



OLF 12/02/2022

Kubeflow:

Bring your ML project into Production

Liang Yan

Sr. Software Engineer, DigitalOcean



Liang Yan

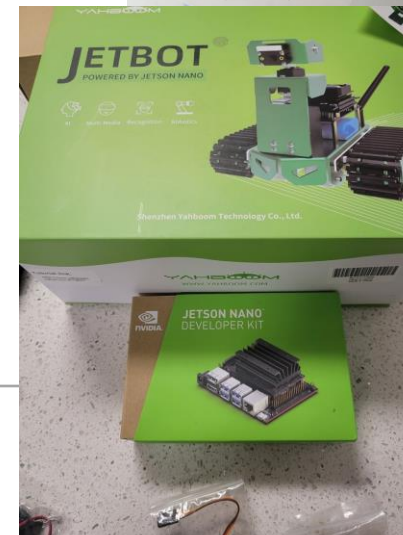
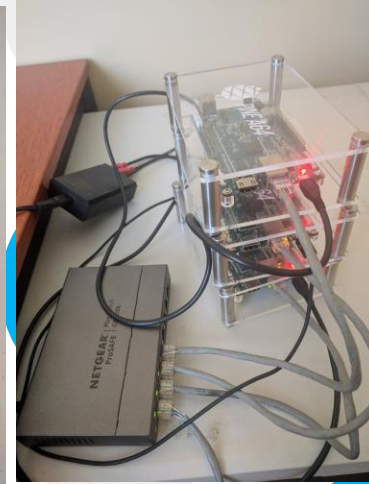
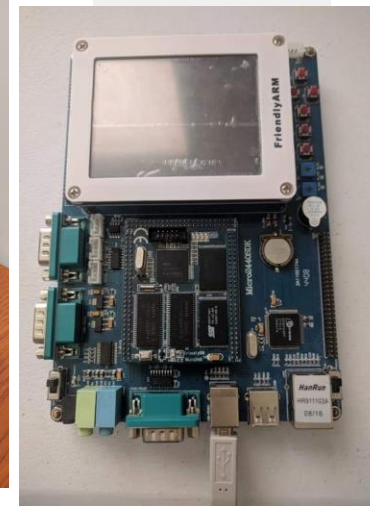
Software Engineer
OpenSUSE Member
Open-Source Advocator(KYOSS)

Louisville, KY

Interests:

- Virtualization
- SysML/Distributed ML
- Infrastructure build and optimization
- ARM64 board Enthusiast
- DevOps

<https://xryan.net>





Outline



Prologue



Kubeflow



Kubeflow Components



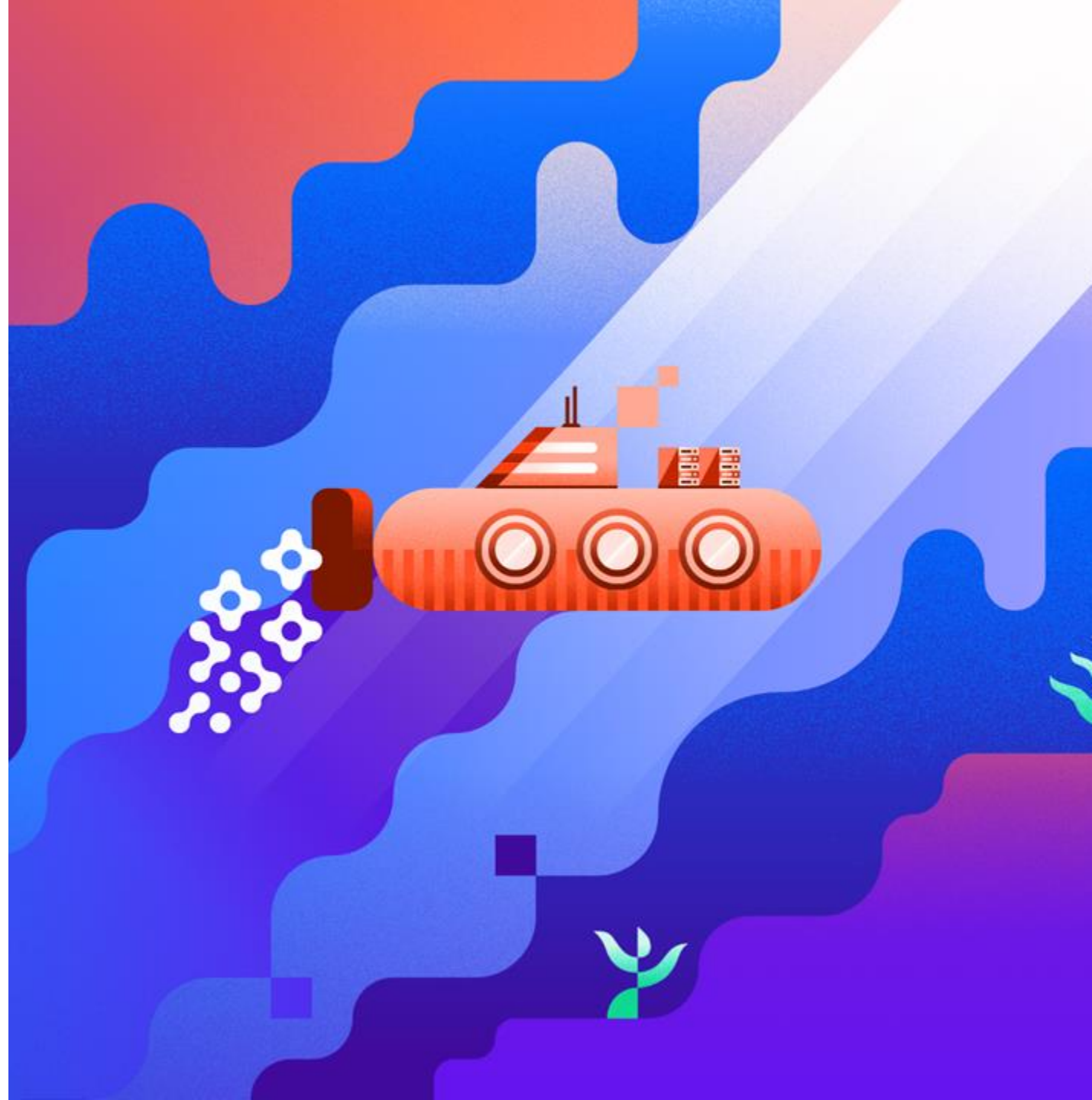
Demo



Beyond



Q&A





Prologue



Kubeflow





Prologue

Flight Delay Predictor:

<https://github.com/xrlyan/Flight-Delay-Prediction-Based-on-Neural-Networks>

Input:

- Flight no
- Flight date

Output:

- Delay possibility

Features:

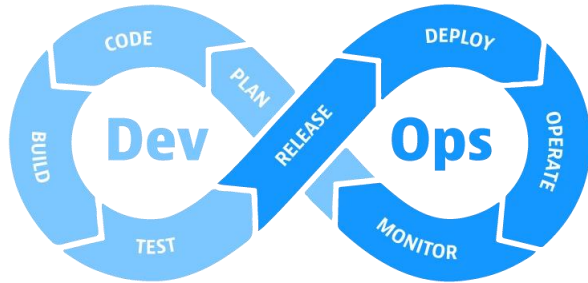
- Depart/Arrive airport
- Depart/Arrive time
- Depart/Arrive city weather
- Flight model
- Flight History delay rate



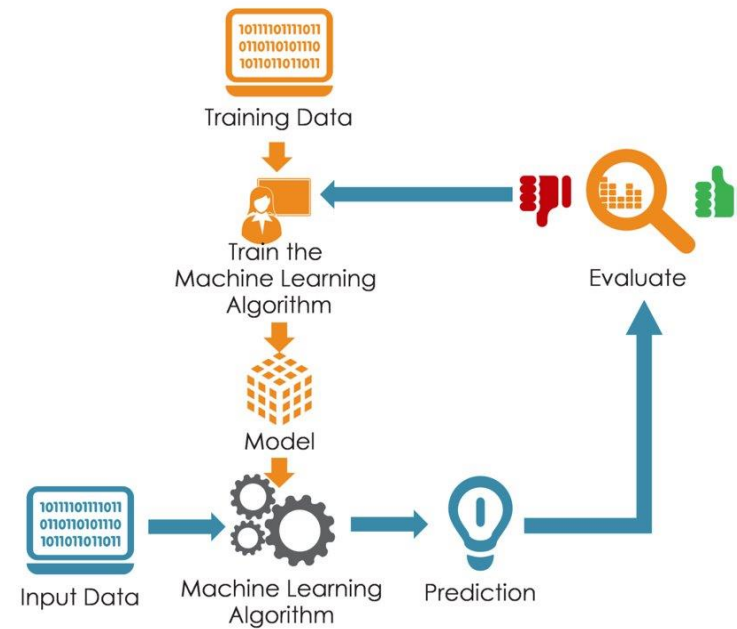


Prologue

As a Software Engineer:

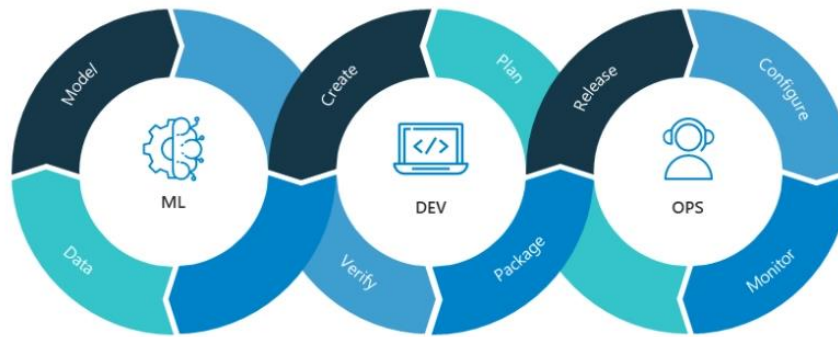
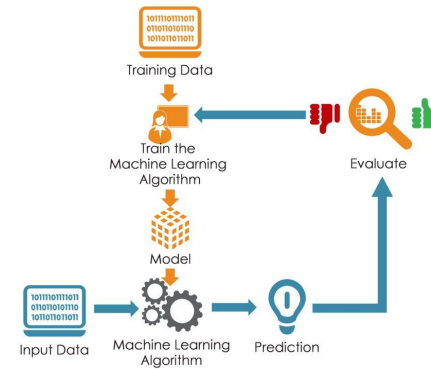
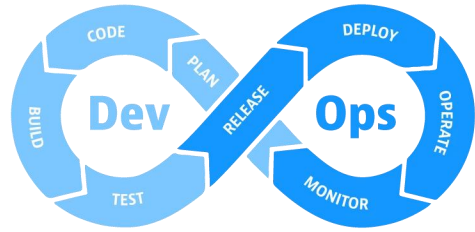


As a Data Scientist/Engineer:





Prologue

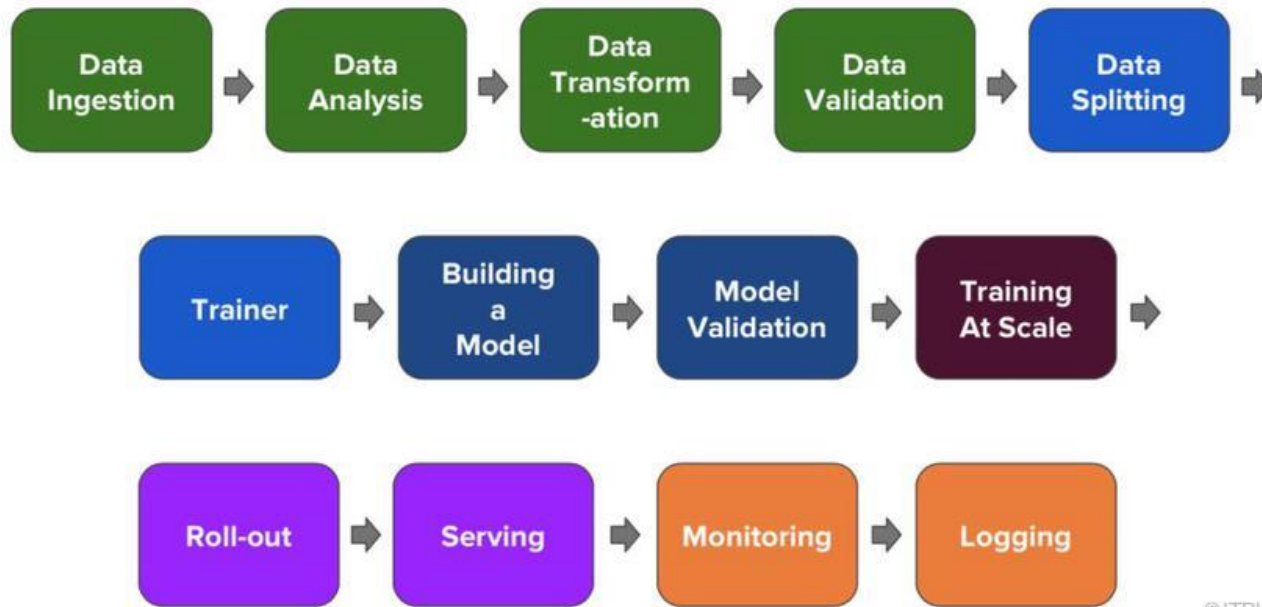


Pic: <https://blogs.nvidia.com/blog/2020/09/03/what-is-mlops/>



Prologue


Eventually, it becomes:



@ITPUB



Kubeflow-Central-dashboard

 **Kubeflow**

Home

Notebooks

Tensorboards

Models

Volumes

Experiments (AutoML)

Experiments (KFP)

Pipelines

Runs

Recurring Runs


Artifacts


Privacy • Usage Reporting
build version dev_local


kubeflow-user (Owner) ▾


DashboardActivity

Quick shortcuts


 **Upload a pipeline**
Pipelines


 **View all pipeline runs**
Pipelines

 **Create a new Notebook server**
Notebook Servers


 **View Katib Experiments**
Katib


Recent Notebooks


 **kale.log**
Accessed 10/12/2021, 2:06:43 PM


 **lost+found**
Accessed 10/12/2021, 2:06:00 PM


Recent Pipelines

 **open-vaccine-model**
Created 5/6/2021, 12:32:25 PM


 **[Tutorial] DSL - Control structures**
Created 5/6/2021, 1:42:51 AM


 **[Tutorial] Data passing in python components**
Created 5/6/2021, 1:42:49 AM


 **[Demo] TFX - Taxi tip prediction model trainer**
Created 5/6/2021, 1:42:48 AM


 **[Demo] XGBoost - Iterative model training**


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 



Kubeflow

Kubeflow is:

K8S + TensorFlow
Application Toolkit
Orchestration
Cloud Native
DevOps/MLOps



Kubeflow is not:

K8S + TensorFlow
Application
Scheduler
Machine Learning Algorithm
Machine Learning Framework

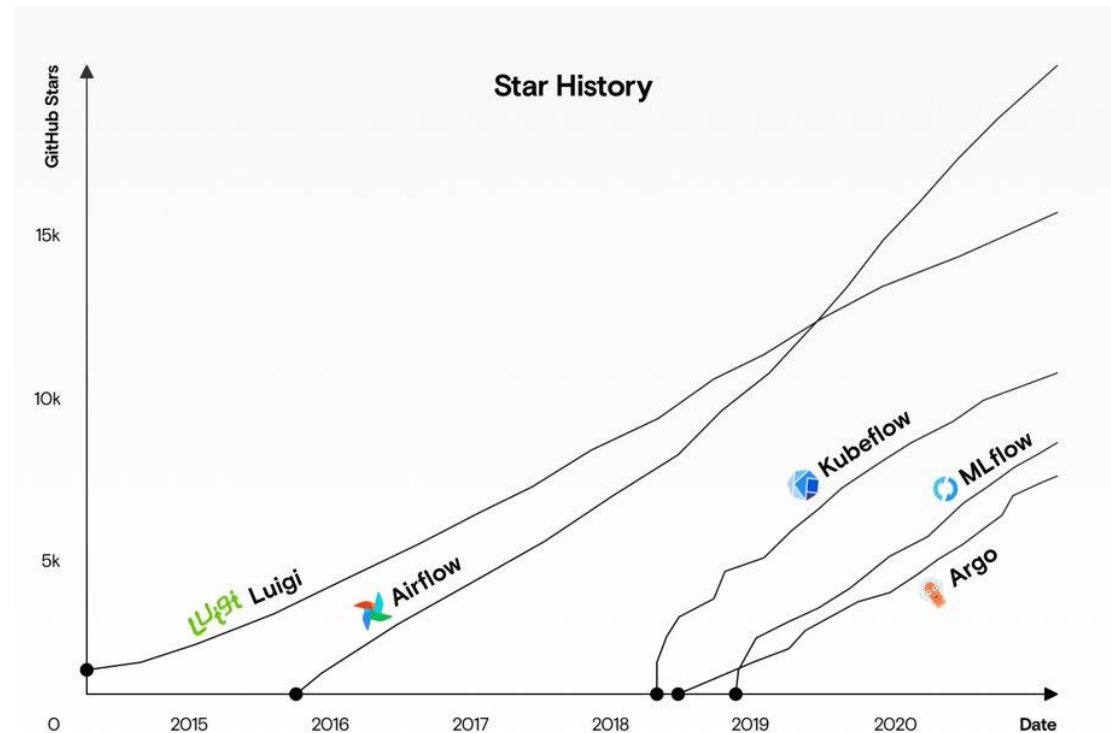


Kubeflow

Machine Learning Orchestration Platform:

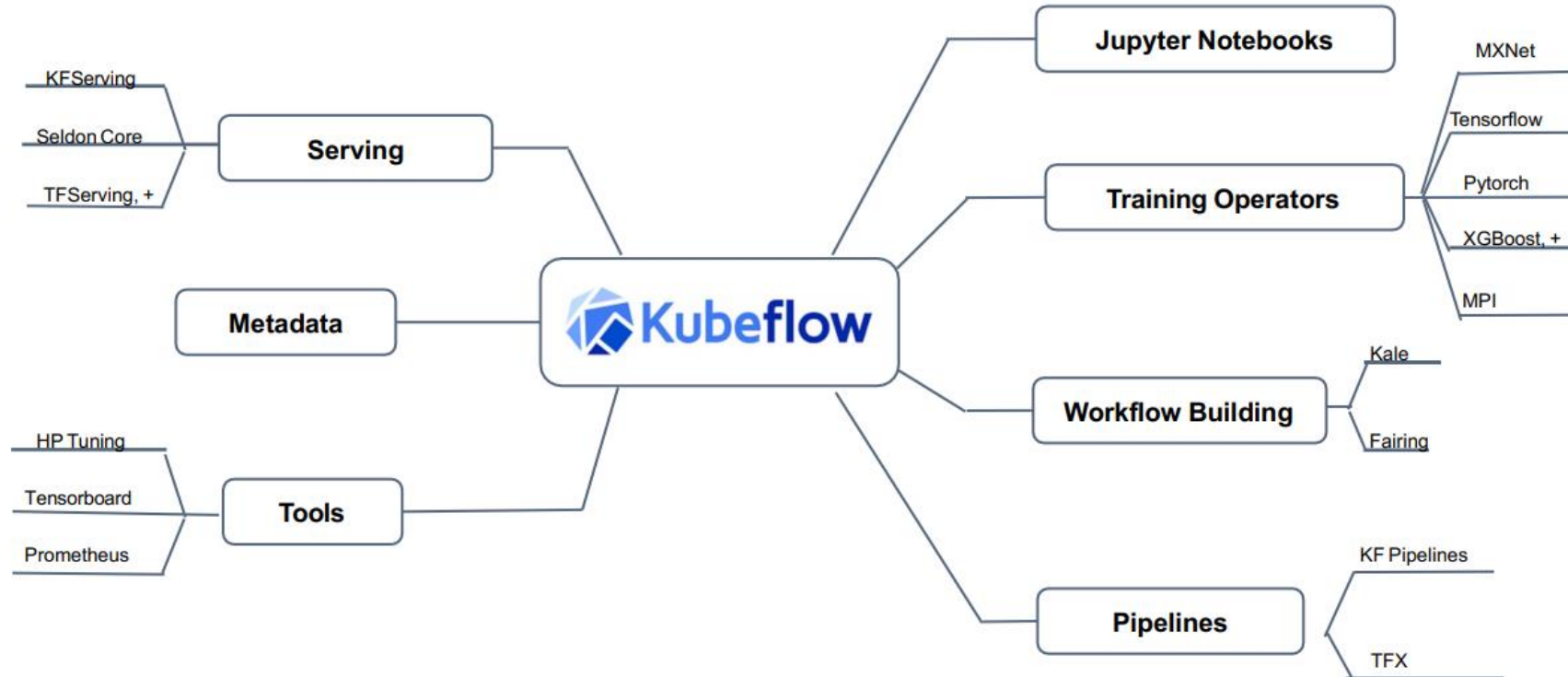
1. Orchestrate pipeline
2. Orchestrate ML task

Great mind think alike!



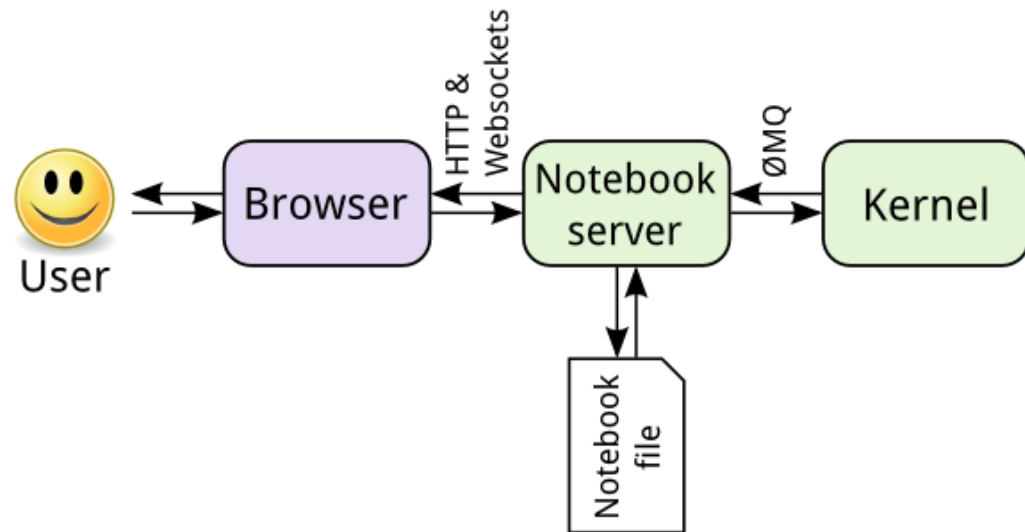


Kubeflow components





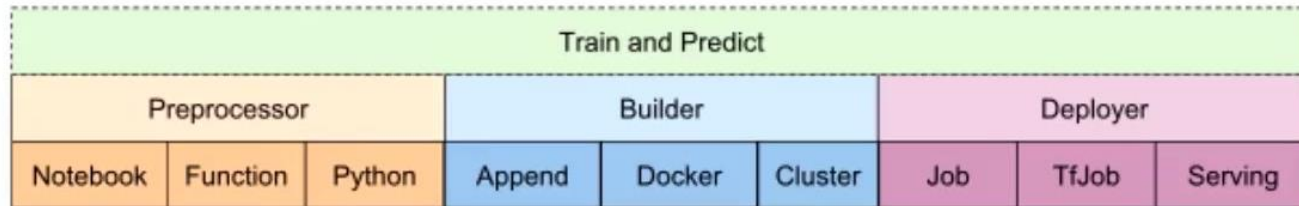
Jupyter-notebook





Fairing

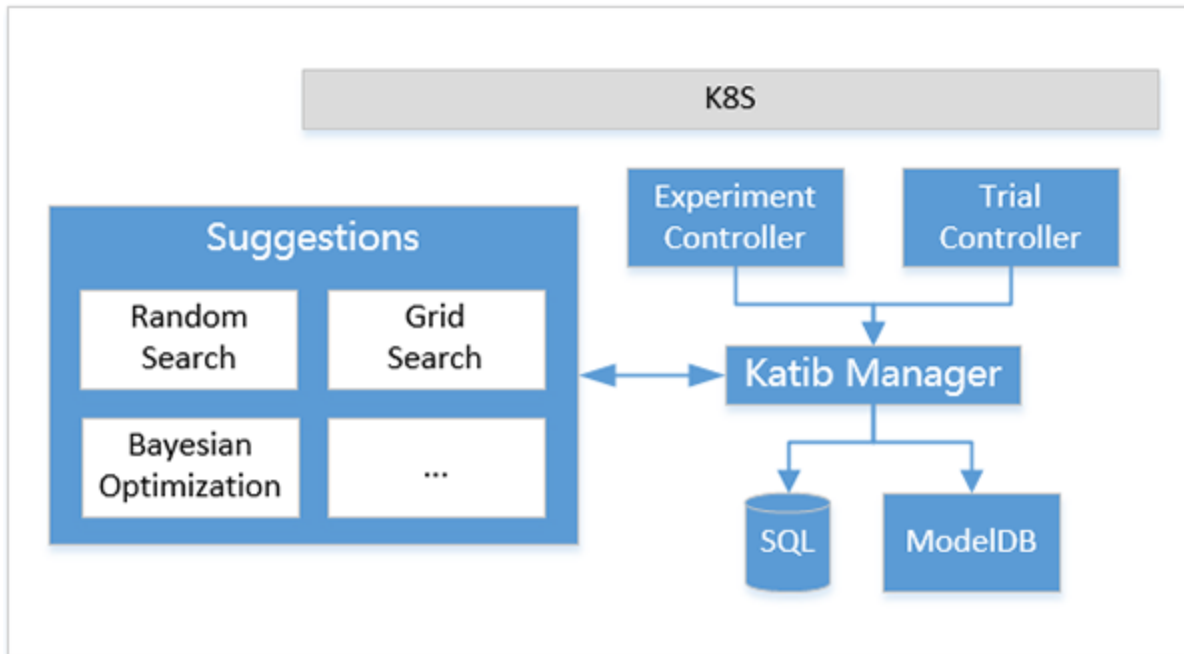
Architecture





Katib

An implementation for AutoML: tune hyperparameter automatically



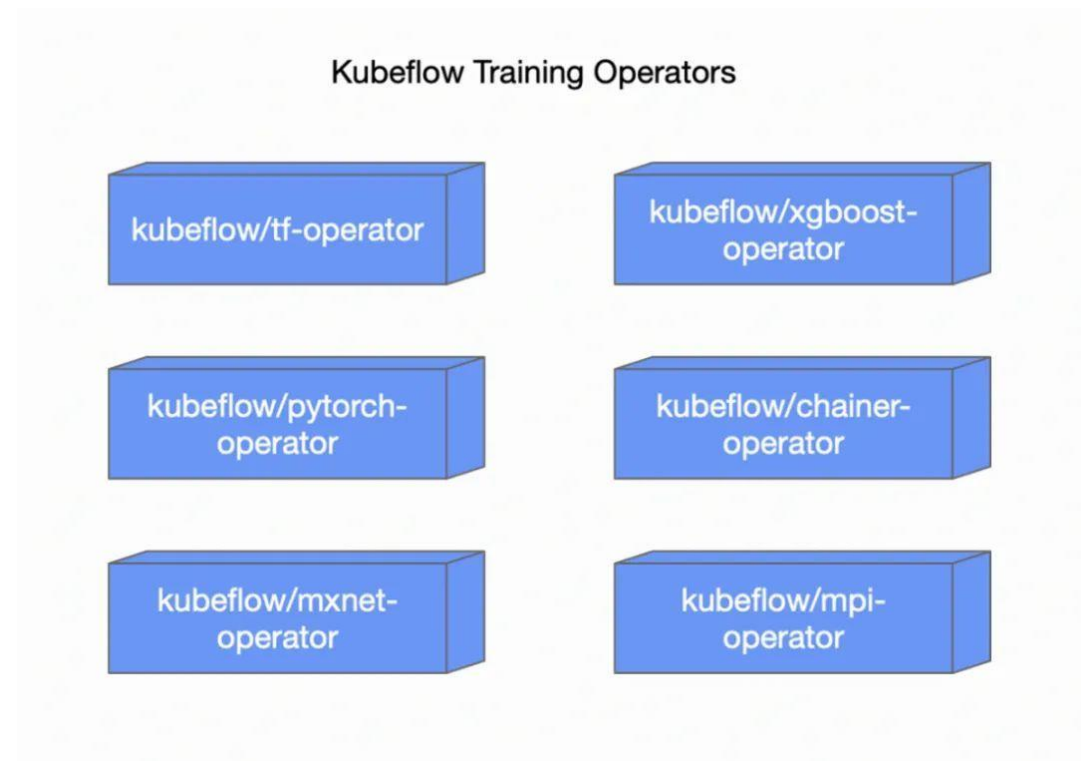
Three CRDs:
experiment
suggestion
trial

The experiment creates
multiple trials based on different
suggestion algorithms.



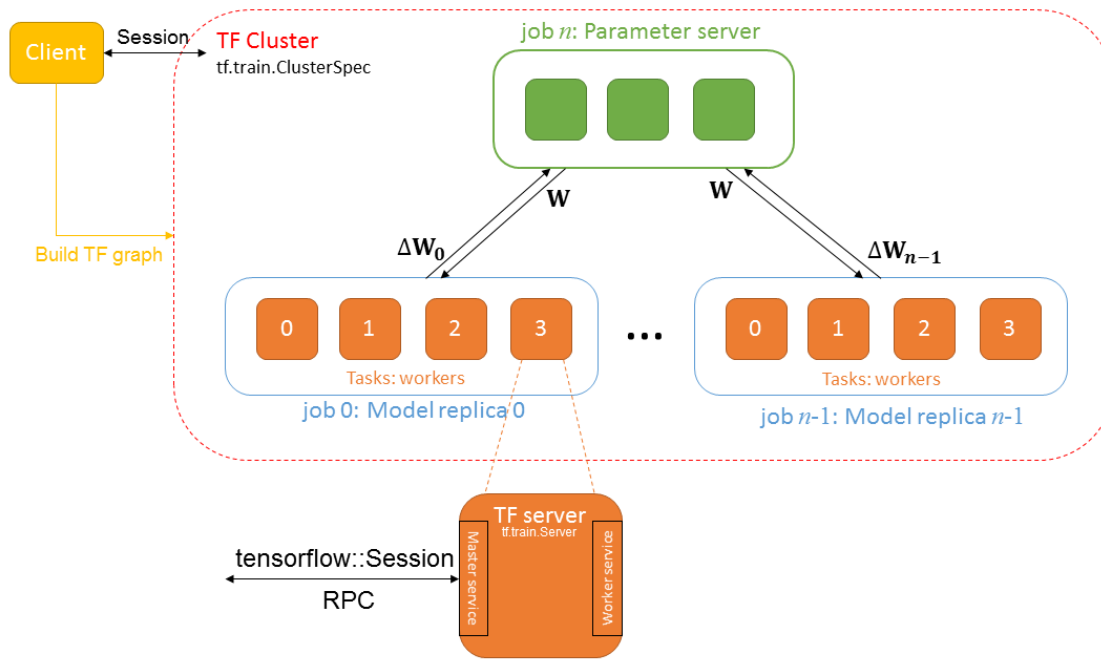
Training Operator

Operator = Controller + CRD + Webhook
Tool: kubernetes





TF-Operator



Chief coordinate training job

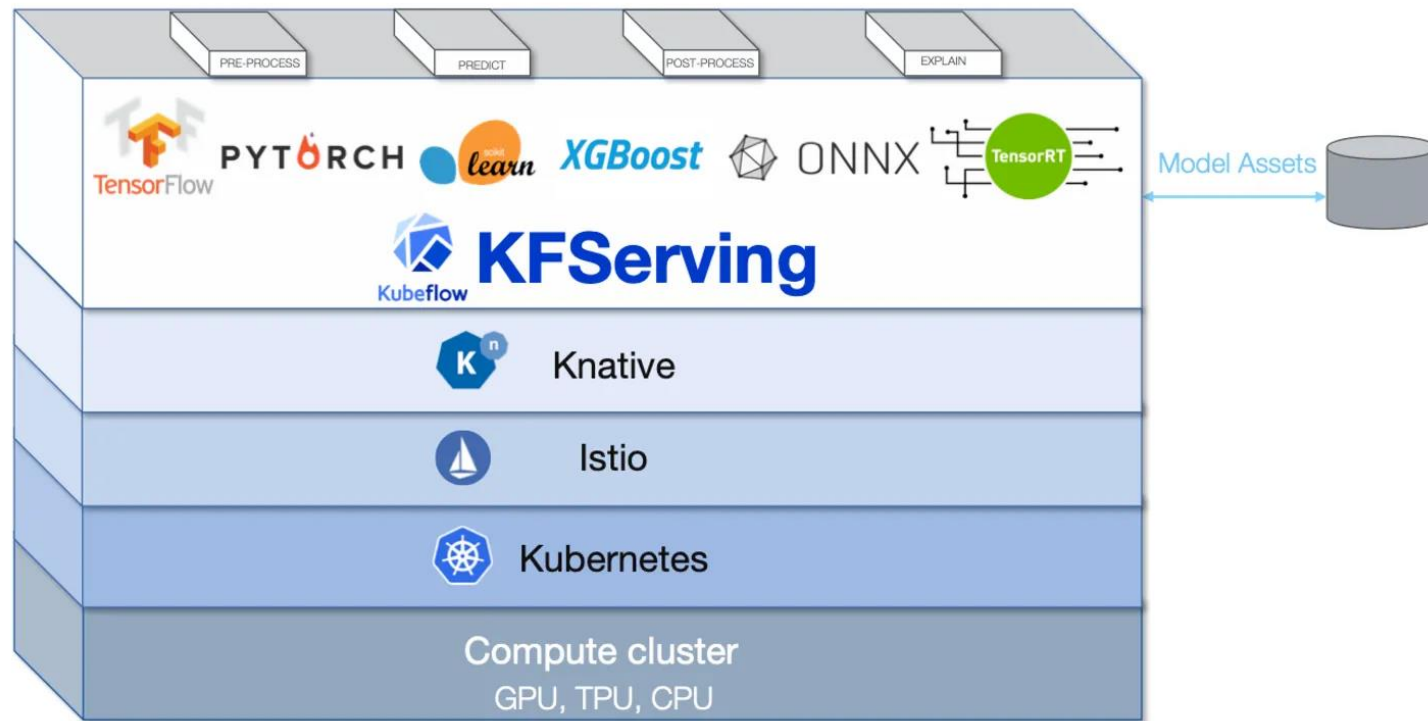
PS server, parameter

Worker

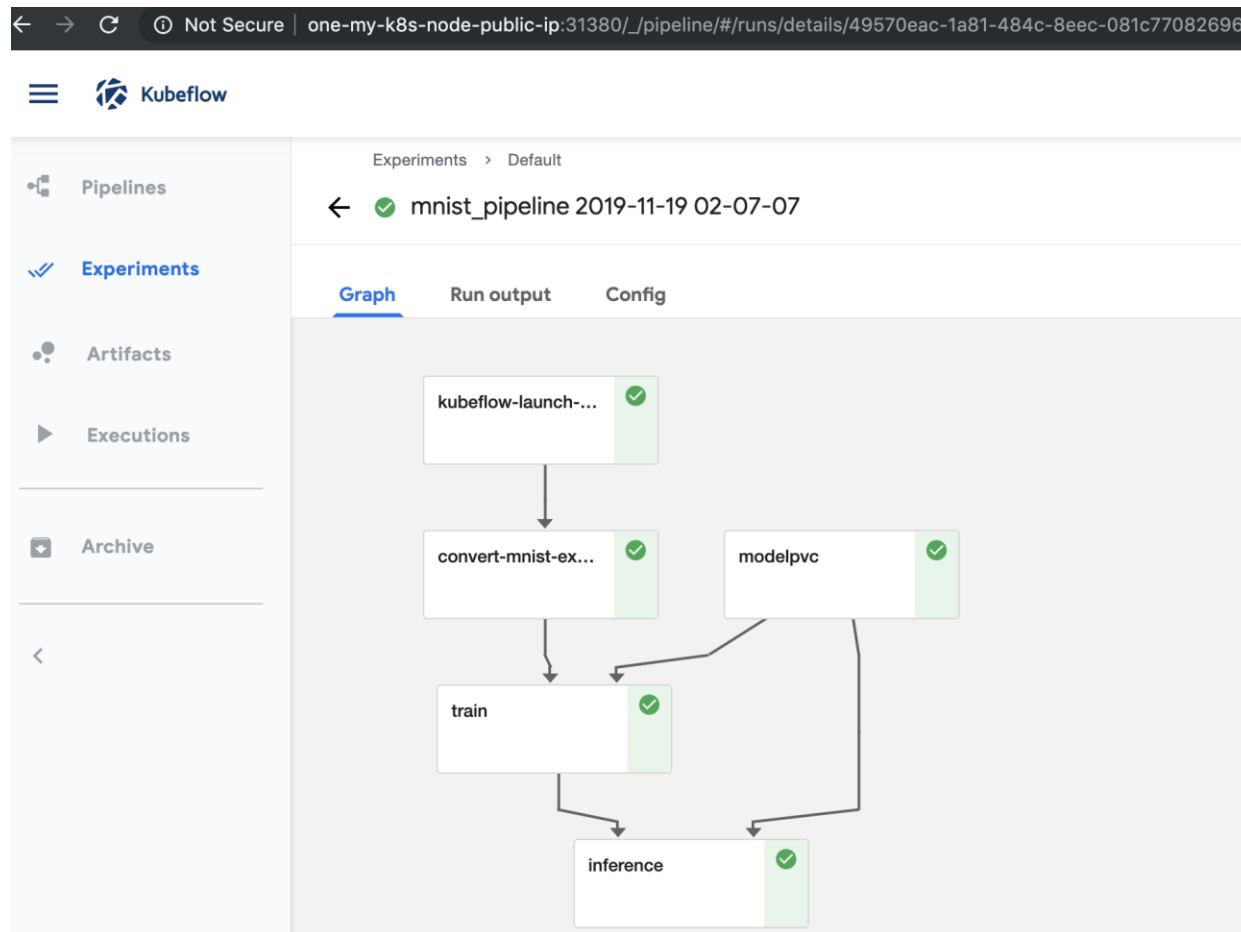
Evaluator

KFServing

The last mile!

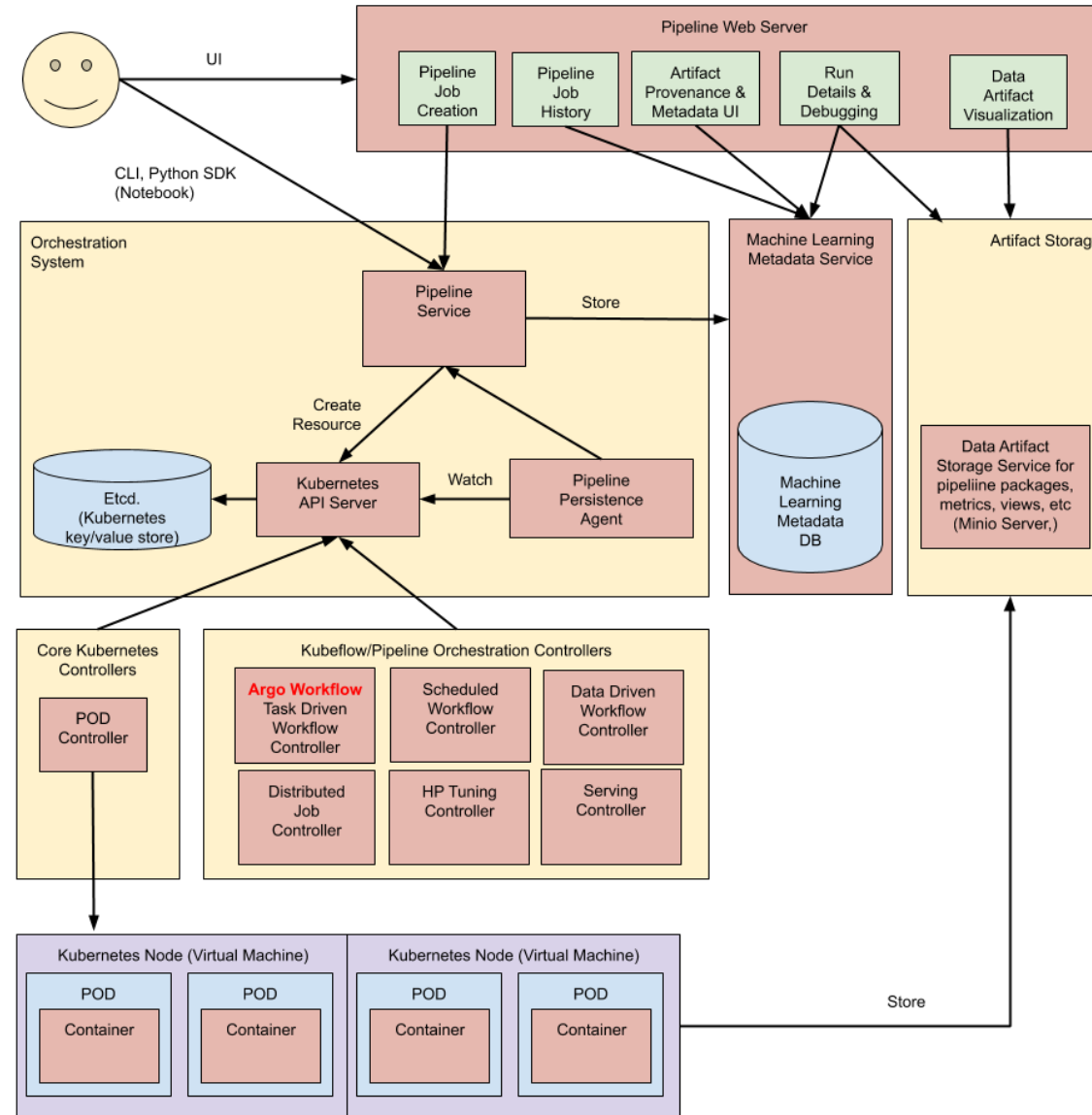


KFP DAG





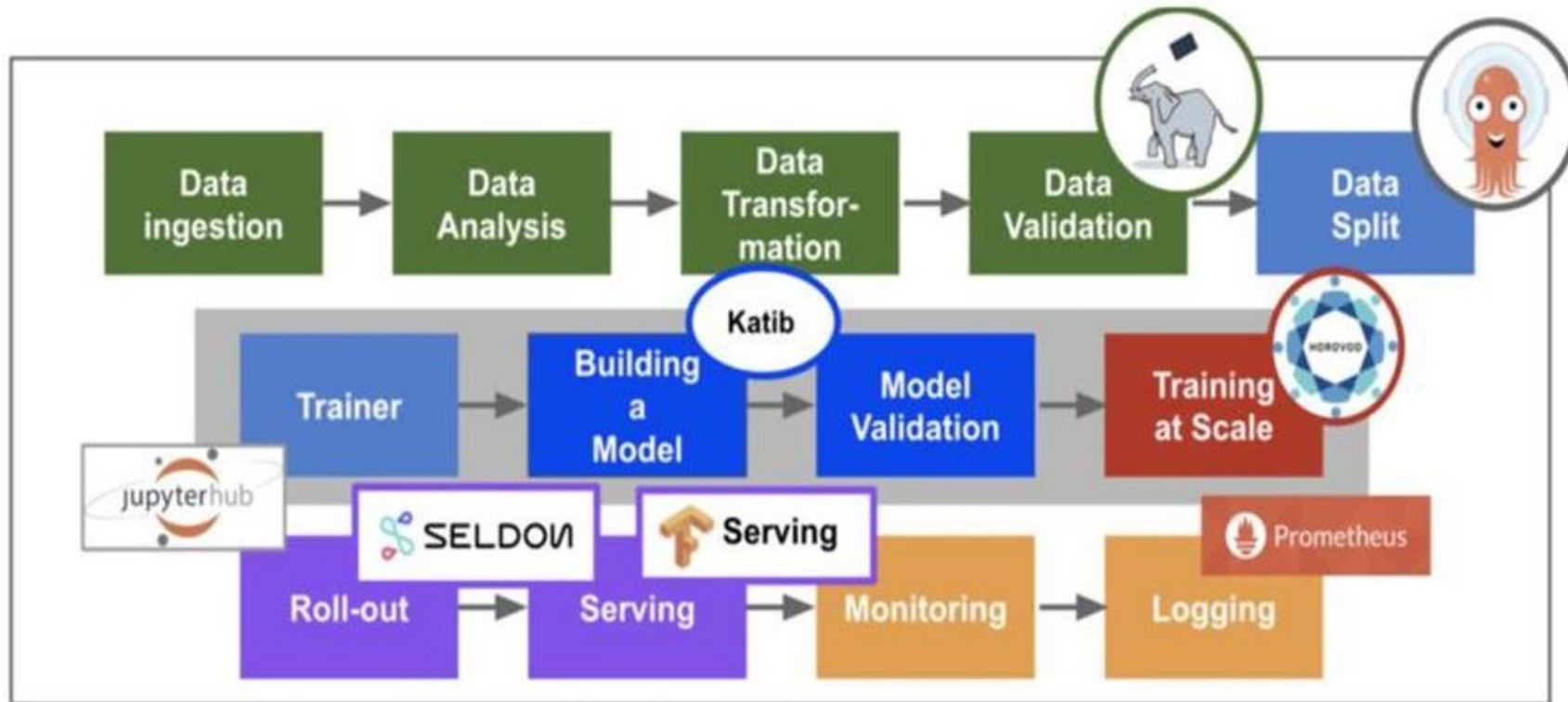
KFP Architecture



<https://shikanon.com/2019/%E8%BF%90%E7%BB%B4/kubeflow%E4%BB%8B%E7%BB%8D/>



Kubeflow





DEMO

Experiment Platform:

DO-DOKS: kubernetes v1.24
Kubeflow: v1.6.1
Linux Distro: Debian 10

Demo:

```
while ! kustomize build example | kubectl apply -f -; do echo "Retrying to apply resources"; sleep 10; done  
kubectl --kubeconfig=/Users/lyan/kubeflow-kubeconfig.yaml port-forward svc/istio-ingressgateway -n istio-system 8080:80
```

Local Setup:

Juju + microk8s: kubernetes v1.22
Kubeflow: v1.6.0
Linux Distro: Ubuntu Jammy
<https://charmed-kubeflow.io/docs/quickstart>



Beyond

Distributed Machine Learning(*)

Why?

Scalability, we really do not need it if it is a small dataset or model or customer base.

What?

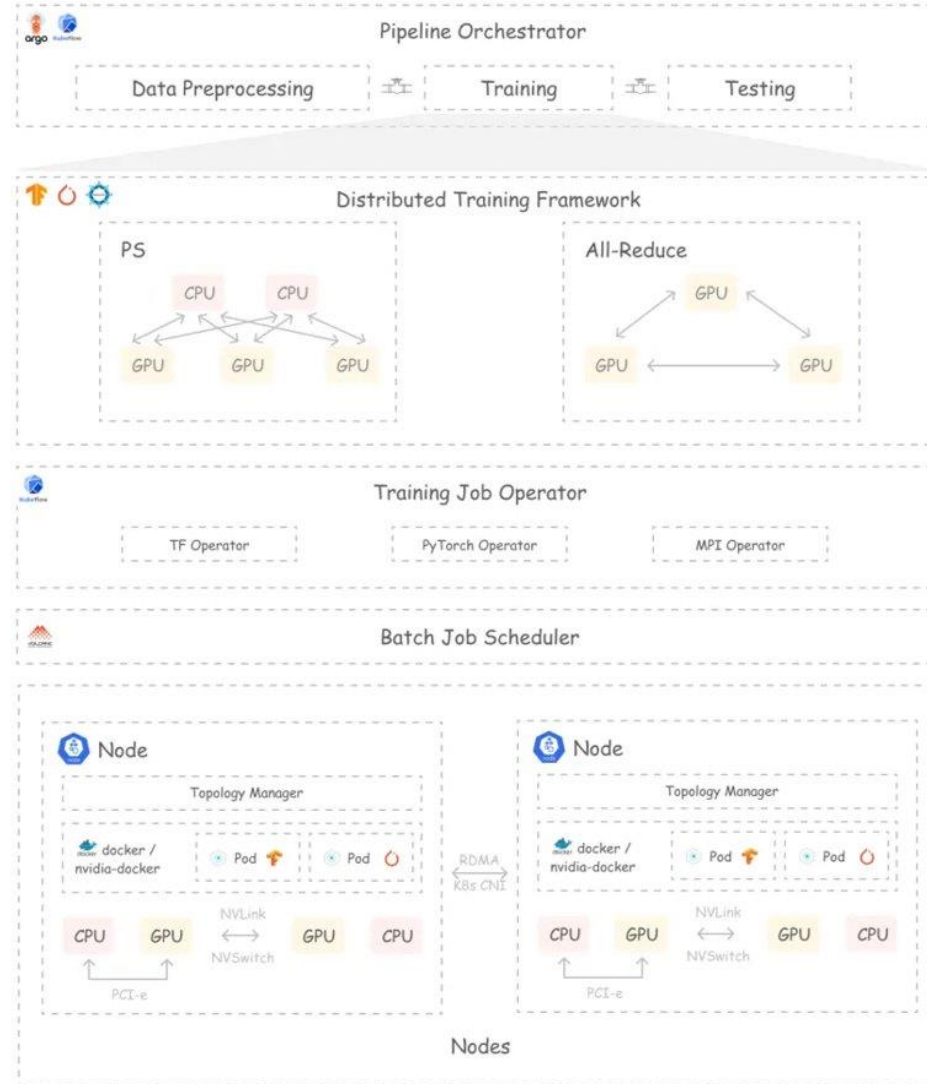
Training Operator Scheduling

Inference Model Optimization

How?

Simulate/predict for scheduler

Model compiler



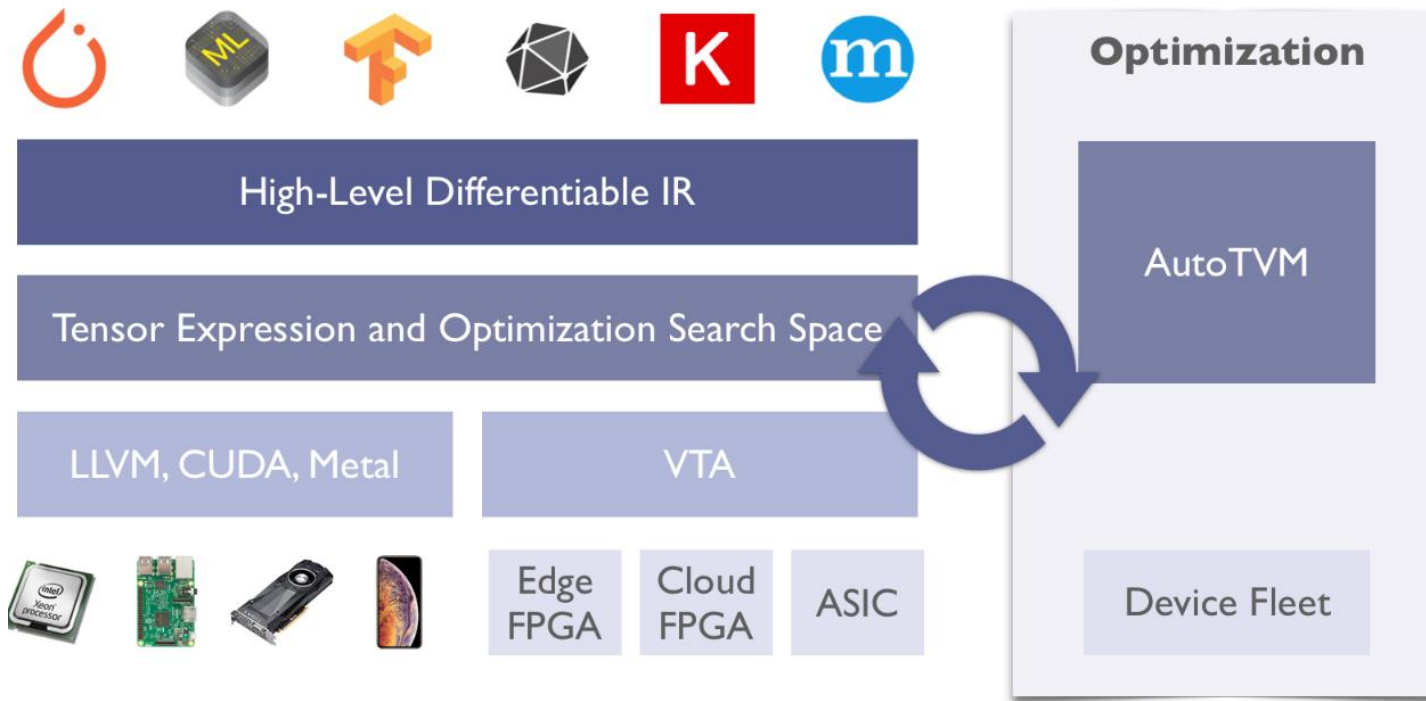
<https://wallpaperaccess.com/to-be-continued>

<https://zhuanlan.zhihu.com/p/548219786>



Beyond

Embedded Model





Lesson learned

1. Deployment

1. Kubevirt 1.22 + kustomerize
2. Disable TLS
3. Setup/Enable StorageClass

2. Running

1. docker runtime re-size
2. docker repository setup

3. Training model

1. ML training requests a lot resources
2. Need to do a lot of experiments
3. Setup environment is time consuming
4. Needs automation/pipeline

4. System Failure / Efficiency

Monitor large scale machine clusters are difficult
Resource Competition

Q & A

Thanks!



Claim:

All the information is based on personal using experience, no preference or commercial advertising. If there are any conflicts, please refer to the statement from providers.



AI Cloud Providers

 Alibaba Cloud




 BAIDU AI CLOUD

 Google Cloud

 IBM **Cloud**

 Microsoft
Azure

ORACLE
Cloud Infrastructure

 Tencent Cloud

 **linode**


Paperspace

 **Lambda**

 **VULTR**



Support Matrix

	M60	P4	P40	P100	T4	RTX 6000	V100	A10	A40	A100	Notes
Aliyun		✓		✓	✓		✓			✓	
AWS	✓		✓		✓		✓	✓		✓	
Baidu		✓			✓		✓			✓	
Google		✓		✓	✓		✓			✓	TPU
IBM	✓			✓			✓				
Microsoft	✓		✓	✓	✓		✓	✓		✓	FPGA/AMD
Oracle				✓			✓			✓	
Tencent		✓	✓		✓		✓			✓	
Linode						✓					
Paperspace							✓				
Lambda						✓	✓				
Vultr										✓	vGPU/MIG

<https://developer.nvidia.com/cuda-gpus>

<https://www.nvidia.com/en-us/data-center/gpu-cloud-computing/>

NVIDIA Data Center Products

GPU	Compute Capability
NVIDIA A100	8.0
NVIDIA A40	8.6
NVIDIA A30	8.0
NVIDIA A10	8.6
NVIDIA A16	8.6
NVIDIA A2	8.6
NVIDIA T4	7.5
NVIDIA V100	7.0
Tesla P100	6.0
Tesla P40	6.1
Tesla P4	6.1
Tesla M60	5.2
Tesla M40	5.2
Tesla K80	3.7
Tesla K40	3.5
Tesla K20	3.5
Tesla K10	3.0



AI Cloud Service

- IAAS

- ML VM Image
- Container:
 - Docker
 - NGC
- Conda/pip3

- PaaS

Help manage data and model
(paperspace, Colaboratory)

- SaaS

Help consume AI solution
(IBM Watson, Google voice)