03/02/24

PySpark

1) Apache Spark (2009) is an open source distributed processing system. used for big data workloads.

History -) 2009. project VC Berkley

2) Spark features:-

-) written in Scala Programming language & runs in JVM.

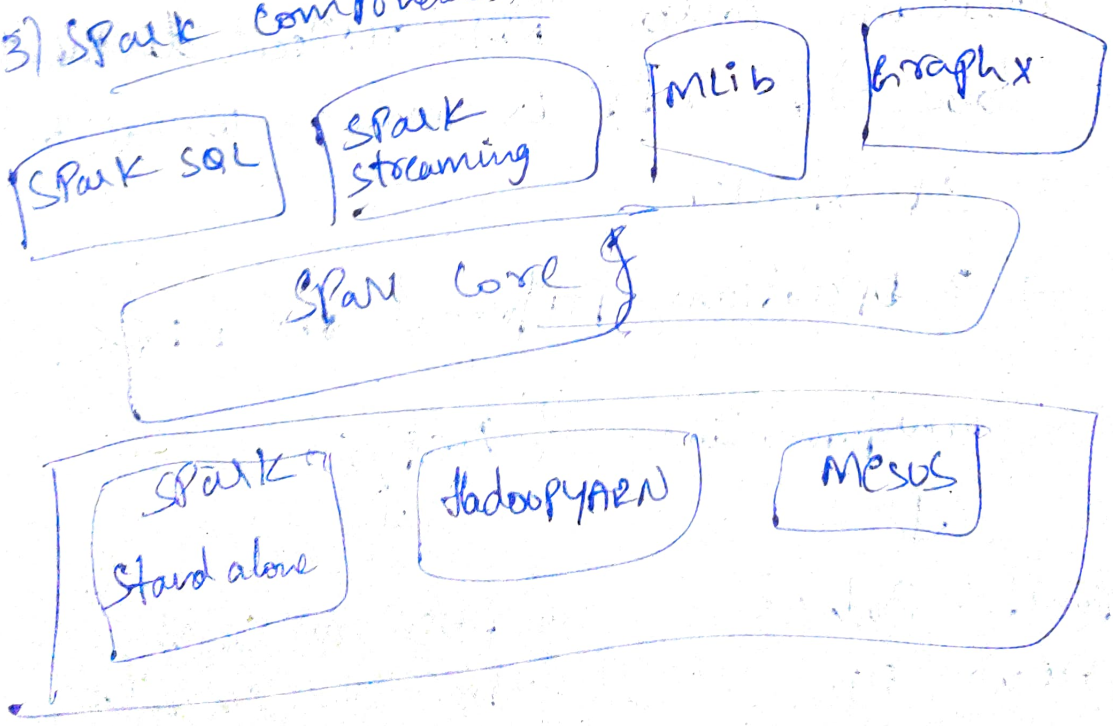
-) API: Scala, Java, Python, R.

-) Interactive shell: Scala & Python

-) data Sources:- SQL, NoSQL, S3, HDFS, Local File system, Etc.

-) Good Fit for iterative tasks like ML algorithm

3) Spark components:-



~~features~~!

i) Spark Core:

-) functionalities are provided by Apache spark are built on top of spark core.

-) delivers speed by providing in-memory computation.

features

-) task dispatching

-) fault recovery

-) overcomes snag of mapreduce by using in-memory computations

* Spark core is Embedded with a special collection called RDD (resilient distributed dataset). RDD is among the abstraction of spark.

-) 2 operations performed on RDDs.

Transformation & Action

Transformation ↓

function produces new RDD from existing RDD's

Action ↓

RDD's are created from each other but when we want to work w actual data set, we use actions

ii) Spark SQL:

-) It is ~~a component~~ distributed framework for structured data processing.

-) using this spark get more info about structure of data & computations.

-) with that info, it can perform extra optimization.

* features:-

-) cost based optimizer

-) Mid query fault - tolerance

-) Data frames.

iii) Spark Streaming:

-) It allows scalable, high-throughput, fault-tolerant stream processing of live data streams.

-) 3 phase of spark streaming.

a) Gathering

↓

2 categories

basic source:- sources available for streaming context

Ex: socket connections

Advanced source:-

Kafka, etc.

b) Processing

Gathered data is processed using complex algorithms

c) Data storage

Processed data is pushed out to file systems, databases & dashboards.

-) known as discretized stream DSt

iv) Spark MLlib

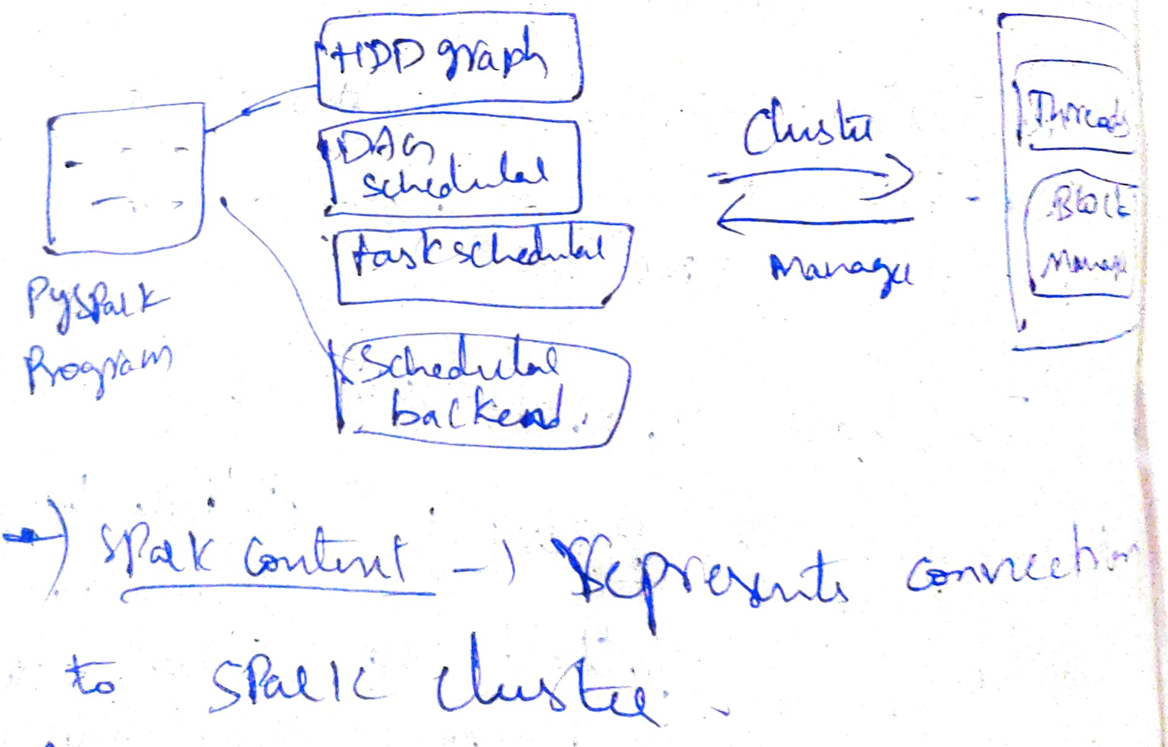
-) MLlib is scalable Machine leaning library that discusses both high-quality algorithm and high speed.

-) motive is make ~~[illegible]~~ ML scalable & Easy.

-) Consists of clustering, regression, classification & collaborative filtering.

Spark component:

DAG schedular
task schedular
schedule Backend.



-) Spark Content -) Represents connection to Spark cluster.

-) DAG Schedular -) Compute a DAG of stages for Each Job.

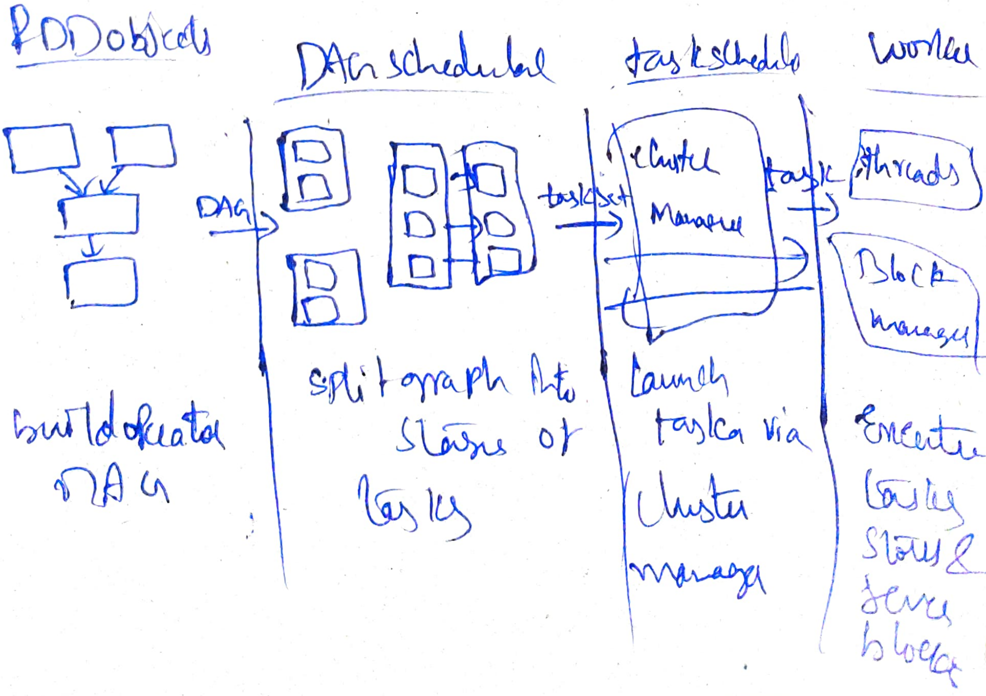
-) Task schedular :- responsible for sending the tasks to cluster, running them, retrying failures & to run the Jobs.

-) Schedule Backend - for scheduling system that allows pluggings in diff implementations (Mesos, YARN, Stand alone, local)

3) **How Spark works!** - Spark has a small code base and system is divided in various layers

8 Layers - Interpreter
Spark create a Operator Graph when it runs an action Graph subbmitted to DAG schedull the DAG divide graph into stages & also optimised the Graph.

-) stages are Passed on task schedular.



RDD objects
DAG schedular
task schedule
worker
DAG
taskset
Cluster Manager
task
threads
Block manager
build operator DAG
split graph into stages of tasks
launch tasks via cluster manager
Execute tasks stores & serves blocks