Databricks ML Workshop

Samir Gupta, Solutions Architect

databricks

Workshop Agenda

Overview of Databricks for Data Science
The Databricks ML Runtime
Koalas - Pandas interface to Spark Dataframes
R on Databricks
DBconnect - use your favourite IDE on Databricks
Mlflow - train, experiment, track and deploy models
Sample notebooks
Q&A

Hardest Part of AI isn't AI, it's Data

"Hidden Technical Debt in Machine Learning Systems," ***Google NIPS 2015***

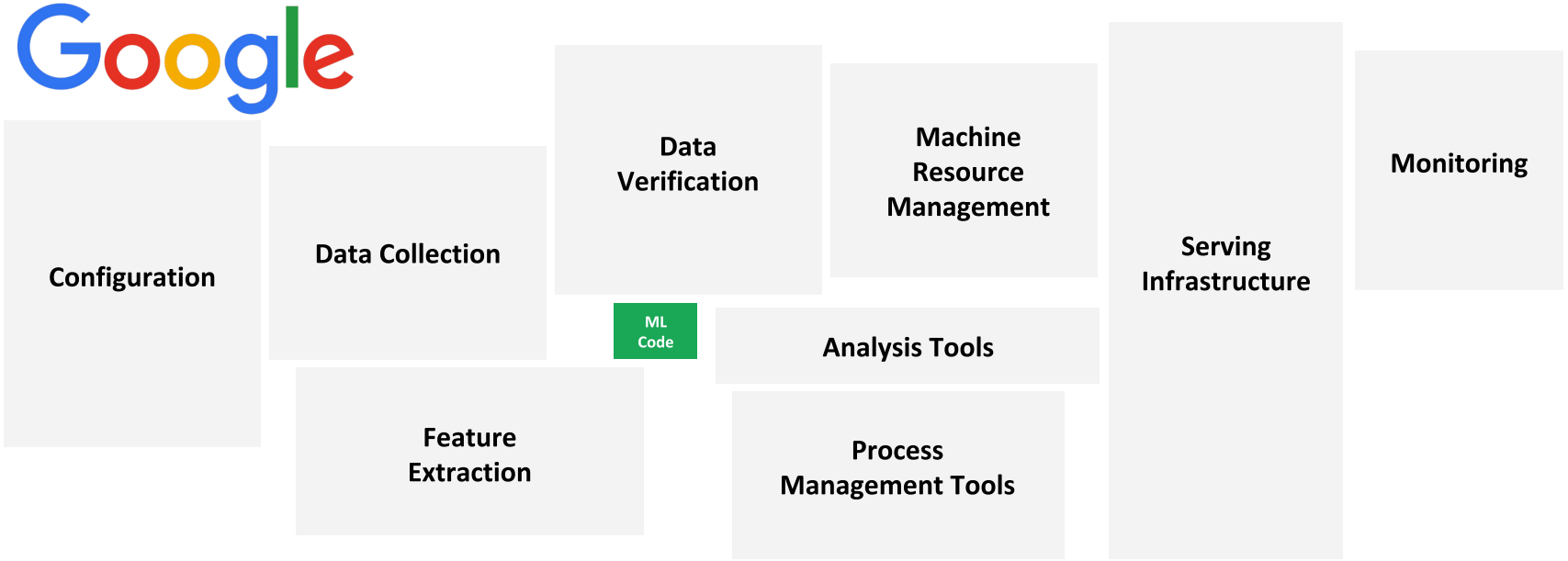
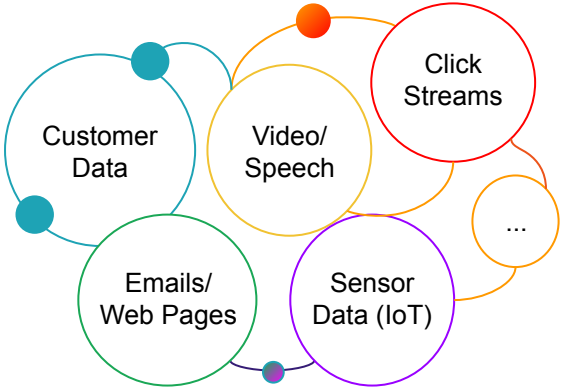


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small green box in the middle. The required surrounding infrastructure is vast and complex.

And…Data & AI Technologies are in Silos

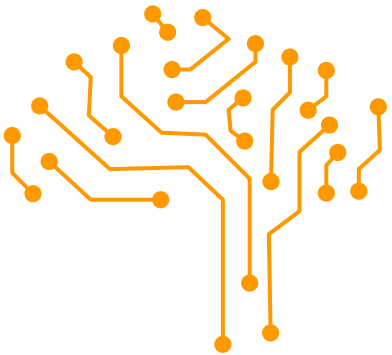


amazon REDSHIFT · S3 · mongoDB · Parquet

kafka · ORACLE · TERADATA

Great for Data, but not AI

X



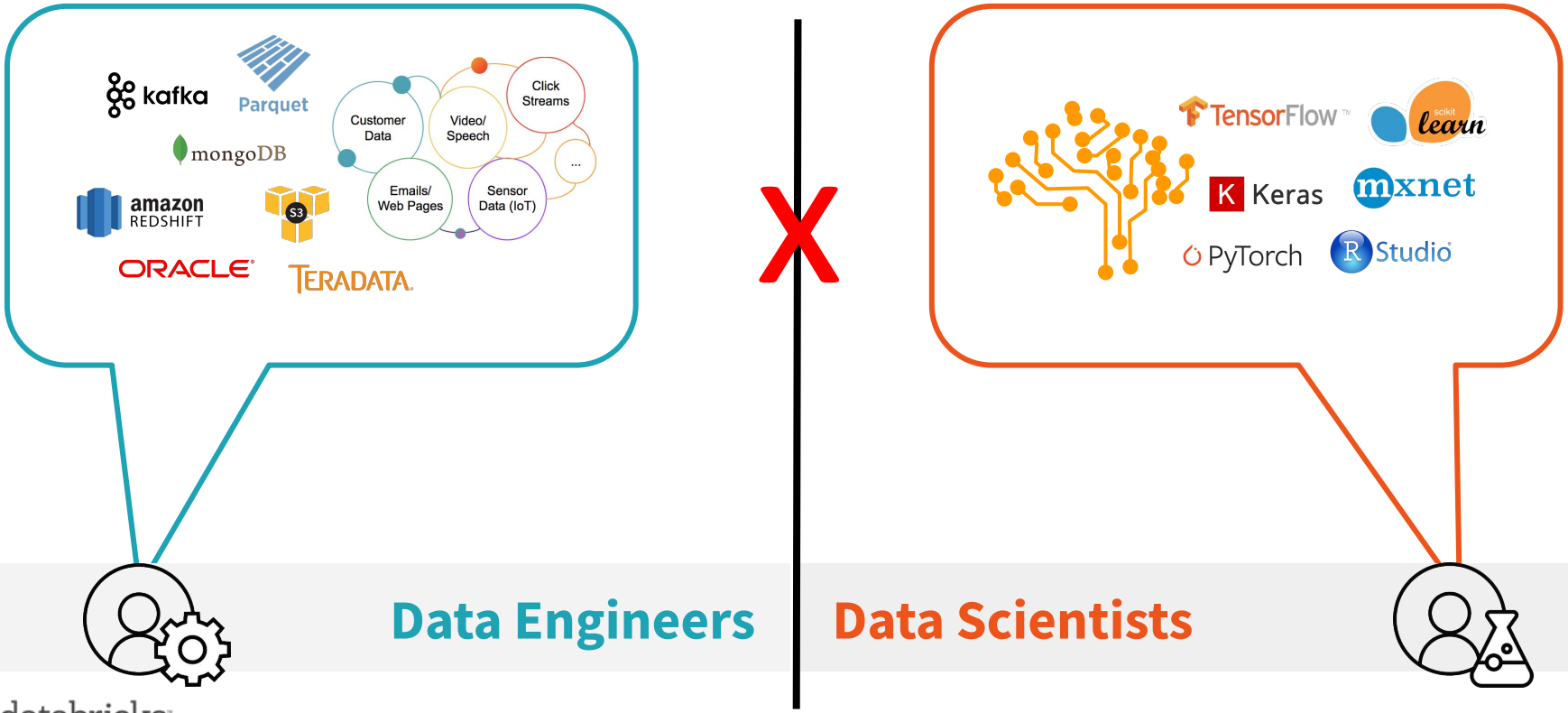
Keras · RStudio · PyTorch

TensorFlow · scikit learn · mxnet

Great for AI, but not for data

databricks

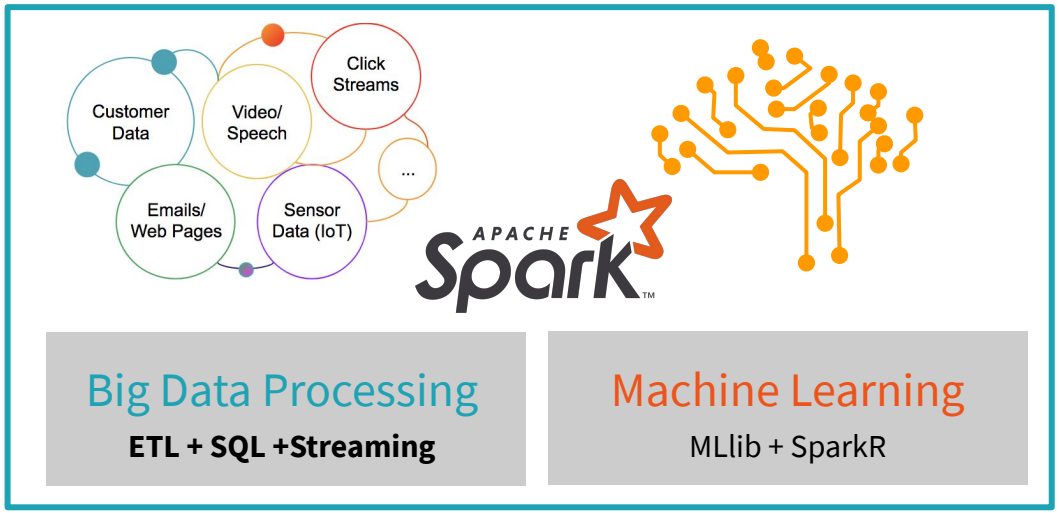
Data & AI People are in Silos



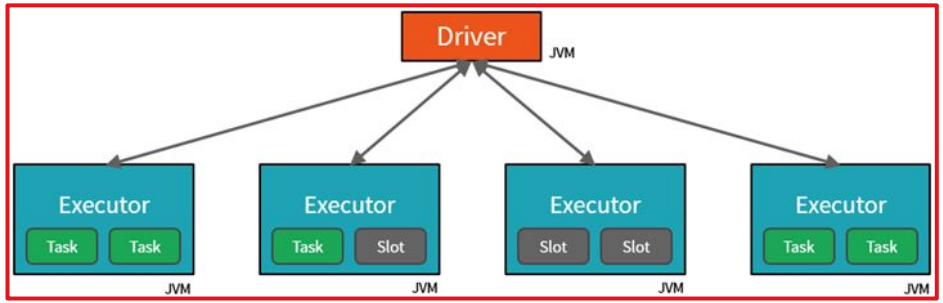
databricks

Apache Spark...

1st Unified Analytics Engine that uniquely combines data & AI technologies...



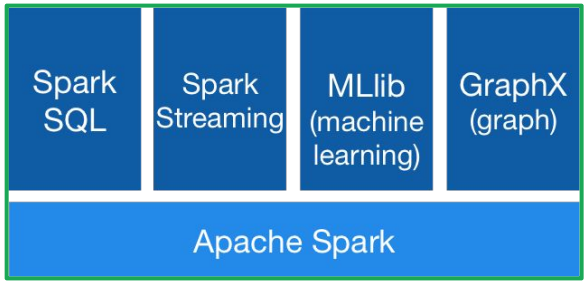

What is Apache Spark?



Fast, distributed processing on **large volumes** of structured and unstructured data

```
df = spark.read.json("logs.json")
df.where("age > 21")
  .select("name.first").show()
```

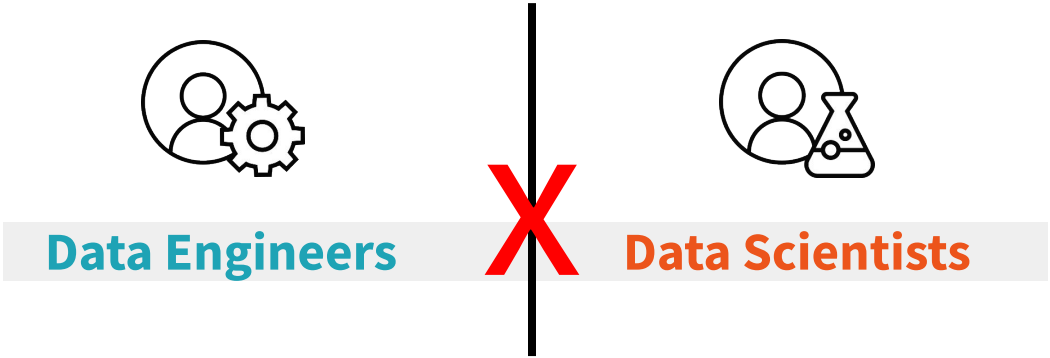
Rich and easy to use API programmable in Python, Scala, R, Java, SQL



Flexible processing for batch, streaming, and data science/ML workloads

Apache Spark...

Was only the first step



Spark Expertise, Data Pipelines, Infrastructure Mgmt...

Unified Analytics Engine

APACHE Spark™

Databricks Unified Analytics Platform

AZURE DATA SOURCES

Blob Storage

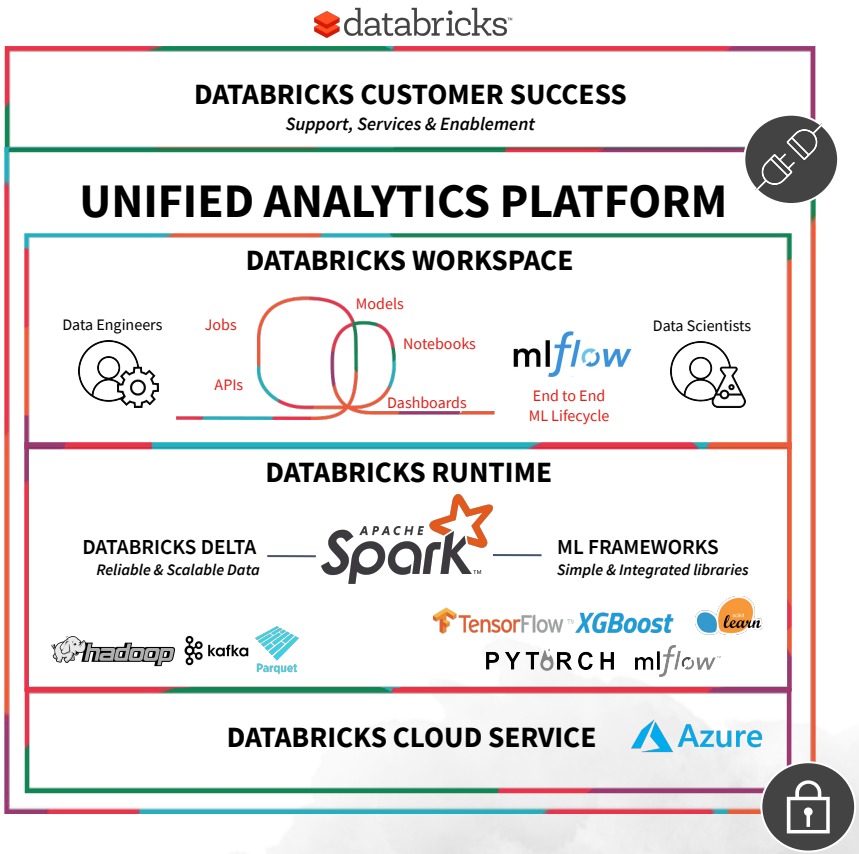
Data Lake Store

SQL Data Warehouse

Cosmos DB

Event Hub

IoT Hub



Microsoft Active Directory

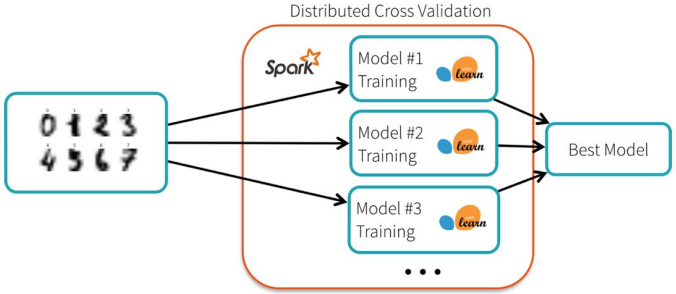
Power BI

Azure Data Factory

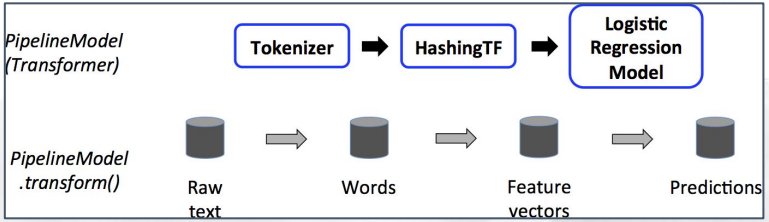
Why Distributed Machine Learning?

1. Speed - to slow to train a model on a single machine
2. Resources - data or models are too big for a single machine

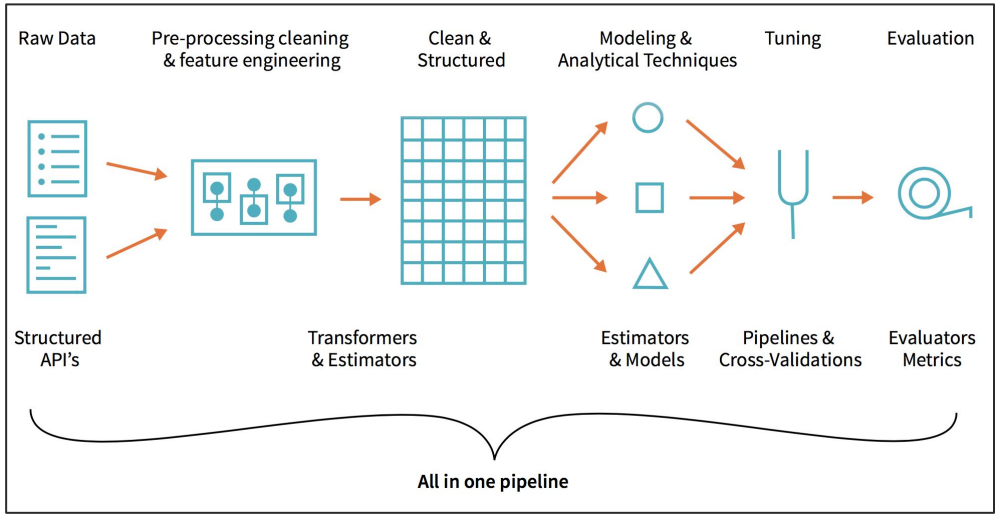
spark-sklearn - train multiple sklearn models on Spark



SparkML - Native Python and Scala API for ML on Spark
Featuring engineering, regression, classification, clustering, and more written for Spark
Uses transformers (.transform()) and estimators (.fit())
Combine many transformers and estimators into a Pipeline



Productionizing Spark Models



Latency Requirement	What to Use
10-100ms	Low latency Sparkless Highly Available Prediction Server
100ms-1m	Spark Structured Streaming
1m-1hr+	Spark Batch Jobs

Later: MLflow + AzureML - End-to-end model lifecycle

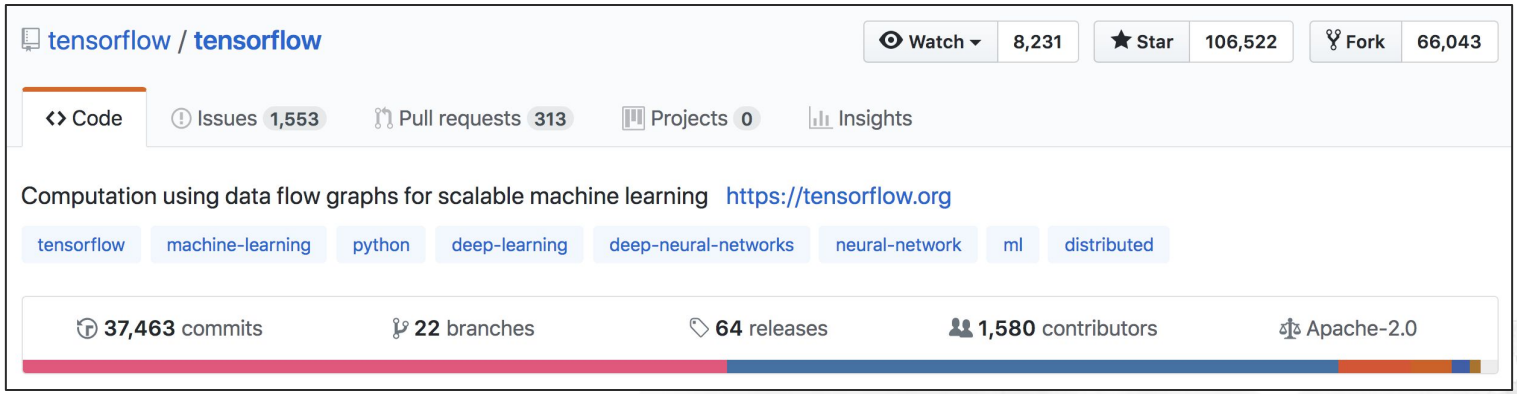
Koalas: Pandas on Spark

Pandas is the *de-facto* Python standard (single-node) for DataFrame data processing
Spark is the de facto standard for big data processing
Koalas let's data scientists:

- Be immediately productive in Spark - with no learning curve if already familiar with pandas
- Have a single codebase that works both with pandas (tests, smaller datasets) and with Spark (large, distributed datasets).

```
from databricks import koalas

data = koalas.read_csv("/databricks-datasets/sai-summit-2019-sf/fire-calls.csv", header=0)
data = data[["CallType", "Neighborhood", "NumAlarms", "OrigPriority", "UnitType", "Delay"]]

# Rename the columns
data.columns = ['a', 'b', 'c', 'd', 'e', 'f']

# Do some operations in place:
data['g'] = data.c * data.f

display(data)
```

We love Spark, but what about Deep Learning?



Databricks ML Runtime

Ready to use clusters with built-in ML Frameworks

including TensorFlow, Keras, Horovod, and more

Horovod Estimator

for **simplified distributed training** on TensorFlow with Horovod using Apache Spark on Databricks

GPU support

Azure GPU (NC) instances pre-loaded with Tesla/CUDA/cuDNN libraries!

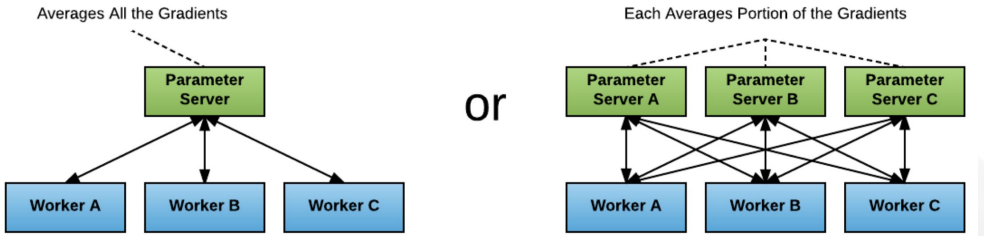
Azure

Options for Deep Learning on Databricks

Option 1: Train on a single node

Option 2: Train on each node of the cluster
- Ex. Distributed hyperparameter tuning & search

Option 3: Distributed Training (ie. HorovodEstimator, Dist-Keras)




```
# Only one line of code change!
optimizer = hvd.DistributedOptimizer(optimizer)
```

R on Databricks

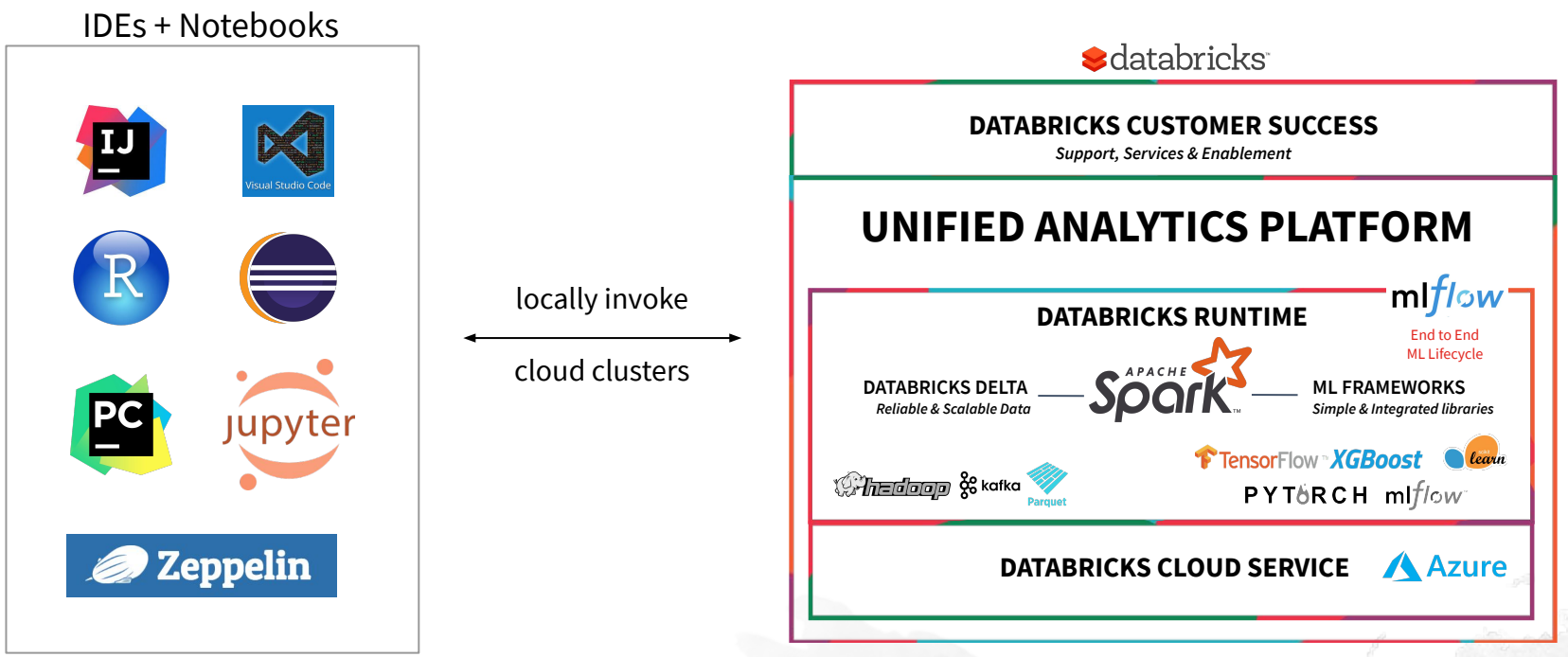
[Databricks RStudio Integration]

SparkR - recommended

- Created by Spark committers
- Optimized performance on Databricks
- Up to 100x faster!
- dapply() to apply in parallel!

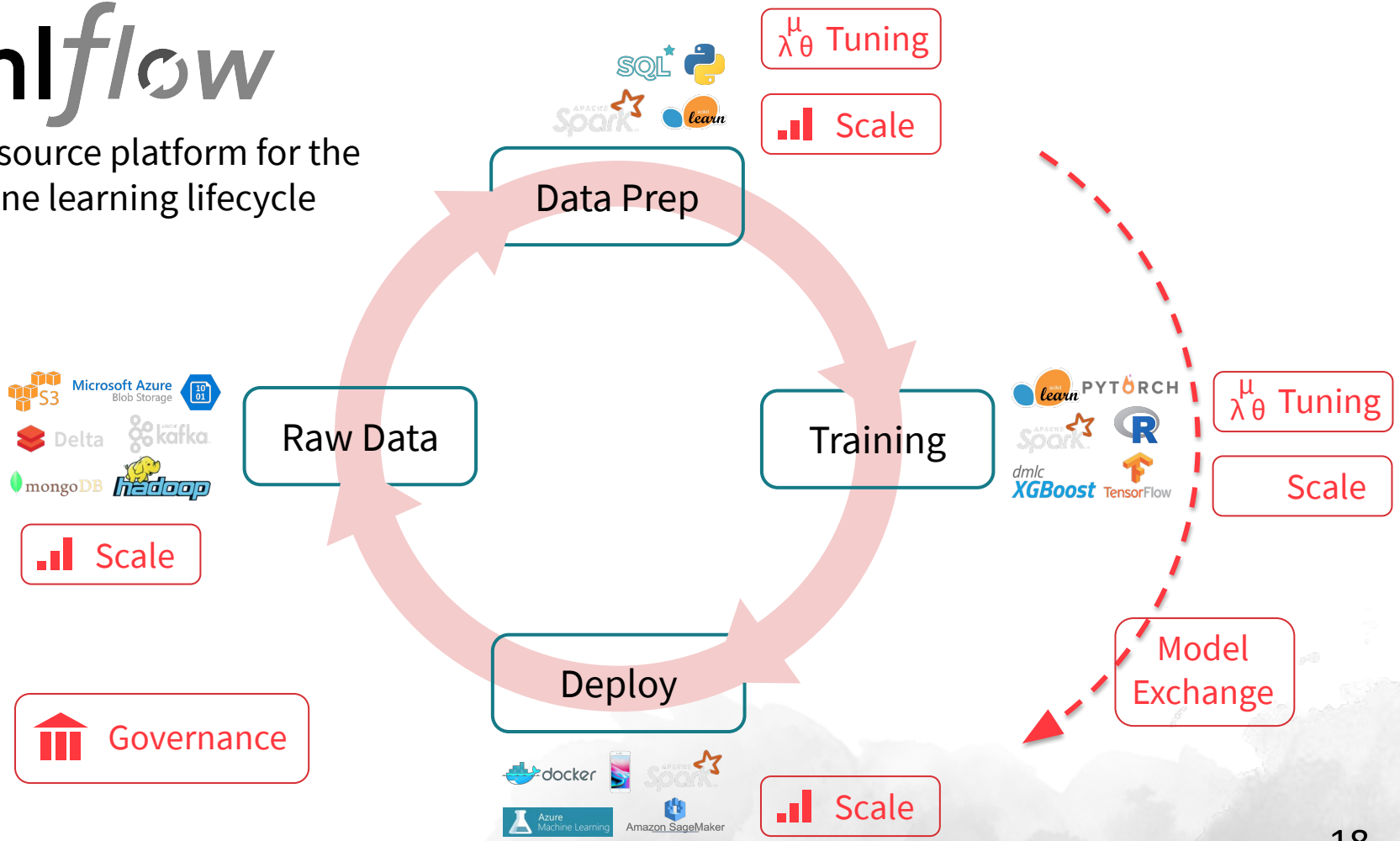
```
# Convert waiting time from hours to seconds.
# Note that we can apply UDF to DataFrame.
schema <- structType(structField("eruptions", "double"), structField("waiting", "double"),
                     structField("waiting_secs", "double"))
df1 <- dapply(df, function(x) { x <- cbind(x, x$waiting * 60) }, schema)
head(collect(df1))
##  eruptions waiting waiting_secs
##1     3.600      79         4740
##2     1.800      54         3240
##3     3.333      74         4440
##4     2.283      62         3720
##5     4.533      85         5100
##6     2.883      55         3300
```

DBconnect: Databricks for Developers



mlflow

An open source platform for the machine learning lifecycle



MLflow Components

mlflow
Tracking

Record and query experiments: code, data, config, results

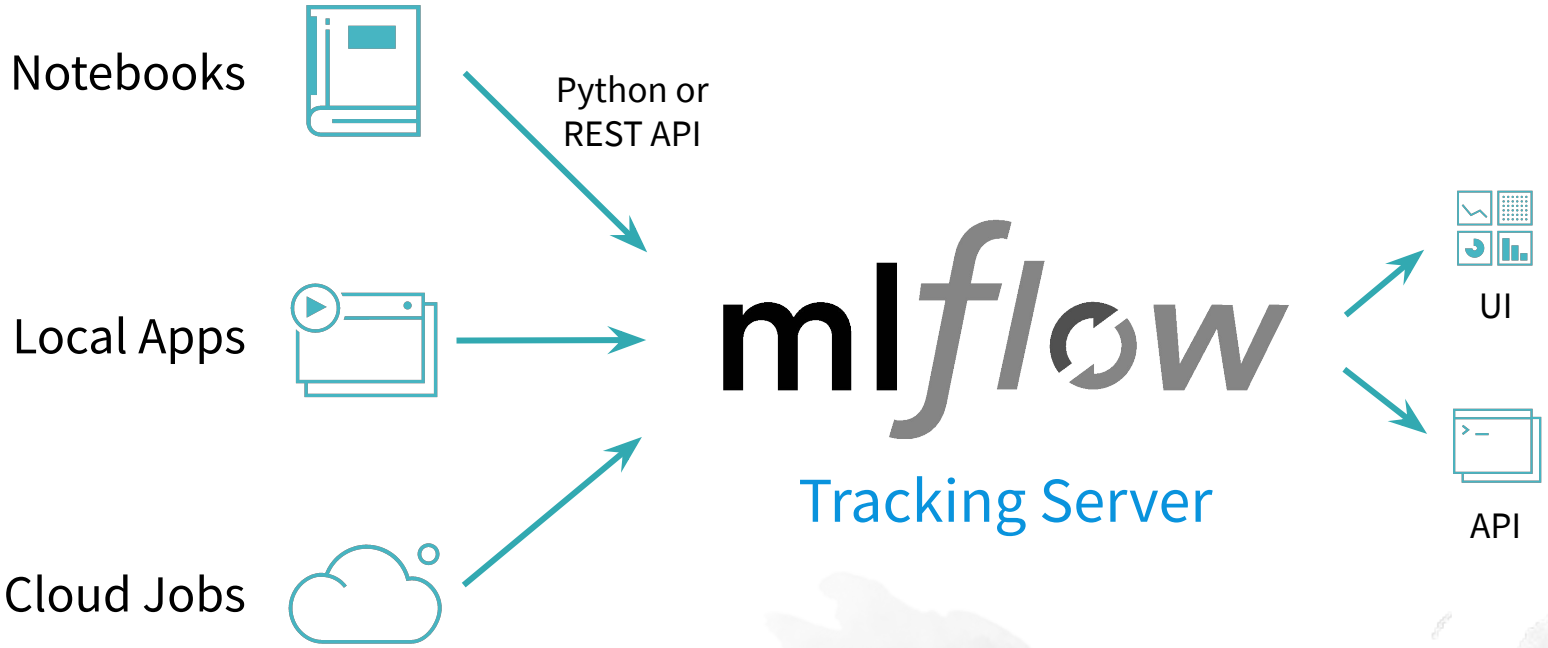
mlflow
Projects

Packaging format for reproducible runs on any platform

mlflow
Models

General model format that supports diverse deployment tools

MLflow Tracking



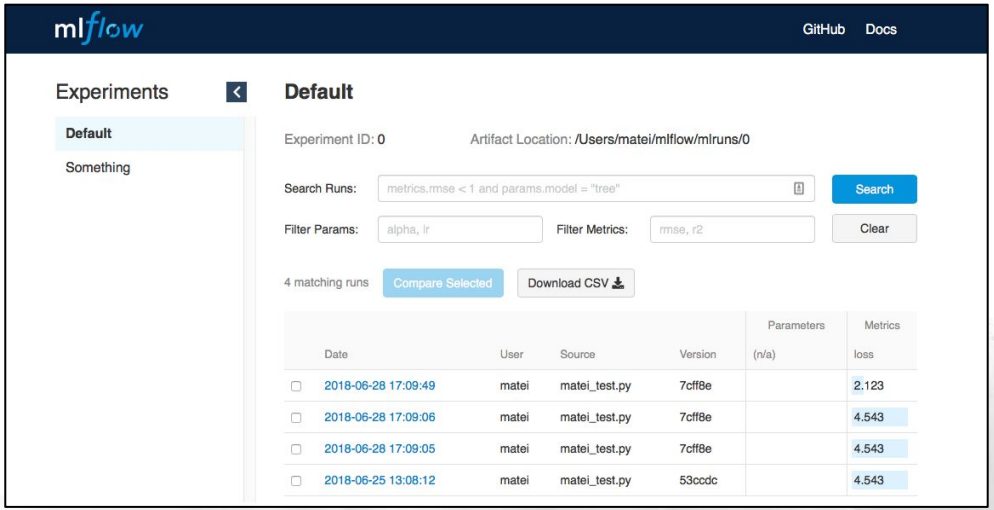
Key Concepts in Tracking

Parameters: key-value inputs to your code

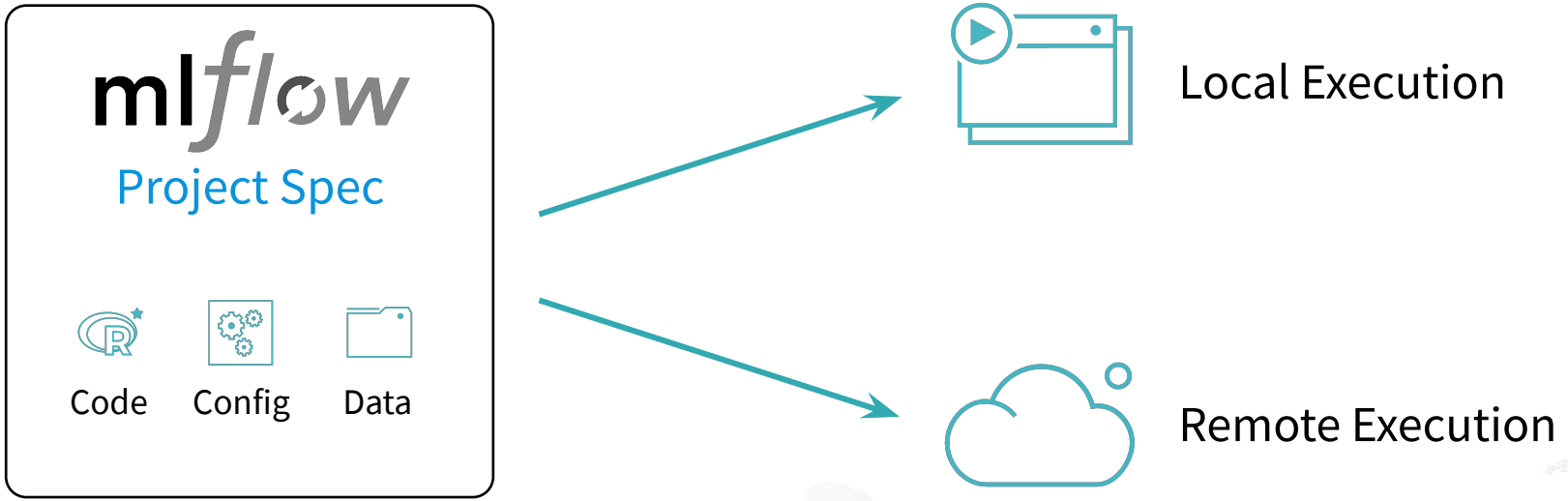
Metrics: numeric values (can update over time)

Artifacts: arbitrary files, including models

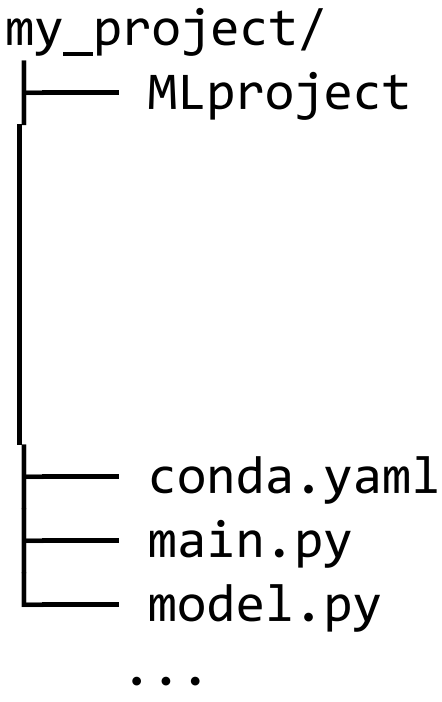
Source: what code ran?



MLflow Projects



Example MLflow Project

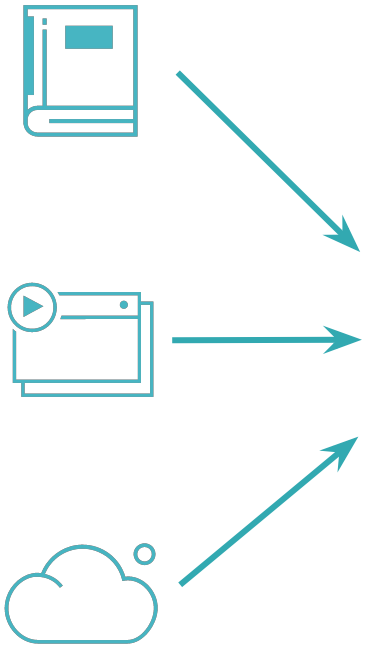



```
conda_env: conda.yaml

entry_points:
  main:
    parameters:
      training_data: path
      lambda: {type: float, default: 0.1}
    command: python main.py {training_data} {lambda}
```

```
$ mlflow run git://<my_project>
```

```
mlflow.run("git://<my_project>", ...)
```

MLflow Models



Run Sources

mlflow

Model Format

Flavor 1

Flavor 2

Simple model flavors
usable by many tools

docker

Inference Code

APACHE Spark

Batch & Stream Scoring

Azure Machine Learning

kubernetes

Cloud Serving Tools

databricks

Example MLflow Model

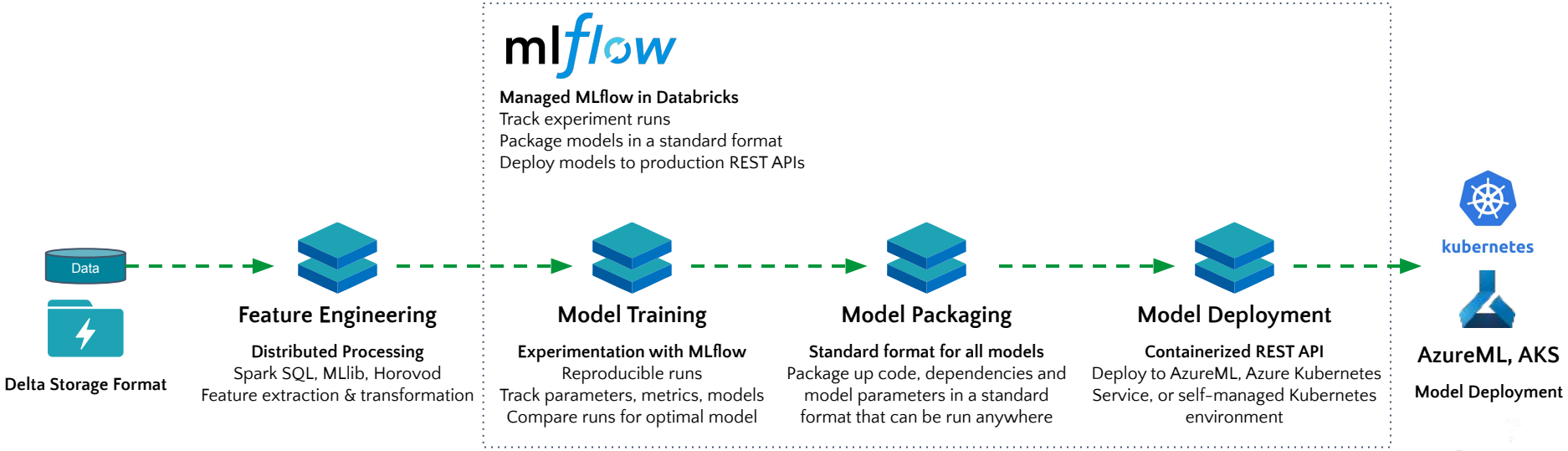
```
my_model/
├── MLmodel
└── estimator/
    ├── saved_model.pb
    └── variables/
        ...
```

```
run_id: 769915006efd4c4bbd662461
time_created: 2018-06-28T12:34
flavors:
  tensorflow:
    saved_model_dir: estimator
    signature_def_key: predict
  python_function:
    loader_module: mlflow.tensorflow
```

Usable by tools that understand TensorFlow model format

Usable by any tool that can run Python (Docker, Spark, etc!)

Machine Learning



Resources

Documentation:
Spark MLlib Guide - docs for feature extraction, ML algorithms, and pipelines in Spark
Databricks ML Guide - best practices for using MLlib on Databricks
MLflow Guide - tracking experiments, saving and deploying models on Databricks
Deep Learning on Databricks - guide to using distributed DL frameworks
Koalas - Pandas interface to Spark

Notebooks:
ML Workshop - GitHub repository with examples on data analysis, ML training, and experiments
More example notebooks - email samir@databricks.com

Demo and Sample Notebooks