Machine Learning Operations (MLOps)

Usage of Pipelines in the ML Lifecycle with Tensor Flow Extended (TFX) and Kubeflow

Prof. Dr. Jan Kirenz
HdM Stuttgart

The Proof of Concept Factory

Most companies...

- … conduct AI experiments and pilots but achieve a low scaling success rate
- … have significant under investment, yielding low returns

accenture

80-85% PoC Factory

Source: Accenture (2019) https://www.accenture.com/us-en/insights/artificial-intelligence/ai-investments

Gartner Top 10 Data and Analytics Trends, 2021

Accelerating Change

1. Smarter, Responsible, Scalable AI
2. Composable Data and Analytics
3. Data Fabric Is the Foundation
4. From Big to Small and Wide Data

Operationalizing Business Value

5. XOps
6. Engineering Decision Intelligence
7. D&A as a Core Business Function

Distributed Everything

8. Graph Relates Everything
9. The Rise of the Augmented Consumer
10. D&A at the Edge

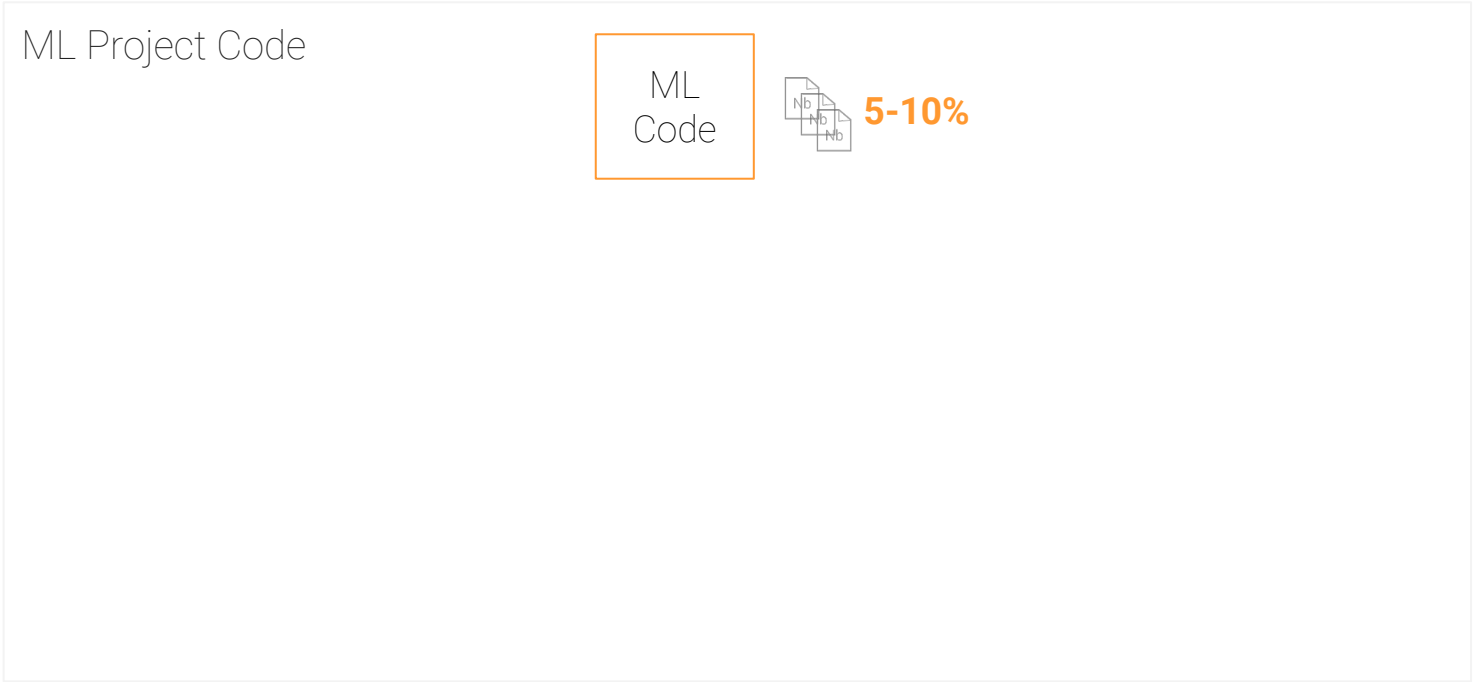
Scalable AI

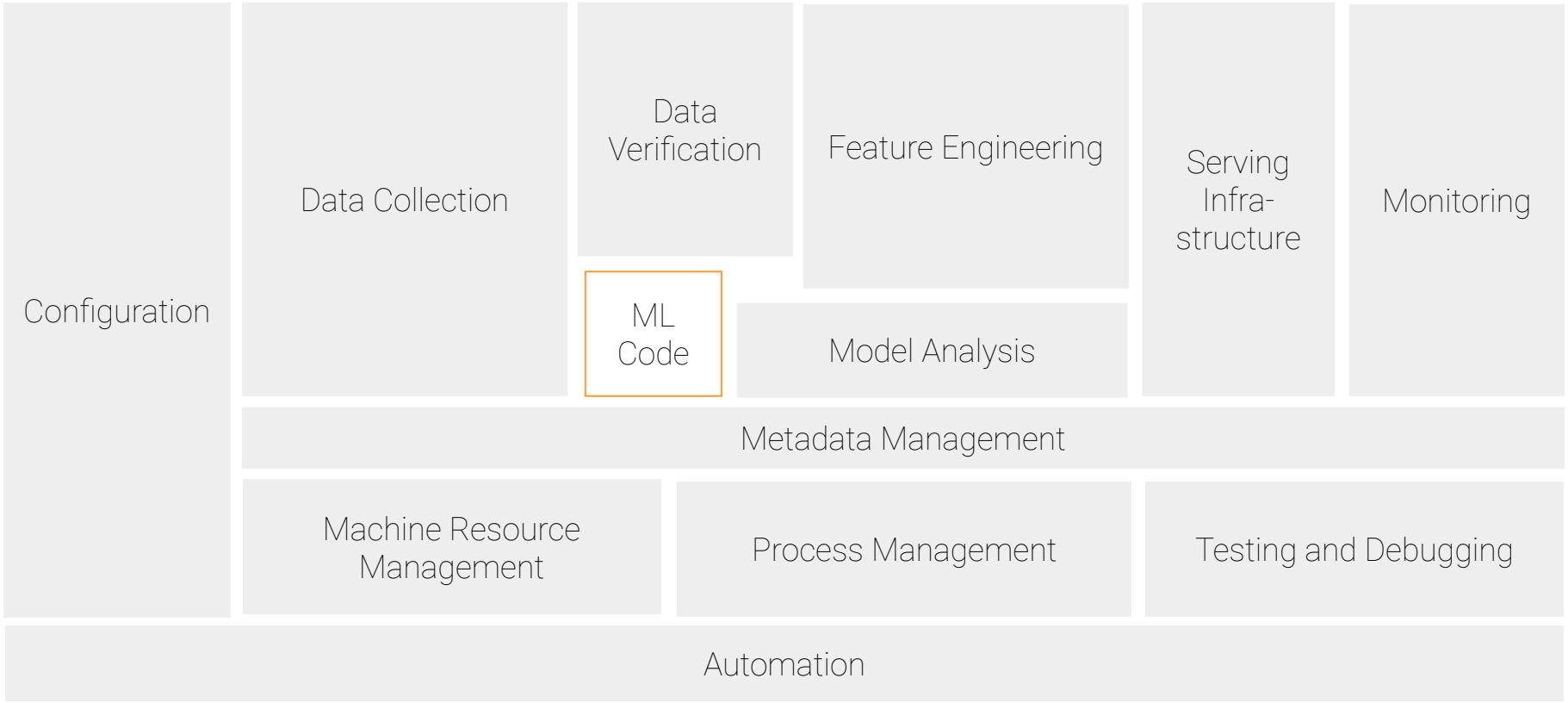
gartner.com/SmarterWithGartner

Gartner

https://www.gartner.com/smarterwithgartner/gartner-top-10-data-and-analytics-trends-for-2021/

The problem with scaling AI: ML code is only a fraction of a production-ready ML project code



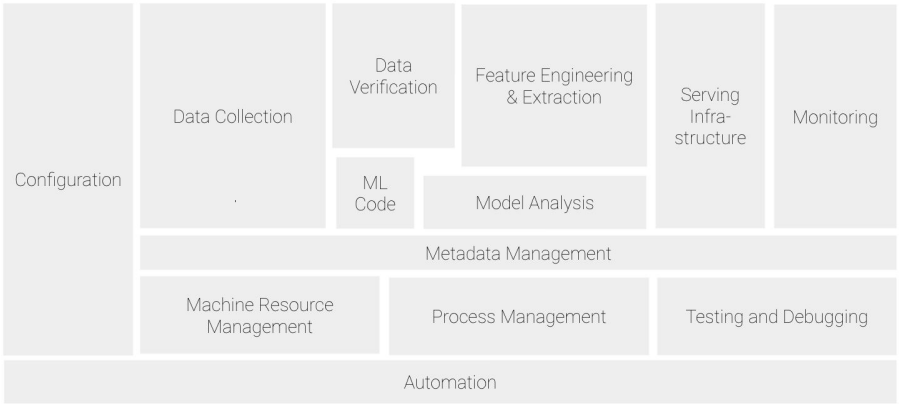


Hidden technical debt in machine learning systems

Sculley, D. et al. (2015). Hidden technical debt in machine learning systems. Advances in neural information processing systems, 28, pp. 2503-2511

Machine learning operations (MLOps)

- ML Engineering culture and practice that aims at **unifying** ML System **development** (Dev) and ML system **operations** (Ops)
- Tools and principles to support workflow **standardization** and **automation** through the ML system lifecycle (e.g. with *pipelines*)

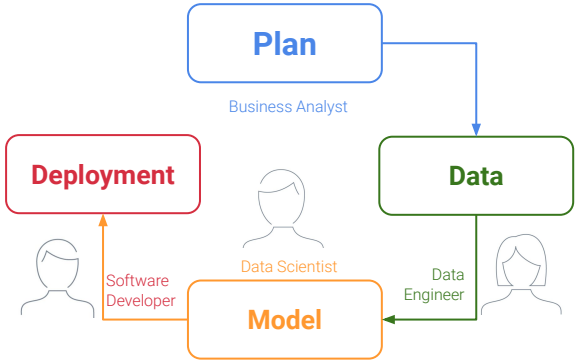


Machine learning **lifecycle**

Prof. Dr. Jan Kirenz

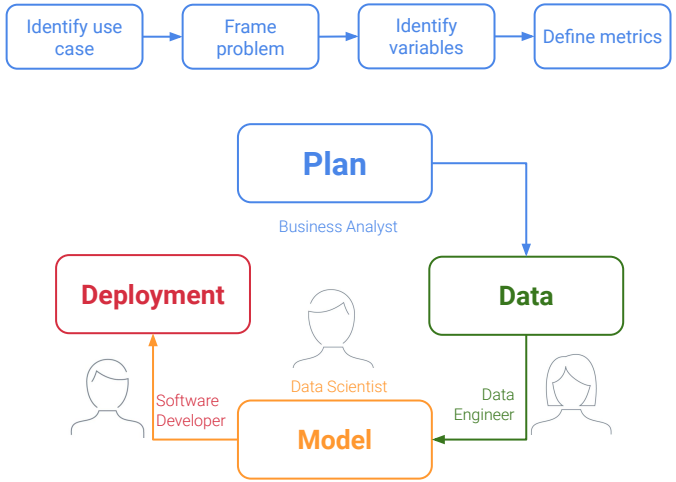
Lifecycle
of an ML System

Plan | Data | Model | Deployment



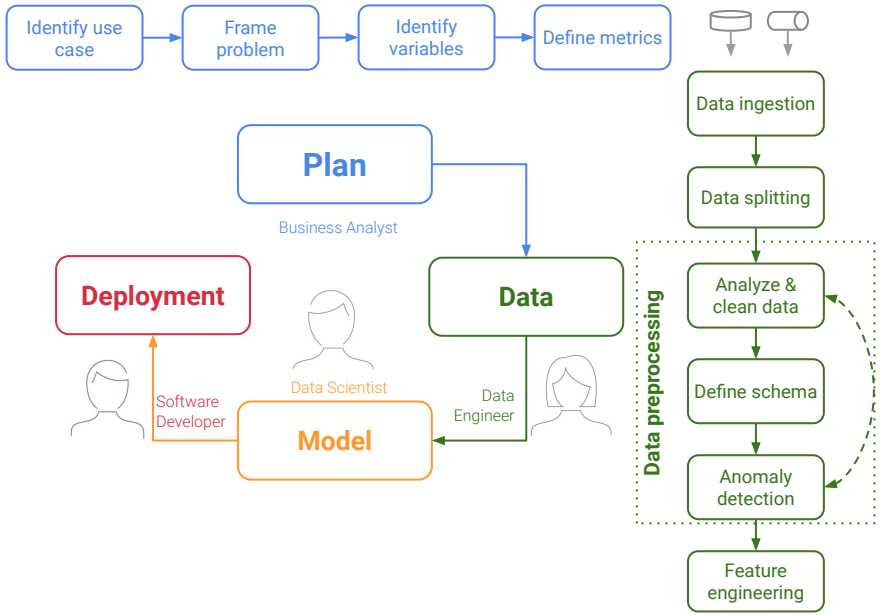
Lifecycle
of an ML System

Plan | Data | Model | Deployment



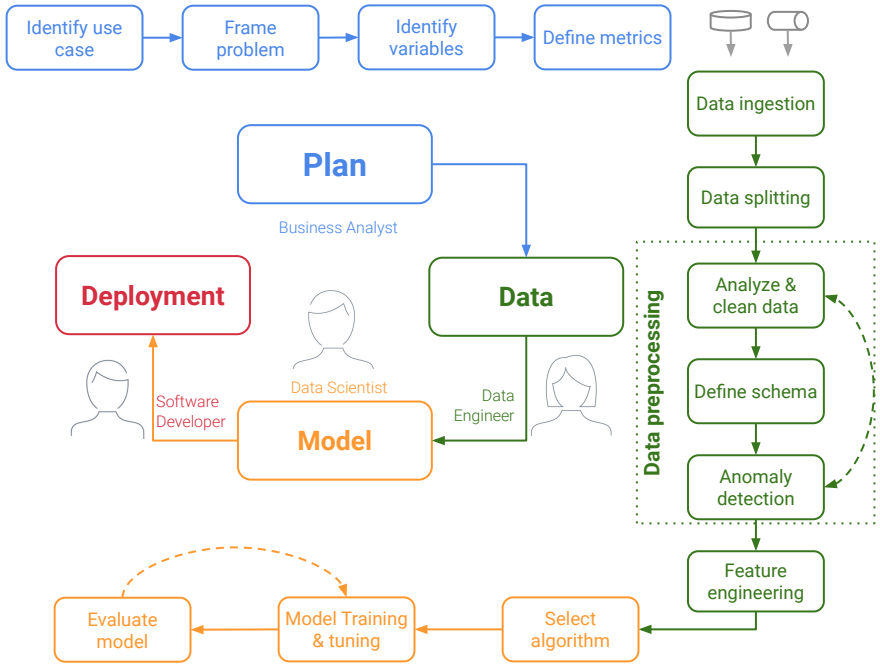
Lifecycle
of an ML System

Plan | **Data** | Model | Deployment



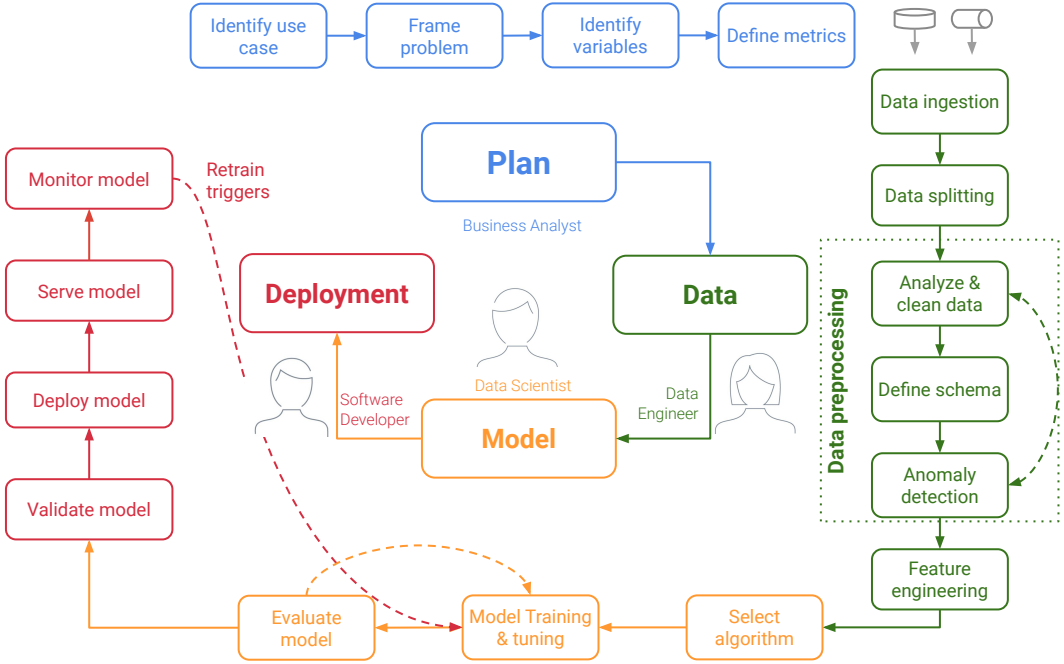
Lifecycle
of an ML System

Plan | Data | **Model** | Deployment



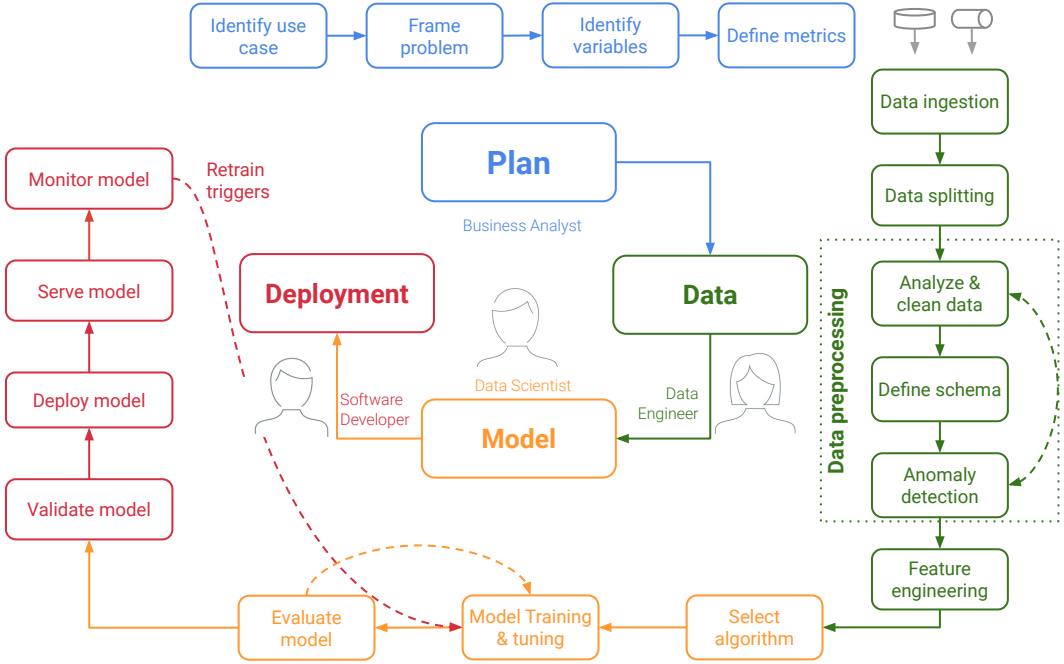
Lifecycle
of an ML System

Plan | Data | Model | **Deployment**



Lifecycle
of an ML System

Plan | Data | Model | Deployment

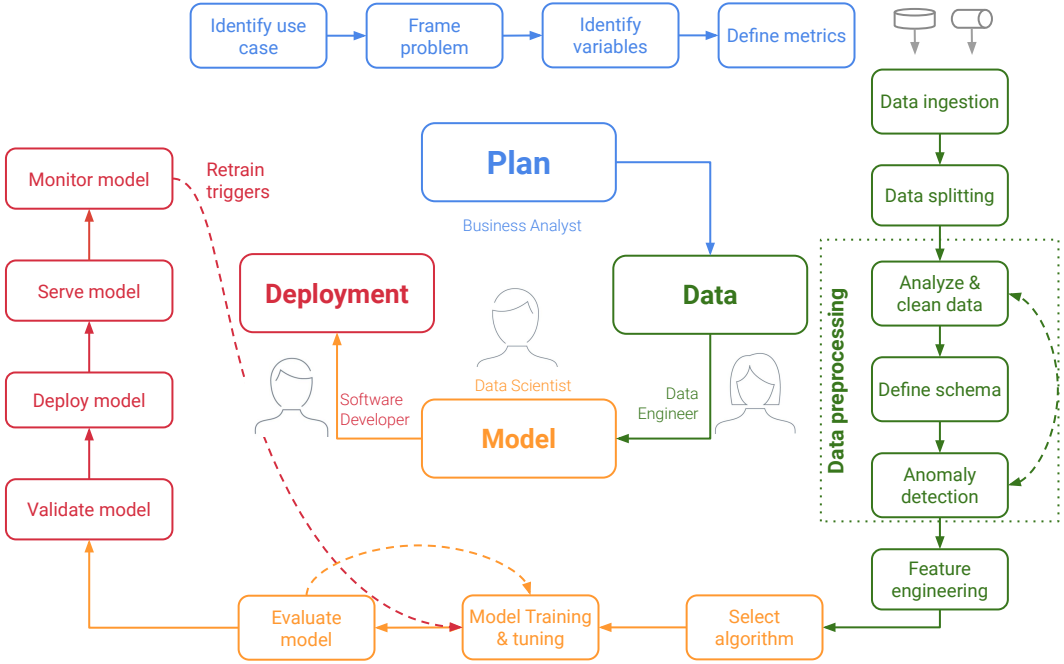


Common issues which lead to a PoC to production gap

- Lack of reuse and duplication
- Inconsistency (data, code, models)
- Manual and slow transition from PoC to production

Lifecycle
of an ML System

Plan | Data | Model | Deployment

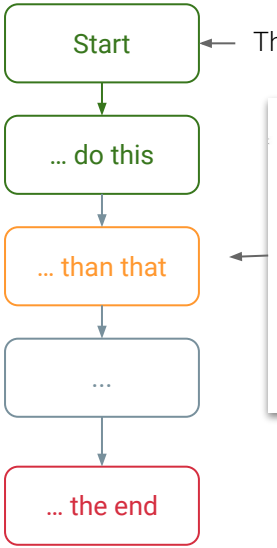


MLOps components

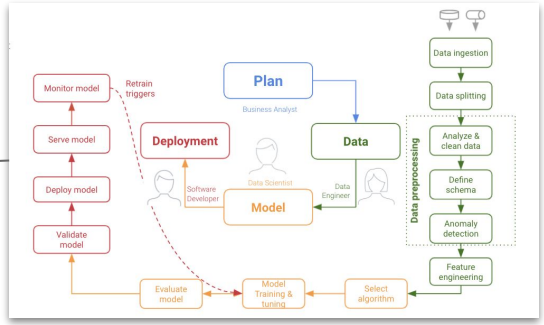
Data and feature management	Model management	Pipeline management	Metadata management
Feature store	Model registry	Pipeline orchestration	Metadata store

What is a pipeline?

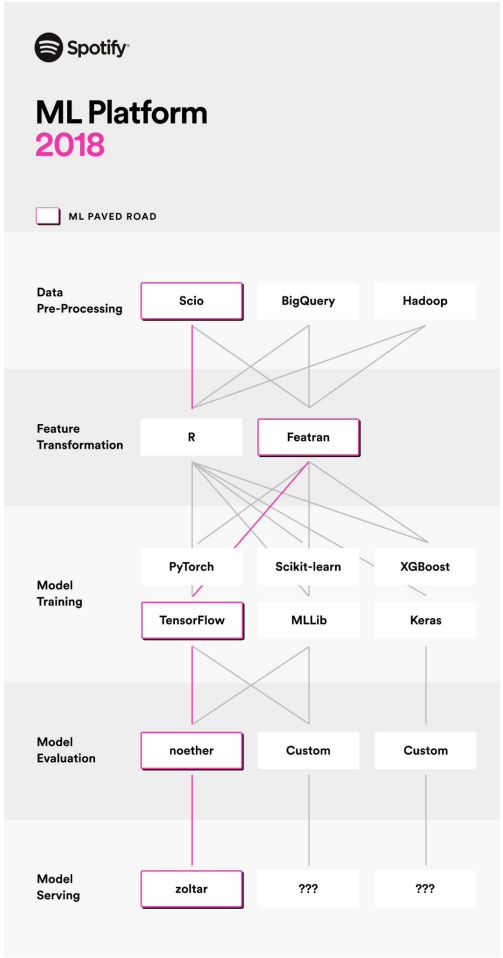
- Description of an ML **workflow**
- A pipeline **component** is a self-contained set of user code that performs one step in the pipeline
- Includes the definition of the **configuration** and **inputs** required to run the pipeline (e.g. model hyperparameters)

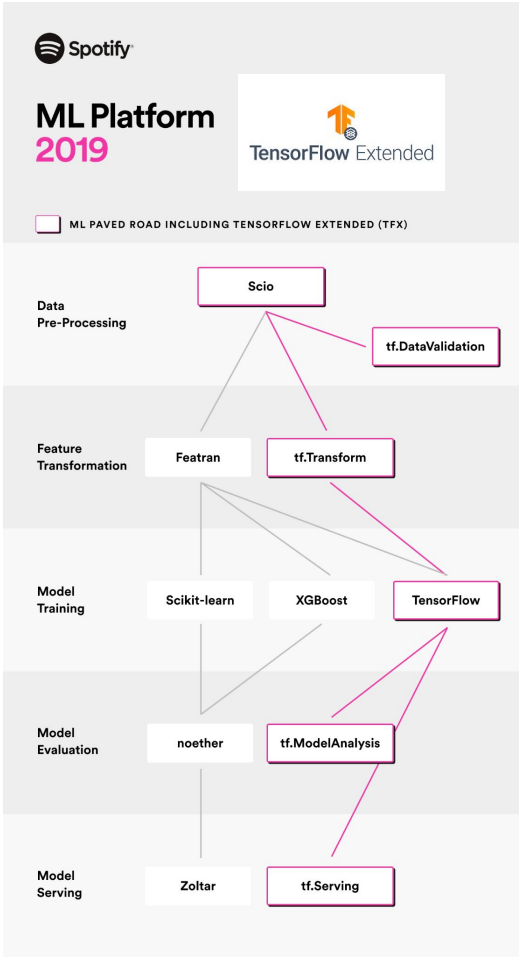


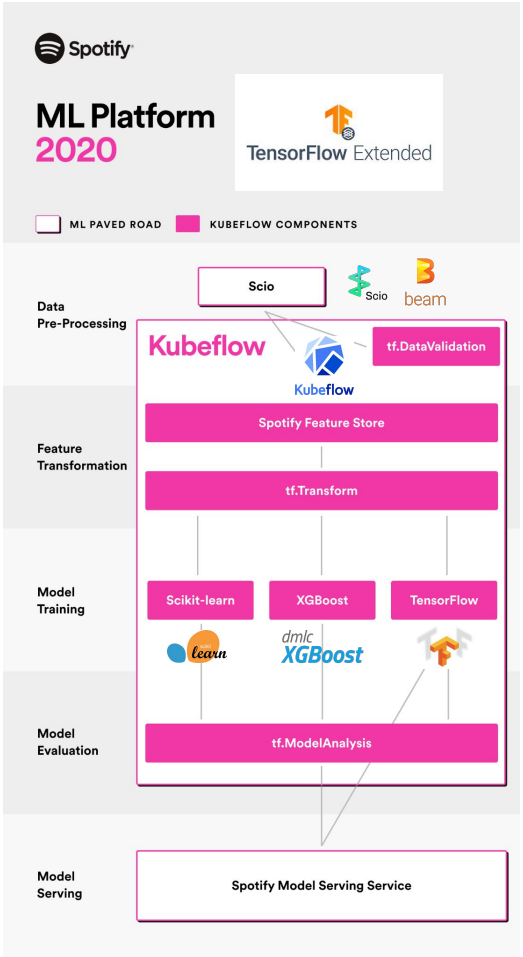
The workflow is also called directed acyclic graph (**DAG**)

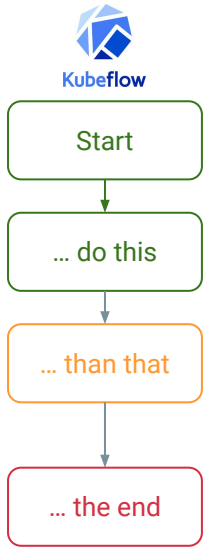


Complete workflow of the ML system lifecycle









Source: Baer & Ngahane (2019)

TensorFlow Extended (TFX)

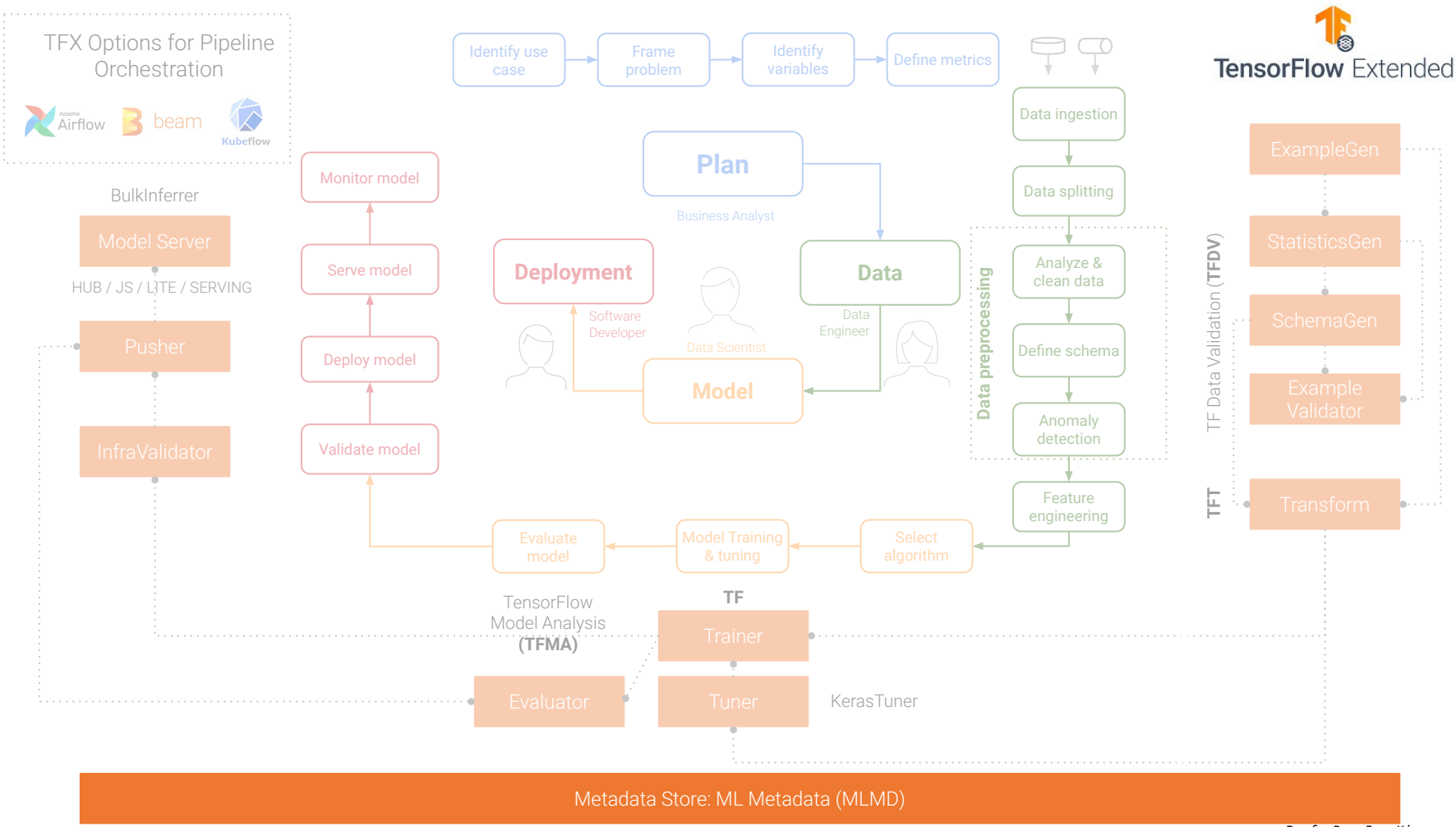
- Google-production-scale machine learning (ML) platform based on TensorFlow
- Portable to multiple environments (Azure, AWS, Google Cloud, IBM, ...)
- Python based toolkit; can be used with notebooks
- Helps you orchestrate your ML process: Apache Airflow, Apache Beam or Kubeflow pipelines

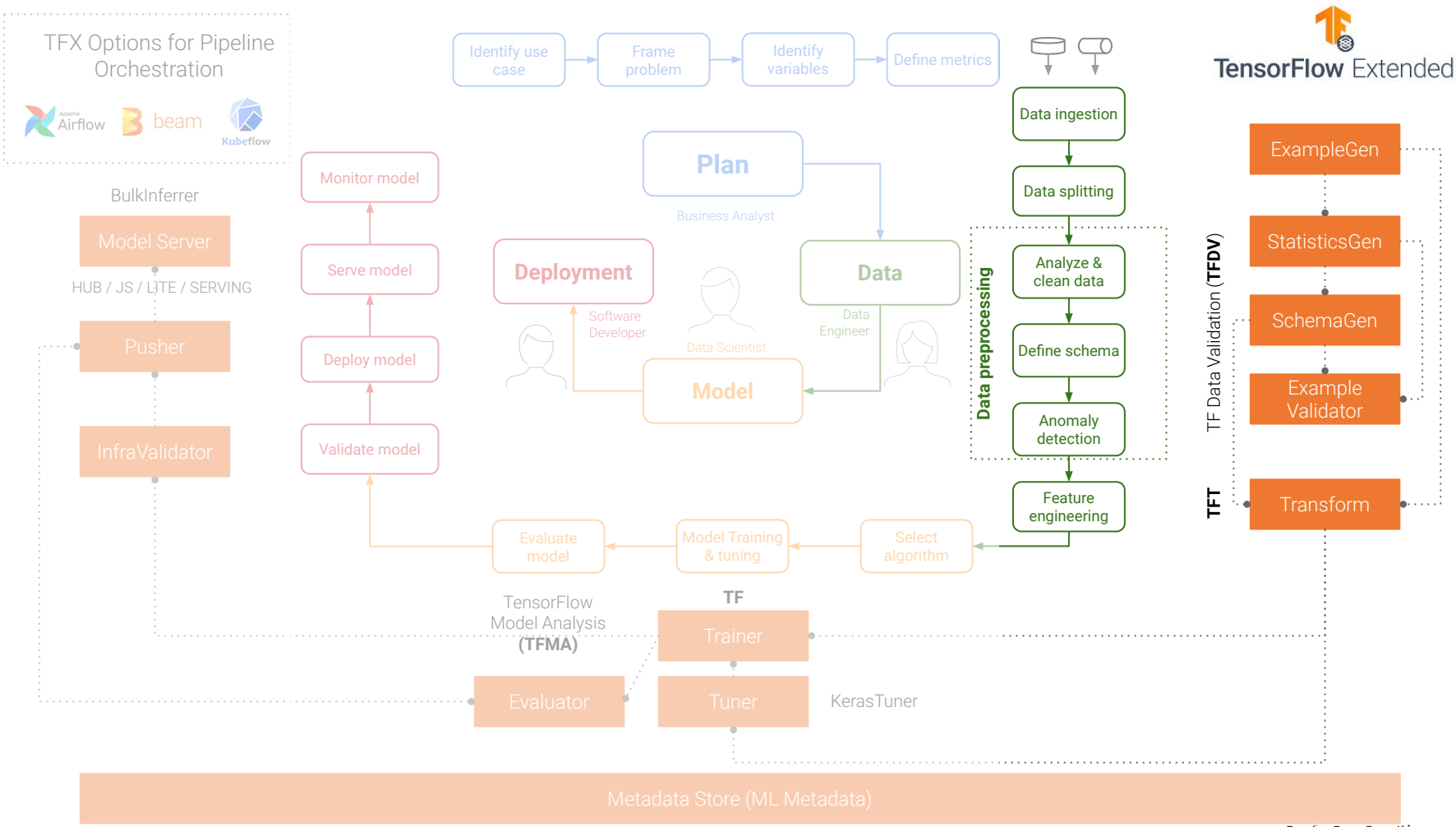
Source: TensorFlow (2021)

TFX 1.0 (19.05.21)

- Enterprise-grade support
- Security patches and select bug fixes for up to three years
- Guaranteed API & Artifact backward compatibility

Source: Google (2021)


TFX Options for Pipeline Orchestration
Apache Airflow
beam
Kubeflow
TensorFlow Extended
Identify use case
Frame problem
Identify variables
Define metrics
Plan
Business Analyst
Data ingestion
Data splitting
Analyze & clean data
Define schema
Anomaly detection
Feature engineering
Data preprocessing
Data
Data Engineer
Deployment
Software Developer
Data Scientist
Model
Monitor model
Serve model
Deploy model
Validate model
BulkInferrer
Model Server
HUB / JS / LITE / SERVING
Pusher
InfraValidator
Evaluate model
Model Training & tuning
Select algorithm
TensorFlow Model Analysis (TFMA)
TF
Trainer
Evaluator
Tuner
KerasTuner
ExampleGen
StatisticsGen
SchemaGen
Example Validator
TF Data Validation (TFDV)
TFT
Transform
Metadata Store: ML Metadata (MLMD)


TFX Options for Pipeline Orchestration
Apache Airflow
beam
Kubeflow
Identify use case
Frame problem
Identify variables
Define metrics
TensorFlow Extended
Data ingestion
Data splitting
Analyze & clean data
Define schema
Anomaly detection
Feature engineering
Data preprocessing
ExampleGen
StatisticsGen
SchemaGen
Example Validator
Transform
TF Data Validation (TFDV)
TFT
Plan
Business Analyst
Data
Data Engineer
Deployment
Software Developer
Data Scientist
Model
Monitor model
Serve model
Deploy model
Validate model
BulkInferrer
Model Server
HUB / JS / LITE / SERVING
Pusher
InfraValidator
Evaluate model
Model Training & tuning
Select algorithm
TensorFlow Model Analysis (TFMA)
TF
Trainer
Tuner
KerasTuner
Evaluator
Metadata Store (ML Metadata)

TFX Options for Pipeline Orchestration

Apache Airflow · beam · Kubeflow

BulkInferrer

Model Server

HUB / JS / LITE / SERVING

Pusher

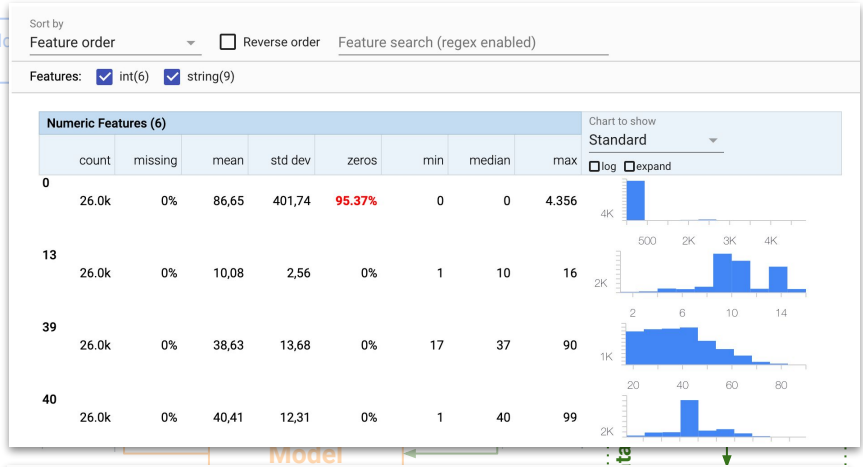
InfraValidator

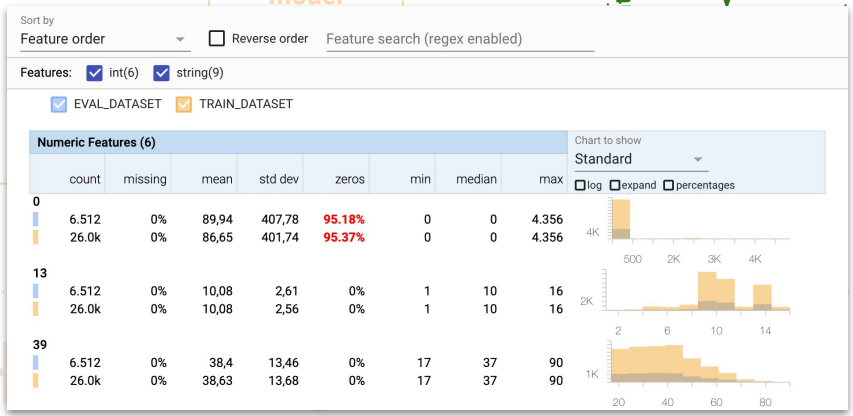
Monitor model

Serve model

Deploy model

Validate model





TensorFlow Extended

ExampleGen

StatisticsGen

SchemaGen

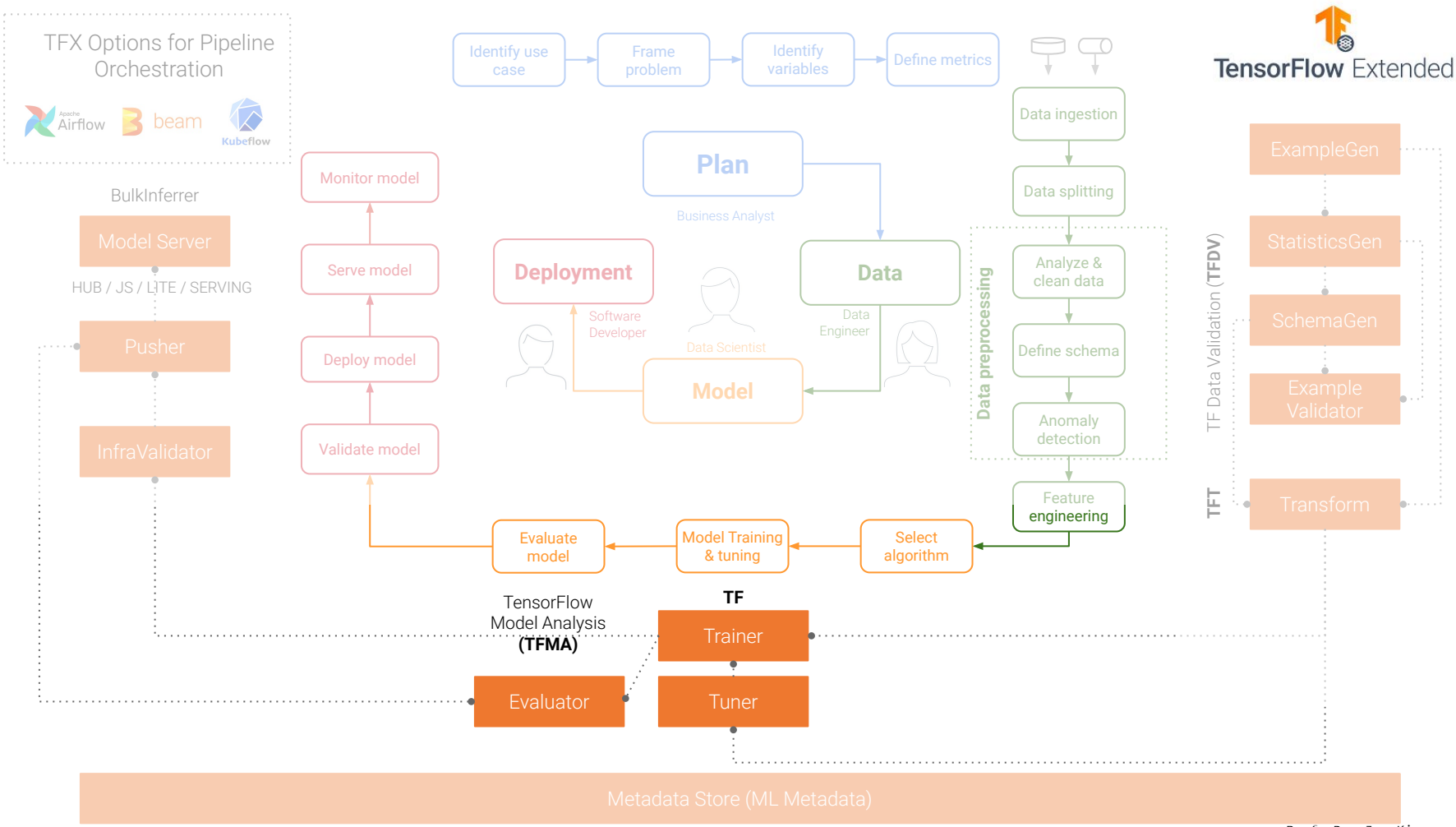
Example Validator

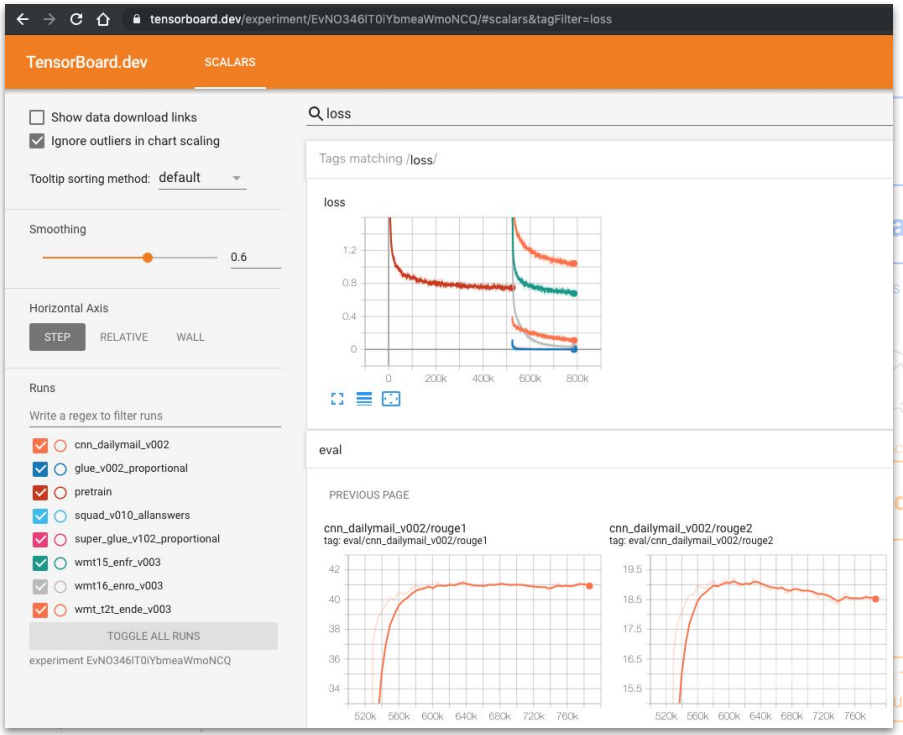
TF Data Validation (TFDV)

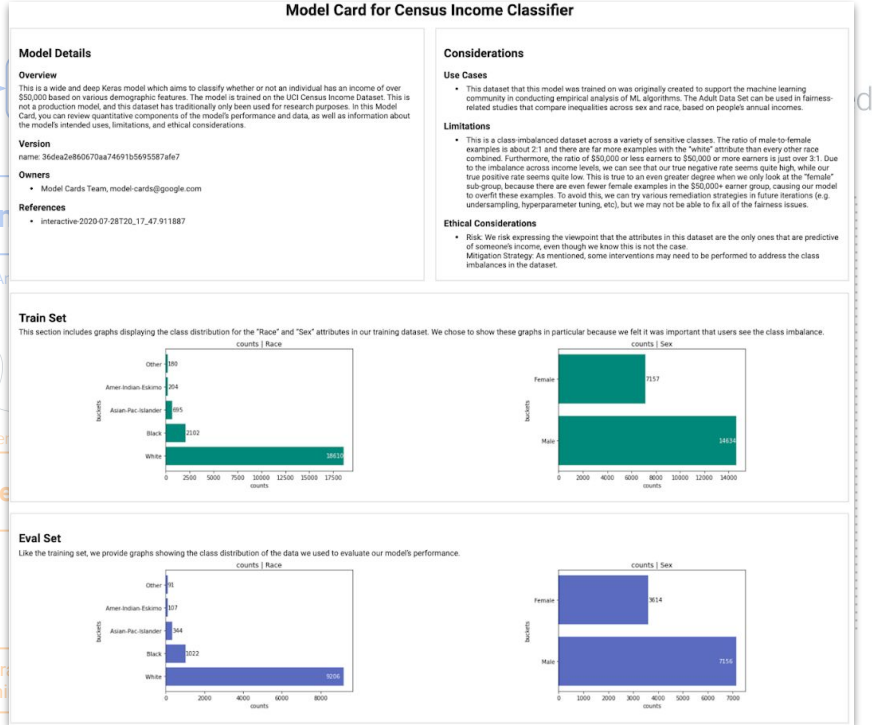
TFT

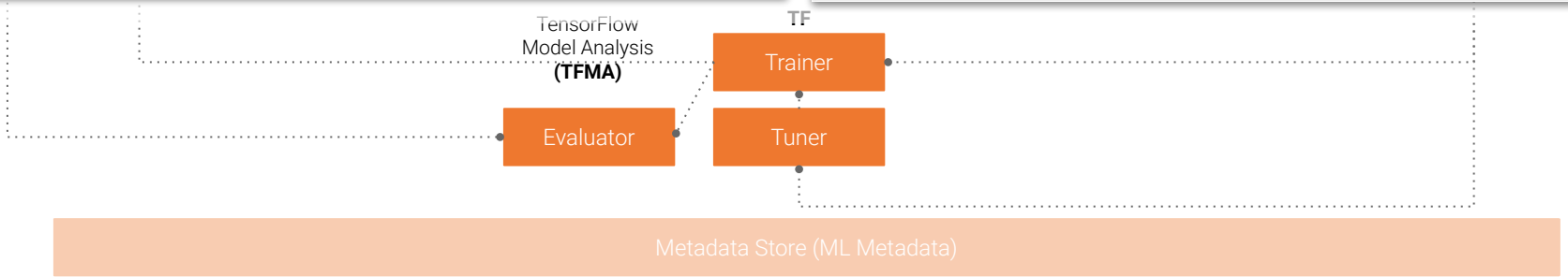
Transform

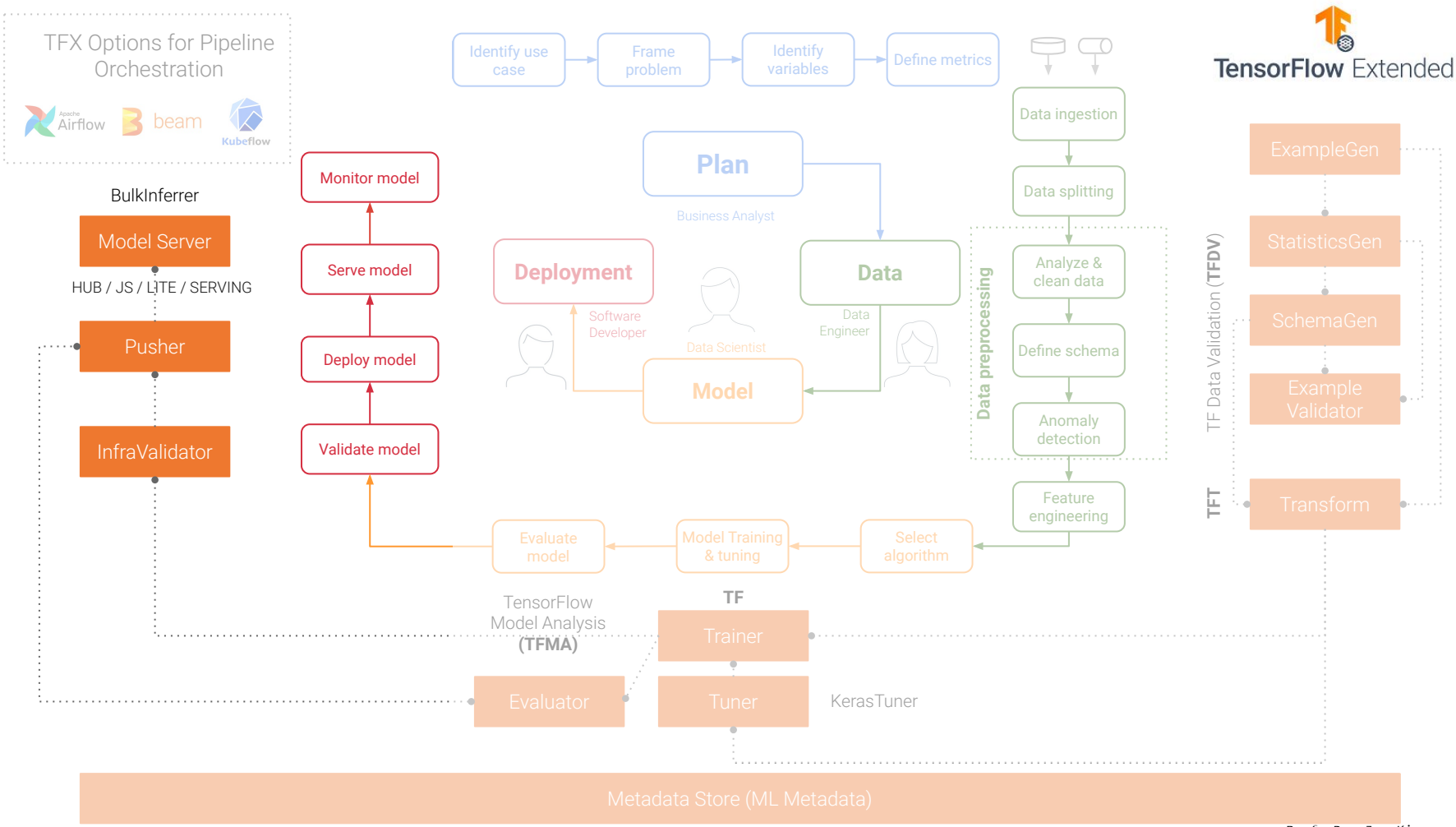
Metadata Store (ML Metadata)


TFX Options for Pipeline Orchestration
Apache Airflow
beam
Kubeflow
TensorFlow Extended
Identify use case
Frame problem
Identify variables
Define metrics
Plan
Business Analyst
Data
Data Engineer
Deployment
Software Developer
Data Scientist
Model
Data ingestion
Data splitting
Analyze & clean data
Define schema
Anomaly detection
Feature engineering
Data preprocessing
ExampleGen
StatisticsGen
SchemaGen
Example Validator
Transform
TF Data Validation (TFDV)
TFT
Select algorithm
Model Training & tuning
Evaluate model
Monitor model
Serve model
Deploy model
Validate model
BulkInferrer
Model Server
HUB / JS / LITE / SERVING
Pusher
InfraValidator
TensorFlow Model Analysis (TFMA)
TF
Trainer
Tuner
Evaluator
Metadata Store (ML Metadata)


tensorboard.dev/experiment/EvNO346IT0iYbmeaWmoNCQ/#scalars&tagFilter=loss
TensorBoard.dev
SCALARS
Show data download links
Ignore outliers in chart scaling
Tooltip sorting method: default
Smoothing
0.6
Horizontal Axis
STEP
RELATIVE
WALL
Runs
Write a regex to filter runs
cnn_dailymail_v002
glue_v002_proportional
pretrain
squad_v010_allanswers
super_glue_v102_proportional
wmt15_enfr_v003
wmt16_enro_v003
wmt_t2t_ende_v003
TOGGLE ALL RUNS
experiment EvNO346IT0iYbmeaWmoNCQ
loss
Tags matching /loss/
loss
eval
PREVIOUS PAGE
cnn_dailymail_v002/rouge1
tag: eval/cnn_dailymail_v002/rouge1
cnn_dailymail_v002/rouge2
tag: eval/cnn_dailymail_v002/rouge2


Model Card for Census Income Classifier
Model Details
Overview
Version
Owners
References
Considerations
Use Cases
Limitations
Ethical Considerations
Train Set
This section includes graphs displaying the class distribution for the "Race" and "Sex" attributes in our training dataset. We chose to show these graphs in particular because we felt it was important that users see the class imbalance.
counts | Race
counts | Sex
Eval Set
Like the training set, we provide graphs showing the class distribution of the data we used to evaluate our model's performance.
counts | Race
counts | Sex


TensorFlow
Model Analysis
(TFMA)
TF
Trainer
Evaluator
Tuner
Metadata Store (ML Metadata)


TFX Options for Pipeline Orchestration
Apache Airflow
beam
Kubeflow
TensorFlow Extended
Identify use case
Frame problem
Identify variables
Define metrics
Data ingestion
Data splitting
Analyze & clean data
Define schema
Anomaly detection
Feature engineering
Data preprocessing
ExampleGen
StatisticsGen
SchemaGen
Example Validator
Transform
TF Data Validation (TFDV)
TFT
Plan
Business Analyst
Data
Data Engineer
Deployment
Software Developer
Data Scientist
Model
Select algorithm
Model Training & tuning
Evaluate model
Monitor model
Serve model
Deploy model
Validate model
BulkInferrer
Model Server
HUB / JS / LITE / SERVING
Pusher
InfraValidator
TensorFlow Model Analysis (TFMA)
TF
Trainer
Tuner
KerasTuner
Evaluator
Metadata Store (ML Metadata)

TFX Options for Pipeline Orchestration

Apache Airflow beam Kubeflow

TensorFlow Extended

TensorFlow Lite is a set of tools that enables on-device machine learning by helping developers run their models on mobile, embedded, and IoT devices.

BulkInferrer

Model Server

HUB / JS / **LITE** / SERVING

Pusher

InfraValidator

ExampleGen

StatisticsGen

SchemaGen

Example Validator

TF Data Validation (**TFDV**)

TFT

Transform

TensorFlow Model Analysis **(TFMA)**

TF

Trainer

Evaluator

Tuner

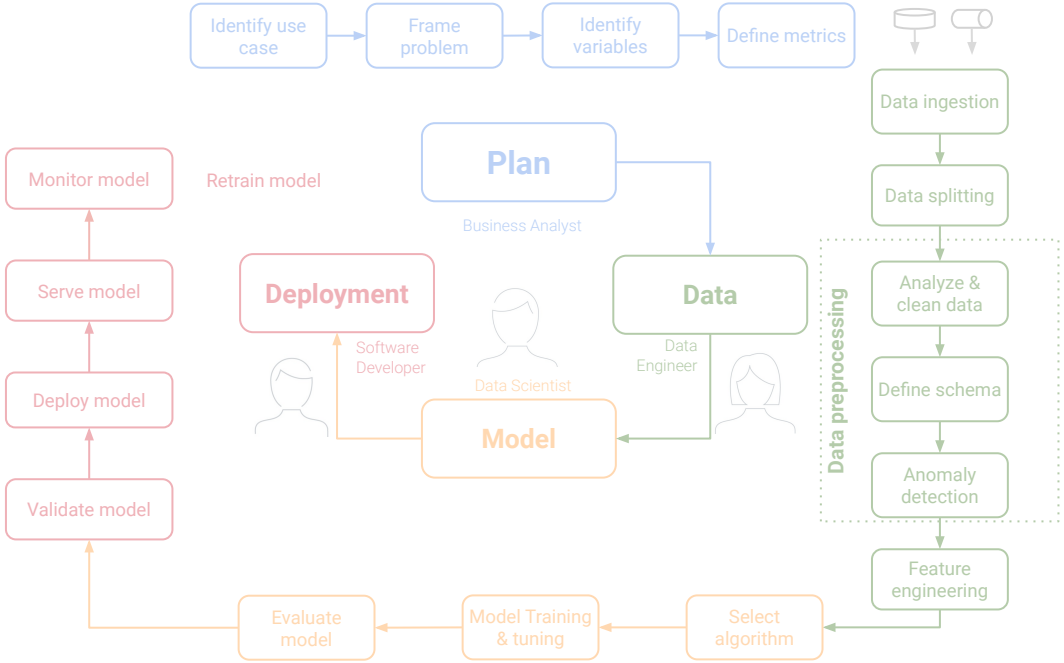
KerasTuner

Metadata Store (ML Metadata)

TFX Options for Pipeline Orchestration

Apache Airflow beam Kubeflow

TensorFlow Extended



Production phase: automate the execution of the ML pipeline based on a schedule or certain triggering conditions.

Data preparation phase: automatically ingest, validate and transform data and provide features to models

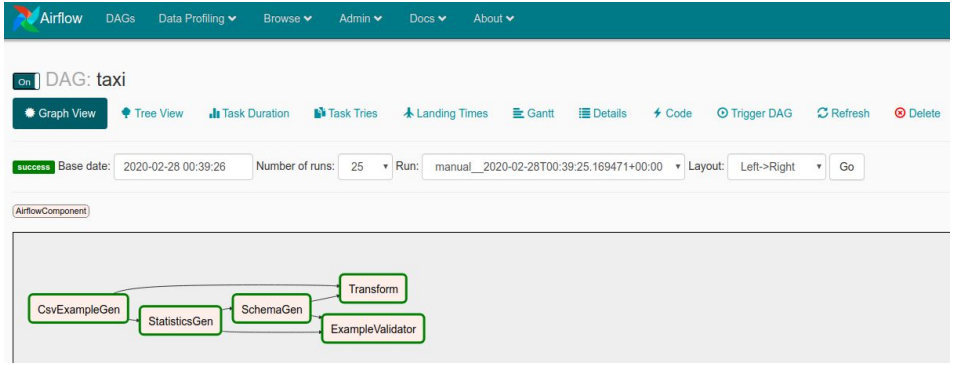
Development phase: run the ML experiment, instead of manually executing each step.

Metadata Store (ML Metadata)

Pipeline orchestration

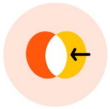
TFX & Apache Airflow

- Programmatically author, schedule and monitor workflows with **Python** code.
- **User interface** to visualize pipelines running in production, monitor progress, and troubleshoot issues.



TFX & Apache Beam

- Provides a framework for running **batch** and **streaming** data processing jobs that run on a variety of runners (Spark, Flink, ...).
- Beam provides an abstraction layer which enables TFX to run on any supported runner without code modifications
- TFX only uses the Beam **Python API**



Unified
Use a single programming model for both batch and streaming use cases.

Extensible
Write and share new SDKs, IO connectors, and transformation libraries.

Portable
Execute pipelines on multiple execution environments.

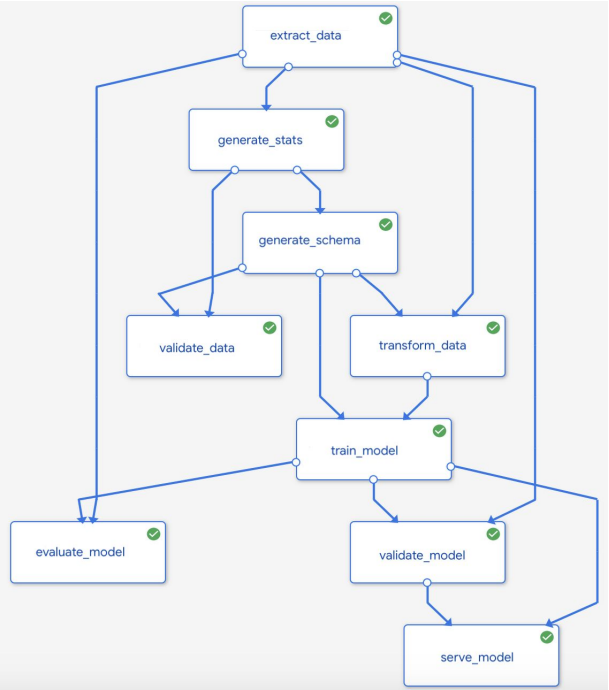
Open Source
Community-based development and support to help evolve your application and use cases.

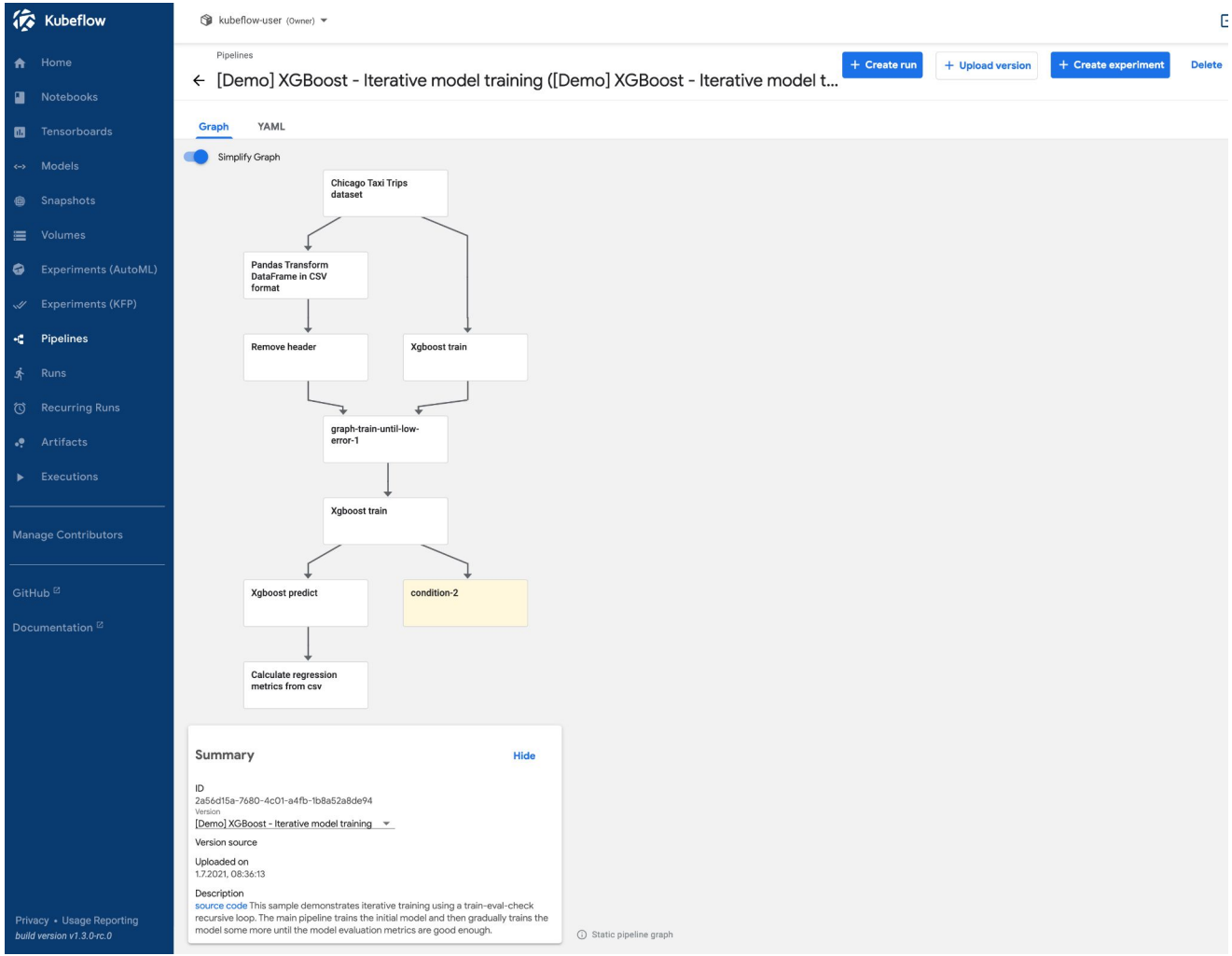
TFX & Kubeflow pipelines

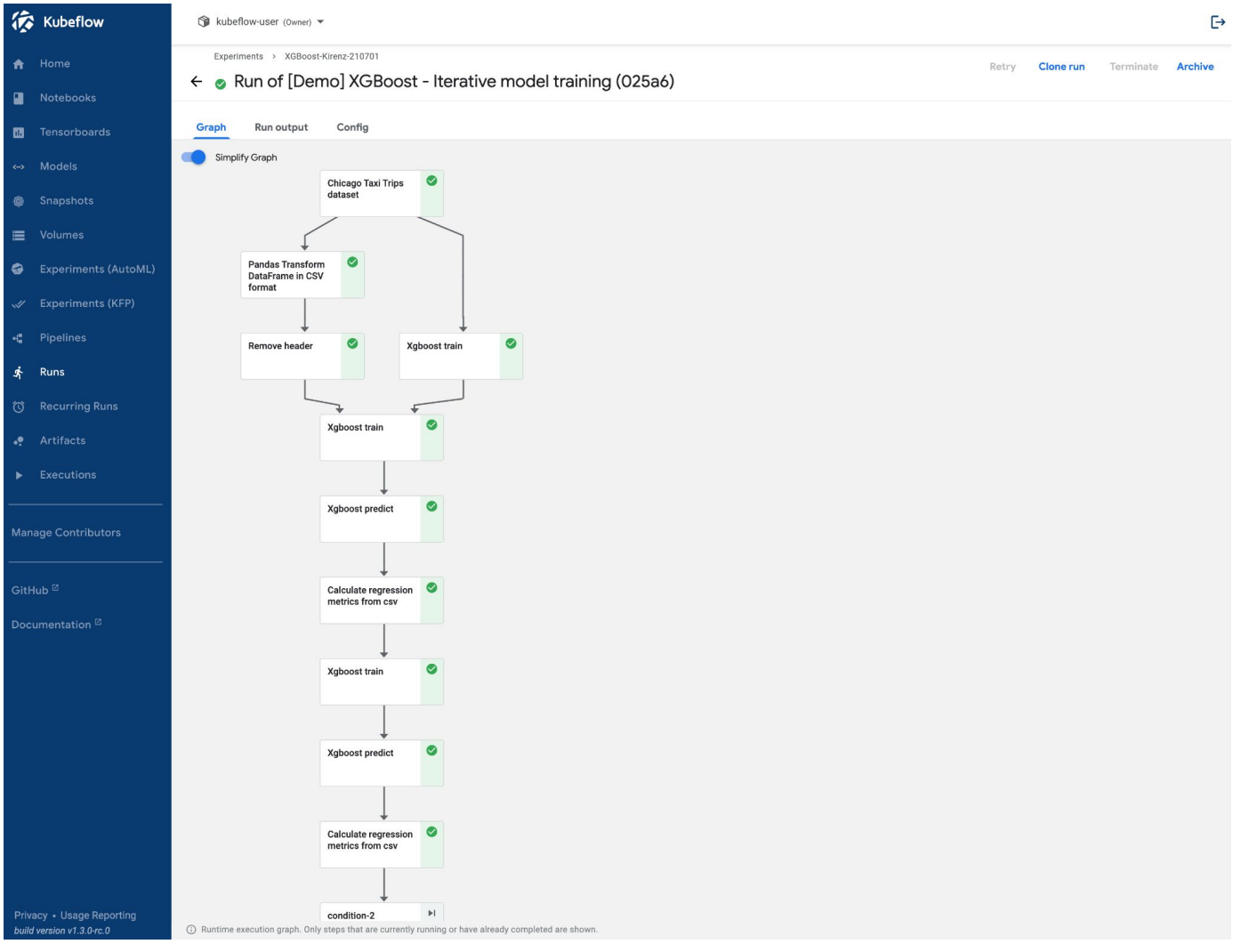
The Kubeflow Pipelines platform consists of:

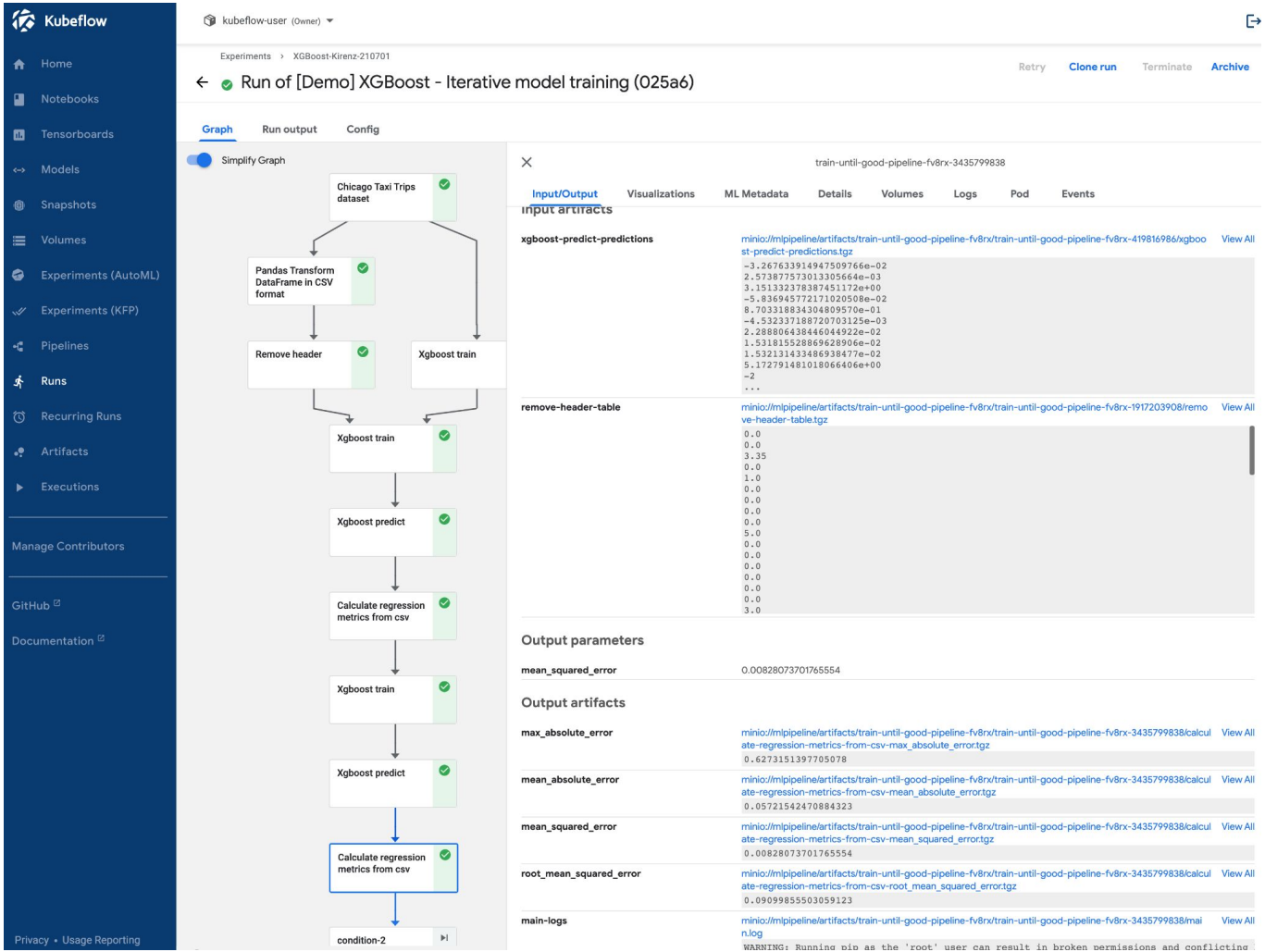
- An engine for scheduling multi-step ML workflows (using **Kubernetes**).
- **User interface** (UI) for managing and tracking experiments, jobs, and runs.
- **Python SDK** for defining and manipulating pipelines and components.
- **Notebooks** for interacting with the system using the SDK

Kubeflow Pipelines is available as a core component of Kubeflow or as a standalone installation.




Kubeflow
kubeflow-user (Owner)
Home
Notebooks
Tensorboards
Models
Snapshots
Volumes
Experiments (AutoML)
Experiments (KFP)
Pipelines
Runs
Recurring Runs
Artifacts
Executions
Manage Contributors
GitHub
Documentation
Privacy • Usage Reporting
build version v1.3.0-rc.0
Pipelines
[Demo] XGBoost - Iterative model training ([Demo] XGBoost - Iterative model t...
+ Create run
+ Upload version
+ Create experiment
Delete
Graph
YAML
Simplify Graph
Chicago Taxi Trips dataset
Pandas Transform DataFrame in CSV format
Remove header
Xgboost train
graph-train-until-low-error-1
Xgboost train
Xgboost predict
condition-2
Calculate regression metrics from csv
Summary
Hide
ID
2a56d15a-7680-4c01-a4fb-1b8a52a8de94
Version
[Demo] XGBoost - Iterative model training
Version source
Uploaded on
1.7.2021, 08:36:13
Description
source code This sample demonstrates iterative training using a train-eval-check recursive loop. The main pipeline trains the initial model and then gradually trains the model some more until the model evaluation metrics are good enough.
Static pipeline graph


Kubeflow
kubeflow-user (Owner)
Home
Notebooks
Tensorboards
Models
Snapshots
Volumes
Experiments (AutoML)
Experiments (KFP)
Pipelines
Runs
Recurring Runs
Artifacts
Executions
Manage Contributors
GitHub
Documentation
Privacy • Usage Reporting
build version v1.3.0-rc.0
Experiments > XGBoost-Kirenz-210701
Run of [Demo] XGBoost - Iterative model training (025a6)
Retry
Clone run
Terminate
Archive
Graph
Run output
Config
Simplify Graph
Chicago Taxi Trips dataset
Pandas Transform DataFrame in CSV format
Remove header
Xgboost train
Xgboost train
Xgboost predict
Calculate regression metrics from csv
Xgboost train
Xgboost predict
Calculate regression metrics from csv
condition-2
Runtime execution graph. Only steps that are currently running or have already completed are shown.


Kubeflow
kubeflow-user (Owner)
Home
Notebooks
Tensorboards
Models
Snapshots
Volumes
Experiments (AutoML)
Experiments (KFP)
Pipelines
Runs
Recurring Runs
Artifacts
Executions
Manage Contributors
GitHub
Documentation
Privacy • Usage Reporting
Experiments › XGBoost-Kirenz-210701
Run of [Demo] XGBoost - Iterative model training (025a6)
Retry
Clone run
Terminate
Archive
Graph
Run output
Config
Simplify Graph
Chicago Taxi Trips dataset
Pandas Transform DataFrame in CSV format
Remove header
Xgboost train
Xgboost train
Xgboost predict
Calculate regression metrics from csv
Xgboost train
Xgboost predict
Calculate regression metrics from csv
condition-2
train-until-good-pipeline-fv8rx-3435799838
Input/Output
Visualizations
ML Metadata
Details
Volumes
Logs
Pod
Events
Input artifacts
xgboost-predict-predictions
minio://mlpipeline/artifacts/train-until-good-pipeline-fv8rx/train-until-good-pipeline-fv8rx-419816986/xgboost-predict-predictions.tgz
View All
-3.2676339149475097666e-02
2.5738775730133056640e-03
3.1513323787387451172e+00
-5.8369457721710205080e-02
8.7033188343048095700e-01
-4.5323371887207031250e-03
2.2888064384460449220e-02
1.5318155288696289060e-02
1.5321314334869384770e-02
5.1727914810180664060e+00
-2
...
remove-header-table
minio://mlpipeline/artifacts/train-until-good-pipeline-fv8rx/train-until-good-pipeline-fv8rx-1917203908/remove-header-table.tgz
View All
0.0
0.0
3.35
0.0
1.0
0.0
0.0
0.0
0.0
5.0
0.0
0.0
0.0
0.0
0.0
0.0
3.0
Output parameters
mean_squared_error
0.00828073701765554
Output artifacts
max_absolute_error
minio://mlpipeline/artifacts/train-until-good-pipeline-fv8rx/train-until-good-pipeline-fv8rx-3435799838/calculate-regression-metrics-from-csv-max_absolute_error.tgz
View All
0.6273151397705078
mean_absolute_error
minio://mlpipeline/artifacts/train-until-good-pipeline-fv8rx/train-until-good-pipeline-fv8rx-3435799838/calculate-regression-metrics-from-csv-mean_absolute_error.tgz
View All
0.05721542470884323
mean_squared_error
minio://mlpipeline/artifacts/train-until-good-pipeline-fv8rx/train-until-good-pipeline-fv8rx-3435799838/calculate-regression-metrics-from-csv-mean_squared_error.tgz
View All
0.00828073701765554
root_mean_squared_error
minio://mlpipeline/artifacts/train-until-good-pipeline-fv8rx/train-until-good-pipeline-fv8rx-3435799838/calculate-regression-metrics-from-csv-root_mean_squared_error.tgz
View All
0.09099855503059123
main-logs
minio://mlpipeline/artifacts/train-until-good-pipeline-fv8rx/train-until-good-pipeline-fv8rx-3435799838/main.log
View All
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting

- Kubeflow on AWS
- Kubeflow on Azure
- Kubeflow on GCP
- Kubeflow on IBM Cloud
- Kubeflow Operator
- Kubeflow on OpenShift

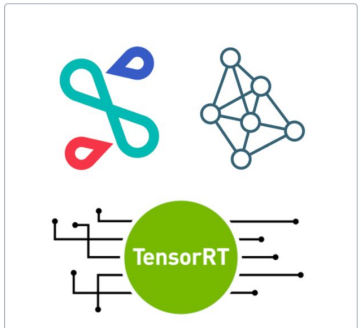
KubeFlow

Google TensorFlow Extended kubernetes Kubeflow

ML toolkit for Kubernetes

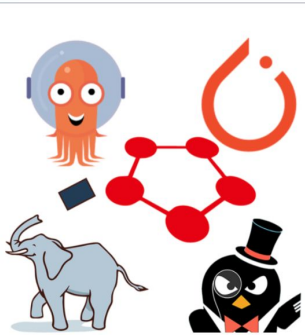
Notebooks

TensorFlow model training



Model serving

Pipelines

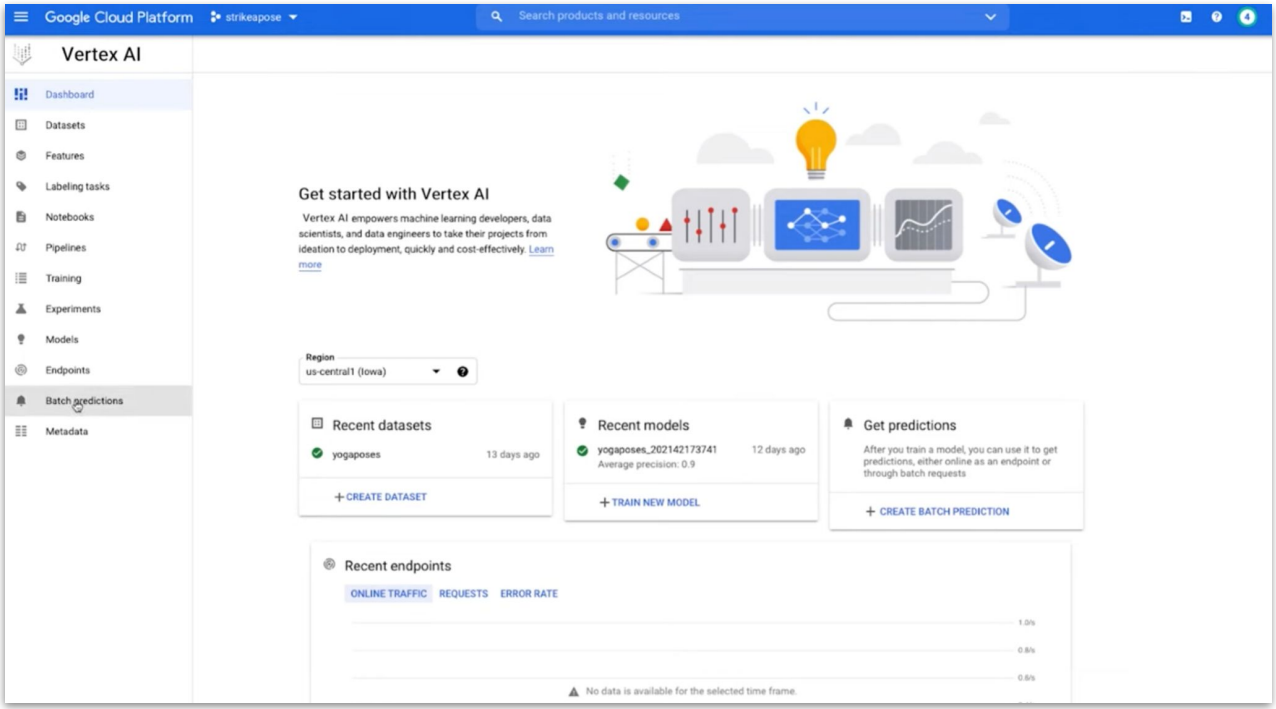


Multi-framework

Google's Vertex AI

Launched in May 2021

Google Cloud



ML Pipelines | wrap-up

By using a ML pipeline, you can:

- Automate your ML process, which lets you regularly retrain, evaluate, and deploy your model.
- Utilize distributed compute resources for processing large datasets and workloads.
- Increase the velocity of experimentation by running a pipeline with different sets of hyperparameters.

To learn more visit the following tutorials @:

https://kirenz.github.io/

MLOps tutorials on how to:

- Install TF and TFX
- Build your first TFX pipeline
- Install Kubeflow
- Build your first Kubeflow pipeline

Jan Kirenz
www.kirenz.com

Source: TensorFlow (2021)