NAME: BHOGADI NAGA ISWARYA LAKSHMI **SUBMISSION:**07-06-2022
BATCH: DXC-262-ANALYTICS-B12-AZURE **ASSESSMENT:**7

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

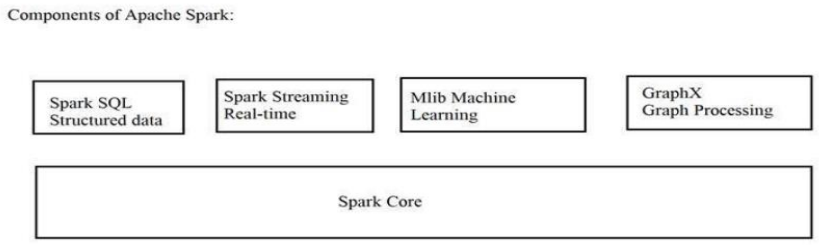
Shark: Shark is one of the Spark Ecosystem components. It is used to perform structured data analysis, especially if the data is too voluminous. Shark also allows running unmodified Hive queries on existing Hadoop deployment.

Spark Streaming: Spark Streaming is one of those unique features, which have empowered Spark to potentially take the role of Apache Storm. Spark Streaming mainly enables you to create analytical and interactive applications for live streaming data. You can do the streaming of the data and then, Spark can run its operations from the streamed data itself.

MLLib:MLLib is a machine learning library like Mahout. It is built on top of Spark, and has the provision to support many machine learning algorithms. But the point difference with Mahout is that it runs almost 100 times faster than MapReduce. It is not yet as enriched as Mahout, but it is coming up pretty well, even though it is still in the initial stage of growth.

GraphX: For graphs and graphical computations, Spark has its own Graph Computation Engine, called GraphX. It is similar to other widely used graph processing tools or databases, like Neo4j, Girafe, and many other distributed graph databases.

SparkR: There are many people from data science track, who must be aware that for statistical analysis, R is among the best. There is already an integration of R with Hadoop. Now, SparkR is a package for R language to enable R users to leverage the power of Spark from R shell.



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

Spark Core is the base of the whole project. It provides distributed task dispatching, scheduling, and basic I/O functionalities. Spark uses a specialized fundamental data structure known as RDD (Resilient Distributed Datasets) that is a logical collection of data partitioned across machines.

RDD was the primary user-facing API in Spark since its inception. At the core, an RDD is an immutable distributed collection of elements of your data, partitioned across nodes in your cluster that can be operated in parallel with a low-level API that offers transformations and actions.

3. Explain various Mlib algorithms Spark is supporting ?

-IT is a low-level machine learning library that is simple to use, is scalable and

compatible with various programming languages.

-MLlib eases the deployment and development of scalable machine learning algorithms

-It contains machine learning libraries that have an implementation of various machine

learning algorithms

1)Clustering

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).

2)Classification

The spark.mllib package supports various methods for binary classification, multiclass classification and regression analysis. Some of the most popular algorithms in classification are Random Forest, Naive Bayes, Decision Tree, etc.

3)Collaborative Filtering

Collaborative filtering is commonly used for recommender systems. These techniques aim to fill in the missing entries of a user-item association matrix. spark.mllib currently supports model-based collaborative filtering, in which users and products are described by a small set of latent factors that can be used to predict missing entries.

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

BENEFITS OF SparkSQL

- High compatibility
- Scalability
- Performance Optimization
- Standard Connectivity
- Unified Dara Access
- Integrated

Apache Spark has multiple ways to read data from different sources like files, databases etc. But when it comes to loading data into RDBMS(relational database management system), Spark supports only Append and Overlay of the data using dataframes. Spark dataframes do not support Updating of data into a database.

- Read data from a CSV file
- Create a database schema and table in MySQL db
- Load spark dataframe data into a database.
- Update database table records using Spark

5. Explain Spark streaming in detail ?

Apache Spark Streaming is a scalable fault-tolerant streaming processing system that natively supports both batch and streaming workloads. Spark Streaming is an extension of the core Spark API that allows data engineers and data scientists to process real-time data from various sources including (but not limited to) Kafka, Flume, and Amazon Kinesis. This processed data can be pushed out to file systems, databases, and live dashboards. Its key abstraction is a Discretized Stream or, in short, a DStream, which represents a stream of data divided into small batches. DStreams are built on RDDs, Spark's core data abstraction. This allows Spark Streaming to seamlessly integrate with any other Spark components like MLlib and Spark SQL. Spark Streaming is different from other systems that either have a processing engine designed only for streaming, or have similar batch and streaming APIs but compile internally to different engines. Spark's single execution engine and unified programming model for batch and streaming lead to some unique benefits over other traditional streaming systems.

Four Major Aspects of Spark Streaming

- Fast recovery from failures and stragglers
- Better load balancing and resource usage
- Combining of streaming data with static datasets and interactive queries
- Native integration with advanced processing libraries (SQL, machine learning, graph processing)

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

The Spark follows the master-slave architecture. Its cluster consists of a single master and multiple slaves.

The Spark architecture depends upon two abstractions:

- Resilient Distributed Dataset (RDD)
- Directed Acyclic Graph (DAG)

Resilient Distributed Datasets (RDD)

The Resilient Distributed Datasets are the group of data items that can be stored in-memory on worker nodes. Here,

- Resilient: Restore the data on failure.
- Distributed: Data is distributed among different nodes.
- Dataset: Group of data.

Directed Acyclic Graph (DAG)

Directed Acyclic Graph is a finite direct graph that performs a sequence of computations on data. Each node is an RDD partition, and the edge is a transformation on top of data. Here, the graph refers the navigation whereas directed and acyclic refers to how it is done.

Master - Slave architecture

Master-slave architectures are used to help stabilize a system. Master is the true data keeper while a slave is a replication of master. Cache/caching is an option but using it as complementary to the master-slave system would be better. Replication is the process of synchronizing data from the master to slave.

7. Explain various cluster managers in SPARK?

Cluster manager is a platform (cluster mode) where we can run Spark. Simply put, cluster manager provides resources to all worker nodes as per need, it operates all nodes accordingly.
We can say there are a master node and worker nodes available in a cluster. That master nodes provide an efficient working environment to worker nodes.

There are three types of Spark cluster manager. Spark supports these cluster manager:

1. Standalone cluster manager
2. Hadoop Yarn
3. Apache Mesos

Apache Spark also supports pluggable cluster management. The main task of cluster manager is to provide resources to all applications. We can say it is an external service for acquiring required resources on the cluster.

Let's discuss all these cluster managers in detail:

1. 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.

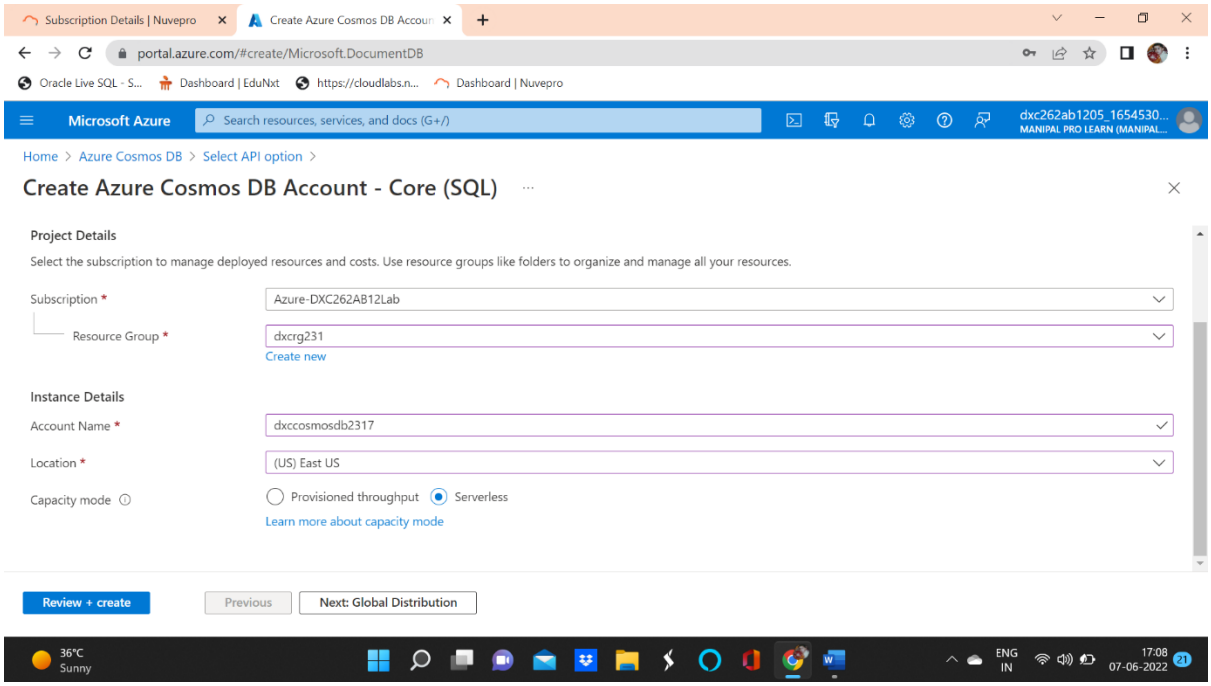
2. Hadoop Yarn

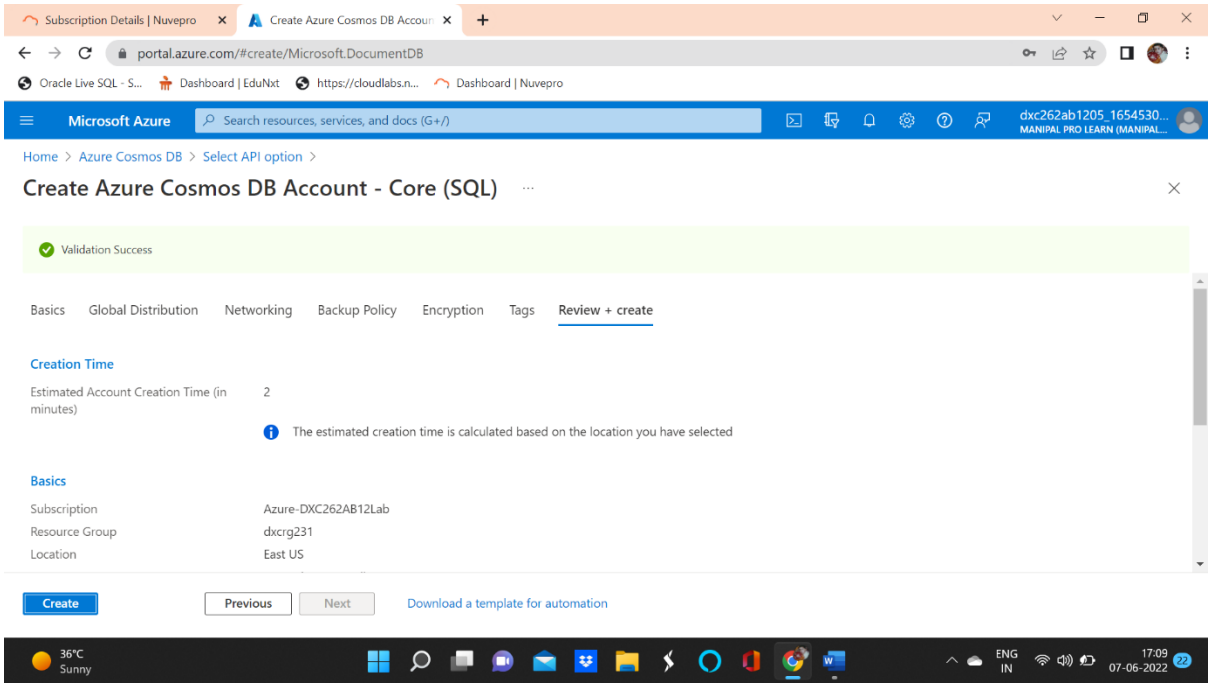
This cluster manager works as a distributed computing framework. It also maintains job scheduling as well as resource management. In this cluster, masters and slaves are highly available for us. We are also available with executors and pluggable scheduler.

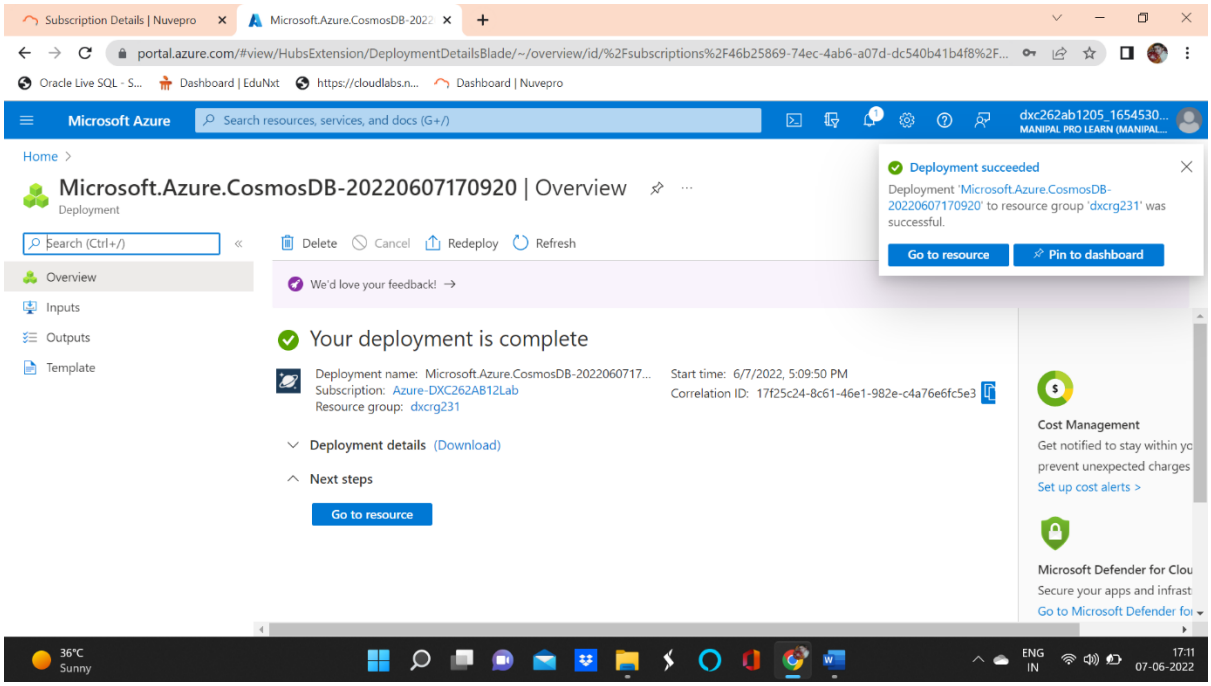
3. Apache Mesos

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

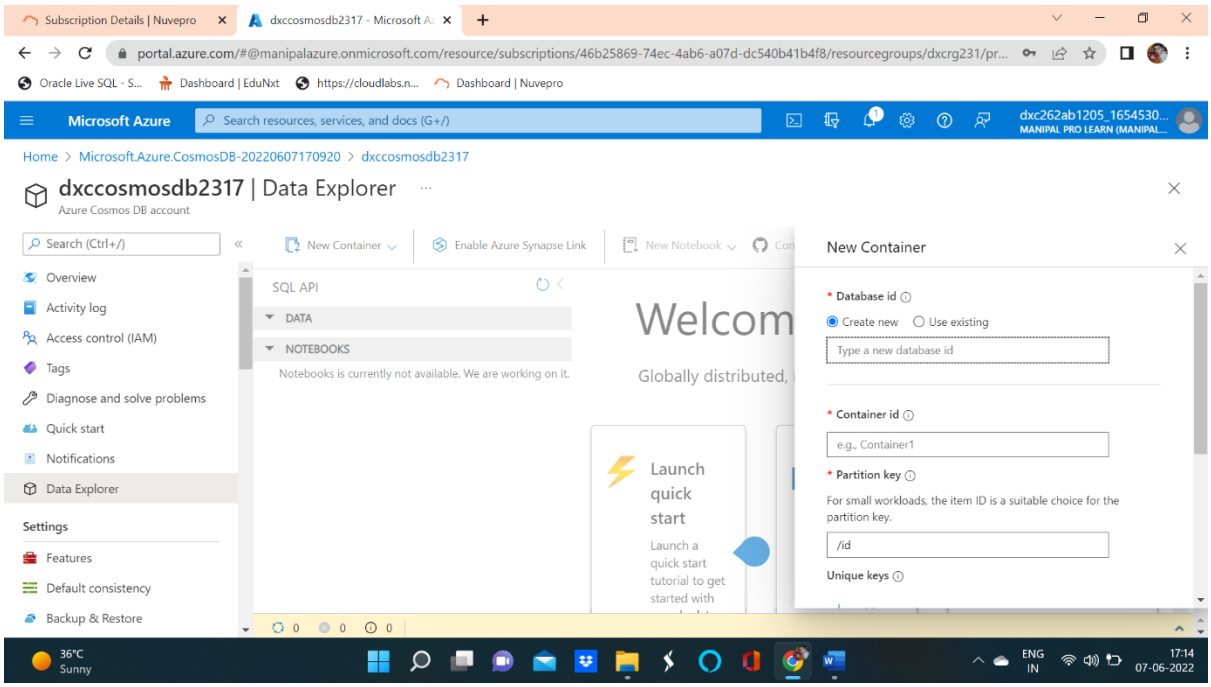
8. Explain with sceenshots & steps how to create Cosmos DB ?

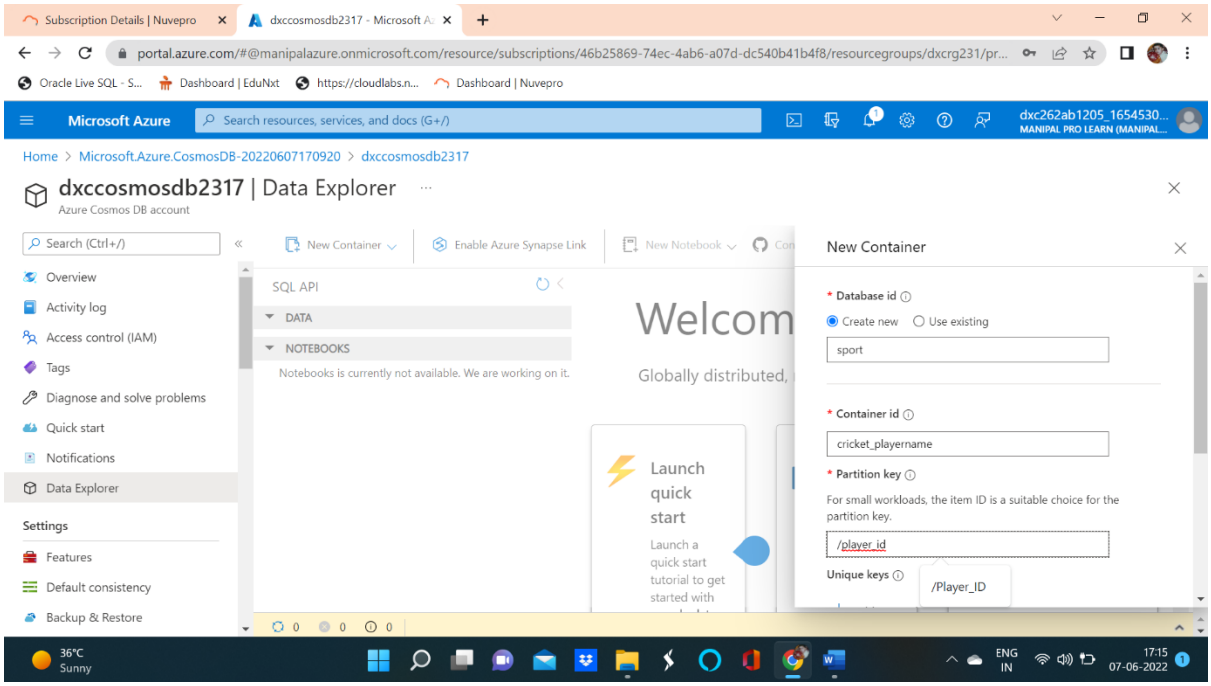

Subscription Details | Nuvepro
Create Azure Cosmos DB Accour
portal.azure.com/#create/Microsoft.DocumentDB
Oracle Live SQL - S...
Dashboard | EduNxt
https://cloudlabs.n...
Dashboard | Nuvepro
Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1205_1654530...
MANIPAL PRO LEARN (MANIPAL...
Home > Azure Cosmos DB > Select API option >
Create Azure Cosmos DB Account - Core (SQL)
Project Details
Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.
Subscription *
Azure-DXC262AB12Lab
Resource Group *
dxcrg231
Create new
Instance Details
Account Name *
dxccosmosdb2317
Location *
(US) East US
Capacity mode
Provisioned throughput
Serverless
Learn more about capacity mode
Review + create
Previous
Next: Global Distribution
36°C
Sunny
ENG
IN
17:08
07-06-2022

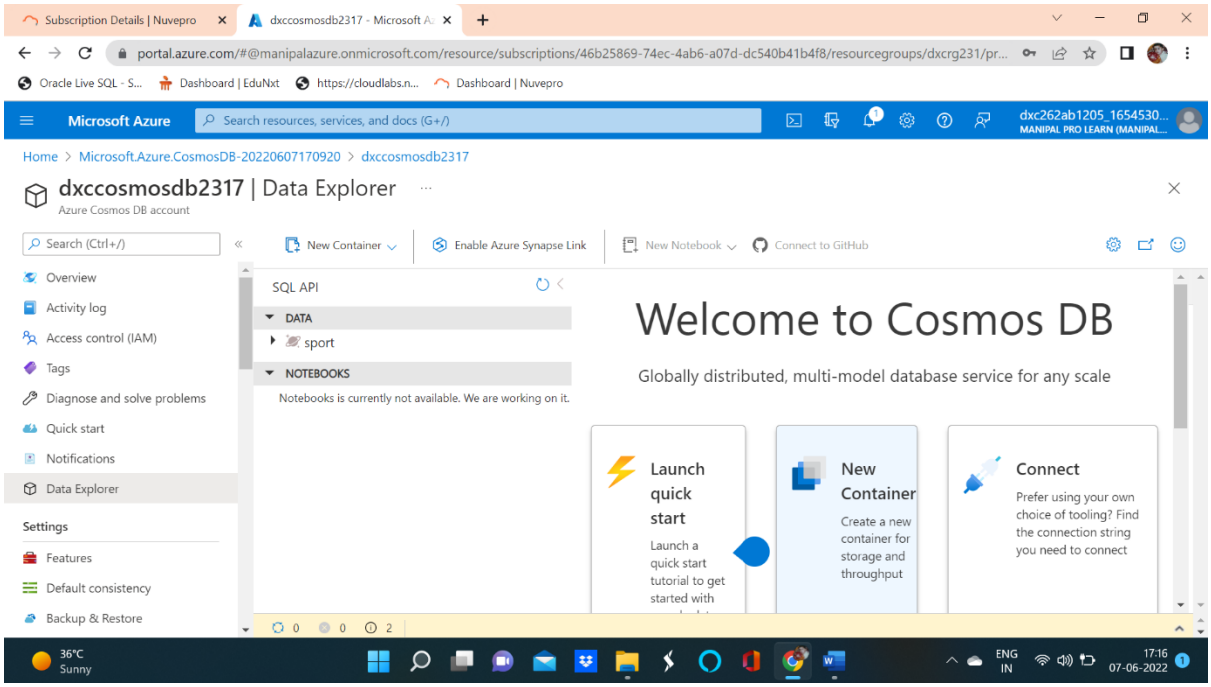

Subscription Details | Nuvepro
Create Azure Cosmos DB Accour
portal.azure.com/#create/Microsoft.DocumentDB
Oracle Live SQL - S...
Dashboard | EduNxt
https://cloudlabs.n...
Dashboard | Nuvepro
Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1205_1654530...
MANIPAL PRO LEARN (MANIPAL...
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
dxcrg231
Location
East US
Create
Previous
Next
Download a template for automation
36°C
Sunny
ENG
IN
17:09
07-06-2022

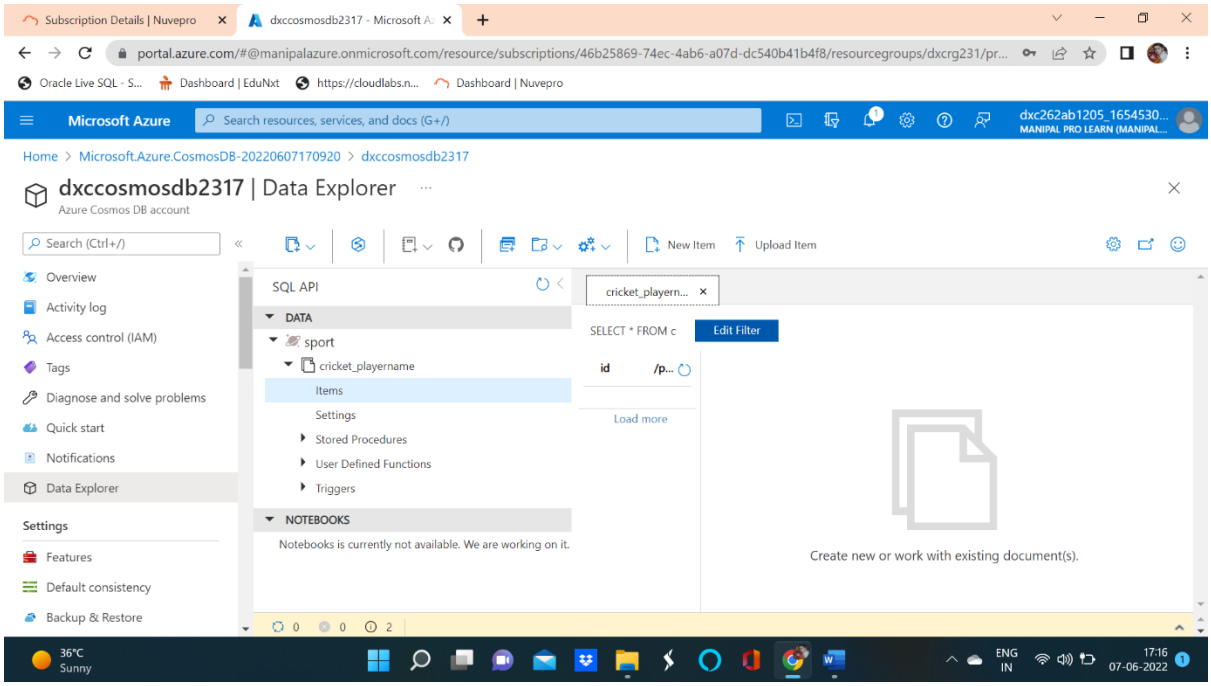


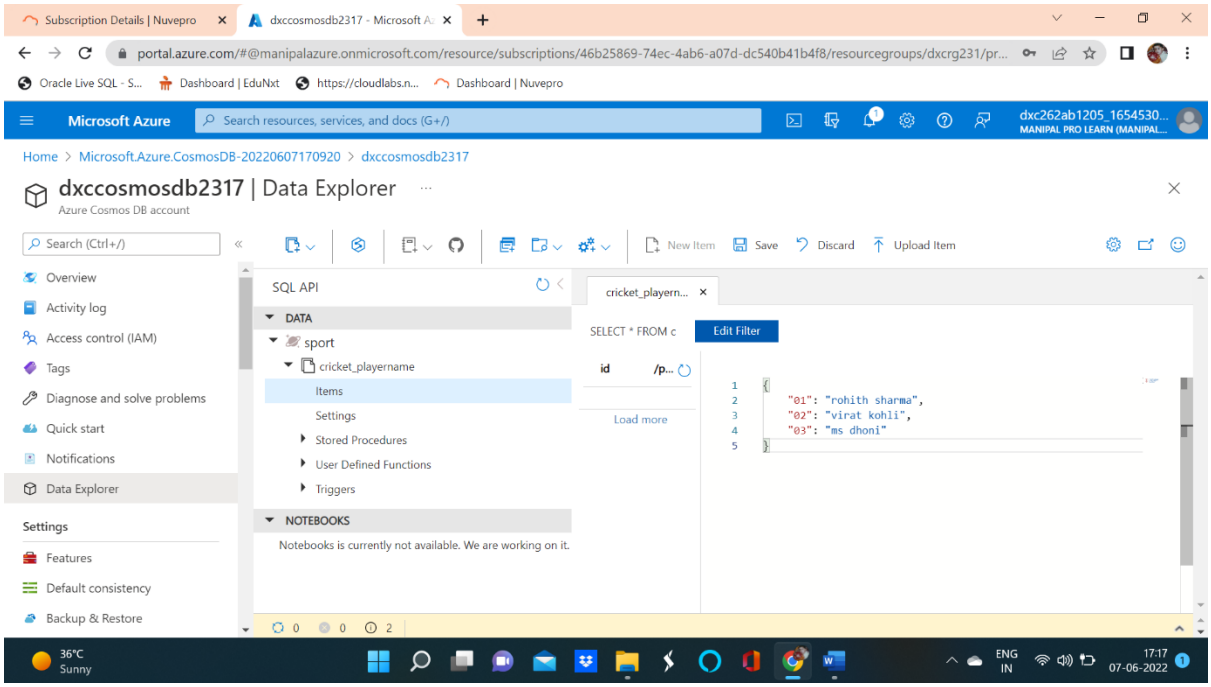
9. Explain with sceenshots & step how to insert data into Cosmos DB?

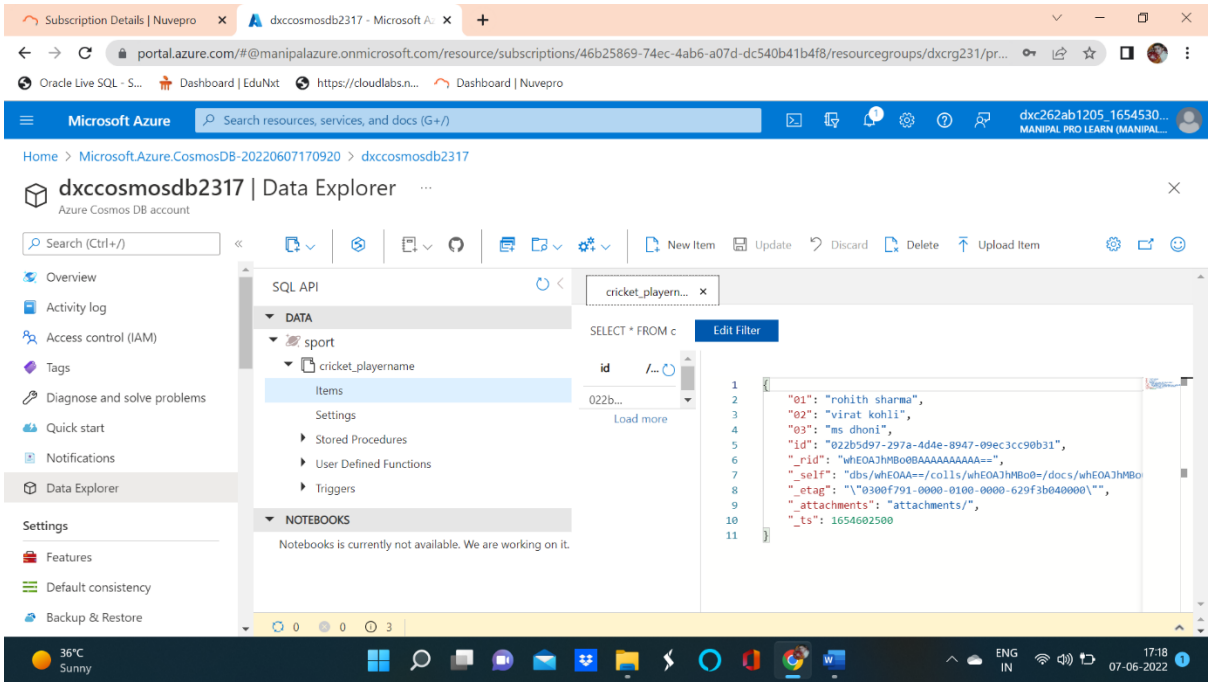



Subscription Details | Nuvepro
dxccosmosdb2317 - Microsoft A
portal.azure.com/#@manipalazure.onmicrosoft.com/resource/subscriptions/46b25869-74ec-4ab6-a07d-dc540b41b4f8/resourcegroups/dxcrg231/pr...
Oracle Live SQL - S...
Dashboard | EduNxt
https://cloudlabs.n...
Dashboard | Nuvepro
Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1205_1654530...
MANIPAL PRO LEARN (MANIPAL...
Home > Microsoft.Azure.CosmosDB-20220607170920 > dxccosmosdb2317
dxccosmosdb2317 | Data Explorer
Azure Cosmos DB account
Search (Ctrl+/)
New Container
Enable Azure Synapse Link
New Notebook
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Quick start
Notifications
Data Explorer
Settings
Features
Default consistency
Backup & Restore
SQL API
DATA
NOTEBOOKS
Notebooks is currently not available. We are working on it.
Welcom
Globally distributed,
Launch quick start
Launch a quick start tutorial to get started with
New Container
* Database id
Create new
Use existing
sport
* Container id
cricket_playername
* Partition key
For small workloads, the item ID is a suitable choice for the partition key.
/player_id
Unique keys
/Player_ID
36°C
Sunny
ENG
IN
17:15
07-06-2022

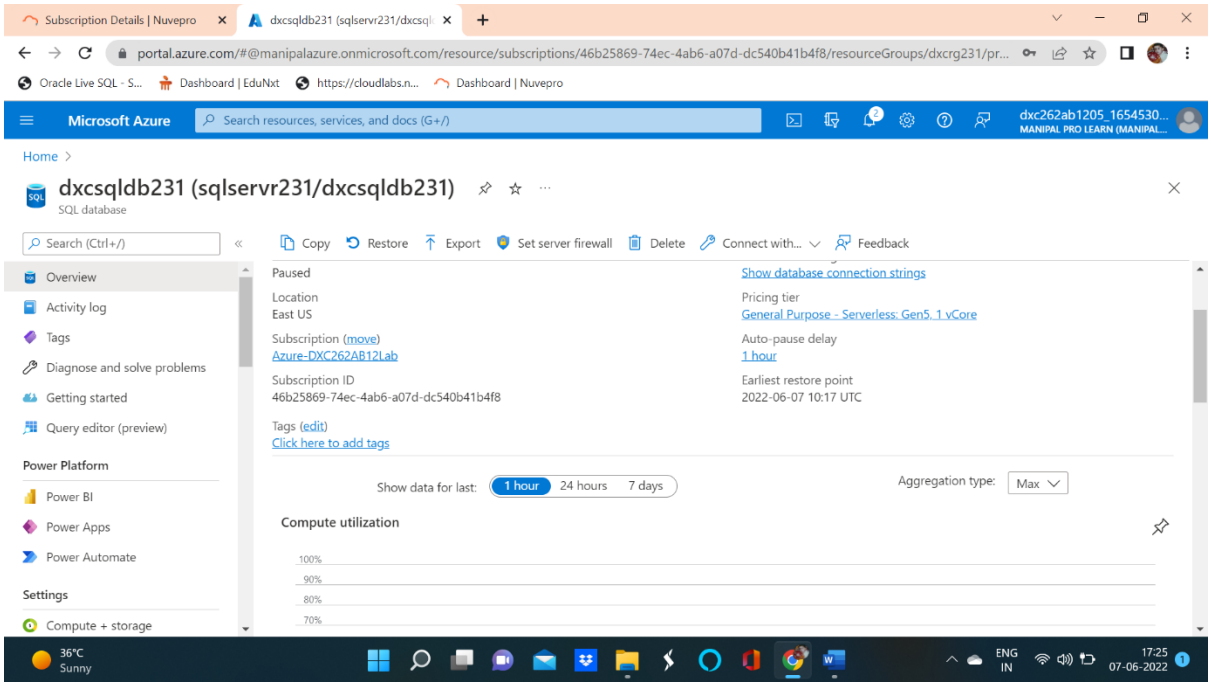

Subscription Details | Nuvepro
dxccosmosdb2317 - Microsoft A
portal.azure.com/#@manipalazure.onmicrosoft.com/resource/subscriptions/46b25869-74ec-4ab6-a07d-dc540b41b4f8/resourcegroups/dxcrg231/pr...
Oracle Live SQL - S...
Dashboard | EduNxt
https://cloudlabs.n...
Dashboard | Nuvepro
Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1205_1654530...
MANIPAL PRO LEARN (MANIPAL...
Home > Microsoft.Azure.CosmosDB-20220607170920 > dxccosmosdb2317
dxccosmosdb2317 | Data Explorer
Azure Cosmos DB account
Search (Ctrl+/)
New Container
Enable Azure Synapse Link
New Notebook
Connect to GitHub
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Quick start
Notifications
Data Explorer
Settings
Features
Default consistency
Backup & Restore
SQL API
DATA
sport
NOTEBOOKS
Notebooks is currently not available. We are working on it.
Welcome to Cosmos DB
Globally distributed, multi-model database service for any scale
Launch quick start
Launch a quick start tutorial to get started with
New Container
Create a new container for storage and throughput
Connect
Prefer using your own choice of tooling? Find the connection string you need to connect
36°C
Sunny
ENG
IN
17:16
07-06-2022

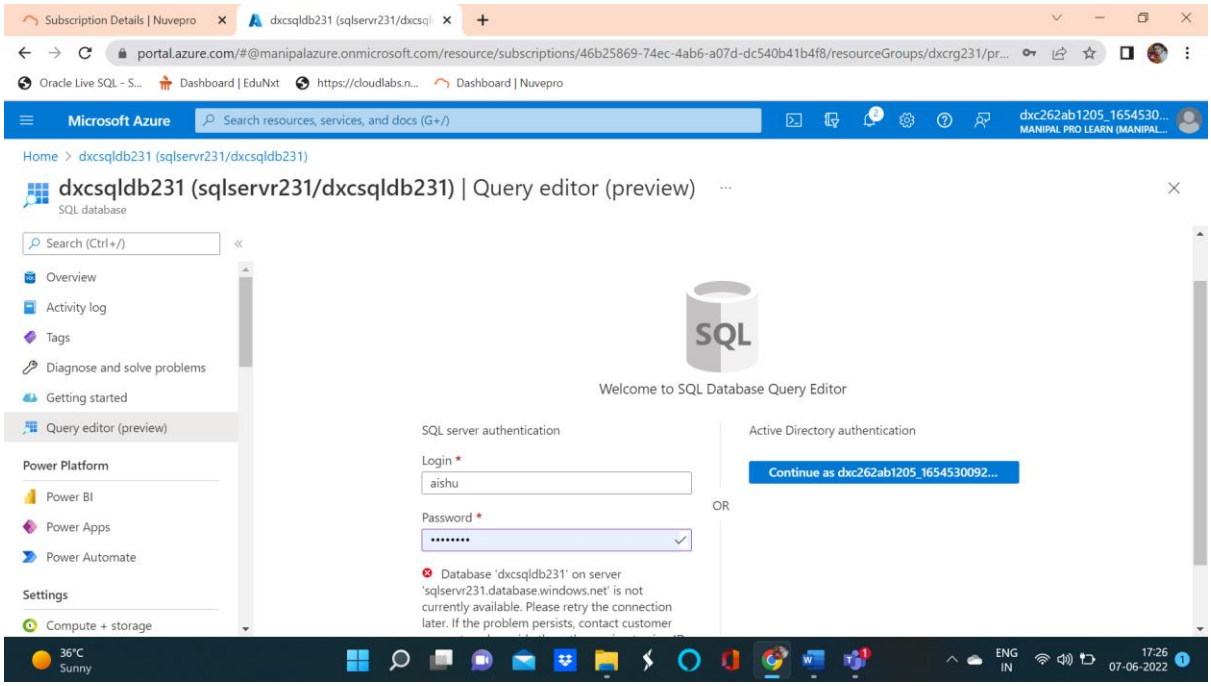

Subscription Details | Nuvepro
dxccosmosdb2317 - Microsoft A
portal.azure.com/#@manipalazure.onmicrosoft.com/resource/subscriptions/46b25869-74ec-4ab6-a07d-dc540b41b4f8/resourcegroups/dxcrg231/pr...
Oracle Live SQL - S...
Dashboard | EduNxt
https://cloudlabs.n...
Dashboard | Nuvepro
Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1205_1654530...
MANIPAL PRO LEARN (MANIPAL...
Home > Microsoft.Azure.CosmosDB-20220607170920 > dxccosmosdb2317
dxccosmosdb2317 | Data Explorer
Azure Cosmos DB account
Search (Ctrl+/)
New Item
Upload Item
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Quick start
Notifications
Data Explorer
Settings
Features
Default consistency
Backup & Restore
SQL API
DATA
sport
cricket_playername
Items
Settings
Stored Procedures
User Defined Functions
Triggers
NOTEBOOKS
Notebooks is currently not available. We are working on it.
cricket_playern...
SELECT * FROM c
Edit Filter
id
/p...
Load more
Create new or work with existing document(s).
36°C
Sunny
ENG
IN
17:16
07-06-2022

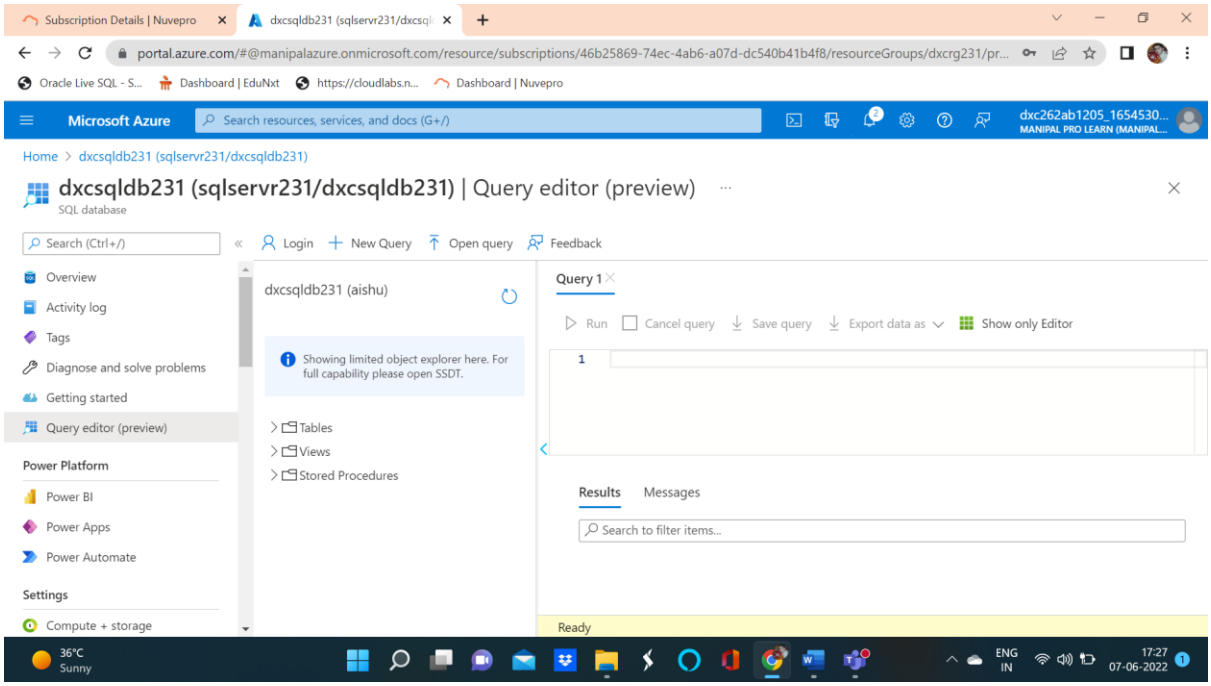

Subscription Details | Nuvepro
dxccosmosdb2317 - Microsoft A
portal.azure.com/#@manipalazure.onmicrosoft.com/resource/subscriptions/46b25869-74ec-4ab6-a07d-dc540b41b4f8/resourcegroups/dxcrg231/pr...
Oracle Live SQL - S...
Dashboard | EduNxt
https://cloudlabs.n...
Dashboard | Nuvepro
Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1205_1654530...
MANIPAL PRO LEARN (MANIPAL...
Home > Microsoft.Azure.CosmosDB-20220607170920 > dxccosmosdb2317
dxccosmosdb2317 | Data Explorer
Azure Cosmos DB account
Search (Ctrl+/)
New Item
Save
Discard
Upload Item
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Quick start
Notifications
Data Explorer
Settings
Features
Default consistency
Backup & Restore
SQL API
DATA
sport
cricket_playername
Items
Settings
Stored Procedures
User Defined Functions
Triggers
NOTEBOOKS
Notebooks is currently not available. We are working on it.
cricket_playern...
SELECT * FROM c
Edit Filter
id
/p...
Load more
{
"01": "rohith sharma",
"02": "virat kohli",
"03": "ms dhoni"
}
36°C
Sunny
ENG
IN
17:17
07-06-2022

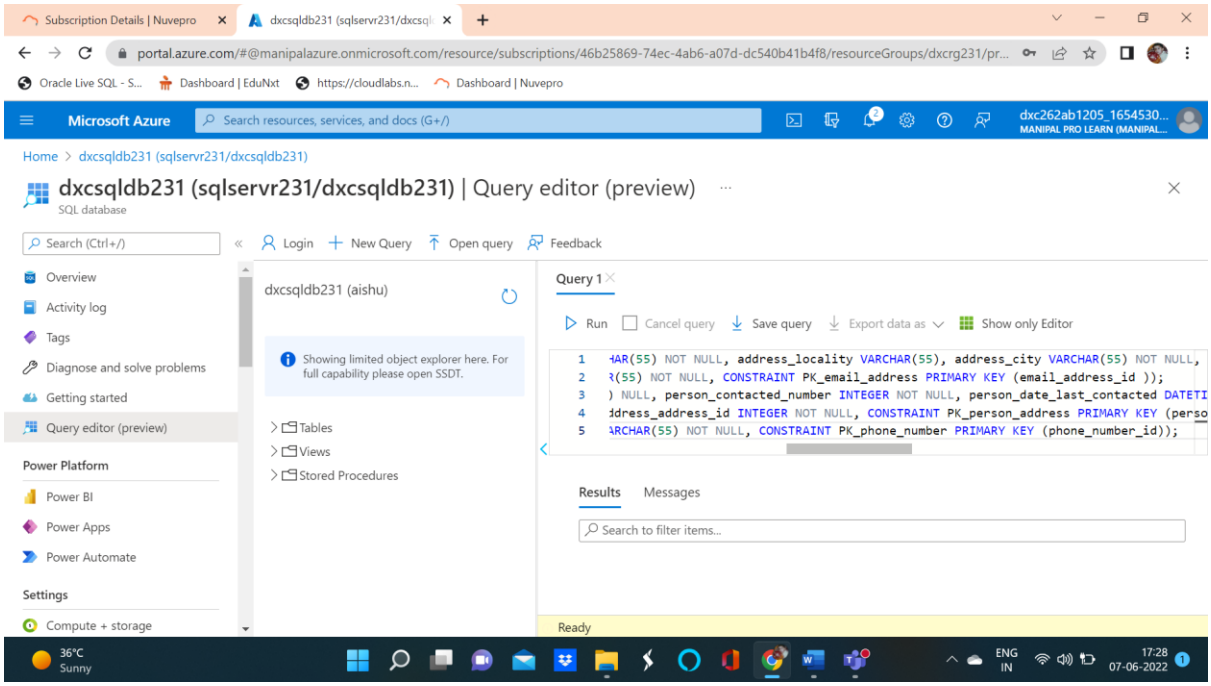


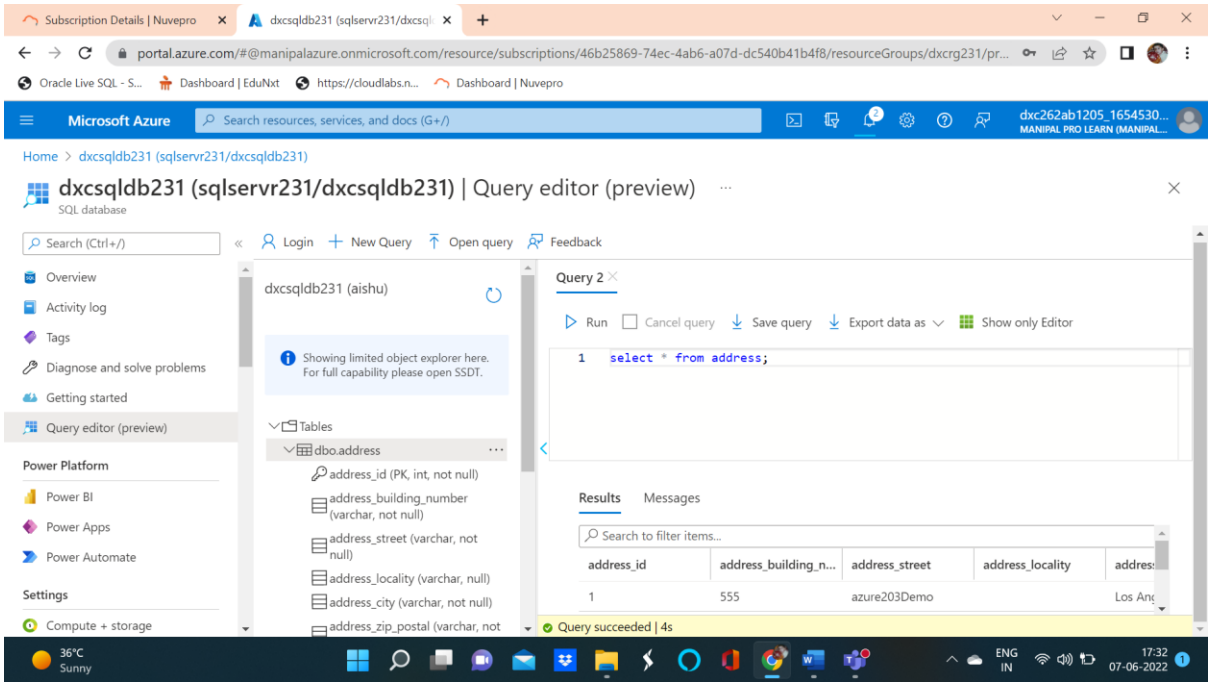
10. Explain with sceenshots & step how to create Azure SQL Db & also explain how to insert data into Azure SQL D?




Subscription Details | Nuvepro
dxcsqldb231 (sqlservr231/dxcsql
portal.azure.com/#@manipalazure.onmicrosoft.com/resource/subscriptions/46b25869-74ec-4ab6-a07d-dc540b41b4f8/resourceGroups/dxcrg231/pr...
Oracle Live SQL - S...
Dashboard | EduNxt
https://cloudlabs.n...
Dashboard | Nuvepro
Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1205_1654530...
MANIPAL PRO LEARN (MANIPAL...
Home > dxcsqldb231 (sqlservr231/dxcsqldb231)
dxcsqldb231 (sqlservr231/dxcsqldb231) | Query editor (preview)
SQL database
Search (Ctrl+/)
Overview
Activity log
Tags
Diagnose and solve problems
Getting started
Query editor (preview)
Power Platform
Power BI
Power Apps
Power Automate
Settings
Compute + storage
SQL
Welcome to SQL Database Query Editor
SQL server authentication
Login *
aishu
Password *
Database 'dxcsqldb231' on server 'sqlservr231.database.windows.net' is not currently available. Please retry the connection later. If the problem persists, contact customer
OR
Active Directory authentication
Continue as dxc262ab1205_1654530092...
36°C
Sunny
ENG
IN
17:26
07-06-2022


Subscription Details | Nuvepro
dxcsqldb231 (sqlservr231/dxcsql
portal.azure.com/#@manipalazure.onmicrosoft.com/resource/subscriptions/46b25869-74ec-4ab6-a07d-dc540b41b4f8/resourceGroups/dxcrg231/pr...
Oracle Live SQL - S...
Dashboard | EduNxt
https://cloudlabs.n...
Dashboard | Nuvepro
Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1205_1654530...
MANIPAL PRO LEARN (MANIPAL...
Home > dxcsqldb231 (sqlservr231/dxcsqldb231)
dxcsqldb231 (sqlservr231/dxcsqldb231) | Query editor (preview)
SQL database
Search (Ctrl+/)
Login
New Query
Open query
Feedback
Overview
Activity log
Tags
Diagnose and solve problems
Getting started
Query editor (preview)
Power Platform
Power BI
Power Apps
Power Automate
Settings
Compute + storage
dxcsqldb231 (aishu)
Showing limited object explorer here. For full capability please open SSDT.
Tables
Views
Stored Procedures
Query 1
Run
Cancel query
Save query
Export data as
Show only Editor
1
Results
Messages
Search to filter items...
Ready
36°C
Sunny
ENG
IN
17:27
07-06-2022


Subscription Details | Nuvepro
dxcsqldb231 (sqlserver231/dxcsql
portal.azure.com/#@manipalazure.onmicrosoft.com/resource/subscriptions/46b25869-74ec-4ab6-a07d-dc540b41b4f8/resourceGroups/dxcrg231/pr...
Oracle Live SQL - S...
Dashboard | EduNxt
https://cloudlabs.n...
Dashboard | Nuvepro
Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1205_1654530...
MANIPAL PRO LEARN (MANIPAL...
Home > dxcsqldb231 (sqlserver231/dxcsqldb231)
dxcsqldb231 (sqlserver231/dxcsqldb231) | Query editor (preview)
SQL database
Search (Ctrl+/)
Login
New Query
Open query
Feedback
Overview
Activity log
Tags
Diagnose and solve problems
Getting started
Query editor (preview)
Power Platform
Power BI
Power Apps
Power Automate
Settings
Compute + storage
dxcsqldb231 (aishu)
Showing limited object explorer here. For full capability please open SSDT.
Tables
Views
Stored Procedures
Query 1
Run
Cancel query
Save query
Export data as
Show only Editor
1 4AR(55) NOT NULL, address_locality VARCHAR(55), address_city VARCHAR(55) NOT NULL,
2 R(55) NOT NULL, CONSTRAINT PK_email_address PRIMARY KEY (email_address_id));
3) NULL, person_contacted_number INTEGER NOT NULL, person_date_last_contacted DATETI
4 ddress_address_id INTEGER NOT NULL, CONSTRAINT PK_person_address PRIMARY KEY (perso
5 HRCHAR(55) NOT NULL, CONSTRAINT PK_phone_number PRIMARY KEY (phone_number_id));
Results
Messages
Search to filter items...
Ready
36°C
Sunny
ENG
IN
17:28
07-06-2022


Subscription Details | Nuvepro
dxcsqldb231 (sqlserver231/dxcsql
portal.azure.com/#@manipalazure.onmicrosoft.com/resource/subscriptions/46b25869-74ec-4ab6-a07d-dc540b41b4f8/resourceGroups/dxcrg231/pr...
Oracle Live SQL - S...
Dashboard | EduNxt
https://cloudlabs.n...
Dashboard | Nuvepro
Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1205_1654530...
MANIPAL PRO LEARN (MANIPAL...
Home > dxcsqldb231 (sqlserver231/dxcsqldb231)
dxcsqldb231 (sqlserver231/dxcsqldb231) | Query editor (preview)
SQL database
Search (Ctrl+/)
Login
New Query
Open query
Feedback
Overview
Activity log
Tags
Diagnose and solve problems
Getting started
Query editor (preview)
Power Platform
Power BI
Power Apps
Power Automate
Settings
Compute + storage
dxcsqldb231 (aishu)
Showing limited object explorer here. For full capability please open SSDT.
Tables
dbo.address
address_id (PK, int, not null)
address_building_number (varchar, not null)
address_street (varchar, not null)
address_locality (varchar, null)
address_city (varchar, not null)
address_zip_postal (varchar, not
Query 2
Run
Cancel query
Save query
Export data as
Show only Editor
1 select * from address;
Results
Messages
Search to filter items...
address_id
address_building_n...
address_street
address_locality
1
555
azure203Demo
Los An
Query succeeded | 4s
36°C
Sunny
ENG
IN
17:32
07-06-2022