MACHINE LEARNING

In **Azure** Databricks

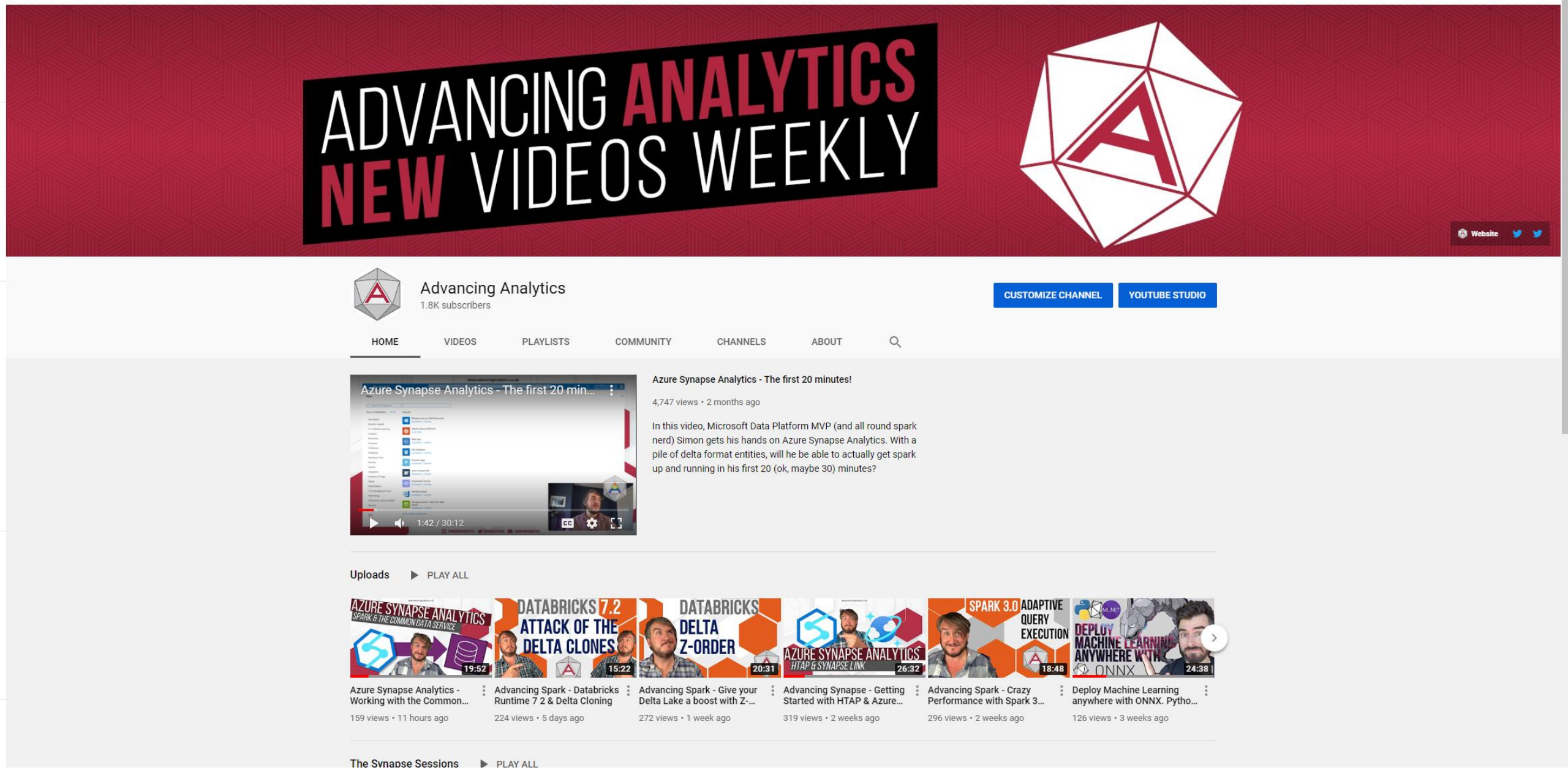
Terry McCann – Director of AI

ADVANCING ANALYTICS

DATA SCIENCE IN PRODUCTION PODCAST

- Data Science
- Dataops
- Data engineering
- Applied AI
- Model management
- Applied Training

ADVANCING ANALYTICS NEW VIDEOS WEEKLY



https://www.youtube.com/c/AdvancingAnalytics

ADVANCING ANALYTICS

Microsoft Partner

Microsoft

Gold Data Analytics
Gold Data Platform
Silver Cloud Platform

ADVANCING ANALYTICS

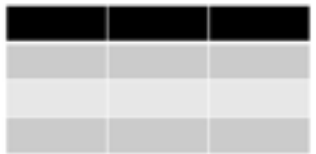
databricks partner

Terry McCann
Director of AI

- 10+ Years Microsoft Data Analytics
- Applied Data Engineer & Scientist
- Pioneer of MLOps
- Leader in Data Science & AI Community

MVP Microsoft® Most Valuable Professional

What is SPARK?



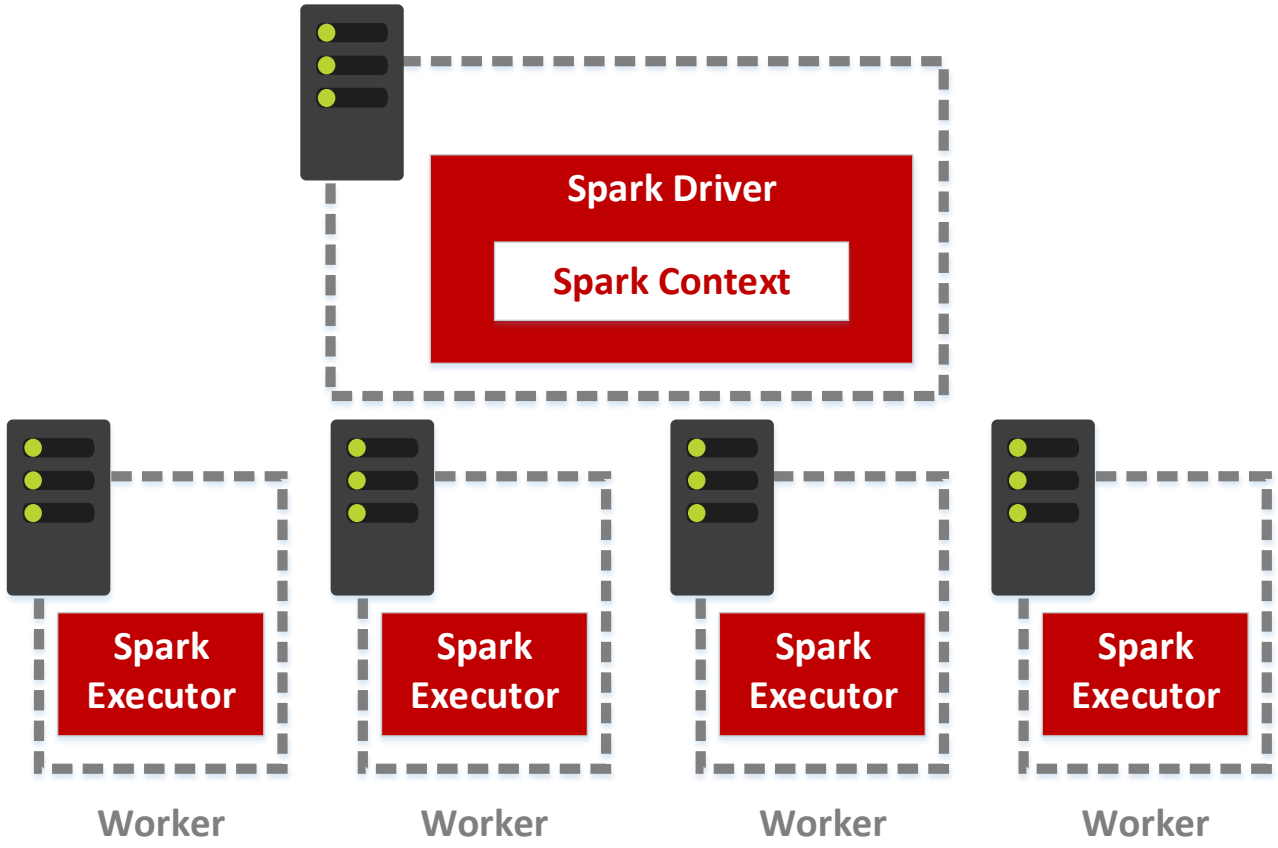
SQL / DataFrames

Machine Learning

Graph

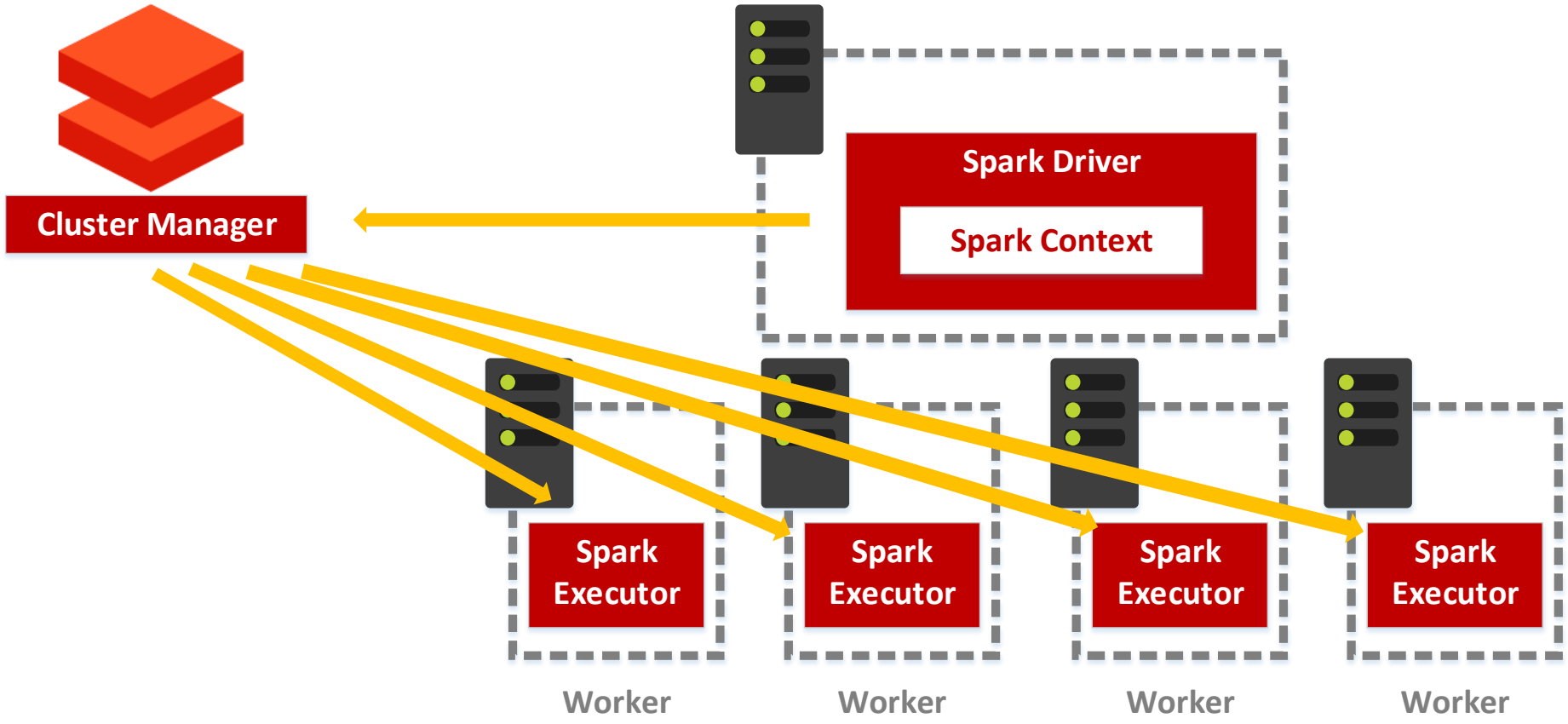
Streaming

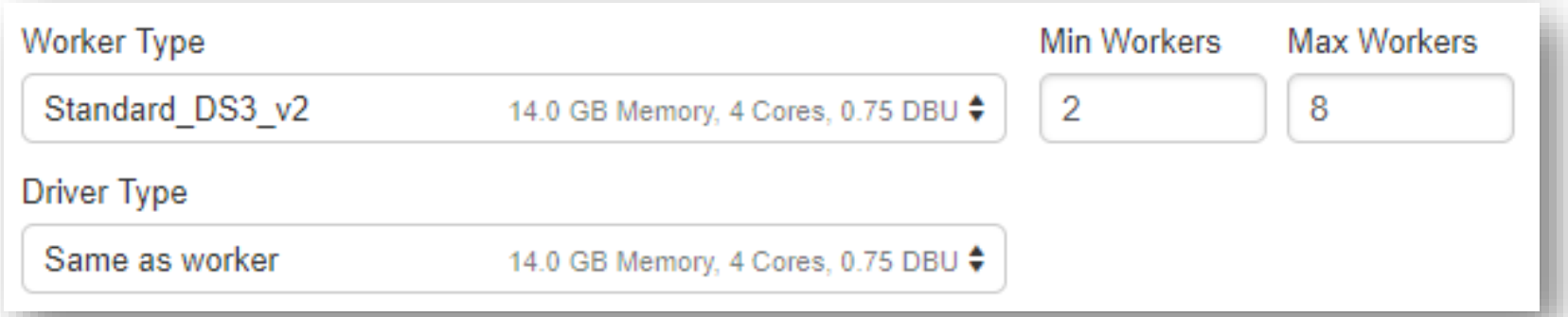
PHYSICAL SPARKITECTURE



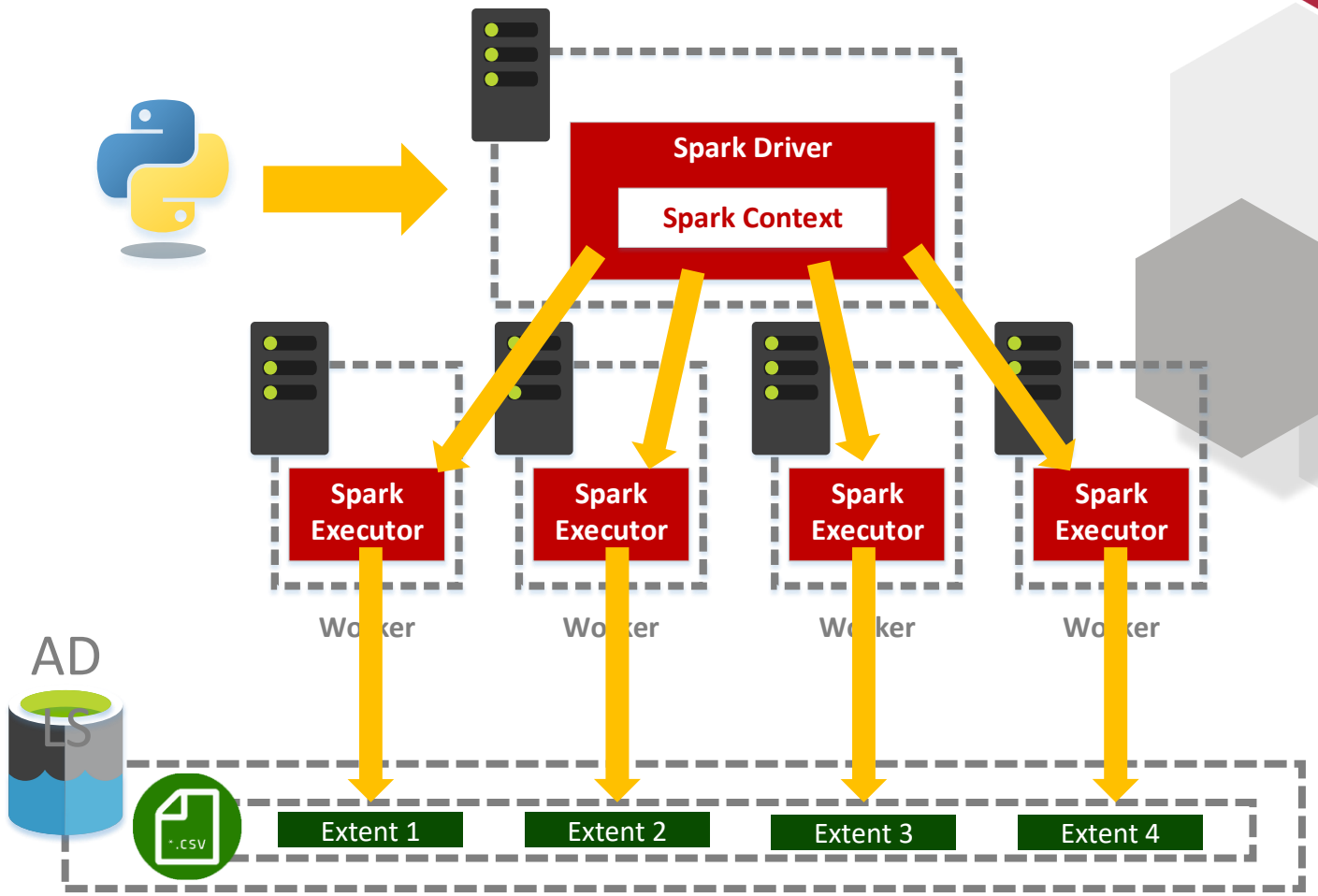
Whatever physical infrastructure under pins it, Spark will always run Java Virtual Machines for each driver / executor !

PHYSICAL SPARKITECTURE

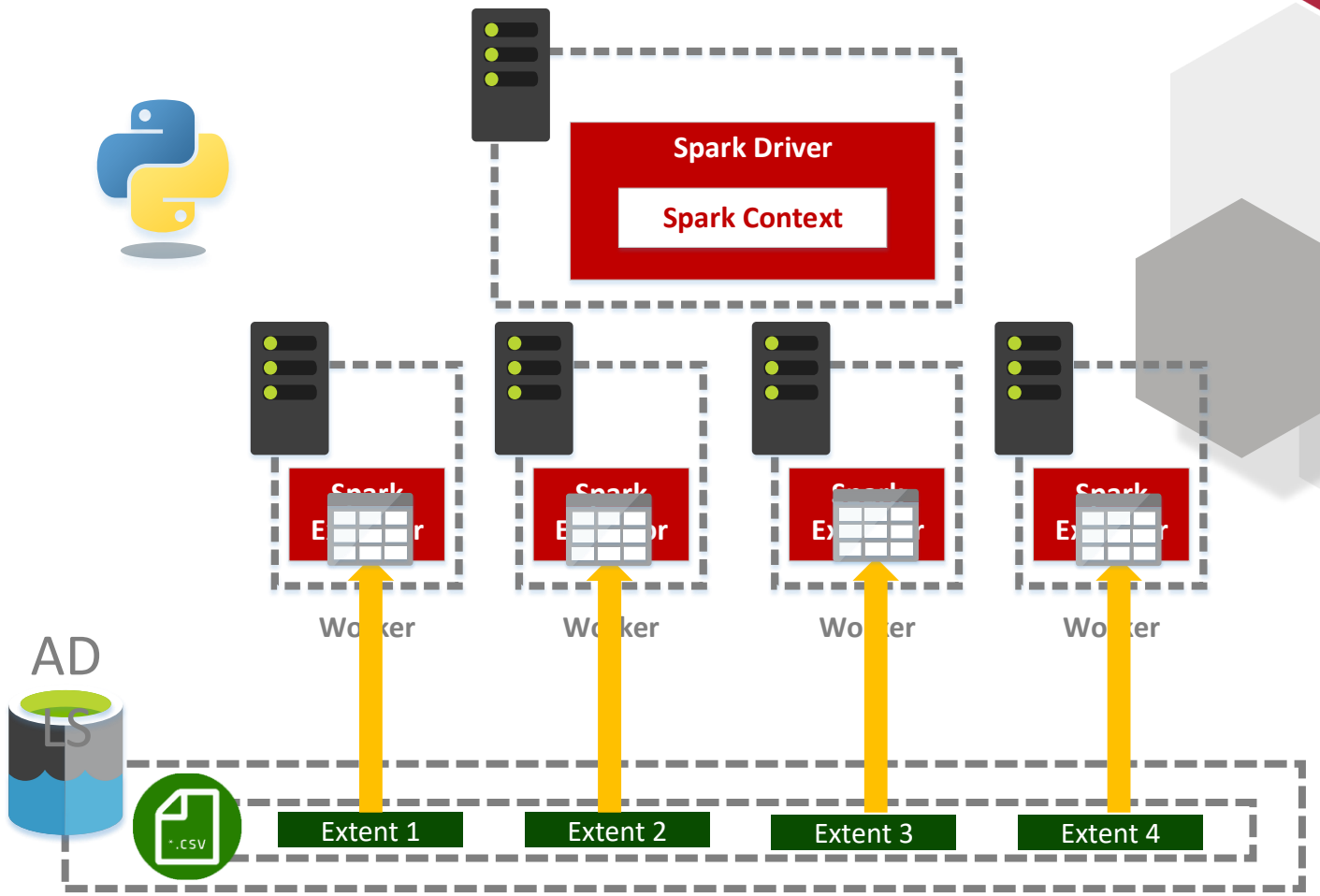




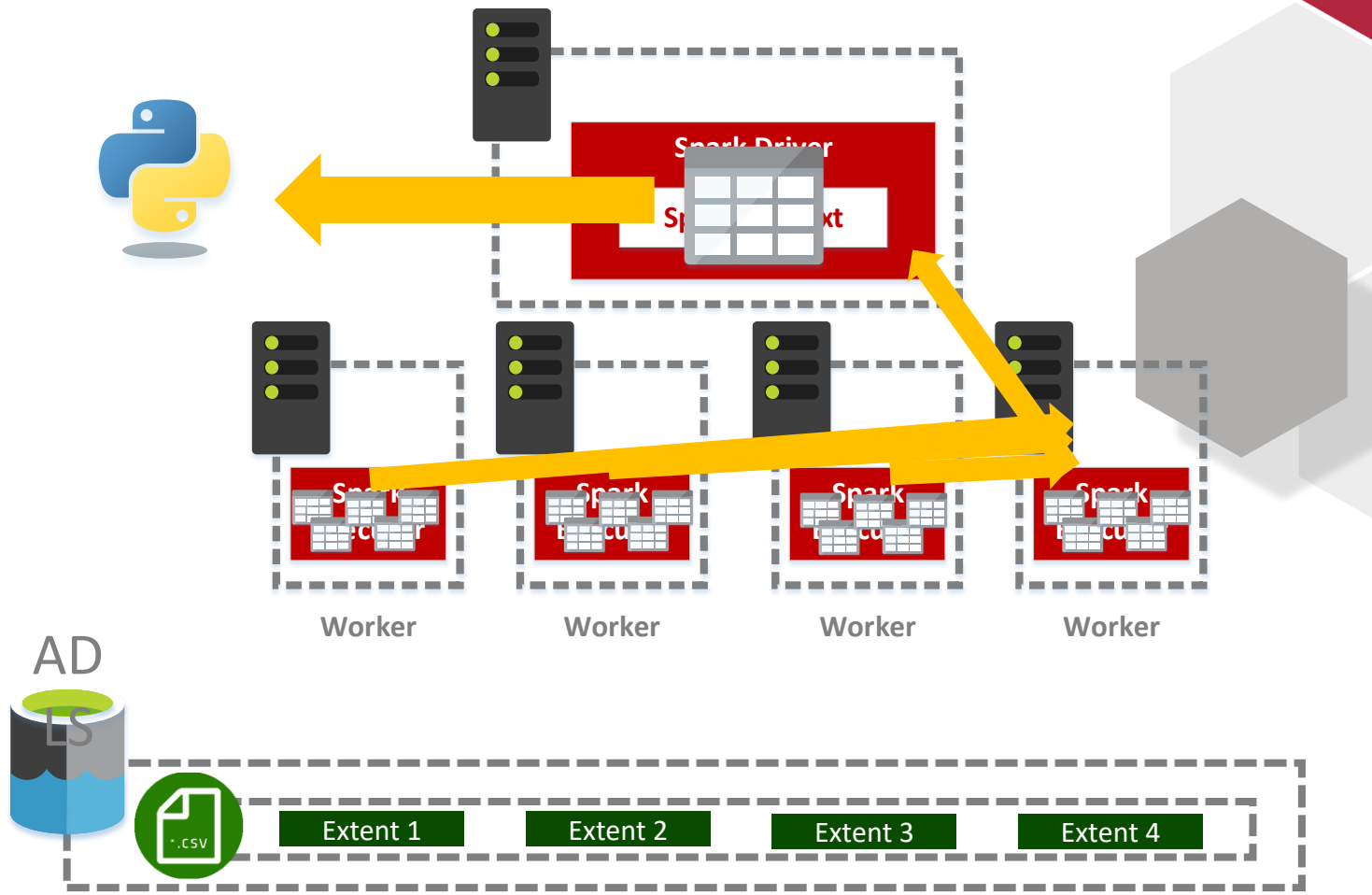
DISTRIBUTED COMPUTE



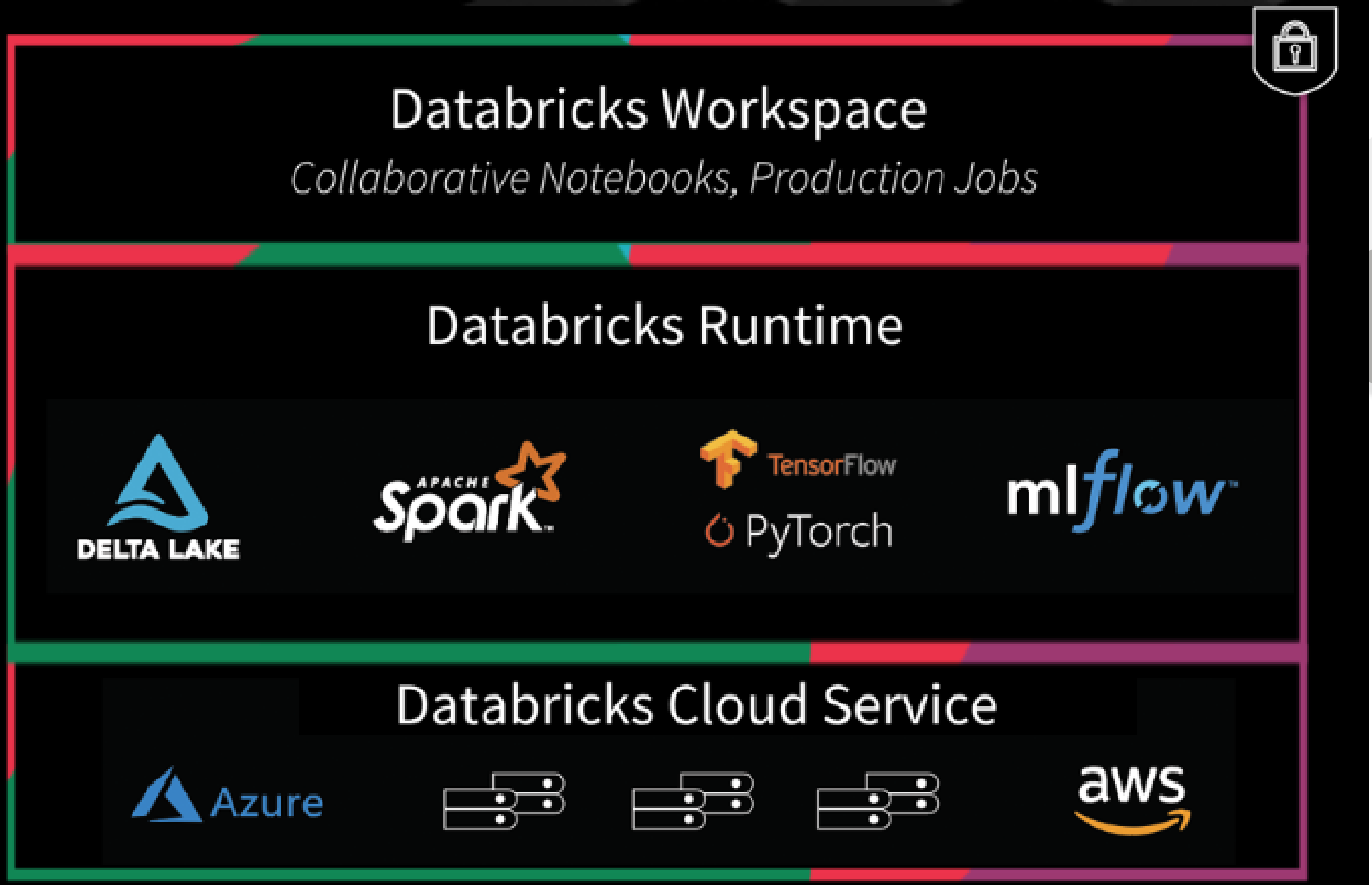
DISTRIBUTED COMPUTE



DISTRIBUTED COMPUTE



So What Is Databricks?


Databricks Workspace
Collaborative Notebooks, Production Jobs
Databricks Runtime
DELTA LAKE
APACHE Spark
TensorFlow
PyTorch
mlflow
Databricks Cloud Service
Azure
aws

Unifying Data Science and Engineering

Databricks Unified Analytics Platform, from the original creators of Apache Spark™, unifies data science and engineering across the Machine Learning lifecycle from data preparation, to experimentation and deployment of ML applications.

Data Science

Collaboratively explore large datasets, build models iteratively and deploy across multiple platforms.

LEARN MORE

Data Engineering

Speed up the preparation of high quality data, essential for best-in-class ML applications, at scale.

Machine learning on Azure Databricks

The Machine Learning Process

1. Gathering data
2. Preparing that data / visualise the data
3. Select an algorithm
4. Train a model (data + algorithm)
5. Evaluate the model
6. Hyperparameter tuning
7. Test the prediction (new data + model)
8. Deploy the model (“Productionised”)

Machine learning provides systems the ability to ***automatically learn*** and improve from experience ***without being explicitly programmed***.

It relies on underlying hypothesis of creating the model and tries to ***improve*** it by ***fitting*** more ***data*** into the model over time.

https://www.quora.com/What-is-the-best-definition-for-machine-learning

Types of Machine Learning

Shallow Learning

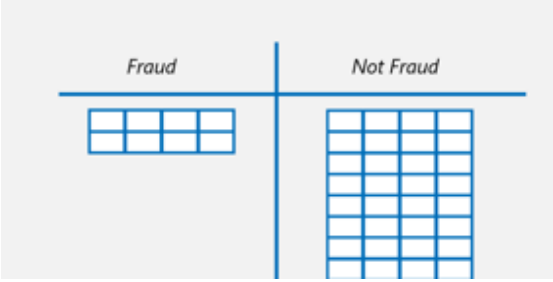
Deep Learning

Shallow Learning

Input Data

Feature Extraction

Classification

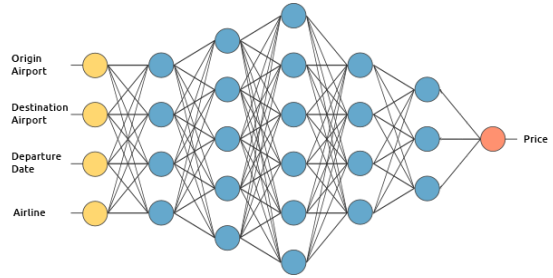


Output

Deep Learning

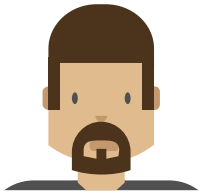
Input Data

Feature Extraction & Classification

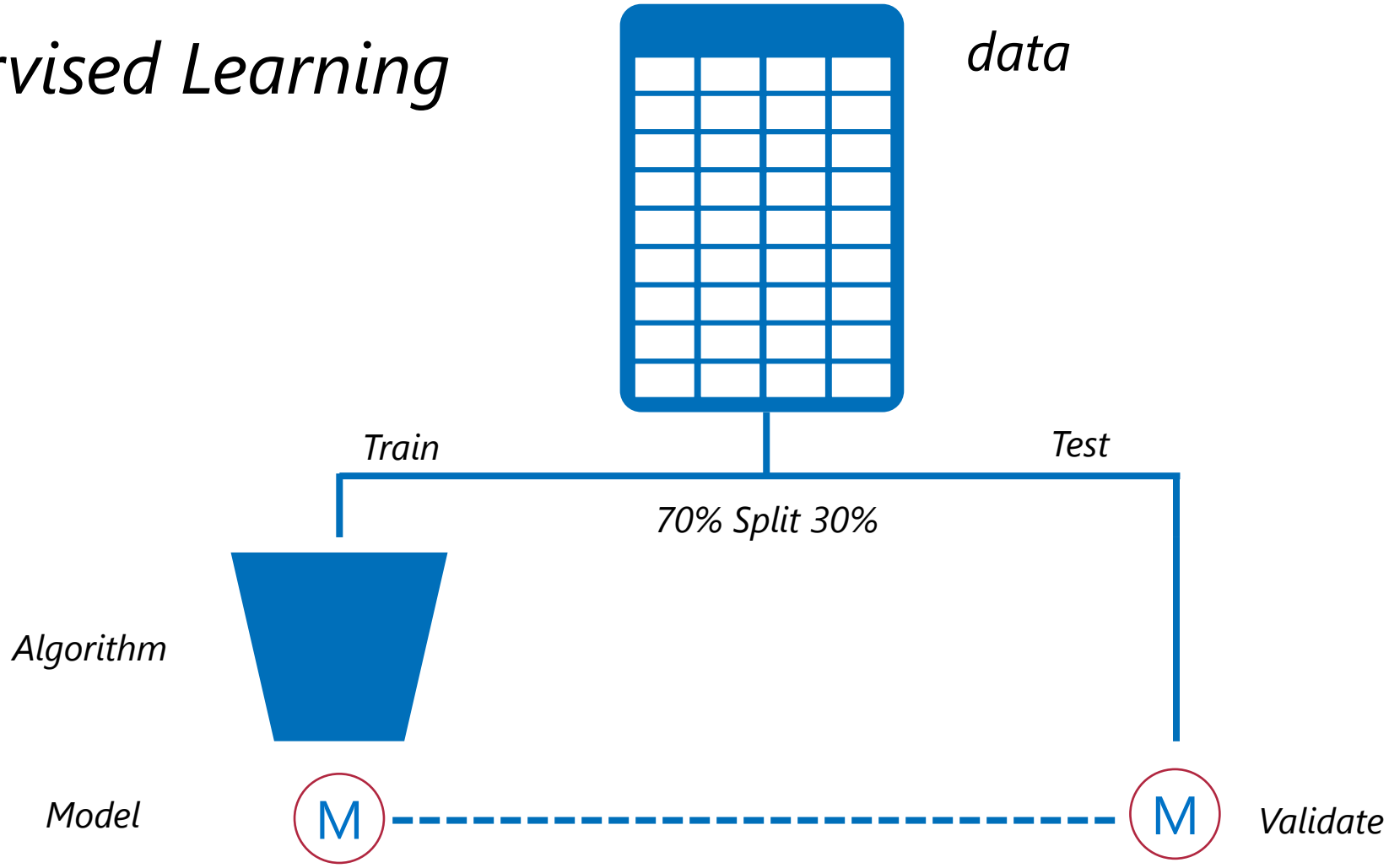


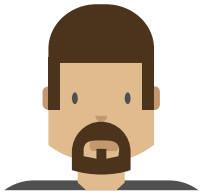
Output

Types of machine learning

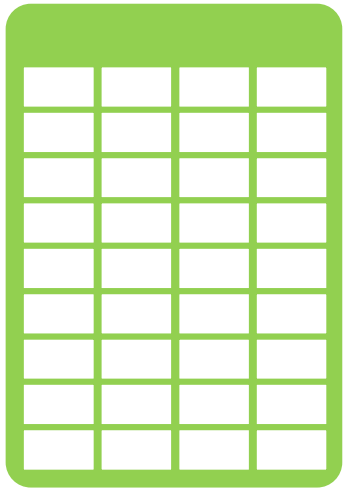


Supervised Learning

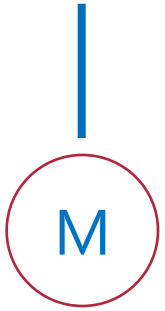




Supervised Learning



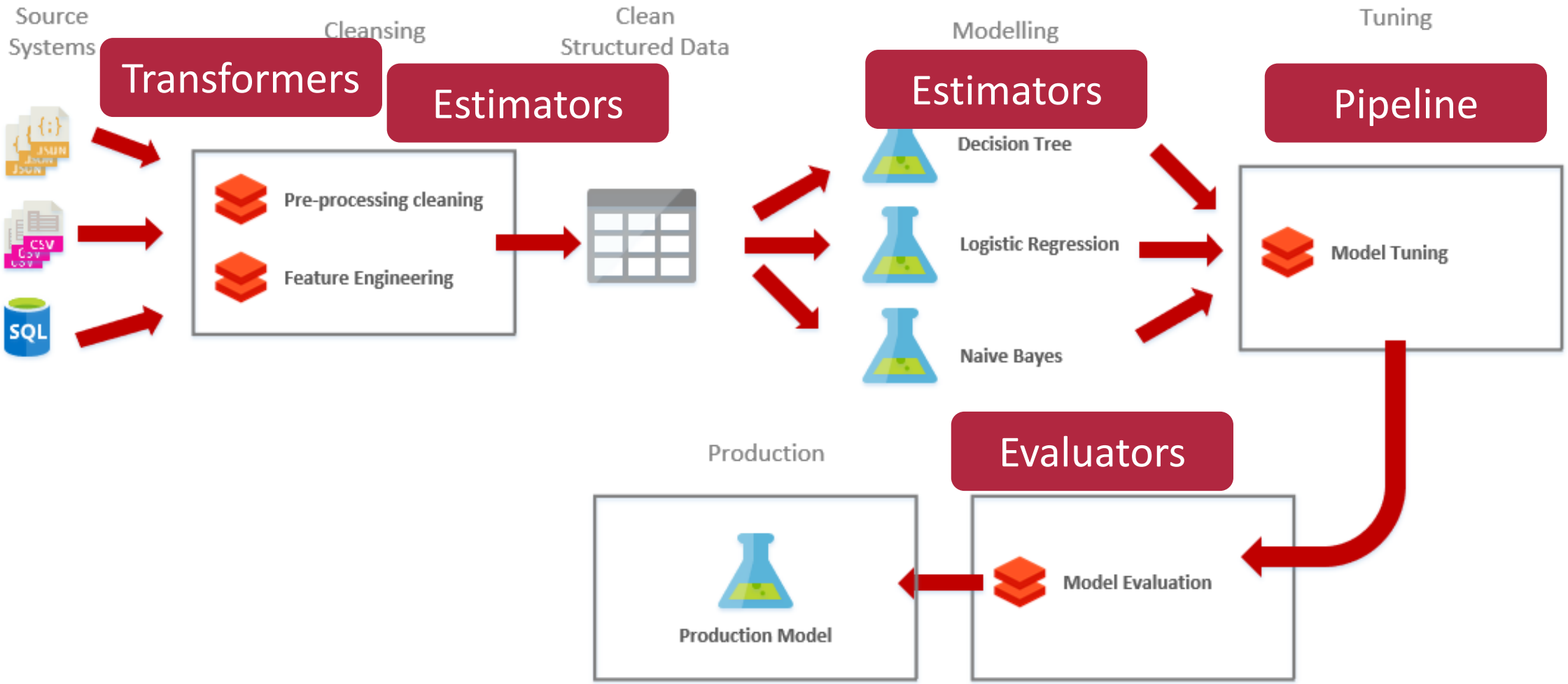
New data



Model

Prediction

SPARK ML


Source Systems
Cleansing
Transformers
Estimators
Clean Structured Data
Modelling
Estimators
Tuning
Pipeline
Pre-processing cleaning
Feature Engineering
Decision Tree
Logistic Regression
Naive Bayes
Model Tuning
Evaluators
Production
Production Model
Model Evaluation

Transformers

Estimators

Pipelines

Evaluators

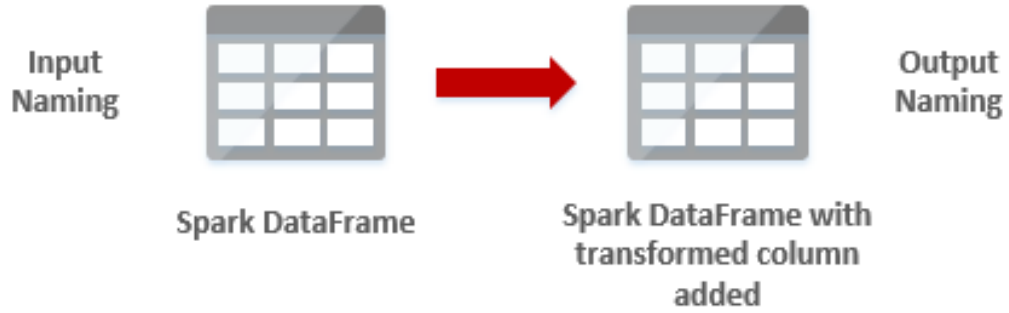
Transformers

Estimators

Pipelines

Evaluators

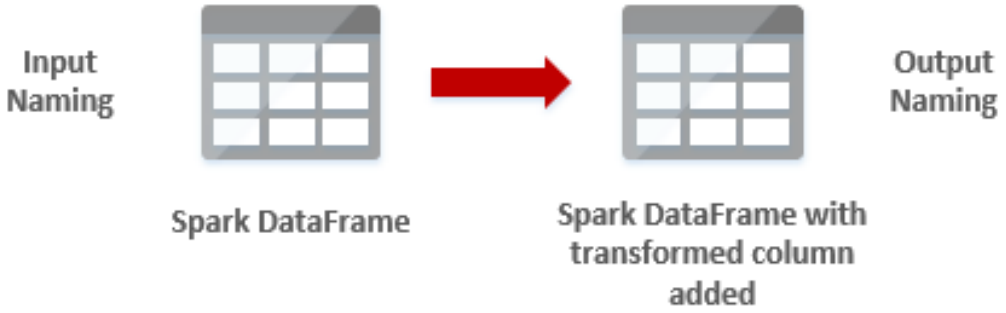
TRANSFORMERS



Primary use is for feature engineering and pre-processing

- Take in one DataFrame and produces another
- They complete a function across the data
  - Convert data type
  - Convert categorical variables to numerical
  - Normalise a column
- We call **transform** on a **transformer**

TRANSFORMERS



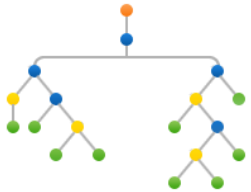
Example:

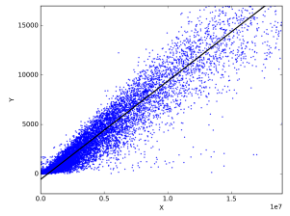
```
from pyspark.ml.feature import RegexTokenizer

tokenizer = (RegexTokenizer()
        .setInputCol("review")
        .setOutputCol("tokens")
        .setPattern("\\W+"))

tokenizedDF = tokenizer.transform(reviewsDF)
display(tokenizedDF.limit(5))
```

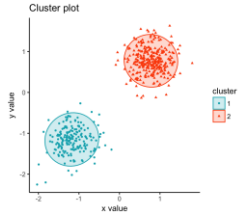
MACHINE LEARNING IN SPARK

Transformers

Estimator



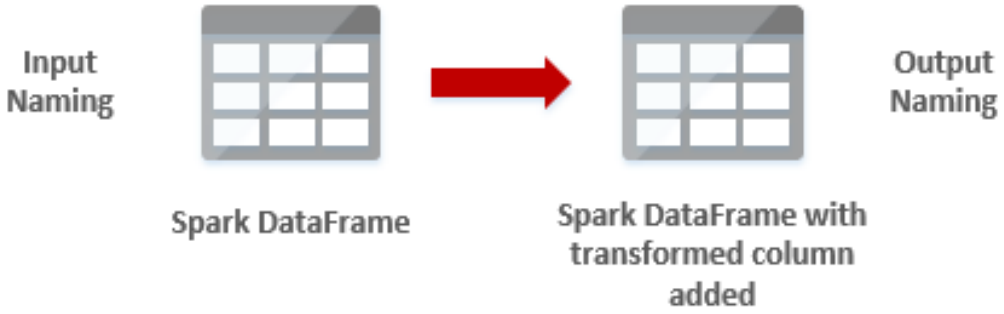




Pipelines

Evaluators

ESTIMATOR



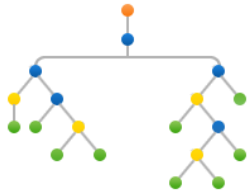
- We **fit** estimators with data at scale
- Once fitted becomes a **Transformer**
- Other types of estimator applies a function over the data (Machine Learning function)
- Examples:
  - Impute missing values

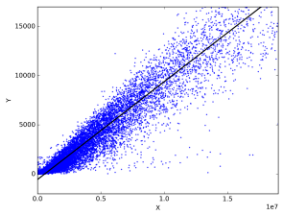
ESTIMATOR

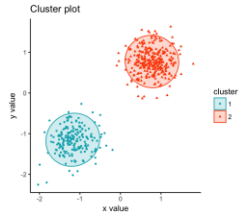
- VectorAssembler - Assemble the feature columns into a feature vector.
- VectorIndexer: Identify columns which should be treated as categorical. This is done heuristically, identifying any column with a small number of distinct values as being categorical.
- Algorithm: Example Decision Tree - This will build a decision tree to learn how to predict rental counts from the feature vectors.

Transformers

Estimator

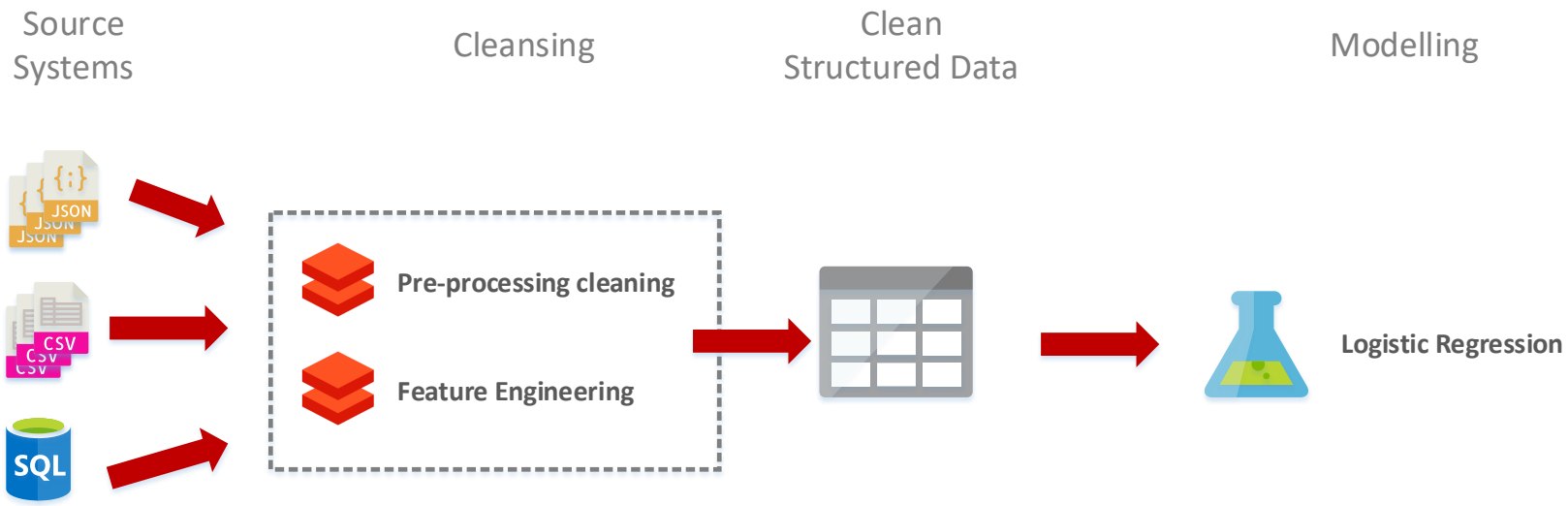






Pipeline

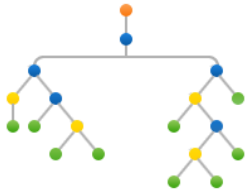
Evaluators

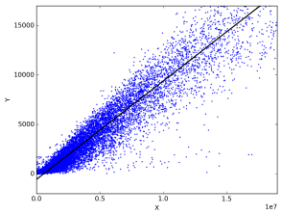


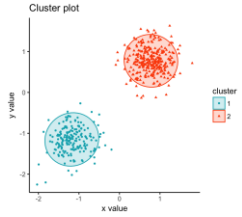
- Encapsulates all our logic in to one flow
- Orchestrates the training of our model

Transformers

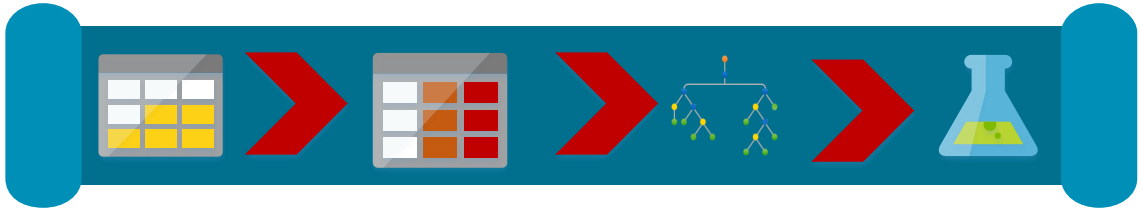
Estimator







Pipeline



Evaluator

Did we get a good prediction?

SPARKML MACHINE LEARNING

- Transformers
- Estimators
- Pipelines
- Bike Regression with MLFlow

When should you use spark?

BATCH

- Performed on a batch of data
- Performed on a schedule (Nightly)
- Time to process is in minutes
- Predictions are calculated over a large amount of data.
  - Customer Churn
  - Product recommendations
  - Micro segmentation

INTERACTIVE

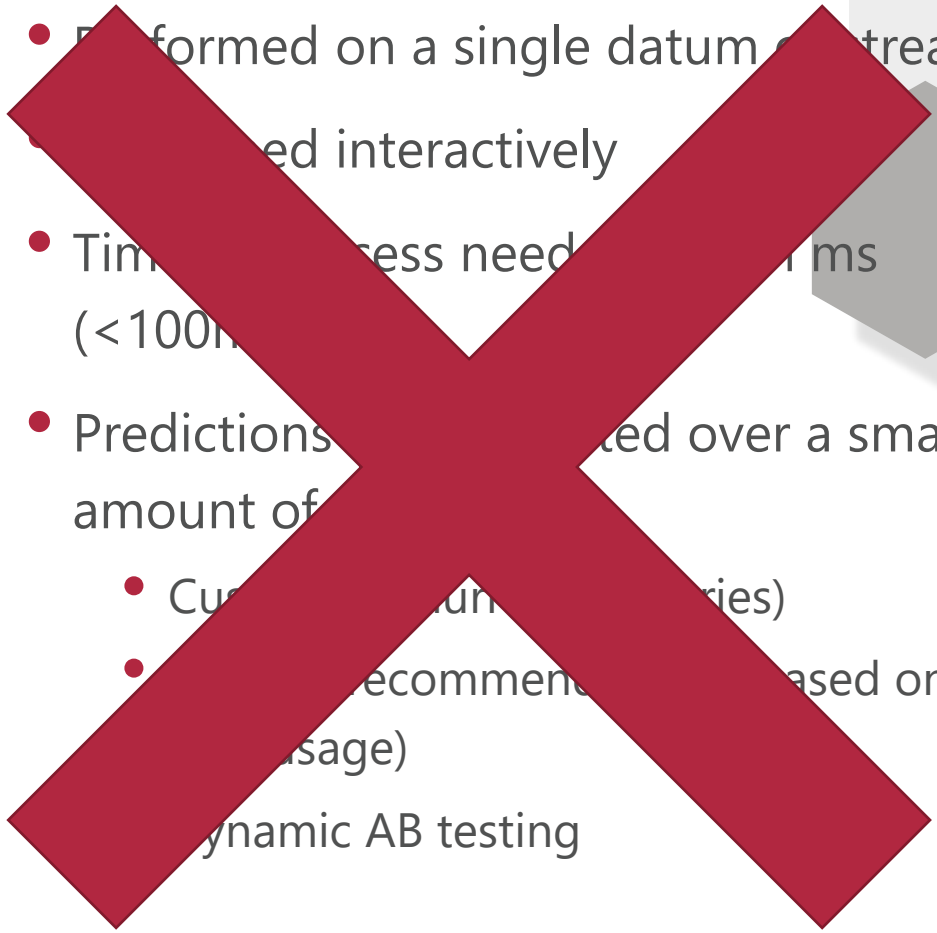
- Performed on a single datum or stream
- Performed interactively
- Time to process needs to be in ms (<100ms)
- Predictions are calculated over a small amount of data
  - Customer Churn (time-series)
  - Product recommendations (based on web usage)
  - Dynamic AB testing

BATCH

- Performed on a batch of data
- Performed on a schedule (Nightly)
- Time to process is in minutes
- Predictions are calculated over a large amount of data.
  - Customer Churn
  - Product recommendations
  - Micro segmentation

INTERACTIVE

- Performed on a single datum [illegible] stream
- [illegible] ed interactively
- Tim[illegible] ess need[illegible] ms (<100[illegible]
- Predictions [illegible] ted over a small amount of
  - Cu[illegible] un[illegible] ries)
  - [illegible] ecommen[illegible] ased on [illegible] sage)
  - [illegible] ynamic AB testing



BATCH

- Performed on a batch of data
- Performed on a schedule (Nightly)
- Time to process is in minutes
- Predictions are calculated over a large amount of data.
  - Customer Churn
  - Product recommendations
  - Micro segmentation

INTERACTIVE

- Performed on a single datum or stream
- Performed interactively
- Time to process needs to be in ms (<100ms)
- Predictions are calculated over a small amount of data
  - Customer Churn (time-series)
  - Product recommendations (based on web usage)
  - Dynamic AB testing

Microsoft Partner

Microsoft

Gold Data Analytics
Gold Data Platform
Silver Cloud Platform

ADVANCING ANALYTICS

databricks
partner

Terry McCann

Apache Spark is great for so many Data Analytics tasks. Where your data is in the **cloud**, it is the go-to for batch **Machine Learning**.

With **MLFlow** as a managed service, Machine Learning on Spark has never been so easy with Databricks!

Terry McCann
Director of Artificial Intelligence

terry@advancinganalytics.co.uk

ADVANCING ANALYTICS