Pyspark

A big data framework having in memory distributed processing engine.

Apache biggest project

Features

🡪Cost Effective

🡪Fault tolerance

🡪Lazy Evalution

🡪In memory computation

🡪Open source

🡪Supports multiple languages ie Python, java, Scala , r

🡪Scalibility

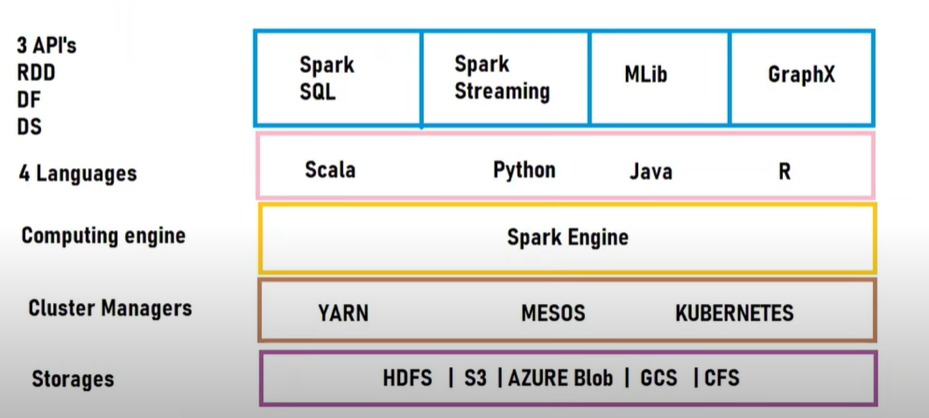
🡪Reusability

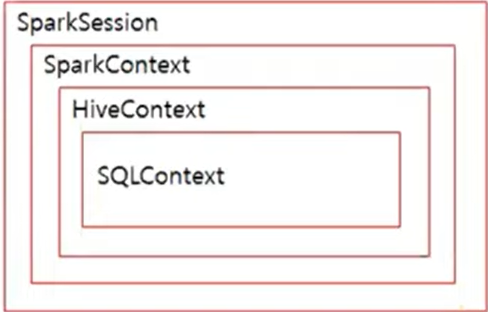
🡪100 times faster than Hadoop

🡪10 times faster in disc

🡪Spark can integrate with Hadoop system and etl tools like talend,informatica,CFS(Cloudera File System)

Spark Components





Spark Session is the combination of spark context, hive context and sql context.

Spark Context was earlier used as an entry point to connect to different clusters . It was used only for RDD creation.

But now apache spark 2.0 has a Session called sparkSession having all 3 different context within one.

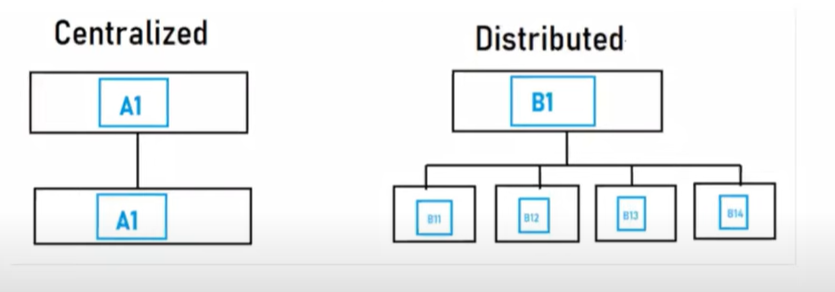
Creating a spark session

from pyspark.sql import SparkSession  
spark= SparkSession.builder.master("local[\*]").getOrCreate()  
spark.conf.set("spark.sql.repl.eagerEval.enabled",True)

Need of distributing os

Centralized

Distributed



IN centralized the mode of flow of data is only from one server to a single server. This is good unless the data is less. When data size increases then we need to have vertical scalability which is by adding external hardware to increase the space for storage or having optimized software.

That’s why distributed system came into the picture as its allows multiple clusters to store data and process it distributed Ly. Thus Spark is the biggest open source for Distributed Storage System.