**07-06-2022**

**BY**

**Sivarathri Brahmaiah**

1. There are 5 components of SPARK

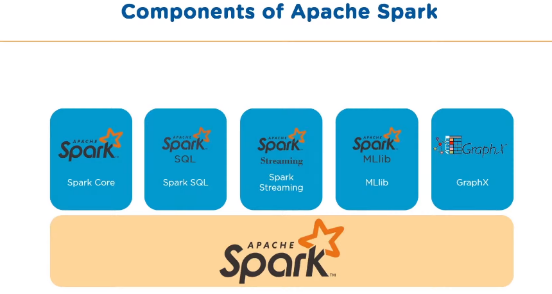
Spark Core

Spark SQL

Spark Streaming

MLlib

Graphx



**Spark core:** It is the base engine for large scale parallel and distributed data processing. It is responsible for

Memory management

Fault recovery

Scheduling

Interacting with storage systems

**Spark SQL**: It is framework component it used for structural and semi structural data processing

**Spark Streaming:** It is light weight API that allows developers to perform batch processing and real time streaming of data.

**Spark MLlib:** It allows a low level ML library that is simple to use. It is scalable and compatible. EX Clustering.

**Spark Graphx:** It is graph computation engine and data store

**2)** **Spark core:** It is the base engine for large scale parallel and distributed data processing. It is responsible for

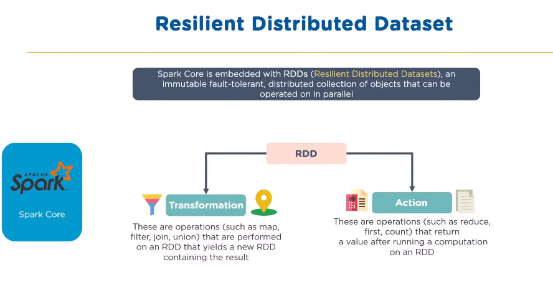
Memory management

Fault recovery

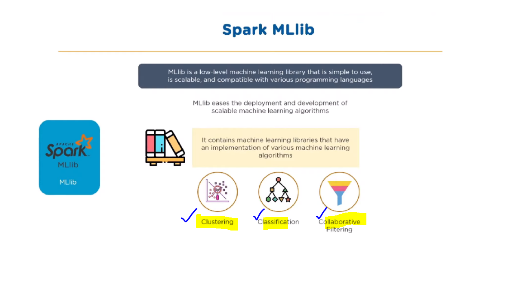
Scheduling

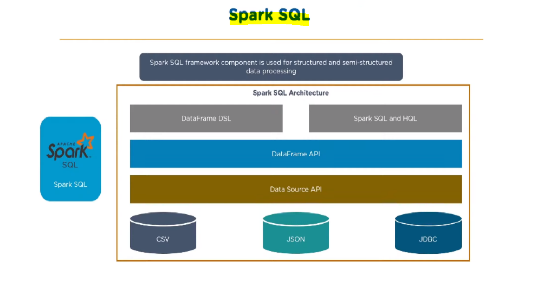
Interacting with storage systems

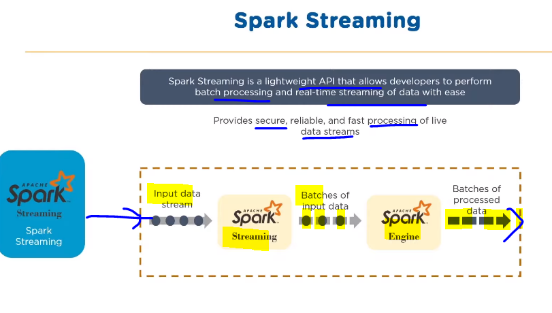
**RDD:** Spark core embedded with RDD an immutable fault tolerant distributed collection of objects that can be operated on in parallel

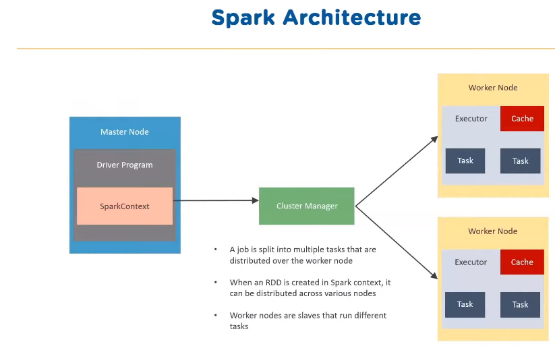


3) MLlib: It allows a low level ML library that is simple to use. It is scalable and compatible. It contains ML libraries. EX Clustering.



**4) Spark sql:** Spark sql framework component used for structural and semi structural data processing.

5)**Spark streaming**: It is light weight API that allows developers to perform batch processing and real time streaming of data with easilly.

6)**Spark Architecture:**

**Master-slave:** One is responsible for giving tasks and another one is responsible for executing tasks.

7) **Various cluster managers in Spark:**

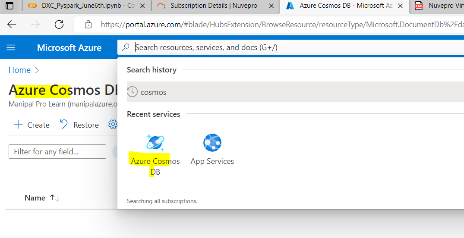
*Standalone mode*: It will follows FIFO(First In First Out) principle and each application will try

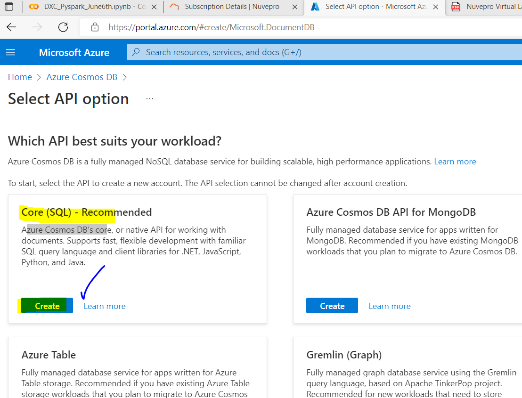
to use all available nodes.

*MESOS:* It is an open source project to manage computer clusters and can also run Hadoop applications.

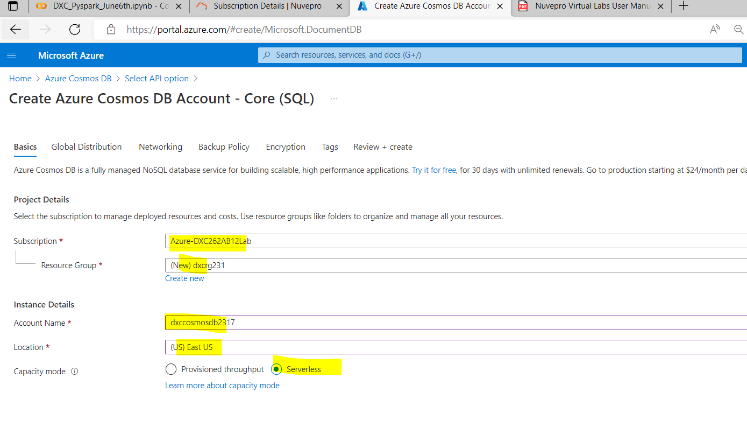
*Apache YARN:* It is cluster resource manager of Hadoop.

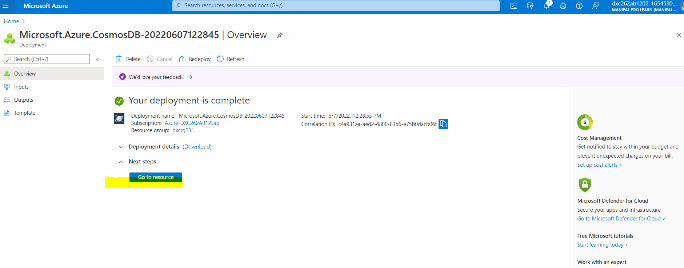
*Kubernetes:* It is an open source system for automating deployment ,scaling and management.

8) At first login to Microsoft Teams. Now search azurecosmosdb. Then display will be shown in below

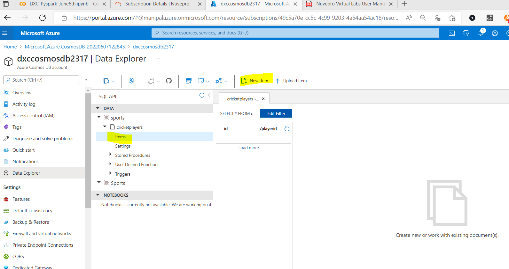
And click on the cosmosdb then following page will be come

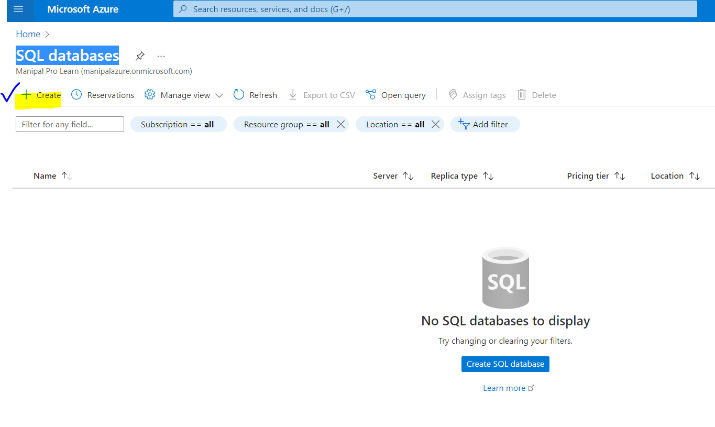
Click on core sql create and some basic information it will be asked which is shown in below image

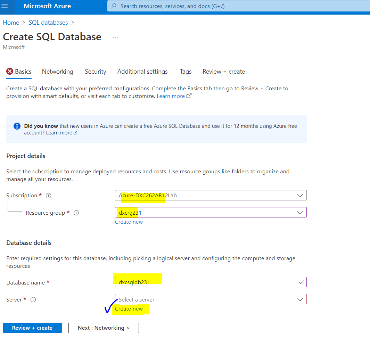


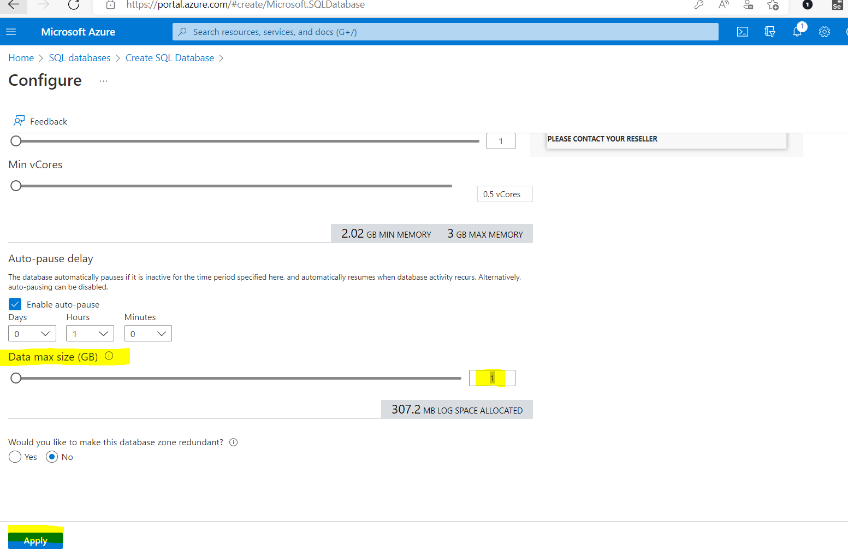
After successful creation of interface details almost deployment is done. After click on the create we can insert information into data base.

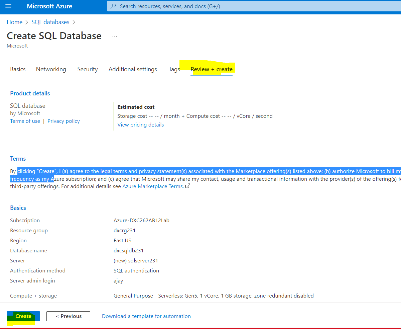
9) If further we want insert the data into data base in cosmosdb goto data> items>

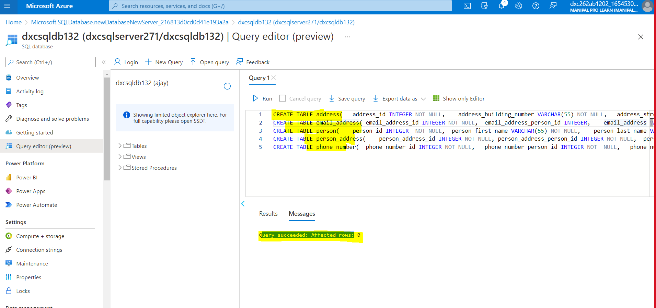
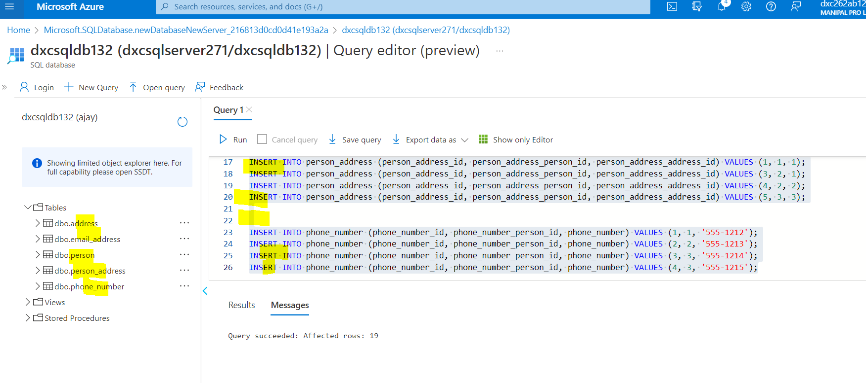
In the items we can add new items into already existing data base which is showing below image.

10) Goto sql data base and create

Then give the database details

Apply the values which is shown in below

Now click on the review+create and create

Deployment is created and login with the credentials which you have given before the crate table and finally execute by inserting values which is shown in below images