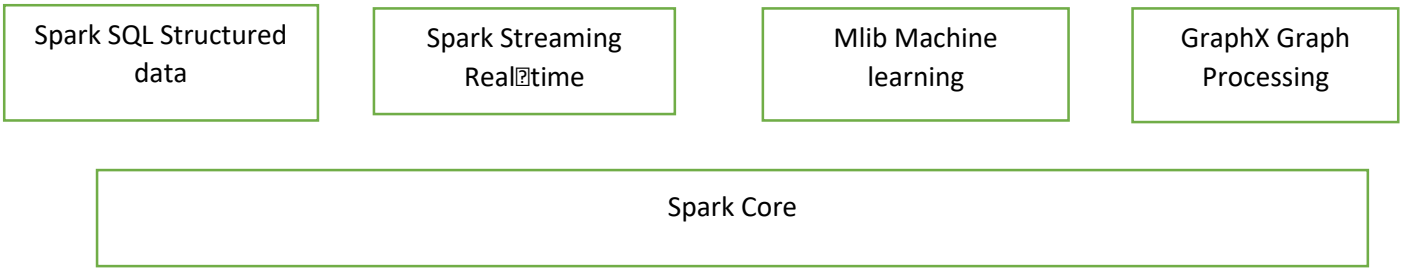
ASSIGNMENT 7

NAME: Aishee Bhattacharya

BATCH: DXC-262-Analytics-B12-Azure

DATE: 07/06/2022

1. Explain what are various components of SPARK with block diagram? explain functionality of every components?



- Spark Core: Spark core is the base engine for large-scale parallel and distributed data processing.
- Spark SQL:Spark SQL framework component is used for structured and semi-structured data processing
- Spark Streaming: Spark streaming is a light weight API that allows developers to perform batch processing and real time streaming of data with ease.
- MLib: M-lib is a low machine library that is simple to use is scalable and compatible with various programming languages.
- GraphX: GraphX is a spark's own graph computation engine and data store.

2. Explain Spark core in details & how RDD is related to Spark core - explain with Spark program ?

Spark core is the base engine for large-scale parallel and distributed data processing.

It is responsible for

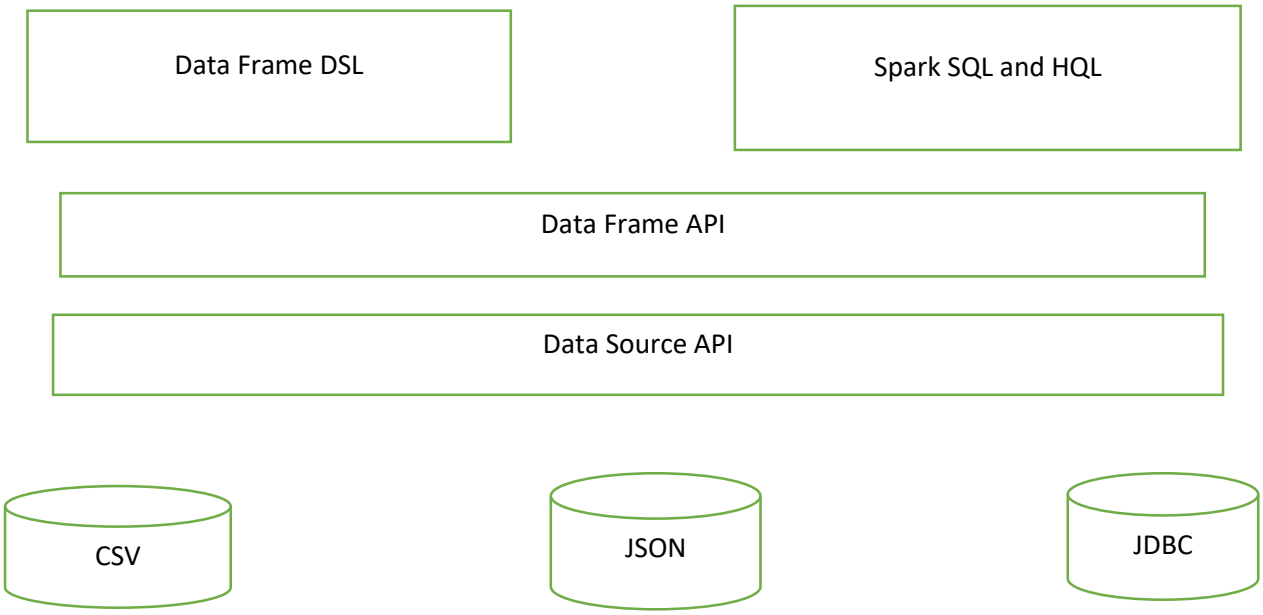
- Memory management
- Fault recovery
- Scheduling, distributing and monitoring jobs on a cluster
- Interacting with storage systems

Spark core is embedded with RDDs an immutable fault tolerant distributed collection of objects that can be operated on in parallel.

3. Explain various Mlib algorithms Spark is supporting ?
   Mlib contains machine learning libraries that have an implementation of various machine learning algorithms.
   - Clustering
   - Classification
   - Collaborative filtering

4. Explain benifits Spark SQL & how relational data will be inserted into SPARK ?

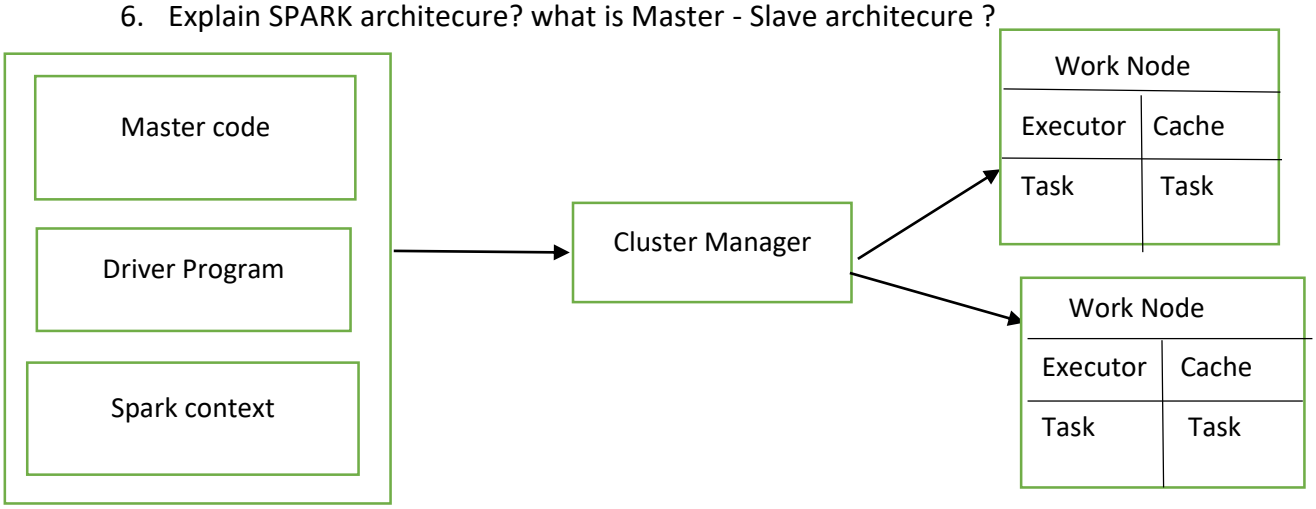
   Spark SQL framework component is used for structured and semi-structured data processing.



5. Explain Spark streaming in detail ?
   Spark streaming is a light weight API that allows developers to perform batch processing and real time streaming of data with ease. It provides secure, reliable and fast processing of live data streams.

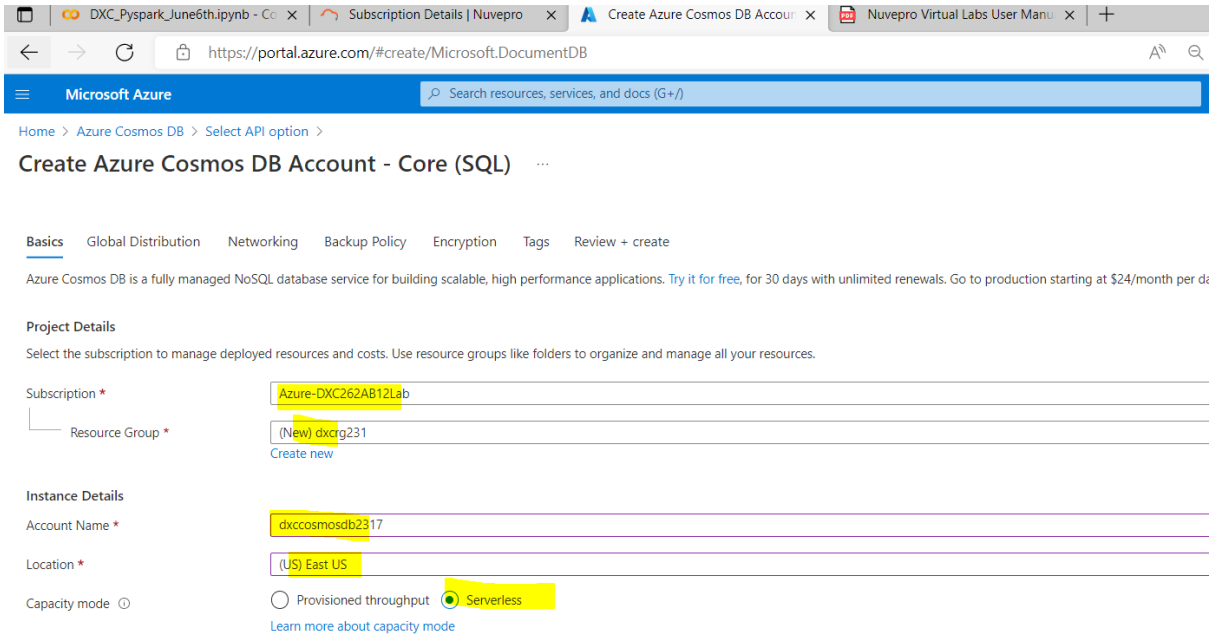
Input data stream ⟶ Batches of input data ⟶ Batches of processed data

6. Explain SPARK architecure? what is Master - Slave architecure ?

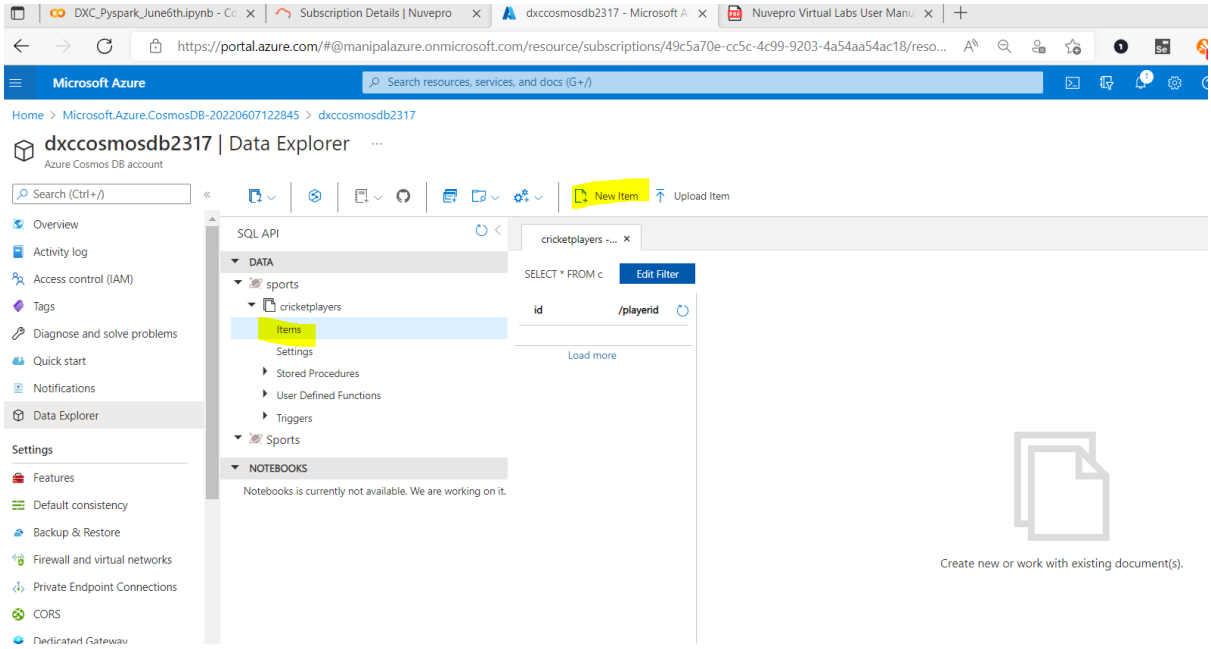


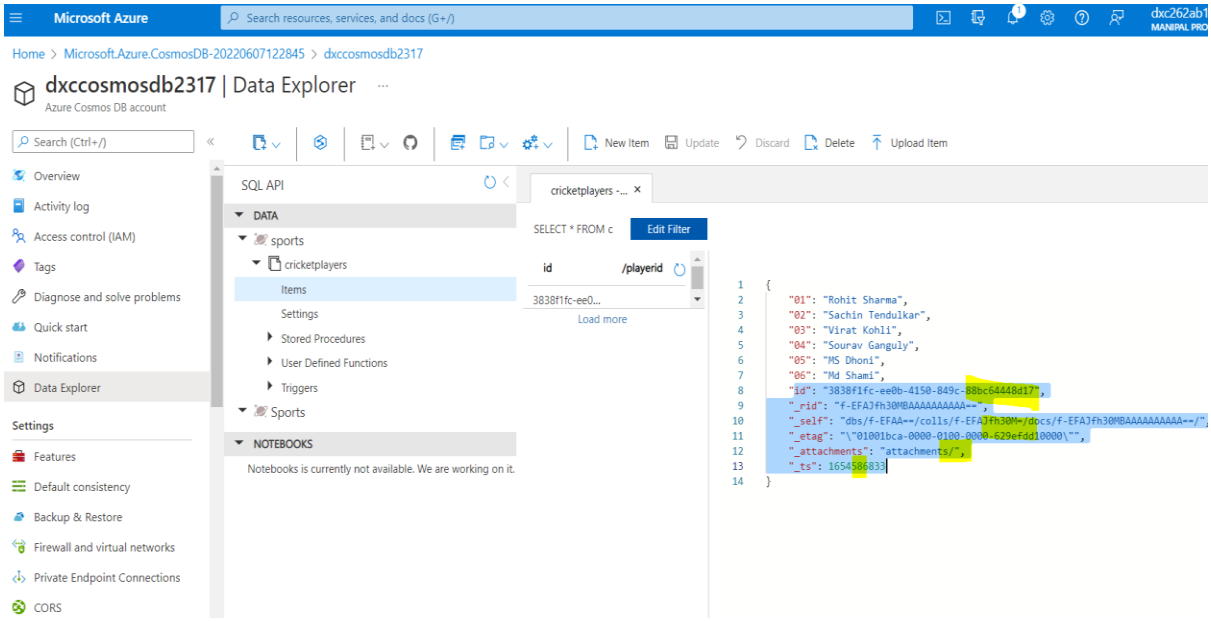
7. Explain various cluster managers in SPARK?
   - Spark Standalone mode- By default applications submitted to the standalone mode cluster will run in FIFO order, and each application will try to use all available nodes.
   - Apache Mesos- It is an open source project to manage computer clusters and also can run Hadoop applications.
   - Hadoop YARN- It is the cluster resource manager of Hadoop 2 Spark can be run on YARN.
   - Kubernetes-Kubernetes is an open-source system for automating deployment,scaling and management of containerized applications.

8. Explain with sceenshots & steps how to create Cosmos DB ?



9. Explain with sceenshots & step how to insert data into Cosmos DB?





10.Explain with sceenshots & step how to create Azure SQL Db & also explain how to insert data into Azure SQL D?