Introduction to Kubeflow and TensorFlow Extended

Kubeflow

TensorFlow Extended

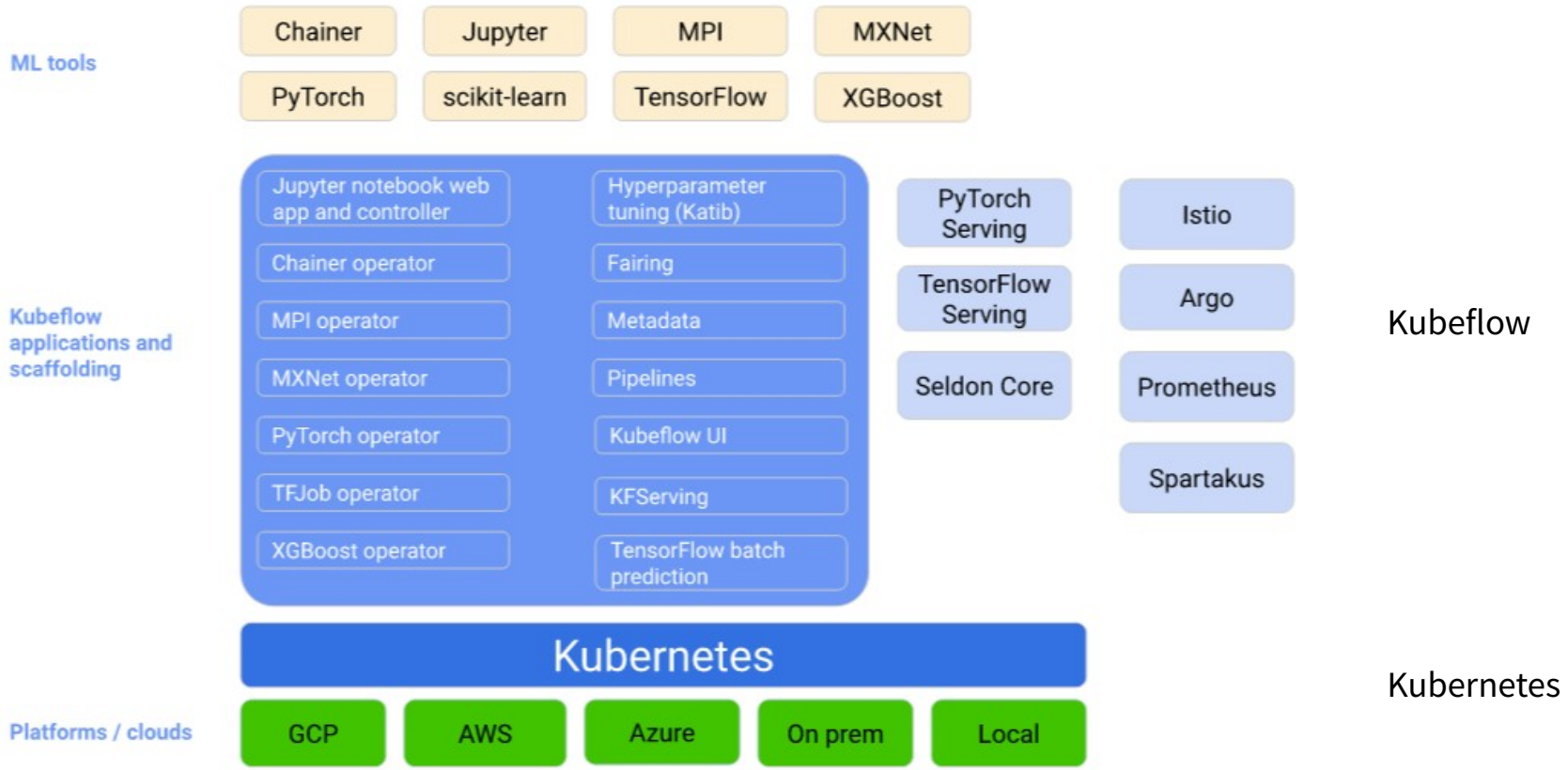
KT AI Service Lab 2020-05-14 (Thu)
Tae-Hyung Kim, Ph.D.
the.kim@kt.com

Kubeflow Overview

Kubeflow = Kubernetes + Machine Learning Flow

The Goal is to provide a straightforward way to deploy **best-of-breed open-source systems for ML** to diverse infrastructures not to recreate other services.

Conceptual overview



https://www.kubeflow.org/docs/started/kubeflow-overview/

Components of Kubeflow

Central Dashboard (Kubeflow UI)

The central user interface (UI) in Kubeflow

Multi-Tenancy in Kubeflow

Multi-user isolation (Admin, User, Profile) & Identity Access Management

Frameworks for Training (ML tools)

Training of ML models in Kubeflow

Miscellaneous

Nuclio - High performance serverless for data processing and ML

Jupyter Notebooks

Using Jupyter notebooks in Kubeflow

Pipelines

ML Pipelines in Kubeflow

Fairing

Streamlines to build, train & deploy in a hybrid cloud environment

Hyperparameter Tuning (Katib)

Hyperparameter tuning of ML models in Kubeflow

Metadata

Tracking and managing metadata of machine learning workflows

Tools for Serving

Serving of ML models in Kubeflow

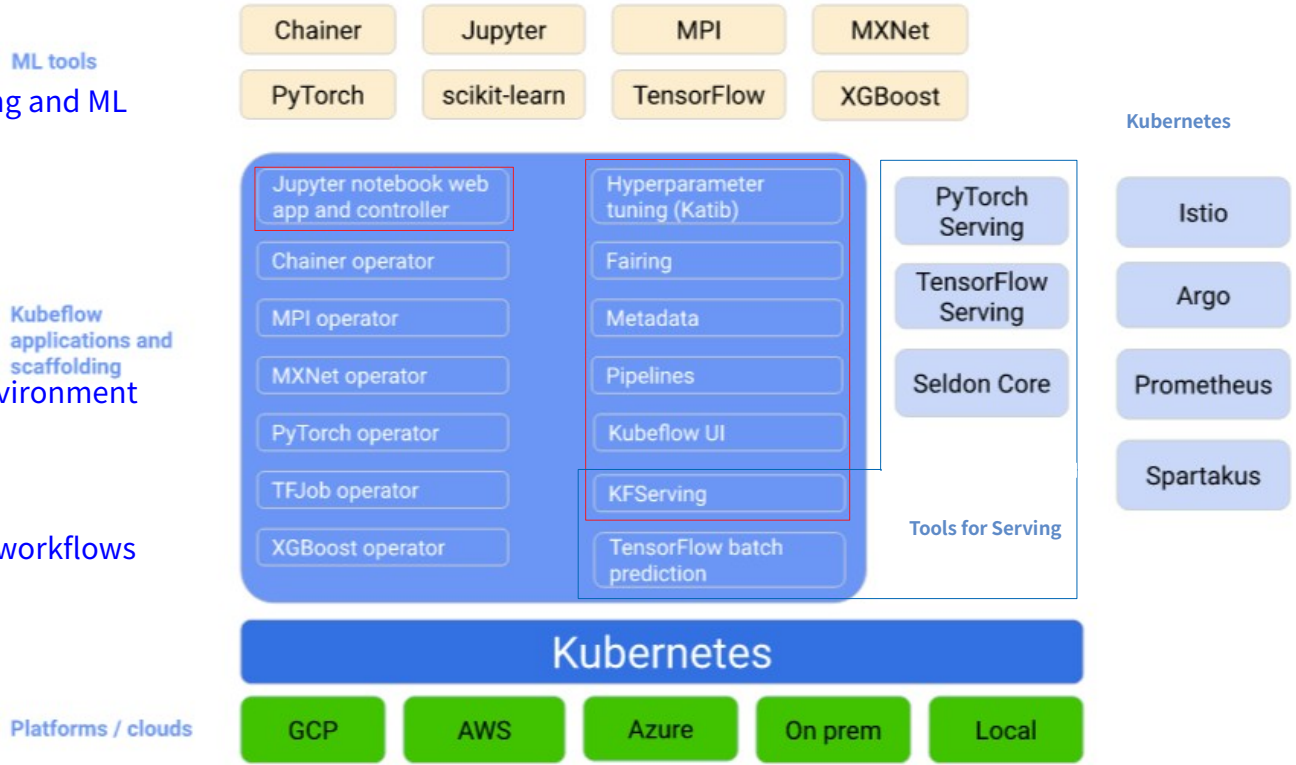
Kubernetes

Kubernetes Istio Connect, secure, control, and observe services

Argo Open source Kubernetes native workflows, events, CI and CD

Prometheus Monitoring system and time series database

Spartakus Collecting usage information about Kubernetes clusters



https://www.kubeflow.org/docs/components/

https://www.kubeflow.org/docs/started/kubeflow-overview/

Kubeflow Central Dashboard or Kubeflow User Interface (KFUI)

Kubeflow

kubeflow-sarahmaddox (...

Namespace configuration can host different team, group, user types, etc.

Home

Pipelines

Notebook Servers

Katib

Artifact Store

Manage Contributors

GitHub

Privacy • Usage Reporting

build version 0.7.0

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 kubeflow-sarahmaddox

Recent Pipelines

[Sample] Basic - Exit Handler
Created 22/12/2019, 06:50:18

[Sample] Basic - Conditional execution
Created 22/12/2019, 06:50:17

[Sample] Basic - Parallel execution
Created 22/12/2019, 06:50:16

[Sample] Basic - Sequential execution
Created 22/12/2019, 06:50:15

[Sample] ML - XGBoost - Training with ...
Created 22/12/2019, 06:50:14

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

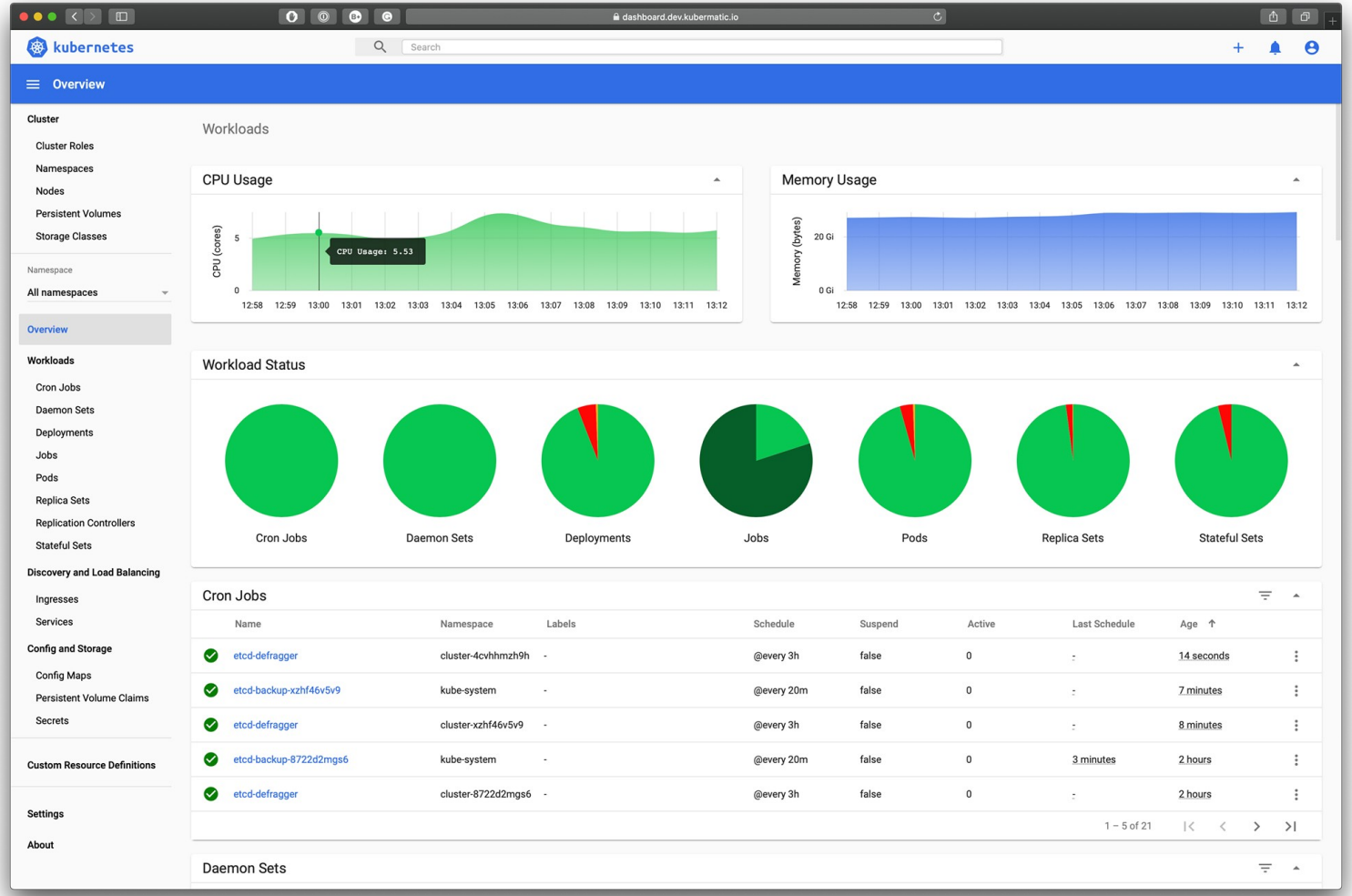
Kubeflow on GCP
Running Kubeflow on Kubernetes Engine and Google Cloud Platform

Kubeflow on AWS
Running Kubeflow on Elastic Container Service and Amazon Web Services

Requirements for Kubeflow

https://www.kubeflow.org/docs/started/kubeflow-overview/#kubeflow-user-interface-ui

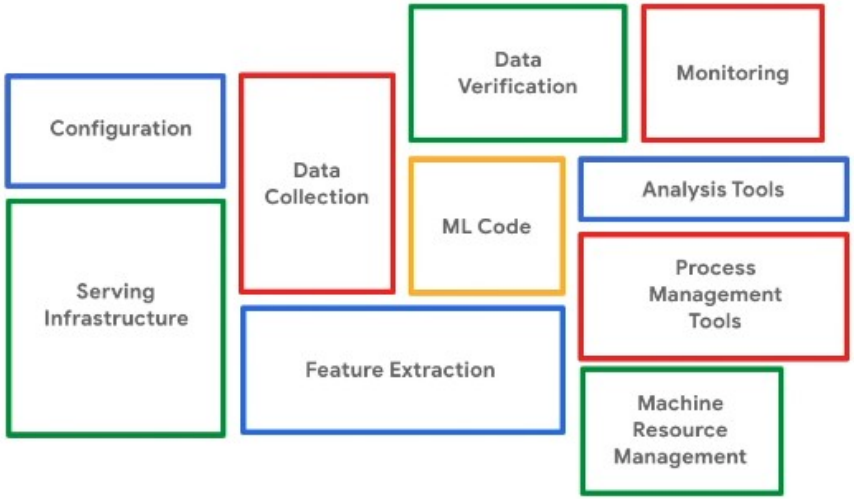
Kubernetes Dashboard

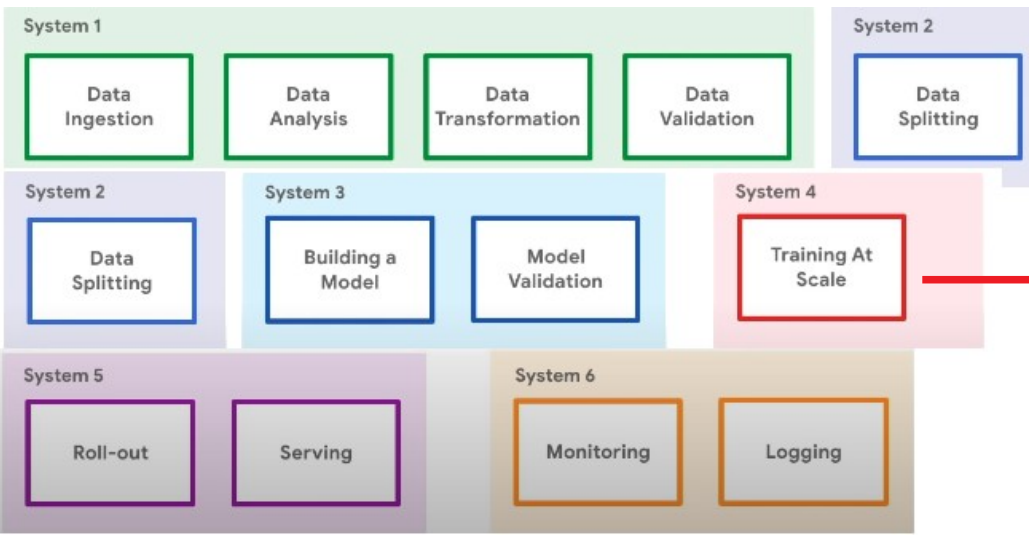


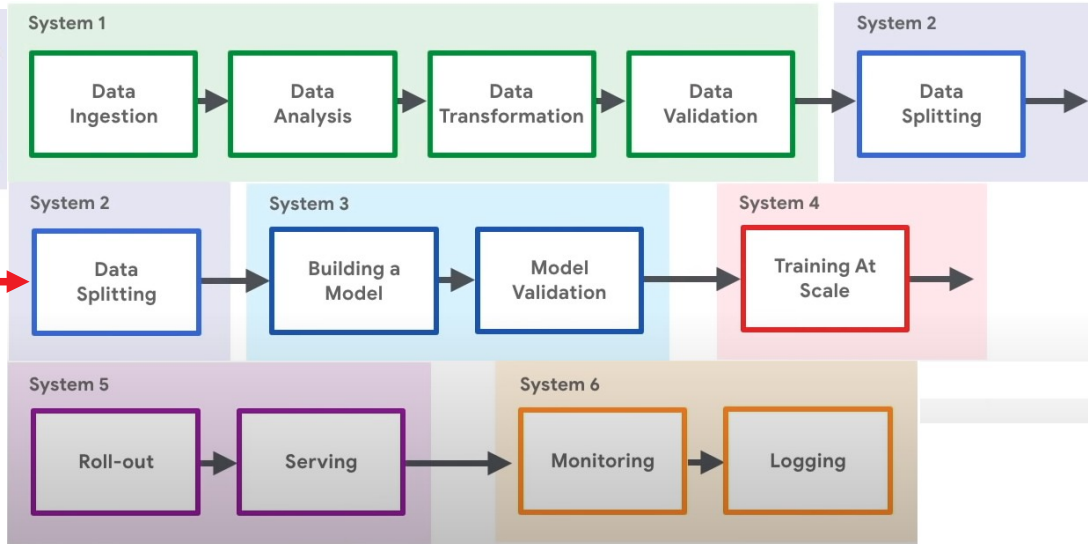
https://github.com/kubernetes/dashboard

Introduction to Kubeflow - Kubeflow 101 (1/3)

https://youtu.be/cTZArDgbIWw [3:46]

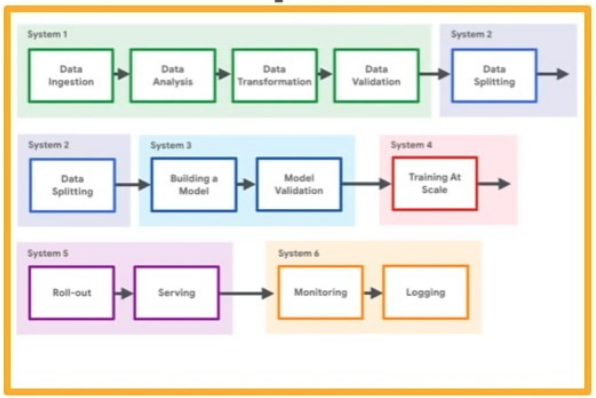






Introduction to Kubeflow - Kubeflow 101 (2/3)

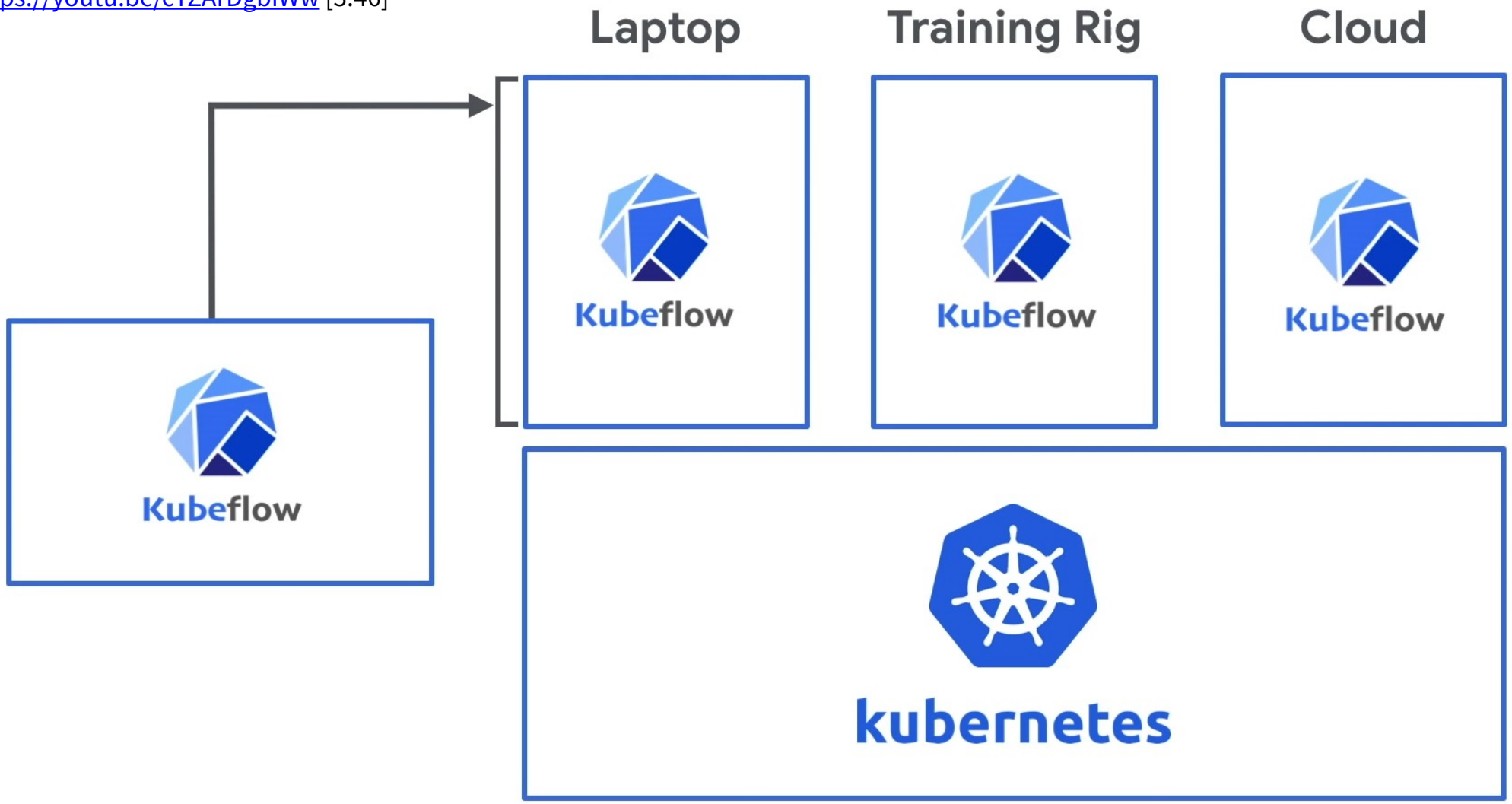
https://youtu.be/cTZArDgbIWw [3:46]



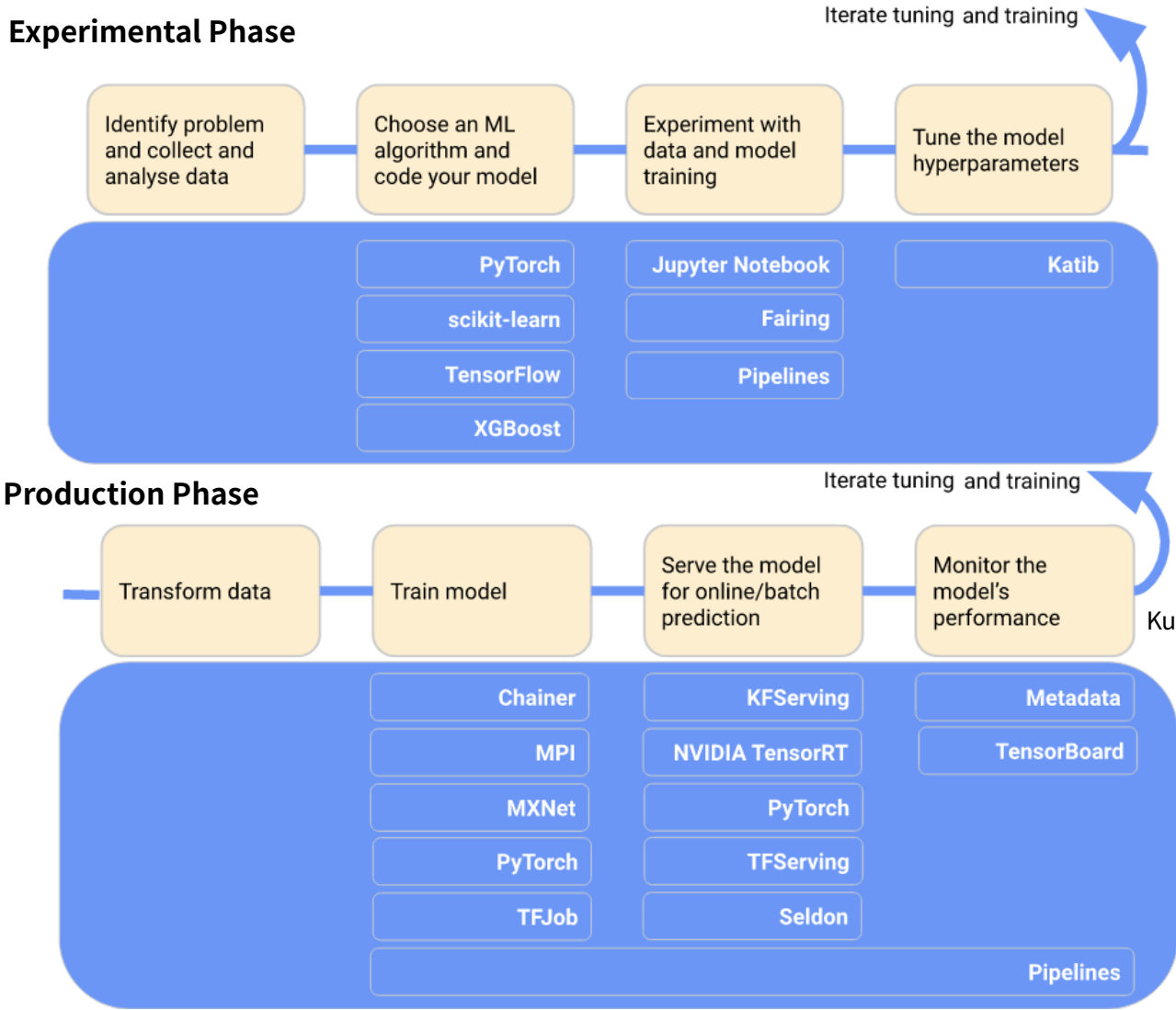
Laptop	Training Rig	Cloud
Model	Model	Model
UX	UX	UX
Tooling	Tooling	Tooling
Framework	Framework	Framework
Storage	Storage	Storage
Runtime	Runtime	Runtime
Drivers	Drivers	Drivers
OS	OS	OS
Accelerator	Accelerator	Accelerator
HW	HW	HW

Introduction to Kubeflow - Kubeflow 101 (3/3)

https://youtu.be/cTZArDgbIWw [3:46]



Kubeflow Components in the Machine Learning Workflow



Spawn & manage Jupyter notebooks

Kubeflow Pipelines

build, deploy, & manage multi-step ML workflows based on Docker containers.

Kubeflow offers several components that you can use to build you

Big Picture: End-to-End Platform for Deploying Production Machine Learning Pipelines

Kubeflow

The Machine Learning Toolkit for Kubernetes

Kubeflow

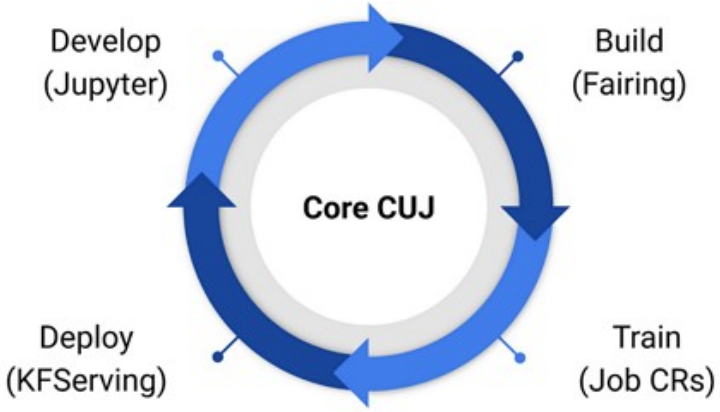
https://www.kubeflow.org/

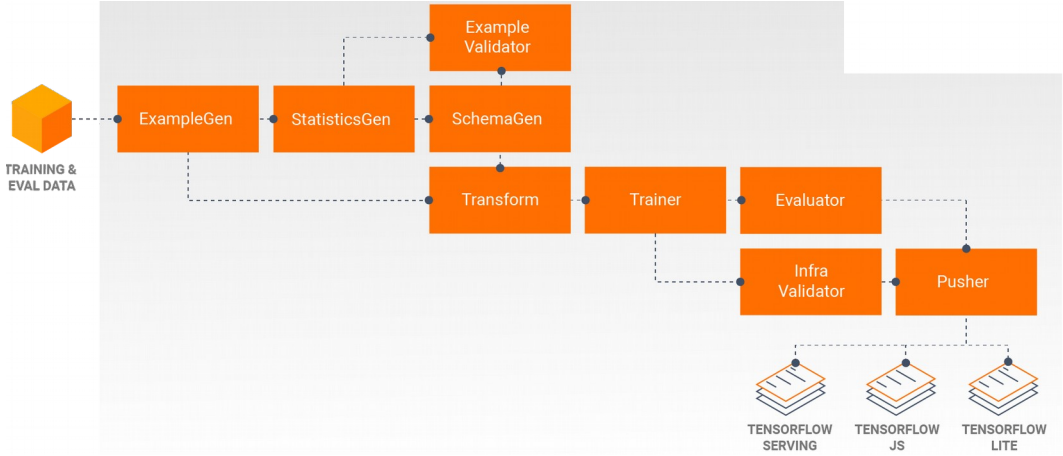
TensorFlow Extended (TFX)

an end-to-end platform for
deploying production ML pipelines

TensorFlow Extended

https://www.tensorflow.org/tfx





Overview of TensorFlow Extended (TFX)

- TensorFlow Extended (TFX) is a Google-production-scale machine learning platform based on TensorFlow.
- It provides a configuration framework and shared libraries to integrate common components needed to define, launch, and monitor your machine learning system.

- TensorFlow 2.x was released at TensorFlow Dev Summit 2019; TFX as an extension package.
  - [15:06] TensorFlow Extended (TFX) Post-training Workflow (TF Dev Summit '19) https://youtu.be/0O201IQlkxc
  - [31:34] TensorFlow Extended (TFX) Overview and Pre-training Workflow (TF Dev Summit '19) https://youtu.be/A5wiwT1qFjc

- TFX is compatible with TensorFlow 2.x and the high-level APIs that existed in TensorFlow 1.x .

Installation

```
pip install tfx
```

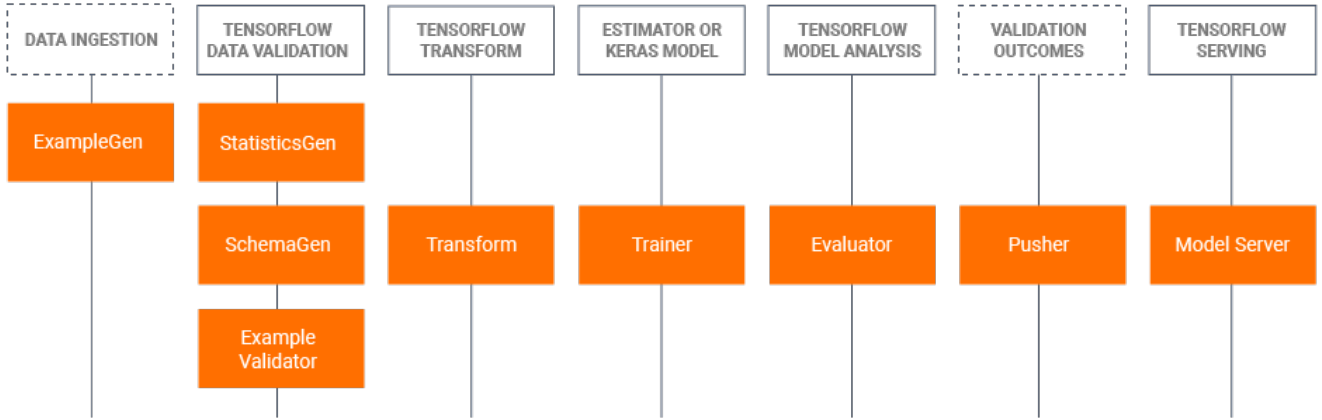
Homepage/Git Repository

https://www.tensorflow.org/tfx

https://github.com/tensorflow/tfx

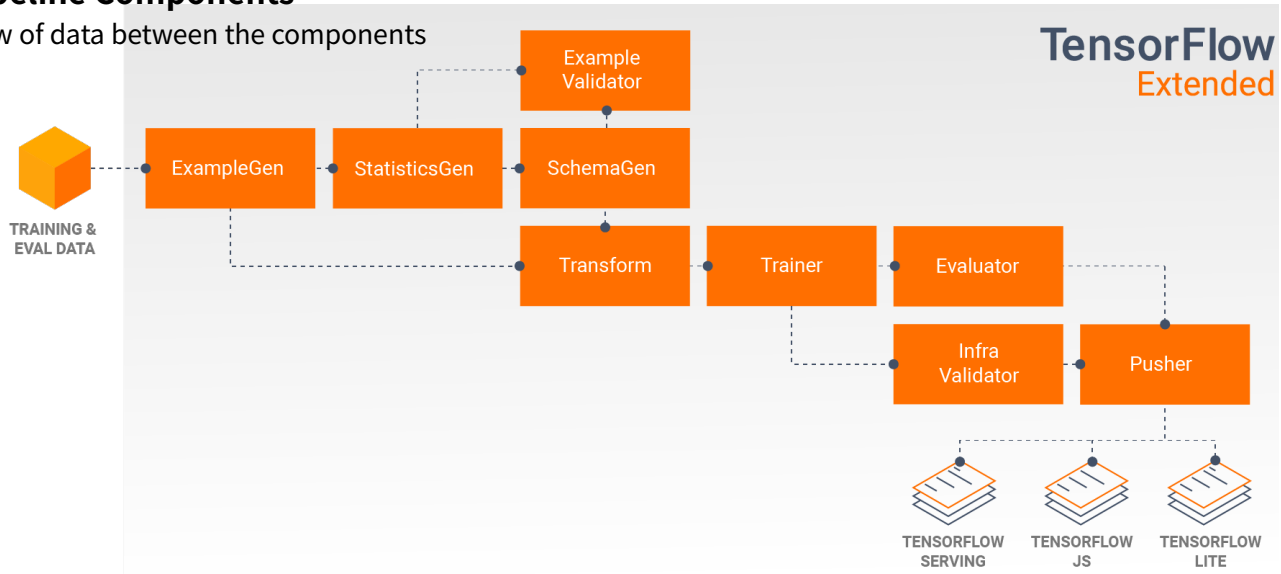
TensorFlow Extended (TFX) Libraries & Pipeline Components

TFX Libraries



TFX Pipeline Components

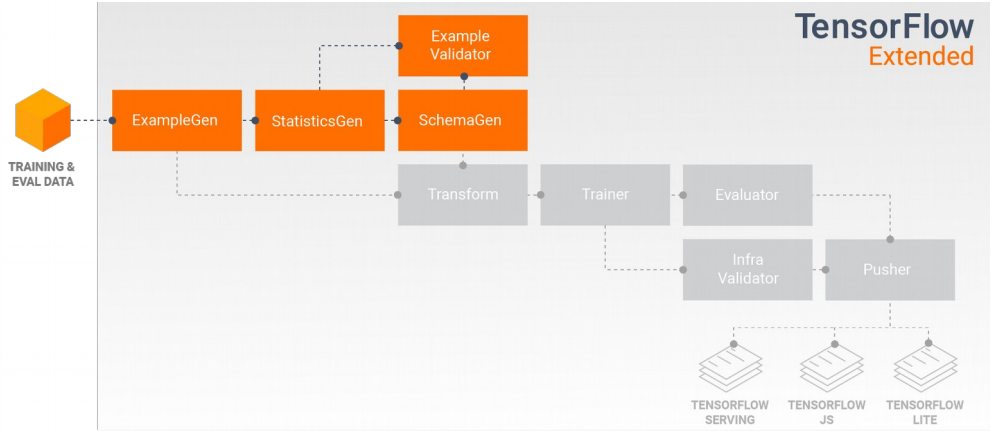
the flow of data between the components



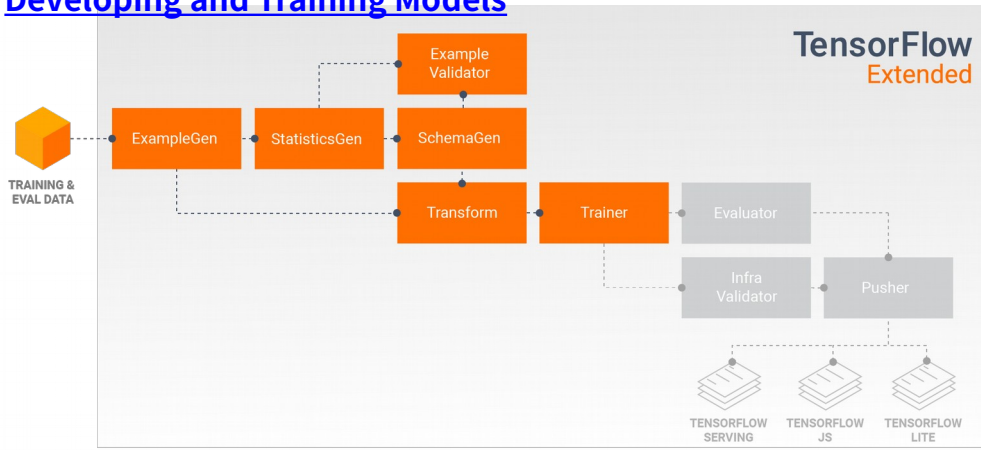
Source: The TFX User Guide, https://www.tensorflow.org/tfx/guide

TensorFlow Extended (TFX) Pipleline Components in Action

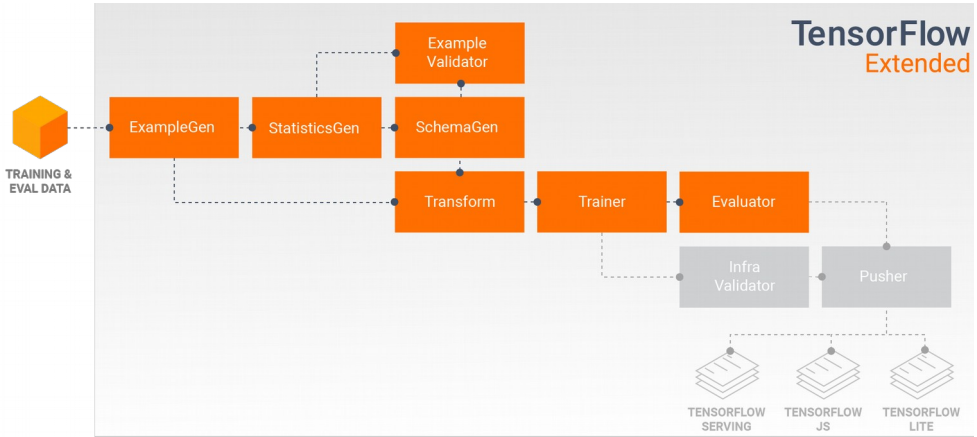
Data Exploration, Visualization, and Cleaning



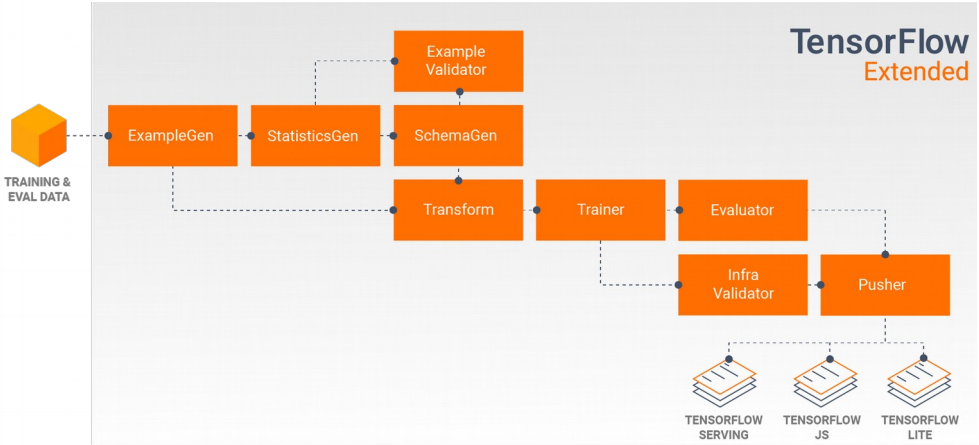
Developing and Training Models



Analyzing and Understanding Model Performance

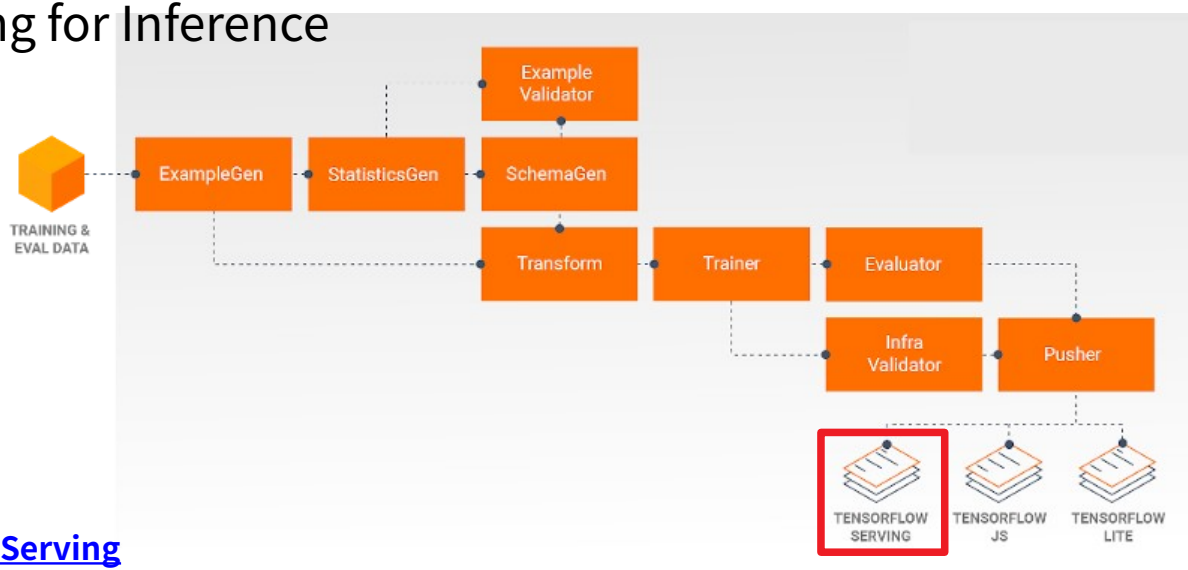


Deployment Targets



Source: The TFX User Guide, https://www.tensorflow.org/tfx/guide

TensorFlow Serving for Inference



Inference: TensorFlow Serving

- TensorFlow Serving (TFS) is a flexible, high-performance serving system for machine learning models, **designed for production environments**.
- It runs as a set of processes on one or more network servers, using one of several advanced architectures to handle synchronization and distributed computation.
- It consumes a SavedModel and will accept inference requests over either **REST or gRPC interfaces**.

- In a typical pipeline, a SavedModel which has been trained in a Trainer component would first be infra-validated in an InfraValidator component.
- InfraValidator launches **a canary TFS model server** to actually serve the SavedModel.
- If validation has passed, a Pusher component will finally deploy the SavedModel to your TFS infrastructure.
- This includes **handling multiple versions and model updates**.

- For details, refer to,
  - Serving Models, https://www.tensorflow.org/tfx/guide/serving
  - **From Research to Production with TFX Pipelines and ML Metadata**, https://blog.tensorflow.org/2019/05/research-to-production-with-tfx-ml.html

TensorFlow Extended (TFX) with respect to Kubeflow

Portability and Interoperability

TFX is designed to be portable to multiple environments and orchestration frameworks, including

- Apache Beam (required)
- Apache Airflow (optional), and
- Kubeflow (optional).

Creating a TFX Pipeline With Kubeflow

Setup

Kubeflow requires a Kubernetes cluster to run the pipelines at scale. See the Kubeflow deployment guideline that guide through the options for deploying the Kubeflow cluster.

Configure and run TFX pipeline

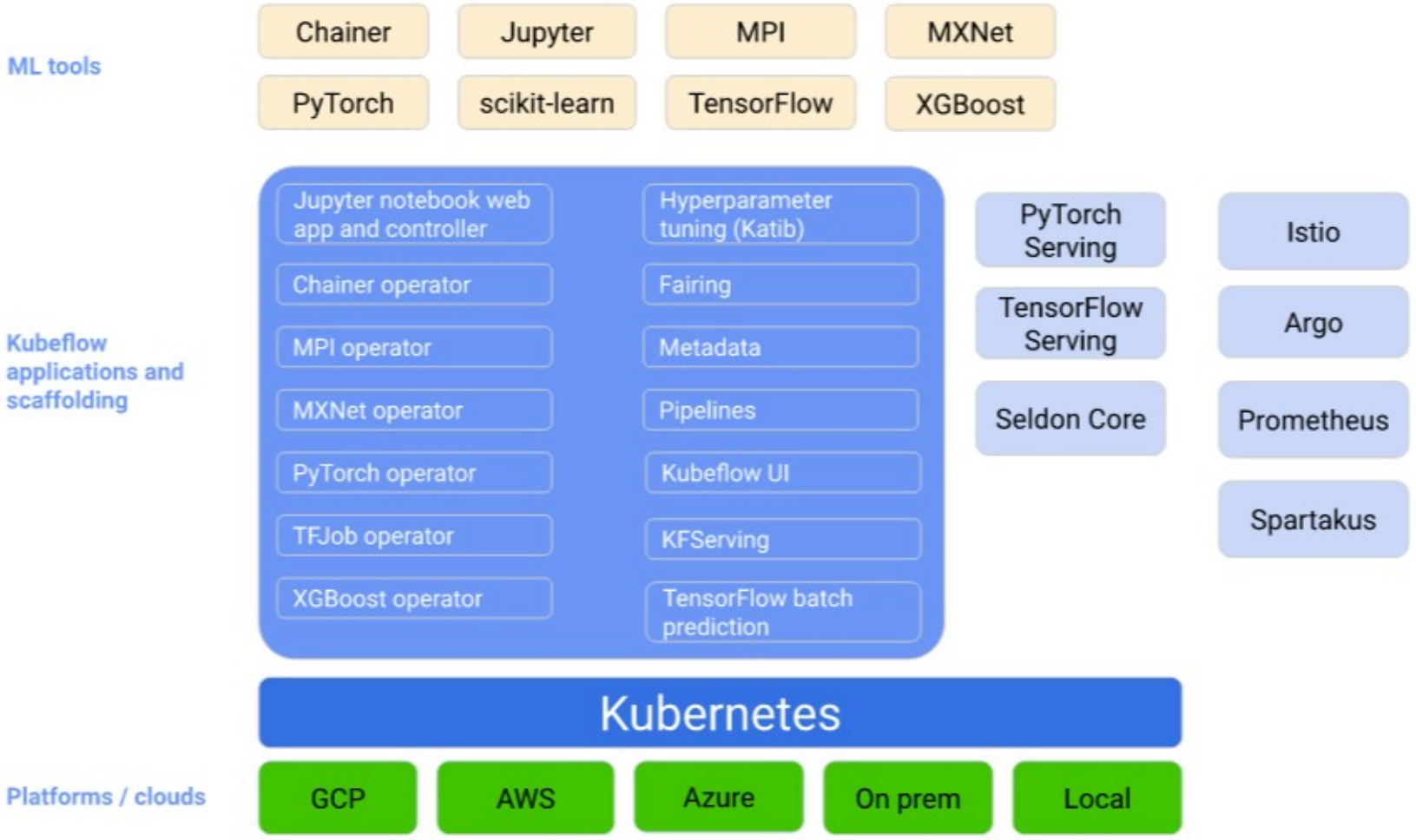
Please follow the TFX on Cloud AI Platform Pipeline tutorial to run the TFX example pipeline on Kubeflow. TFX components have been containerized to compose the Kubeflow pipeline and the sample illustrates the ability to configure the pipeline to read large public dataset and execute training and data processing steps at scale in the cloud.

Kubeflow is dedicated to making deployments of machine learning (ML) workflows on Kubernetes simple, portable and scalable. Kubeflow's goal is not to recreate other services, but to **provide a straightforward way to deploy best-of-breed open-source systems for ML to diverse infrastructures.**
Kubeflow Pipelines enable composition and execution of reproducible workflows on Kubeflow, integrated with experimentation and notebook based experiences. Kubeflow Pipelines services on Kubernetes include the hosted Metadata store, container based orchestration engine, notebook server, and UI to help users develop, run, and manage complex ML pipelines at scale. The Kubeflow Pipelines SDK allows for creation and sharing of components and composition of pipelines programmatically.

Source: The TFX User Guide, https://www.tensorflow.org/tfx/guide

Kubeflow Overview

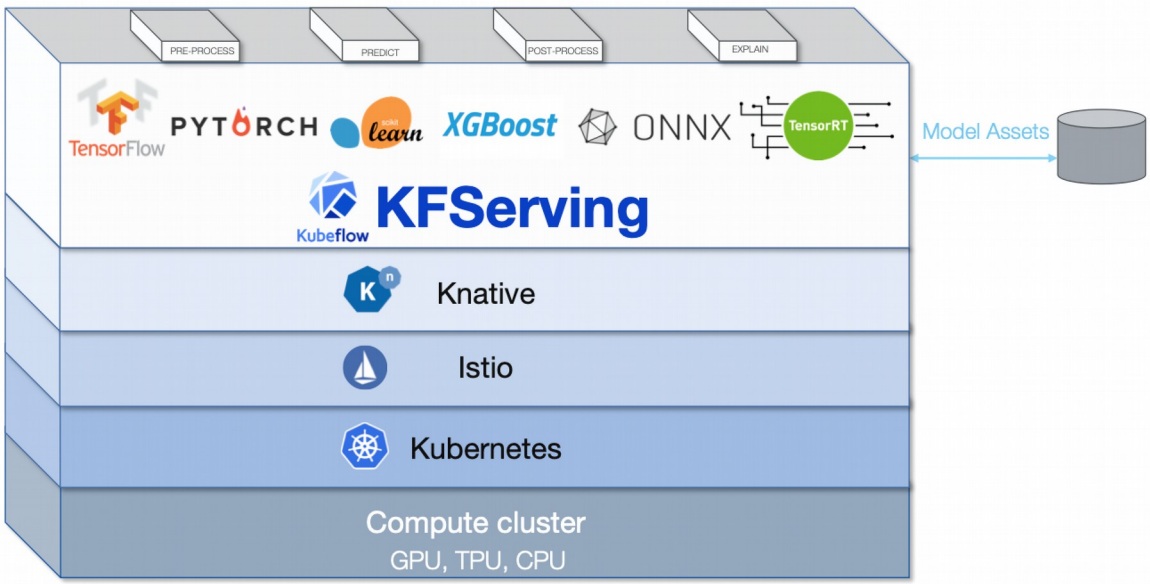
Conceptual overview

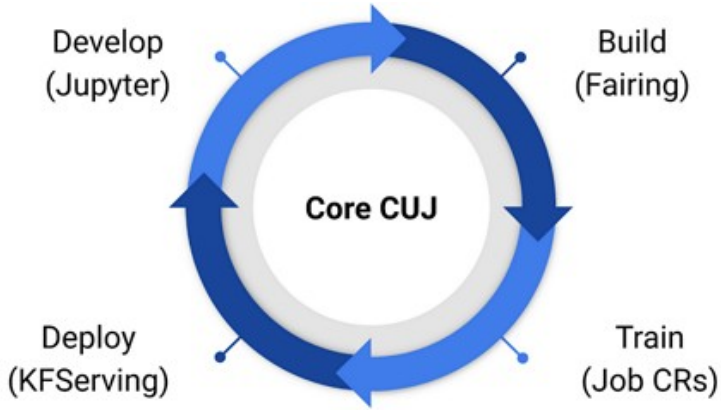


https://www.kubeflow.org/docs/started/kubeflow-overview/

KFServing

KFServing provides a Kubernetes Custom Resource Definition for serving machine learning (ML) models on arbitrary frameworks.
It aims to solve production model serving use cases by providing performant, high abstraction interfaces for common ML frameworks like Tensorflow, XGBoost, ScikitLearn, PyTorch, and ONNX.
It encapsulates the complexity of autoscaling, networking, health checking, and server configuration to bring cutting edge serving features like GPU Autoscaling, Scale to Zero, and Canary Rollouts to your ML deployments.
It enables a simple, pluggable, and complete story for Production ML Serving including prediction, pre-processing, post-processing and explainability.





https://github.com/aimldl/computing_environments/blob/master/kubeflow/temp.md

Thanks