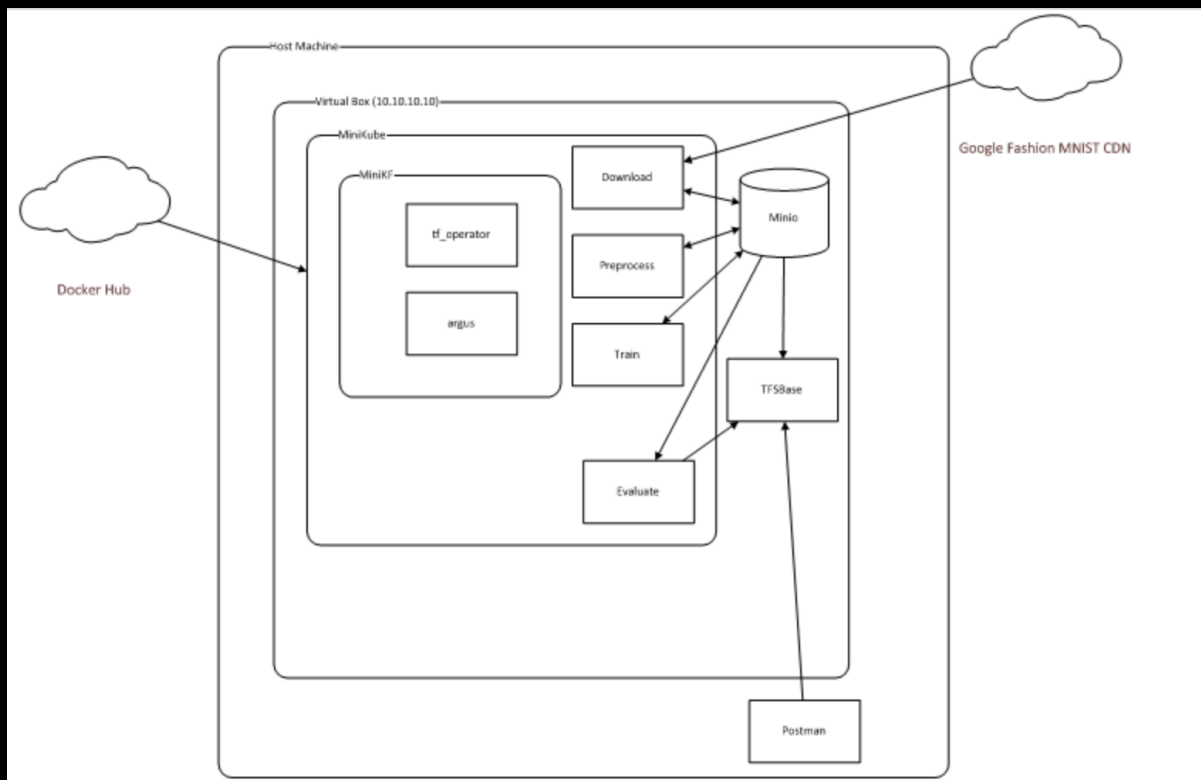
MiniKF

- A fast and easy way to deploy Kubeflow onto your desktop
- Full-fledged, production-ready Kubeflow instance that deploys in minutes
- Few clicks and you're up for experimentation, including running full Kubeflow pipelines
- To train at scale, move to a public cloud deployment with one click and no recoding
- <http://kubeflow.org/docs/started/getting-started-minikf>
- Discussion on #minikf Slack channel. Ask questions, request features and get Support.
- Contributed by Arrikto - core contributors to Kubeflow

The Demo

- Download: pull 70k images and labels from Internet
- Pre-process: turn arrays into 28x28 matrixes with values 0 to 1
- Train: pass 60k images and labels through a neural network
- Evaluate: test our model against 10k images and their labels
- Serve: push model up to Tensorflow Serving

Demo Architecture



Demo: Pipelines

Additional Points of Note

- Securing the prediction layer (Tensorflow Serving, for example)
- Securing S3
- Common K8S On Premises Issues
 - Container lifespan limitations
 - K8S action limitations (e.g. creating and configuring namespaces)
 - Procurement challenges
 - Live installation directly from the Internet
 - Insecure Docker images in distribution

In Conclusion

- All source is at <https://github.com/JavaDerek/FashionMnistKF>
- Please contact me with any comments, questions or concerns...
 - derek.ferguson@jpmorgan.com

Questions?