



Lab Manual- Azure Data Bricks Provisioning and Data Ingestion Part1

Prepared for:

Date: 18th March 2022

Prepared by:

Document Name: Lab Manual

Document Number AZLabn990

Contributor:

Contents

1.	Introduction.....	3
2.	Lab 1: Provision Azure Data Bricks	3
3.	Lab 2: Configure Databricks Cluster with Apace Spark	5
4.	Lab 3: Data Ingestion using Delta lake Table.....	8

1. Introduction

Systems are working with massive amounts of data in petabytes or even more and it is still growing at an exponential rate. Big data is present everywhere around us and comes in from different sources like social media sites, sales, customer data, transactional data, etc

[Apache Spark](#) is an open-source, fast cluster computing system and a highly popular framework for big data analysis. This framework processes the data in parallel that helps to boost the performance. It is written in [Scala](#), a high-level language, and also supports APIs for Python, SQL, Java and R.

Azure Databricks is the implementation of Apache Spark on Azure. With fully managed Spark clusters, it is used to process large workloads of data and also helps in data engineering, data exploring and also visualizing data using Machine learning.

In this Lab, you do the following:

- Provision Azure Data Bricks
- Configure Azure Data bricks Cluster
- Ingest Data to Azure Databricks with Delta Lake

2. Lab 1: Provision Azure Data Bricks

1. Sign in to Azure Portal
2. In the Search Bar type Databricks and select Azure Databricks
3. Click Create to Create Azure Databricks workspace
4. Use below parameter in the wizard
 - Resource Group : **Your Resource Group**
 - Virtual Machine Name : **Anyname**
 - Pricing Tier : **Standard**
 - Size : **Default**
 - Username : **VMAdmin**
 - Password : **Password@123**
 - Ports : **80,3389**

[Home](#) > [Azure Databricks](#) >

Create an Azure Databricks workspace ...

Basics Networking Advanced Tags Review + create

Project Details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * ⓘ	<input type="text" value="Visual Studio Enterprise Subscription"/>
Resource group * ⓘ	<input type="text" value="bipeen7676"/>
	Create new

Instance Details

Workspace name *	<input type="text" value="bipeen7676"/>
Region *	<input type="text" value="East US"/>
Pricing Tier * ⓘ	<input type="text" value="Standard (Apache Spark, Secure with Azure AD)"/>

Review + create

< Previous

Next : Networking >

5. Leave all option default and click Next

[Home](#) > [Azure Databricks](#) >

Create an Azure Databricks workspace ...

Basics **Networking** Advanced Tags Review + create

Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP) ⓘ ☐ Yes ☒ No

Deploy Azure Databricks workspace in your own Virtual Network (VNet) ☐ Yes ☒ No

6. Leave all option default and click Next

[Home](#) > [Azure Databricks](#) >

Create an Azure Databricks workspace ...

Basics Networking **Advanced** Tags Review + create

Enable Infrastructure Encryption ⓘ

☐

⚠The current pricing tier does not support infrastructure encryption.

7. Click **Review and create**

[Home](#) > [Azure Databricks](#) >

Create an Azure Databricks workspace ...

✓ Validation Succeeded

Basics Networking Advanced Tags **Review + create**

Summary

Basics

Workspace name	bipeendb
Subscription	Visual Studio Enterprise Subscription
Resource group	bipeen7676
Region	East US
Pricing Tier	standard

Networking

Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP)	No
Deploy Azure Databricks workspace in your own Virtual Network (VNet)	No

Advanced

Enable Infrastructure Encryption	No
----------------------------------	----

3. **Lab 2: Configure Databricks Cluster with Apace Spark**

- Click your workspace to open it.

bipeendb
Azure Databricks Service

Search (Ctrl+/) Delete

Overview

Activity log

Access control (IAM)

Tags

Settings

Virtual Network Peerings

Encryption

Properties

Locks

Automation

Tasks (preview)

Export template

Support + troubleshooting

New Support Request

Essentials

Status : Active

Resource group : [bipeen7676](#)

Location : East US

Subscription : [Visual Studio Enterprise Subscription](#)

Subscription ID : 463fbf22-369d-445d-b8c3-c9dbb477ee76


Tags ([edit](#)) : [Click here to add tags](#)

Managed Resource Group : [databricks-rg-bipeendb-pns3jgkzoufia](#)

URL : <https://adb-5375180450477845.5.azuredatabricks.net>

Pricing Tier : standard

[JSON View](#)



[Launch Workspace](#)

Documentation

Getting Started

Import Data from File

Import Data from Azure Storage

Notebook

Admin Guide

Link Azure ML workspace

- Click Launch workspace to open **Databricks studio**

Microsoft Azure | Databricks

Portal | shruti.sinha@hotmail.com

Get started

This is your home for all data science and engineering work.

We'll show you how to set up clusters, data and users.

Set up your workspace

Create a cluster

Ingest data

Invite your team

Next steps

Explore Notebook gallery

Read documentation

Data Science & Engineering

Notebook
Create a new notebook for querying, data processing, and machine...
[Create a notebook](#)

Data import
Quickly import data, preview its schema, create a table, and query it...
[Browse files](#)

Partner Connect
Fivetran, dbt
Tableau, Power BI
[View all partners](#)

Guide: Quickstart tutorial
Spin up a cluster, run queries on preloaded data, and display result
[Start tutorial](#)

Recents

Name

Last viewed

There are no recents yet

Documentation

Get started guide
This tutorial gets you going with Azure Databricks Data Science & Engineering

Best practices
Get the best performance when using Azure Databricks

Data guide
How to work with data in Azure Databricks

[More documentation](#)

Release notes

Runtime release notes

[Azure Databricks preview releases](#)

Platform release notes

[More release notes](#)

Blog posts

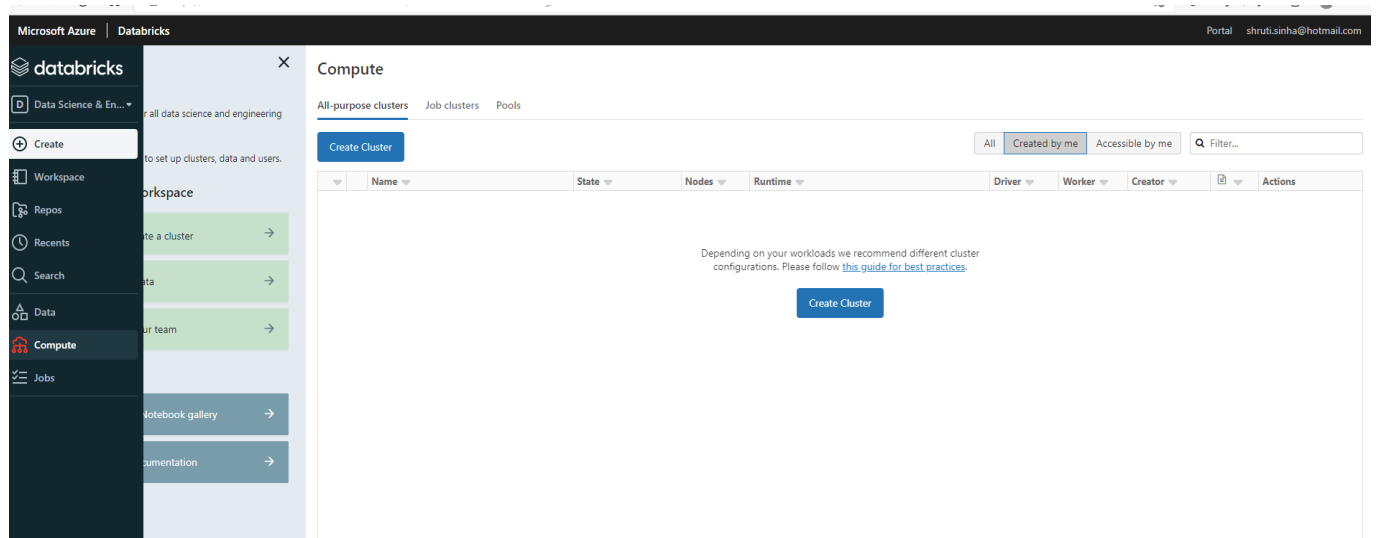
Speed Up Streaming Queries With Asynchronous State Checkpointing
May 2, 2022

Build Data and ML Pipelines More Easily With Databricks and Apache Airflow
April 30, 2022

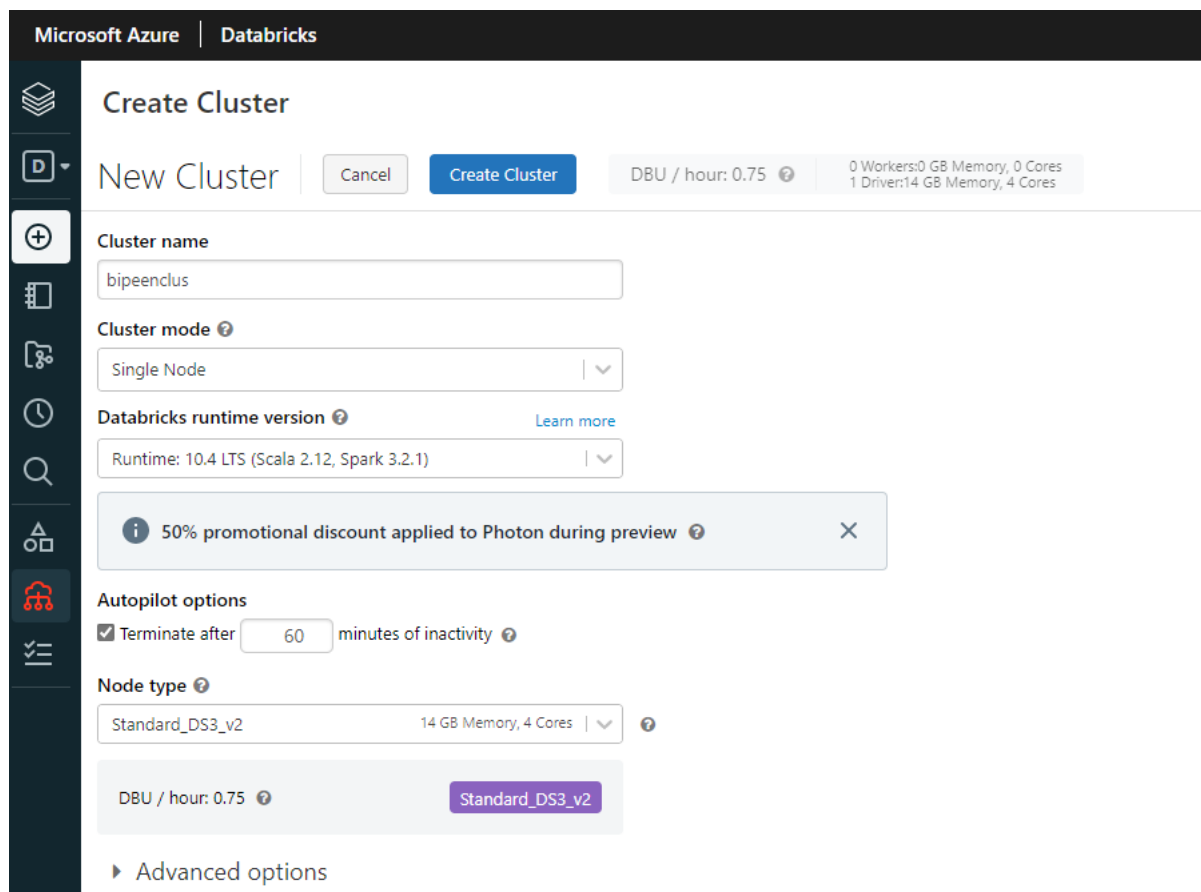
Implementing the GDPR 'Right to be Forgotten' in Delta Lake
March 23, 2022

[More blog posts](#)

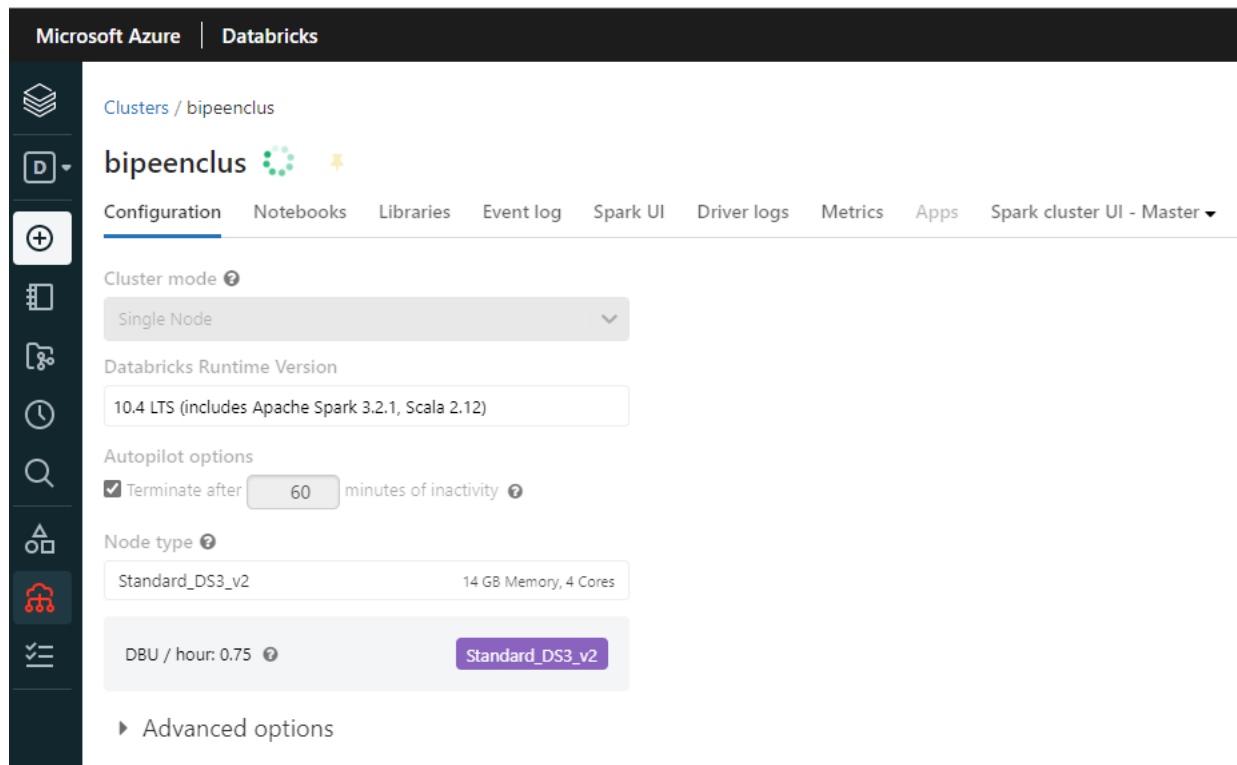
- In the Databricks Studio, Click **Compute** from left hand side menu



- Click Create Cluster and type your **cluster name**
- In the **cluster Node** select **Single Node**
- In the **terminate** after type **60**

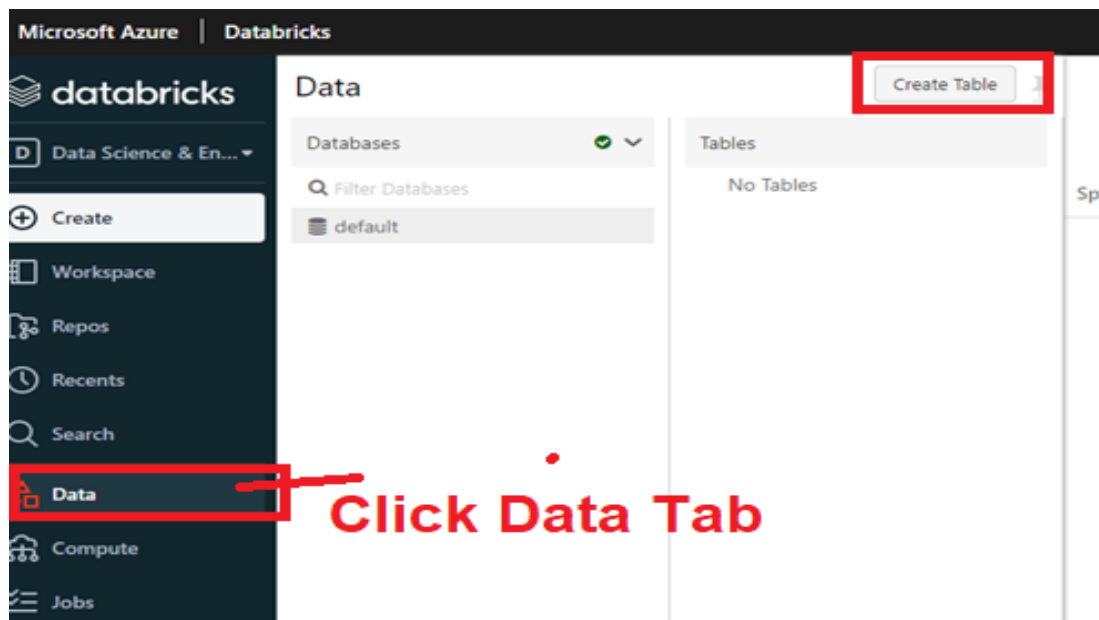


- Click **Create Cluster**. It will take 5-10 Minute

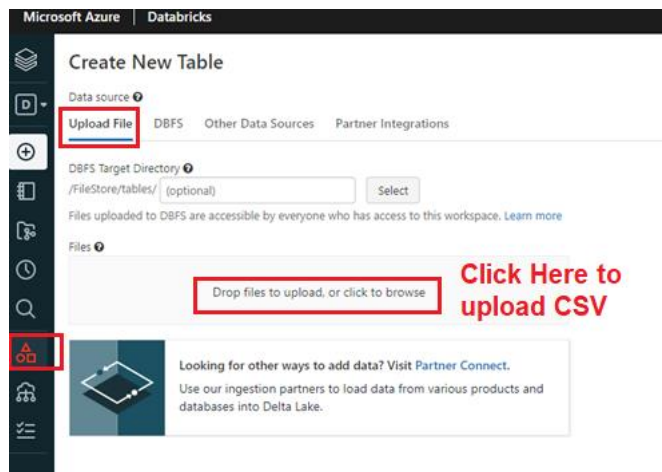


4. Lab 3: Data Ingestion using Delta lake Table

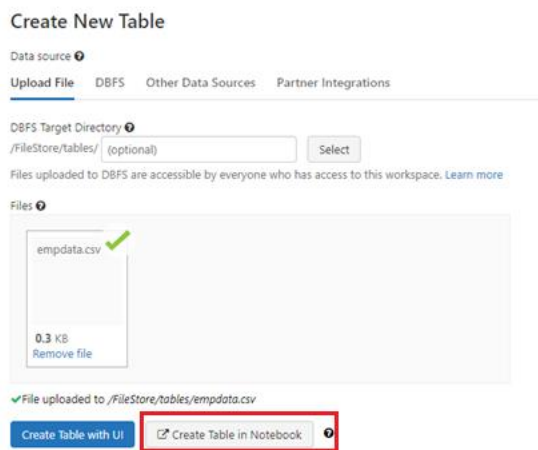
- Click **Data** menu from Left side Pane and Click **Create Table**



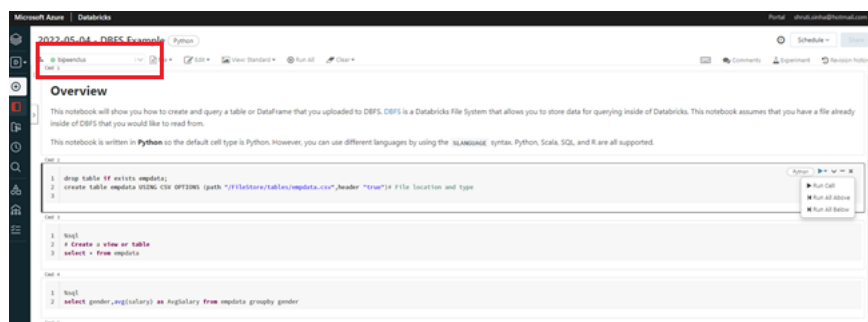
- Upload your CSV File



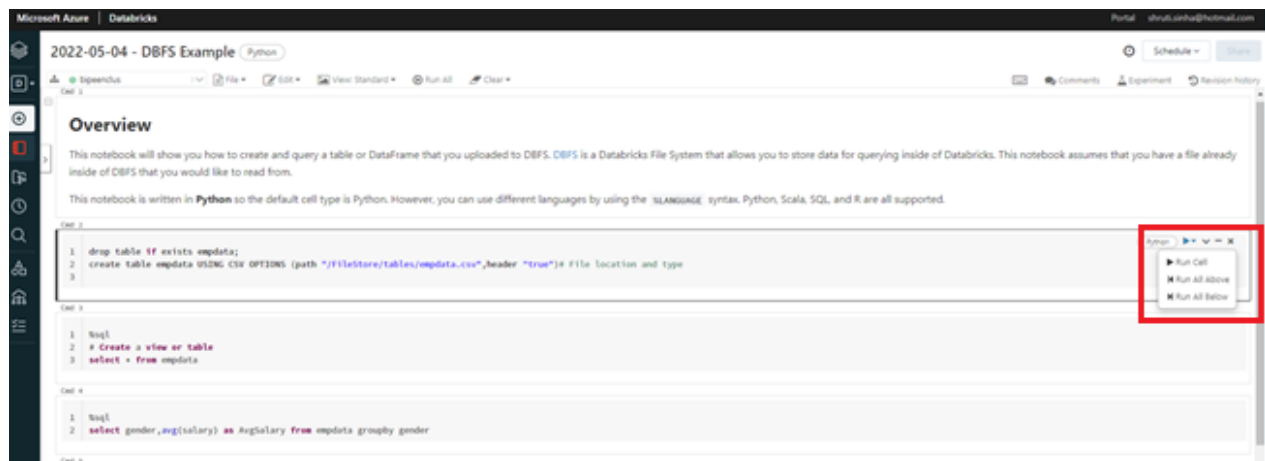
- Once your CSV Uploaded click **“Create Table in Notebook”**



- In The notebook first **attach your cluster**



- In The notebook in each cell, change the type from **Python** to **SQL**



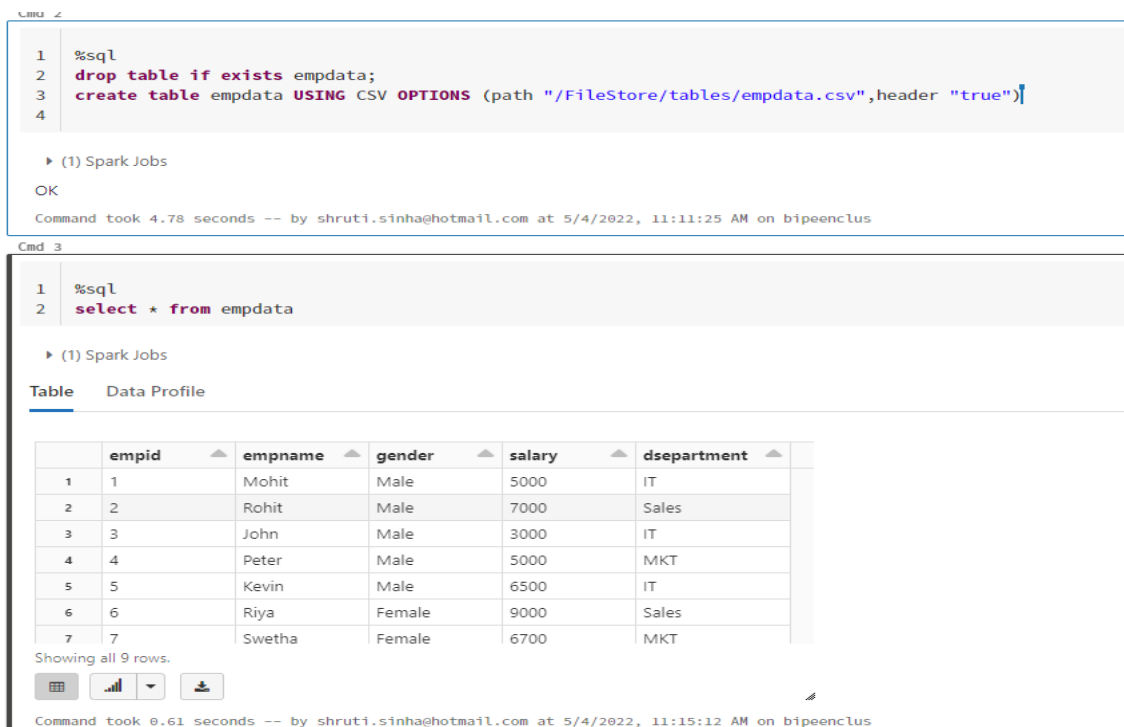
- In The notebook in **First Cell** Remove existing code and type below **SQL Script** to create Table and Click **Run Cell**

drop table if exists empdata;
create table empdata USING CSV OPTIONS (path "/FileStore/tables/empdata.csv",header "true")



- In The notebook in **2nd Cell** Remove existing code and type below SQL Script to show the data from just created table and click run cell. It will show the content of your CSV Data as Table

select * from empdata



- In The notebook in **3rd Cell** Remove existing code and type below SQL Script to show the data from just created table and click run cell. It will show the content of your CSV Data as Table with condition

select * from empdata where dsepartment='IT'

Cmd 4

```
1 %sql
2 select * from empdata where dsepartment='IT'
```

▶ (1) Spark Jobs

Table Data Profile

	empid	empname	gender	salary	dsepartment
1	1	Mohit	Male	5000	IT
2	3	John	Male	3000	IT
3	5	Kevin	Male	6500	IT
4	9	Jonson	Male	3000	IT

Showing all 4 rows.

Command took 0.45 seconds -- by shruti.sinha@hotmail.com at 5/4/2022, 11:27:23 AM on bipeenclus