

# Digital Talent Scholarship 2022

## Data Pipelines with TF Data Services

Lead a sprint through the Machine Learning Track

# Agenda

- TFDS Data Pipelines
- Split and Slice API for Dataset in TF
- Exporting your data into the training pipeline
- Parallelization with TFDS

# **Are your students ML-ready?**

# Recap

# Data tidak seperti apa yang terlihat

- Dataset memiliki berbagai kumpulan data dalam format yang berbeda sehingga diperlukan custom code untuk memproses data tersebut.
- Kumpulan data tertentu, perlu kita download terlebih dahulu agar kita yakin berapa banyak data yang disimpan.
- Kita perlu mengubah data ke dalam format yang mudah digunakan oleh sebuah model.
- Setiap sumber data dapat berbeda dan memiliki pertimbangan yang berbeda.



# What is TFDS?

Tensorflow Dataset (TFDS) is a collection of datasets ready to use, with TensorFlow or other Python ML frameworks, such as Jax. All datasets are exposed as `tf.data.Datasets`, enabling easy-to-use and high-performance input pipelines. To get started see the guide and our list of datasets.



# Example Code

```
import tensorflow as tf
import tensorflow_datasets as tfds

# Construct a tf.data.Dataset
ds = tfds.load('mnist', split='train', shuffle_files=True)

# Build your input pipeline
ds = ds.shuffle(1024).batch(32).prefetch(tf.data.AUTOTUNE)
for example in ds.take(1):
    image, label = example["image"], example["label"]
```

# Working with Tensorflow Datasets



## Demo : TFDS Dataset

**Demo : TFDS Rock, Paper, Scissors**

# Transfer Learning and Splits API

# Splits Tensorflow

All TFDS datasets expose various data splits (e.g. 'train', 'test') which can be explored in the catalog.

In addition of the "official" dataset splits, TFDS allow to select slice(s) of split(s) and various combinations.

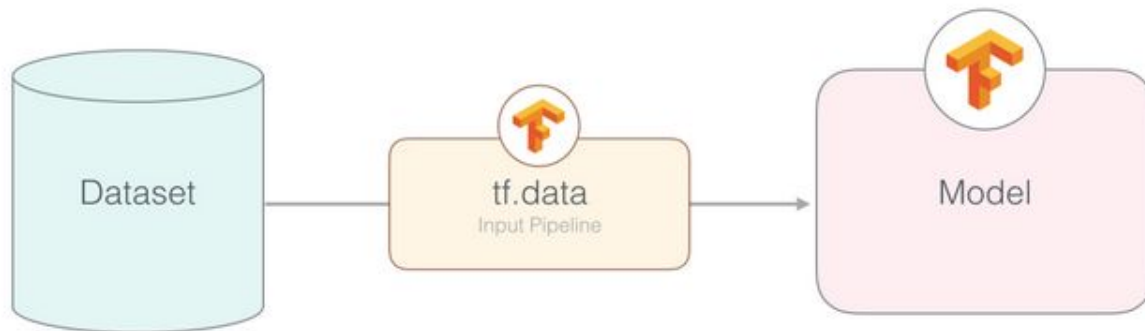
# What is data pipeline in TensorFlow?

The tf. data API enables you to build complex input pipelines from simple, reusable pieces. For example, the pipeline for an image model might aggregate data from files in a distributed file system, apply random perturbations to each image, and merge randomly selected images into a batch for training.



tf.data

# tf.data: Build TensorFlow input pipelines



Demo : [tf.data Tensorflow Input Pipelines](#)



tf.data



# TFDS Parallelization

# Model Parallelism

## Parallization with TFDS

# Q & A

# Thank You