**ASSIGNMENT NO. 07**

**Name – Atharva Shivaji Jadhav Roll Number - DXC262AB12035**

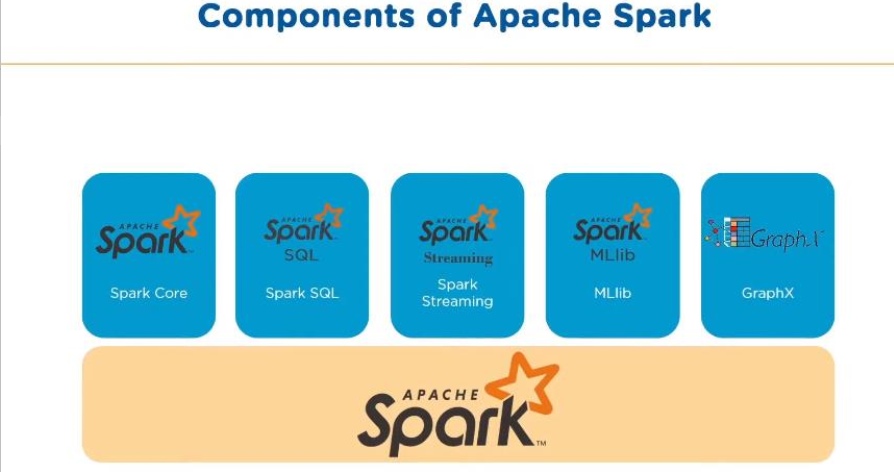
**Batch – Dxc-262-Analytics-B12-Azure Company – DXC Technology**

**Employee Domain –Azure Analytics Training Under – Manipal Pro Learn**

**Trainer Name – Mr. Ajay Kumar**

**Q1- 1. Explain what are various components of SPARK with block diagram?**

**explain functionality of every components?**

****

**1**.**SparkSQL** – The Spark SQL is built on the top of Spark Core. It provides support for structured data. It allows to query the data via SQL, It also supports various sources of data like Hive tables, Parquet, and JSON.

**2. Spark Streaming** is a Spark component that supports scalable and fault-tolerant processing of streaming data. The log files generated by web servers can be considered as a real-time example of a data stream.

**3. The MLlib** is a Machine Learning library that contains various machine learning algorithms.

**4. The GraphX** is a library that is used to manipulate graphs and perform graph-parallel computations.

It facilitates to create a directed graph with arbitrary properties attached to each vertex and edge.

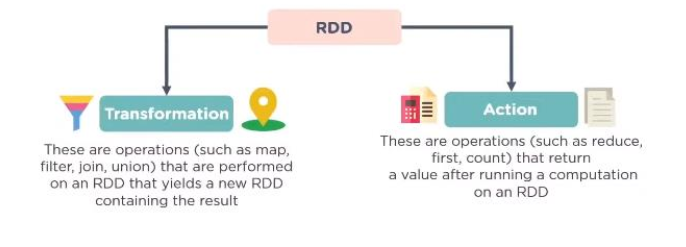
**2. Explain Spark core in details & how RDD is related to Spark core - explain with Spark program ?**

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.

**Spark core**  is responsible for –

1. Memory management
2. Fault recovery
3. Schedulling, distributing and monitor jobs on a cluster.
4. Interacting with storage systems

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



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

MLlib 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—

**Clustering**

**Classification**

**Collaborative Filtering**

Also,  regression, , dimensionality reduction, as well as underlying optimization primitives

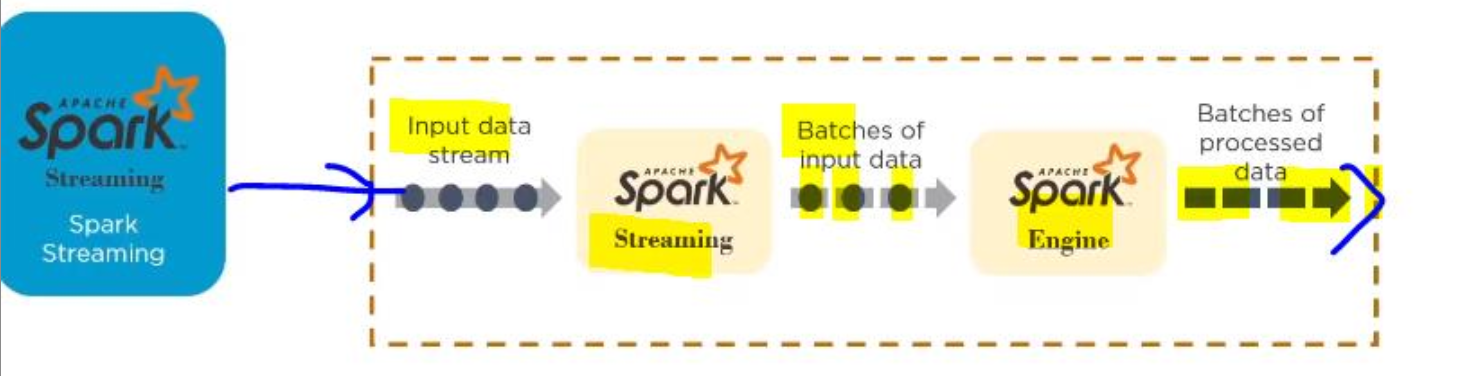
**4.Explain benifits Spark SQL & how relational data will be inserted into SPARK ?**

1. It helps in easy data querying. The SQL queries are mixed with Spark programs for querying structured data as a distributed dataset (RDD). Also, the SQL queries are run with analytic algorithms using Spark SQL’s integration property.
2. Another important advantage of Spark SQL is that the loading and querying can be done for data from different sources. Hence, the data access is unified.
3. It offers standard connectivity as Spark SQL can be connected through JDBC or ODBC.
4. It can be used for faster processing of Hive tables.
5. Another important offering of Spark SQL is that it can run unmodified Hive queries on existing warehouses as it allows easy compatibility with existing Hive data and queries.

**5.Explain Spark streaming in detail ?**

**Spark Streaming**  is lightweight api that allows developers to perform batch processing and realtime streaming of data with ease.

Provides **secure**, **reliable** and **fast** **processing** of live data streams.



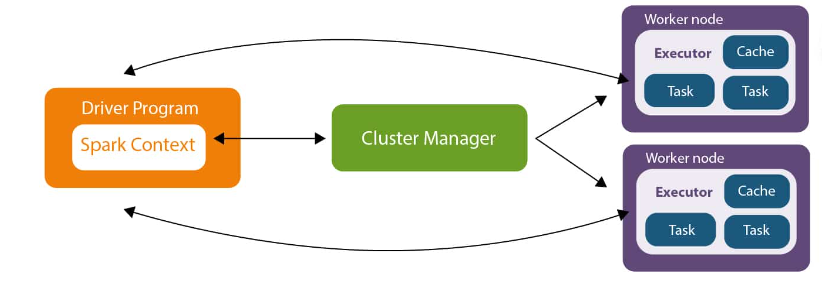
Apache Spark Streaming is a scalable fault-tolerant streaming processing system that natively supports both batch and streaming workloads.

* 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 architecure? what is Master - Slave architecure ?**

The Apache Spark framework uses a **master-slave architecture** that consists of a **driver**, which runs as a **master** node, and many executors that run across as **worker** nodes in the **cluster**.

Apache Spark can be used for **batch processing** and **real-time processing** as well.



Driver Program in the Apache Spark architecture calls the main program of an application and creates SparkContext.

A SparkContext consists of all the basic functionalities. Spark Driver contains various other components such as DAG Scheduler, Task Scheduler, Backend Scheduler, and Block Manager, which are responsible for translating the user-written code into jobs that are actually executed on the cluster.

Spark Driver and SparkContext collectively watch over the job execution within the cluster. Spark Driver works with the Cluster Manager to manage various other jobs. The cluster Manager does the resource allocating work. And then, the job is split into multiple smaller tasks which are further distributed to worker nodes.

Whenever an RDD is created in the SparkContext, it can be distributed across many worker nodes and can also be cached there. Worker nodes execute the tasks assigned by the Cluster Manager and return it back to the Spark Context. An executor is responsible for the execution of these tasks. The lifetime of executors is the same as that of the Spark Application. If we want to increase the performance of the system, we can increase the number of workers so that the jobs can be divided into more logical portions.

**7.Explain various cluster managers in SPARK?**

There are three types of Spark cluster manager.

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

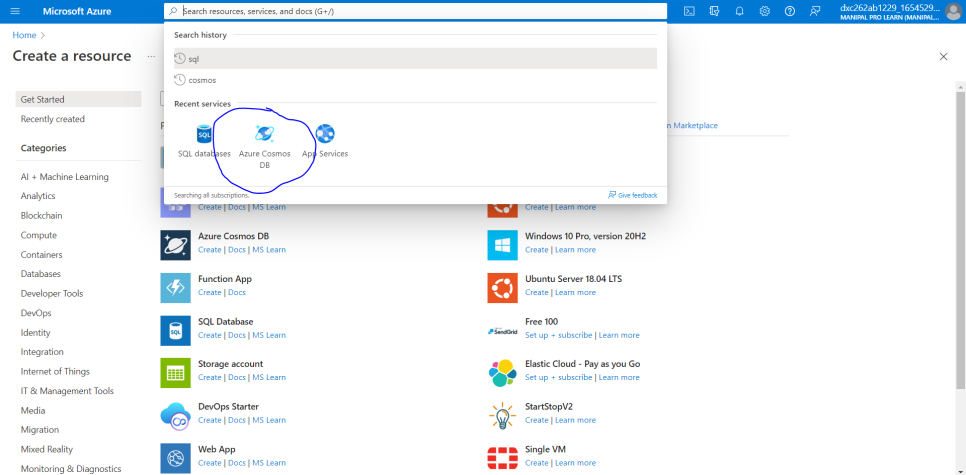
**1.Standalone Cluster Manager** It is a part of spark distribution and available as a simple cluster manager to us. Standalone cluster manager is resilient in nature, it can handle work failures. It has capabilities to manage resources according to the requirement of applications.

**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.

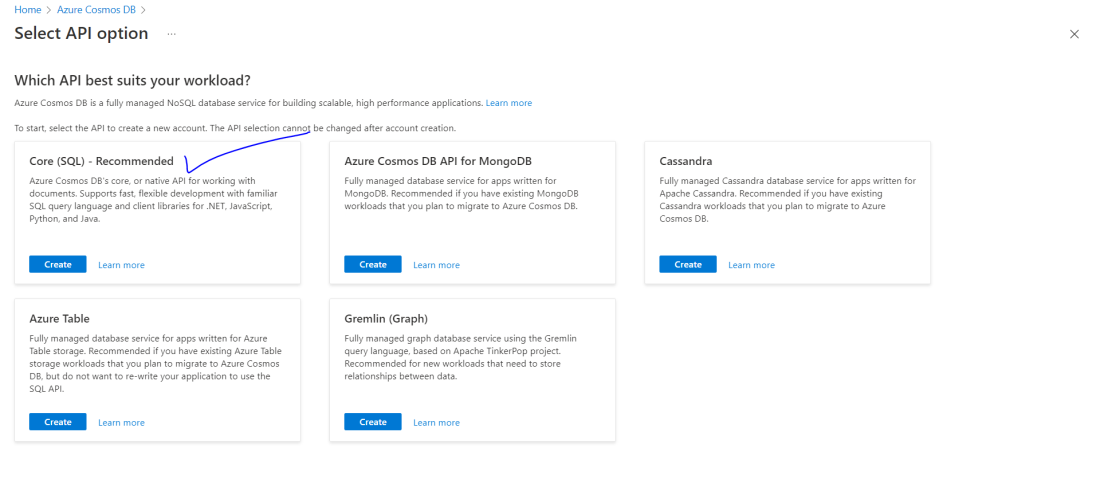
**3. Apache Mesos** It is a distributed cluster manager. As like yarn, it is also highly available for master and slaves. It can also manage resource per application. We can run spark jobs, Hadoop MapReduce or any other service applications easily.

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

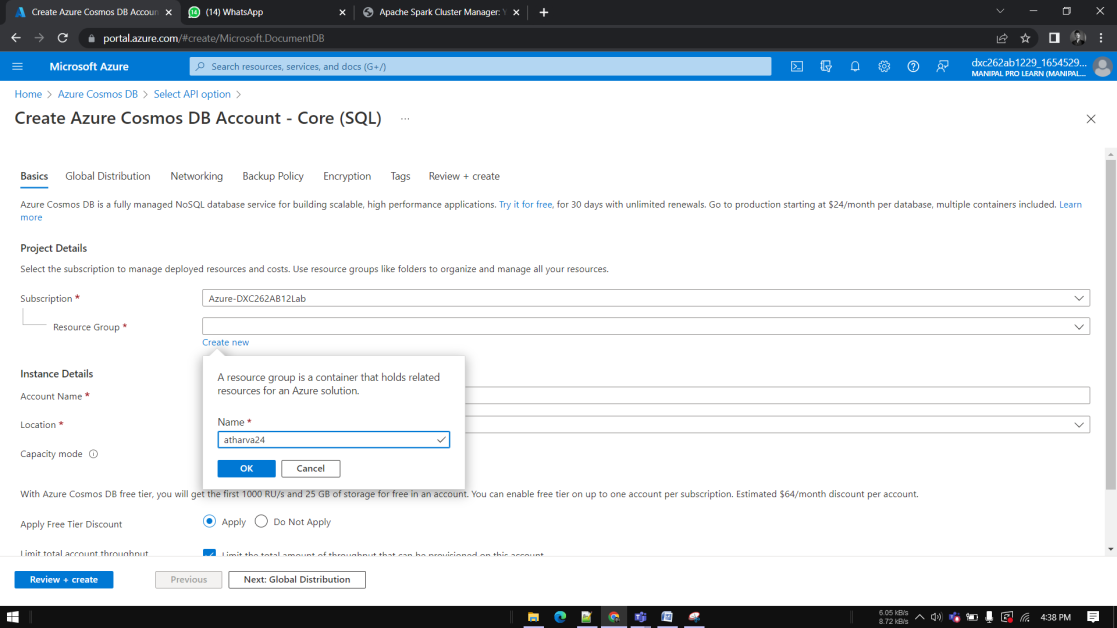
**Step1—**

****

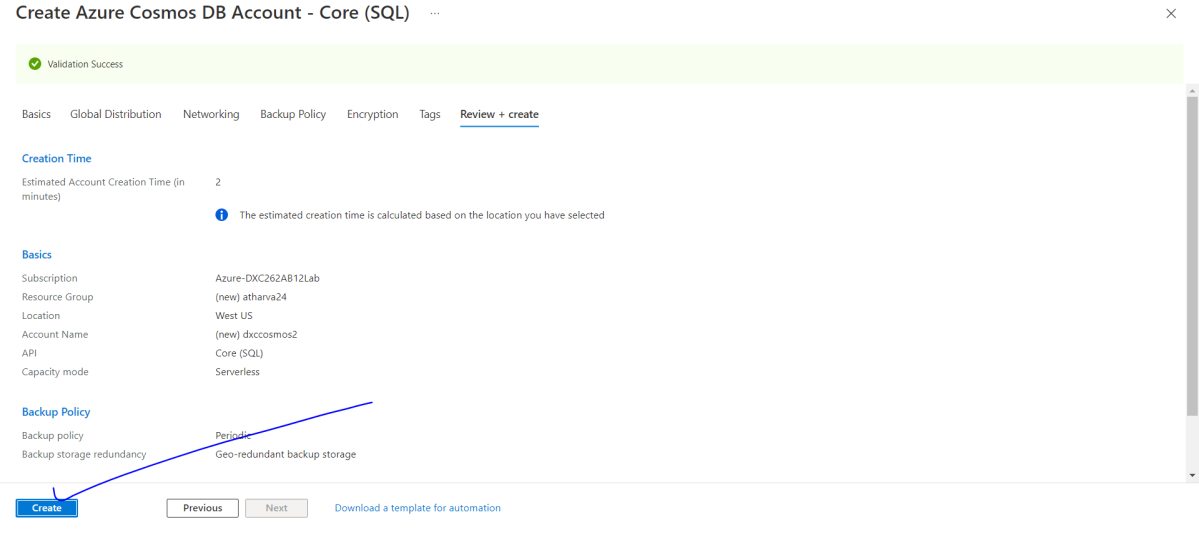
**Step2—**

****

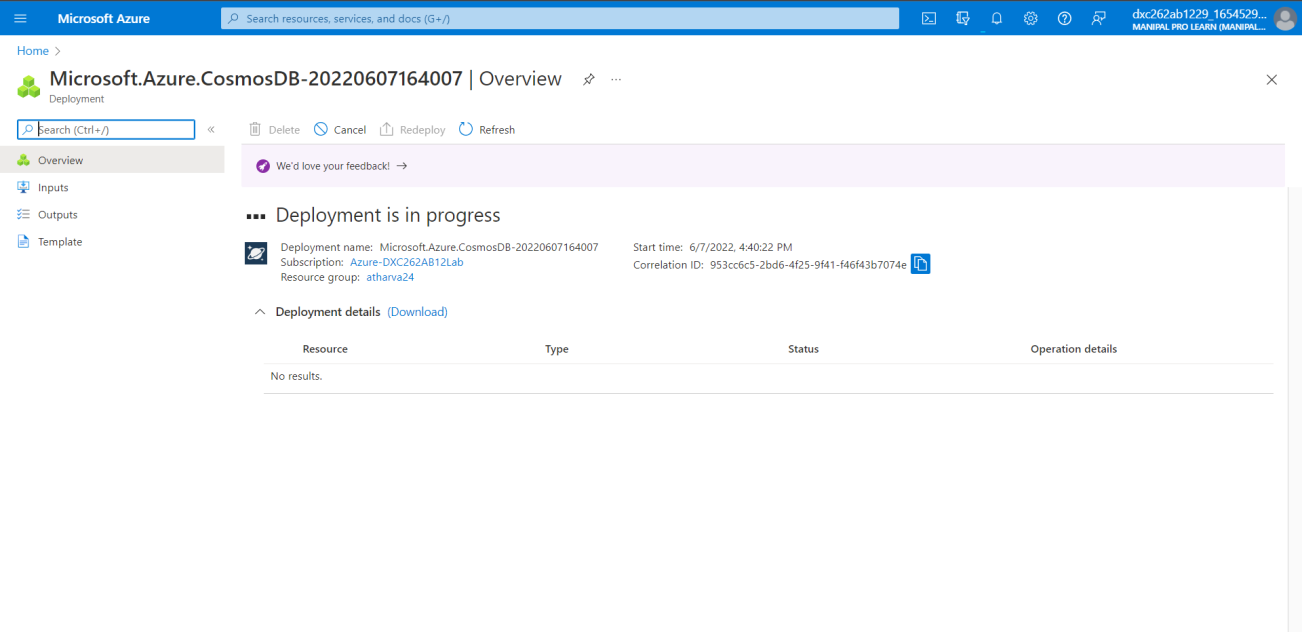
**Step 3—**

****

**Step 4—**

****

**Step5—**

****

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

**Step 1—**

**Go to dataexplorer**

**Step 2—**

**Click on new container**

**Step 3—**

**Give database id, container id**

**Step 4—**

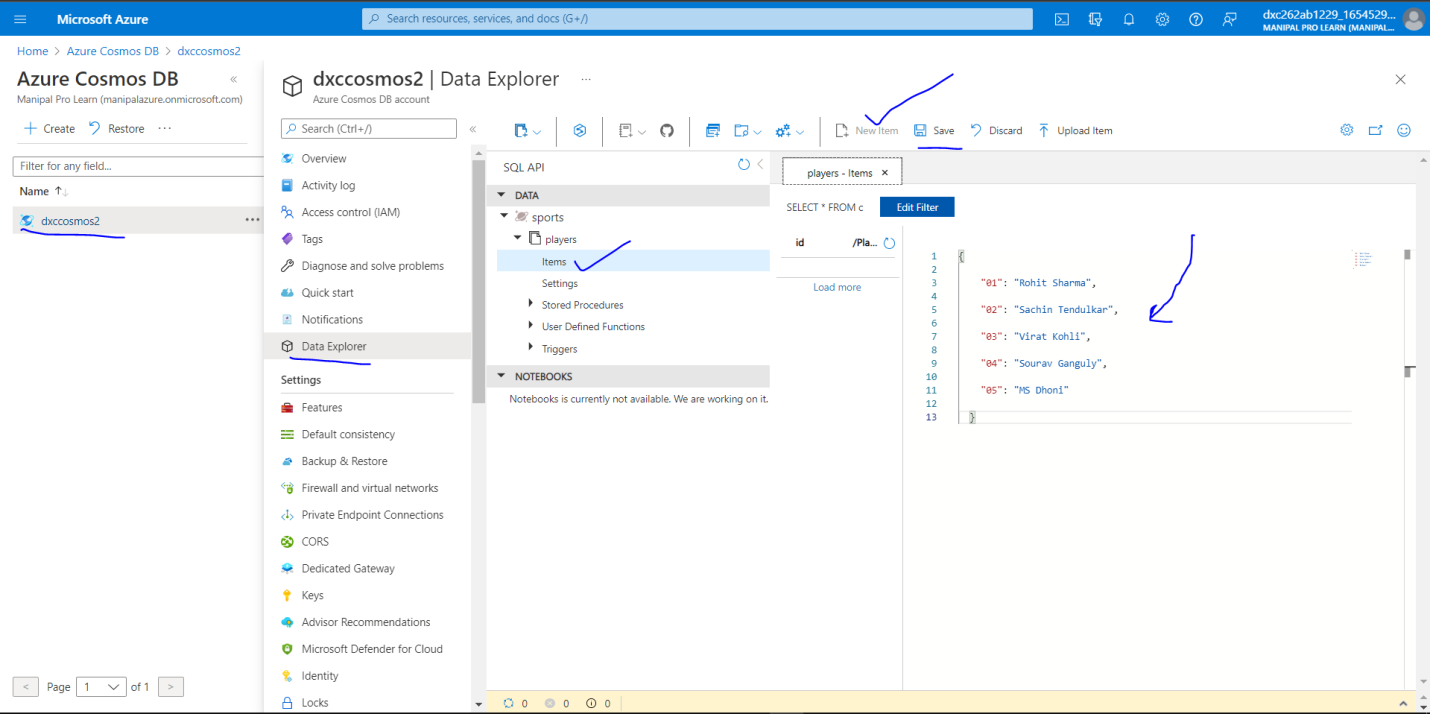
**Go to dropdown – data, click on items**

**Step5—**

**Click on new item**

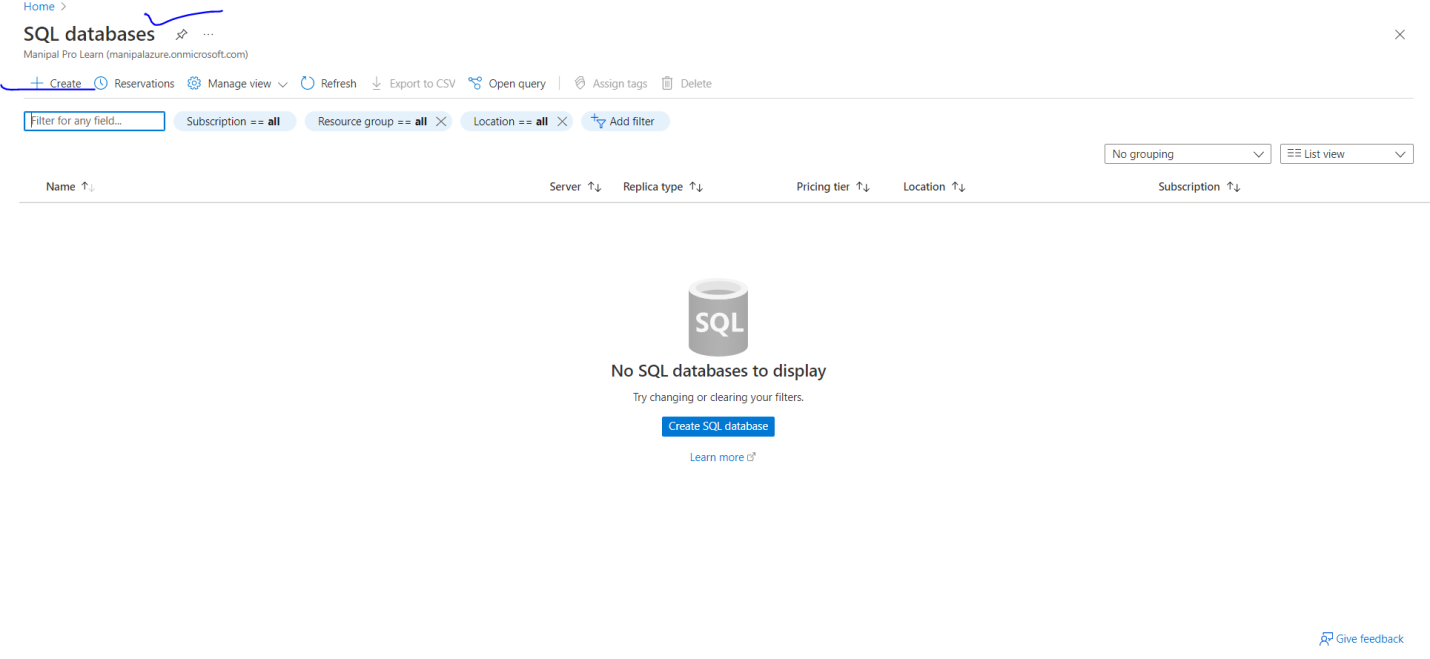
**Step 6 –**

**Add data and click on save**

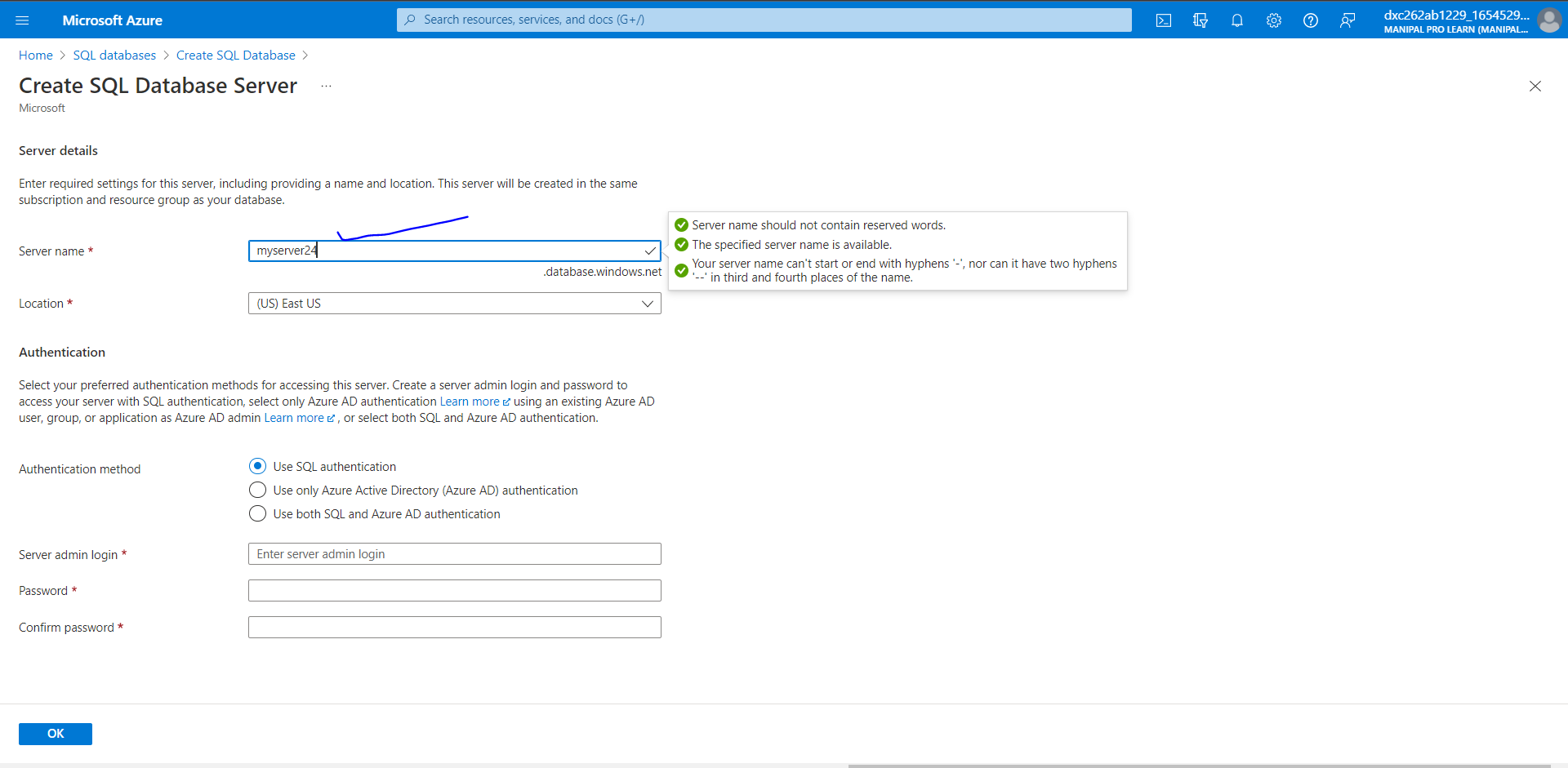
****

**10. Explain with sceenshots & step how to create Azure SQL Db & also explain how to**

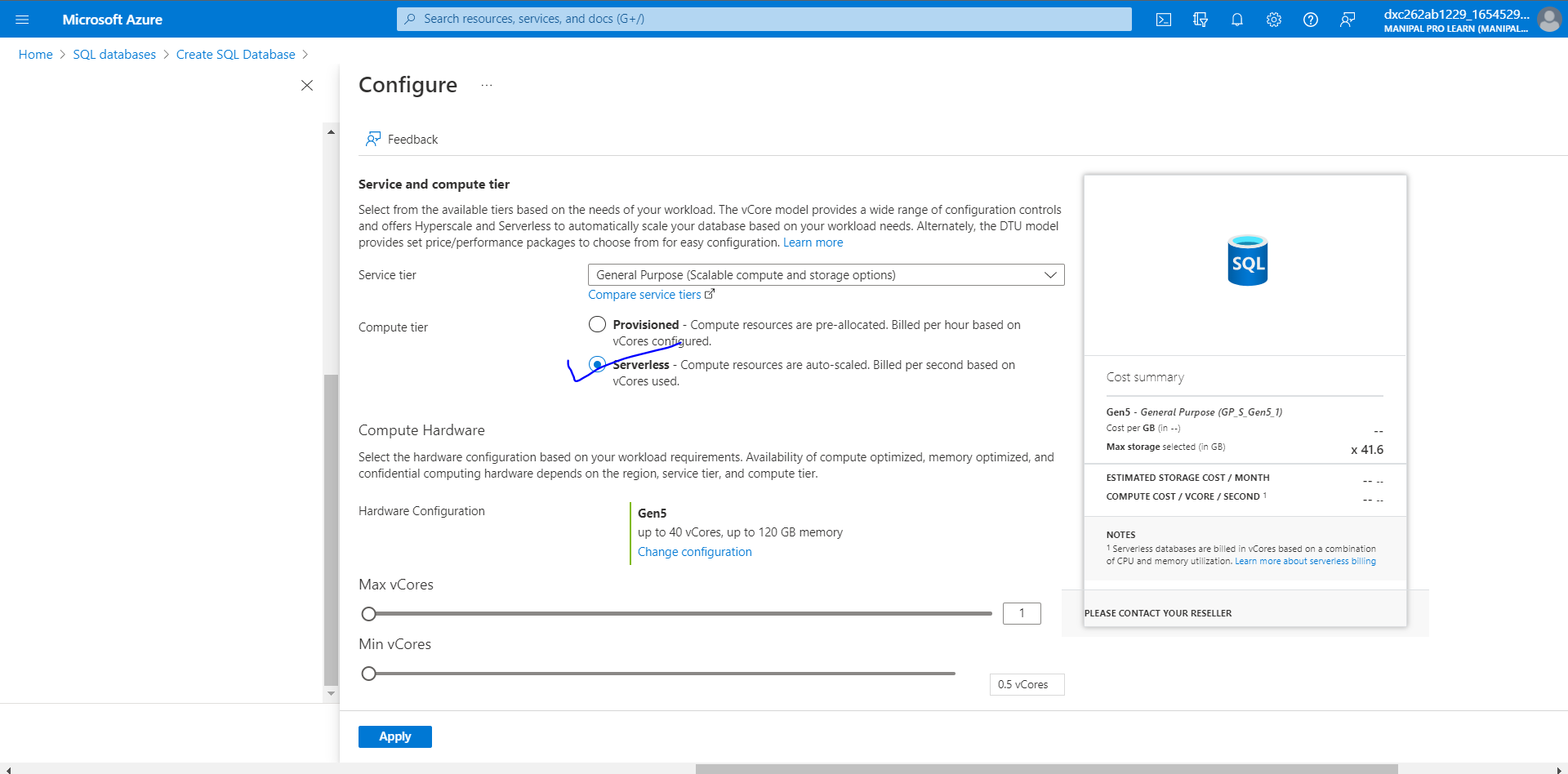
**insert data into Azure SQL D?**

**step 1 – **

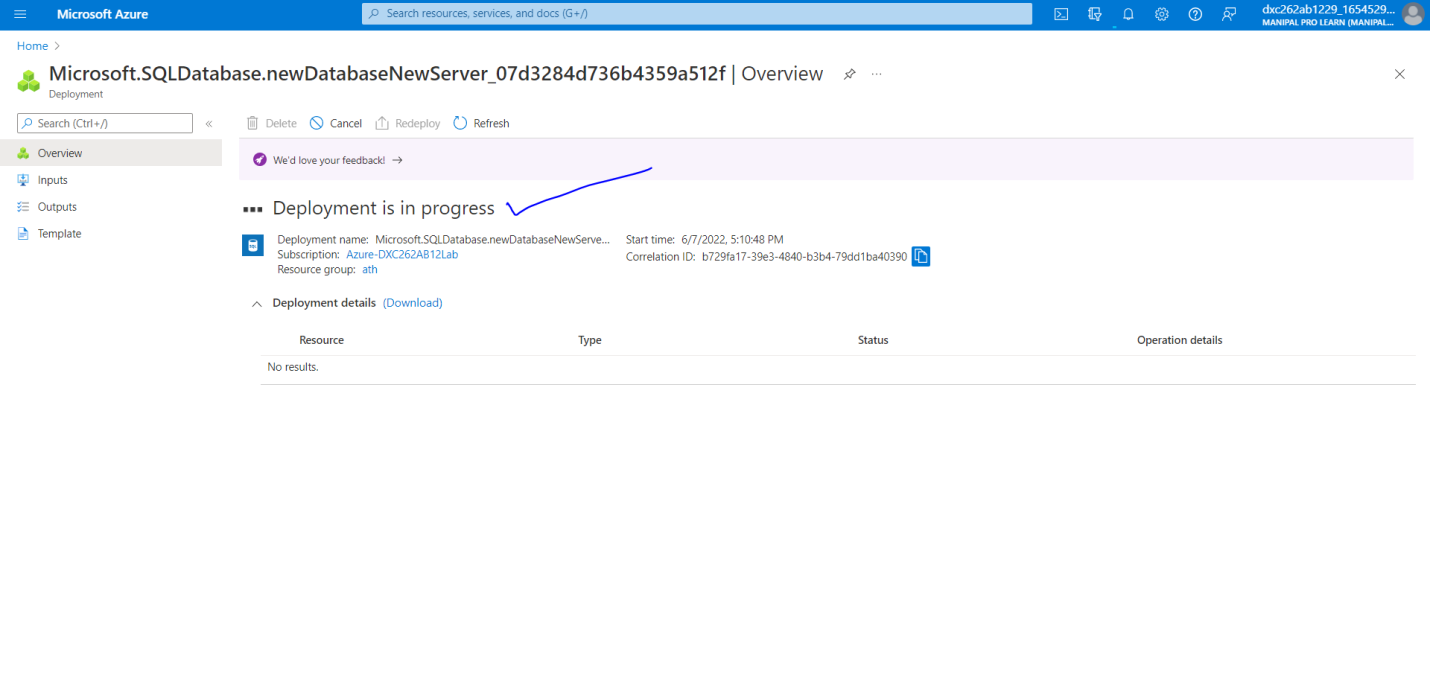
**step2—**

****

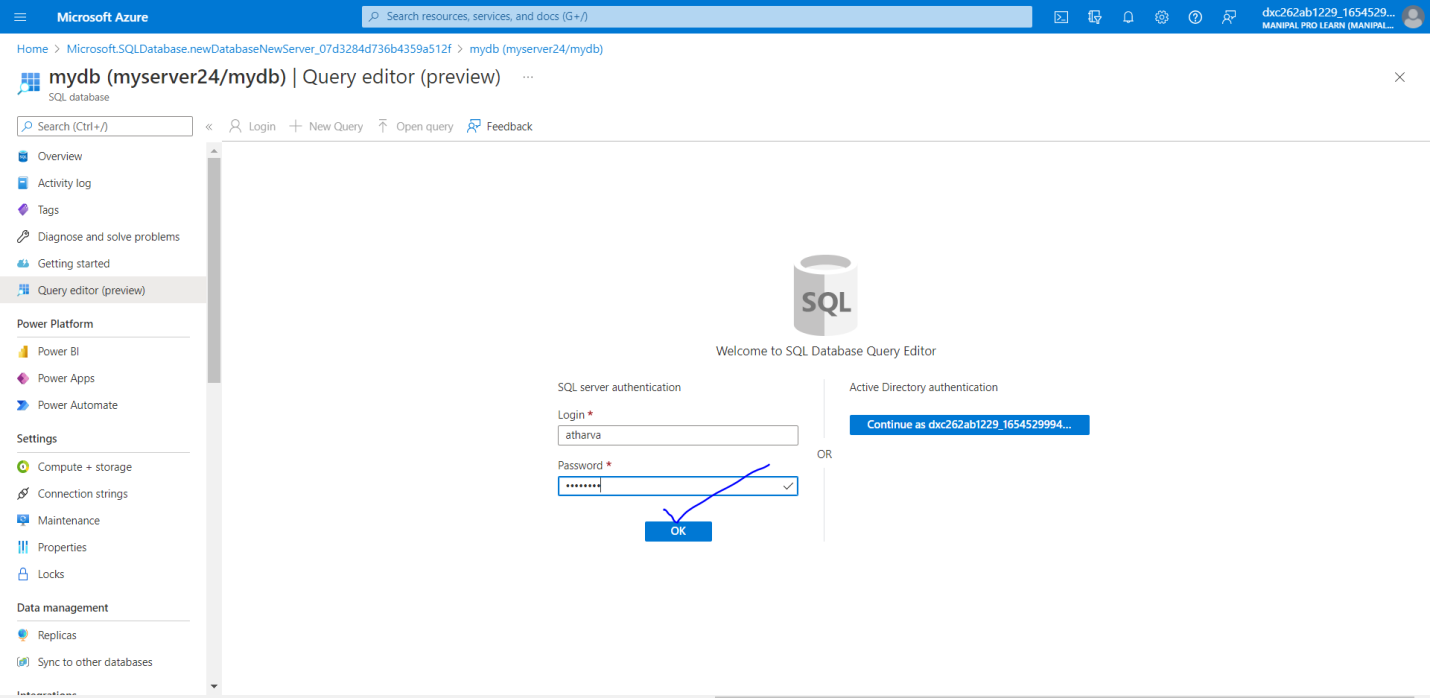
**Step 3—**

****

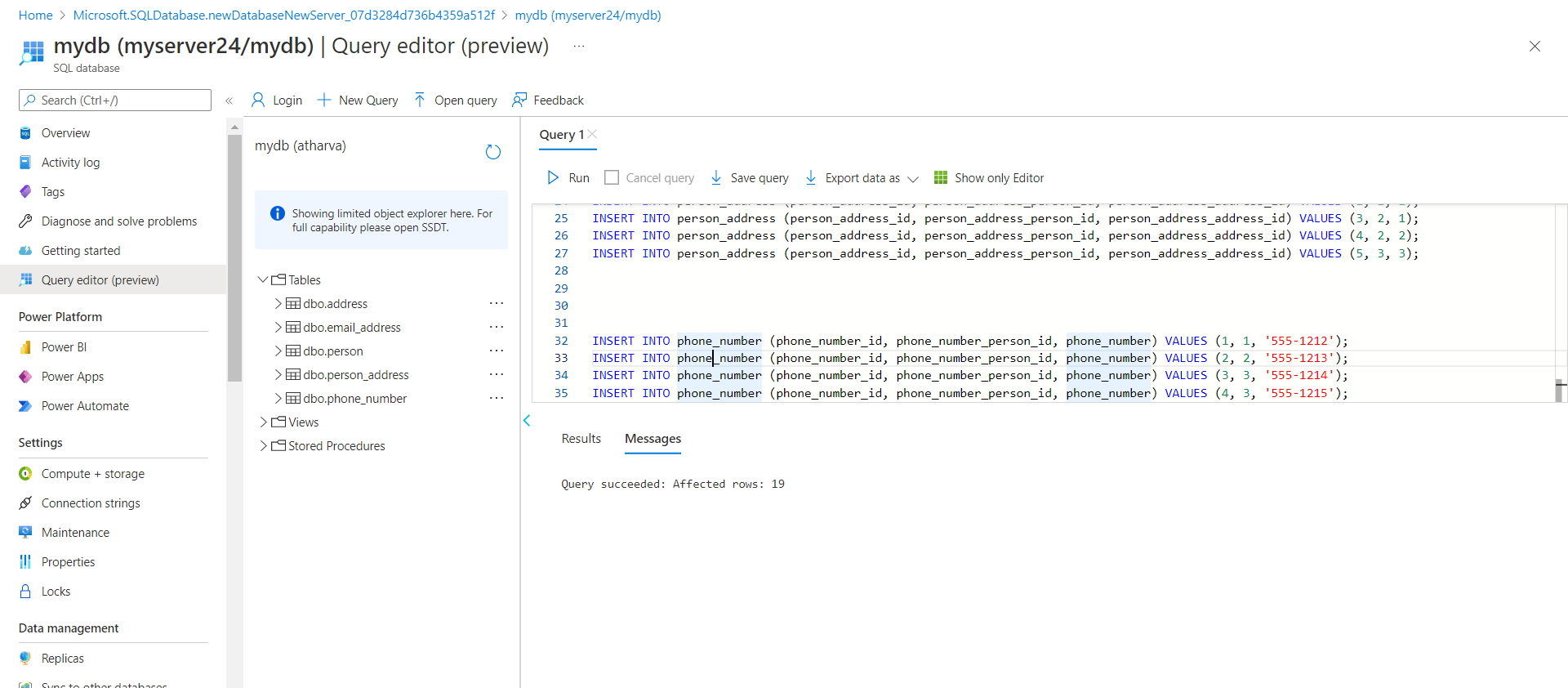
**Step4—**

****

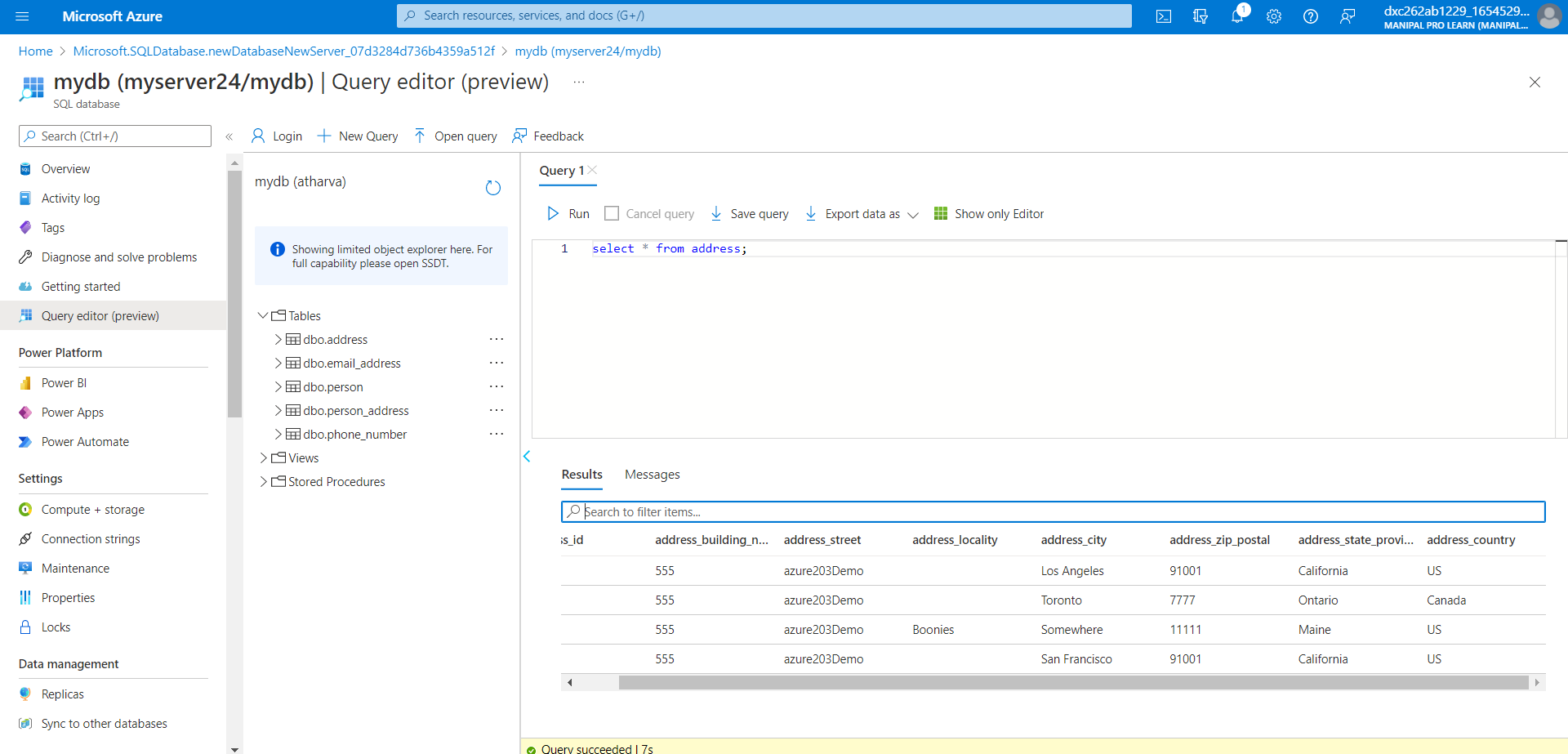
**Step – 5**

****

**Step 6 –**

****

**Step 7—**

****