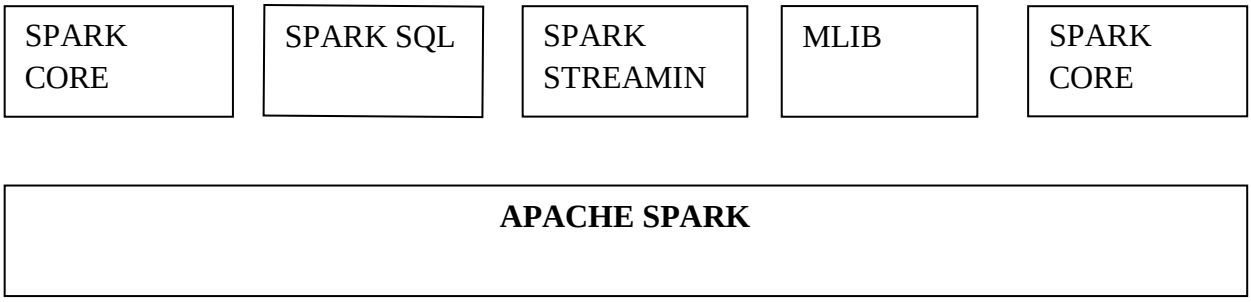
ASSIGNMENT 7

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

COMPONENTS OF SPARK:



FUNCTIONALITIES OF COMPONENTS:

SPARK CORE:

- ✓ Spark Core is the heart of the Apache Spark framework.
- ✓ Spark Core provides the execution engine for the Spark platform which is required and used by other components which are built on top of Spark Core as per the requirement.
- ✓ Spark Core provides the in-built memory computing and referencing datasets stored in external storage systems.
- ✓ It is Spark's core responsibility to perform all the basic I/O functions, scheduling, monitoring, etc. Also, fault recovery and effective memory management are Spark Core's other important functions.

SPARK SQL:

- ✓ Spark SQL is built on top of Shark which was the first interactive SQL on the Hadoop system.
- ✓ Shark was built on top of Hive codebase and achieved performance improvement by swapping out the physical execution engine part of the Hive.
- ✓ But due to the limitations of Hive, Shark was not able to achieve the performance it was supposed to.
- ✓ So the Shark project was stopped and Spark SQL was built with the knowledge of Shark on top of Spark Core Engine to leverage the power of Spark.
- ✓ You can read more about Shark in the following blog by Reynold Xin, one of the Spark SQL code maintainers.

SQL STREAMING:

- ✓ This Spark library is primarily maintained by Tathagat Das and helped by MatieZaharia.
- ✓ As the name suggests this library is for Streaming data.
- ✓ This is a very popular Spark library as it takes Spark's big data processing power and cranks up the speed.
- ✓ Spark Streaming has the ability to Stream gigabytes per second.
- ✓ This capability of big and fast data has a lot of potentials. Spark Streaming is used for analyzing a continuous stream of data.
- ✓ A common example is processing log data from a website or server.

MLIB:

- ✓ MLlib is a low-level machine learning library.
- ✓ It can be called from Java, Scala and Python programming languages.
- ✓ It is simple to use, scalable and can be easily integrated with other tools and frameworks.
- ✓ MLlib eases the deployment and development of scalable machine learning pipelines.
- ✓ Machine learning in itself is a subject and it may not be possible to get into details here.
- ✓ But these are some of the important features and capabilities Spark MLLib offers:

    - Linear regression, logistic regression
    - Support Vector Machines
    - Naive Bayes classifier
    - K-Means clustering
    - Decision trees

GRAPH X:

- ✓ GraphX is useful in giving overall information about the graph network like it can tell how many triangles appear in the graph and apply the PageRank algorithm to it.
- ✓ It can measure things like "connectedness", degree distribution, average path length and other high-level measures of a graph.
- ✓ It can also join graphs together and transform graphs quickly. It also supports the Pregel API for traversing a graph. Spark GraphX provides Resilient Distributed Graph (RDG- an abstraction of Spark RDD's).
- ✓ RDG's API is used by data scientists to perform several graph operations through various computational primitives.
- ✓ Similar to RDDs basic operations like map, filter, property graphs also consist of basic operators.

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

SPARK CORE:

- ✓ Spark Core is the underlying general execution engine for the Spark platform that all other functionality is built on top of.
- ✓ It provides in-memory computing capabilities to deliver speed, a generalized execution model to support a wide variety of applications, and Java, Scala, and Python APIs for ease of development.

RDD- RESILLENT DISTRIBUTED DATASET:

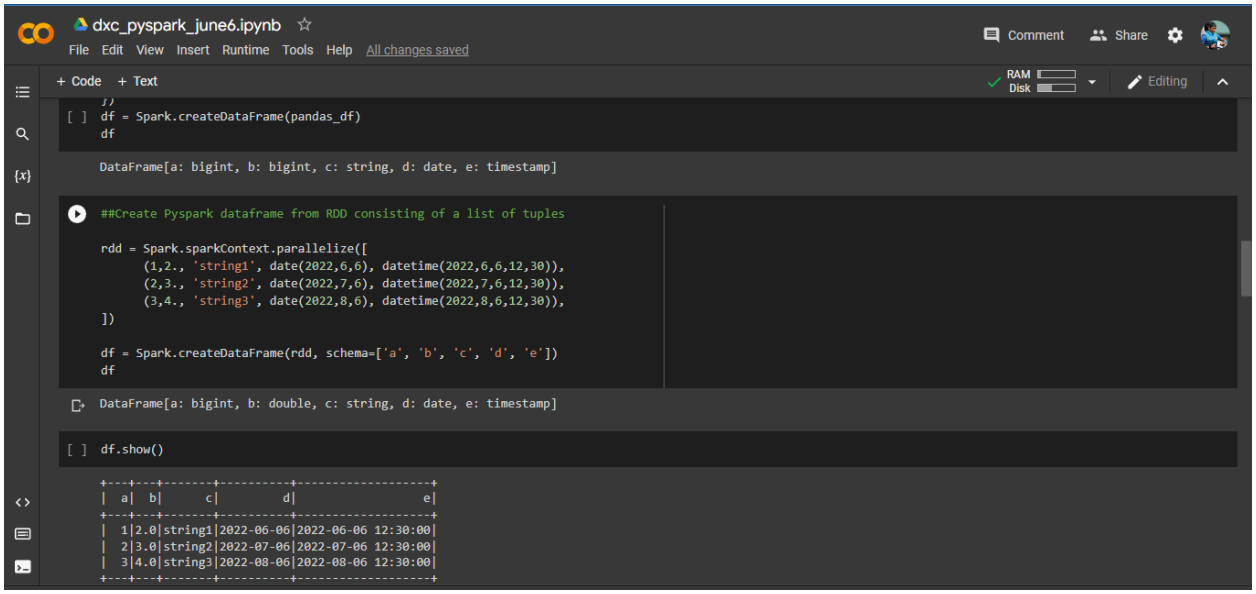
Spark core is embedded with RDDS an immutable fault- tolerant, distributed collection of objects that can be operated on in parallel.

TRANSFORMATION:

These are operations (such as map, filter, join, union) that are performed on a RDD that yields a new RDD containing the result.

ACTION:

These are operations (such as reduce first count) that return a value after running a computation on an RDD.



3. Explain various Mlib algorithms Spark is supporting ?

SPARK LIB

It contains machine learning libraries that have an implementation of various machine learning algorithm.

CLUSTERING

CLASSIFICATION

COLLABORATIVE FILTERING

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

BENIFITS SPARK SQL:

- ✓ Integrated
- ✓ Unified data Access
- ✓ High compatibility
- ✓ Standard Connectivity
- ✓ Scalability
- ✓ Performance optimization
- ✓ Batch processing of hive tables

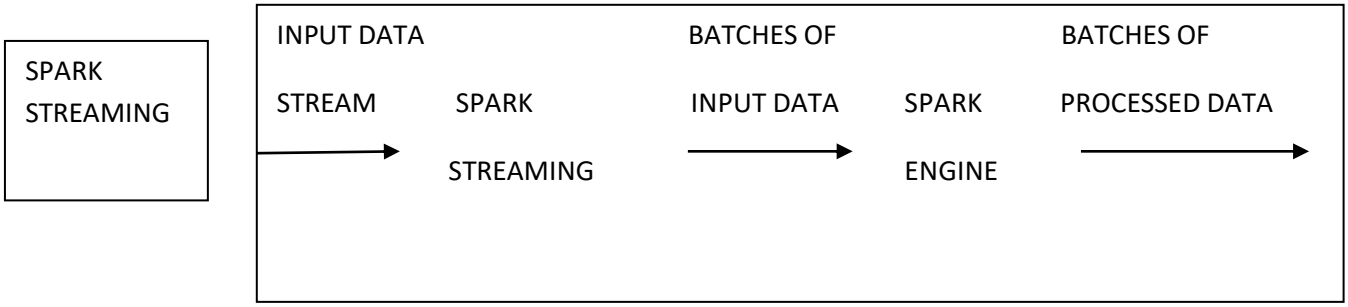
```
[ ] df.show()

+---+---+-------+----------+-------------------+
|  a|  b|      c|         d|                  e|
+---+---+-------+----------+-------------------+
|  1|2.0|string1|2022-06-06|2022-06-06 12:30:00|
|  2|3.0|string2|2022-07-06|2022-07-06 12:30:00|
|  3|4.0|string3|2022-08-06|2022-08-06 12:30:00|
+---+---+-------+----------+-------------------+
```

5. Explain Spark streaming in detail ?

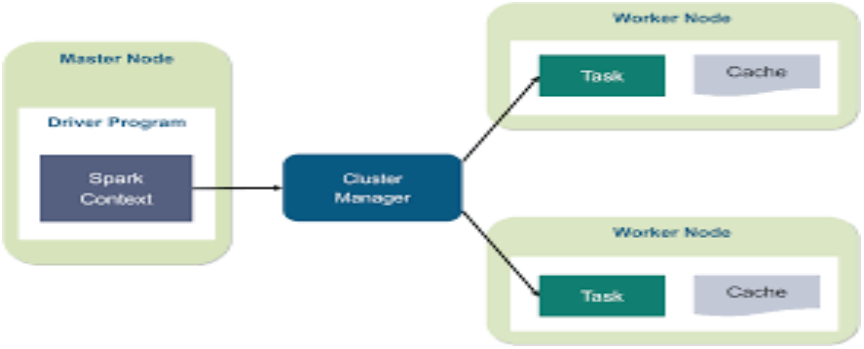
SPARK STREAMING:

- ✓ Spark streaming is a lightweight API that allow developers to perform batch processing and real time streaming of data with ease.
- ✓ Provides Secure,Reliable and fast processing of live data streams.

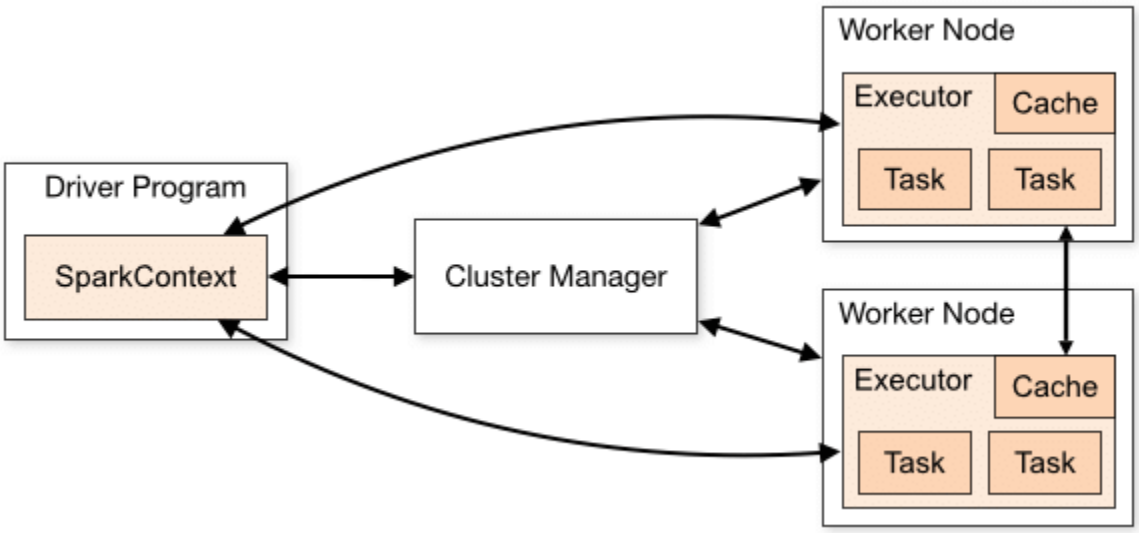


6. Explain SPARK architecure? what is Master - Slave architecure ?

SPARK ARCHITECTURE:



MASTER SLAVE ARCHITECTURE:



7. Explain various cluster managers in SPARK?

CLUSTER MANAGERS IN SPARK:

- ✓ **STANDALONE CLUSTER MANAGER**

It is a part of spark distribution and available as a simple cluster manager to us. Standalone cluster manager is resilient in nature, it can handle work failures. It has capabilities to manage resources according to the requirement of applications.

We can easily run it on *Linux, Windows, or Mac*. It can also access HDFS (Hadoop Distributed File System) data. This is the easiest way to run Apache spark on this cluster. It also has high availability for a master.

- ✓ **HADOOP YARN**

This cluster manager works as a distributed computing framework. It also maintains job scheduling as well as resource management. In this cluster, masters and slaves are highly available for us. We are also available with executors and pluggable scheduler.

We can also run it on Linux and even on windows. Hadoop yarn is also known as MapReduce 2.0. It also bifurcates the functionality of resource manager as well as job scheduling.

- ✓ **APACHE MESOS**

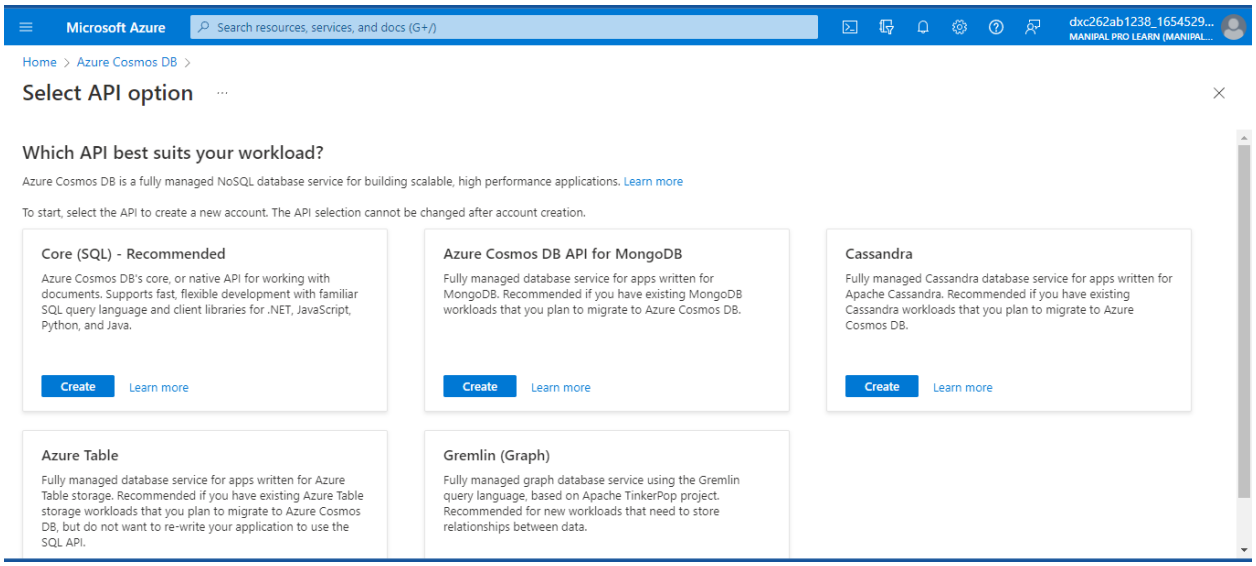
It is a distributed cluster manager. As like yarn, it is also highly available for **maste**r and **slaves**. It can also manage resource per application. We can run spark jobs, Hadoop MapReduce or any other service applications easily.

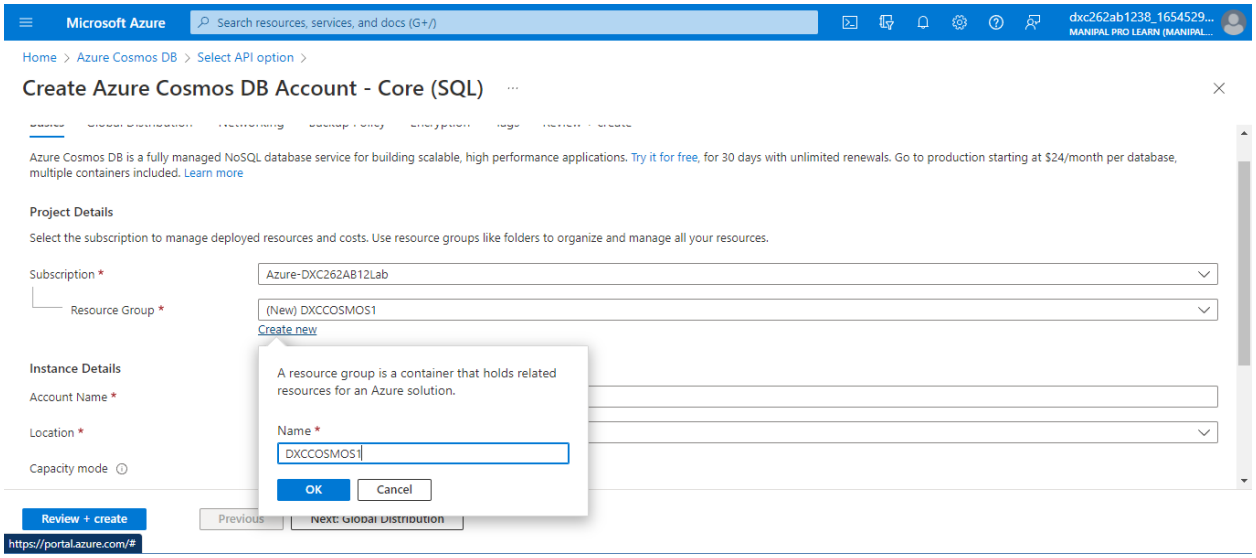
Apache has API's for Java, Python as well as c++. We can run Mesos on Linux or Mac OSX also.

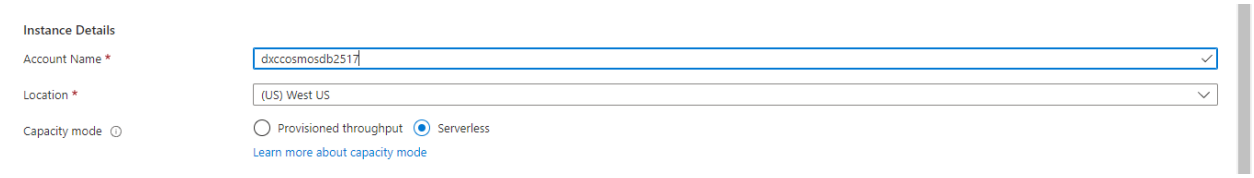
- ✓ **KUBERNETES**

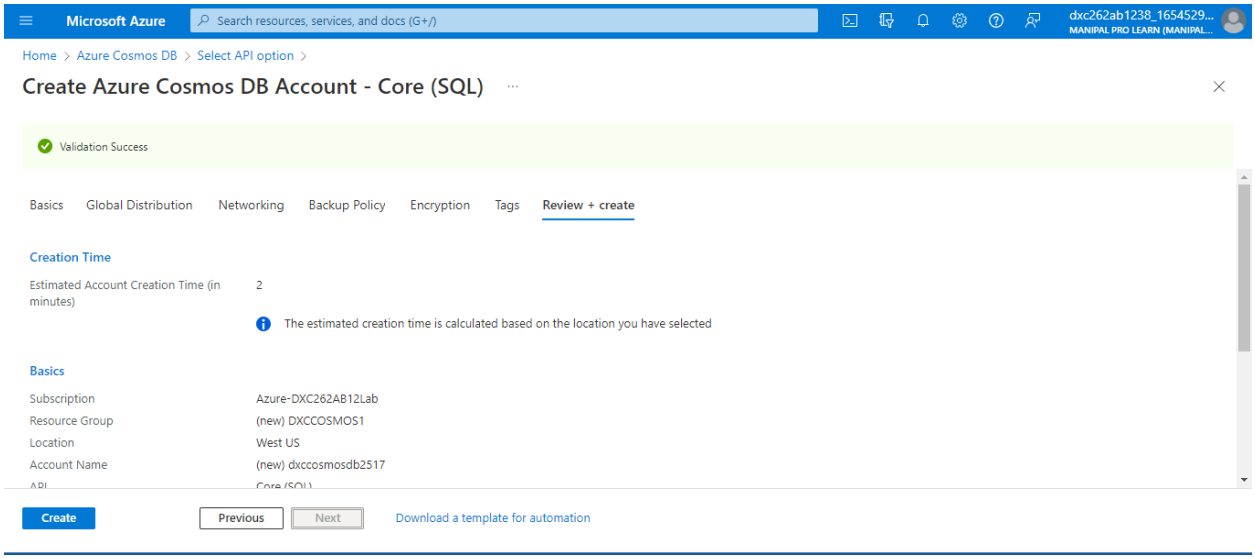
Kubernetes, also known as K8s, is an open-source system for automating deployment, scaling, and management of containerized applications.

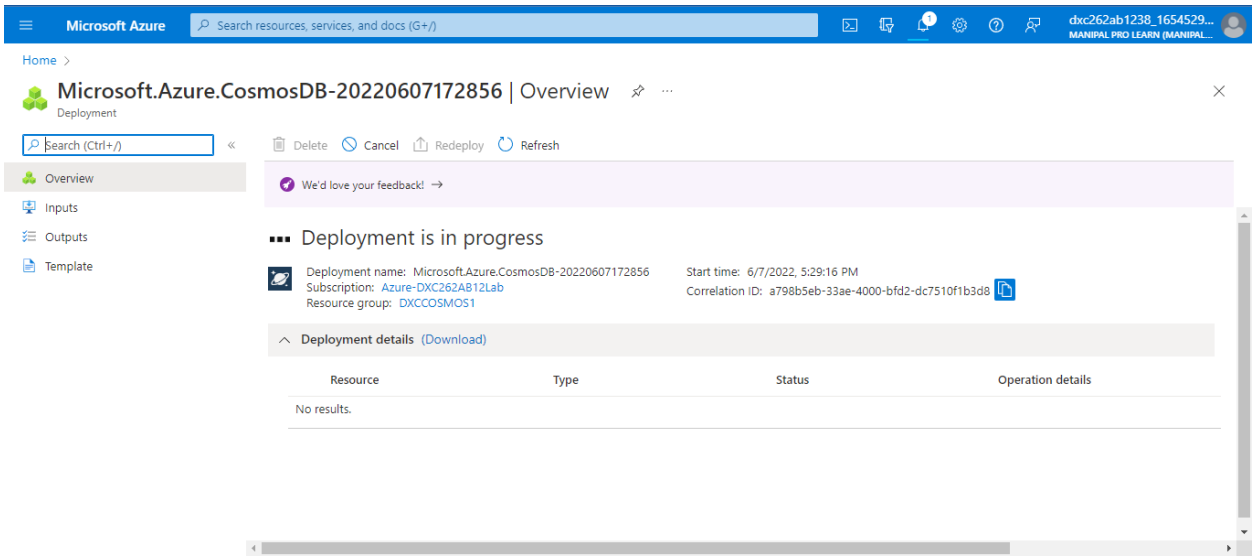
8. Explain with sceenshots & steps how to create Cosmos DB ?

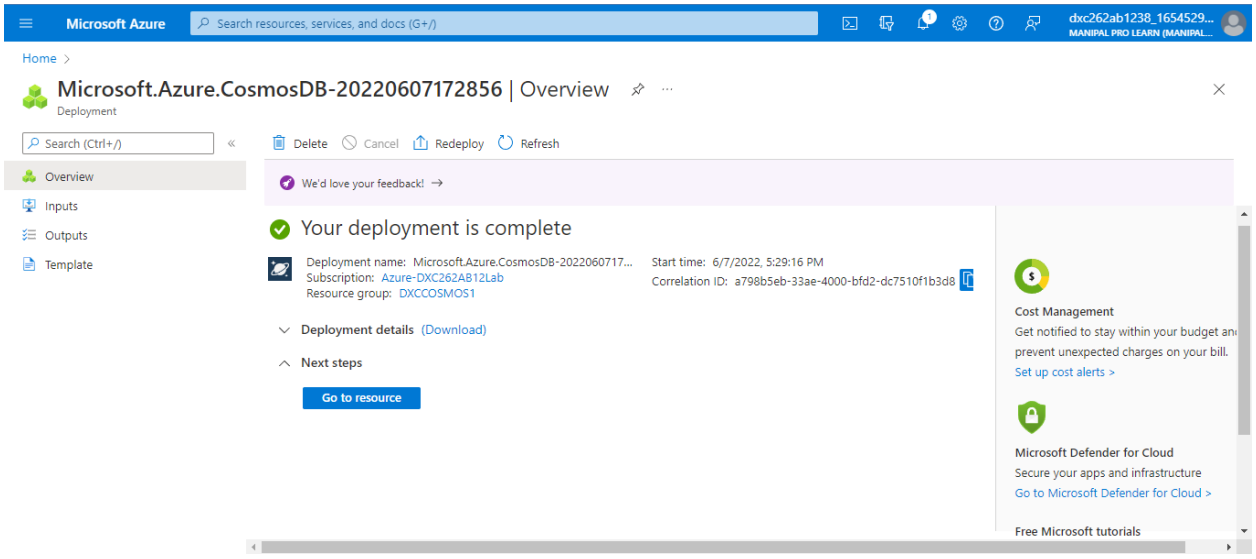


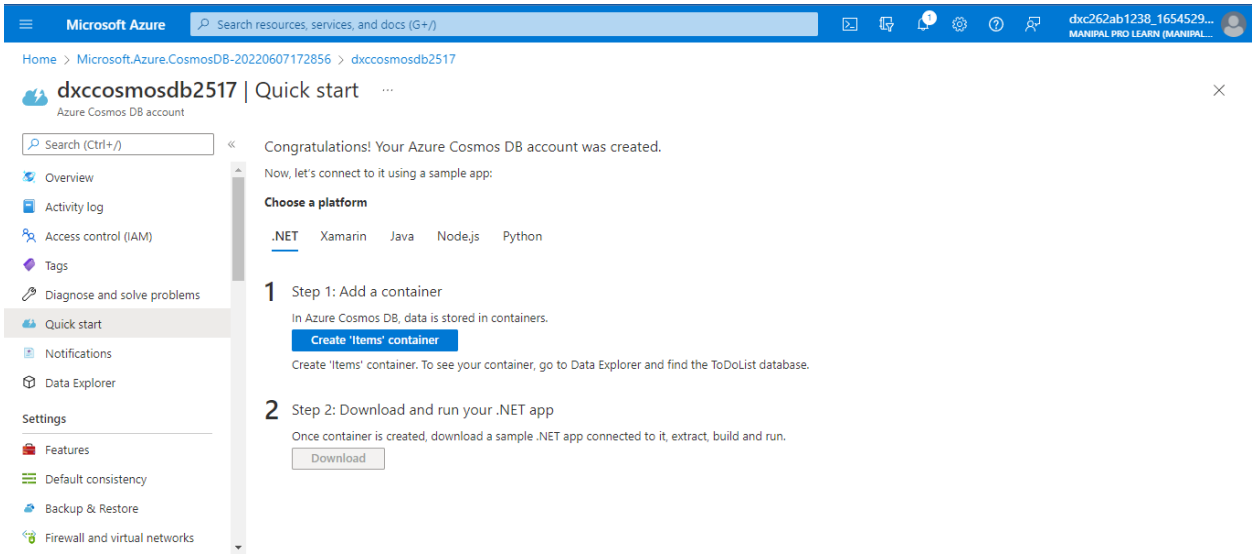



Instance Details
Account Name *
dxccosmosdb2517
Location *
(US) West US
Capacity mode
Provisioned throughput
Serverless
Learn more about capacity mode

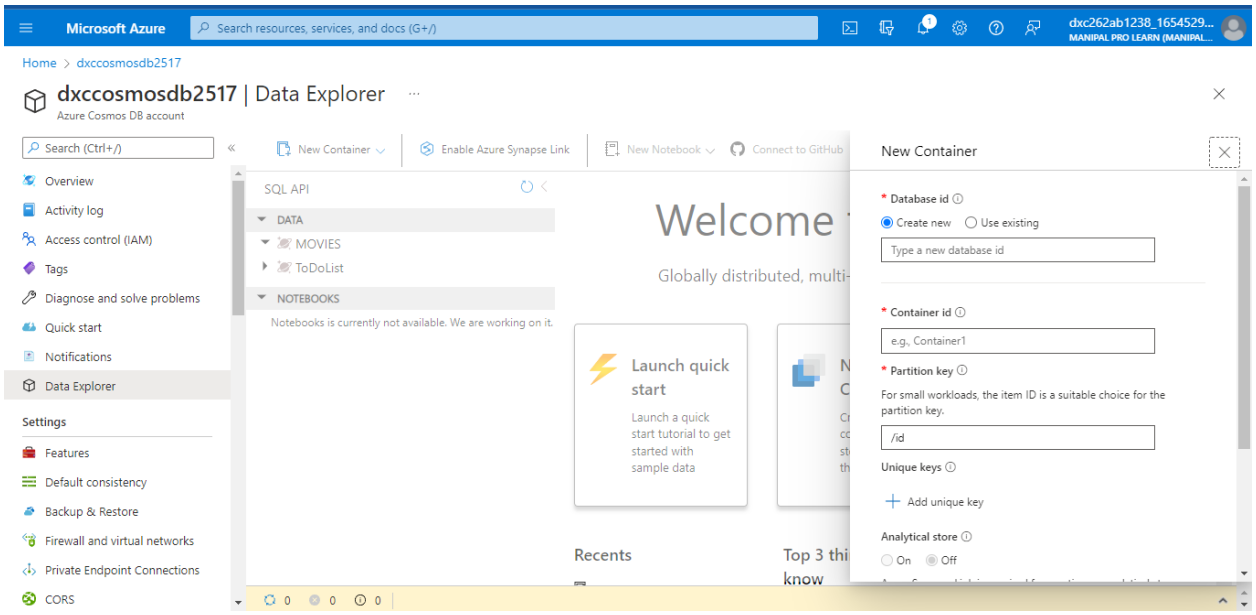

Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1238_1654529...
MANIPAL PRO LEARN (MANIPAL...
Home > Azure Cosmos DB > Select API option >
Create Azure Cosmos DB Account - Core (SQL)
Validation Success
Basics
Global Distribution
Networking
Backup Policy
Encryption
Tags
Review + create
Creation Time
Estimated Account Creation Time (in minutes)
2
The estimated creation time is calculated based on the location you have selected
Basics
Subscription
Azure-DXC262AB12Lab
Resource Group
(new) DXCCOSMOS1
Location
West US
Account Name
(new) dxccosmosdb2517
Create
Previous
Next
Download a template for automation


Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1238_1654529...
MANIPAL PRO LEARN (MANIPAL...
Home >
Microsoft.Azure.CosmosDB-20220607172856 | Overview
Deployment
Search (Ctrl+/)
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Refresh
We'd love your feedback!
Deployment is in progress
Deployment name: Microsoft.Azure.CosmosDB-20220607172856
Subscription: Azure-DXC262AB12Lab
Resource group: DXCCOSMOS1
Start time: 6/7/2022, 5:29:16 PM
Correlation ID: a798b5eb-33ae-4000-bfd2-dc7510f1b3d8
Deployment details (Download)
Resource
Type
Status
Operation details
No results.

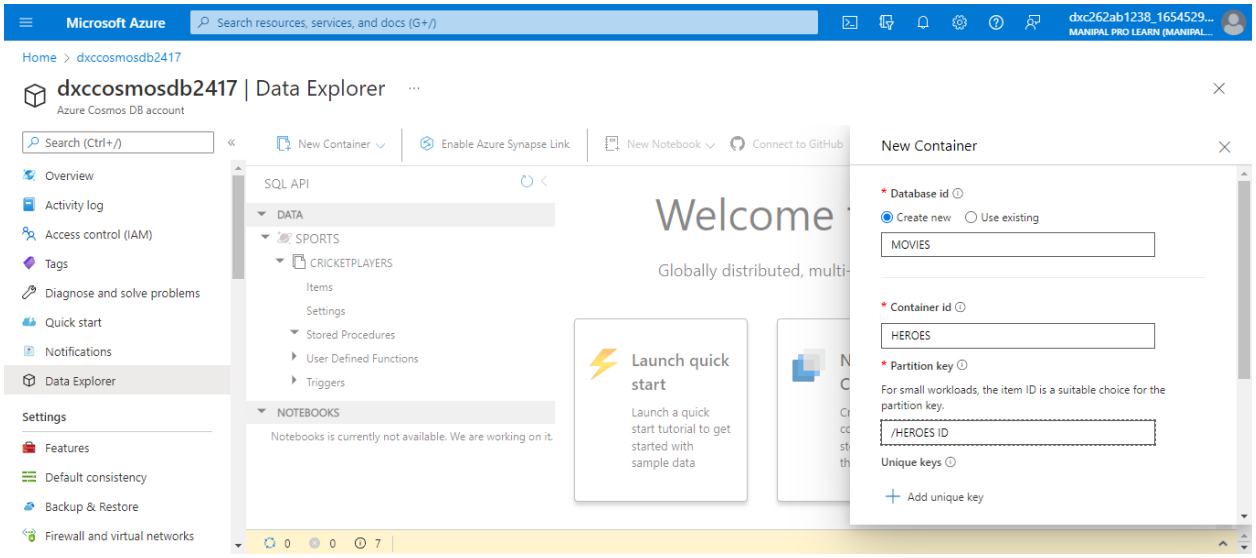

Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1238_1654529...
MANIPAL PRO LEARN (MANIPAL...
Home
Microsoft.Azure.CosmosDB-20220607172856 | Overview
Deployment
Search (Ctrl+/)
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Refresh
We'd love your feedback!
Your deployment is complete
Deployment name: Microsoft.Azure.CosmosDB-2022060717...
Subscription: Azure-DXC262AB12Lab
Resource group: DXCCOSMOS1
Start time: 6/7/2022, 5:29:16 PM
Correlation ID: a798b5eb-33ae-4000-bfd2-dc7510f1b3d8
Deployment details (Download)
Next steps
Go to resource
Cost Management
Get notified to stay within your budget and prevent unexpected charges on your bill.
Set up cost alerts >
Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >
Free Microsoft tutorials

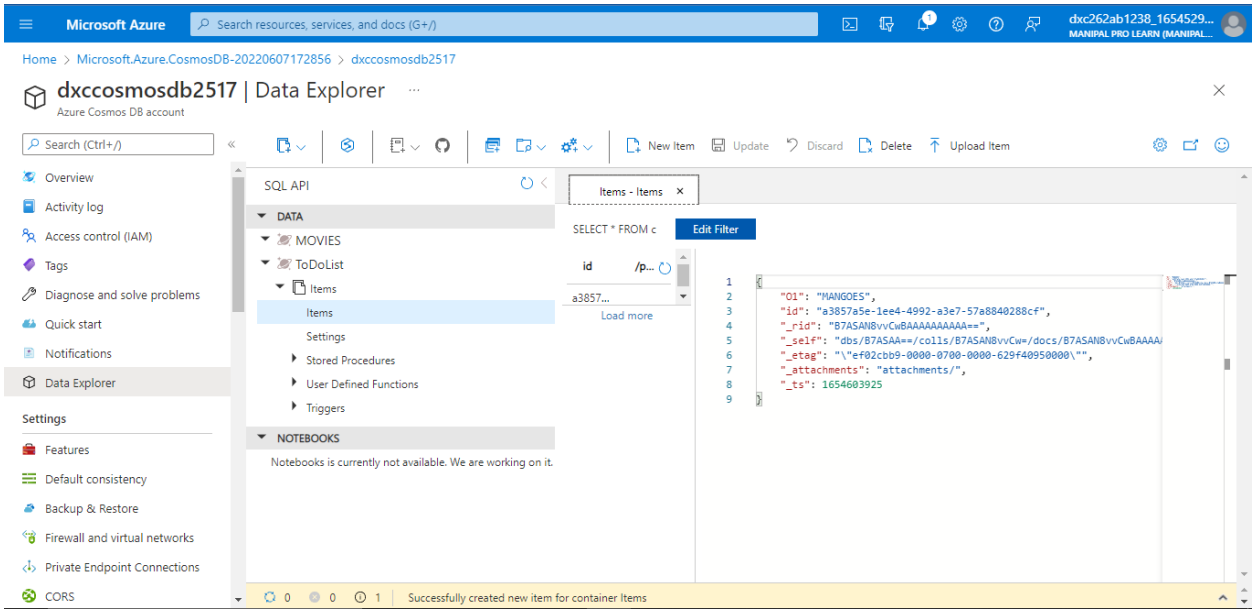

Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1238_1654529...
MANIPAL PRO LEARN (MANIPAL...
Home > Microsoft.Azure.CosmosDB-20220607172856 > dxccosmosdb2517
dxccosmosdb2517 | Quick start
Azure Cosmos DB account
Search (Ctrl+/)
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Quick start
Notifications
Data Explorer
Settings
Features
Default consistency
Backup & Restore
Firewall and virtual networks
Congratulations! Your Azure Cosmos DB account was created.
Now, let's connect to it using a sample app:
Choose a platform
.NET
Xamarin
Java
Node.js
Python
1 Step 1: Add a container
In Azure Cosmos DB, data is stored in containers.
Create 'Items' container
Create 'Items' container. To see your container, go to Data Explorer and find the ToDoList database.
2 Step 2: Download and run your .NET app
Once container is created, download a sample .NET app connected to it, extract, build and run.
Download

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?




Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1238_1654529...
MANIPAL PRO LEARN (MANIPAL...
Home > Microsoft.Azure.CosmosDB-20220607172856 > dxccosmosdb2517
dxccosmosdb2517 | Data Explorer
Azure Cosmos DB account
Search (Ctrl+/)
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Quick start
Notifications
Data Explorer
Settings
Features
Default consistency
Backup & Restore
Firewall and virtual networks
Private Endpoint Connections
CORS
New Item
Update
Discard
Delete
Upload Item
SQL API
DATA
MOVIES
ToDoList
Items
Items
Settings
Stored Procedures
User Defined Functions
Triggers
NOTEBOOKS
Notebooks is currently not available. We are working on it.
Items - Items
SELECT * FROM c
Edit Filter
id
/p...
a3857...
Load more
{
"O1": "MANGOES",
"id": "a3857a5e-1ee4-4992-a3e7-57a8840288cf",
"_rid": "B7ASAN8vvCwBAAAAAAAAAA==",
"_self": "dbs/B7ASAA==/colls/B7ASAN8vvCw=/docs/B7ASAN8vvCwBAAAA
"_etag": "\"ef02cbb9-0000-0700-0000-629f40950000\"",
"_attachments": "attachments/",
"_ts": 1654603925
}
0
0
1
Successfully created new item for container Items