TRAINING ASSIGNMENT

(7th JUNE 2022)

NAME – ROHIT ARORA ROLLNO – DXC-262AB-1209

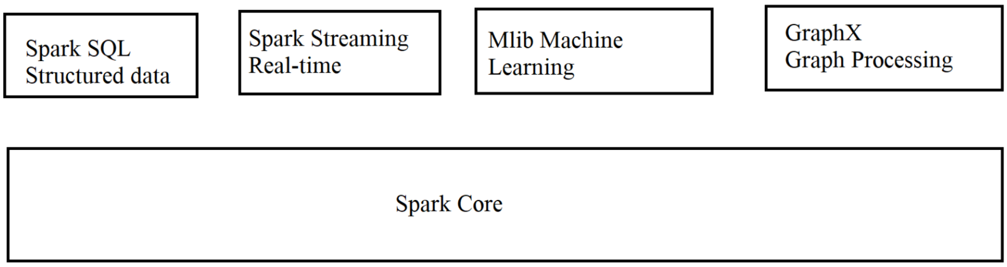
BATCH – DXC-262-ANALYTICS-B12-AZURE COMPANY – DXC TECHNOLOGY

EMPLOYEE DOMAIN – AZURE ANALYTICS 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??

Components of Apache Spark:



Spark Core - It is the base engine for large scale distributed and parallel data processing in Spark. It is responsible for memory management , fault recovery , job scheduling to different clusters and interacting with storage systems.

Spark SQL - Spark SQL is embedded with RDD , resilient distributed database with fault tolerance and is immutable with parallel processing features.

Spark Streaming - Spark streaming provides real time data processing because of a lightning fast clustering algorithm.

Spark Mlib - It provides and executes most of all the machine learning algorithms like clustering , classification and regression.

Spark GraphX - It provides all the data analysis features like ETL , EDA , reporting using graphs.

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

All the functionalities being provided by Apache Spark are built on the highest of the Spark Core. It delivers speed by providing in-memory computation capability. Spark Core is the foundation of parallel and distributed processing of giant dataset. It is the main backbone of the essential I/O functionalities and significant in programming and observing the role of the spark cluster. It holds all the components related to scheduling, distributing and monitoring jobs on a cluster, Task dispatching, Fault recovery. The functionalities of this component are:

· It contains the basic functionality of spark. (Task scheduling, memory management, fault recovery, interacting with storage systems).
· Home to API that defines RDDs.

Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. It is an immutable distributed collection of objects. Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster.

· TRANFORMATION – operations like map, filter performed on RDD yielding new RDD. Eg- val(x)=sc.textfile -àval y=x.map()

· ACTION – Operations such as reduce,first,count that returns a value after running a computation RDD. Eg- z.count()

3. Explain various Mlib algorithms Spark is supporting ?

Spark MLlib is used to perform machine learning in Apache Spark. MLlib consists of popular algorithms and utilities. MLlib in Spark is a scalable Machine learning library that discusses both high-quality algorithm and high speed. The machine learning algorithms like regression, classification, clustering, pattern mining, and collaborative filtering. Lower level machine learning primitives like generic gradient descent optimization algorithm are also present in MLlib.

The popular algorithms and utilities in Spark MLlib are:

· Basic Statistics.
· Regression.
· Classification.
· Recommendation System.
· Clustering.
· Dimensionality Reduction.
· Feature Extraction.
· Optimization.

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

In reference to Spark SQL, it is a module of Apache Spark that analyses the structured data. It provides Scalability, it ensures high compatibility of the system. It has standard connectivity through JDBC or ODBC. Thus, it provides the most natural way to express the Structured Data. It provides a programming abstraction called DataFrames and can also act as a distributed SQL query engine. It enables unmodified Hadoop Hive queries to run up to 100x faster on existing deployments and data.

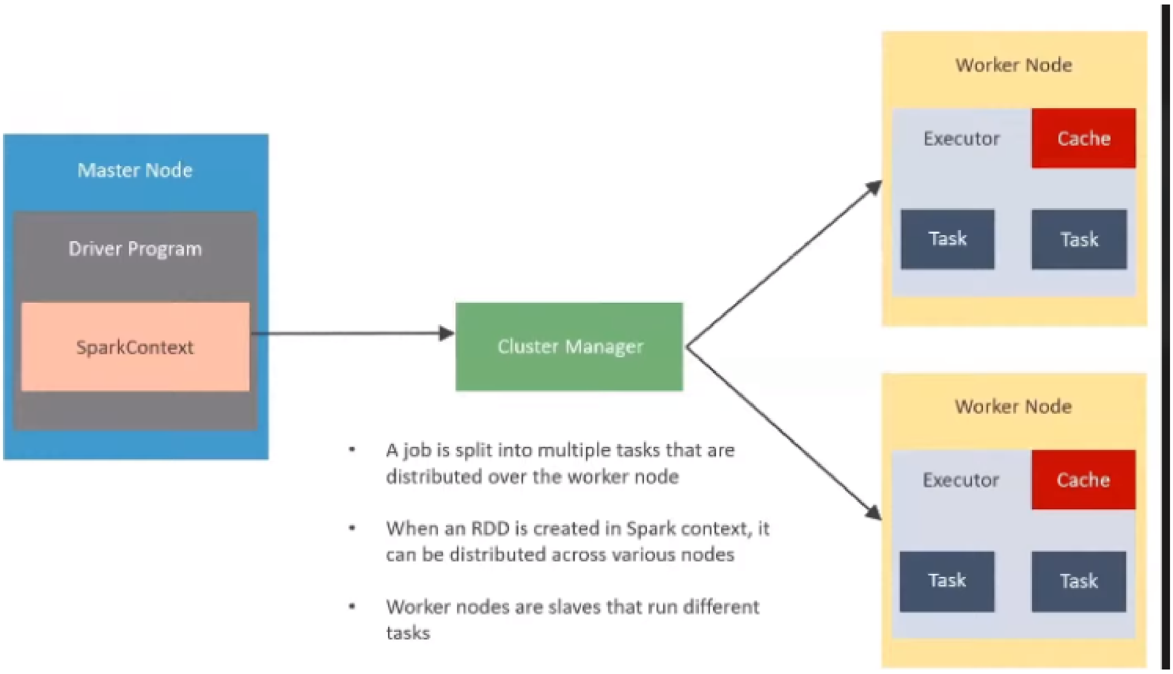
There are multiple methods to insert relational data in spark. For that first we need to open Google Collabs. Import PySpark. Then type the following commands and press Shift+Enter

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.
INPUT DATA STREAM àSPARK STREAMING àBATCH OF INPUT DATA àSPARK ENGINE à BATCHES OF PROCESSED DATA .

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

> The master slave architecture is there is a master node and in that master node all the worker nodes work



- A job is split into multiple tasks that are distributed over the worker node
- When an RDD is created in Spark context, it can be distributed across various nodes
- Worker nodes are slaves that run different tasks

7. Explain various cluster managers in SPARK?

Apache Standalone - Application runs on FIFO mode when run in standalone and consume all resources available.

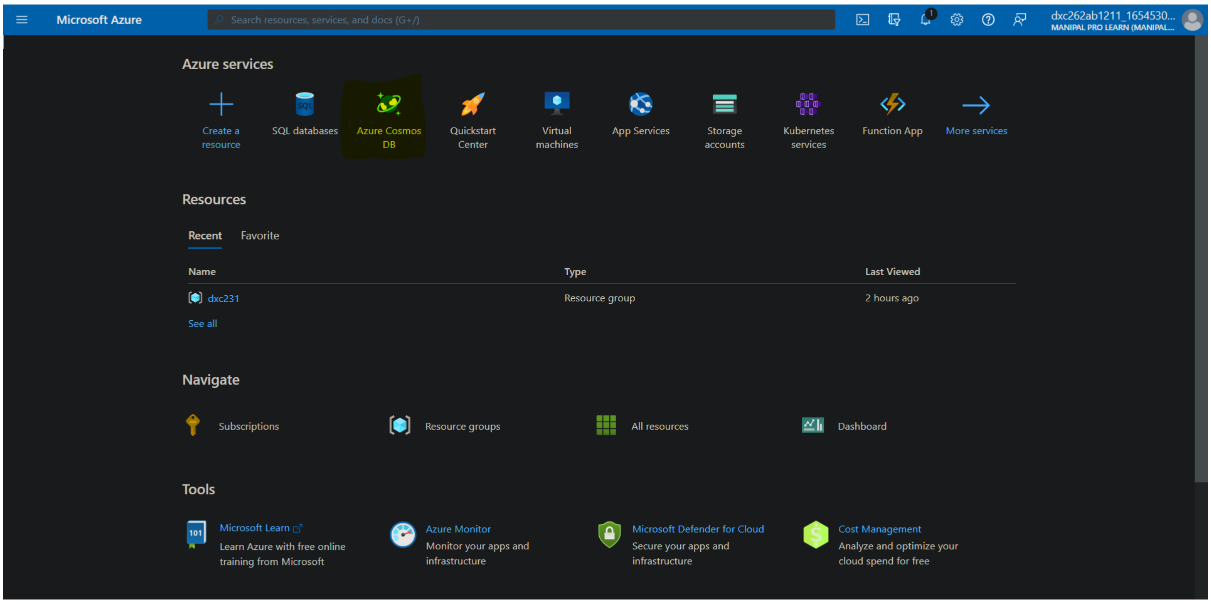
Apache Mesos - is an open source project to manage computer clusters

Apache Yarn - It is a cluster manager in hadoop 2 , spark can run on yarn.

Kubernetes - It is used for automating deployment , scaling and managing the containerized applications.

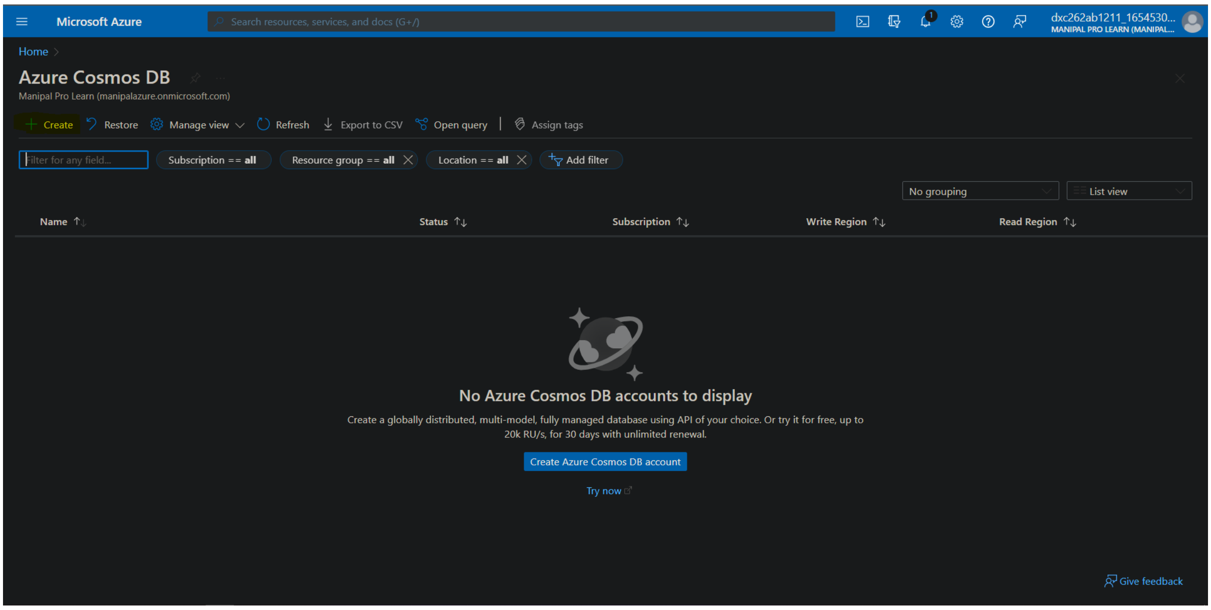
8. Explain with sceenshots & steps how to create Cosmos DB ?

To create Cosmos DB firstly we need to login into the azure cloud platform after that we need to select Azure Cosmos DB from resources



Name	Type	Last Viewed
dxc231	Resource group	2 hours ago

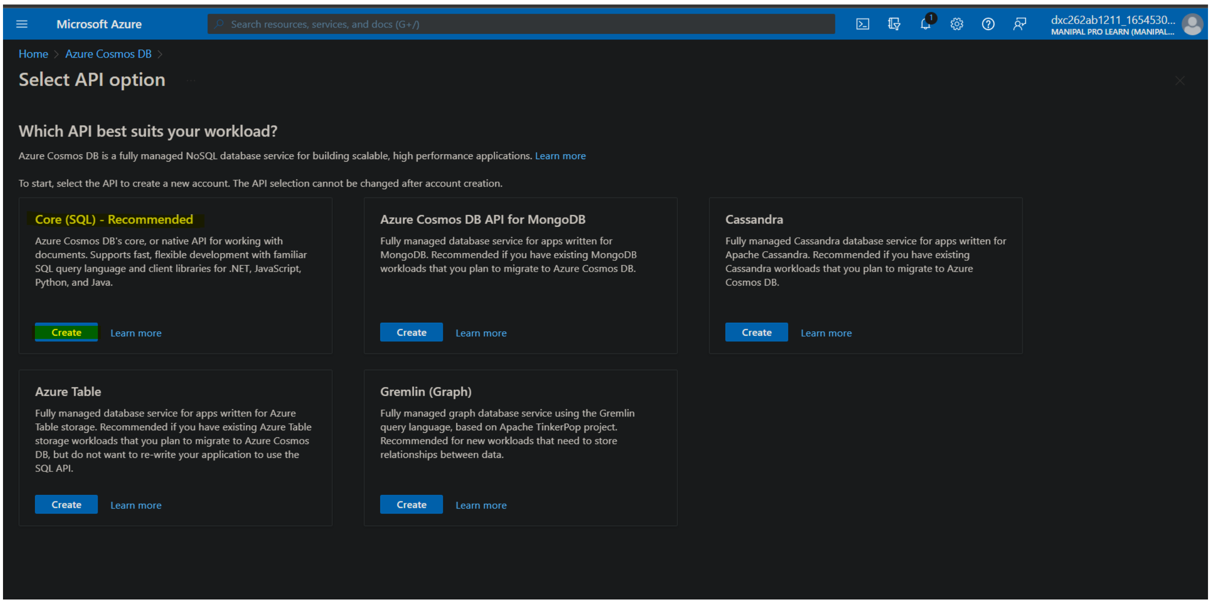
After selecting that we need to click on +Create in order to create a new CDB database



No Azure Cosmos DB accounts to display

Create a globally distributed, multi-model, fully managed database using API of your choice. Or try it for free, up to 20k RU/s, for 30 days with unlimited renewal.

After we will click on the kind of database you want to create here we will create SQL supported Document database


Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1211_1654530...
MANIPAL PRO LEARN (MANIPAL...
Home > Azure Cosmos DB >

Select API option

Which API best suits your workload?

Azure Cosmos DB is a fully managed NoSQL database service for building scalable, high performance applications. Learn more

To start, select the API to create a new account. The API selection cannot be changed after account creation.

Core (SQL) - Recommended

Azure Cosmos DB's core, or native API for working with documents. Supports fast, flexible development with familiar SQL query language and client libraries for .NET, JavaScript, Python, and Java.

Azure Cosmos DB API for MongoDB

Fully managed database service for apps written for MongoDB. Recommended if you have existing MongoDB workloads that you plan to migrate to Azure Cosmos DB.

Cassandra

Fully managed Cassandra database service for apps written for Apache Cassandra. Recommended if you have existing Cassandra workloads that you plan to migrate to Azure Cosmos DB.

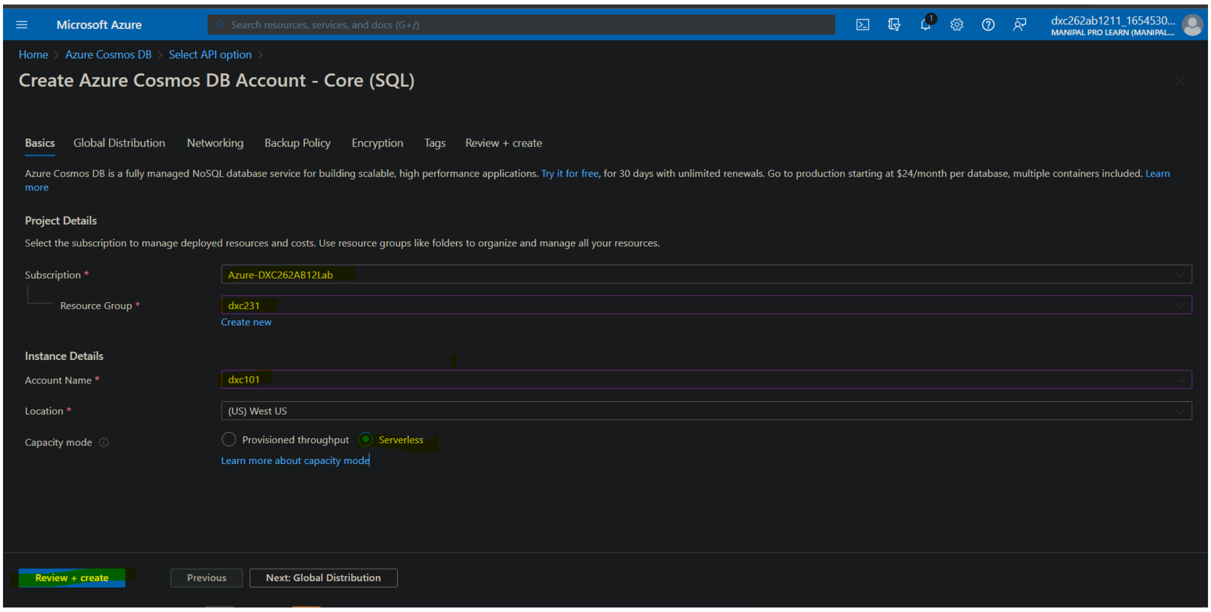
Azure Table

Fully managed database service for apps written for Azure Table storage. Recommended if you have existing Azure Table storage workloads that you plan to migrate to Azure Cosmos DB, but do not want to re-write your application to use the SQL API.

Gremlin (Graph)

Fully managed graph database service using the Gremlin query language, based on Apache TinkerPop project. Recommended for new workloads that need to store relationships between data.

After that we need to enter the details regarding the database


Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1211_1654530...
MANIPAL PRO LEARN (MANIPAL...
Home > Azure Cosmos DB > Select API option >

Create Azure Cosmos DB Account - Core (SQL)

Basics
Global Distribution
Networking
Backup Policy
Encryption
Tags
Review + create

Azure Cosmos DB is a fully managed NoSQL database service for building scalable, high performance applications. Try it for free, for 30 days with unlimited renewals. Go to production starting at $24/month per database, multiple containers included. Learn more

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 * dxc231

Create new

Instance Details

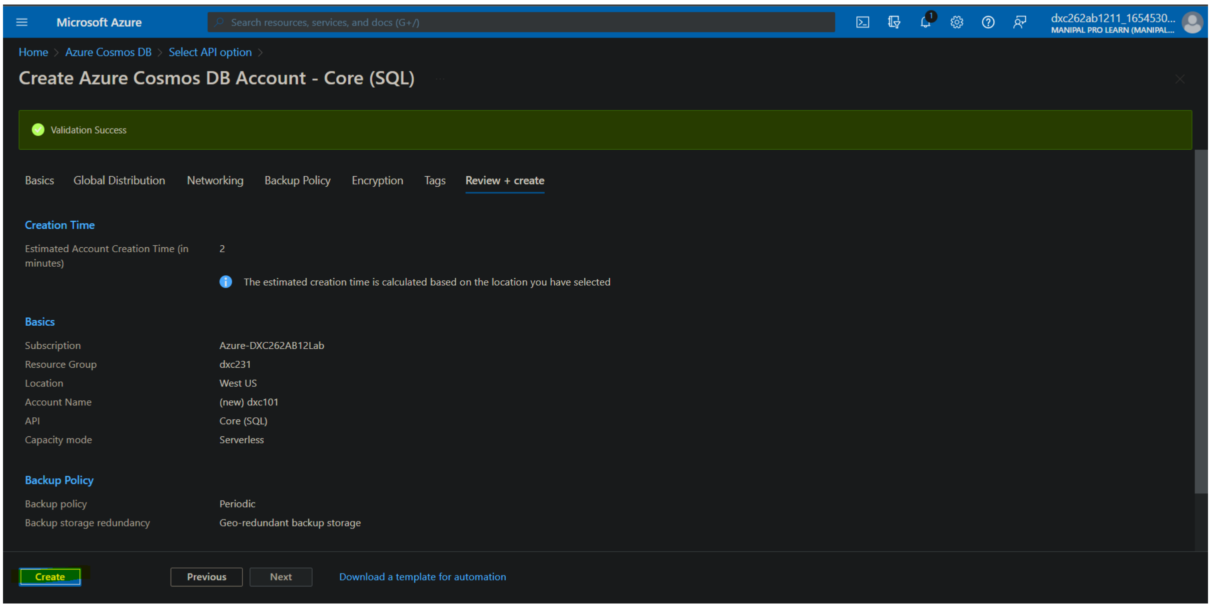
Account Name * dxc101

Location * (US) West US

Capacity mode ○ Provisioned throughput ● Serverless

Learn more about capacity mode

After that validate and review the details and click of create after validating


Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1211_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

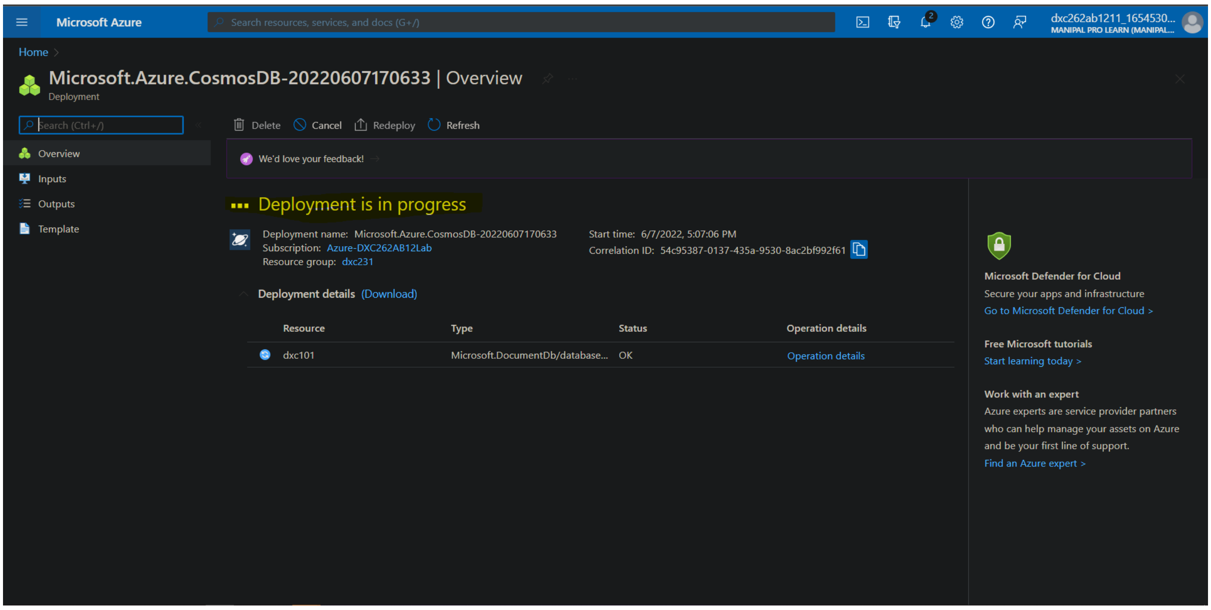
Basics

Subscription	Azure-DXC262AB12Lab
Resource Group	dxc231
Location	West US
Account Name	(new) dxc101
API	Core (SQL)
Capacity mode	Serverless

Backup Policy

Backup policy	Periodic
Backup storage redundancy	Geo-redundant backup storage

Through this we have created a Cosmos DB database with name as dxc101



Microsoft.Azure.CosmosDB-20220607170633 | Overview

Deployment

We'd love your feedback!

Deployment is in progress

Deployment name: Microsoft.Azure.CosmosDB-20220607170633
Subscription: Azure-DXC262AB12Lab
Resource group: dxc231

Start time: 6/7/2022, 5:07:06 PM
Correlation ID: 54c95387-0137-435a-9530-8ac2bf992f61

Deployment details (Download)

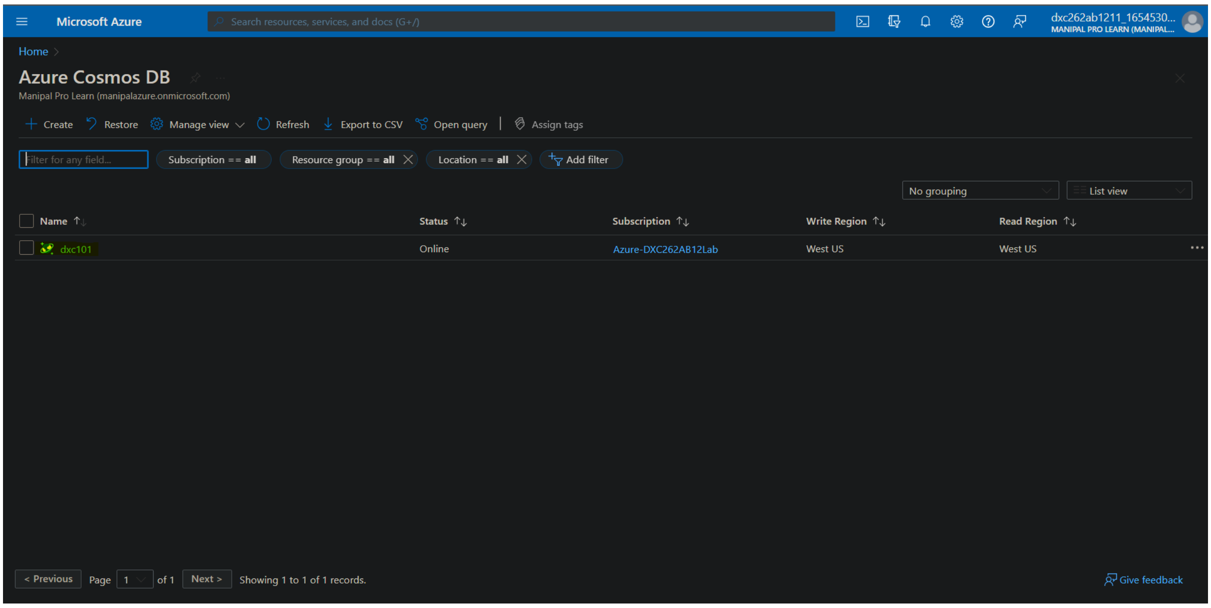
Resource	Type	Status	Operation details
dxc101	Microsoft.DocumentDb/database...	OK	Operation details

Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >
Free Microsoft tutorials
Start learning today >
Work with an expert
Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support.
Find an Azure expert >

After this our Database will be ready

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

To insert data into the database of cosmos DB , open the azure cloud platform and navigate towards the Azure cosmos DB and click on the database you want to work with



Azure Cosmos DB

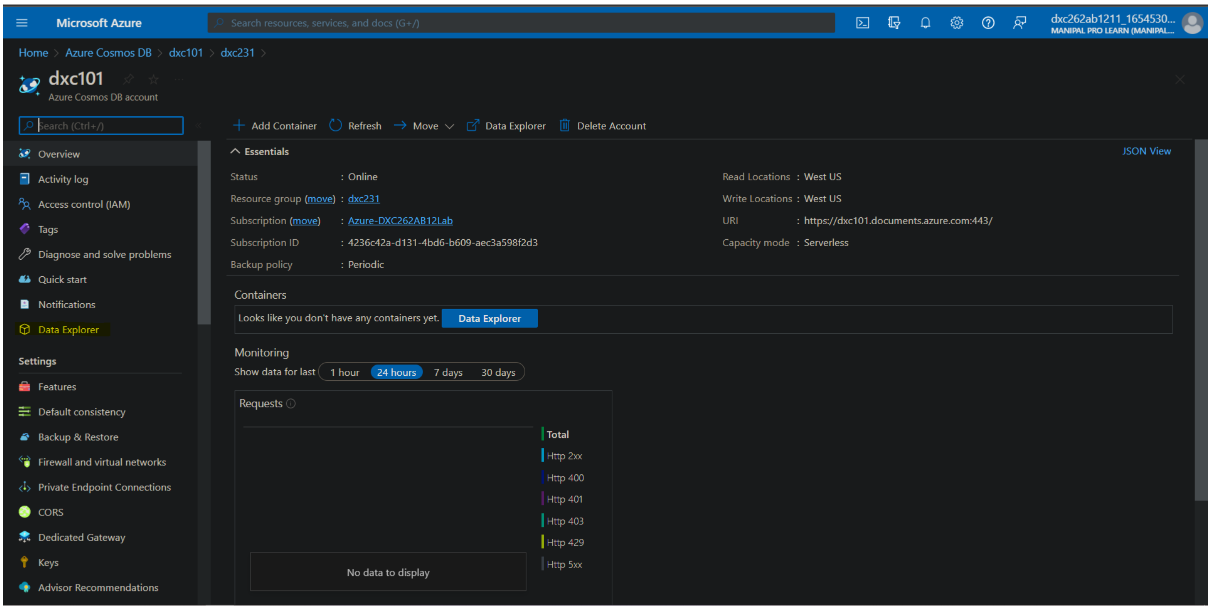
Manipal Pro Learn (manipalazure.onmicrosoft.com)

Subscription == all | Resource group == all | Location == all

Name	Status	Subscription	Write Region	Read Region
dxc101	Online	Azure-DXC262AB12Lab	West US	West US

Page 1 of 1 Showing 1 to 1 of 1 records.

After choosing the database click on data explorer



dxc101

Azure Cosmos DB account

Essentials

Status	: Online	Read Locations	: West US
Resource group (move)	: dxc231	Write Locations	: West US
Subscription (move)	: Azure-DXC262AB12Lab	URI	: https://dxc101.documents.azure.com:443/
Subscription ID	: 4236c42a-d131-4bd6-b609-aec3a598f2d3	Capacity mode	: Serverless
Backup policy	: Periodic		

Containers

Looks like you don't have any containers yet. Data Explorer

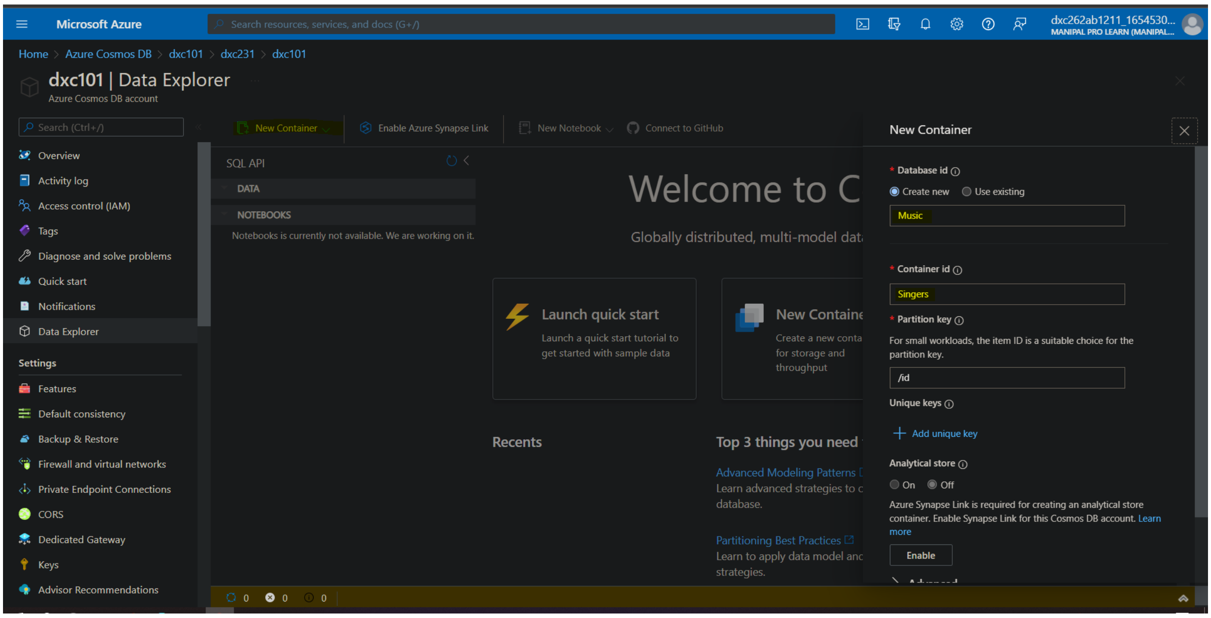
Monitoring

Show data for last 1 hour 24 hours 7 days 30 days

Requests

No data to display

After clicking on that click on the new container and enter the database you want to create and the tables in that



Notebooks is currently not available. We are working on it.

Launch a quick start tutorial to get started with sample data

New Containe
Create a new conta for storage and throughput

Advanced Modeling Patterns
Learn advanced strategies to c database.

Partitioning Best Practices
Learn to apply data model anc strategies.

New Container

* Database id

● Create new ○ Use existing

Music

* Container id

Singers

* Partition key

For small workloads, the item ID is a suitable choice for the partition key.

/id

Unique keys

+ Add unique key

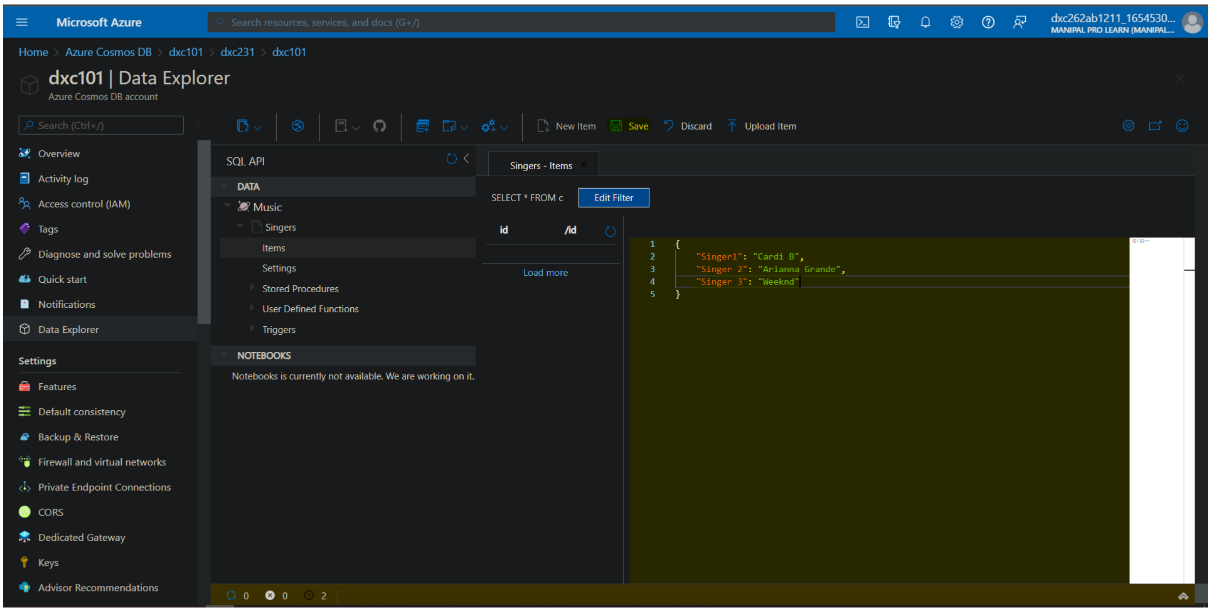
Analytical store

○ On ● Off

Azure Synapse Link is required for creating an analytical store container. Enable Synapse Link for this Cosmos DB account. Learn more

Enable

After the tables are created you need to navigate to the table and click on + new item

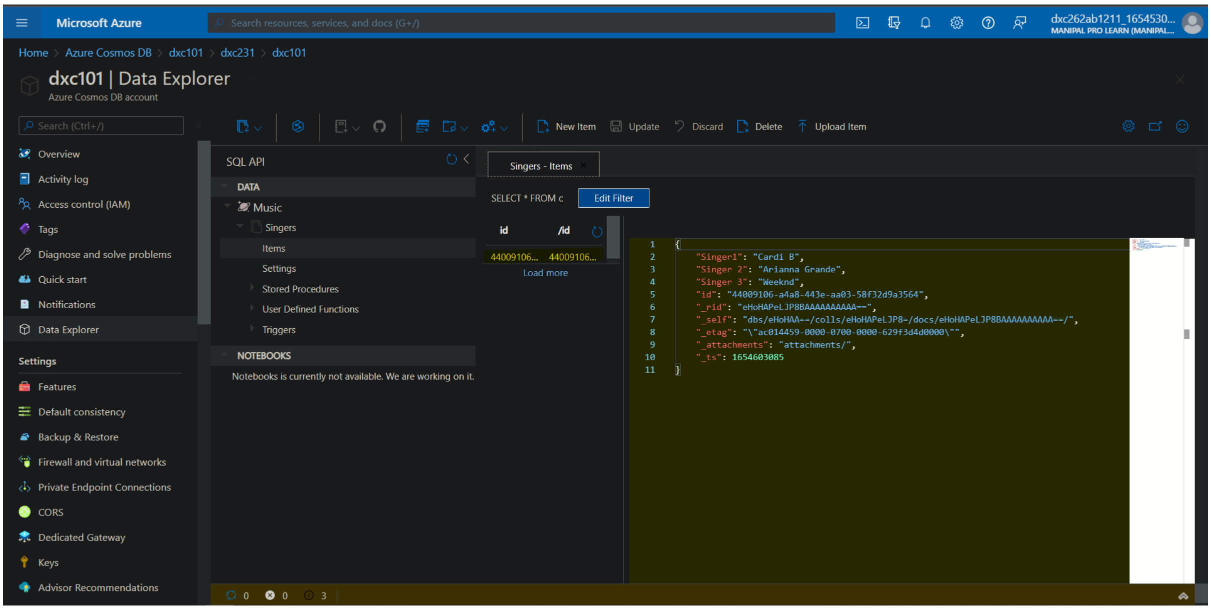


Notebooks is currently not available. We are working on it.

SELECT * FROM c

```
{
    "Singer1": "Cardi B",
    "Singer 2": "Arianna Grande",
    "Singer 3": "Weeknd"
}
```

Add the data you want to insert and click on Save



Notebooks is currently not available. We are working on it.

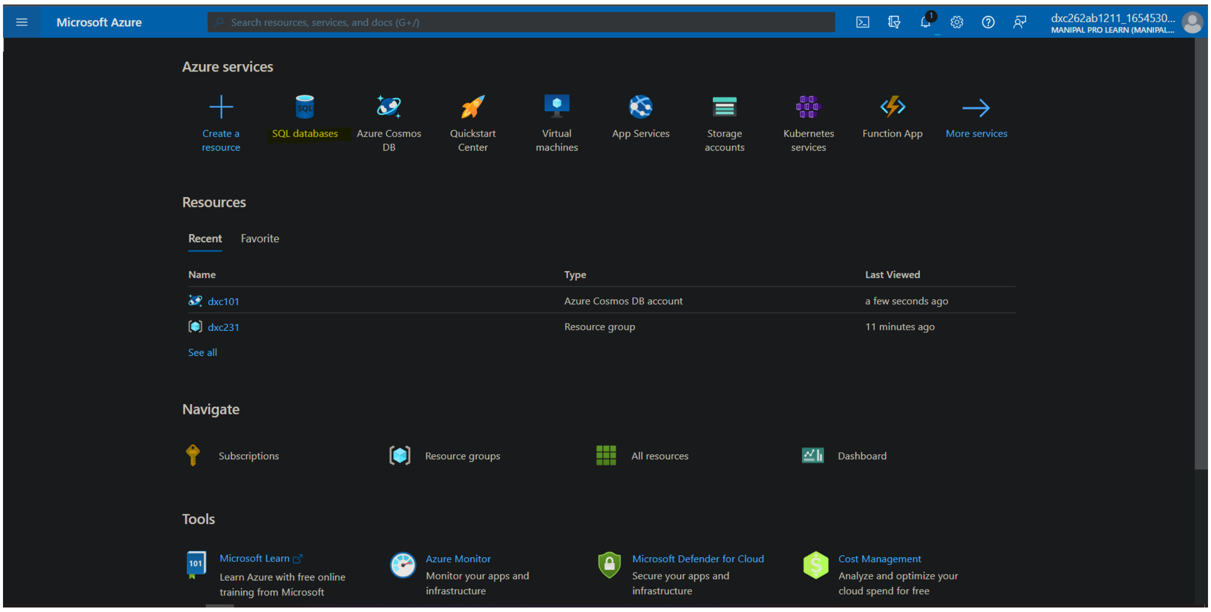
SELECT * FROM c

```
{
    "Singer1": "Cardi B",
    "Singer 2": "Arianna Grande",
    "Singer 3": "Weeknd",
    "id": "44009106-a4a8-443e-aa03-58f32d9a3564",
    "_rid": "eHoHAPeLJP8BAAAAAAAAAA==",
    "_self": "dbs/eHoHAA==/colls/eHoHAPeLJP8=/docs/eHoHAPeLJP8BAAAAAAAAAA==/",
    "_etag": "\"ac014459-0000-0700-0000-629f3d4d0000\"",
    "_attachments": "attachments/",
    "_ts": 1654603085
}
```

Now through this you have successfully insert a data into Cosmos DB

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

To create a SQL database and insert data into it we need to first navigate to the SQL database on Azure Cloud Platform



Resources

Recent Favorite

Name	Type	Last Viewed
dxc101	Azure Cosmos DB account	a few seconds ago
dxc231	Resource group	11 minutes ago

See all

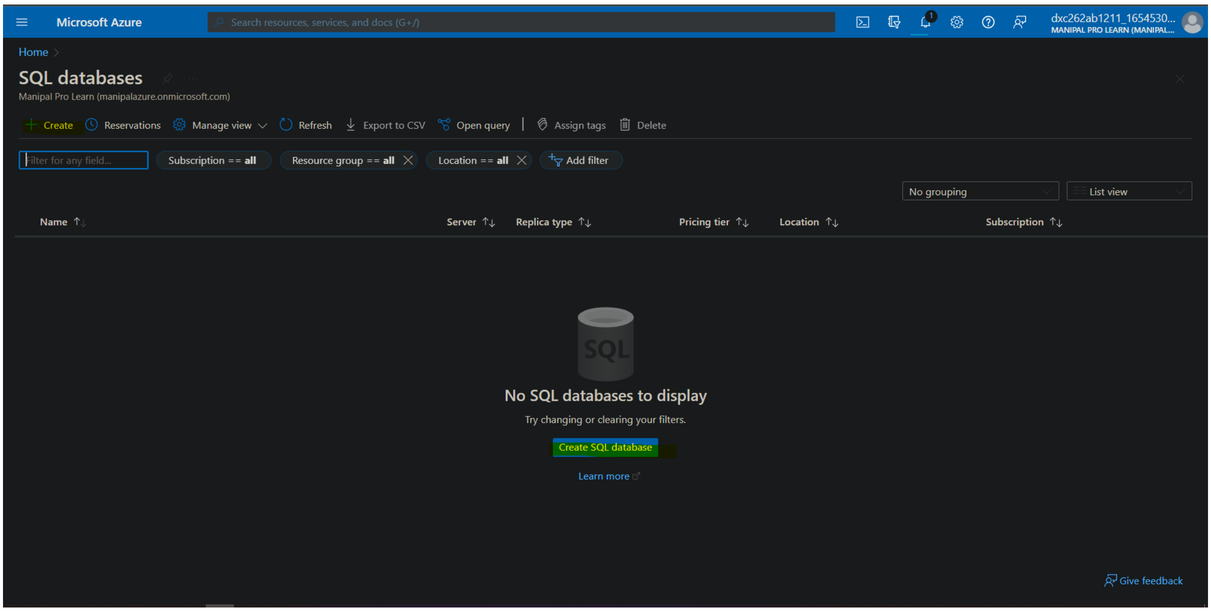
Microsoft Learn
Learn Azure with free online training from Microsoft

Azure Monitor
Monitor your apps and infrastructure

Microsoft Defender for Cloud
Secure your apps and infrastructure

Cost Management
Analyze and optimize your cloud spend for free

After you are into the SQL database you need to click + create



SQL databases

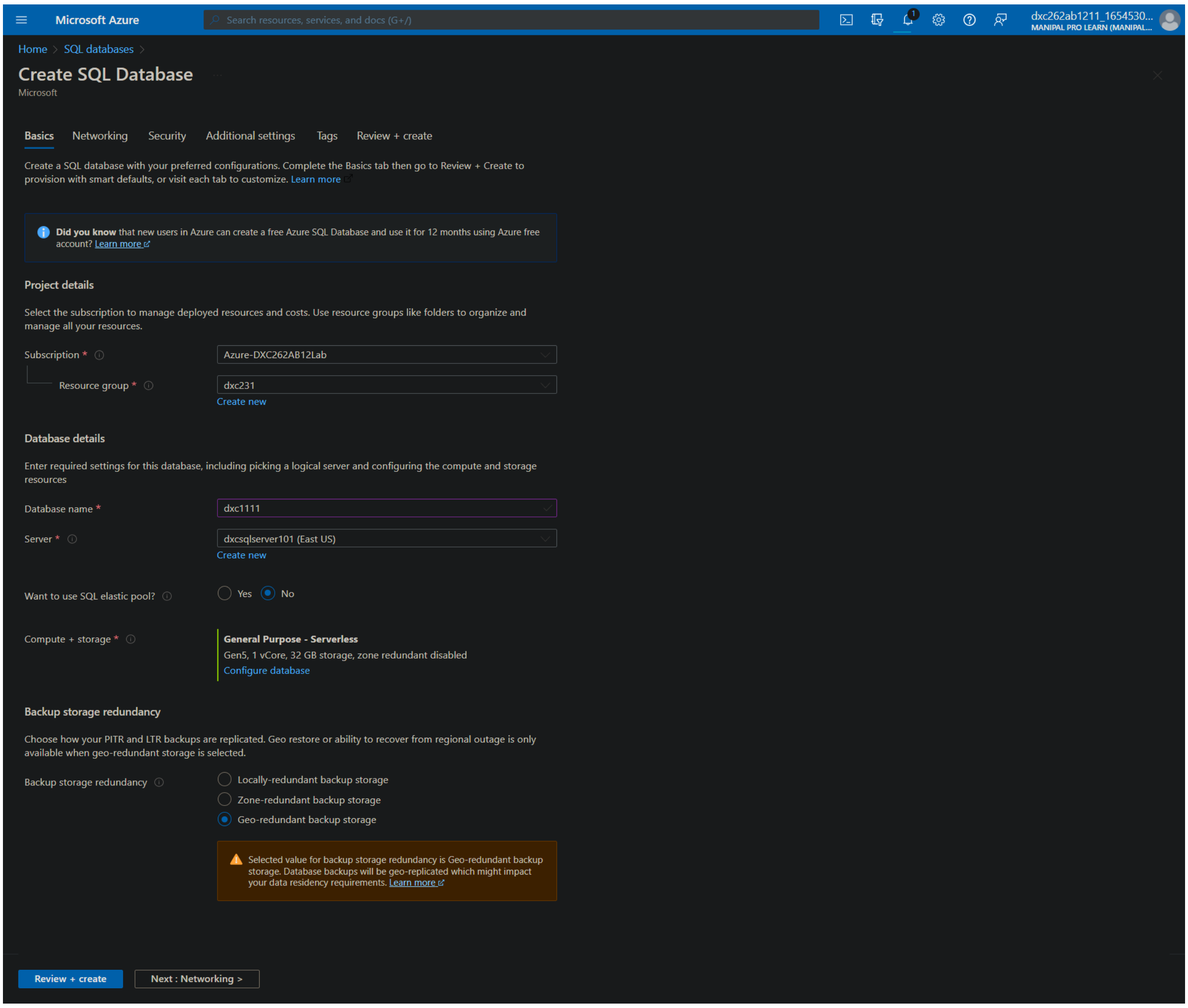
Manipal Pro Learn (manipalazure.onmicrosoft.com)

Filter for any field...   Subscription == **all**   Resource group == **all**   Location == **all**

No SQL databases to display

Try changing or clearing your filters.

After clicking on create enter the details regarding the resource group , server and database id , name and click on create



Create SQL Database

Microsoft

Create a SQL database with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults, or visit each tab to customize. Learn more

Did you know that new users in Azure can create a free Azure SQL Database and use it for 12 months using Azure free account? Learn more

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 * : dxc231

Database details

Enter required settings for this database, including picking a logical server and configuring the compute and storage resources

Database name * : dxc1111

Server * : dxcsqlserver101 (East US)

Want to use SQL elastic pool? ○ Yes ● No

Compute + storage * : **General Purpose - Serverless**
Gen5, 1 vCore, 32 GB storage, zone redundant disabled

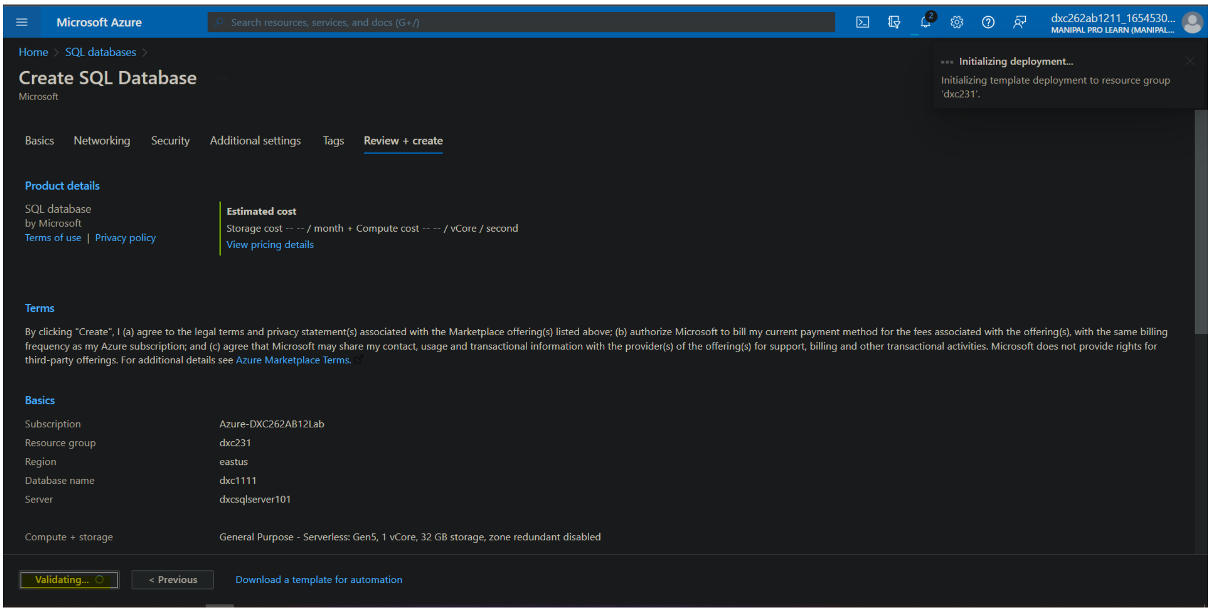
Backup storage redundancy

Choose how your PITR and LTR backups are replicated. Geo restore or ability to recover from regional outage is only available when geo-redundant storage is selected.

Backup storage redundancy
○ Locally-redundant backup storage
○ Zone-redundant backup storage
● Geo-redundant backup storage

Selected value for backup storage redundancy is Geo-redundant backup storage. Database backups will be geo-replicated which might impact your data residency requirements. Learn more

After that you need the validate the data that you have entered proceed and click create after validating



Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1211_1654530...
MANIPAL PRO LEARN (MANIPAL...
Home > SQL databases >

Initializing deployment...

Initializing template deployment to resource group 'dxc231'.

Create SQL Database

Microsoft

Product details

SQL database
by Microsoft
Terms of use | Privacy policy

Estimated cost
Storage cost -- -- / month + Compute cost -- -- / vCore / second
View pricing details

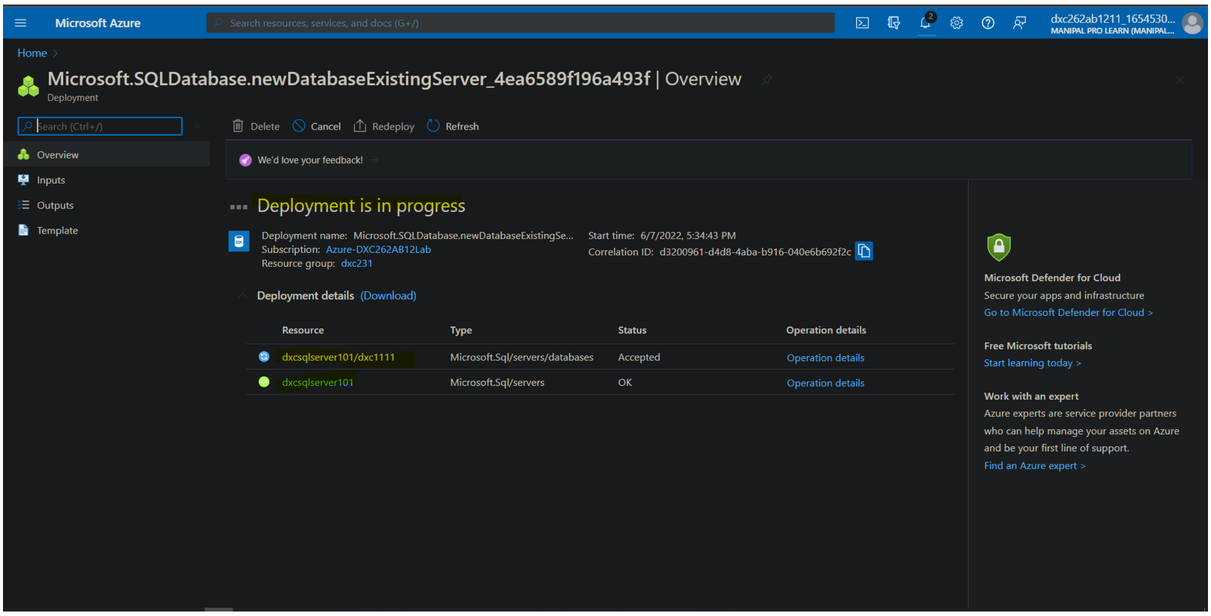
Terms

By clicking "Create", I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. For additional details see Azure Marketplace Terms.

Basics

Subscription	Azure-DXC262AB12Lab
Resource group	dxc231
Region	eastus
Database name	dxc1111
Server	dxcsqlserver101
Compute + storage	General Purpose - Serverless: Gen5, 1 vCore, 32 GB storage, zone redundant disabled

It wil; take some time to deploy the database



Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1211_1654530...
MANIPAL PRO LEARN (MANIPAL...
Home >

Microsoft.SQLDatabase.newDatabaseExistingServer_4ea6589f196a493f | Overview

Deployment

Deployment is in progress

Deployment name: Microsoft.SQLDatabase.newDatabaseExistingSe...
Subscription: Azure-DXC262AB12Lab
Resource group: dxc231

Start time: 6/7/2022, 5:34:43 PM
Correlation ID: d3200961-d4d8-4aba-b916-040e6b692f2c

Deployment details (Download)

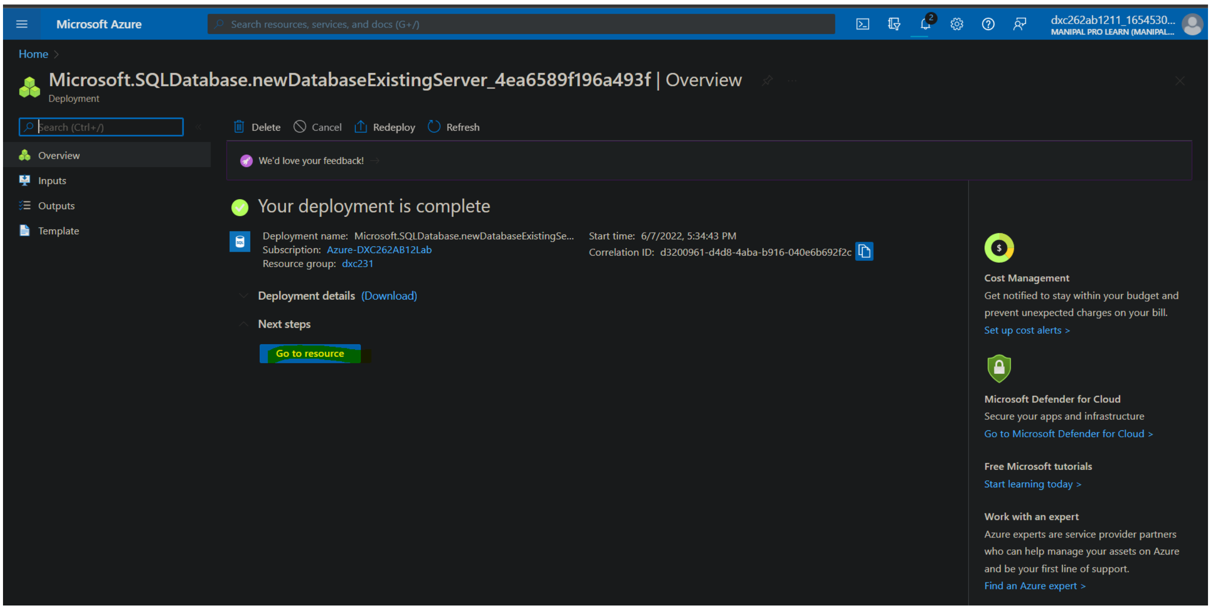
Resource	Type	Status	Operation details
dxcsqlserver101/dxc1111	Microsoft.Sql/servers/databases	Accepted	Operation details
dxcsqlserver101	Microsoft.Sql/servers	OK	Operation details

Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >

Free Microsoft tutorials
Start learning today >

Work with an expert
Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support.
Find an Azure expert >

After the deployment is done click on Go to resource



Microsoft.SQLDatabase.newDatabaseExistingServer_4ea6589f196a493f | Overview

Deployment

We'd love your feedback! →

Your deployment is complete

Deployment name: Microsoft.SQLDatabase.newDatabaseExistingSe...
Subscription: Azure-DXC262AB12Lab
Resource group: dxc231

Start time: 6/7/2022, 5:34:43 PM
Correlation ID: d3200961-d4d8-4aba-b916-040e6b692f2c

Deployment details (Download)

Next steps

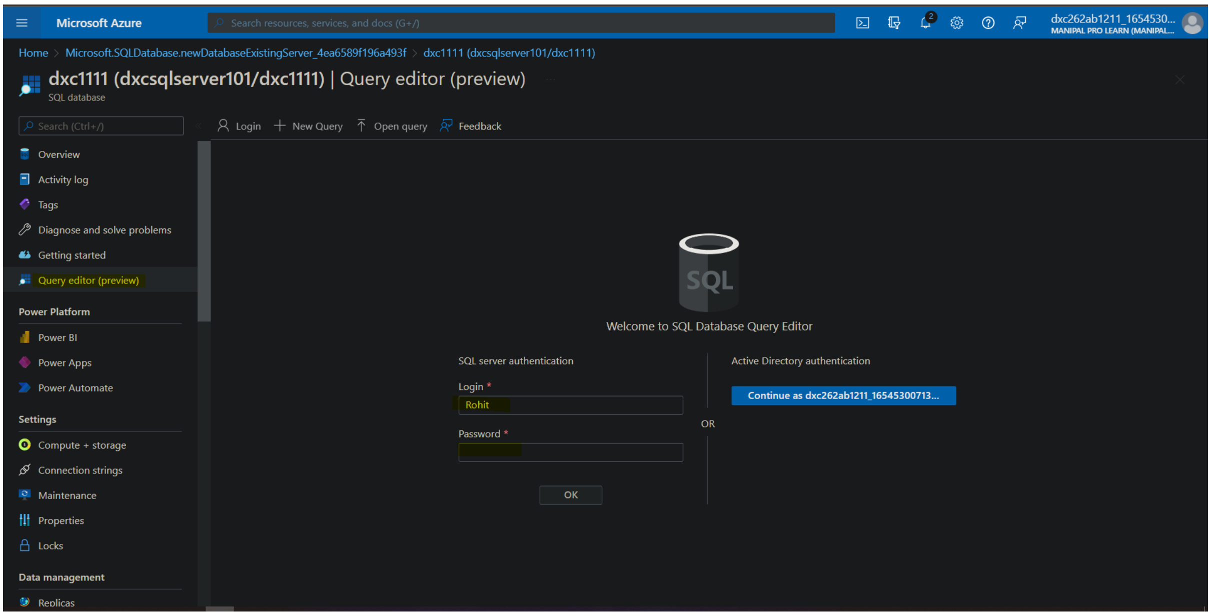
Cost Management
Get notified to stay within your budget and prevent unexpected charges on your bill.
Set up cost alerts >

Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >

Free Microsoft tutorials
Start learning today >

Work with an expert
Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support.
Find an Azure expert >

Proceed to Query Editor



dxc1111 (dxcsqlserver101/dxc1111) | Query editor (preview)

SQL database

Welcome to SQL Database Query Editor

SQL server authentication

Login *
Rohit

Password *

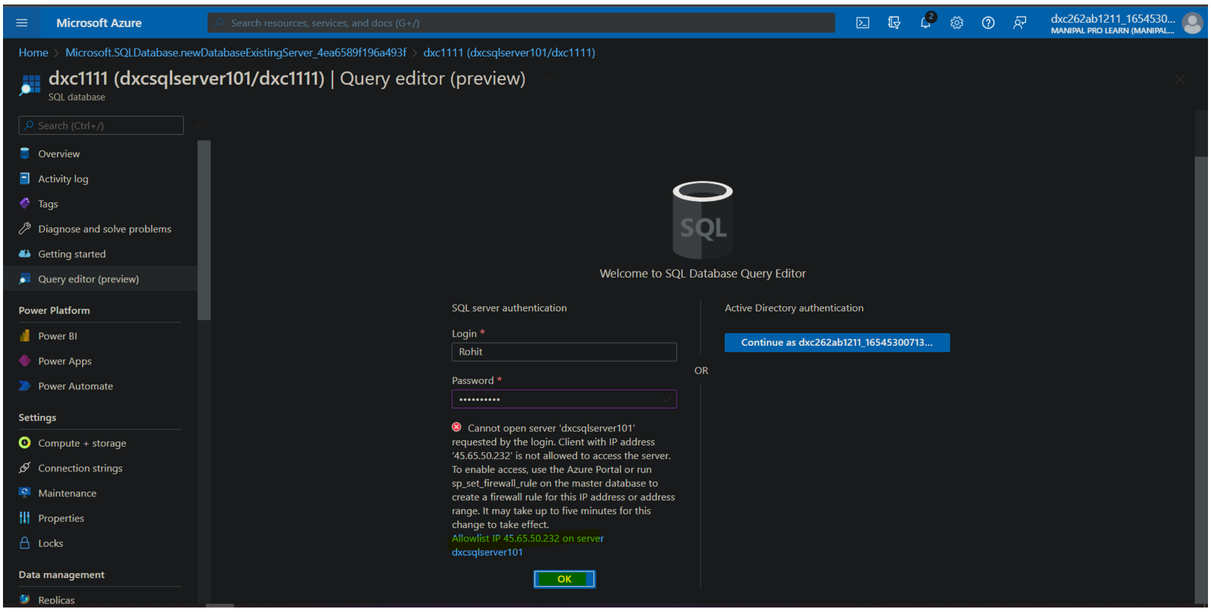
OK

OR

Active Directory authentication

Continue as dxc262ab1211_16545300713...

Type in your server name and password

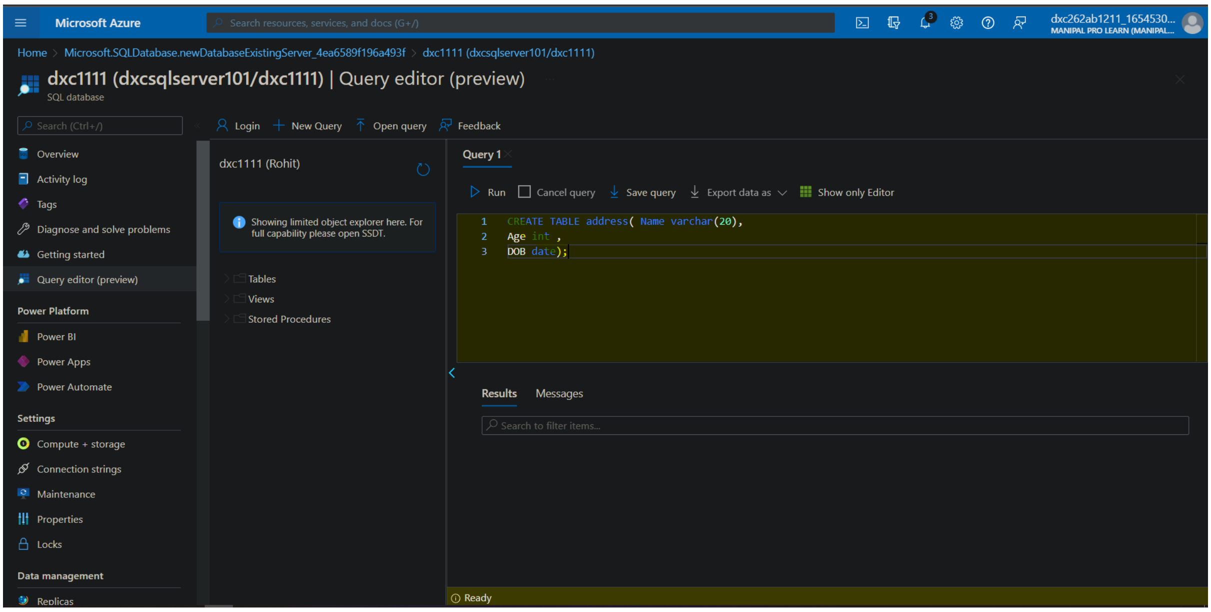

Microsoft Azure
Home > Microsoft.SQLDatabase.newDatabaseExistingServer_4ea6589f196a493f > dxc1111 (dxcsqlserver101/dxc1111)
dxc1111 (dxcsqlserver101/dxc1111) | Query editor (preview)
SQL database
Overview
Activity log
Tags
Diagnose and solve problems
Getting started
Query editor (preview)
Power Platform
Power BI
Power Apps
Power Automate
Settings
Compute + storage
Connection strings
Maintenance
Properties
Locks
Data management
Replicas
Welcome to SQL Database Query Editor
SQL server authentication
Active Directory authentication
Continue as dxc262ab1211_16545300713...
OR
OK

Login *
Rohit

Password *
••••••••••

Cannot open server 'dxcsqlserver101' requested by the login. Client with IP address '45.65.50.232' is not allowed to access the server. To enable access, use the Azure Portal or run sp_set_firewall_rule on the master database to create a firewall rule for this IP address or address range. It may take up to five minutes for this change to take effect.
Allowlist IP 45.65.50.232 on server dxcsqlserver101

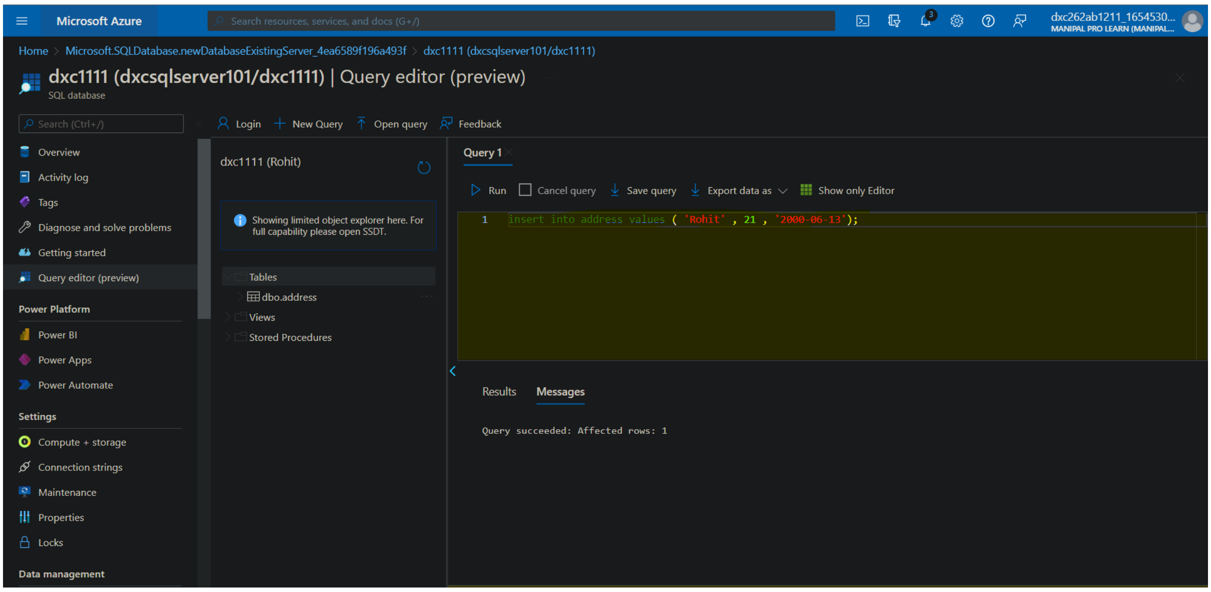
Allow the access for the IP and click on OK and you will now see a screen where you can now create table and insert data into it. Create a sample table


Microsoft Azure
Home > Microsoft.SQLDatabase.newDatabaseExistingServer_4ea6589f196a493f > dxc1111 (dxcsqlserver101/dxc1111)
dxc1111 (dxcsqlserver101/dxc1111) | Query editor (preview)
SQL database
Login
New Query
Open query
Feedback
dxc1111 (Rohit)
Tables
Views
Stored Procedures
Query 1
Run
Cancel query
Save query
Export data as
Show only Editor
Results
Messages
Search to filter items...
Ready

Showing limited object explorer here. For full capability please open SSDT.

```
CREATE TABLE address( Name varchar(20),
Age int ,
DOB date);
```

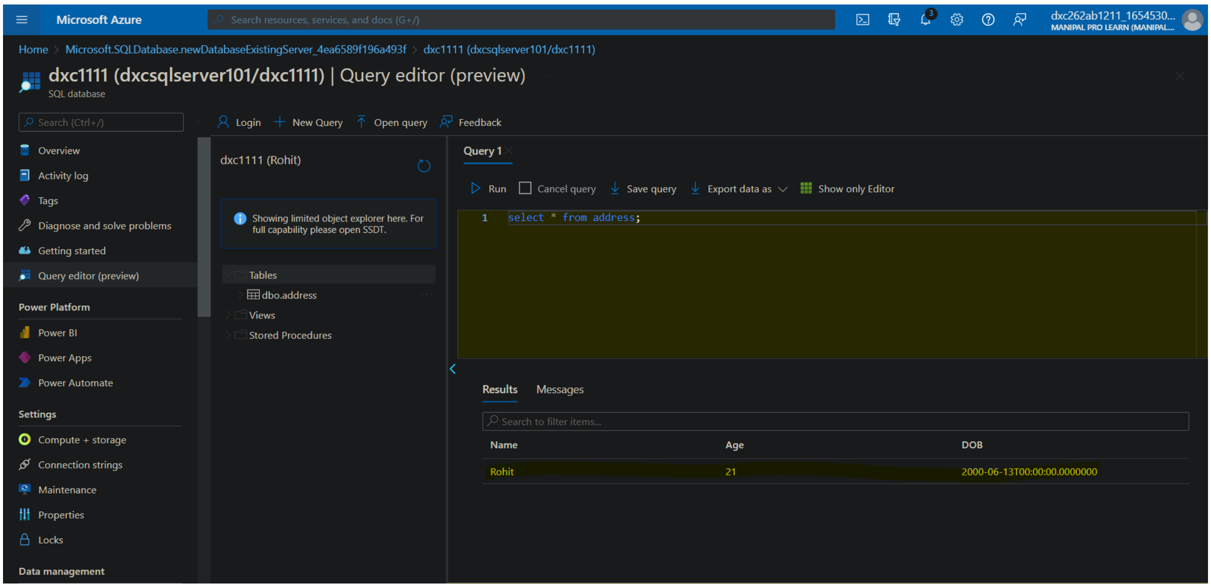
Click on run and insert a sample data in to the table address



```
insert into address values ( 'Rohit' , 21 , '2000-06-13');
```

```
Query succeeded: Affected rows: 1
```

So now the data is inserted we can view the inserted data using select command



```
select * from address;
```

Name	Age	DOB
Rohit	21	2000-06-13T00:00:00.0000000

Tada! We have successfully created and inserted data into the Azure Cloud platform SQL database