LIXAR

DRIVEN BY DATA

Fueled by BDO

# Databricks Deep Dive for Public Sector

May 27, 2021

# Speaker Intros

Eric Rops

I am a Data Engineer, coming from seven years of industry experience as a Geophysicist harnessing multi-terabyte seismic, well log, and mineral exploration datasets to target the most economic prospects (both oil and mining sectors). During that time I became highly involved in data analytics, completing projects in workflow automation, machine learning, data pipelines, and dashboards for executives.

At Lixar BDO, I am currently building production data pipelines, data warehouses, and data lakes, enabling our clients to make the best possible decisions with their datasets.

I love turning large and complex data into something that is understandable and useable, and look forward to giving this workshop and hopefully passing some of that on!

# Speaker Intros

LIXAR
Fueled by BDO

Tom Walsh

I am a Solutions Architect with Lixar Fuelled by BDO who specializes in the architecture and implementation of Data Analytics solutions in Azure.

I come from a consulting background with a decade of experience delivering to a variety of industries including Energy, Law, Telcom, Dentistry, Investments and Sports. For these projects I have leveraged many Azure tools including of course Azure Databricks.

In addition, I have achieved the following certifications Microsoft Azure Solutions Architect, Databricks Associate Developer and Microsoft Data Azure Data Scientist Associate. I am looking forward to showing you what is possible with these tools during this workshop!

# AGENDA

LIXAR
Fueled by BDO

## Morning Session

9:30 – 10:00 AM | Environment Setup

10:00 – 12:30 PM | Data Engineering in Azure Databricks

- Databricks Overview
- Databricks Notebooks
- Databricks leveraging Azure Storage and Azure Key Vault
- Delta Lake Architecture and Delta Tables

12:30 - 1:00 PM | LUNCH BREAK

## Afternoon Session

1:00 – 2:30 PM | Spark Machine Learning in Azure Databricks

- Machine Learning Overview
- Linear Regression with Spark Machine Learning
- MLflow Tracking
- Classification Machine Learning - Fraud Detection

2:30 – 3:00 PM | Q&A

# Azure Setup: Setup01.md

Ensure the following Azure resources are setup:

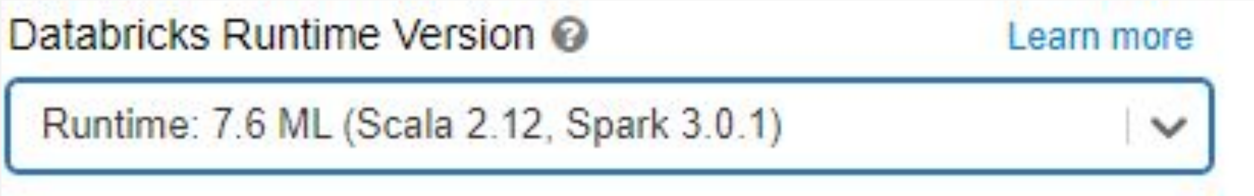
- Storage Account (with Data Lake Storage gen2 enabled)
- Azure Key Vault
- Data Factory V2
- Databricks

# Create Databricks Cluster: Setup02.md

Follow-along session to do the following:

- Create Databricks cluster
- Import the Databricks DBC archive

Choose the following Runtime



Databricks Runtime Version

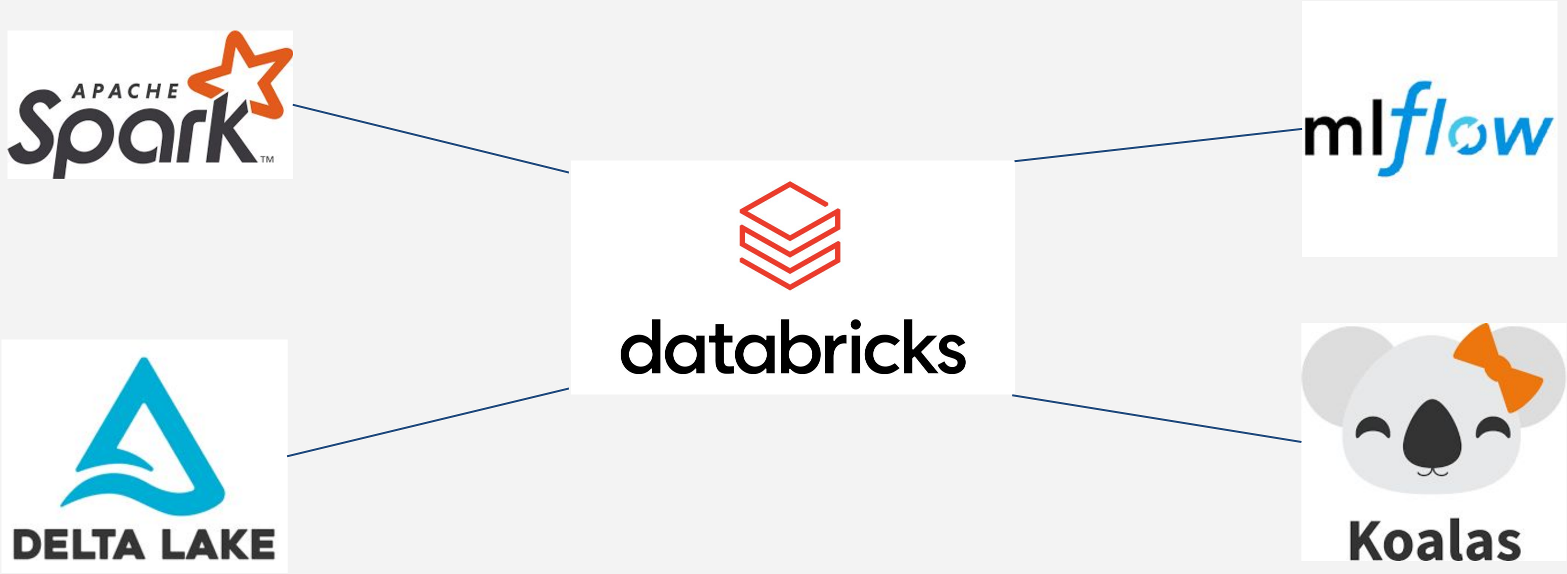
Learn more

Runtime: 7.6 ML (Scala 2.12, Spark 3.0.1)

# Databricks Overview

Databricks is a unified data platform that makes it easy to collaborate on data engineering, analytics, and machine learning workflows.

It is built on top of Apache Spark, and three other extremely popular open source projects.



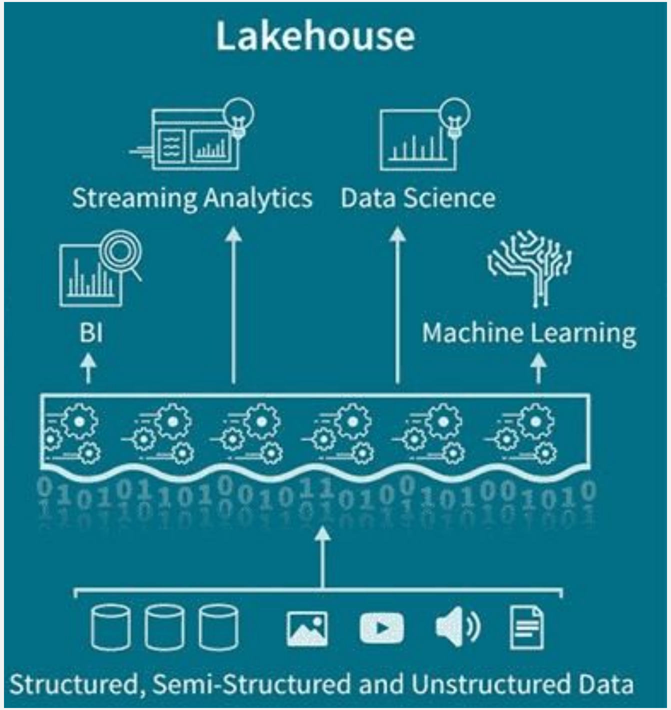
# Databricks Overview

Databricks provides a smooth implementation of the **Lakehouse Architecture,** which combines the best features of Data Warehouses and Data Lakes.

**One platform for every use case**

**Structured transactional layer**

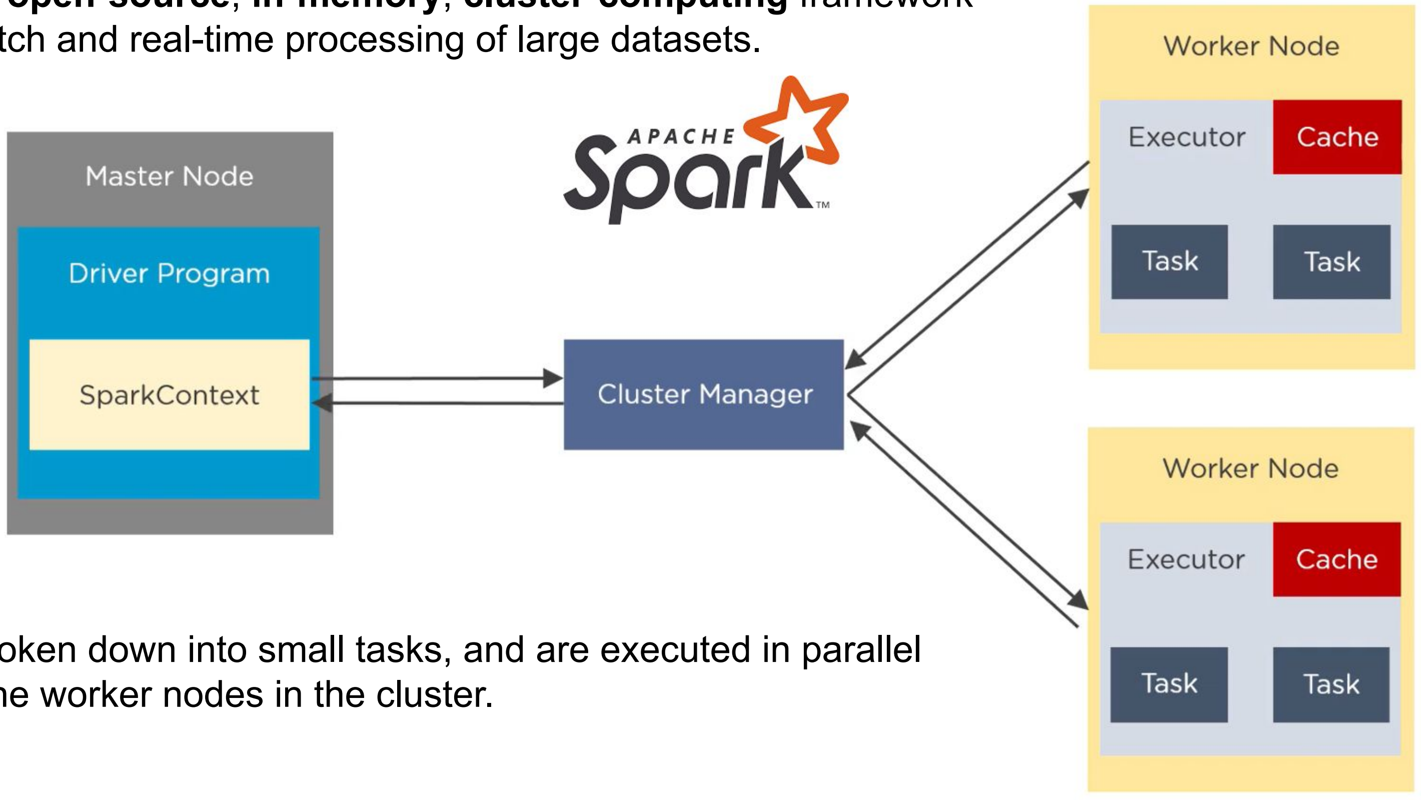
**Data Lake for all your data**



https://databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html

# Apache Spark Overview (Source: SimpliLearn)

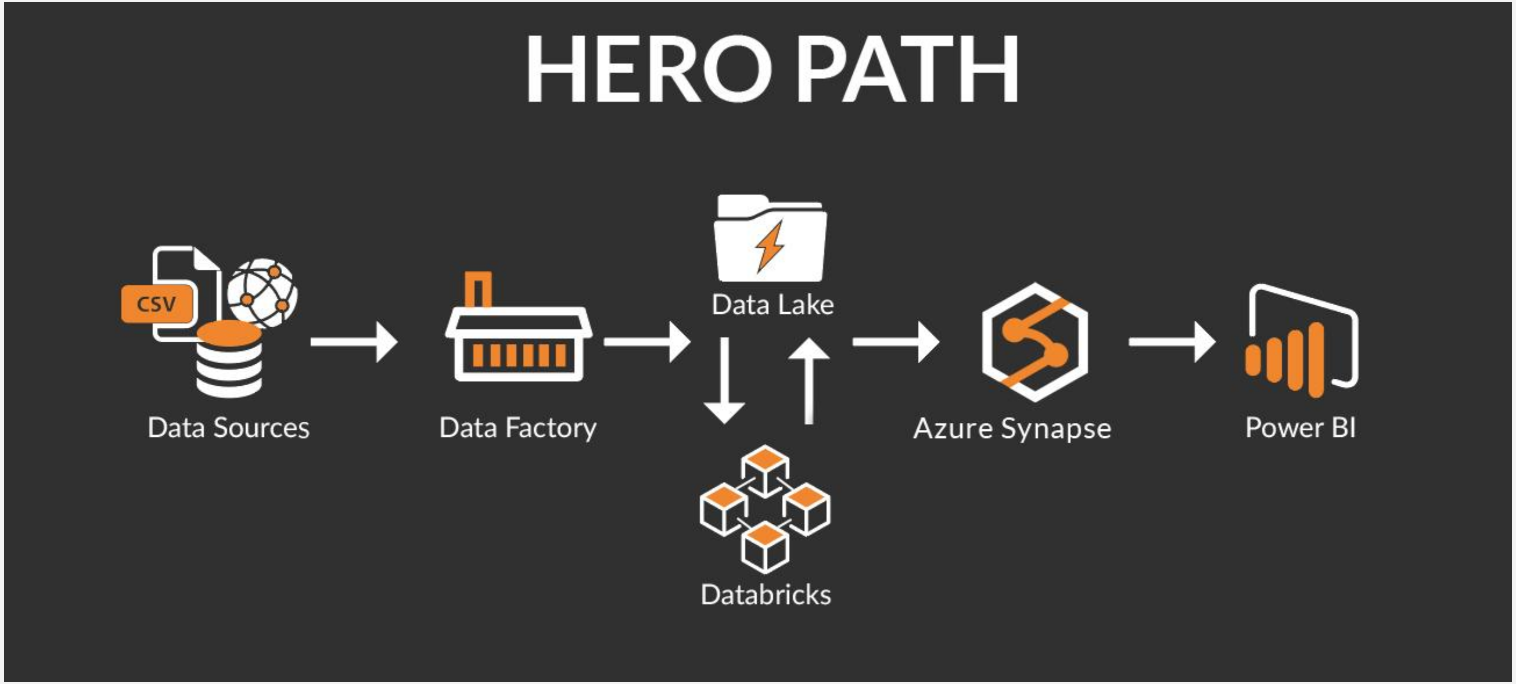
Spark is an **open-source**, **in-memory**, **cluster-computing** framework used for batch and real-time processing of large datasets.



Jobs are broken down into small tasks, and are executed in parallel across all the worker nodes in the cluster.

# Data Analytics with Databricks at the Heart

LIXAR
Fueled by BDO



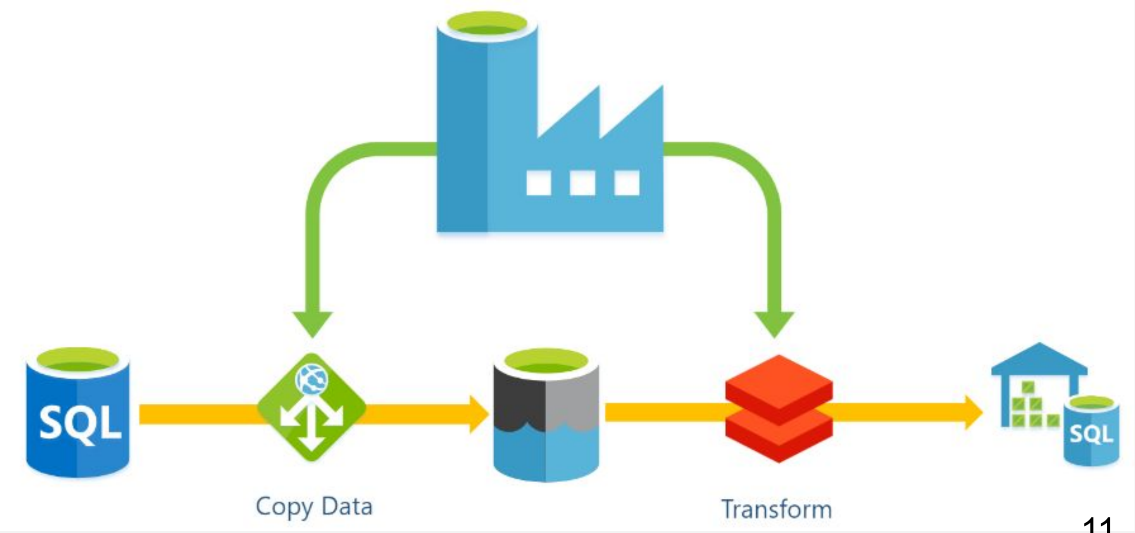
# Azure Data Factory

- Azure Data Factory (ADF) is a **serverless, data orchestration** service

- Data pipelines can be managed to run on a schedule, or based on triggers (such as new data arriving)

- ADF can ingest data from multiple sources, transform it, and load it to almost any data store

- ADF can also trigger Databricks notebooks, pass parameters and environment variables



https://github.com/mrpaulandrew/CommunityEvents/blob/master/DataPlatformDiscoveryDay-2020/Slides.pdf

# Exercise: Intro Notebook

Exercise in notebook file: **01_Intro**

**NOTE:**

- **Running the notebooks yourselves is optional**
- **It’s a good idea to Clear State & Results before running a new notebook**

# Exercise: Basic Databricks notebook

Exercise in notebook file: **02_Basic_Notebook**

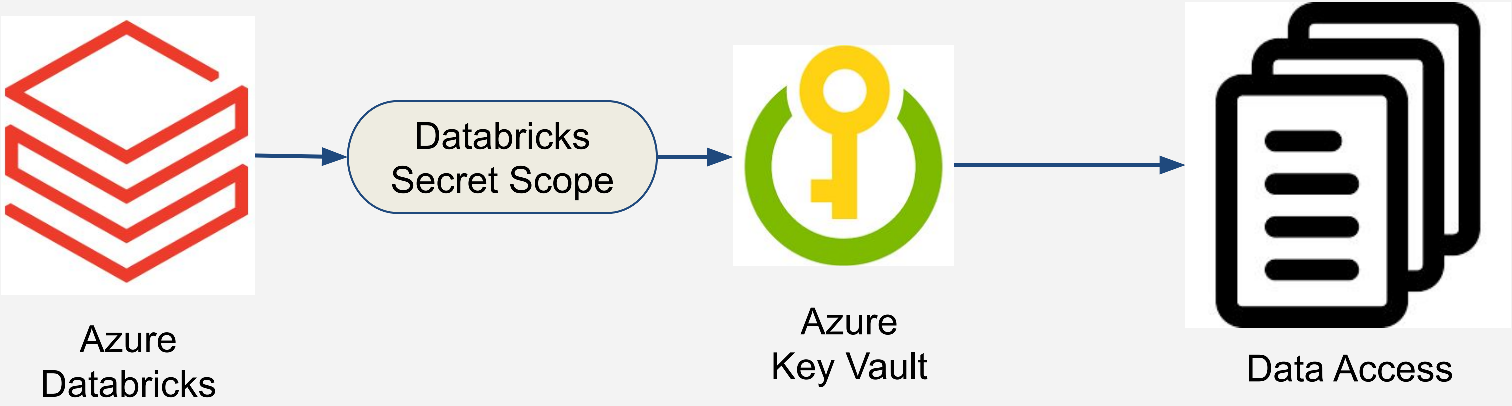
# Azure Storage Account and Secrets

LIXAR
Fueled by BDO

- Before we start loading data, we need to securely store our Storage Account access credentials to prevent outsiders from accessing the data!

- We can store our access credentials as secrets inside the **Azure Key Vault**.

- To access the secrets from Databricks, we create a **Secret Scope** link to the Key Vault.



# Setup Azure Storage Account Key Vault: Setup03.md and Setup04.md

Follow-along session to do the following:

- Azure Storage Container (with Data Lake Storage gen2 enabled)
- Azure Access Policies
- Add the Storage primary access key as a secret
- Create Databricks Secret Scope back to the Key Vault

**Please raise your hand if you run into issues!**

# Exercise: Secrets and Azure Storage

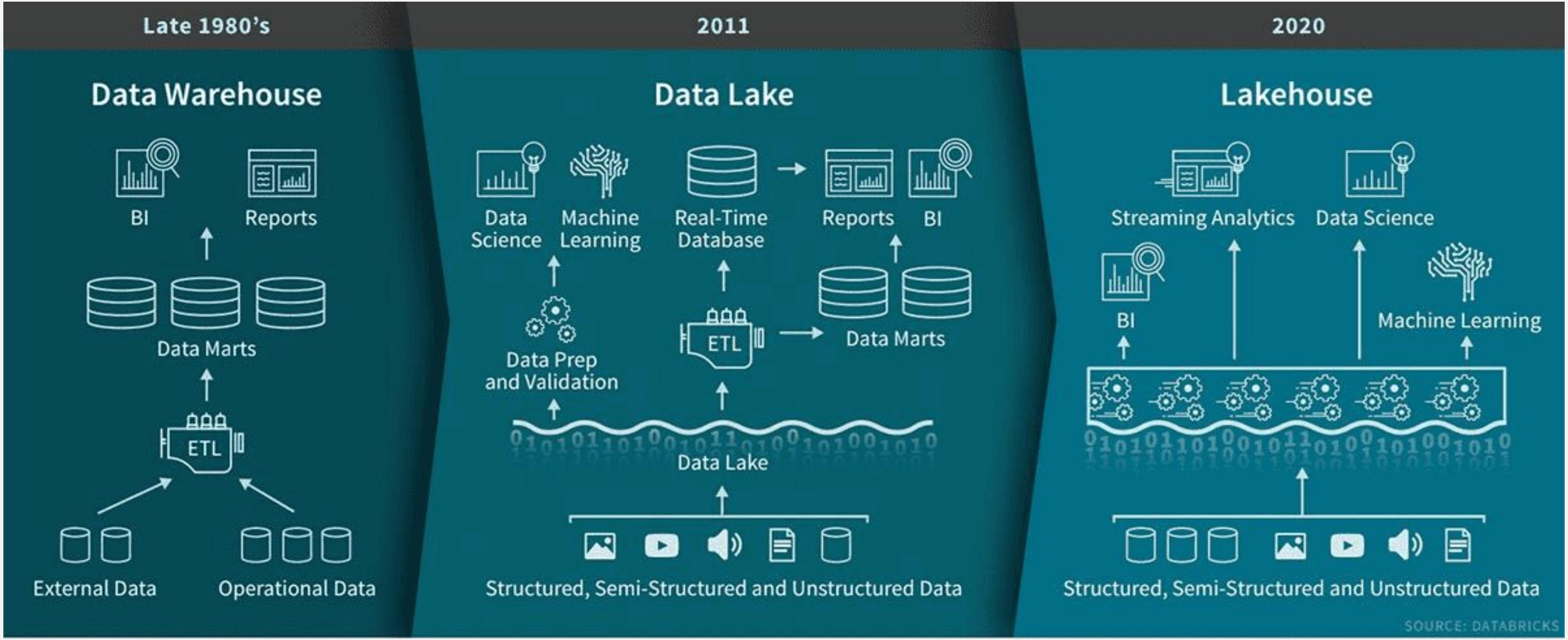
LIXAR
Fueled by BDO

Exercise in notebook file: **03_Secrets_And_Azure_Storage**

# Fundamentals of the Delta Lake Architecture

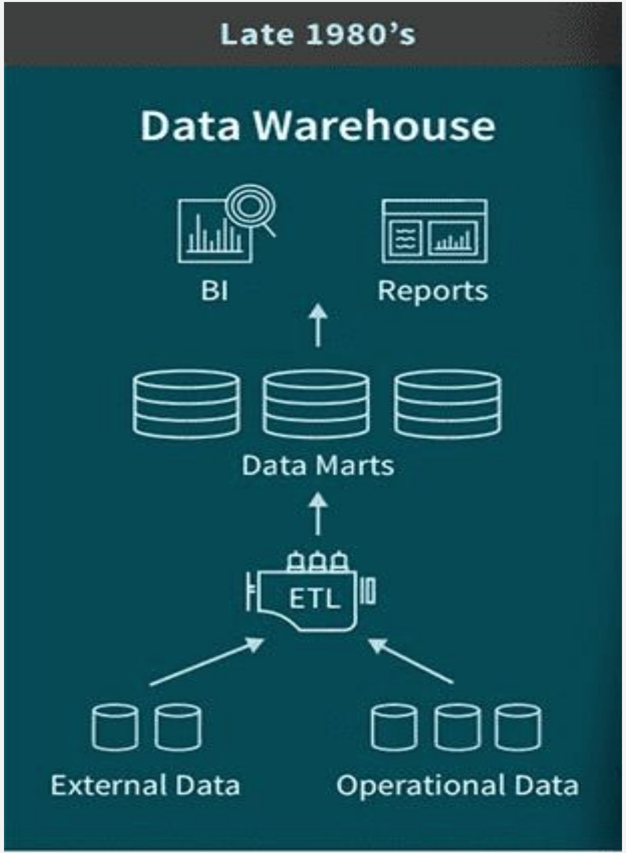
# Evolution: Data Warehouse to Delta “Lakehouse”

LIXAR Fueled by BDO



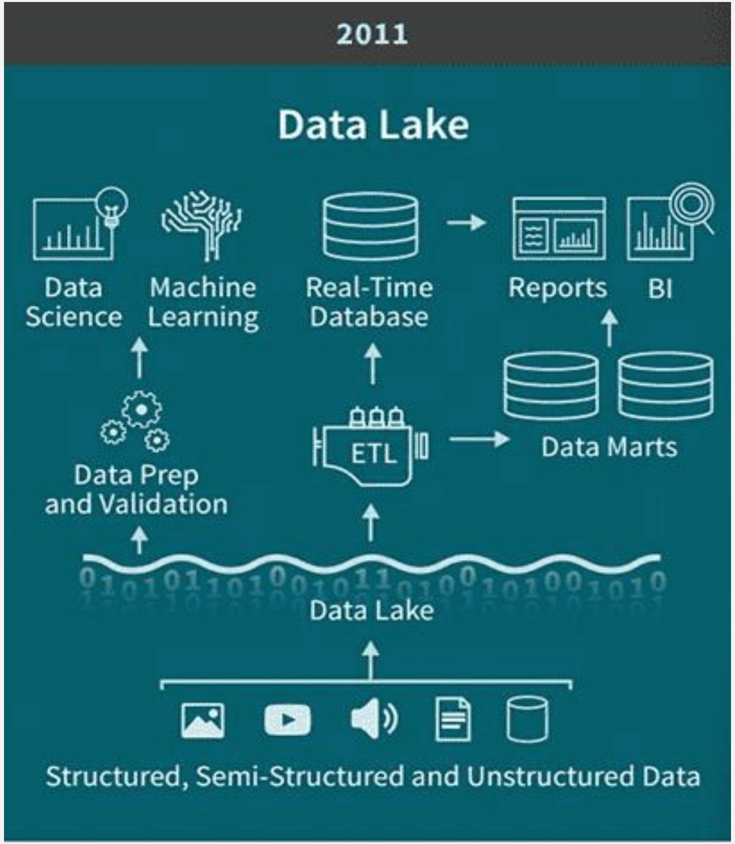
https://databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html

# Data Warehouse



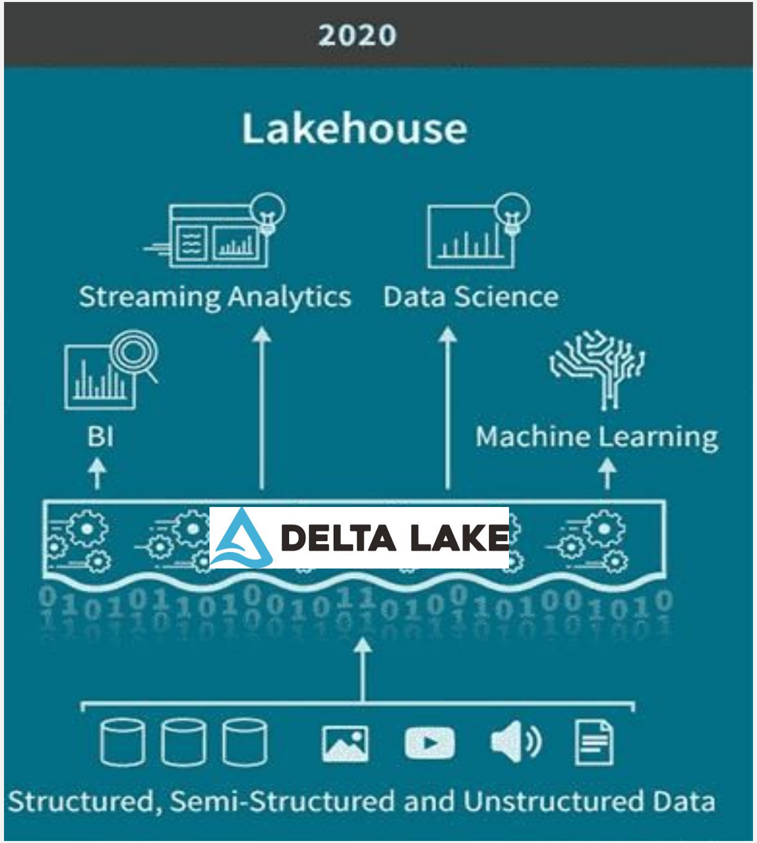
- Built for business intelligence and reporting
- Support for data consistency and quick ad-hoc queries.
- However, they're unable to store unstructured raw data (which are crucial for modern machine learning uses)

# Data Lake



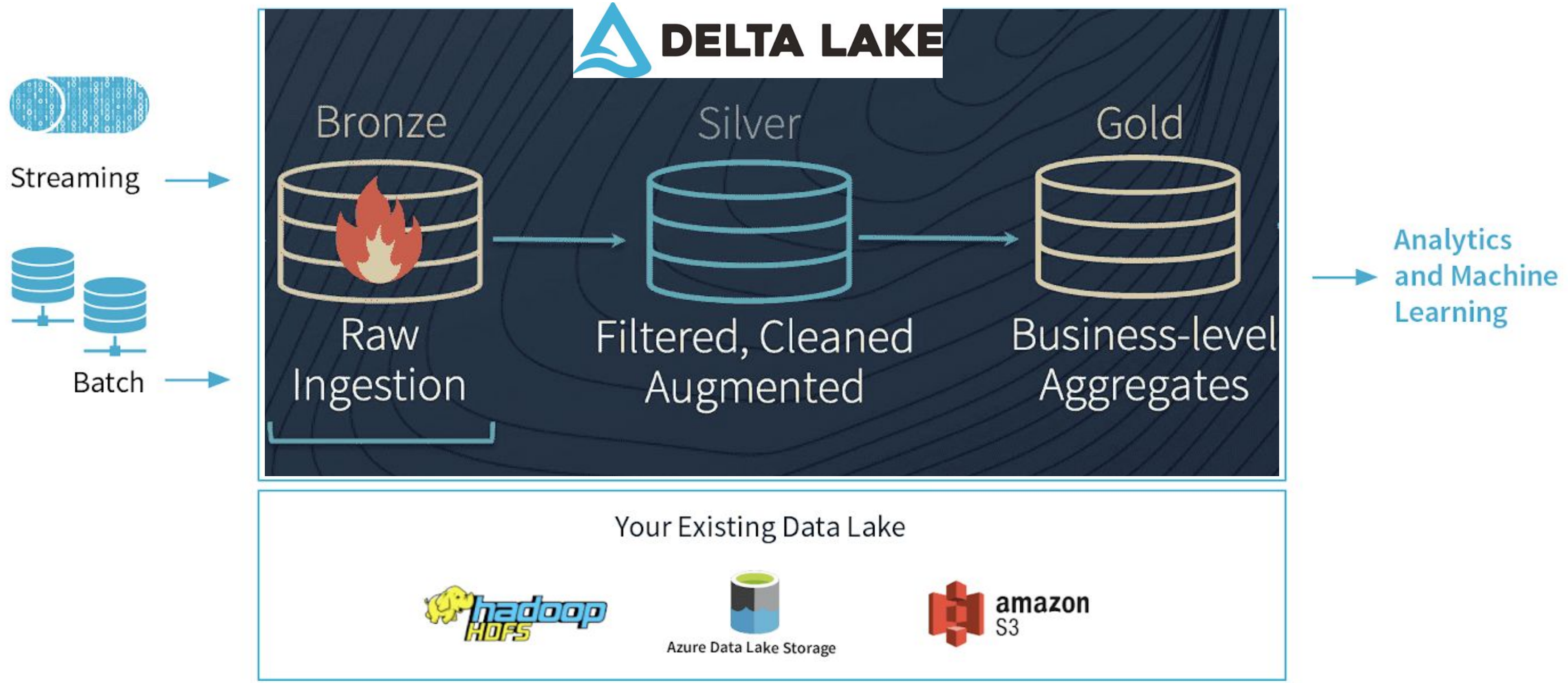
- Can store structured and unstructured data in a variety of formats (Parquet and ORC are popular formats)
- Easy to access without the need of additional data stores
- However, the lack of data governance causes data corruption, inconsistent queries, and overall confusion!

# Delta "Lakehouse"



- It used a **structured transaction layer** built on top of a Data Lake, using Delta Lake technology
- It enables database features, such as ACID transactions, data versioning, and schema management
- The "Lakehouse" provides Data Warehousing performance with Data Lake low-costs
- Builds trust in the data

# Delta Architecture

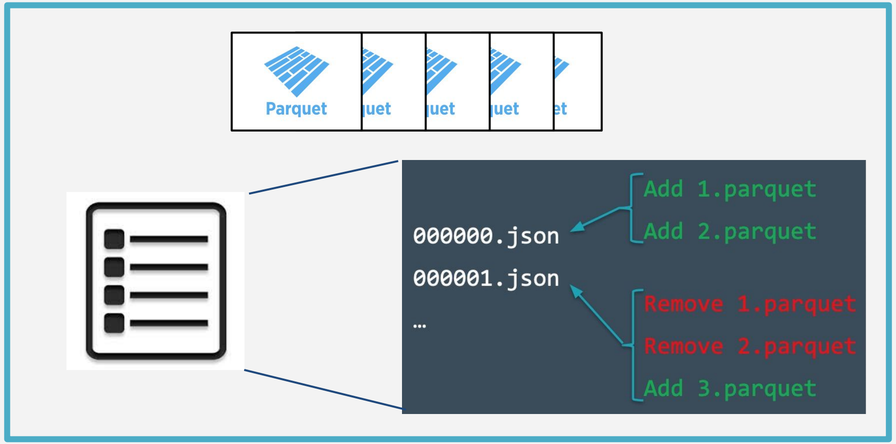


https://databricks.com/blog/2019/08/14/productionizing-machine-learning-with-delta-lake.html

# Delta Tables

A **Delta Table** is a collection of data kept using the Delta Lake technology and consists of three things:

- **Parquet files** containing the data inside object storage
- Delta **transaction log** kept with the Delta files in object storage
- A table registered in the **Metastore**



# Delta Lake Further Reading:

[Delta Lake Comprehensive and Readable documentation]

# Exercise: Delta Tables

Exercise in notebook files: **04_Delta_Tables**

# Lunch Break - 30 mins

# AGENDA

LIXAR
Fueled by BDO

## Morning Session

9:30 – 10:00 AM | Environment Setup

10:00 – 12:30 PM | Data Engineering in Azure Databricks

- Databricks Overview
- Databricks Notebooks
- Databricks leveraging Azure Storage and Azure Key Vault
- Delta Lake Architecture and Delta Tables

12:30 - 1:00 PM | LUNCH BREAK

## Afternoon Session

1:00 – 2:30 PM | Spark Machine Learning in Azure Databricks

- Machine Learning Overview
- Linear Regression with Spark Machine Learning
- MLflow Tracking
- Classification Machine Learning - Fraud Detection

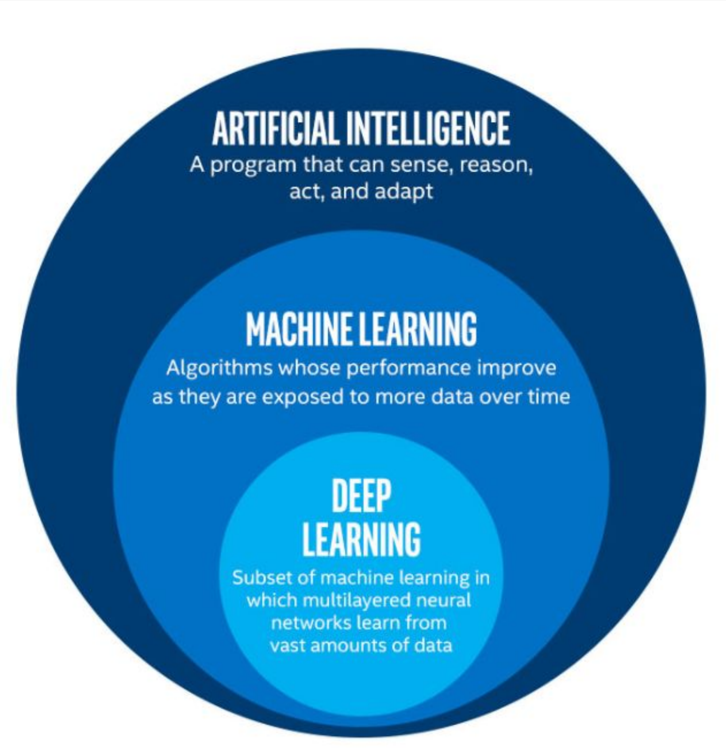
2:30 – 3:00 PM | Q&A

## Reminder - restart Databricks cluster if it’s gone to sleep

# Machine Learning overview

**Machine Learning** (ML) is a type of **Artificial Intelligence** (AI) that enables a system to learn from data rather than through explicit programming.

There needs to a signal in your data on how to predict what you are looking for


ARTIFICIAL INTELLIGENCE
A program that can sense, reason, act, and adapt
MACHINE LEARNING
Algorithms whose performance improve as they are exposed to more data over time
DEEP LEARNING
Subset of machine learning in which multilayered neural networks learn from vast amounts of data

https://www.edureka.co/blog/ai-vs-machine-learning-vs-deep-learning/

# Supervised vs Unsupervised Learning

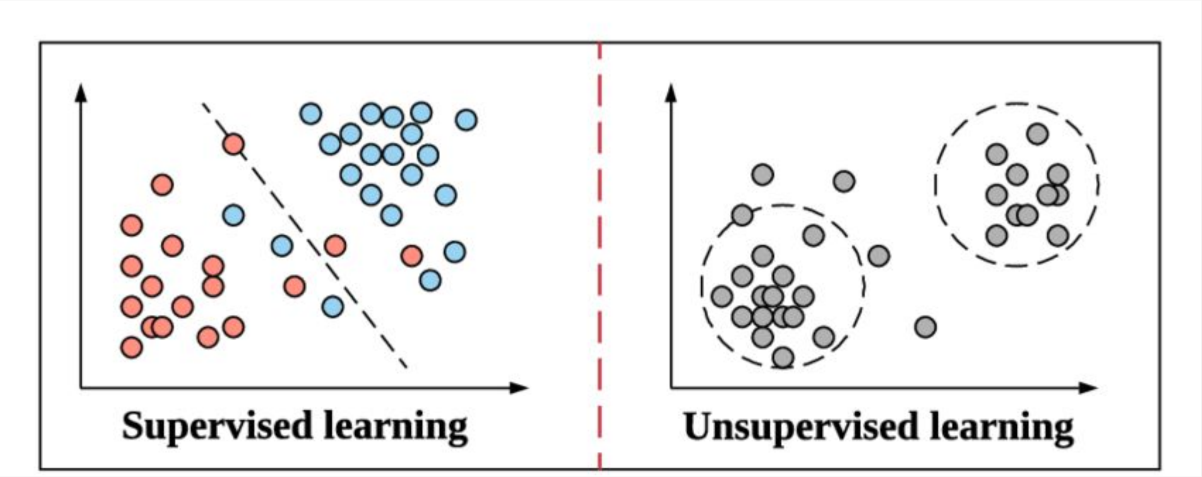
LIXAR Fueled by BDO

**Supervised learning:** Uses a training dataset to learn how to predict a desired output. An iterative process until the prediction error has been sufficiently minimized.

- Examples: Image recognition, fraud detection, predicting sports outcomes

**Unsupervised learning:** Analyze and cluster vast unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention

- Examples: clustering, categorize news articles, product recommendations

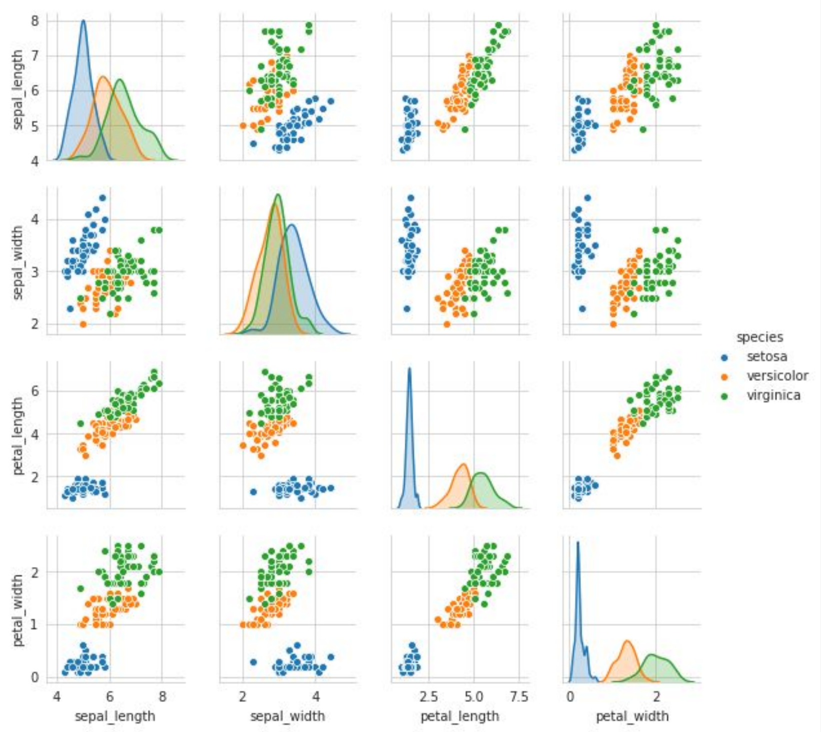


https://www.researchgate.net/figure/Examples-of-Supervised-Learning-Linear-Regression-and-Unsupervised-Learning_fig3_336642133

# Data Exploration

Before we decide which Machine Learning method to use, we need to first understand the data

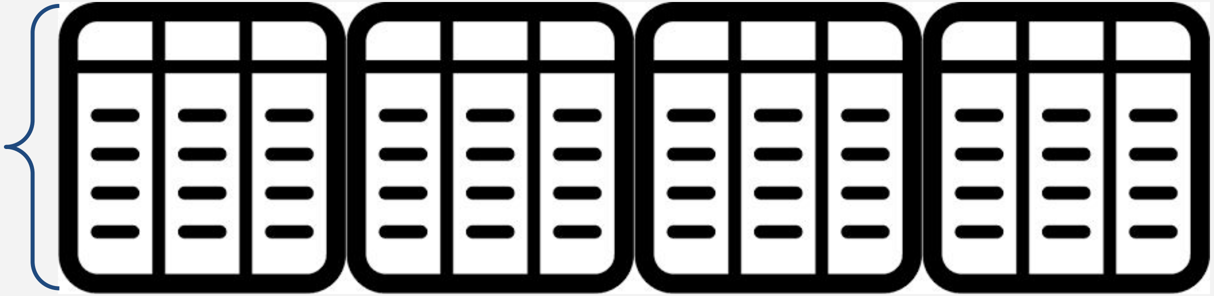
For supervised, what are relevant features for predicting a label?



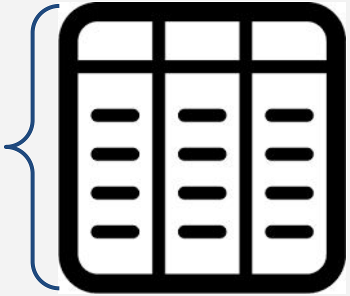
https://manisha-sirsat.blogspot.com/2019/12/exploratory-data-analysis.html

# Train vs Test split

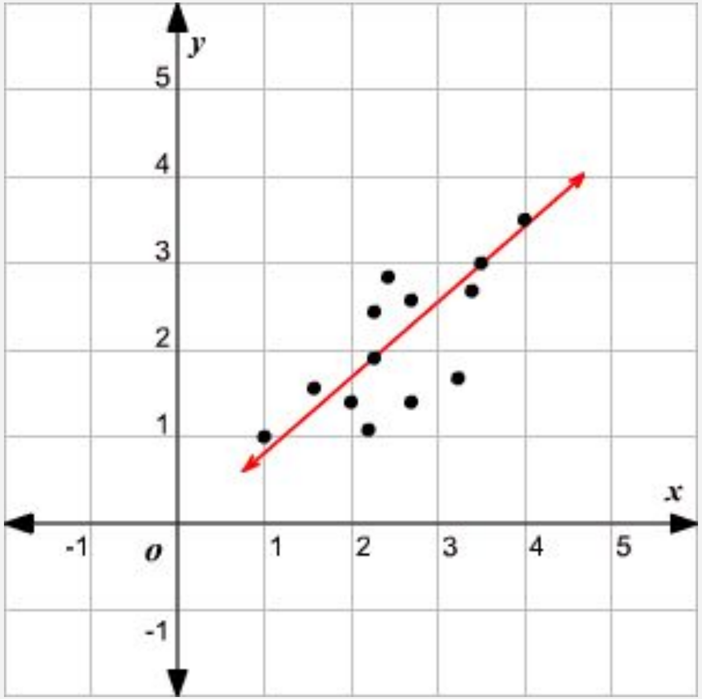
**Training Dataset:**
A larger portion of the data is used to train the model



**Test Dataset:**
A smaller portion of the data is used to test the model performance



# Linear Regression: Line of Best Fit



**Black dots:** True values
**Red line:** Line of best fit
Distance between is residuals

**The goal is to draw a line that minimizes the sum of the squared residuals, or Root Mean Squared Error**

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(observed_t - predicted_t)^2}$$

https://www.varsitytutors.com/hotmath/hotmath_help/topics/line-of-best-fit-eyeball-method

# Machine Learning Libraries

scikit learn

**Scikit-learn** is a popular single-node machine learning library

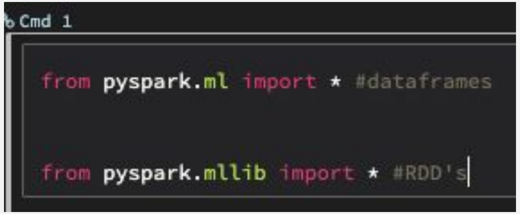
**But what if our data or model gets too big?**

APACHE Spark™

# Machine Learning in Spark

**Scale Out** and **Speed Up**

Machine learning in Spark allows us to work with bigger data and train models faster by **distributing the data and computations across multiple workers.**



```
from pyspark.ml import * #dataframes

from pyspark.mllib import * #RDD's
```

**Spark ML (older version called MLlib)**

ML API

Based on DataFrames

Supported API (MLlib in maintenance)

# Exercise: Linear Regression

Exercise in notebook files:
**05_Machine_Learning_Overview**

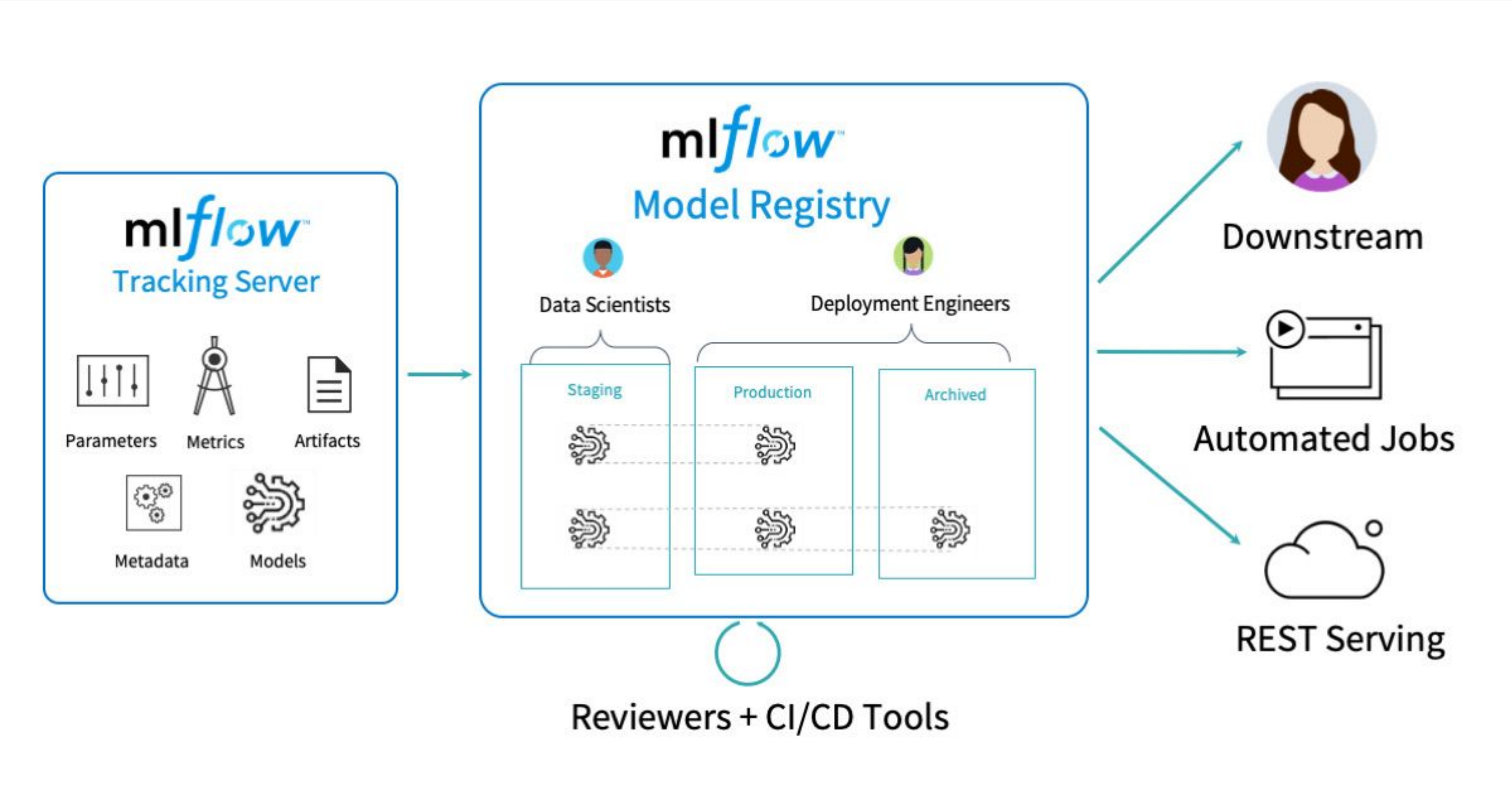
# Exercise: Feature Scaling and Pipelines

LIXAR

Fueled by BDO

Exercise in notebook files:
**06_Feature_Scaling_And_Pipelines**

# MLFlow: Platform for the Machine Learning lifecycle

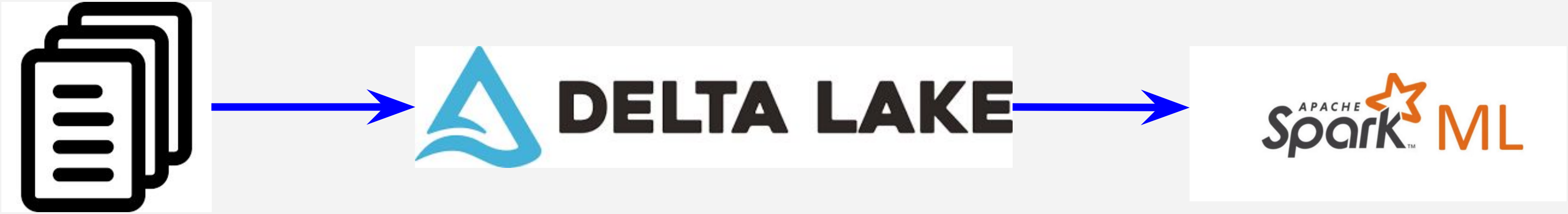


https://databricks.com/blog/2020/04/15/databricks-extends-mlflow-model-registry-with-enterprise-features.html

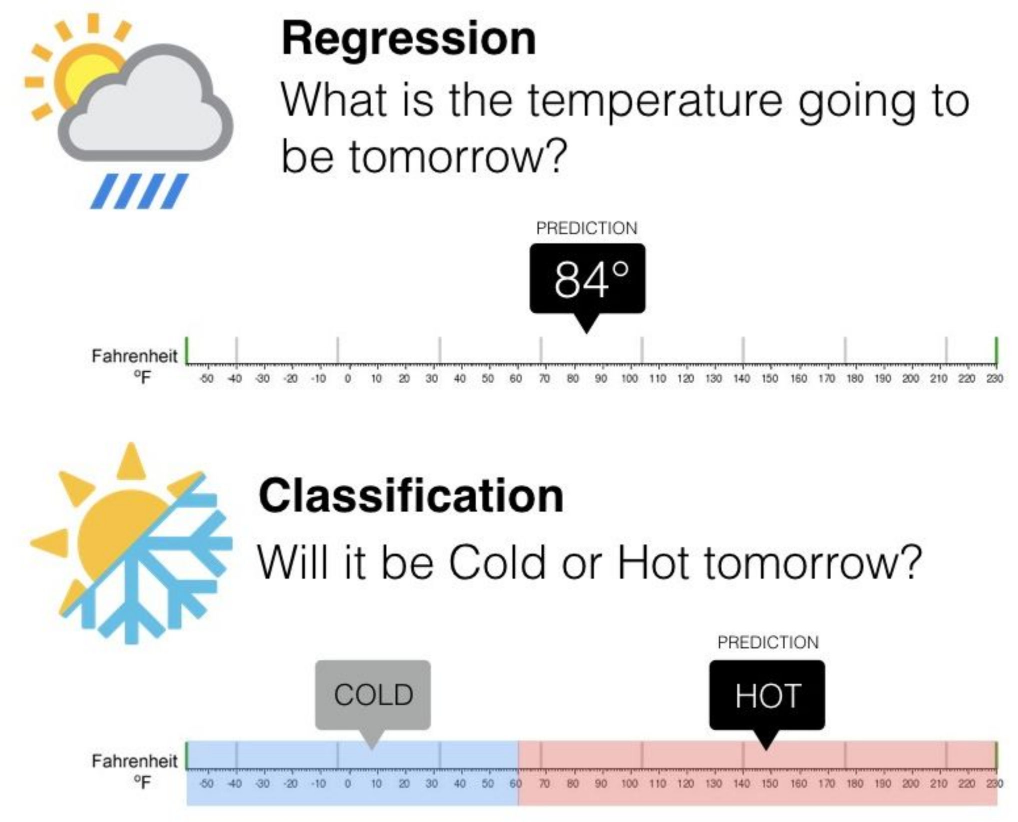
# Exercise: MLFlow

Exercise in notebook files:
**07_ML_Flow**

# Pathway for this workshop



# Regression vs Classification

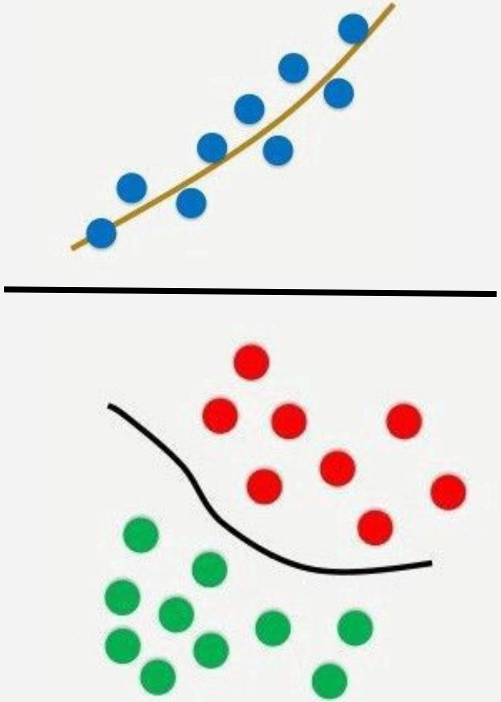


## Regression

What is the temperature going to be tomorrow?

## Classification

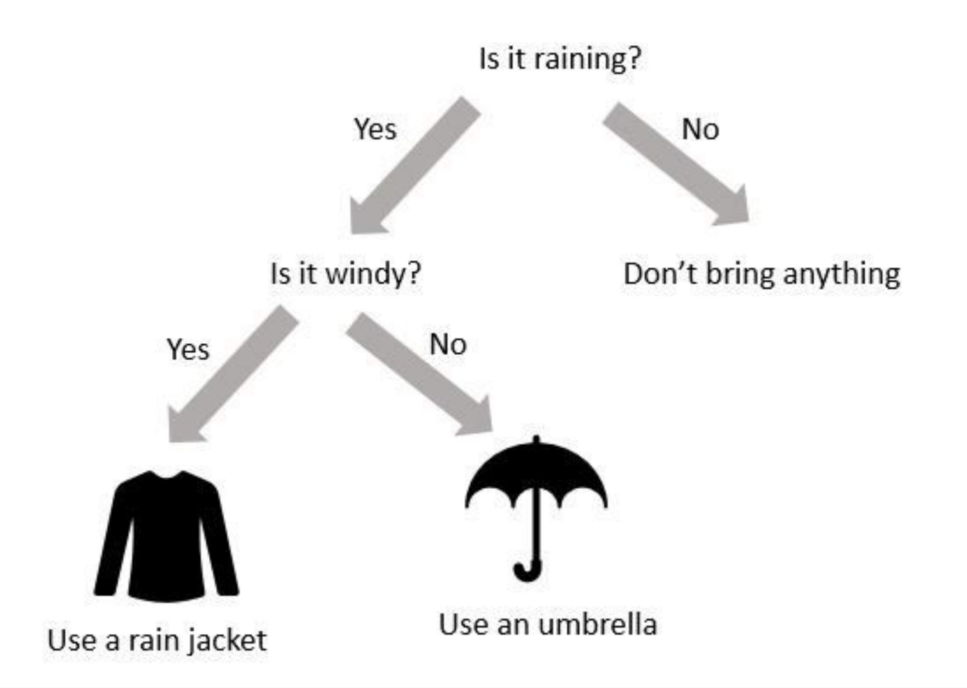
Will it be Cold or Hot tomorrow?



https://towardsdatascience.com/regression-or-classification-linear-or-logistic-f093e8757b9c

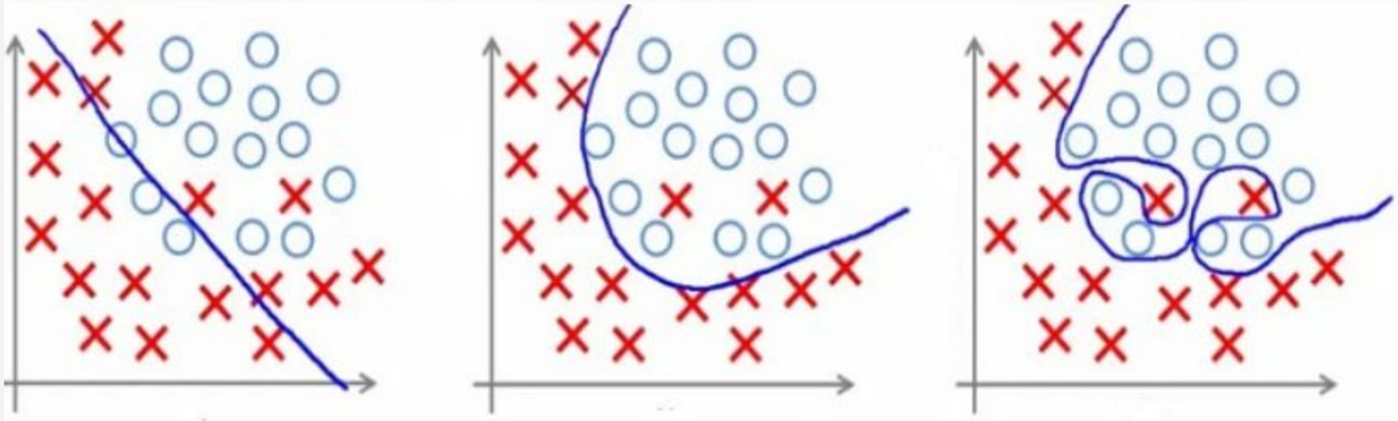
# Decision Trees

**Make a decision based on a set of criteria:**



https://medium.com/x8-the-ai-community/decision-trees-an-intuitive-introduction-86c2b39c1a6c

# Under-fitting vs Over-fitting



**Under-fitting**

(too simple to explain the variance)

**Appropriate-fitting**

**Over-fitting**

(forcefitting -- too good to be true)

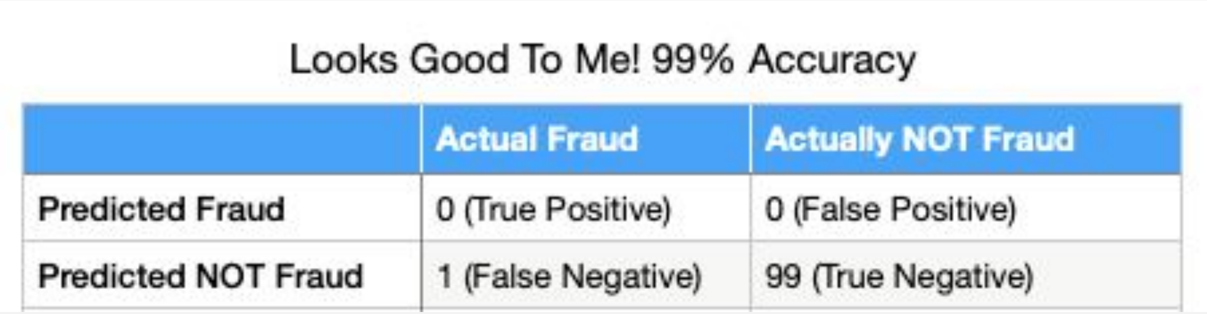
https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

# Classification ML - Fraud Detection

**Classification**: Using ML to predict a class, out of one or many classes

Can no longer use Root Mean Squared Error

In this example we have 100 data points, 1 one which is fraud, 99 which are not fraud



Looks Good To Me! 99% Accuracy

| | Actual Fraud | Actually NOT Fraud |
|---|---|---|
| Predicted Fraud | 0 (True Positive) | 0 (False Positive) |
| Predicted NOT Fraud | 1 (False Negative) | 99 (True Negative) |

# Recall And Precision (Subtle difference)

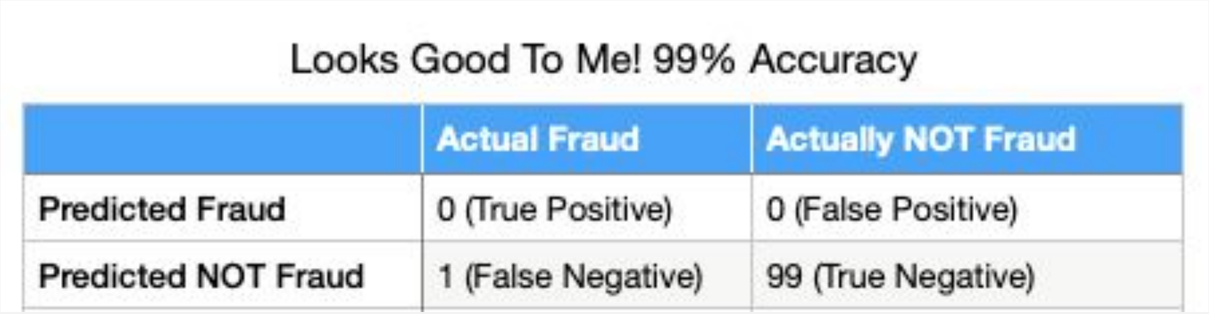
**Recall** = True Positive / (True Positive + **False Negative**) = 0/(0+1) = 0

Considering actual is positive, how often did we predict positive

**Precision** = True Positive / (True Positive + **False Positive**) = 0 / 0+0 = undetermined

We can determine it’s not very good

Considering our predicted positives, how often did we predicts positive.



Looks Good To Me! 99% Accuracy

| | Actual Fraud | Actually NOT Fraud |
|---|---|---|
| Predicted Fraud | 0 (True Positive) | 0 (False Positive) |
| Predicted NOT Fraud | 1 (False Negative) | 99 (True Negative) |

# Exercise: Classification

Exercise in notebook files:
**08_Classification_Fraud_Detection**

# Koalas - For a Specific Persona

Implementation of the pandas Dataframe API on top of Apache Spark

**Similar** to pandas API not exactly the same, however Koalas is much faster with large data sets

Not quite as performant as Spark ML due to Internal Frame overhead

Smaller Delta to refactor non-performant Pandas ML pipelines to Koalas

# Extra Notebooks Provided

Extra notebook files:

- **01_Update_Delta_Tables**
- **02_Azure_Data_Factory**
- **03_SQL_Database**
- **04_Koalas_API_for_Pandas**

# Future Learning With Databricks & Azure

Databricks Academy: http://academy.databricks.com/catalog

MS Learn: https://docs.microsoft.com/en-us/learn/

LIXAR

Fueled by BDO

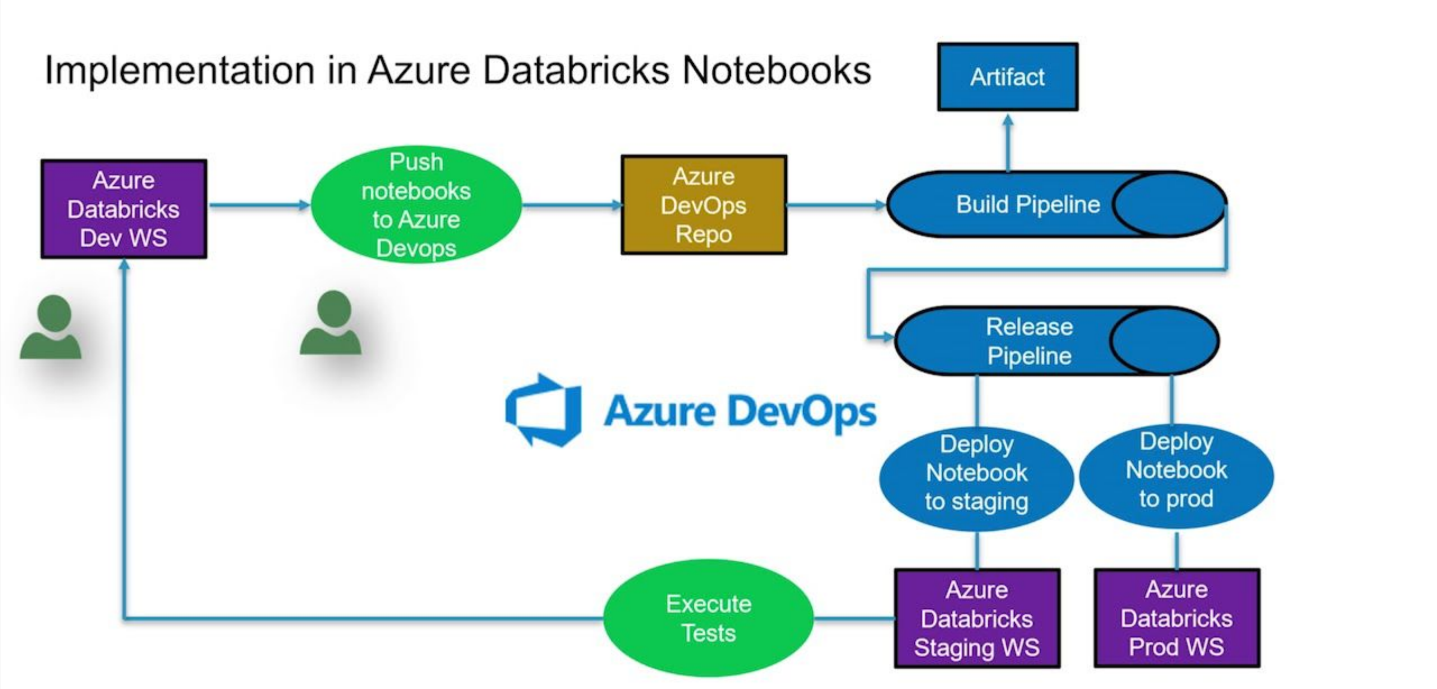
# DRIVEN BY DATA

WE LOVE WORKING WITH PEOPLE
WHO WANT TO BE INNOVATIVE

# Extra Slides

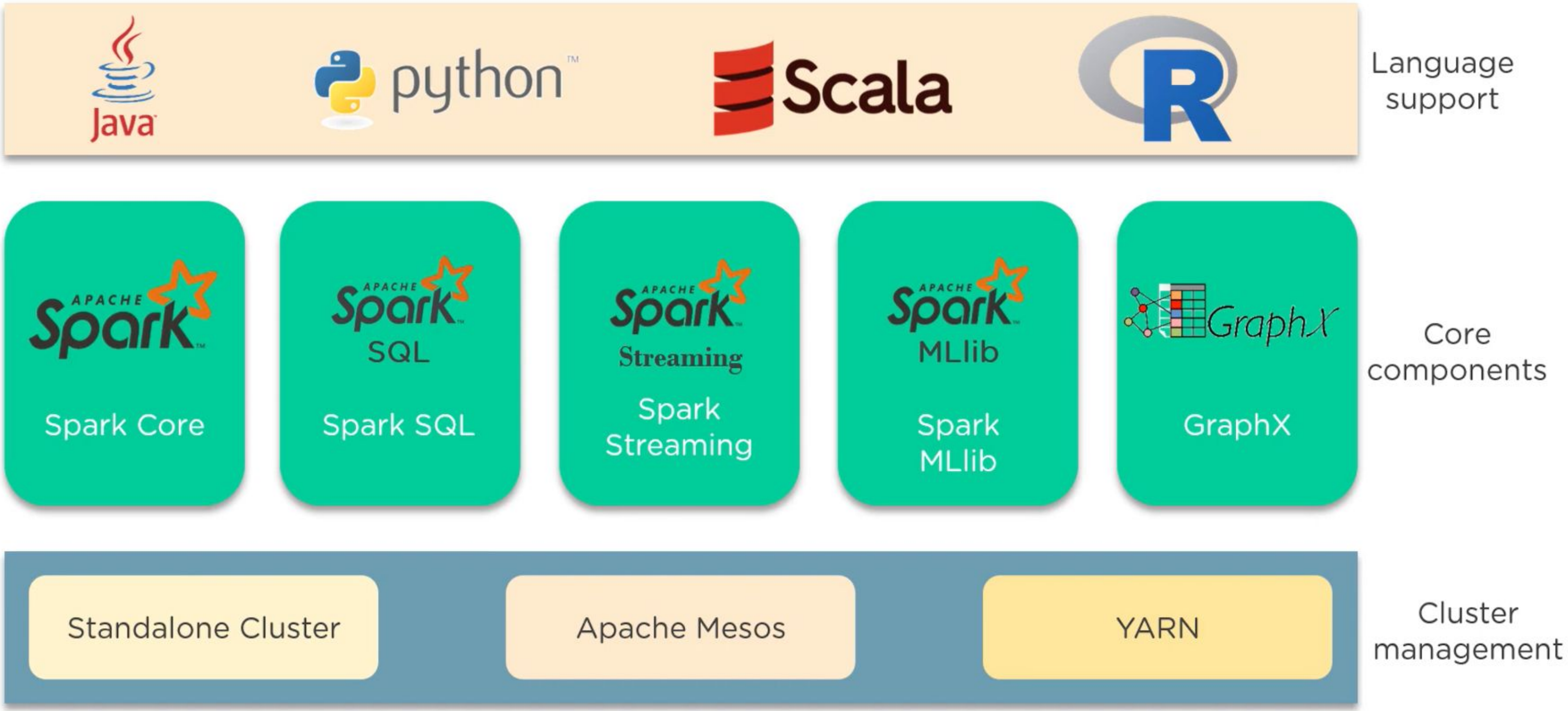
# Honorable Mention: Azure DevOps

In a production environment, you can connect your Databricks workspace to an Azure DevOps repo, and implement an automated release pipeline to test and deploy updates to your Databrick notebooks.



https://databricks.com/session/devops-for-applications-in-azure-databricks-creating-continuous-integration-pipelines-on-azure-using-azure-databricks-and-azure-devops

# Components of the Spark Ecosystem (Source: SimpliLearn)



Language support

Spark Core

Spark SQL

Spark Streaming

Spark MLlib

GraphX

Core components

Standalone Cluster

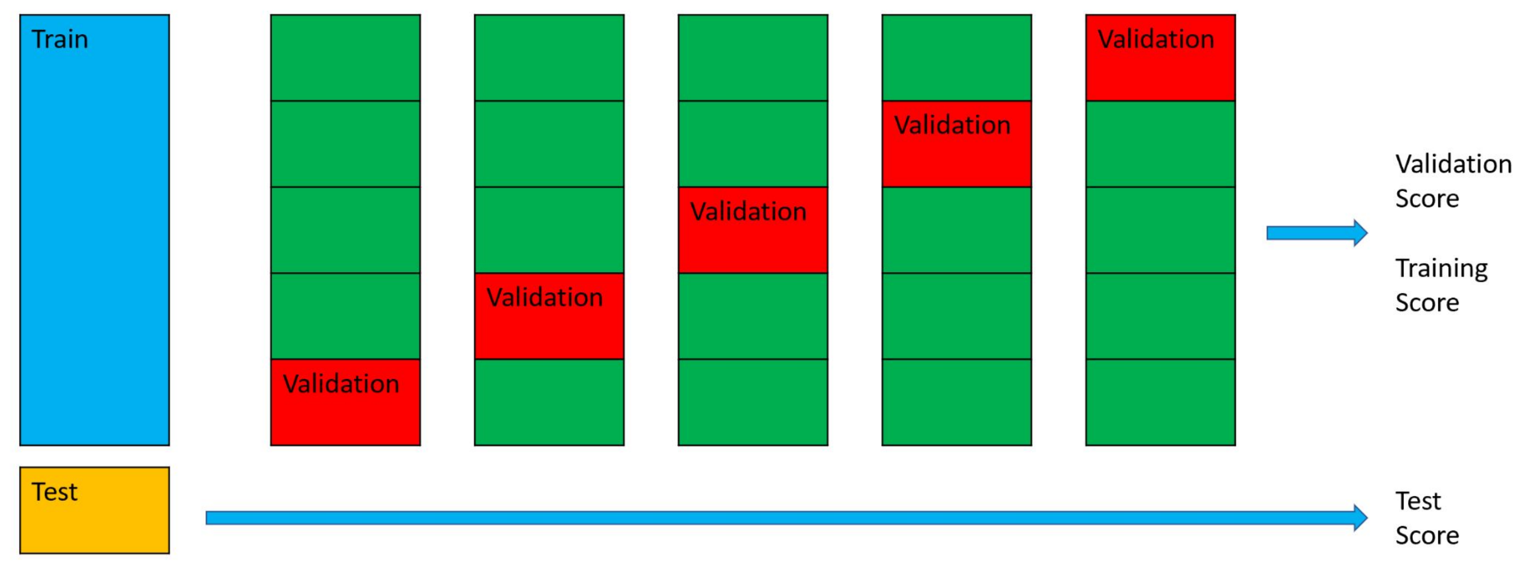
Apache Mesos

YARN

Cluster management

# Cross Validation

- Split the training dataset into smaller chunks
- Iterate through the chunks, each time leaving one out for the model training
  - Use the validation set to calculate the error of each iteration
- The optimal model parameters are the ones with the lowest average validation error



https://medium.com/analytics-vidhya/training-validation-and-test-set-in-machine-learning-7fab555c1080