Mallu Sravana Sandhya **Assessment-21**
28/12/23

Databricks

Databricks is a unified, open analytics platform for building, deploying, sharing, and maintaining enterprise-grade data, analytics, and AI solutions at scale. The Databricks Data Intelligence Platform integrates with cloud storage and security in your cloud account, and manages and deploys cloud infrastructure on your behalf.

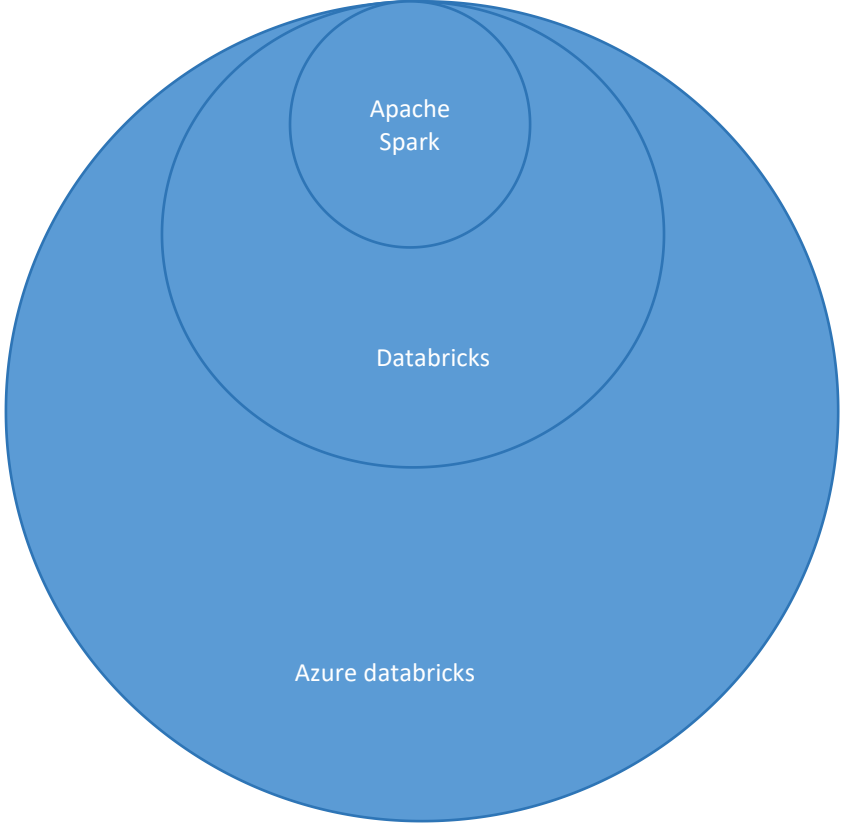
Apache Spark:

It is a unified anaytics engine for processing large amount data.It provides high level APIs in java,python , scala and R programming languages.It is also a optimised engine which supports graphs and machine learning. It works 100 times faster than any other traditional engine

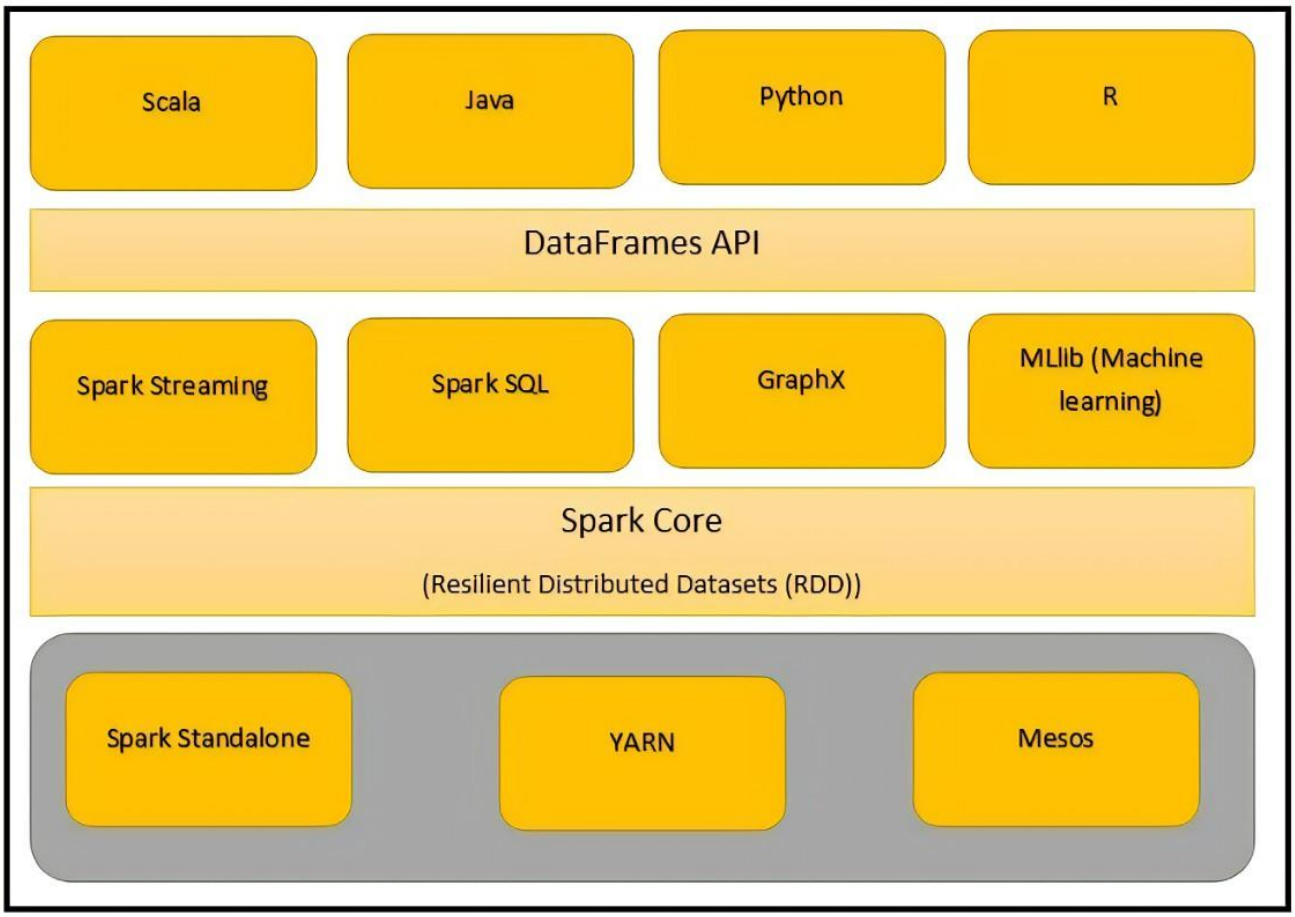
- It contains cluster managers and perform the actions and transformations using master-slave technique where drivers acts as master and executers as slaves
- It is 100% open source under required licience
- Simple and Easy to use APIs
- In-memory processing engine
- Distributed computing platform for different APIs
- Unified engine which supports ML and GraphX

Azure databricks :

It is online platform where we can run apache spark
It provides a space to execute the databricks which is integrated by apache spark



Apache spark Architecture



Here Standalone,yarn,Mesos are the cluster managers on which any spark is functioning

Resilent Distributed Datasets (RDD) are the files which are present in spark core which performs two functions actions and transformation that produces new items to data set, returns RDDs based on requirements respectively
Apache Sparks contains many libraries on SQL,Steam processing,GraphX and Machine learning libraries
It supports different APIs like python, java, scala and R programming languages among them it works efficiently on scala

While dealing with Azure databricks we have some knowledge on basic python programming language and sql language and we must have azure account