

## ASSIGNMENT 7

NAME – JYOTHI CHANDANA VOLETI  
BATCH – DXC-262-ANALYTICS-B12-AZURE  
EMPLOYEE DOMAIN –AZURE ANALYTICS  
TRAINING UNDER – MANIPAL PRO LEARN  
DATE OF SUBMISSION – 7TH JUNE 2022

ROLL NUMBER – DXC-262-AB-1218  
COMPANY – DXC TECHNOLOGY  
TRAINER NAME – MR. AJAY KUMAR  
NO.OF QUESTIONS: 10

**1. Explain what are various components of SPARK with block diagrams? Explain the functionality of every component?**

**Ans: Components of Spark:**

**Block Diagram:**

### Components of Apache Spark



**Spark Core:** Spark core is the base engine for large scale parallel and distributed data processing.

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

**Spark Streaming:** Spark Streaming is a lightweight API that allows developers to perform batch processing and real-time streaming of data with ease.

**Spark MLlib:** MLlib is a low-level machine learning library that is simple to use, is scalable and compatible with various programming languages.

**GraphX:** GraphX is Spark's own Graph Computation Engine and data store.

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

**Ans: Spark Core:**

Spark core is the base engine for large scale parallel and distributed data processing.

It is responsible for:

- Memory Management
- Interacting with storage systems
- Fault Recovery
- Scheduling, distributing and monitoring jobs on a cluster

Spark Core is embedded with RDDs( Resilient Distributed Dataset), an immutable fault-tolerant, distributed collection of objects that can be operated on in parallel.

There are two ways to create RDDs – parallelizing an existing collection in the driver program, and referencing a dataset in an external storage system.

- *Resilient*: Fault tolerant and is capable of rebuilding data on failure
- *Distributed*: Distributed data among the multiple nodes in a cluster
- *Dataset*: Collection of partitioned data with values

## **3. Explain various MLLib algorithms Spark is supporting ?**

**Ans: Spark MLLib:** MLLib is a low-level machine learning library that is simple to use, is scalable and compatible with various programming languages.

**Algorithms MLLib Spark is supporting are:**

- Clustering
- Classification
- Collaborative Filtering
- Regression

MLlib supports various methods for binary classification, multiclass classification, and regression analysis.

1. Binary Classification supports linear SVMs, logistic regression, decision trees, random forests, gradient-boosted trees, naive Bayes methods.
2. Multiclass Classification supports logistic regression, decision trees, random forests, naive Bayes methods.
3. Regression supports linear least squares, Lasso, ridge regression, decision trees, random forests, gradient-boosted trees, and isotonic regression methods.
- 4.

Collaborative Filtering is commonly used for recommender systems. These techniques aim to fill in the missing entries of a user-item association matrix.

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).

MLlib supports the following models:

- K-Means
- Gaussian mixture

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

**Ans: Benefits of Spark SQL:**

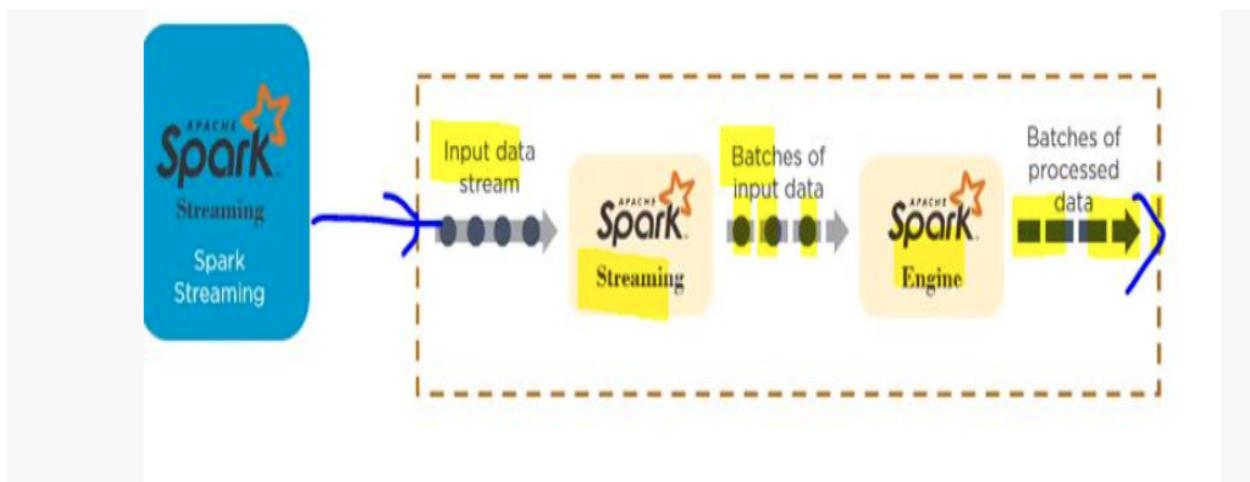
- Integrated
- Unified Data Access
- High Compatibility
- Scalability
- Standard Connectivity
- Performance Optimization

We can Use pandas and read the files using reading csv files by pandas.read\_csv(). And then can import the data files.

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

**Ans: Spark Streaming:** Spark Streaming is a lightweight API that allows developers to perform batch processing and real-time streaming of data with ease.

It provides secure, reliable and fast processing of live data streaming.



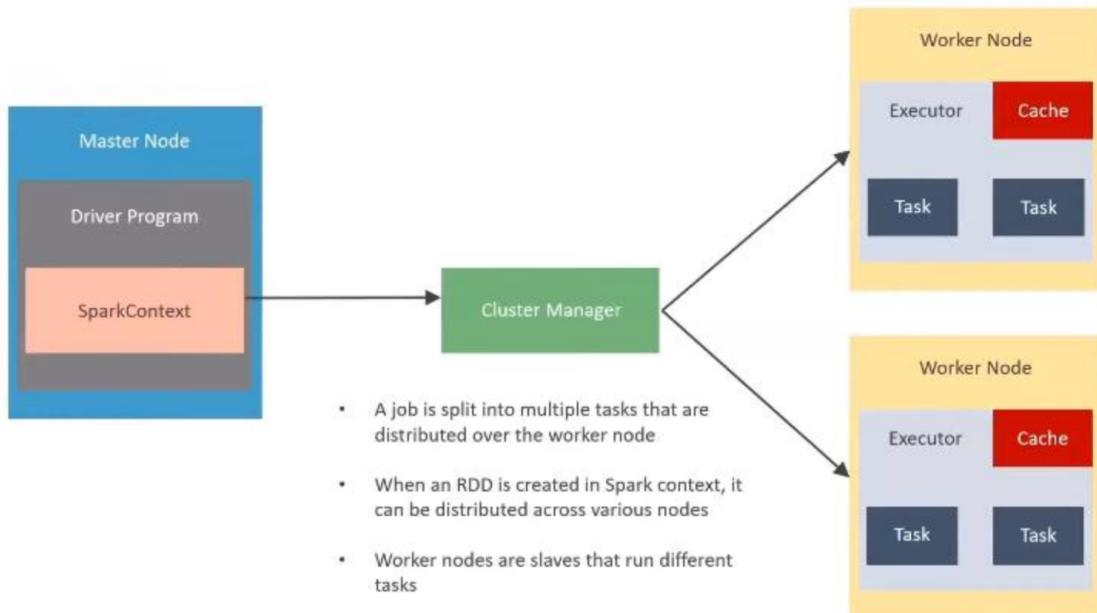
## 6. Explain SPARK architecture? what is Master - Slave architecure ?

Ans: SPARK Architecture:



### Master-Slave Spark Architecture:

## Spark Architecture



In master node, the *driver program*, drives your application. *Worker nodes* are the slave nodes whose job is to basically execute the tasks. These tasks are then executed on the partitioned RDDs in the worker node and hence returns back the result to the Spark Context.

## 7. Explain various cluster managers in SPARK?

**Ans:** **Spark Cluster Managers:**

**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.

**Mesos:** Apache Mesos is an open-source project to manage computer clusters, and can also run Hadoop applications.

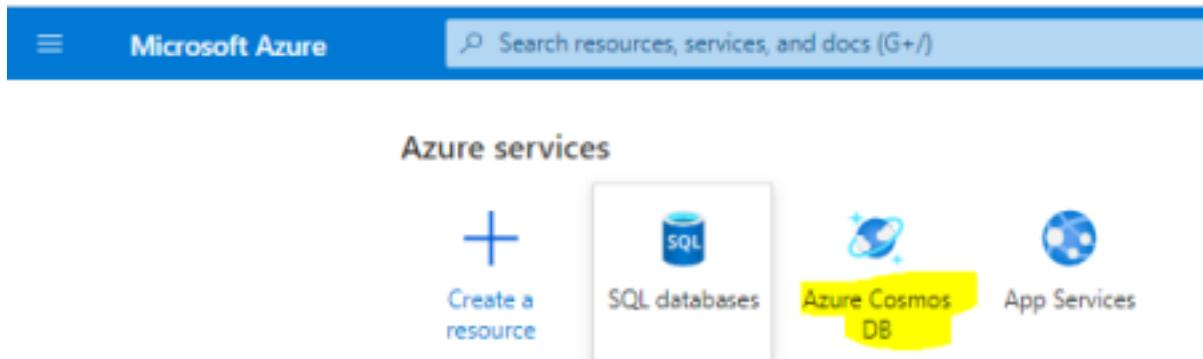
**Hadoop YARN:** Apache YARN is the cluster resource manager of Hadoop 2. Spark can be run on YARN.

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

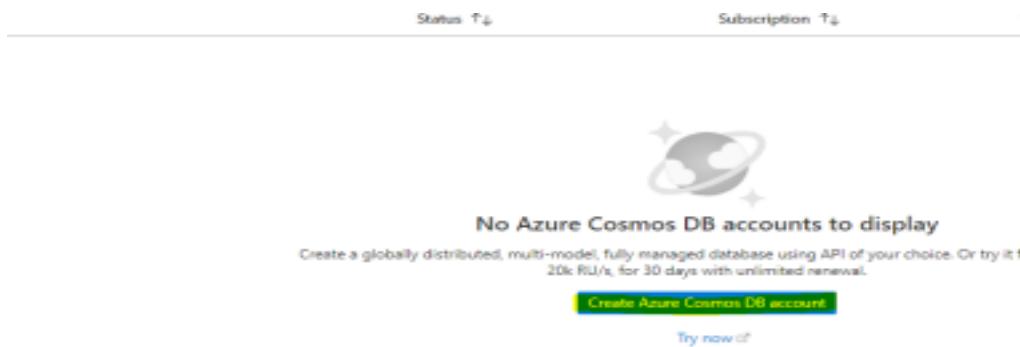
## 8. Explain with screenshots & steps how to create Cosmos DB ?

**Ans:**

**Step 1:** Go to Azure, and select “Azure Cosmos DB”.

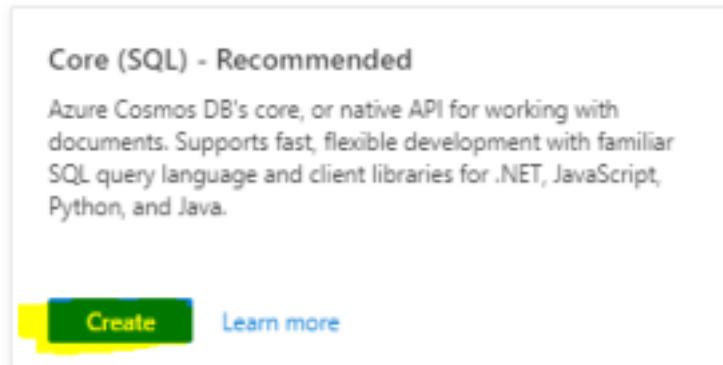


**Step 2:** Click on “Azure Cosmos DB”.



**Step 3:** Click on “Core(SQL)” “Create”.

To start, select the API to create a new account. The API selection cannot



**Step 4:** Follow steps in the screenshot.

The screenshot shows the "Project Details" step of the creation wizard. It includes fields for "Subscription" (selected: "Azure-DXC262AB12Lab"), "Resource Group" (selected: "(New) dcrcrgp235" with a yellow box around it), "Account Name" ("dcosmosdb235" with a yellow box around it), "Location" ("(US) East US" with a yellow box around it), "Capacity mode" (radio button selected: "Provisioned throughput" with a yellow box around it), and "Free tier" settings. At the bottom are buttons for "Review + create" (highlighted with a yellow box), "Previous", and "Next: Global Distribution".

## Step 5: Deployment is being done.

The screenshot shows the Azure portal interface for a deployment. At the top, there are buttons for Delete, Cancel, Redeploy, and Refresh. A feedback link is present. The main area displays deployment details: Deployment name: Microsoft.Azure.CosmosDB-20220607161248, Subscription: Azure-DXC262AB12Lab, Resource group: dxcrg235. The deployment status is "Deployment is in progress". Below this, a table shows deployment details with one row: "No results.". A yellow box highlights the deployment status message.

## Step 6: Go to Resources.

The screenshot shows the Azure portal interface after deployment completion. At the top, there are buttons for Delete, Cancel, Redeploy, and Refresh. A feedback link is present. The main area displays deployment details: Deployment name: Microsoft.Azure.CosmosDB-20220607161248, Subscription: Azure-DXC262AB12Lab, Resource group: dxcrg235. The deployment status is "Your deployment is complete". Below this, a table shows deployment details with one row: "No results.". A yellow box highlights the "Go to resource" button, which is blue and underlined.

## Step 7: Go to “Data Explorer”.

The screenshot shows the Azure portal interface for a Cosmos DB account named "dxcosmosdb235". The left sidebar lists various management options: Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Cost Management, Quick start, Notifications, and Data Explorer. The "Data Explorer" option is highlighted with a yellow box. The main content area is titled "dxcosmosdb235 | Data Explorer" and shows the SQL API blade. It includes a search bar, a "New Container" button, and an "Enable Azure Synapse" link. The "DATA" section is expanded, showing "CONTAINERS" and "NOTBOOKS". A note states: "Notebooks is currently not available. We are working on it." A yellow box highlights the "Data Explorer" link in the sidebar.

**Step 8:** Creating Database with details as shown.

The screenshot shows the 'Create database' wizard. The 'Database id' field is set to 'Cricket'. The 'Autoscale' option is selected for throughput. The 'Database Max RU/s' input field contains '1000'. Below it, a note states: 'Your database throughput will automatically scale from 100 RU/(10% of max RU/s) - 1000 RU/s based on usage.' Estimated monthly cost is listed as '\$8.76 - \$87.60'. The 'Container id' field is set to 'Player\_Name'. The 'Indexing' section shows 'Automatic' indexing is selected.

**9. Explain with screenshots & step how to insert data into Cosmos DB?**  
**Ans:Inserting data into Cosmos DB:**

**Step 1:** Select container “Cricket” “Player\_Name”.

The screenshot shows the Azure Cosmos DB portal. On the left, the navigation sidebar under 'DATA' shows 'Cricket' and 'Player\_Name' containers. The main area displays the message 'Welcome to Cosmos DB' and 'Globally distributed, multi-model database service for any scale'. At the bottom, there are three buttons: 'Launch quick start', 'New Container', and 'Connect'.

**Step 2:** Click on “items” and “New Item”.

The screenshot shows the Azure Cosmos DB SQL API Data Explorer interface. On the left, the navigation pane shows a database named 'axccosmosadb2317' and a container named 'Cricket'. Under 'Cricket', there is a 'Player\_name' collection with its 'Items' blade selected. A yellow box highlights the 'New Item' button at the top right of the main area. The results pane displays a table with columns 'id' and '/Playe...'. The first three rows of data are shown, with the third row's '\_self' field highlighted in red and containing the value 'replace\_with\_new\_document\_id'.

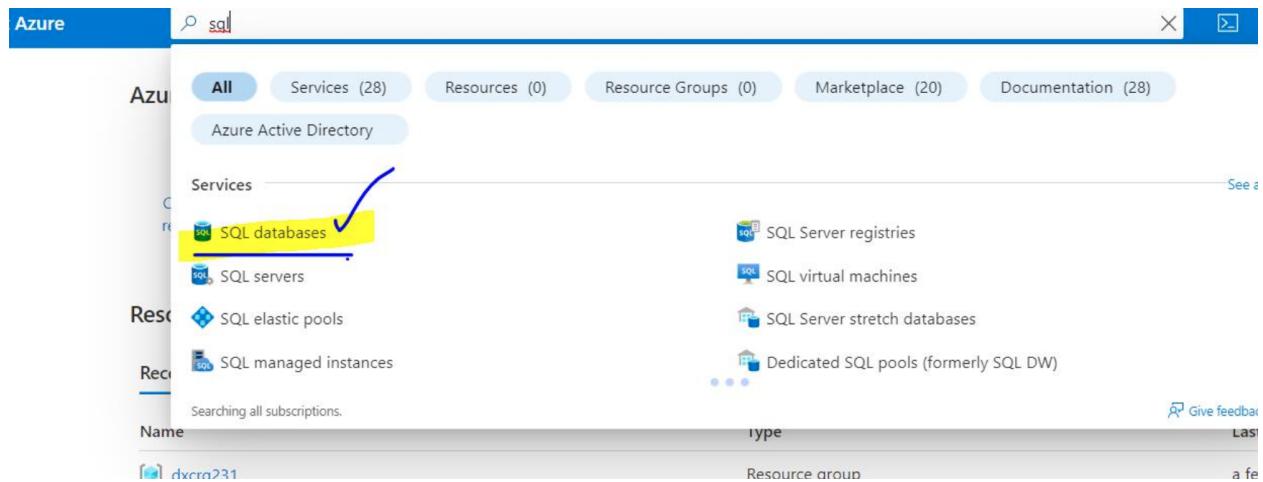
**Step 3:** Write the script shown below and then click on “Save”. And the data will be inserted.

The screenshot shows the Azure Cosmos DB SQL API Data Explorer interface. On the left, the navigation pane shows the same database and container setup as the previous screenshot. The 'cricketplayers' collection's 'Items' blade is selected. A blue arrow points from the 'Items' blade to the 'Execute Selection' button at the top right. The results pane shows a JSON array of player names. The first two items of the array are highlighted in yellow.

```
[{"id": "3838f1fc-ee0b-4150-849c-88bc64448d17", "name": "Rohit Sharma"}, {"id": "f-EFAJfh30MBAAAAAAA=", "name": "Sachin Tendulkar"}, {"id": "VlraN Kohli", "name": "Virat Kohli"}, {"id": "Sourav Ganguly", "name": "Sourav Ganguly"}, {"id": "MS Dhoni", "name": "MS Dhoni"}, {"id": "Md Shami", "name": "Md Shami"}, {"id": "3838f1fc-ee0b-4150-849c-88bc64448d17", "name": "Aaron Bird"}, {"id": "f-EFAJfh30MBAAAAAAA=", "name": "Aaron Finch"}, {"id": "Adam Gilchrist", "name": "Adam Gilchrist"}, {"id": "Andrew Symonds", "name": "Andrew Symonds"}]
```

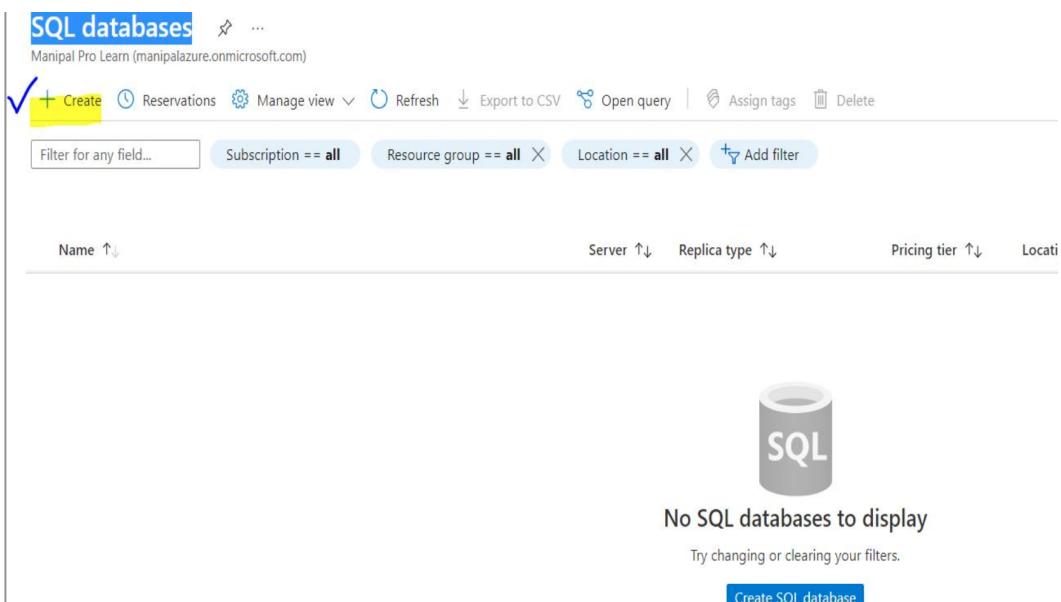
## 10. Explain with screenshots & step how to create Azure SQL Db & also explain how to insert data into Azure SQL Db?

**Ans:** Step1: Click on SQL Databases.



Azure search results for 'sql'. The 'SQL databases' option is highlighted with a yellow box and a blue arrow. Other options shown include 'SQL servers', 'SQL elastic pools', 'SQL managed instances', 'SQL Server registries', 'SQL virtual machines', 'SQL Server stretch databases', and 'Dedicated SQL pools (formerly SQL DW)'. A 'Give feedback' link is visible at the bottom right.

**Step2:** Click on Create.



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.				
<a href="#">Create SQL database</a>				

**Step3:** Click on all the options and review and create after configuring it to 1GB and Click on Create Db.

The screenshot shows the 'Create SQL Database' wizard in Microsoft Azure. The 'Review + create' tab is selected. The 'Product details' section shows an SQL database by Microsoft. The 'Estimated cost' section provides storage and compute details. The 'Terms' section contains legal terms and privacy statements. The 'Basics' section lists subscription, resource group, region, database name, server, authentication method, and server admin login. Compute and storage details are also listed. At the bottom, there are 'Create', 'Previous', and 'Download a template for automation' buttons.

**Step4:** Enter Credentials and login.

The screenshot shows the 'Query editor (preview)' interface for the newly created database. A red arrow points to the 'Query editor (preview)' link in the left sidebar. The main area displays a welcome message and authentication options for SQL server and Active Directory. The 'Login' field is filled with 'ajay' and the 'Password' field is highlighted with a yellow box and a red arrow. A blue 'OK' button is visible at the bottom right.

**Step5:** Click on ‘Quick Preview’ and Write a query and then click on “Run”. And the data is inserted.

The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface. The left sidebar displays the database structure for 'dxcsqlDb132 (ajay)'. The main area is titled 'Query 1' and contains the following SQL code:

```
1 CREATE TABLE address( address_id INTEGER NOT NULL, address_building_number VARCHAR(55) NOT NULL, address_line_1 VARCHAR(55) NOT NULL, address_line_2 VARCHAR(55), address_postal_code VARCHAR(55) NOT NULL, address_town VARCHAR(55) NOT NULL, address_type VARCHAR(55) NOT NULL ) 2 CREATE TABLE email_address( email_address_id INTEGER NOT NULL, email_address_person_id INTEGER, email_address_type VARCHAR(55) NOT NULL ) 3 CREATE TABLE person( person_id INTEGER NOT NULL, person_first_name VARCHAR(55) NOT NULL, person_last_name VARCHAR(55) NOT NULL, person_middle_name VARCHAR(55), person_nationality VARCHAR(55) ) 4 CREATE TABLE person_address( person_address_id INTEGER NOT NULL, person_address_person_id INTEGER NOT NULL, person_address_type VARCHAR(55) NOT NULL ) 5 CREATE TABLE phone_number( phone_number_id INTEGER NOT NULL, phone_number_person_id INTEGER NOT NULL, phone_number_type VARCHAR(55) )
```

The 'Messages' tab at the bottom shows the message: 'Query succeeded: Affected rows: 0'.