

Machine Learning for Real Time Streaming Data with Kafka and TensorFlow

Yong Tang

Maintainer & SIG IO Lead, TensorFlow

GitHub: yongtang

Stream Data Processing & Machine Learning

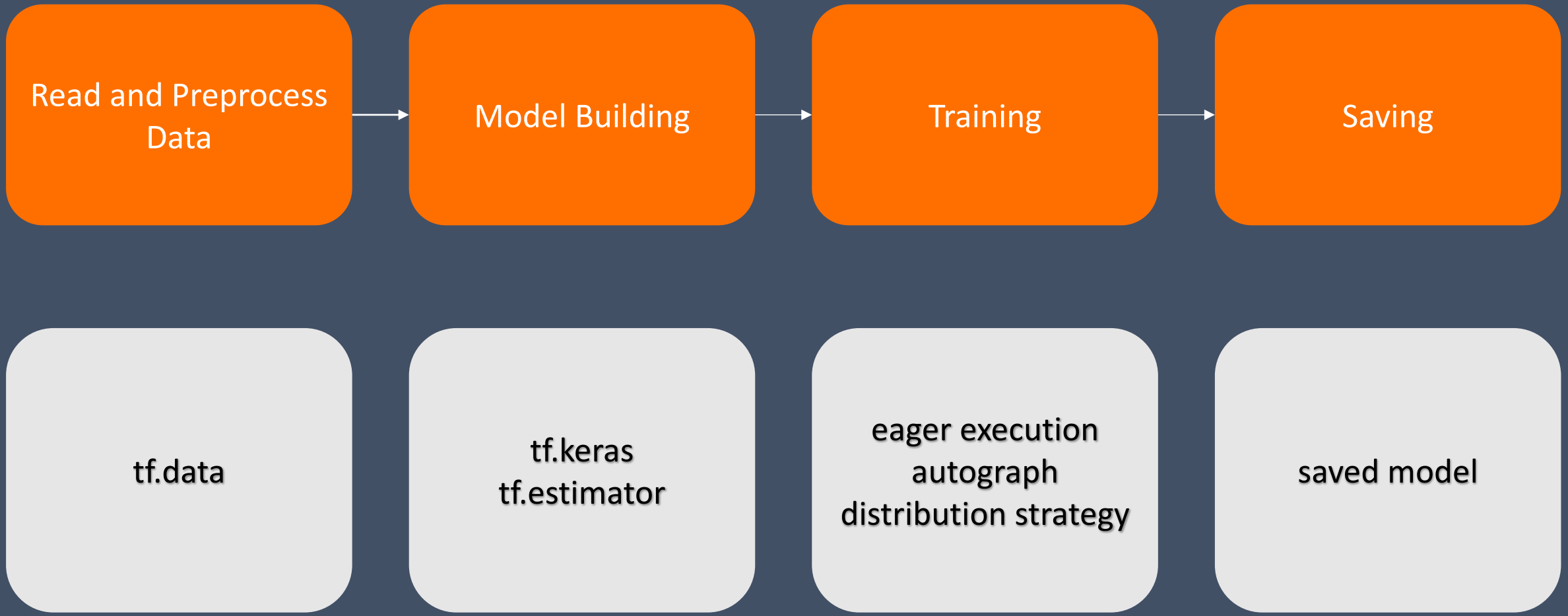
Apache Kafka & Event Driven Microservices

TensorFlow & 2.0

- Most popular machine learning framework
- 1.x
- 2.0
  - Keras high level API
  - Eager execution
  - Data processing: tf.data
  - Distribution Strategy

TensorFlow

Training Workflow (TensorFlow 2.0)



```
# <= read data into tf.data
dataset = …. , test_dataset = ….

# <= build keras model
model = tf.keras.models.Sequential([
    tf.keras.layers.Flatten(input_shape=(28, 28)),
    tf.keras.layers.Dense(512, activation=tf.nn.relu),
    tf.keras.layers.Dropout(0.2),
    tf.keras.layers.Dense(10, activation=tf.nn.softmax) ])
model.compile(optimizer='adam', loss='sparse_categorical_crossentropy',
                        metrics=['accuracy'])

# <= train the model
model.fit(dataset, epochs=5)

# <= inference
predictions = model.predict(test_dataset, callbacks=[callback])
```

Steam Data + Machine Learning: Challenges

Streaming & big data

- Hadoop/Zookeeper
- Kafka, Spark, Flink
- Java/Scala
- Parquet/Avro/Arrow

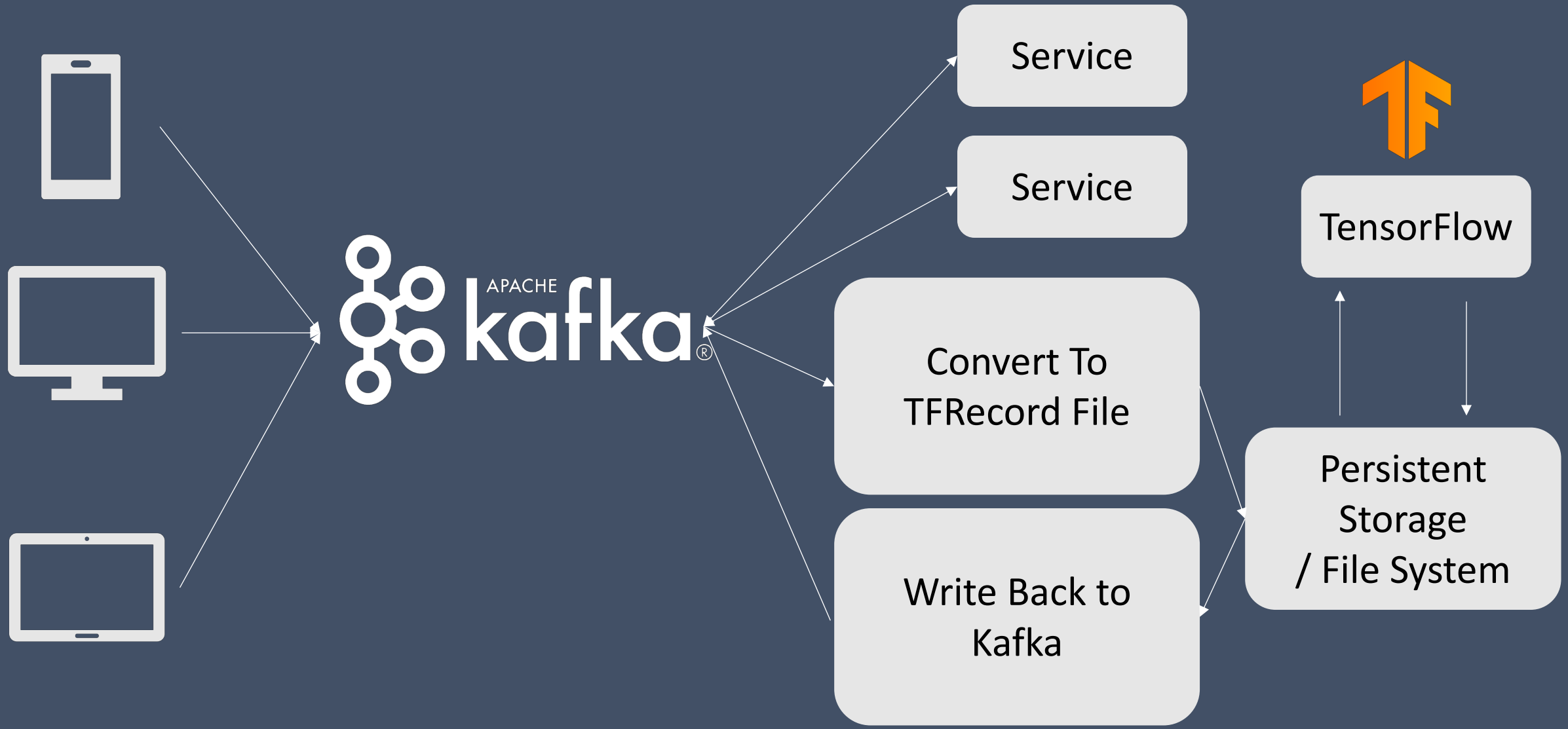
Machine leaning

- CPU/GPU/TPU
- Python (C++, CUDA)
- Numpy
- TFRecord/CSV/Text/JPEG/PNG

TensorFlow TFRecord Format

- Sequence of binary strings
- Protocol buffer for serialization
- Flexible and simple
- ***ONLY used by TensorFlow***
- *NO ecosystem support like parquet, etc.*

Infrastructure Solution



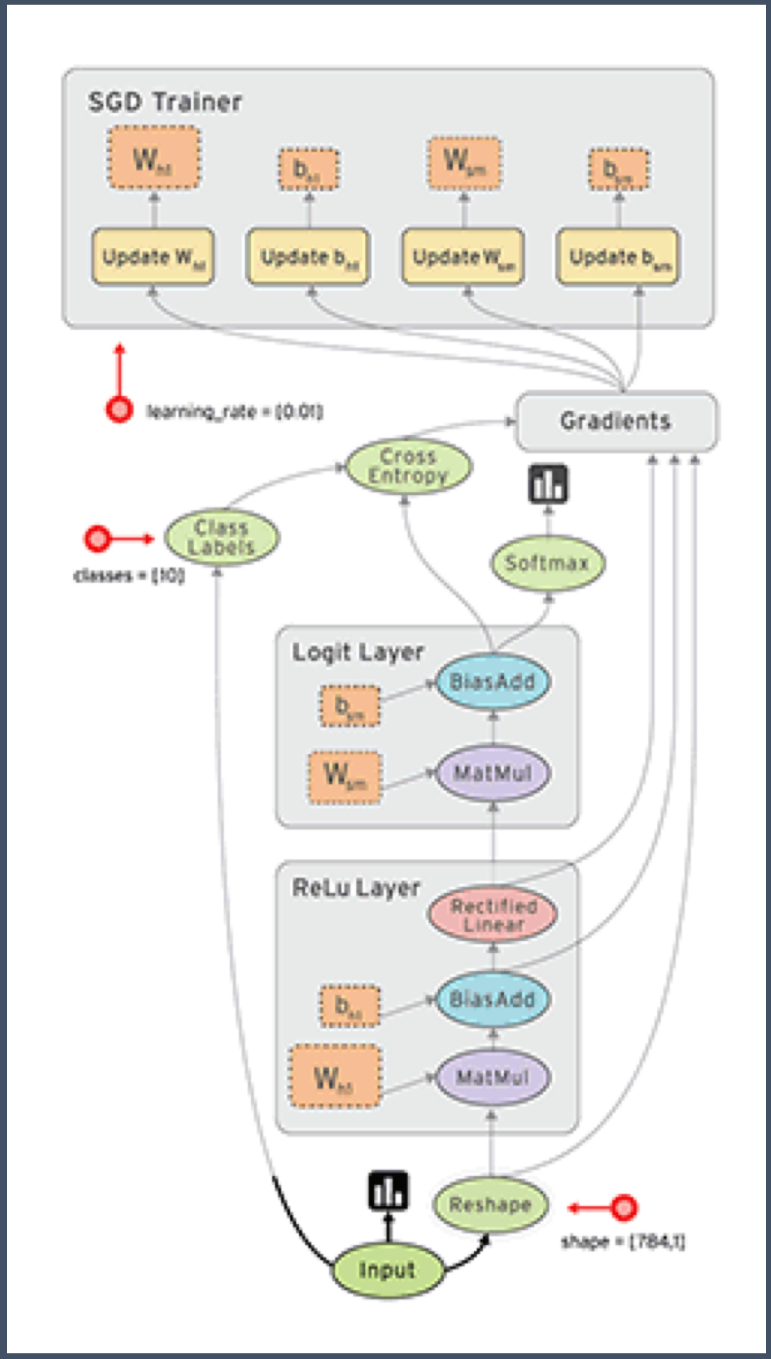
tf.data in TensorFlow

- Part of the graph in TensorFlow
- Support Eager and Graph mode
- Focus on data transformation

tf.data.TFRecordDataset

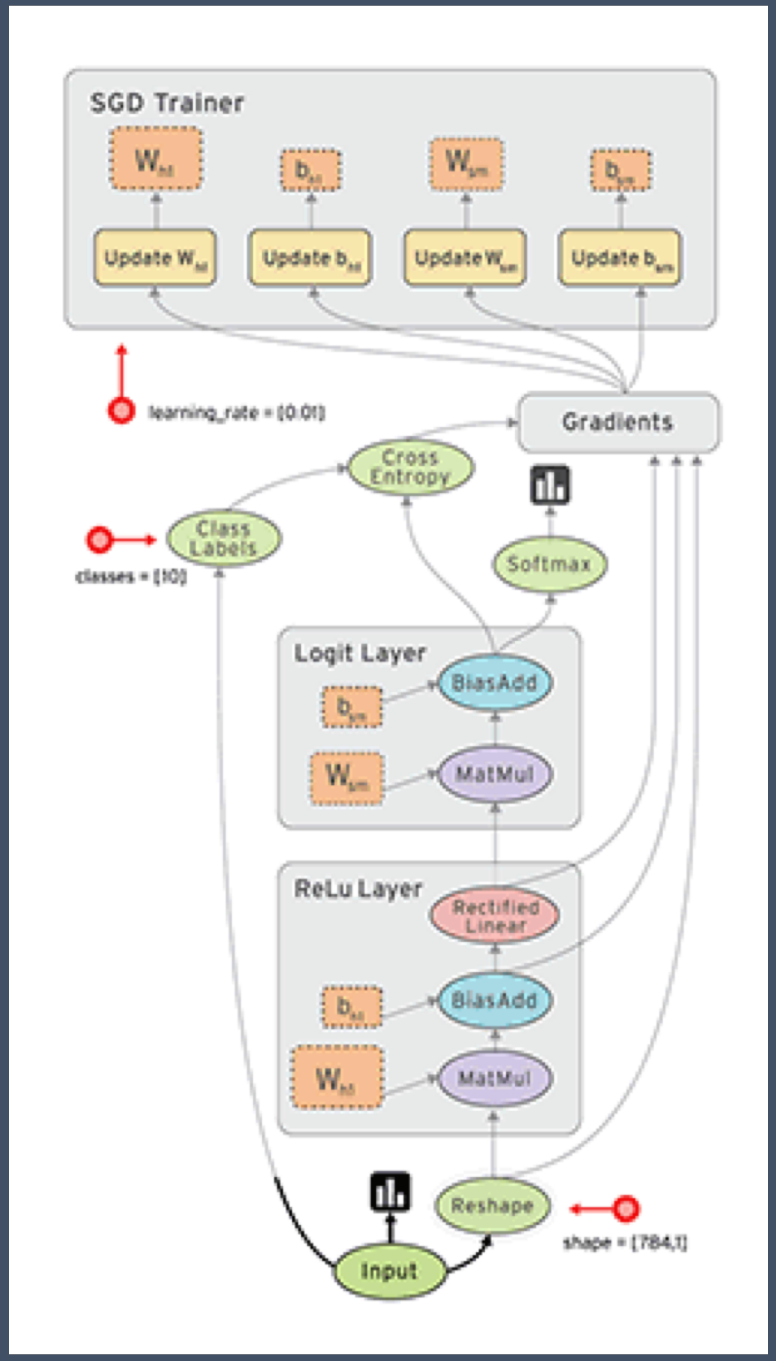
tf.data.TextLineDataset

tf.data.FixedLengthRecordDataset



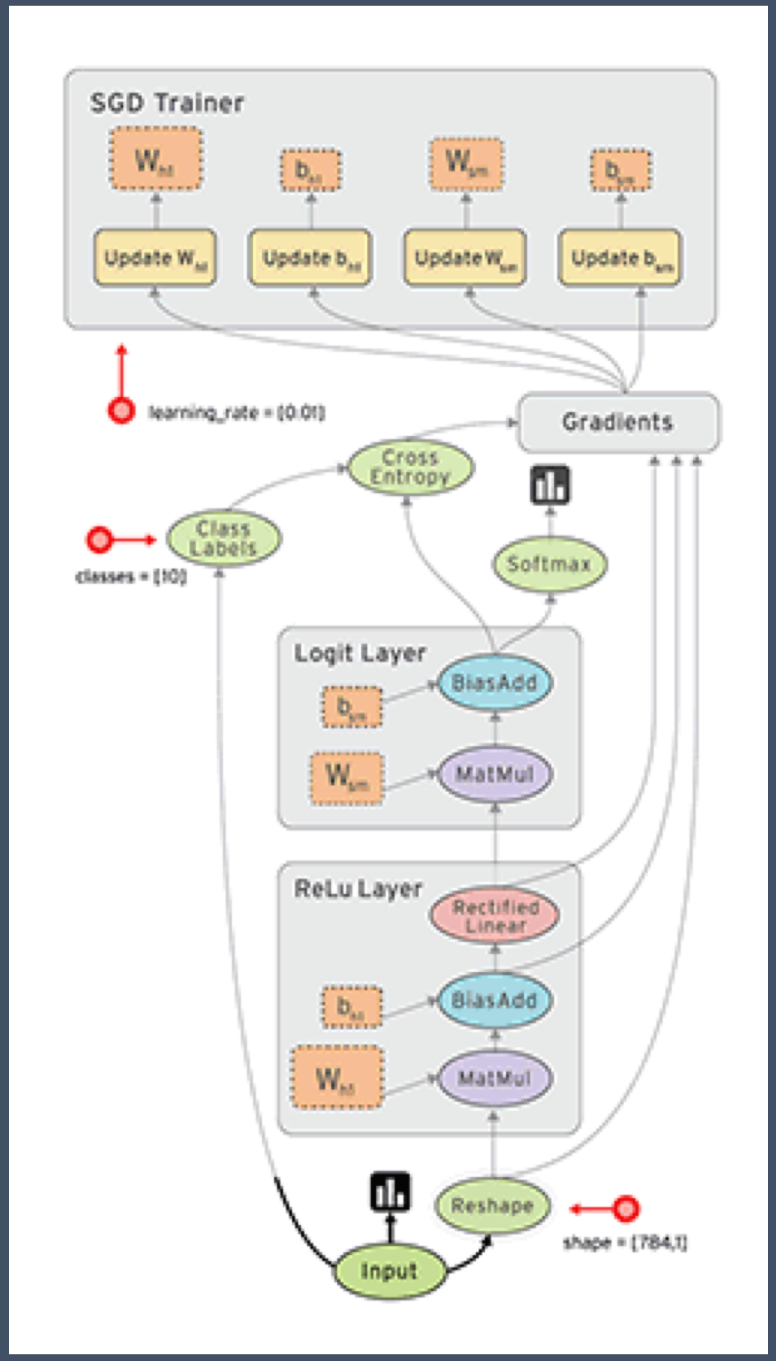
KafkaDataset for TensorFlow

- Subclass of tf.data.Dataset
- Written in C++ (linked with librdkafka)
- Part of the graph in TensorFlow
- Support Eager and Graph mode



KafkaDataset for TensorFlow

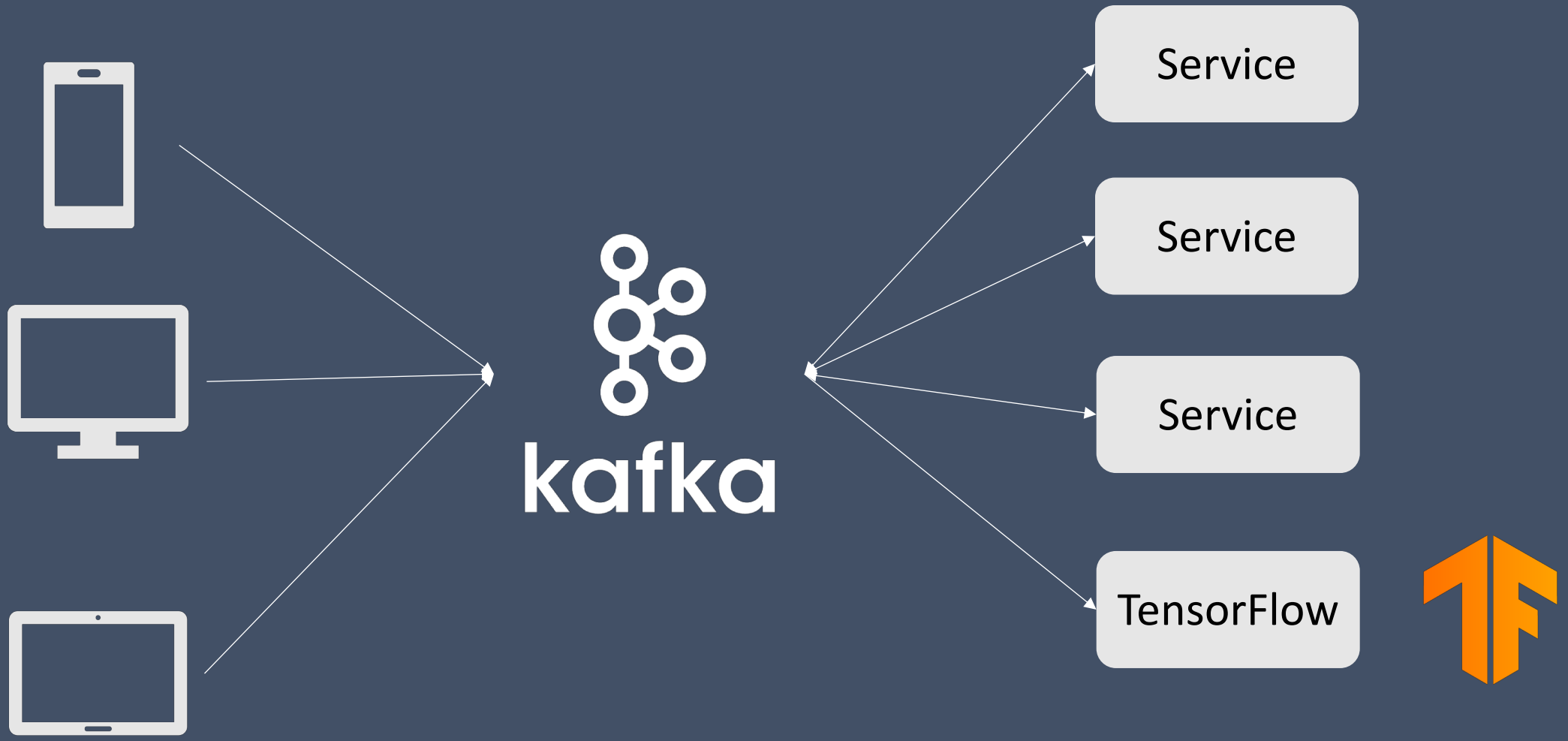
- Part of tensorflow package in TF 1.x
  - tf.contrib.kafka
- Part of tensorflow-io package in TF 2.0
- KafkaDataset (Read)
- KafkaOutputSequence (Write)
- Python and R (other language support possible)



```
$ pip install tensorflow-io

import tensorflow_io.kafka as kafka_io

dataset = kafka_io.KafkaDataset( 'topic', server='localhost', group= '')

# <= pre-process data (if needed), batch/repeat/etc.
dataset = dataset.map(
   lambda x: ……)

# <= build keras model
model = tf.keras.models...
model.compile(...)

# <= train the model
model.fit(dataset, epochs=5)
```

```
# continue with inference...

# build keras callback
class OutputCallback(tf.keras.callbacks.Callback):

    def __init__(self, batch_size, topic, servers):
        self._sequence = kafka_io.KafkaOutputSequence(
            topic=topic, servers=servers)
        self._batch_size = batch_size

    def on_predict_batch_end(self, batch, logs=None):
        ...
        self._sequence.setitem(index, class_names[np.argmax(output)])

# <= inference with callback for streaming input and output
model.predict(
    test_dataset, callbacks=[OutputCallback(32, 'topic', 'localhost')])
```

KafkaDataset for TensorFlow



TensorFlow SIG I/O

- https://github.com/tensorflow/io
- Special Interest Group under TensorFlow org
  - Focus on I/O, streaming, and data formats
- Streaming:
  - Apache Kafka, AWS Kinesis, Google Cloud PubSub
- Memory, caching and serialization
  - Apache Ignite, Apache Arrow, Apache Parquet
- File formats:
  - MNIST, WebP/TIFF/etc, Video, Audio
- Cloud Storage
  - GCS, S3, Azure

TensorFlow

THANK YOU