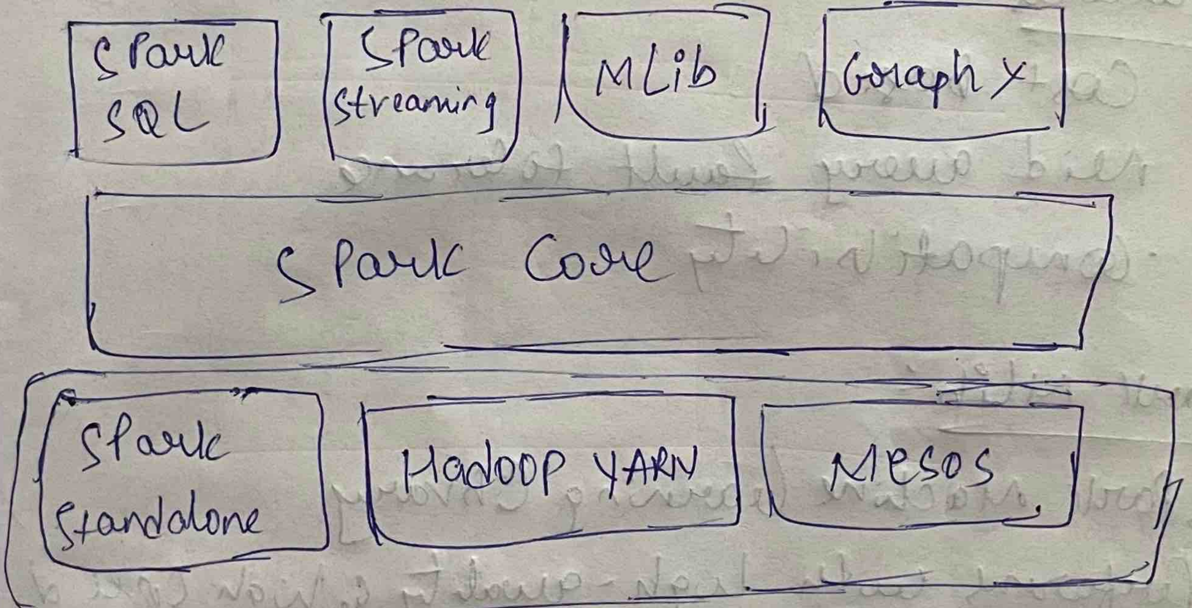
APache spark:

- cluster computing system
- supports execution graph
- In Python 'Scala' is used.
- it can run alone or by any existing cluster manager.

Spark Components

Spark core, Spark SQL, Spark streaming, Spark MLib, Spark Graph x. & Spark R.



Spark SQL

Spark streaming

MLib

Graph x

Spark Core

Spark Standalone

Hadoop YARN

Mesos

→ All functionalities are built on "spark core"

features:-

① essential Input/output functionalities

② fault recovery

③ Task dispatching

④ Observing "spark cluster"

Spark Streaming:-

→ It is an add on to core spark API

→ 'Scalable, fault tolerant stream processing

Spark SQL:-

→ distributed framework for "structured data"

→ This is can perform extra optimization

→ It is spark module for data processin

Features:

1. Cost based
2. Mid query fault tolerance
3. Compatibility

Spark MLib:-

→ Spark Machine learning library

→ Performs with high-quality & high speed

→ It can perform various Implementation of machine learning platforms.

Core Concepts

① **Job**:- It is piece of code which reads input from HDFS.

② **Stages**:- It is divided into stages

③ **Tasks**: every task has some work.

④ **DAG**: "Directed Acyclic Graph".

⑤ **Executor**: Executing task

Spark Components

① Spark Driver

② Spark Context

③ DAG Scheduler

④ Task Scheduler