Introduction to
Big Data

Date : 26-09-2020 | Speaker : Ayon Roy | Event : Hack the Mountain

Visit - AYONROY.ML

Hello **Buddy!**

I am **Ayon Roy**

B.Tech CSE (2017-2021)

Data Science Intern @ **Lulu International Exchange**, Abu Dhabi

(**World's Leading Financial Services Company**)

Brought **Kaggle Days Meetup** Community in India for the 1st time

If you haven't heard about me yet, you might have been living under the rocks. Wake up !!

Agenda (26-09-2020)

- What is Big Data ?
- Why should we focus on Big Data now ?
- Properties of Big Data
- Applications of Big Data
- Introduction to Apache Spark (A very famous name in Big Data Ecosystem)
- How Apache Spark's architecture looks like?
- How we can do Machine Learning with Big Data?
- Resources to get started with Big Data

What is Big Data ?

Defining Big Data

Big data is a domain that analyzes, extracts information from huge datasets which maybe beyond the ability of general tools to manage, process data.

Volume : Scale of Data

Variety : Different types of Data

Velocity : Speedy Ingestion of new Data

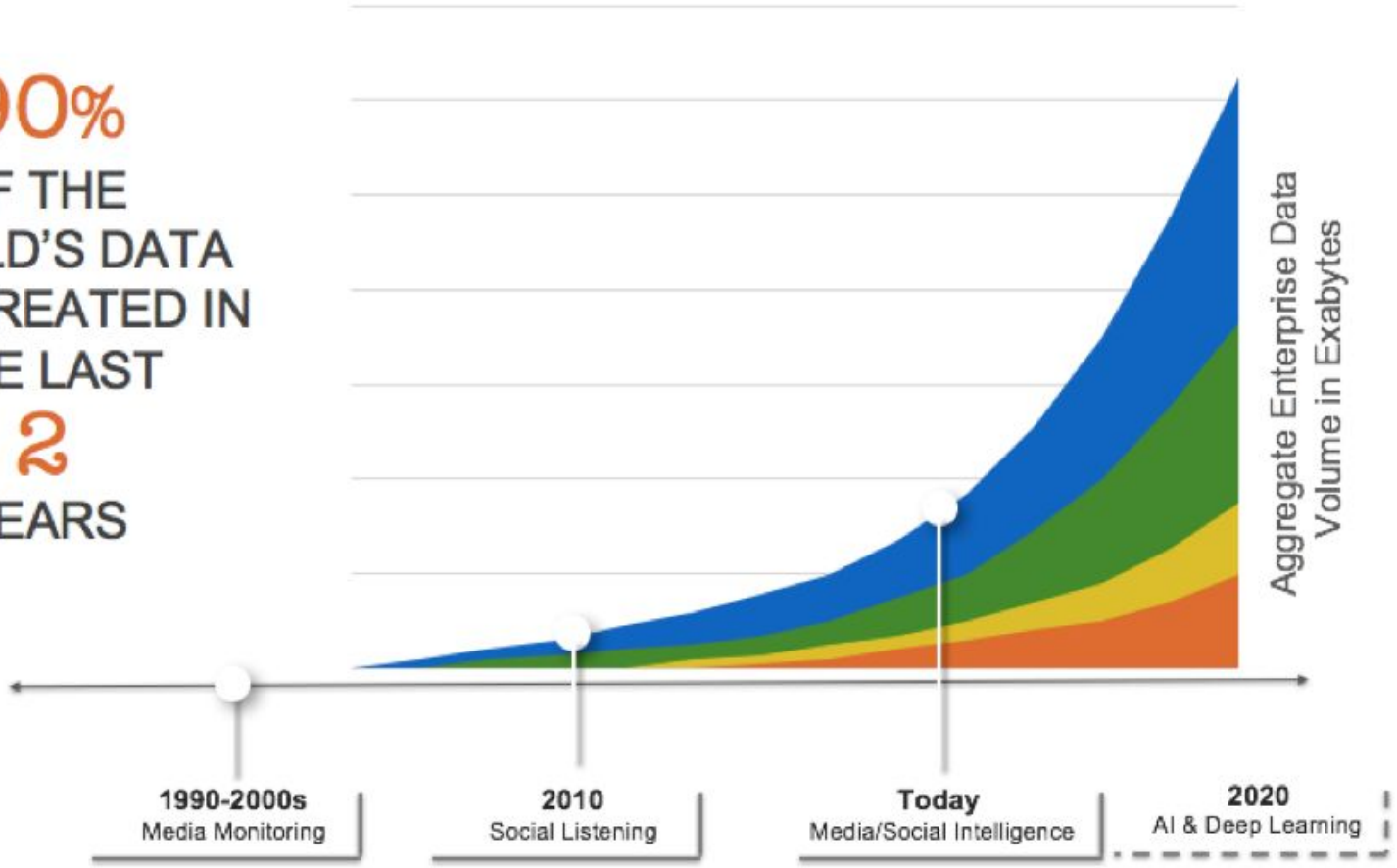
Veracity : Uncertainty in the Data

90%

OF THE WORLD'S DATA WAS CREATED IN THE LAST

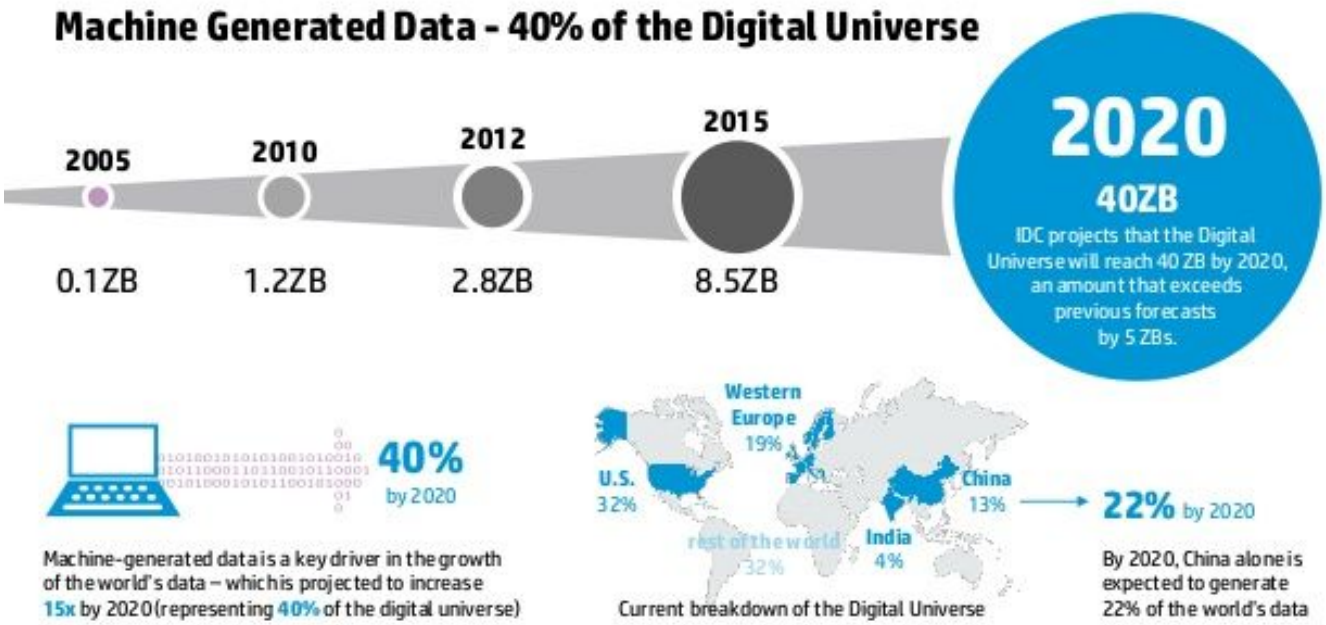
2

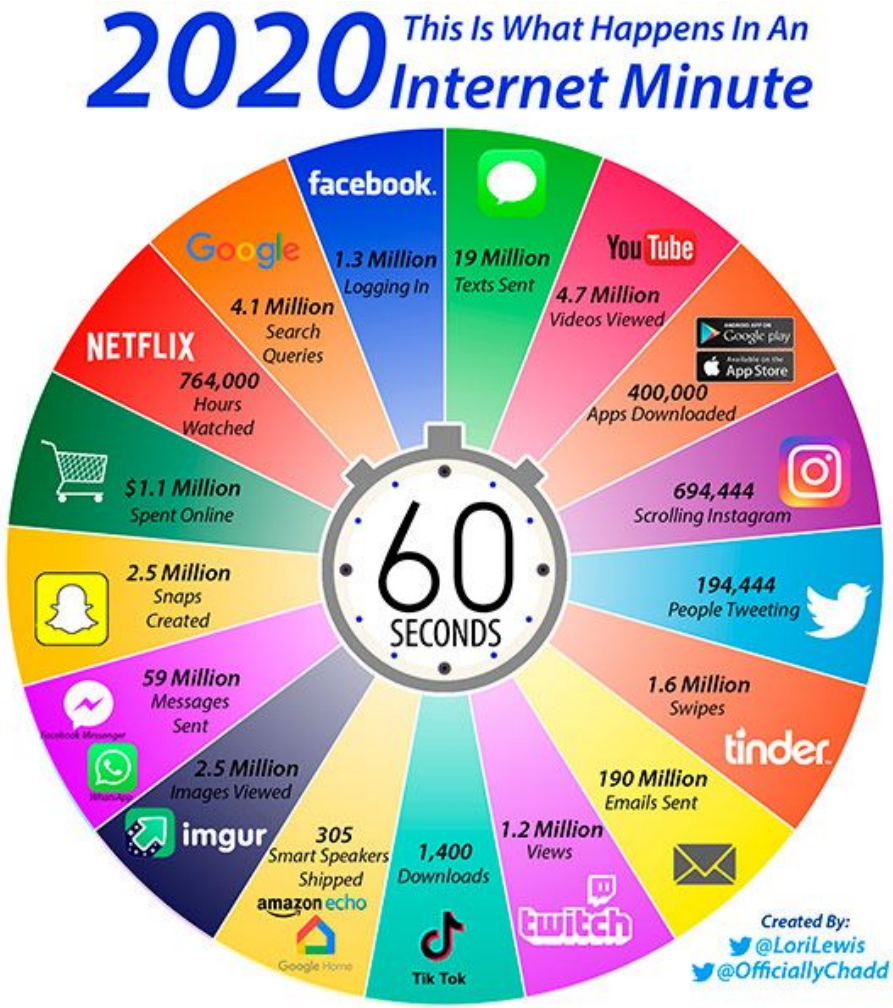
YEARS



Properties of
Big Data

Volume





Velocity

Variety

Veracity

Applications of Big Data

Retail/Consumer

- Merchandizing and market basket analysis
- Campaign management and customer loyalty programs
- Supply-chain management and analytics
- Event- and behavior-based targeting
- Market and consumer segmentations

Finances & Frauds Services

- Compliance and regulatory reporting
- Risk analysis and management
- Fraud detection and security analytics
- Credit risk, scoring and analysis
- High speed arbitrage trading
- Trade surveillance
- Abnormal trading pattern analysis

Web and Digital media

- Large-scale clickstream analytics
- Ad targeting, analysis, forecasting and optimization
- Abuse and click-fraud prevention
- Social graph analysis and profile segmentation
- Campaign management and loyalty programs

Health & Life Sciences

- Clinical trials data analysis
- Disease pattern analysis
- Campaign and sales program optimization
- Patient care quality and program analysis
- Medical device and pharmacy supply-chain management
- Drug discovery and development analysis

Telecommunications

- Revenue assurance and price optimization
- Customer churn prevention
- Campaign management and customer loyalty
- Call detail record (CDR) analysis
- Network performance and optimization
- Mobile user location analysis

Ecommerce & customer service

- Cross-channel analytics
- Event analytics
- Recommendation engines using predictive analytics
- Right offer at the right time
- Next best offer or next best action

A lot of things can be done using Machine Learning ?

What is Apache Spark ?

Apache Spark is a unified analytics engine for large-scale data processing. It provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs.

It also supports a rich set of higher-level tools including **Spark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Structured Streaming for incremental computation and stream processing.**

History of Spark APIs

Distribute collection of JVM objects

Functional Operators (map, filter, etc.)

Distribute collection of Row objects

Expression-based operations and UDFs

Logical plans and optimizer

Fast/efficient internal representations

Internally rows, externally JVM objects

Almost the "Best of both worlds": **type safe + fast**

But slower than DF
Not as good for interactive analysis, especially Python

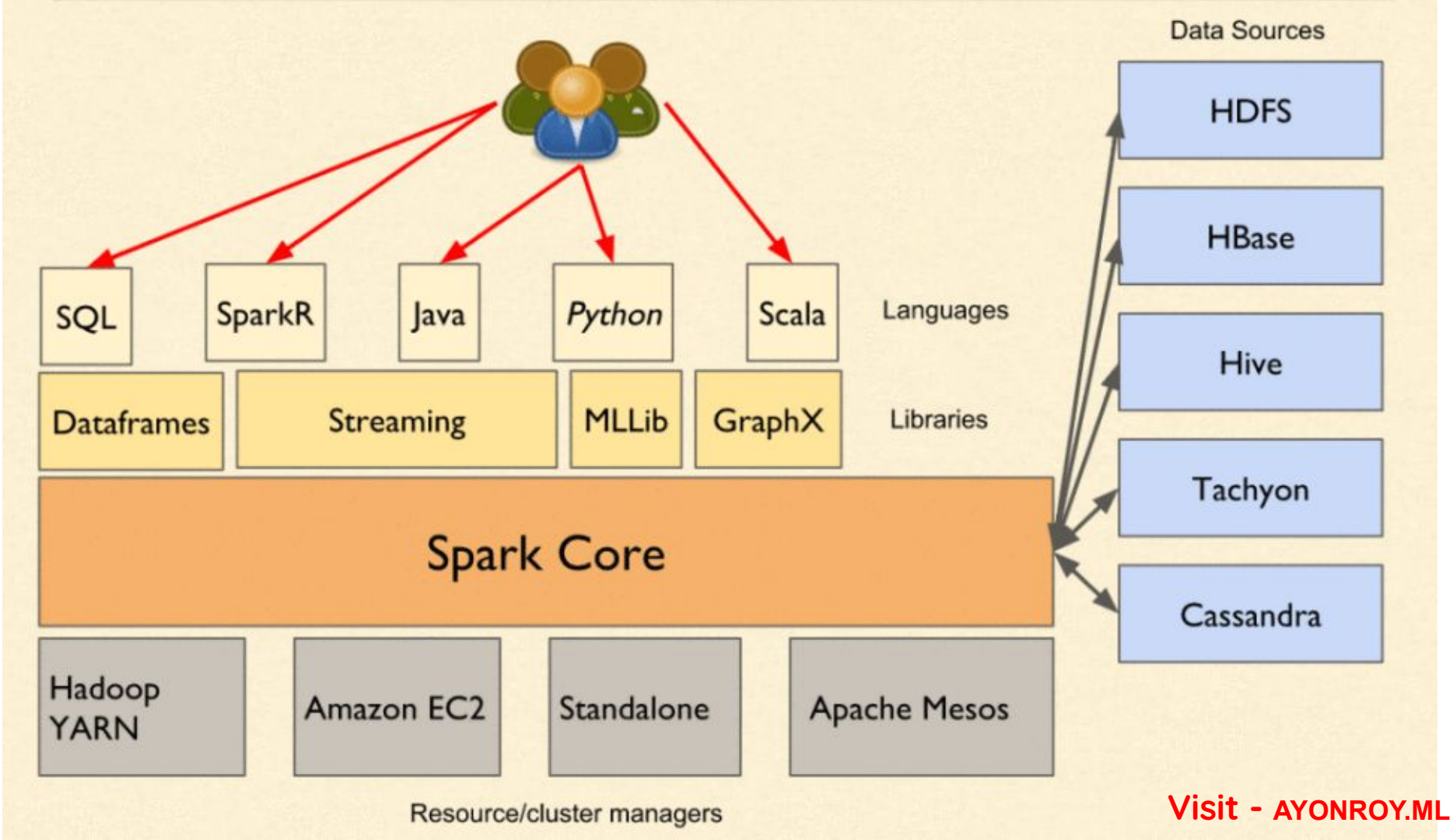
databricks

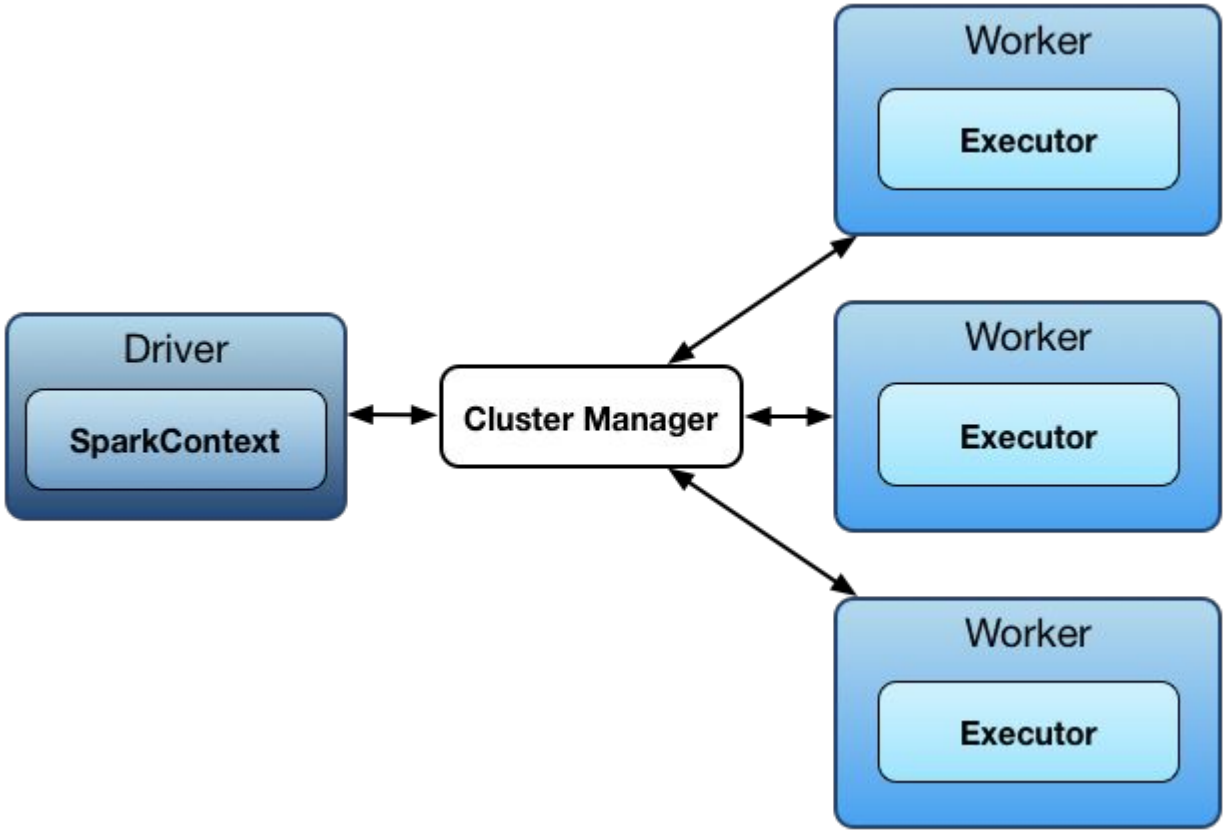
Feature	RDD	DataFrame	Dataset
Immutable	Yes	Yes	Yes
Fault Tolerant	Yes	Yes	Yes
Type-Safe	Yes	No	Yes
Schema	No	Yes	Yes
Execution Optimization	No	Yes	Yes
Optimizer Engine	N/A	Catalyst Engine	Catalyst Engine
API Level for manipulating distributed collection of data	Low	High	High
language Support	Java, Scala, Pyt	Java, Scala, Python, R	Java, Scala

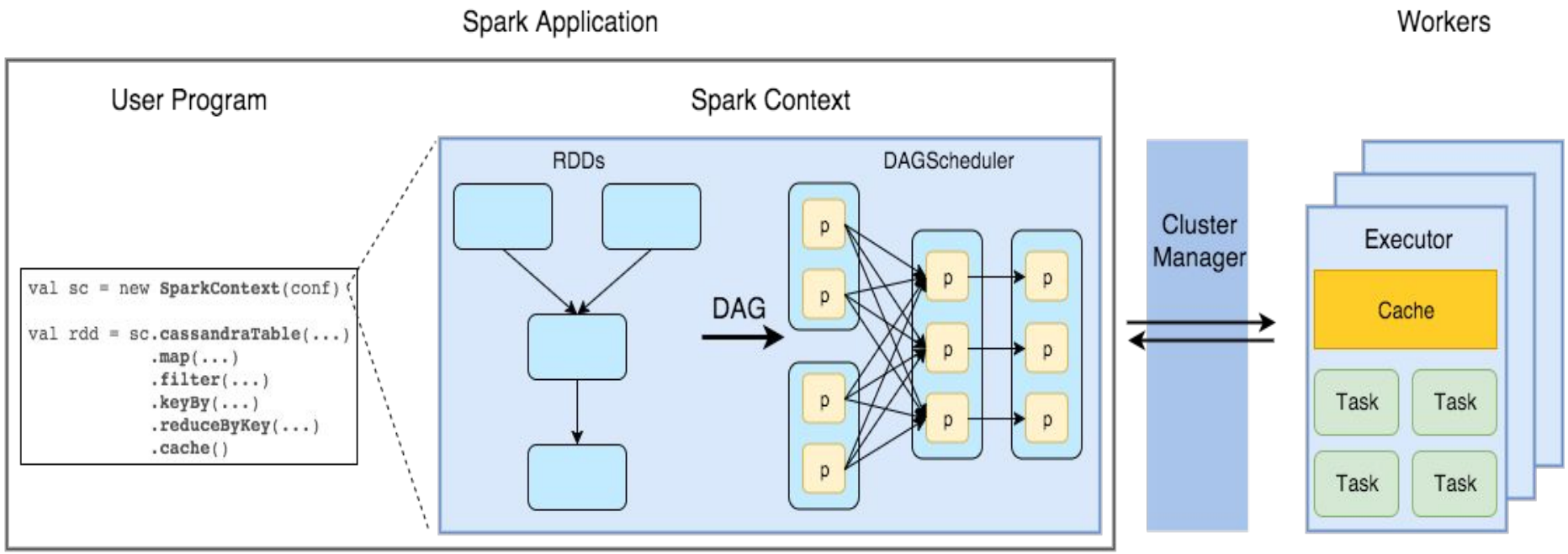
Here's how the

Spark's Architecture

looks like







- **Spark Context:** It holds a connection with Spark cluster manager. All Spark applications run as independent set of processes, coordinated by a SparkContext in a program.

- **Driver :** A driver is incharge of the process of running the main() function of an application and creating the SparkContext.

- **Executor** : Executors are worker nodes' processes in charge of running individual tasks in a given Spark job. They are launched at the beginning of a Spark application and typically run for the entire lifetime of an application.

- **Worker** : A worker, on the other hand, is any node that can run program in the cluster. If a process is launched for an application, then this application acquires executors at worker node.

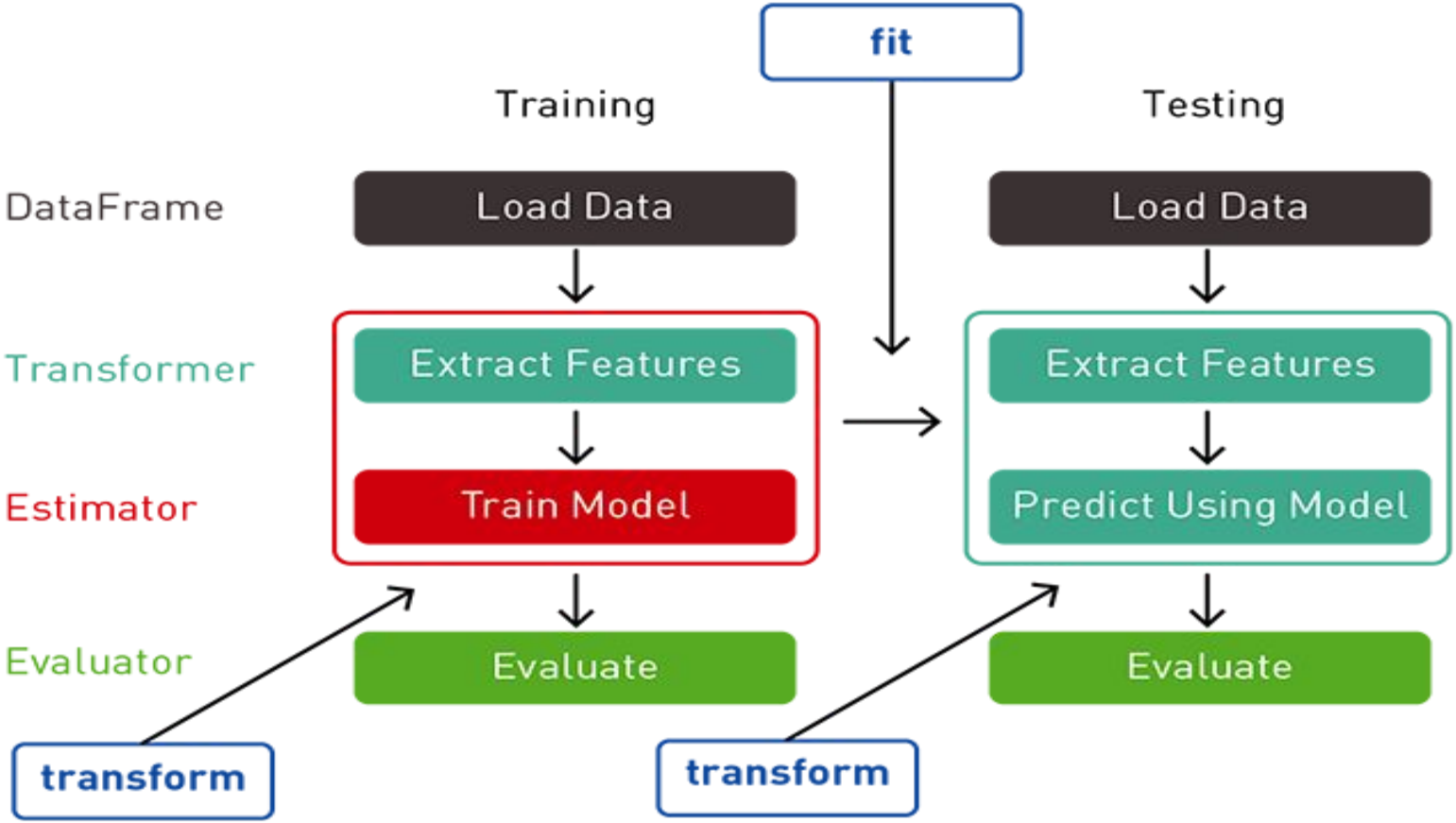
- **Cluster Manager:** Cluster manager allocates resources to each application in driver program. There are three types of cluster managers supported by Apache Spark – Standalone, Mesos and YARN.

How Spark's ML library will help us achieve our goal to fuse Big Data & Machine Learning?

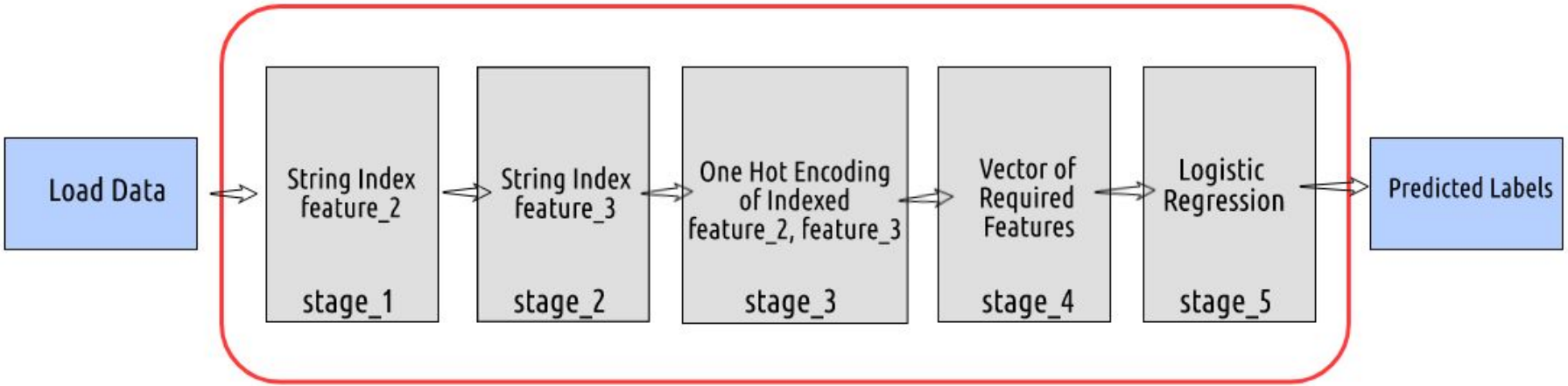
MLlib is Spark's machine learning (ML) library. Its goal is to make practical machine learning scalable and easy. At a high level, it provides tools such as:

- **ML Algorithms** : Common learning algorithms such as classification, regression, clustering, and collaborative filtering
- **Featurization** : Feature extraction, transformation, dimensionality reduction, and selection
- **Pipelines** : Tools for constructing, evaluating, and tuning ML Pipelines
- **Persistence** : Saving and load algorithms, models, and Pipelines
- **Utilities** : Linear algebra, statistics, data handling, etc.

Spark ML Workflow



- **DataFrame**: This ML API uses DataFrame from Spark SQL as an ML dataset, which can hold a variety of data types. E.g., a DataFrame could have different columns storing text, feature vectors, true labels, and predictions.
- **Transformer**: A Transformer is an algorithm which can transform one DataFrame into another DataFrame. E.g., an ML model is a Transformer which transforms a DataFrame with features into a DataFrame with predictions.
- **Estimator**: An Estimator is an algorithm which can be fit on a DataFrame to produce a Transformer. E.g., a learning algorithm is an Estimator which trains on a DataFrame and produces a model.
- **Pipeline**: A Pipeline chains multiple Transformers and Estimators together to specify an ML workflow.



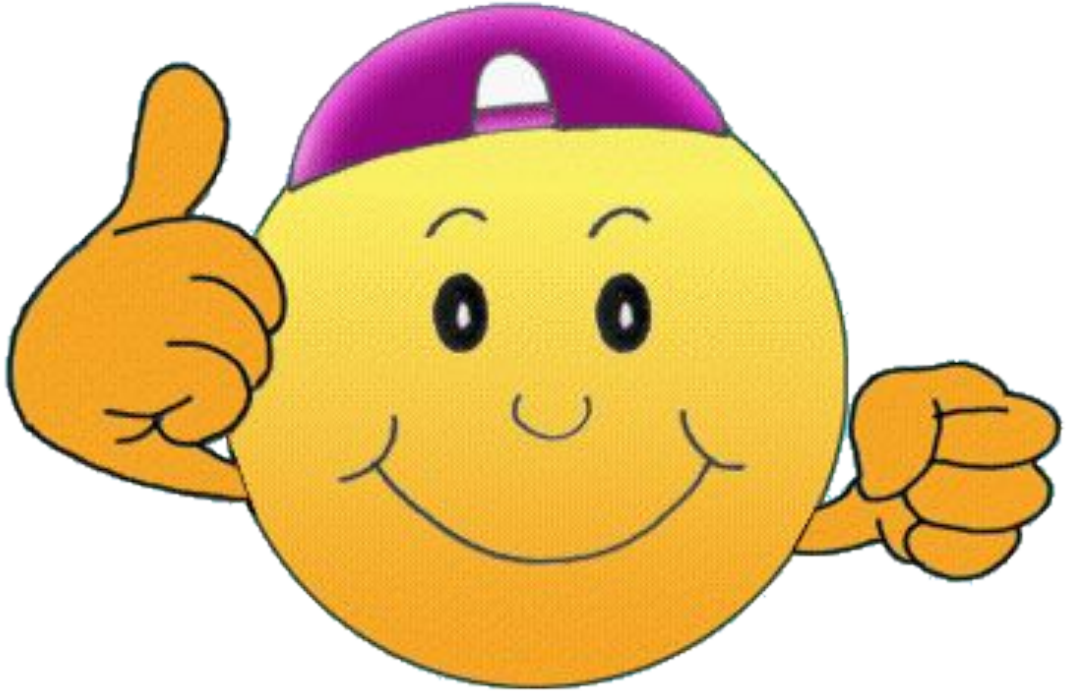
Spark Machine Learning library MLlib contains the following applications –

- Collaborative Filtering for Recommendations – Alternating Least Squares
- Logistic Regression, Lasso Regression, Ridge Regression, Linear Regression and Support Vector Machines (SVM).
- Linear Discriminant Analysis, K-Mean and Gaussian,
- Naïve Bayes, Ensemble Methods, and Decision Trees.
- PCA (Principal Component Analysis) and Singular Value Decomposition (SVD).

A few useful resources

1. https://spark.apache.org/
2. https://spark.apache.org/mllib/
3. https://docs.databricks.com/getting-started/spark/machine-learning.html
4. https://www.coursera.org/specializations/big-data
5. https://www.edx.org/course/big-data-analytics-using-spark
6. https://www.datacamp.com/community/tutorials/apache-spark-tutorial-machine-learning

GO FOR IT !



GOOD LUCK !

Let me answer your Questions now.

Finally, it’s your time to speak.

Danke Schoen

Questions ? Any Feedbacks ? Did you like the talk?
Tell me about it.

If you think I can help you,
connect with me via

Email : ayon.roy2000@gmail.com

LinkedIn / Github / Telegram Username : ayonroy2000

Website : https://AYONROY.ML/