Machine Learning at Scale: Introduction

Performing optimal Business Decisions using Large Data Sets

https://github.com/amjadraza/

https://www.datafy2ai.com/

LOUNGE

TABLE OF CONTENTS

Who am I and What am I doing?

Data Scientist/ Quant / Consultant

Muhammad Amjad Raza (PhD)

Muhammad Amjad Raza is an Electrical Engineer and Data Scientist with more than 10 years of professional experience in various industries. He is worked with The Portland House group as quant for four years.

.ai is the Internet country code top-level domain for Anguilla. It is administered by the government of Anguilla. It is popular with companies in the artificial intelligence industry

---

AI is not new

---

https://medium.com/@amjadraza24

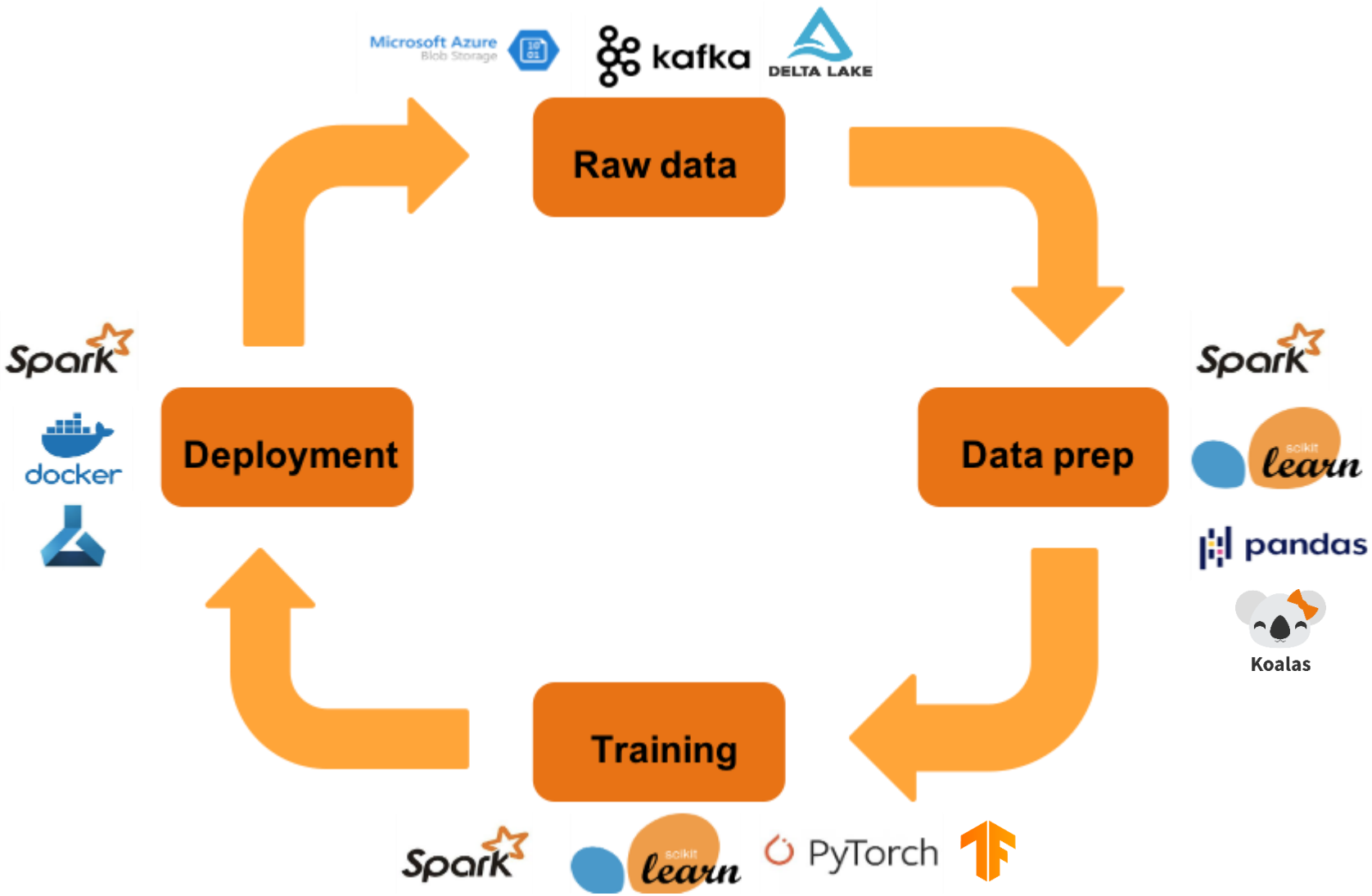
A bigger picture of a Problem

Data Pipelines

- Huge amount of data
- Data Storage
- Cleaning and Processing

ML & AI

- Preprocess of data
- Build Models
- Train the models
- Test Models
- Deploy the Model



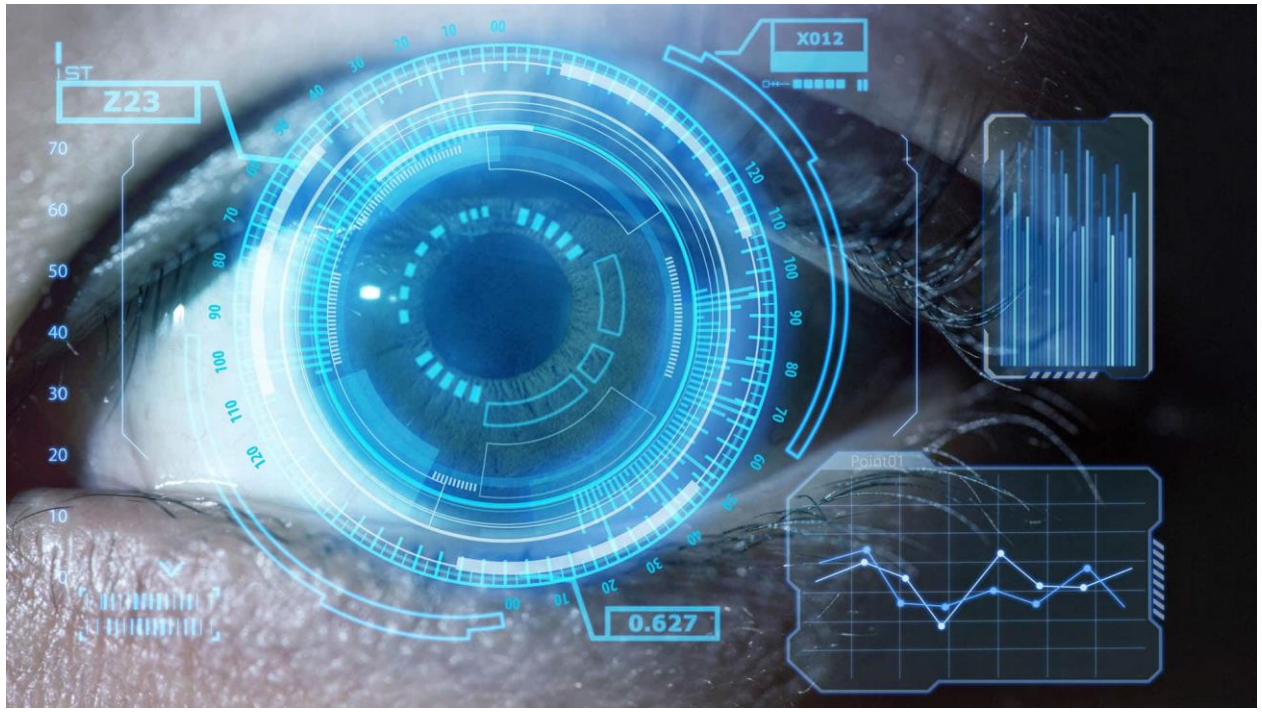
https://images.app.goo.gl/WAdpaX6RNZ9SfEvH9

Objective: A production ready, Scalable ML/AI pipelines for Big Data

Machine Learning & Artificial Intelligence

Machine Learning

- Finding the relationship using data
- Data is the key starting point
- Many libraries existed (SK-learn, TensorFlow, PyTorch, PyCaret)
- AutoML Frameworks
- Application of machine learning in various business domains



https://images.app.goo.gl/sUSQdrGsA5PwUsxe6

Artificial Intelligence

- Applications of ML to solve complex problems based on intuition of human brain
- Neural Networks based solutions
- A network is learned using data
- Data Quality vs quantity

Huge Amount of data is required to train a NETWORK

A Data Platform for ML/AI?

01. Data Generation

- Data is being generated at a rapid speed
- Various sources of data, text, pictures, video, tabular.
- Web, apps, clients
- Private vs Public data

02. Data Cleaning

- Cleaning the data
- Auxiliary data
- Data Governance
- Model the outliers
- Relevant data for a problem at hand

03. Data Storage

- Store the data securely
- Data bases
- Data Lakes
- Data Storage formats (HDFS, pickle, binary, csv, RDB)
- Local vs Cloud storage

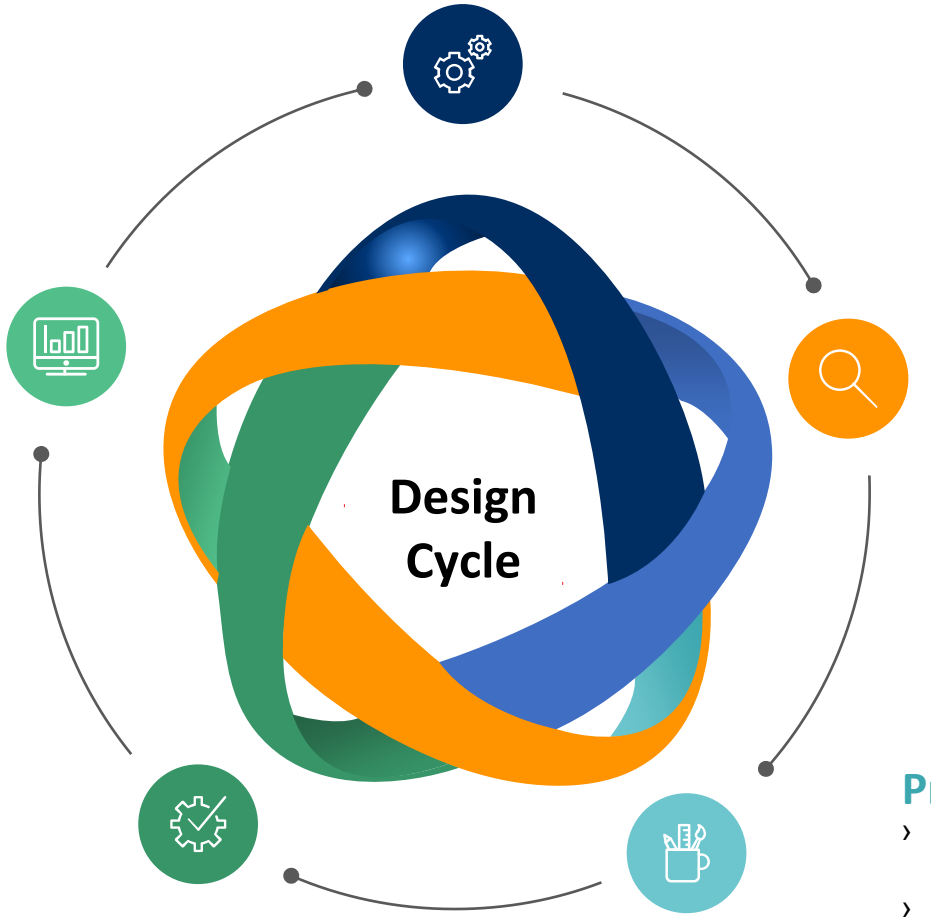
04. Data Ingestion

- Prepare the data compatible with Machine Learning models
- In-Memory data loading
- Load on demand
- Extract, transform and load pipeline

An Overview of ML Framework

Exploratory Data Analysis

- Data Engineering & Governess
- Distributions of data
- Can we believe the data?



**Feature Engineering

- Relationship between data tables
- Extracting the features from existing data
- Creating new features using domain specific knowledge

Preparing Data for Model

- Preparing the training, validation and test data
- Imputing missing values if necessary, with some logic

Model Training

- Train the model
- Find optimal/best model
- Save the model in pickle

Deploy the Model for Production

- Deploy the Model on Cloud or in-house for production.
- Keep tracking of real time performance of model

Feature Engineering

Understanding the Training Data

1. Size of training data
   - Billions of rows of data
   - Can't handle on simple home computers
2. Identify Target variable
   - Identify the target variable.
   - Study the distribution of target and features.
   - Transformations
3. Problem Identified
   - Handling missing data
   - Handling text/other data

Preparing the Data For Training

1. Training data set
   - Prepare the training data 80:20 or any other split
   - Future information Leakage
2. Validation set
   - Validation set is used to choose the best model
   - Usually, part of training set
3. Test Set
   - Unseen data to test the model out of sample
   - Representative samples of training data

Machine Learning Algorithms

Supervised Learning

1 Labeled Data

- A column with Target Variable
- Labels of targets available
- A data is collected and labeled

2 Features

- Defining problem is simple
- Selection of the performance measure is easier
- Wrong labels may lead to misleading results

3 Example Problems

- Linear Regression
- Multi Layer Perceptron
- GBT/Light GBM/XGBoost
- LSTM etc

Un-supervised Learning

1 Labeled Data

- Target Variables are not given
- Data Collection is probably easier
- A challenging problem

2 Features

- Defining problem is Challenging
- Selection of the performance measure is case dependent
- Optimal solution

3 Example Problems

- K-means Clustering
- Auto-Encoders
- Gaussian Mixtures

Multiprocessing vs Distributed Framework?

Multiprocessing Framework

1 Setting up manually

- A lot of overhead in setting up
- Not optimal, often issues
- Python multiprocessing libraries

2 Issues

- Not scalable
- Bad memory management
- Resources are not fully utilized

3 Examples

- `joblib` python package
- `Fibre` by Uber engineering
- `multiprocessing` python

Distributed Processing Framework

1 Distributed Processing

- A native solution for parallel processing
- Better task distribution and management
- A scalable solution

2 Issues

- Learning the tools
- Infrastructure building
- Debugging

3 Examples

- MapReduce
- Spark (Python, Scala, R)
- Hadoop Storage layer

Why We need Big Data Platforms?

Distributed Data Processing

1 Amount of data

- Billions of rows of data
- Can't handle on simple home computers

2 Distributed Storage

- Store data in a distributed way
- Loading would be easier

3 Distributed Processing

- Process the data on demand.
- Better memory/processor utilization
- Scala, Spark etc.

Distributed Machine Learning

1 Distributed training Platforms

- ML algorithms to handle distributed data
- Support platform

2 Models handle big data

- Do not break and handle the batches of data
- Can restart training
- GPU training

3 Deploying the Models

- Deploying models to handle huge traffic
- Dedicated solutions
- Monitoring the models

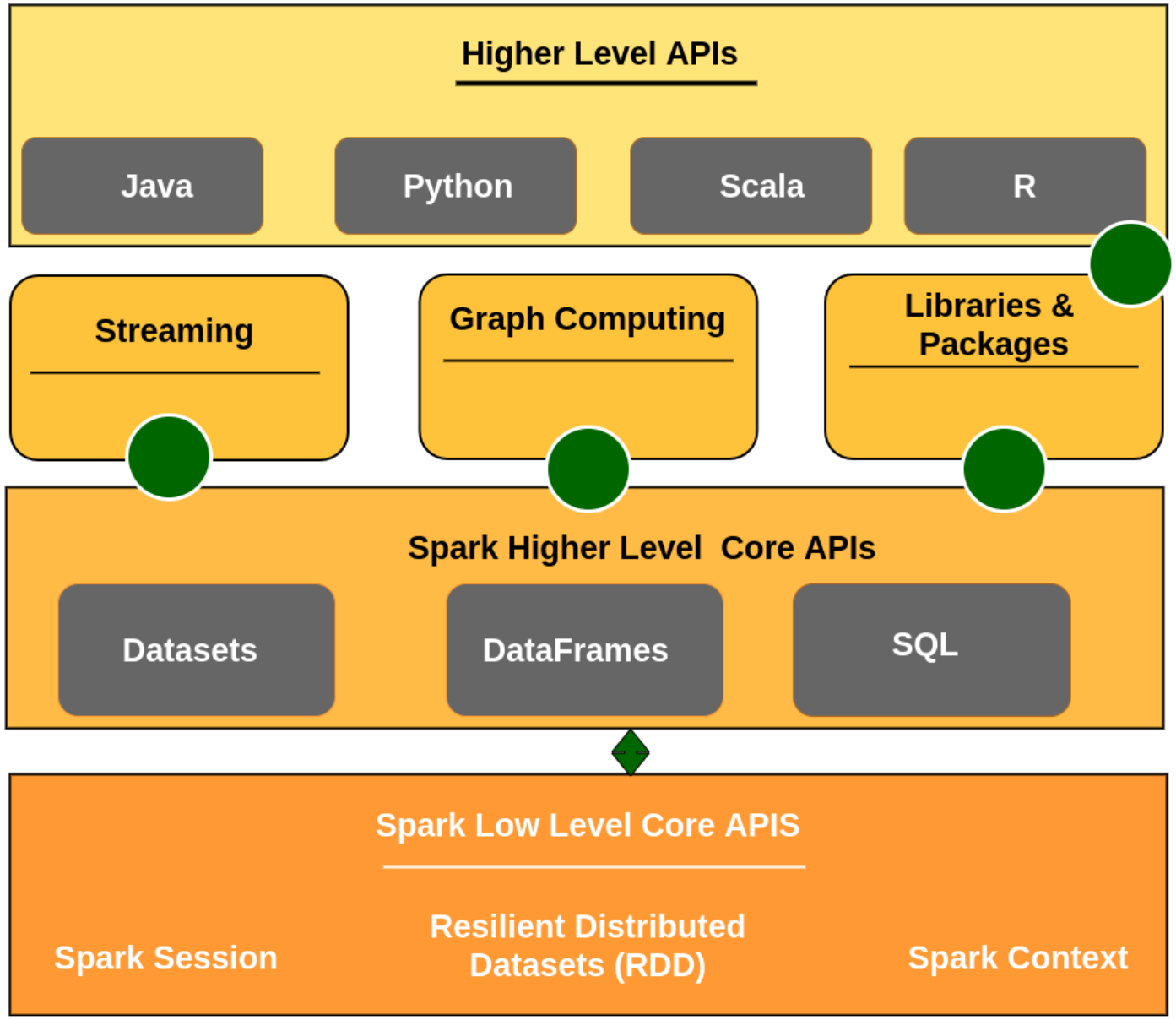
SPARK: Distributed Processing Framework

Distributed Processing

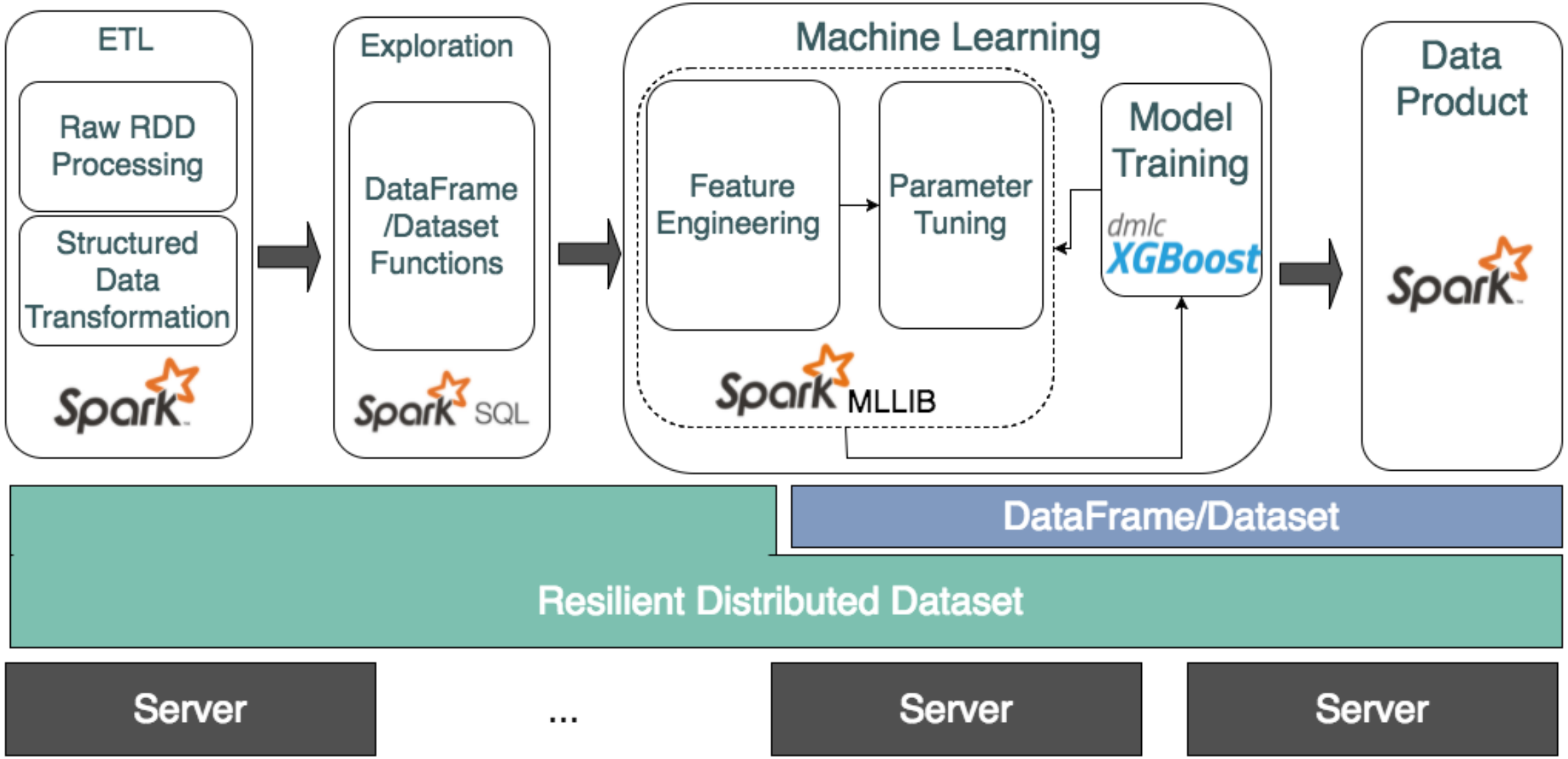
- A unified Computing Engine
- Local vs Cluster mode processing
- Scalable distributed processing
- The most actively developed open-source projects
- A de-facto engine for distributed processing

Distributed ML & AI

- Spark Machine Learning Library
- TensorFlow training on Spark
- Serving applications using Spark
- On-demand resource management for training and deploying
- Supports PYTHON API
- HDFS data storage solutions
- Data Transformation

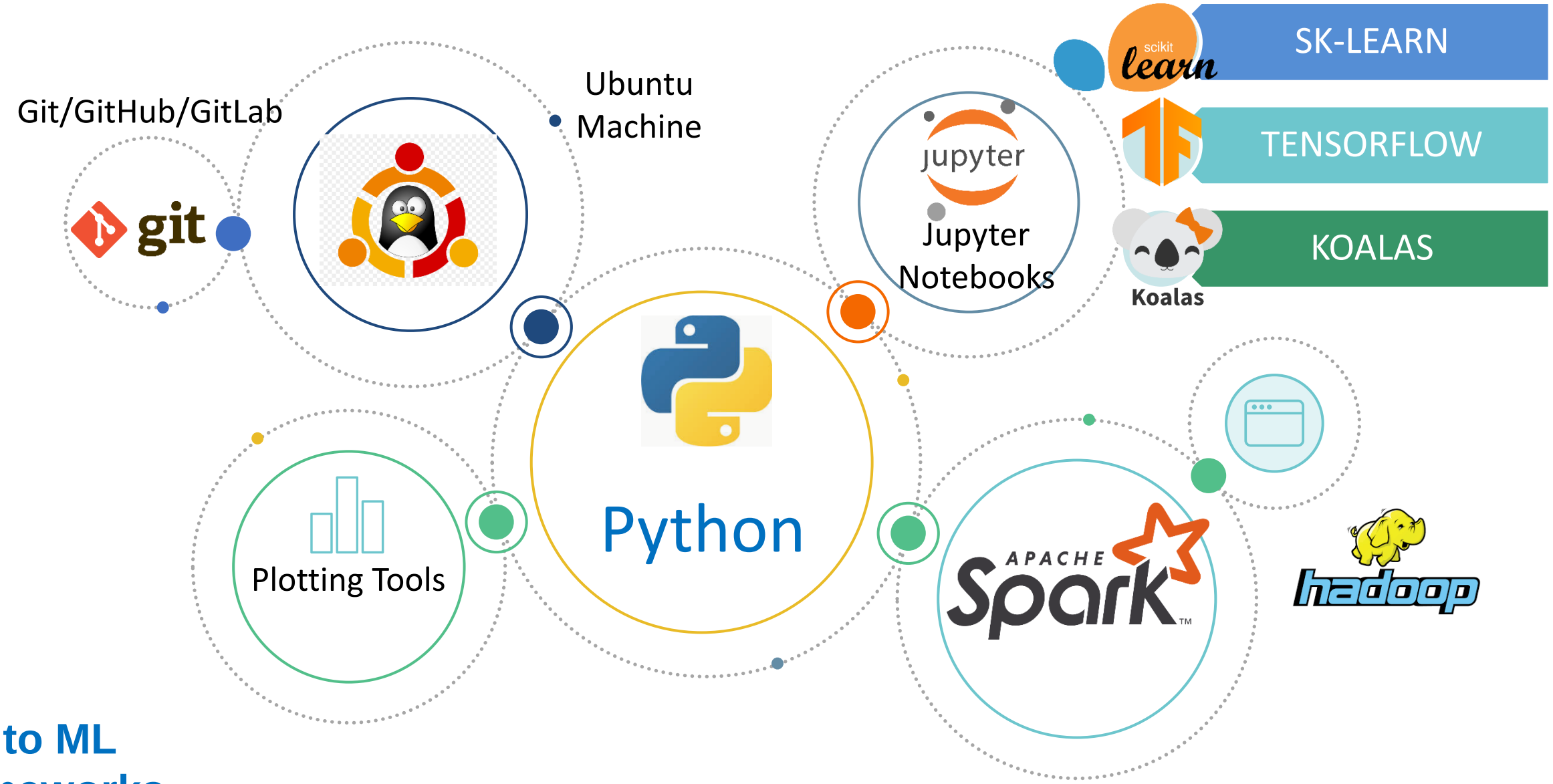


SPARK: Distributed Processing Framework



https://dmlc.github.io/2016/10/26/a-full-integration-of-xgboost-and-spark.html

Tools/Technologies for Big Data Processing



**Auto ML Frameworks

Applications: AI & ML with Spark

01

Stock Market / Financial Data Modelling

Processing historical data using spark and machine learning

02

Image Classification

Training large scale vision models using spark & TensorFlow (CNN, RNN)

03

Processing Text Data

Processing text data require distributed data processing to build models. GPT-3 used GPU and Distributed training

04

Developing Outlier Detection System

Outlier detection system with online feed training

05

Recommendation System

Serving Millions of customers and updating models on the fly.

06

Retail Analytics

Supply Chain, Racking, Sales Forecast, Customer behavior, .

Learning Resources: AI at Scale

MS	AI at Scale by MICROSOFT	https://www.microsoft.com/en-us/research/project/ai-at-scale/
GOOG	Training at Scale	https://cloud.google.com/ai-platform/training/docs/training-at-scale
SPARK	Distributed Processing with Spark	https://spark.apache.org/docs/latest/
Python	Python Ecosystem for ML/AI	Outlier detection system with online feed training
Kaggle	Real World Problem Solving	https://www.kaggle.com/
Data Bricks	Spark De-Factor Cloud Solution	https://databricks.com/

Hands-ON

Asking a right question is an art. IMO, right question contains half the solution

Happy to Connect on LinkedIn: https://www.linkedin.com/in/amjadraza/