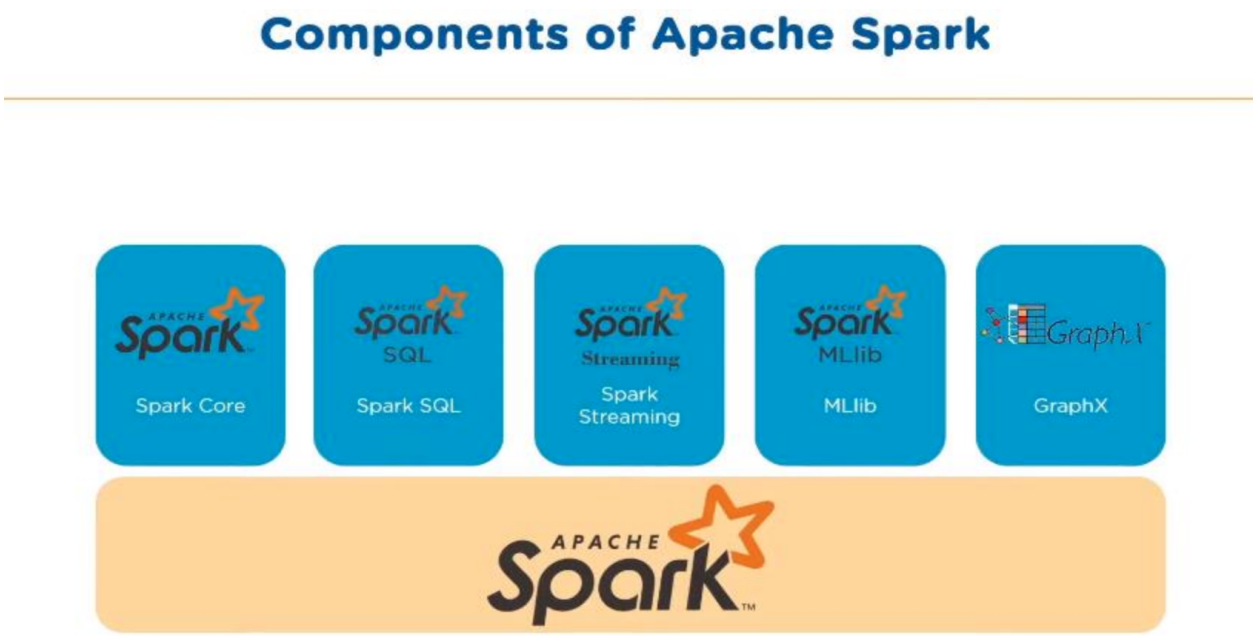
# ASSIGNMENT 7

NAME – JYOTHI CHANDANA VOLETI
ROLL NUMBER – DXC-262-AB-1218
BATCH – DXC-262-ANALYTICS-B12-AZURE
COMPANY – DXC TECHNOLOGY
EMPLOYEE DOMAIN –AZURE ANALYTICS
TRAINING UNDER – MANIPAL PRO LEARN
TRAINER NAME – MR. AJAY KUMAR
DATE OF SUBMISSION – 7TH JUNE 2022
NO.OF QUESTIONS: 10

**1. Explain what are various components of SPARK with block diagrams? Explain the functionality of every component?**
**Ans: Components of Spark:**
**Block Diagram:**




**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 lightweight API that allows developers to perform batch processing and real-time streaming of data with ease.

**Spark MLlib:** MLlib is a low-level machine learning library that is simple to use, is scalable and compatible with various programming languages.

**GraphX:** GraphX is 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 ?**

**Ans: Spark Core:**

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

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

Spark Core is embedded with RDDs( Resilient Distributed Dataset), an immutable fault-tolerant, distributed collection of objects that can be operated on in parallel.
There are two ways to create RDDs − parallelizing an existing collection in the driver program, and referencing a dataset in an external storage system.

- *Resilient:* Fault tolerant and is capable of rebuilding data on failure
- *Distributed:* Distributed data among the multiple nodes in a cluster
- *Dataset:* Collection of partitioned data with values

**3. Explain various Mllib algorithms Spark is supporting ?**

**Ans: Spark MLlib:** MLlib is a low-level machine learning library that is simple to use, is scalable and compatible with various programming languages.

**Algorithms MLlib Spark is supporting are:**

- Clustering
- Classification
- Collaborative Filtering
- Regression

MLlib supports various methods for binary classification, multiclass classification, and regression analysis.

1. Binary Classification supports linear SVMs, logistic regression, decision trees, random forests, gradient-boosted trees, naive Bayes methods.
2. Multiclass Classification supports logistic regression, decision trees, random forests, naive Bayes methods.
3. Regression supports linear least squares, Lasso, ridge regression, decision trees, random forests, gradient-boosted trees, and isotonic regression methods.
4.

Collaborative Filtering is commonly used for recommender systems. These techniques aim to fill in the missing entries of a user-item association matrix.

Clustering is an unsupervised learning problem whereby we aim to group subsets of entities with one another based on some notion of similarity. Clustering is often used for exploratory analysis and/or as a component of a hierarchical supervised learning pipeline (in which distinct classifiers or regression models are trained for each cluster).

MLlib supports the following models:

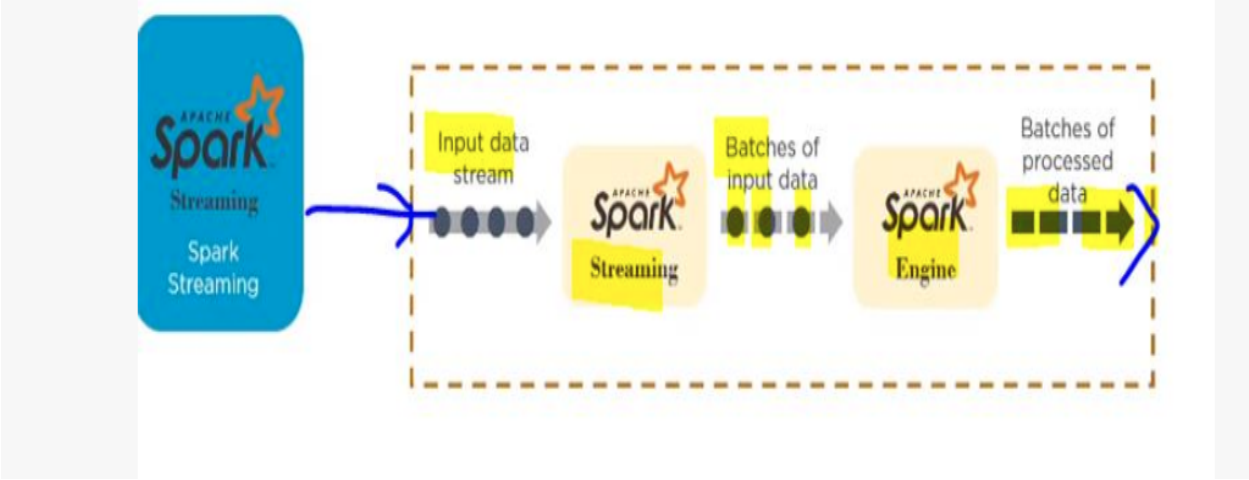
- K-Means
- Gaussian mixture

**4. Explain benefits of Spark SQL & how relational data will be inserted into SPARK ?**
**Ans: Benefits of Spark SQL:**

- Integrated
- Unified Data Access
- High Compatibility
- Scalability
- Standard Connectivity
- Performance Optimization

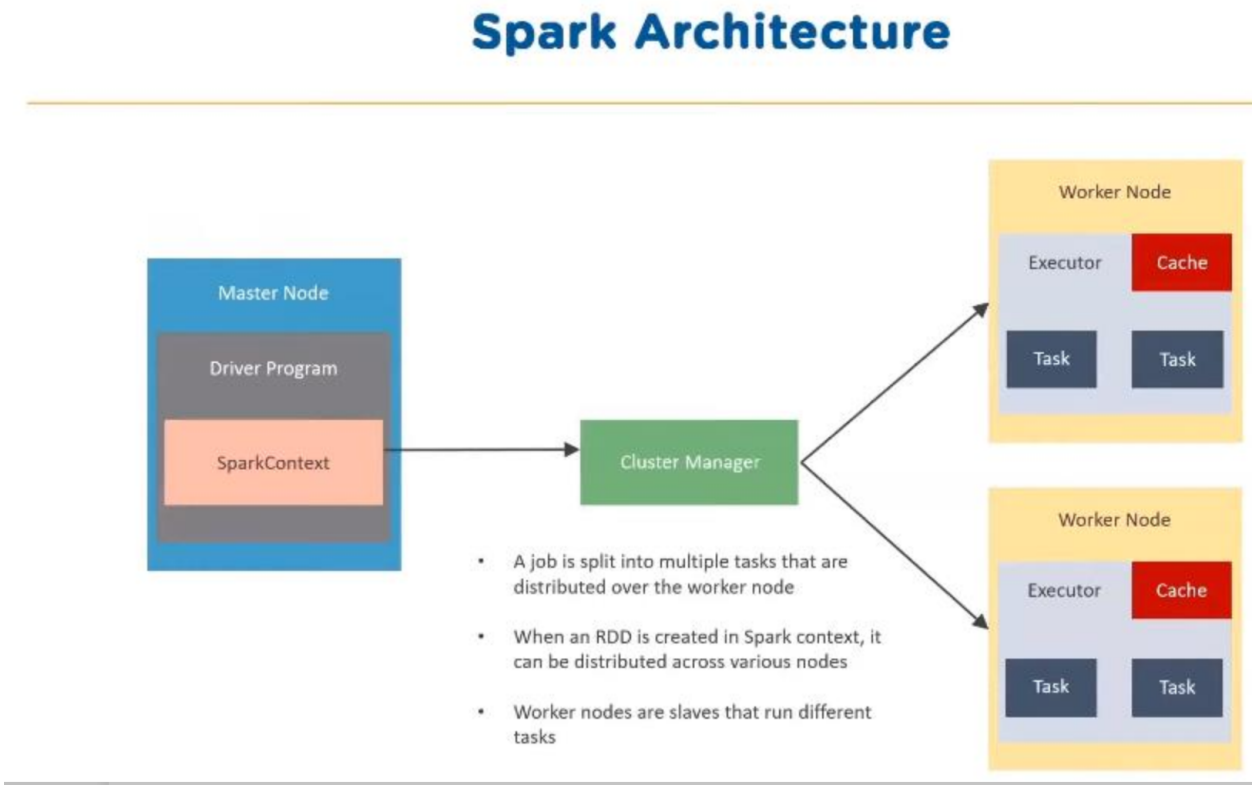
We can Use pandas and read the files using reading csv files by pandas.read_csv(). And then can import the data files.

**5. Explain Spark streaming in detail ?**
**Ans: Spark Streaming:** Spark Streaming is a lightweight 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 streaming.



**6. Explain SPARK architecture? what is Master - Slave architecure ?**
**Ans: SPARK Architecture:**


**Master-Slave Spark Architecture:**



In master node, the *driver program*, drives your application. W*orker nodes* are the slave nodes whose job is to basically execute the tasks. These tasks are then executed on the partitioned RDDs in the worker node and hence returns back the result to the Spark Context.

**7. Explain various cluster managers in SPARK?**
**Ans: Spark Cluster Managers:**
**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.

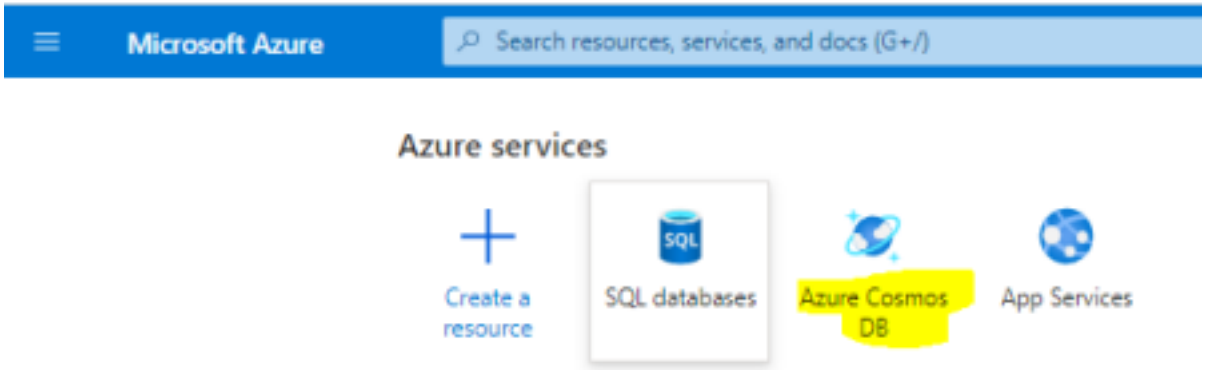
**Mesos:** Apache Mesos is an open-source project to manage computer clusters, and can also run Hadoop applications.

**Hadoop YARN:** Apache YARN is the cluster resource manager of Hadoop 2. Spark can be run on YARN.

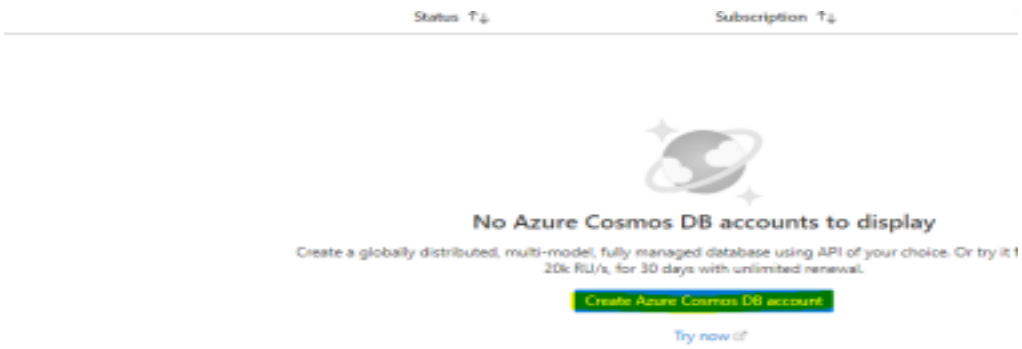
**Kubernetes:** Kubernetes is an open source system for automating deployment, scaling, and management of containerized applications.

**8. Explain with screenshots & steps how to create Cosmos DB ?**
**Ans:**

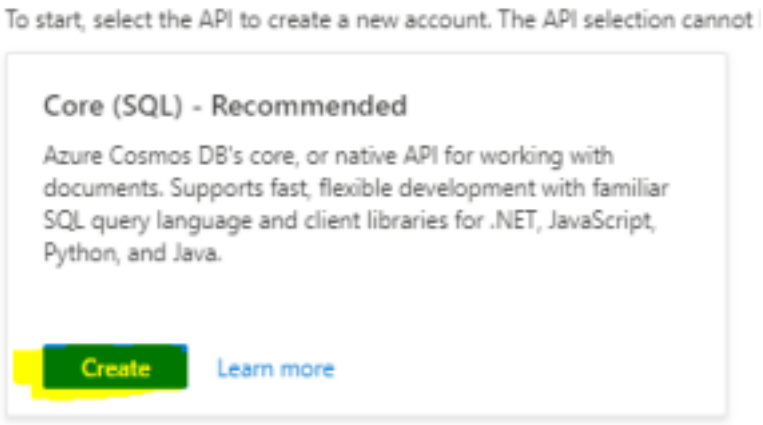
**Step 1:** Go to Azure,and select "Azure Cosmos DB".



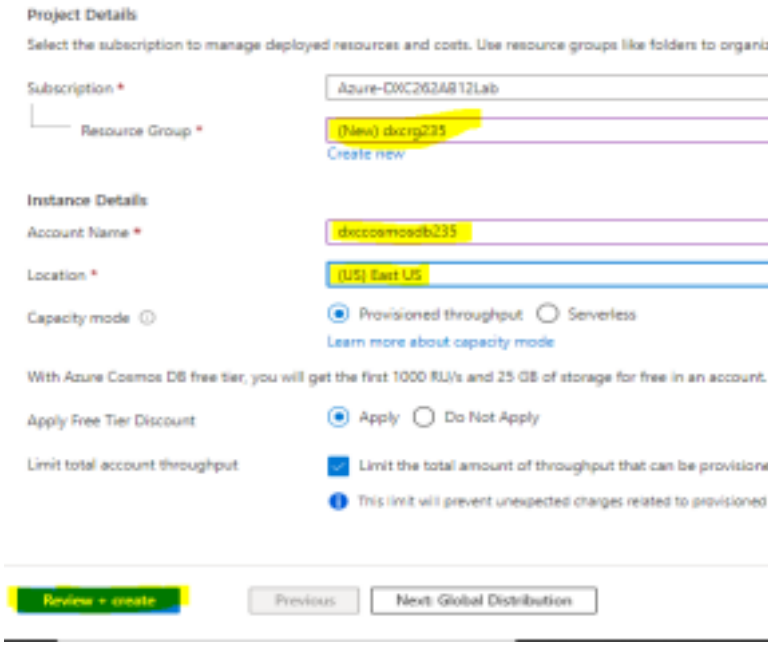

**Step 2:** Click on "Azure Cosmos DB".



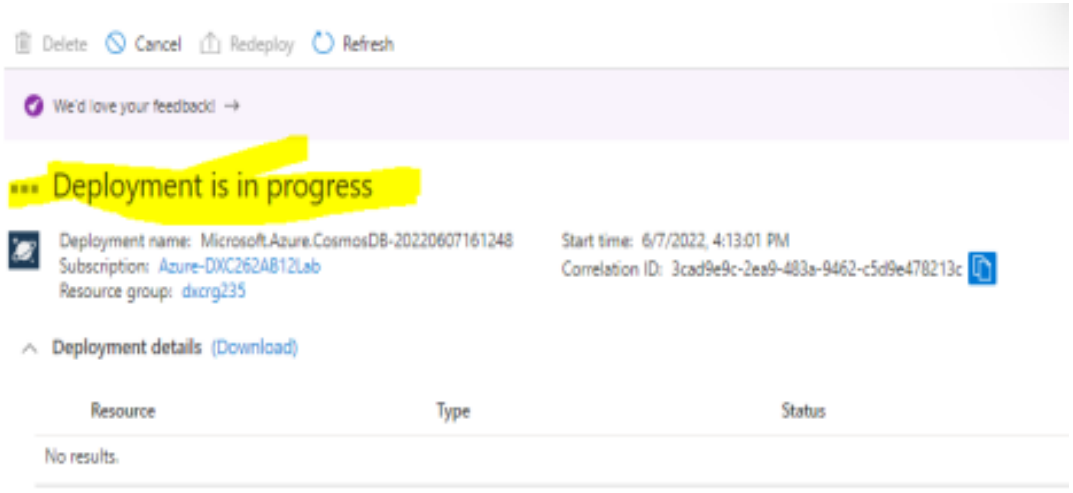
**Step 3:** Click on “Core(SQL)”  ”Create”.



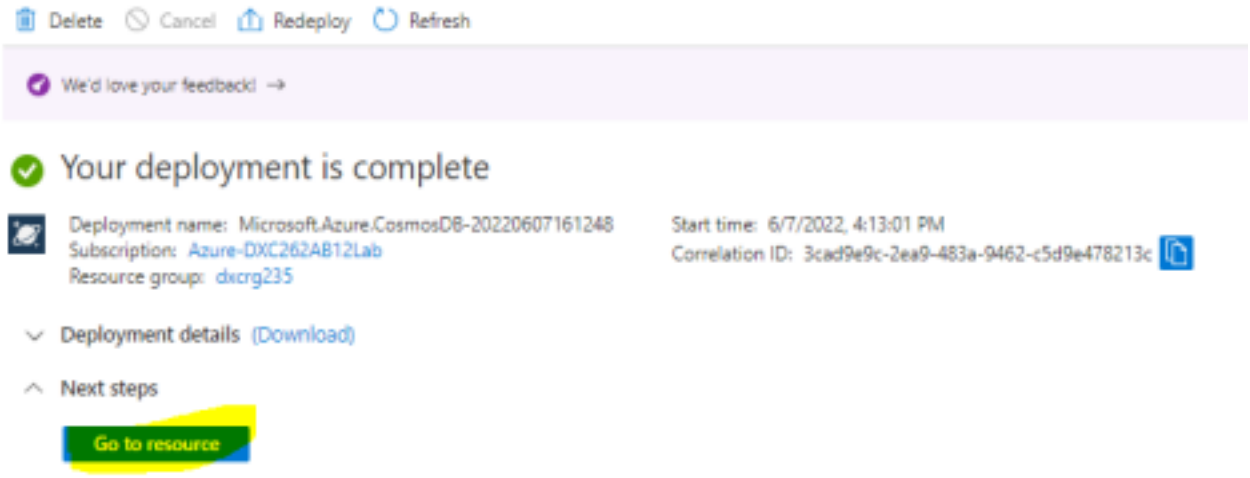

**Step 4:** Follow steps in the screenshot.



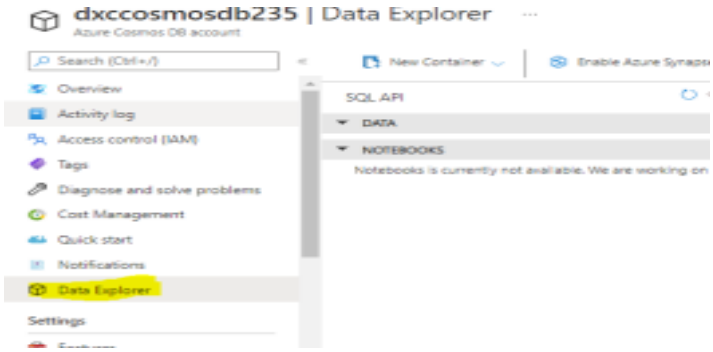
**Step 5:** Deployment is being done.




**Step 6:** Go to Resources.



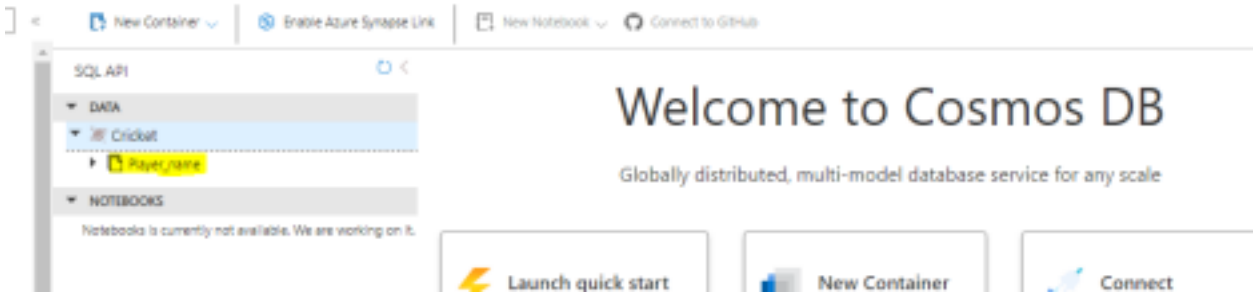

**Step 7:** Go to "Data Explorer".



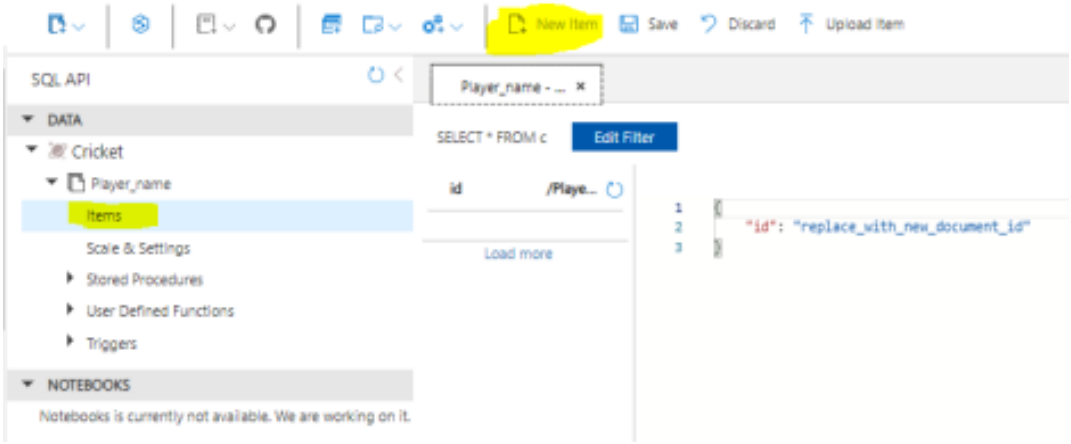
**Step 8:** Creating Database with details as shown.


**9. Explain with screenshots & step how to insert data into Cosmos DB?**
**Ans:Inserting data into Cosmos DB:**

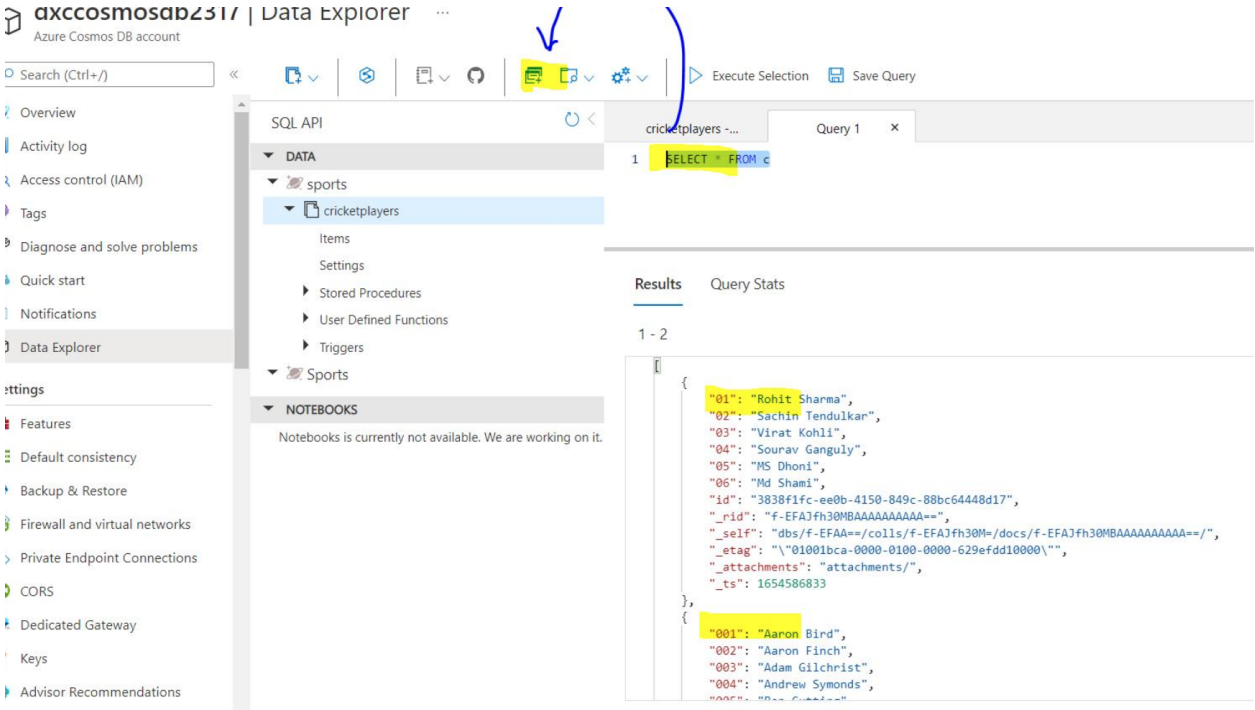
**Step 1:** Select container “Cricket” ”Player_Name”.



**Step 2:** Click on “items” and “New Item”.

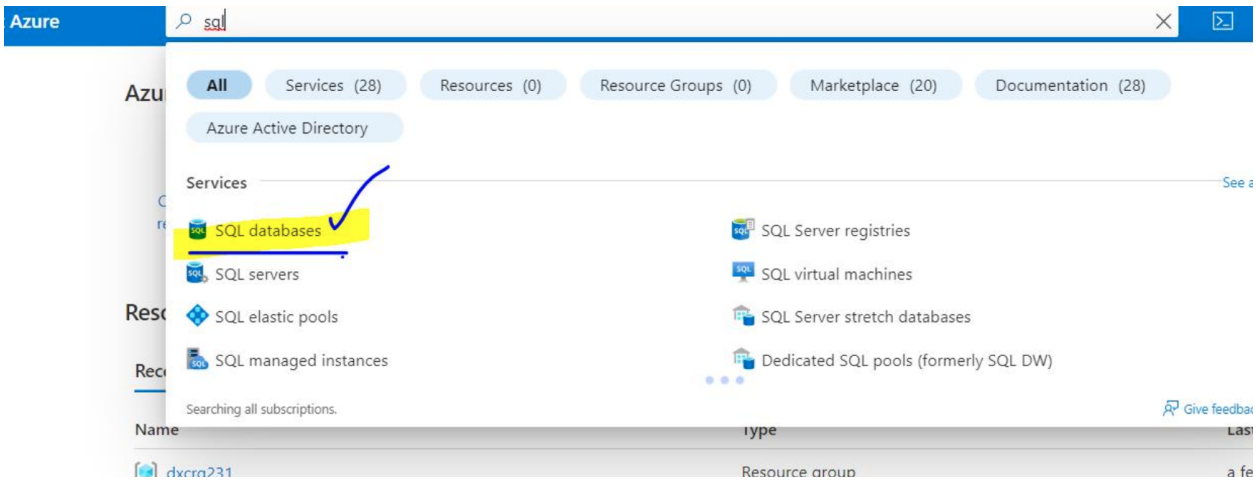



**Step 3:** Write the script shown below and then click on “Save”. And the data will be inserted.

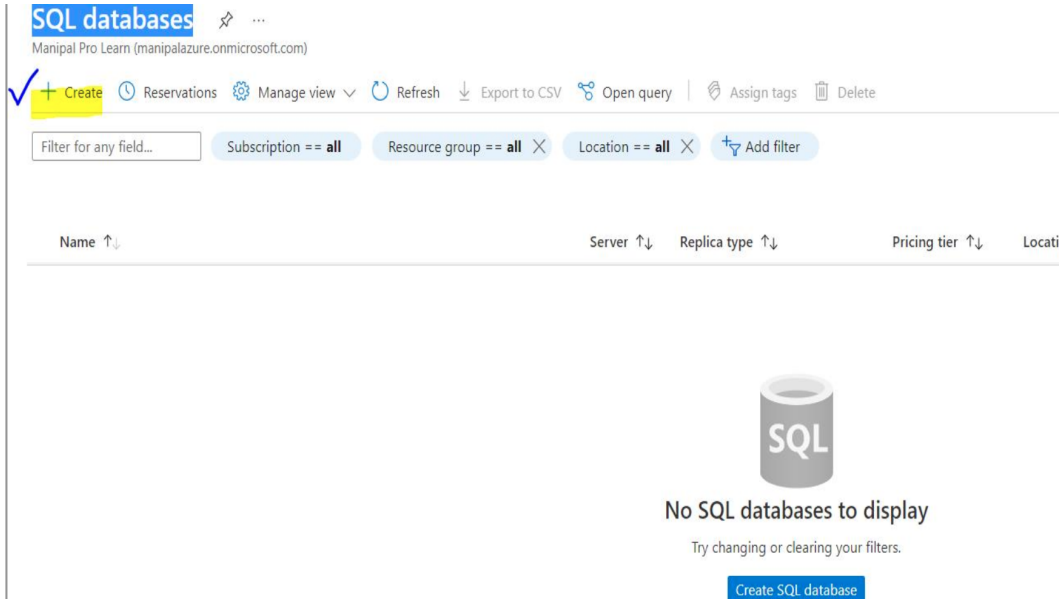


**10. Explain with screenshots & step how to create Azure SQL Db & also explain how to insert data into Azure SQL Db?**

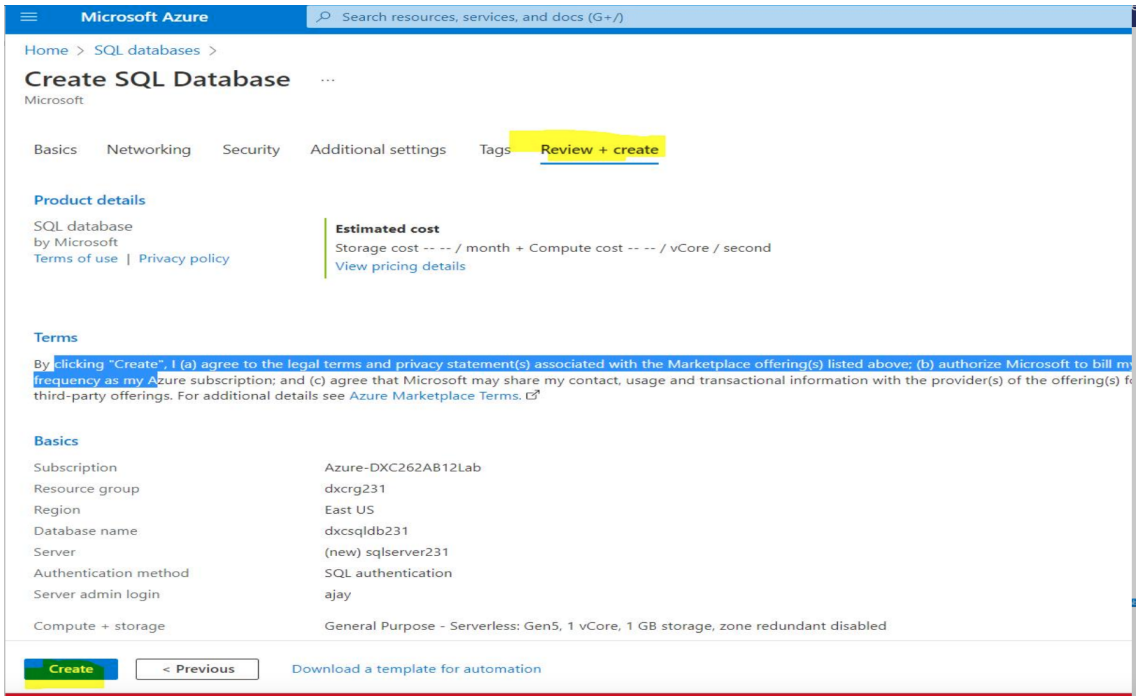
**Ans: Step1:** Click on SQL Databases.



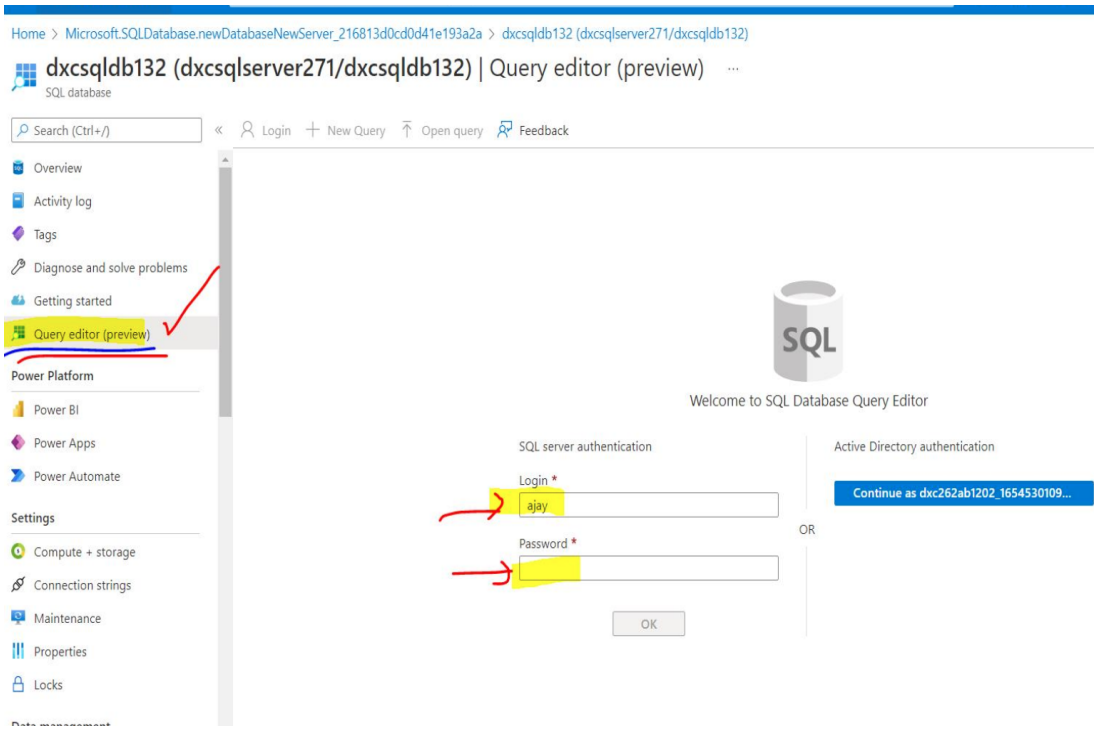

**Step2:** Click on Create.



**Step3:** Click on all the options and review and create after configuring it to 1GB and Click on Create Db.




**Step4:** Enter Credentials and login.



**Step5:** Click on ‘Quick Preview’ and Write a query and then click on “Run”. And the data is inserted.