NAME: SHAIK ABDUL KHADAR JILANI ROLLNO: DXC262AB12038

BATCH: DXC-262-Analytics-B12-Azure SUBMISSION: 7-6-2022

COMPANY: DXC TECHNOLOGY

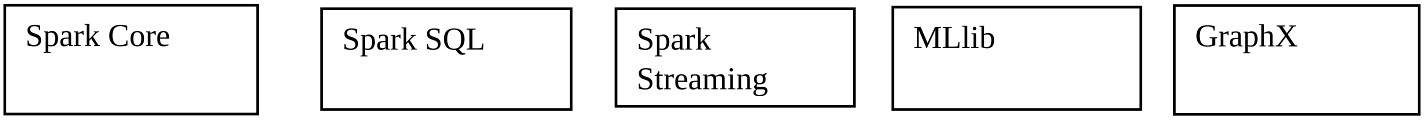
DAY 7

1) Explain what are various components of SPARK with block diagram?

explain functionality of every components?

Components of Apache Spark

GraphX



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

1)Memory management

2)Fault recovery

3)scheduling distributing and monitoring jobs on a cluster

4)Interacting with storage system

SPARK SQL

-Spark SQL framework component is used for structured and semi-structured data processing

Spark streaming

-It is a lightweight API that allows developers to perform batch processing and real-time streaming of data with ease

-provides secure, reliable and fast processing of live data streams

MLlib

-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

GraphX

-GraphX is Spark's own Graph Computation Engine and data store

-Provides a uniform tool for ETL

-Exploratory data analysis

-Interactive graph computations

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

SPARK CORE

-It is the base engine for large-scale parallel and distributed data processing

-It is responsible for

1)Memory management

2)Fault recovery

3)scheduling distributing and monitoring jobs on a cluster

4)Interacting with storage system

RDD

Spark Core is embedded with RDDs(Resilient Distributed Dataset) an immutable

fault-tolerant, distributed collection of objects that can be operated on in parallel

RDD

1)transformation

These are operations (such as map, filter, join, union) that are performed on an RDD that

yields a new RDD containing the result

E.g.,

val x=sc.textfile (RDD will be created)

val Y=x.map

val z=y.filter

2)Action

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

Eg: z.count() ,x.count()

3) Explain various MLlib algorithms Spark is supporting ?

MLlib

-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 benefits Spark SQL & how relational data will be inserted into SPARK?

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.

1) Integrated

Apache Spark SQL mixes SQL queries with Spark programs. With the help of Spark SQL, we can query structured data as a distributed dataset (RDD). We can run SQL queries alongside complex analytic algorithms using tight integration property of Spark SQL.

2) Unified Data Access

Using Spark SQL, we can load and query data from different sources. The Schema-RDDs lets single interface to productively work structured data. For example, Apache Hive tables, parquet files, and JSON files.

3) High compatibility

In Apache Spark SQL, we can run unmodified Hive queries on existing warehouses. It allows full compatibility with existing Hive data, queries and UDFs, by using the Hive fronted and MetaStore.

4) Standard Connectivity

It can connect through JDBC or ODBC. It includes server mode with industry standard JDBC and ODBC connectivity.

5) Scalability

To support mid-query fault tolerance and large jobs, it takes advantage of RDD model. It uses the same engine for interactive and long queries.

6) Performance Optimization

The query optimization engine in Spark SQL converts each SQL query to a logical plan. Further, it converts to many physical execution plans. Among the entire plan, it selects the most optimal physical plan for execution. Read more about Apache Spark performance tuning techniques in detail.

7) For batch processing of Hive tables

We can make use of Spark SQL for fast batch processing of Hive tables.

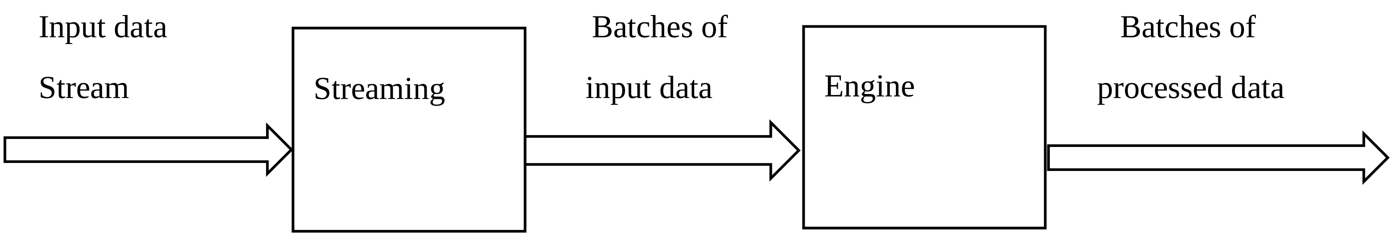
Apache Spark has multiple ways to read data from different sources like files, databases etc. But when it comes to loading data into RDBMS(relational database management system), Spark supports only Append and Overlay of the data using dataframes. Spark dataframes do not support Updating of data into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

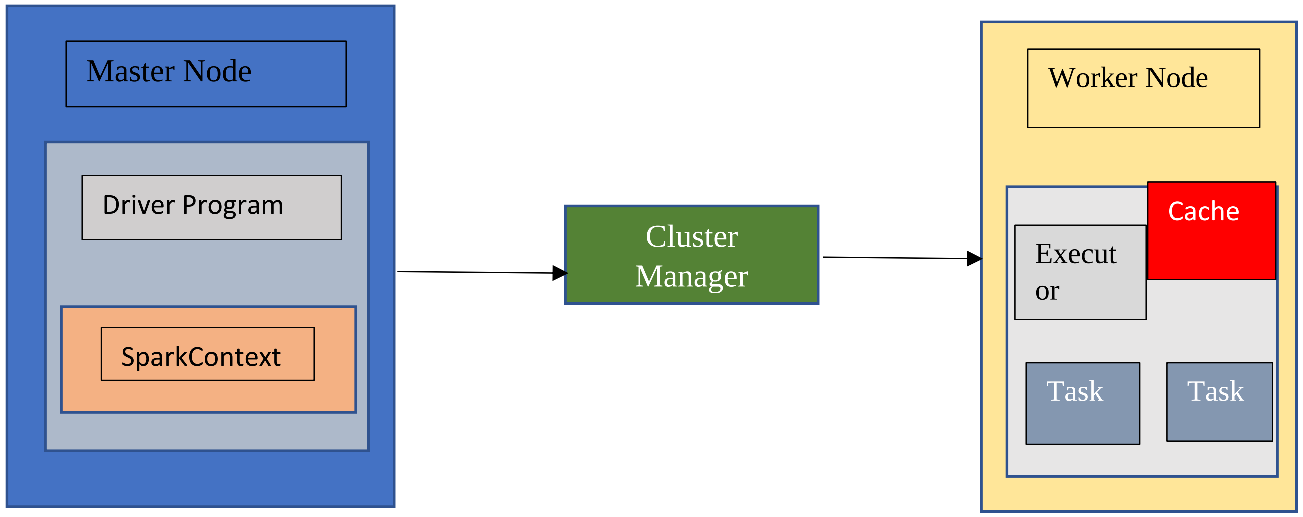
-It is a lightweight API that allows developers to perform batch processing and real-time streaming of data with ease

-provides secure, reliable, and fast processing of live data streams



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

SPARK ARCHITECTURE



As the name suggests, the master-slave is a database architecture divided into a master database and slave databases. The slave database serves as the backup for the master database. The master database is the keeper of the data resources and also the place where all the writing requests are performed.

7)Explain various cluster managers in SPARK?

-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 task

Spark Cluster Managers

1)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

2)MESOS

Apache Mesos is an open-source project to manage computer clusters and can also run Hadoop application

3)Hadoop Yarn

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

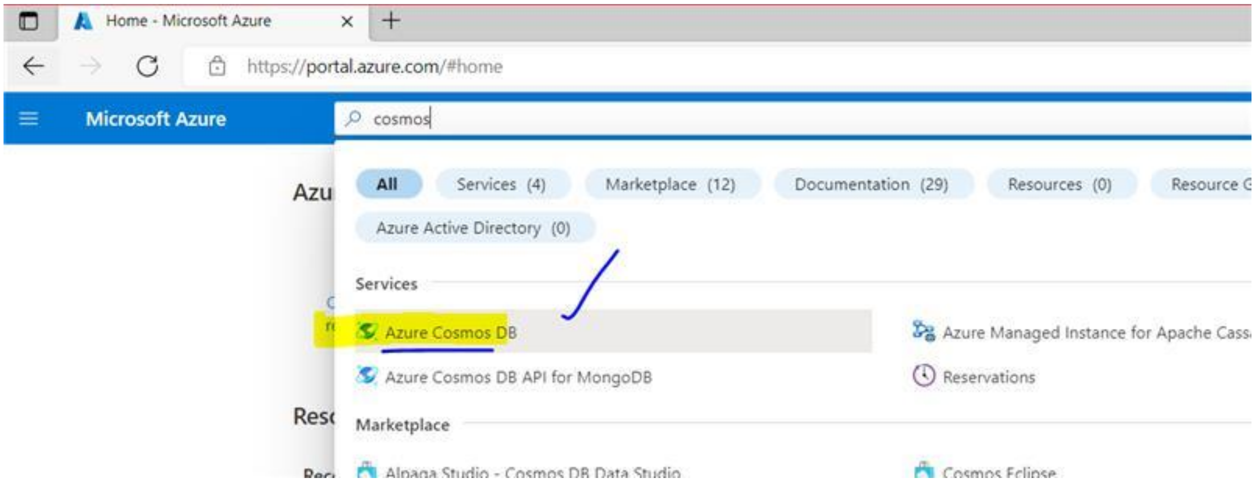
4)kubernetes

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

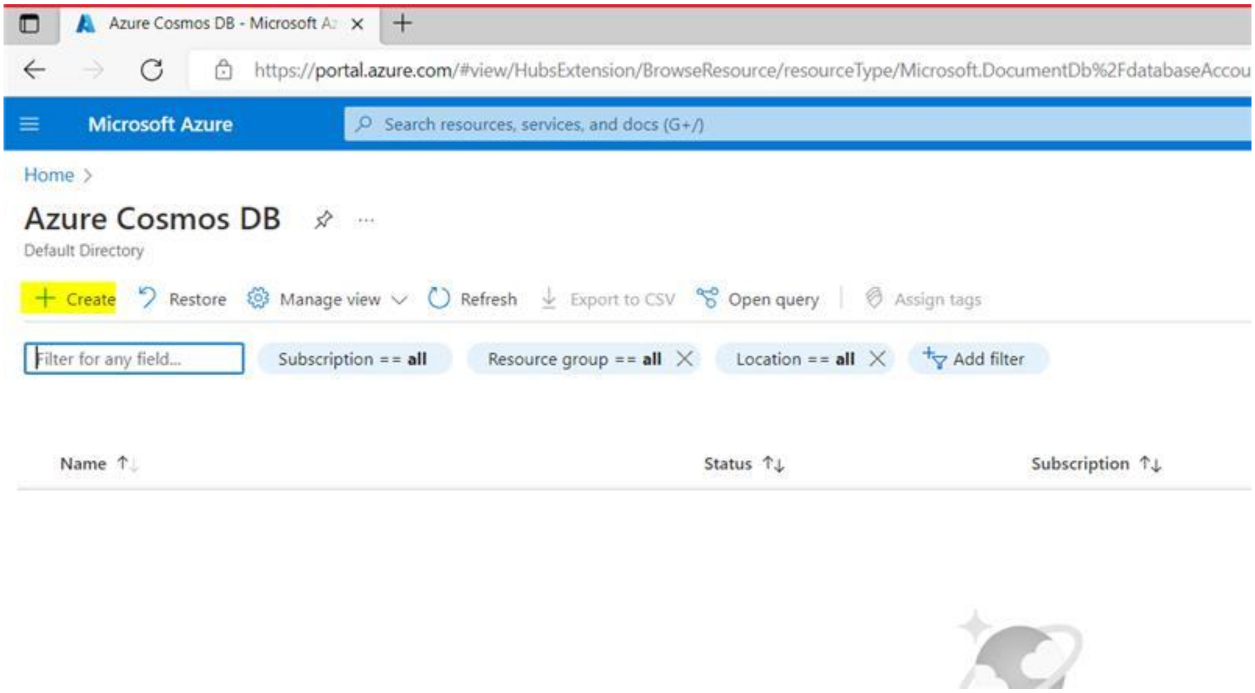
8) Explain with screenshots & steps how to create Cosmos DB?

Go to portal.azure.com

Search for cosmosDB and select Azure Cosmos DB

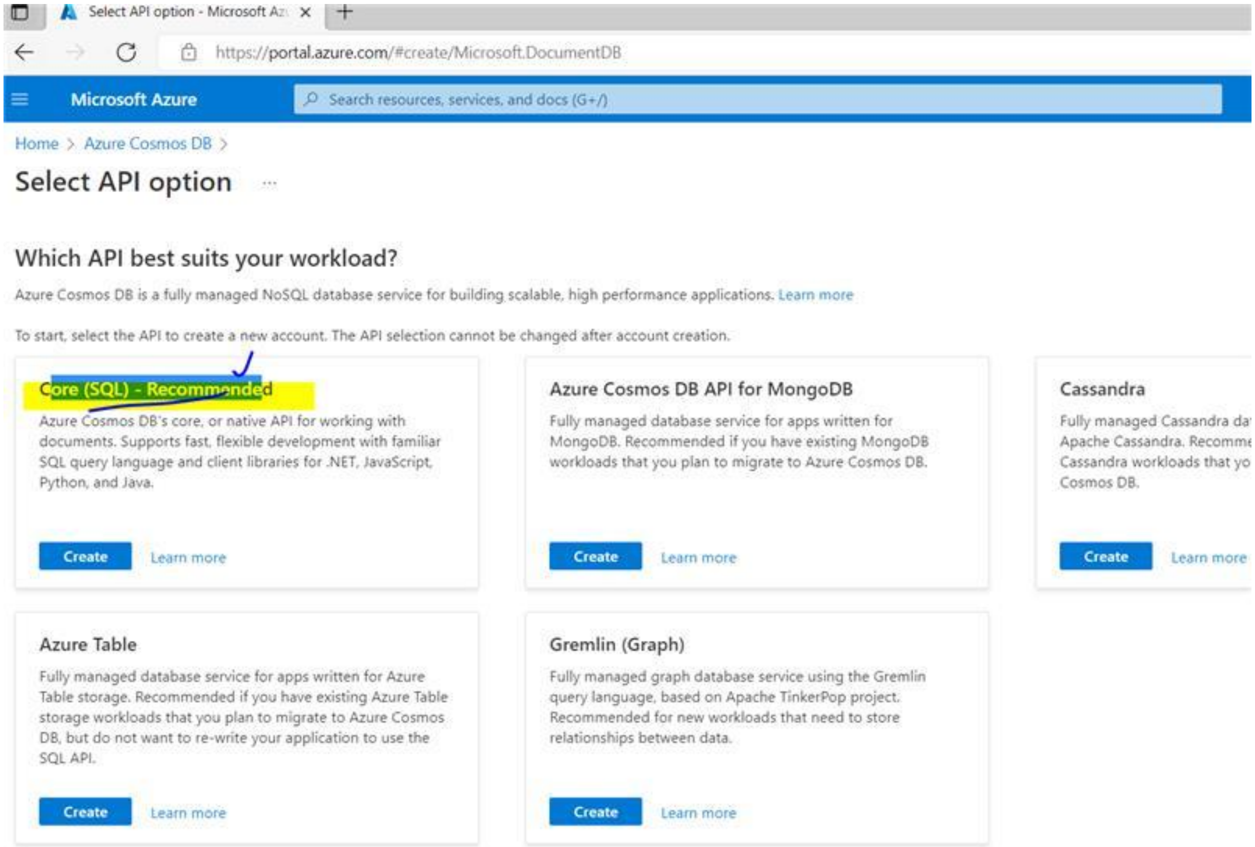


Click on create



-it will show Core(SQL),Azure Cosmos DB API for MongoDB,Cassandra,Azure Table, Gremlin(Graph) options to create

In that select Core(SQL) to create


Select API option - Microsoft Az
https://portal.azure.com/#create/Microsoft.DocumentDB
Microsoft Azure
Search resources, services, and docs (G+/)
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 da Apache Cassandra. Recomme Cassandra workloads that yo Cosmos DB.

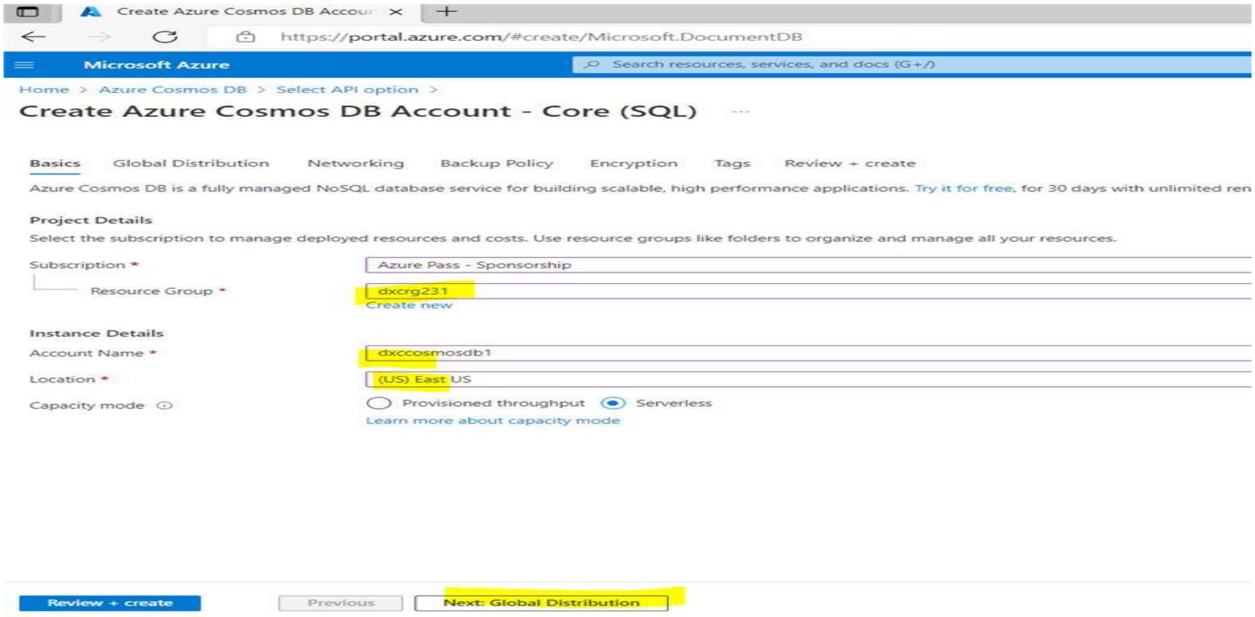
Azure Table

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.

-give accountname dxc and create


Create Azure Cosmos DB Accour
https://portal.azure.com/#create/Microsoft.DocumentDB
Microsoft Azure
Search resources, services, and docs (G+/)
Home > Azure Cosmos DB > Select API option >
Basics
Global Distribution
Networking
Backup Policy
Encryption
Tags
Review + create

Create Azure Cosmos DB Account - Core (SQL) ···

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 ren

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 Pass - Sponsorship

Resource Group * dxcrg231

Create new

Instance Details

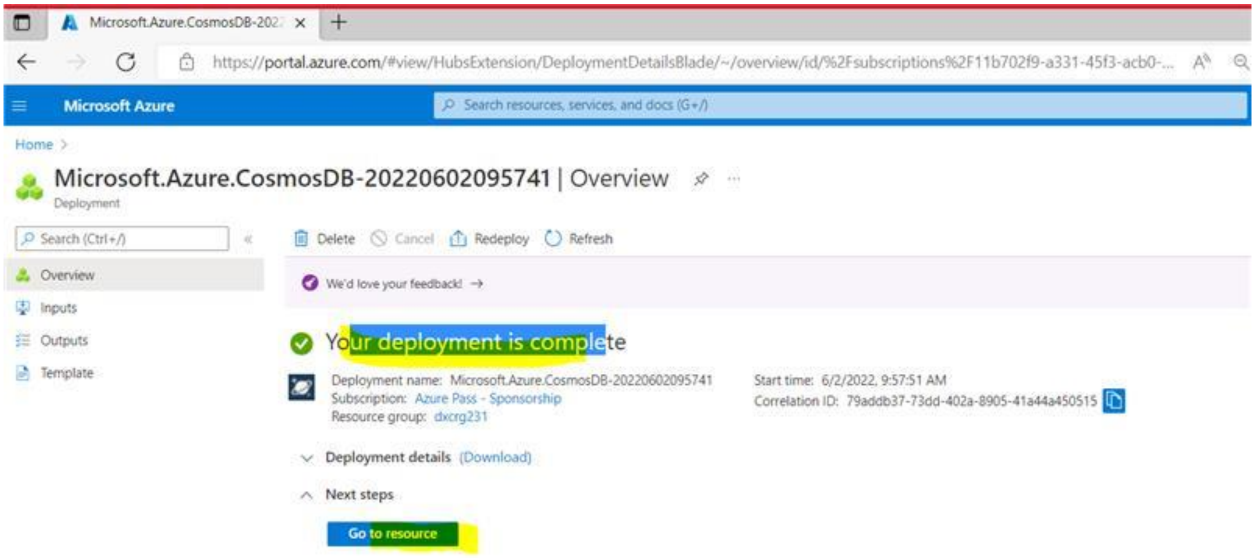
Account Name * dxccosmosdb1

Location * (US) East US

Capacity mode ○ Provisioned throughput ● Serverless

Learn more about capacity mode

Click next till you get this page keep it all default and click next we will get Your deployment is complete


Microsoft.Azure.CosmosDB-202…
https://portal.azure.com/#view/HubsExtension/DeploymentDetailsBlade/~/overview/id/%2Fsubscriptions%2F11b702f9-a331-45f3-acb0-...
Microsoft Azure
Search resources, services, and docs (G+/)
Home >
Microsoft.Azure.CosmosDB-20220602095741 | Overview
Deployment
Search (Ctrl+/)
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Refresh
We'd love your feedback! →
Deployment details (Download)
Next steps
Go to resource

Your deployment is complete

Deployment name: Microsoft.Azure.CosmosDB-20220602095741
Subscription: Azure Pass - Sponsorship
Resource group: dxcrg231

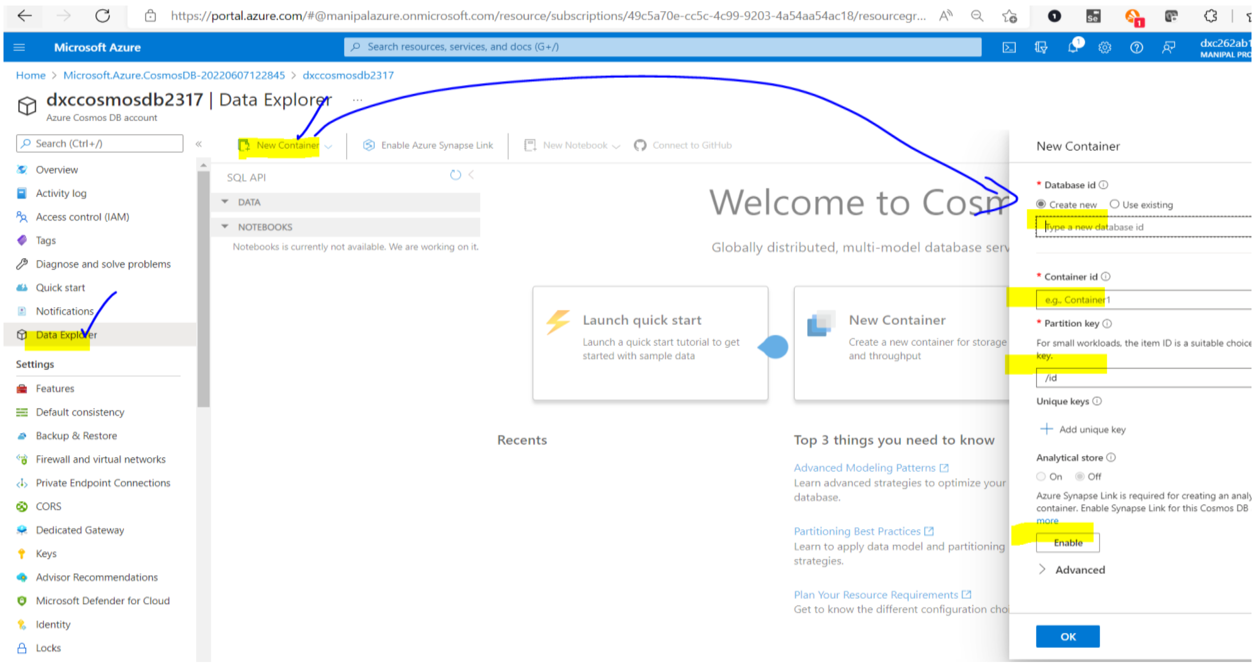
Start time: 6/2/2022, 9:57:51 AM
Correlation ID: 79addb37-73dd-402a-8905-41a44a450515

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

After our deployment is complete go to resource

-click on Data Explorer

And the click on New Container


https://portal.azure.com/#@manipalazure.onmicrosoft.com/resource/subscriptions/49c5a70e-cc5c-4c99-9203-4a54aa54ac18/resourcegr...
Microsoft Azure
Search resources, services, and docs (G+/)
dxc262ab1
MANIPAL PRO
Home > Microsoft.Azure.CosmosDB-20220607122845 > dxccosmosdb2317
dxccosmosdb2317 | 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
Dedicated Gateway
Keys
Advisor Recommendations
Microsoft Defender for Cloud
Identity
Locks
New Container
Enable Azure Synapse Link
New Notebook
Connect to GitHub
SQL API
DATA
NOTEBOOKS
Welcome to Cosm
Globally distributed, multi-model database serv
Launch quick start
New Container
Recents
Top 3 things you need to know
Advanced Modeling Patterns
Partitioning Best Practices
Plan Your Resource Requirements
New Container
Database id
Create new
Use existing
Type a new database id
Container id
e.g., Container1
Partition key
/id
Unique keys
Add unique key
Analytical store
On
Off
Enable
Advanced
OK

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

Launch a quick start tutorial to get started with sample data

Create a new container for storage and throughput

Learn advanced strategies to optimize your database.

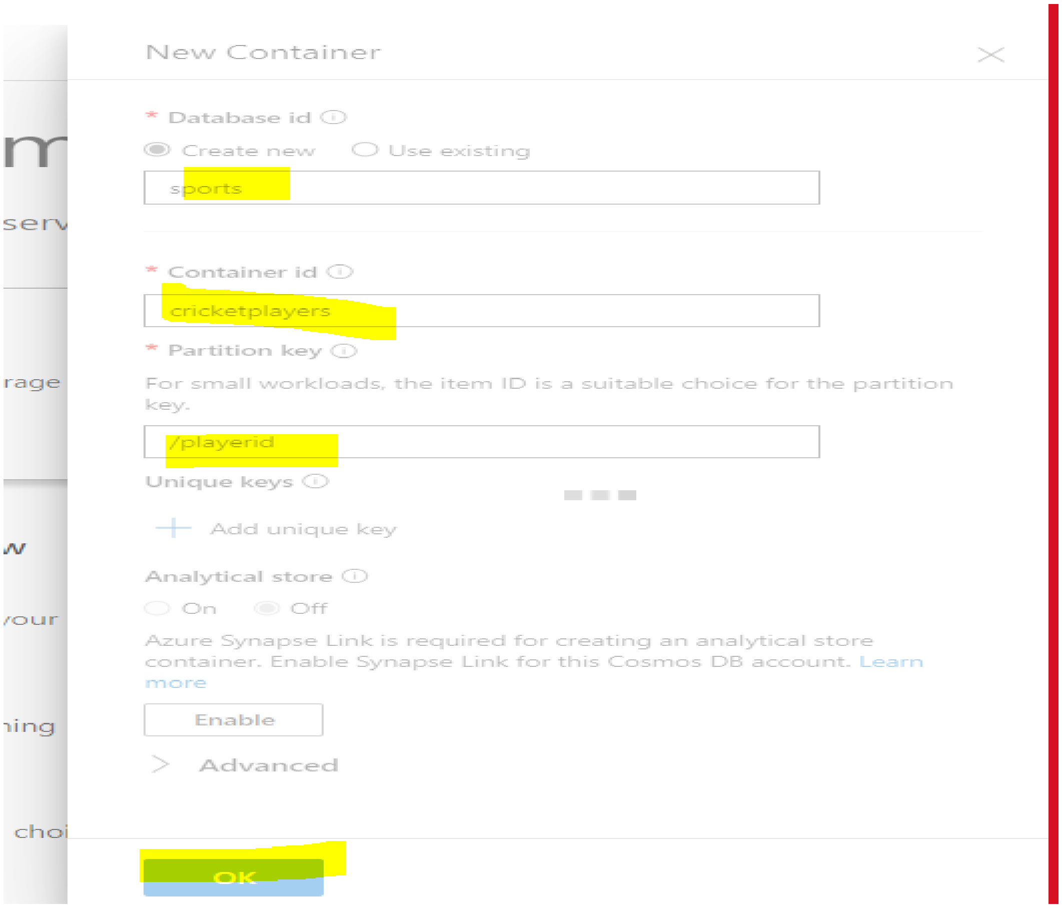
Learn to apply data model and partitioning strategies.

Get to know the different configuration cho

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

Azure Synapse Link is required for creating an analy container. Enable Synapse Link for this Cosmos DB more

-Fill the New Container details and click on OK



New Container

* Database id

● Create new ○ Use existing

sports

* Container id

cricketplayers

* Partition key

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

/playerid

Unique keys

+ Add unique key

Analytical store

○ On ● Off

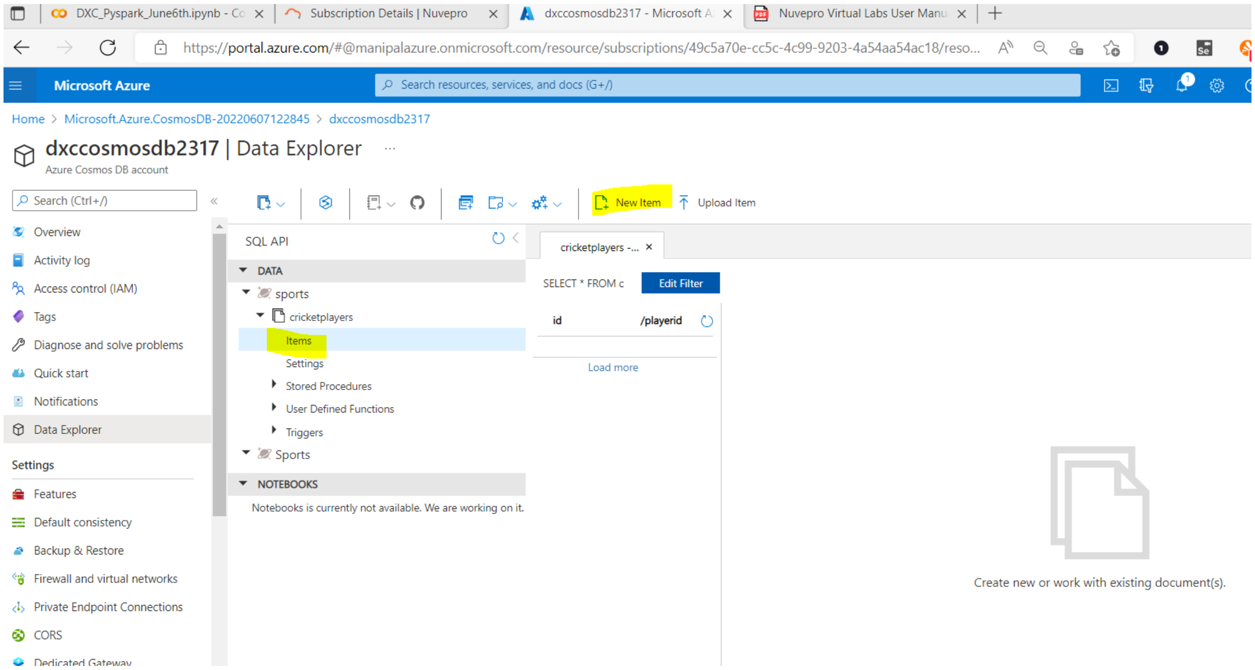
Azure Synapse Link is required for creating an analytical store container. Enable Synapse Link for this Cosmos DB account. Learn more

Enable

> Advanced

OK

-After that go to sports->items and click New item



https://portal.azure.com/#@manipalazure.onmicrosoft.com/resource/subscriptions/49c5a70e-cc5c-4c99-9203-4a54aa54ac18/reso...

Microsoft Azure

Home > Microsoft.Azure.CosmosDB-20220607122845 > dxccosmosdb2317

dxccosmosdb2317 | Data Explorer

Azure Cosmos DB account

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

cricketplayers -... ×

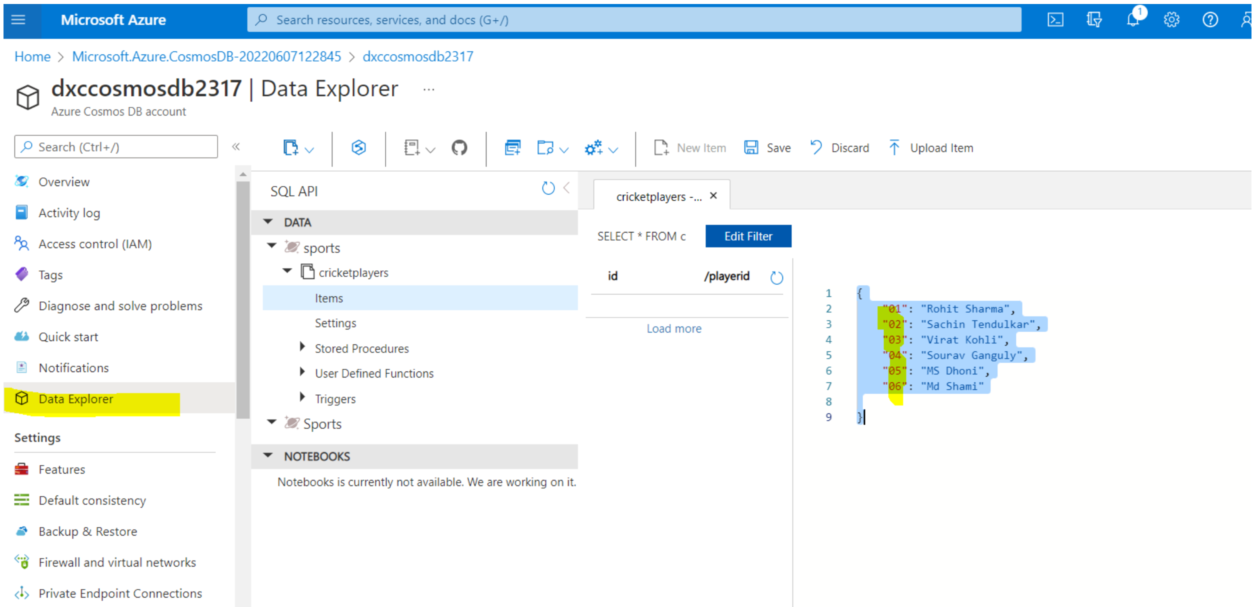
SELECT * FROM c Edit Filter

id	/playerid

Load more

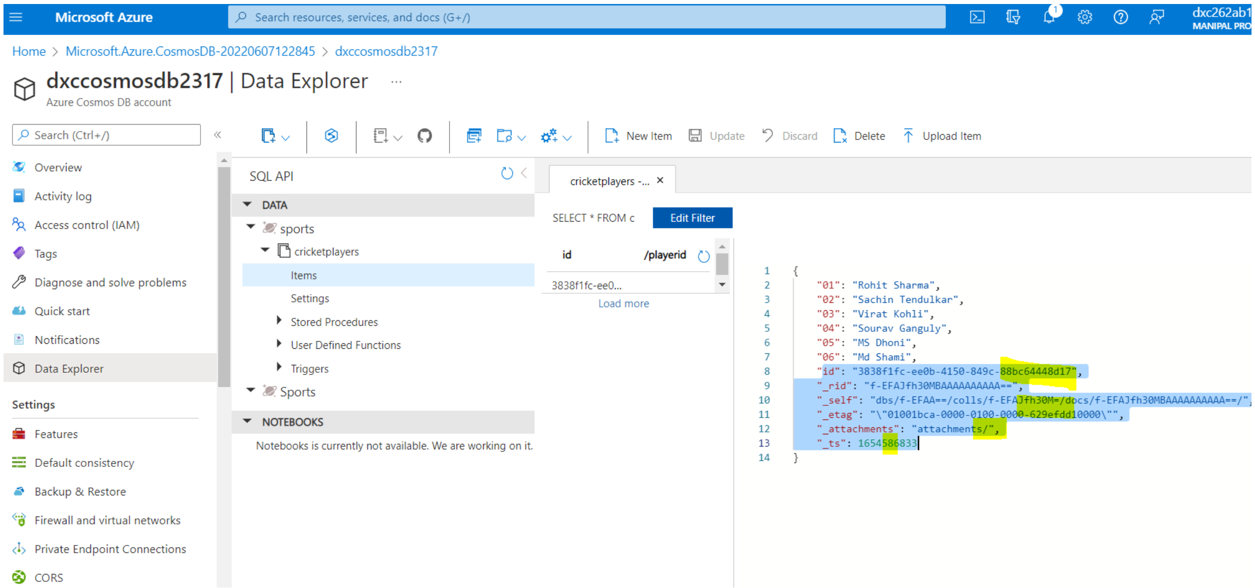
Create new or work with existing document(s).

-Write the data with key and value like the below



```
{
    "01": "Rohit Sharma",
    "02": "Sachin Tendulkar",
    "03": "Virat Kohli",
    "04": "Sourav Ganguly",
    "05": "MS Dhoni",
    "06": "Md Shami"
}
```

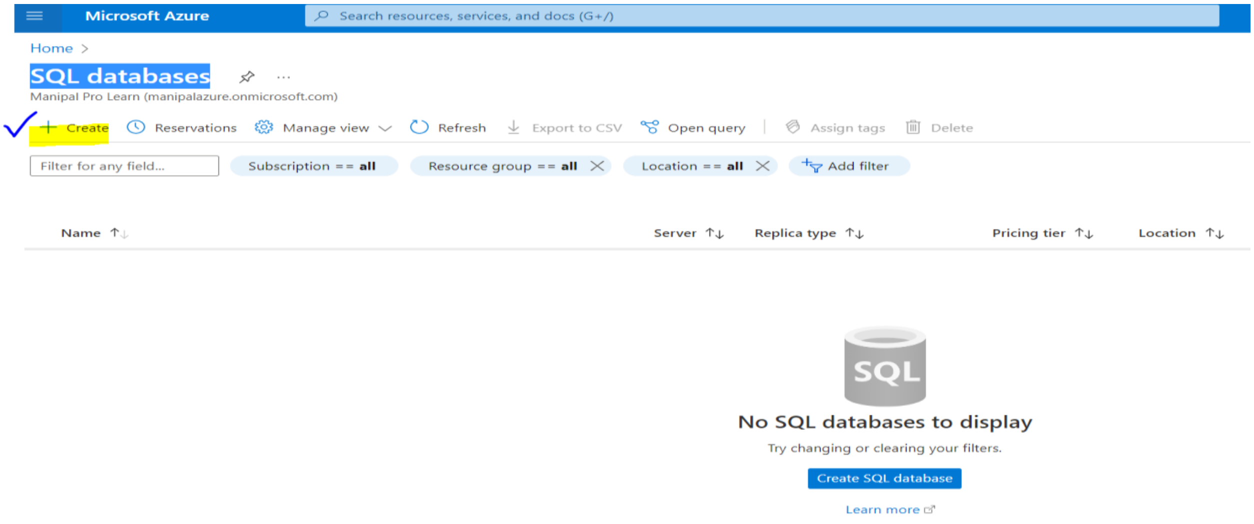
And then save it the data is stored



```
{
    "01": "Rohit Sharma",
    "02": "Sachin Tendulkar",
    "03": "Virat Kohli",
    "04": "Sourav Ganguly",
    "05": "MS Dhoni",
    "06": "Md Shami",
    "id": "3838f1fc-ee0b-4150-849c-88bc64448d17",
    "_rid": "f-EFAJfh30MBAAAAAAAAAAA==",
    "_self": "dbs/f-EFAA==/colls/f-EFAJfh30M=/docs/f-EFAJfh30MBAAAAAAAAAAA==/",
    "_etag": "\"01001bca-0000-0100-0000-629efdd10000\"",
    "_attachments": "attachments/",
    "_ts": 1654586833
}
```

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

-First search the sql database and then we will get the home page like the below one



Microsoft Azure

Search resources, services, and docs (G+/)

Home >

SQL databases

Manipal Pro Learn (manipalazure.onmicrosoft.com)

+ Create | Reservations | Manage view | Refresh | Export to CSV | Open query | Assign tags | Delete

Filter for any field... | Subscription == all | Resource group == all | Location == all | Add filter

Name ↑↓	Server ↑↓	Replica type ↑↓	Pricing tier ↑↓	Location ↑↓

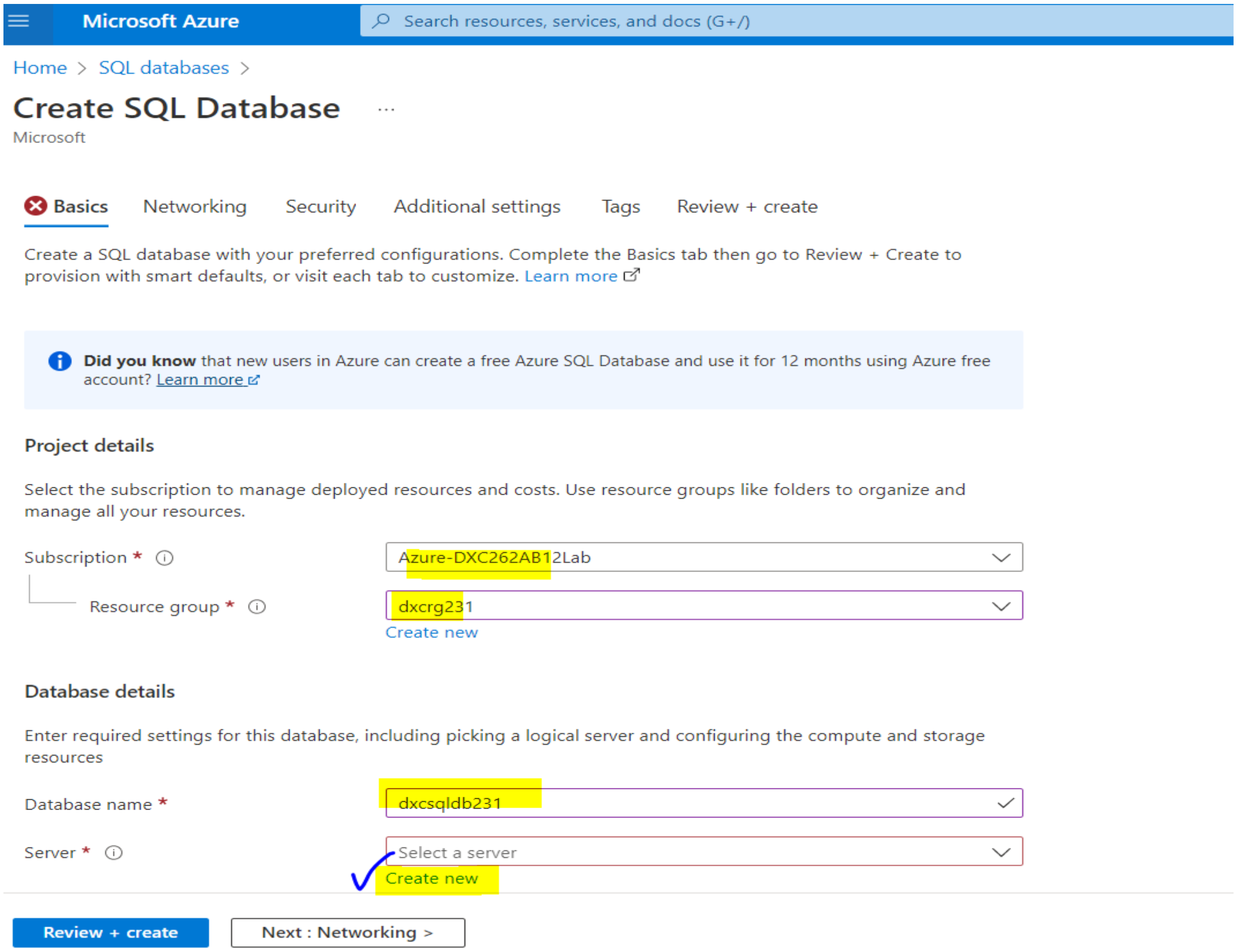
No SQL databases to display

Try changing or clearing your filters.

Create SQL database

Learn more

-Click on Create option we will go to the basics where we will give the details



Microsoft Azure

Search resources, services, and docs (G+/)

Home > SQL databases >

Create SQL Database ...

Microsoft

Basics | Networking | Security | Additional settings | Tags | Review + create

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 * : dxcrg231

Create new

Database details

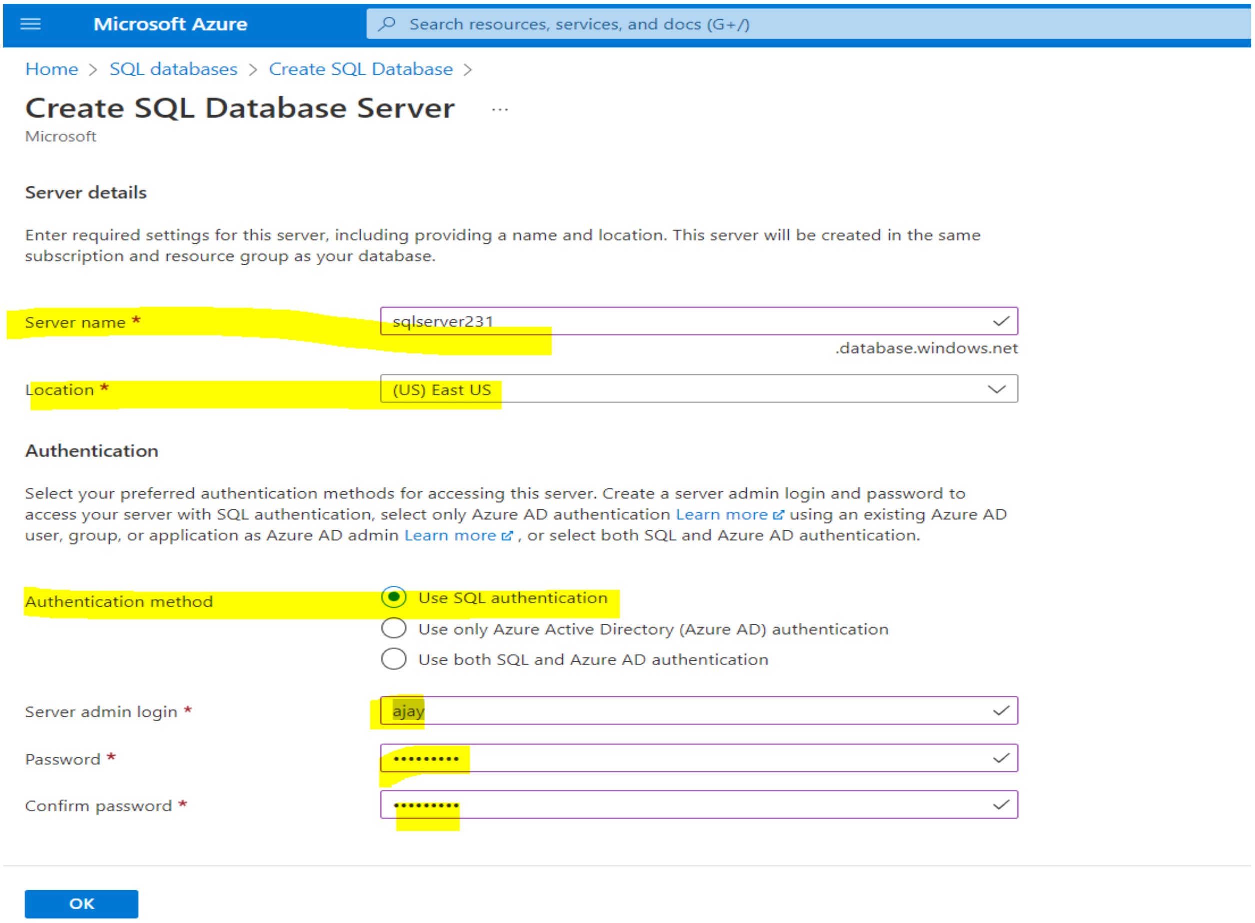
Enter required settings for this database, including picking a logical server and configuring the compute and storage resources

Database name * : dxcsqldb231

Server * : Select a server

Create new

Review + create | Next : Networking >



Microsoft Azure
Search resources, services, and docs (G+/)
Home > SQL databases > Create SQL Database >

Create SQL Database Server ...

Microsoft

Server details

Enter required settings for this server, including providing a name and location. This server will be created in the same subscription and resource group as your database.

Server name * sqlserver231 .database.windows.net

Location * (US) East US

Authentication

Select your preferred authentication methods for accessing this server. Create a server admin login and password to access your server with SQL authentication, select only Azure AD authentication Learn more using an existing Azure AD user, group, or application as Azure AD admin Learn more , or select both SQL and Azure AD authentication.

Authentication method
● Use SQL authentication
○ Use only Azure Active Directory (Azure AD) authentication
○ Use both SQL and Azure AD authentication

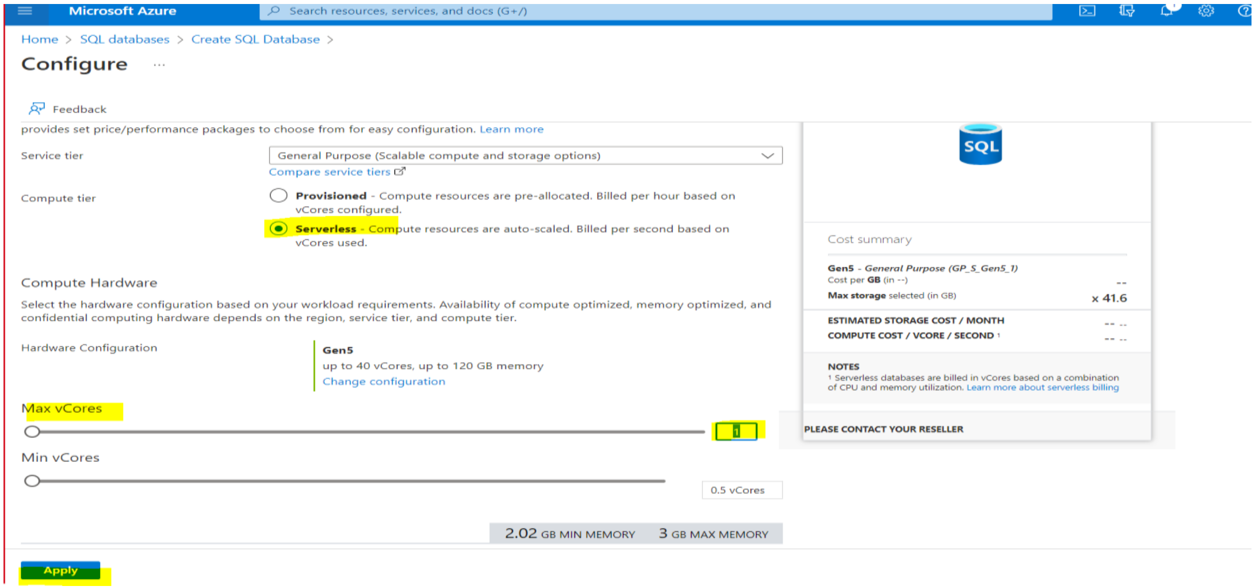
Server admin login * ajay

Password * •••••••••

Confirm password * •••••••••

-give server name and location

-click Configure database give Max vCore 1 GB and Apply



Microsoft Azure
Search resources, services, and docs (G+/)
Home > SQL databases > Create SQL Database >

Configure ...

provides set price/performance packages to choose from for easy configuration. Learn more

Service tier: General Purpose (Scalable compute and storage options)
Compare service tiers

Compute tier
○ **Provisioned** - Compute resources are pre-allocated. Billed per hour based on vCores configured.
● **Serverless** - Compute resources are auto-scaled. Billed per second based on vCores used.

Compute Hardware

Select the hardware configuration based on your workload requirements. Availability of compute optimized, memory optimized, and confidential computing hardware depends on the region, service tier, and compute tier.

Hardware Configuration: **Gen5**
up to 40 vCores, up to 120 GB memory
Change configuration

Max vCores: 1

Min vCores: 0.5 vCores

2.02 GB MIN MEMORY 3 GB MAX MEMORY

Cost summary

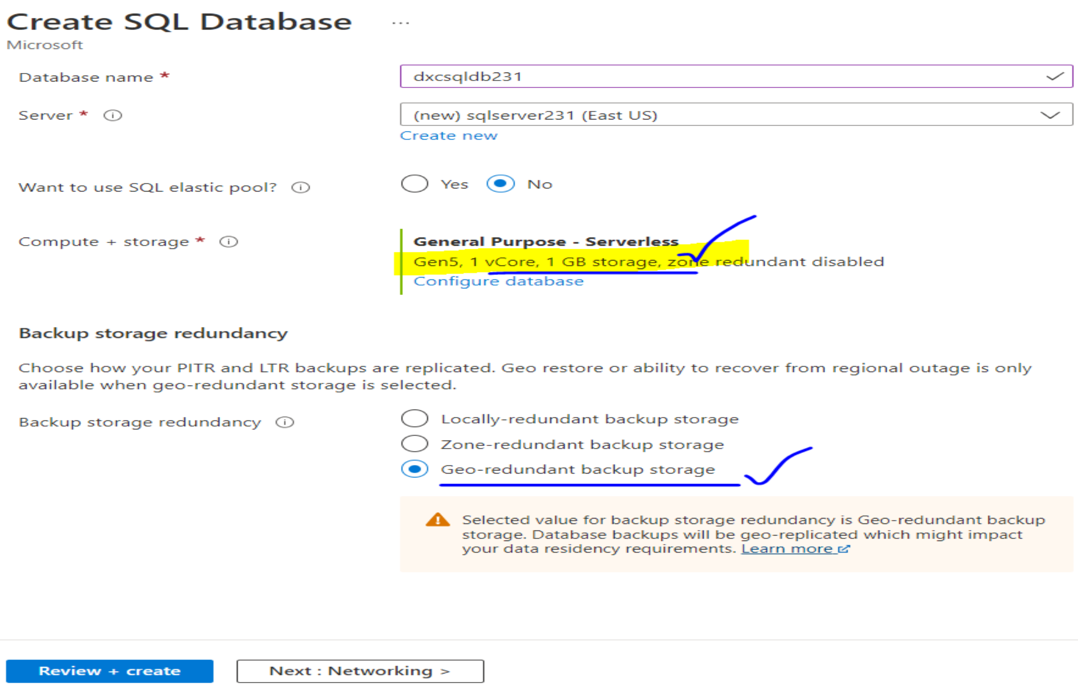
Gen5 - *General Purpose (GP_S_Gen5_1)*
Cost per **GB** (in --) --
Max storage selected (in GB) x 41.6

ESTIMATED STORAGE COST / MONTH -- --
COMPUTE COST / VCORE / SECOND [1] -- --

NOTES
[1] Serverless databases are billed in vCores based on a combination of CPU and memory utilization. Learn more about serverless billing

PLEASE CONTACT YOUR RESELLER

Keep backup storage redundancy as Geo-redundant backup storage



Create SQL Database ···

Microsoft

Database name * dxcsqldb231

Server * (new) sqlserver231 (East US)

Create new

Want to use SQL elastic pool? ○ Yes ● No

Compute + storage * General Purpose - Serverless

Gen5, 1 vCore, 1 GB storage, zone redundant disabled

Configure database

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

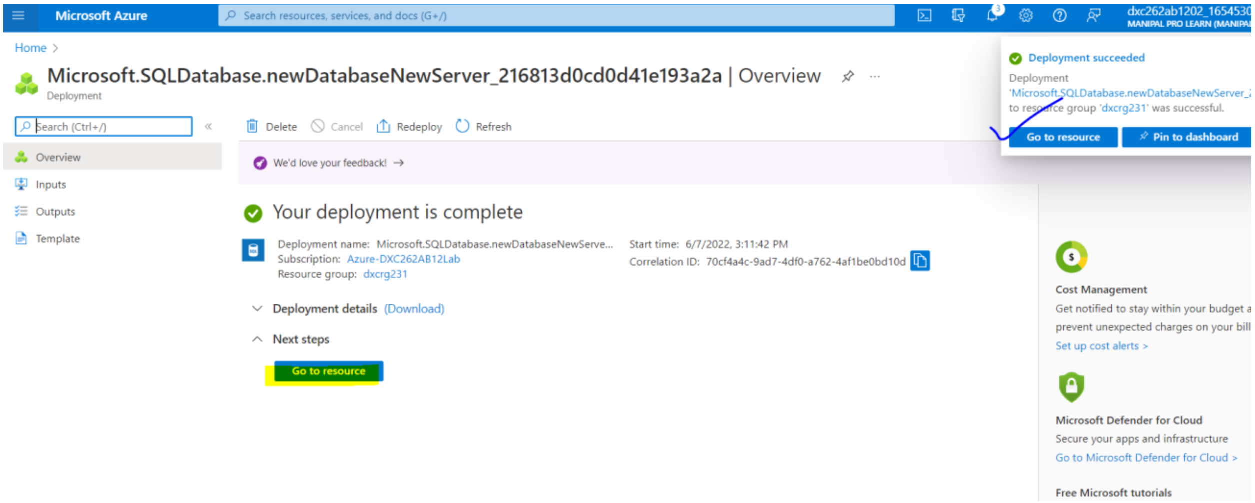
○ Locally-redundant backup storage

○ 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 review + create and Our Deployment is complete



Microsoft.SQLDatabase.newDatabaseNewServer_216813d0cd0d41e193a2a | Overview

Deployment

Deployment succeeded

Deployment 'Microsoft.SQLDatabase.newDatabaseNewServer_ to resource group 'dxcrg231' was successful.

Your deployment is complete

Deployment name: Microsoft.SQLDatabase.newDatabaseNewServe...

Subscription: Azure-DXC262AB12Lab

Resource group: dxcrg231

Start time: 6/7/2022, 3:11:42 PM

Correlation ID: 70cf4a4c-9ad7-4df0-a762-4af1be0bd10d

Deployment details (Download)

Next steps

Go to resource

Cost Management

Get notified to stay within your budget a prevent unexpected charges on your bill

Set up cost alerts >

Microsoft Defender for Cloud

Secure your apps and infrastructure

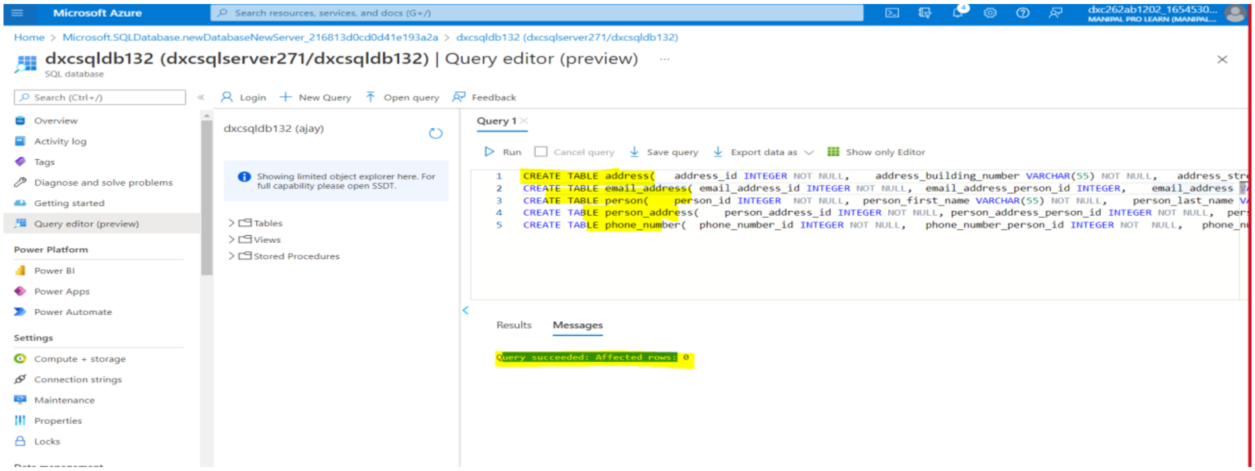
Go to Microsoft Defender for Cloud >

Free Microsoft tutorials

To insert data in Azure Sql after deployment go to

-Query editor (preview)

Give your login id and password and click ok

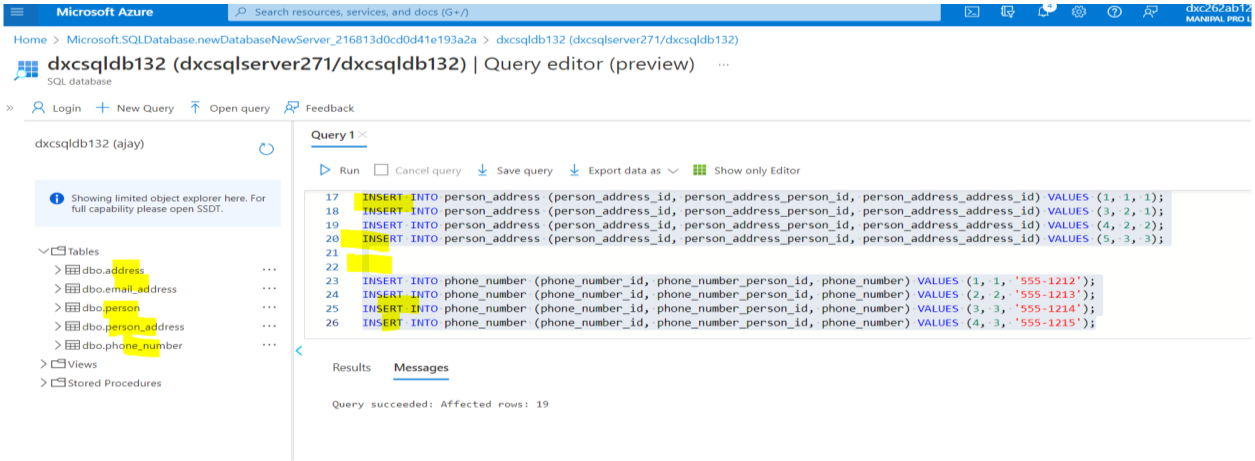


```
CREATE TABLE address(    address_id INTEGER NOT NULL,    address_building_number VARCHAR(55) NOT NULL,    address_stre
CREATE TABLE email_address( email_address_id INTEGER NOT NULL,  email_address_person_id INTEGER,    email_address V
CREATE TABLE person(    person_id INTEGER  NOT NULL,  person_first_name VARCHAR(55) NOT NULL,    person_last_name V
CREATE TABLE person_address(    person_address_id INTEGER NOT NULL, person_address_person_id INTEGER NOT NULL,  per
CREATE TABLE phone_number(  phone_number_id INTEGER NOT NULL,   phone_number_person_id INTEGER NOT  NULL,   phone_n
```

```
Query succeeded: Affected rows: 0
```

Create the tables whatever we want and run the query tables will be created

After that insert the values we need in the table we get tables with inserted values

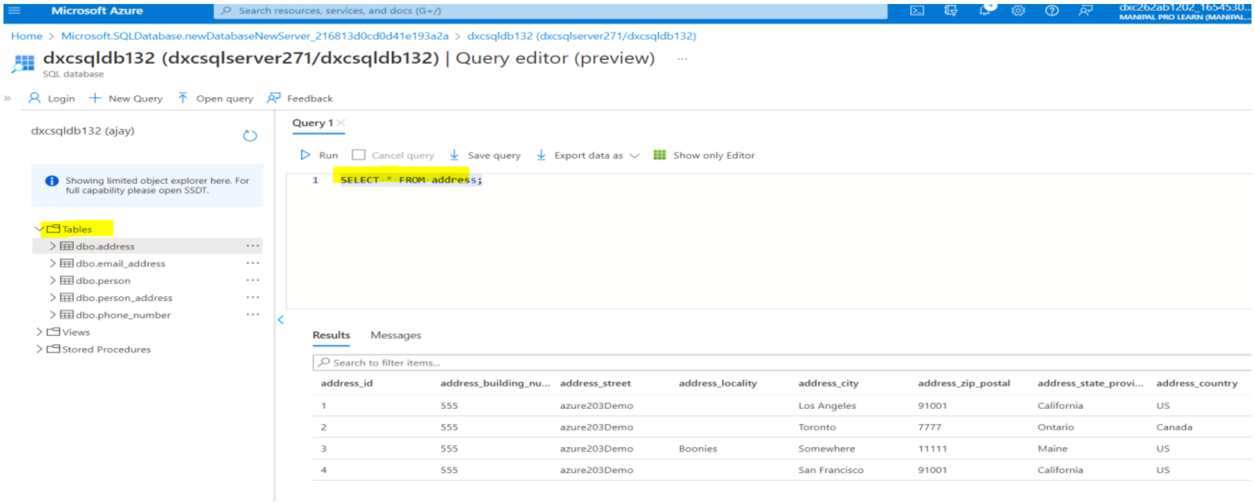

Microsoft Azure
Home > Microsoft.SQLDatabase.newDatabaseNewServer_216813d0cd0d41e193a2a > dxcsqldb132 (dxcsqlserver271/dxcsqldb132)
dxcsqldb132 (dxcsqlserver271/dxcsqldb132) | Query editor (preview)
SQL database
Login
New Query
Open query
Feedback
dxcsqldb132 (ajay)
Showing limited object explorer here. For full capability please open SSDT.
Tables
dbo.address
dbo.email_address
dbo.person
dbo.person_address
dbo.phone_number
Views
Stored Procedures
Query 1
Run
Cancel query
Save query
Export data as
Show only Editor
Results
Messages

```
INSERT INTO person_address (person_address_id, person_address_person_id, person_address_address_id) VALUES (1, 1, 1);
INSERT INTO person_address (person_address_id, person_address_person_id, person_address_address_id) VALUES (3, 2, 1);
INSERT INTO person_address (person_address_id, person_address_person_id, person_address_address_id) VALUES (4, 2, 2);
INSERT INTO person_address (person_address_id, person_address_person_id, person_address_address_id) VALUES (5, 3, 3);


INSERT INTO phone_number (phone_number_id, phone_number_person_id, phone_number) VALUES (1, 1, '555-1212');
INSERT INTO phone_number (phone_number_id, phone_number_person_id, phone_number) VALUES (2, 2, '555-1213');
INSERT INTO phone_number (phone_number_id, phone_number_person_id, phone_number) VALUES (3, 3, '555-1214');
INSERT INTO phone_number (phone_number_id, phone_number_person_id, phone_number) VALUES (4, 3, '555-1215');
```

```
Query succeeded: Affected rows: 19
```

To see the inserted values in the table go to tables and write the query which is select * from table_name


Microsoft Azure
Home > Microsoft.SQLDatabase.newDatabaseNewServer_216813d0cd0d41e193a2a > dxcsqldb132 (dxcsqlserver271/dxcsqldb132)
dxcsqldb132 (dxcsqlserver271/dxcsqldb132) | Query editor (preview)
SQL database
Login
New Query
Open query
Feedback
dxcsqldb132 (ajay)
Showing limited object explorer here. For full capability please open SSDT.
Tables
dbo.address
dbo.email_address
dbo.person
dbo.person_address
dbo.phone_number
Views
Stored Procedures
Query 1
Run
Cancel query
Save query
Export data as
Show only Editor
Results
Messages
Search to filter items...

```
SELECT * FROM address;
```

address_id	address_building_nu...	address_street	address_locality	address_city	address_zip_postal	address_state_provi...	address_country
1	555	azure203Demo		Los Angeles	91001	California	US
2	555	azure203Demo		Toronto	7777	Ontario	Canada
3	555	azure203Demo	Boonies	Somewhere	11111	Maine	US
4	555	azure203Demo		San Francisco	91001	California	US