

**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](#_Toc102632083)

[2. Lab 1: Provision Azure Data Bricks 3](#_Toc102632084)

[3. Lab 2: Configure Databricks Cluster with Apace Spark 5](#_Toc102632085)

[4. Lab 3: Data Ingestion using Delta lake Table 8](#_Toc102632086)

# 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](https://spark.apache.org/) 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](https://spark.apache.org/docs/0.9.1/scala-programming-guide.html), a high-level language, and also supports APIs for Python, SQL, Java and R.

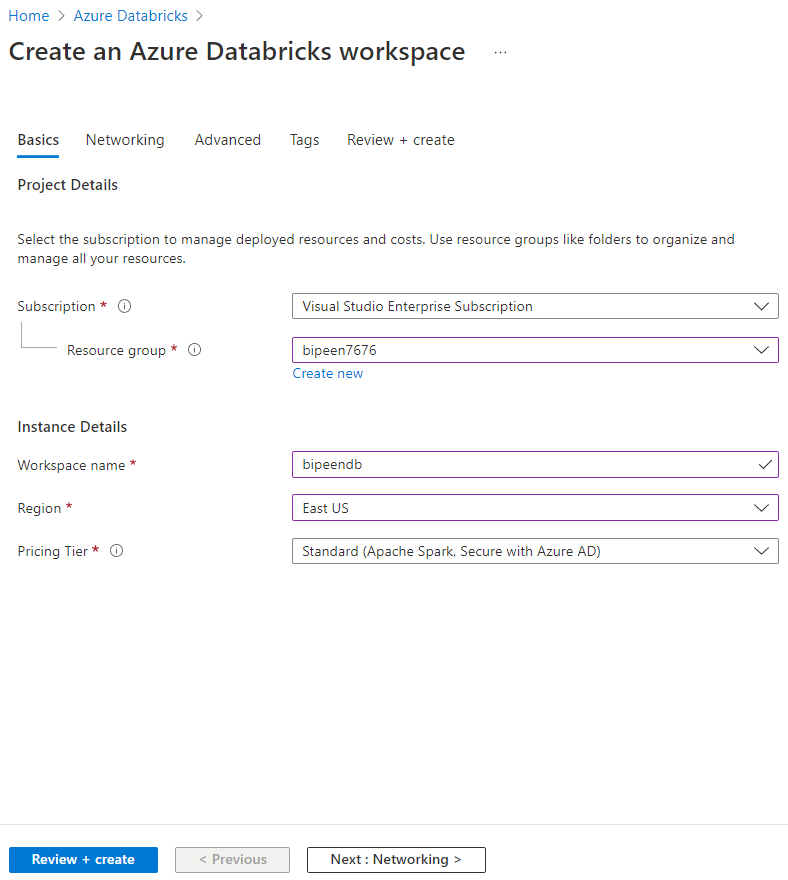
Azure Databricks is the implementation of Apache Spark on Azure. With fully managed Spark clusters, it is used to process large workloads of data and also helps in data engineering, data exploring and also visualizing data using Machine learning.

In this Lab, you do the following:

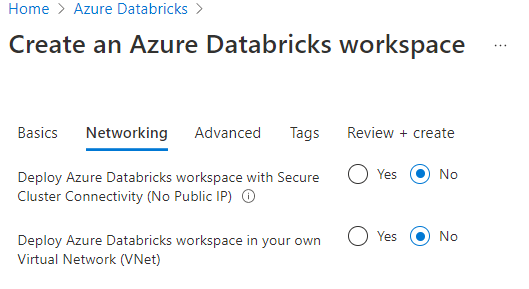
* Provision Azure Data Bricks
* Configure Azure Data bricks Cluster
* Ingest Data to Azure Databricks with Delta Lake

# Lab 1: Provision Azure Data Bricks

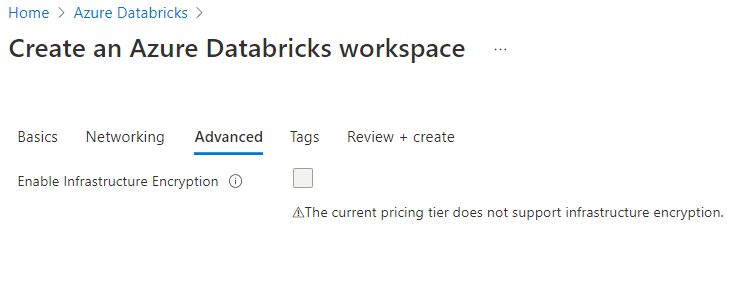
1. Sign in to Azure Portal
2. In the Search Bar type Databricks and select Azure Databricks
3. Click Create to Create Azure Databricks workspace
4. Use below parameter in the wizard
   * **Resource Group :** **Your Resource Group**
   * **Virtual Machine Name :** **Anyname**
   * **Pricing Tier :** 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



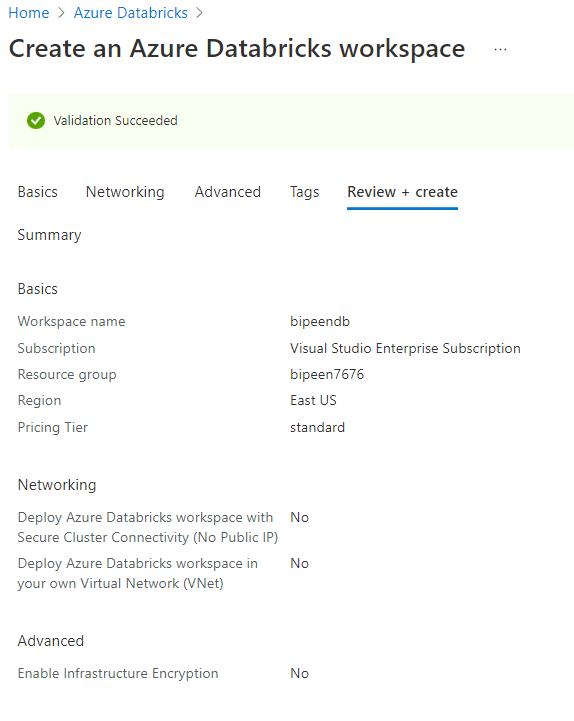
1. Leave all option default and click Next



1. Leave all option default and click Next

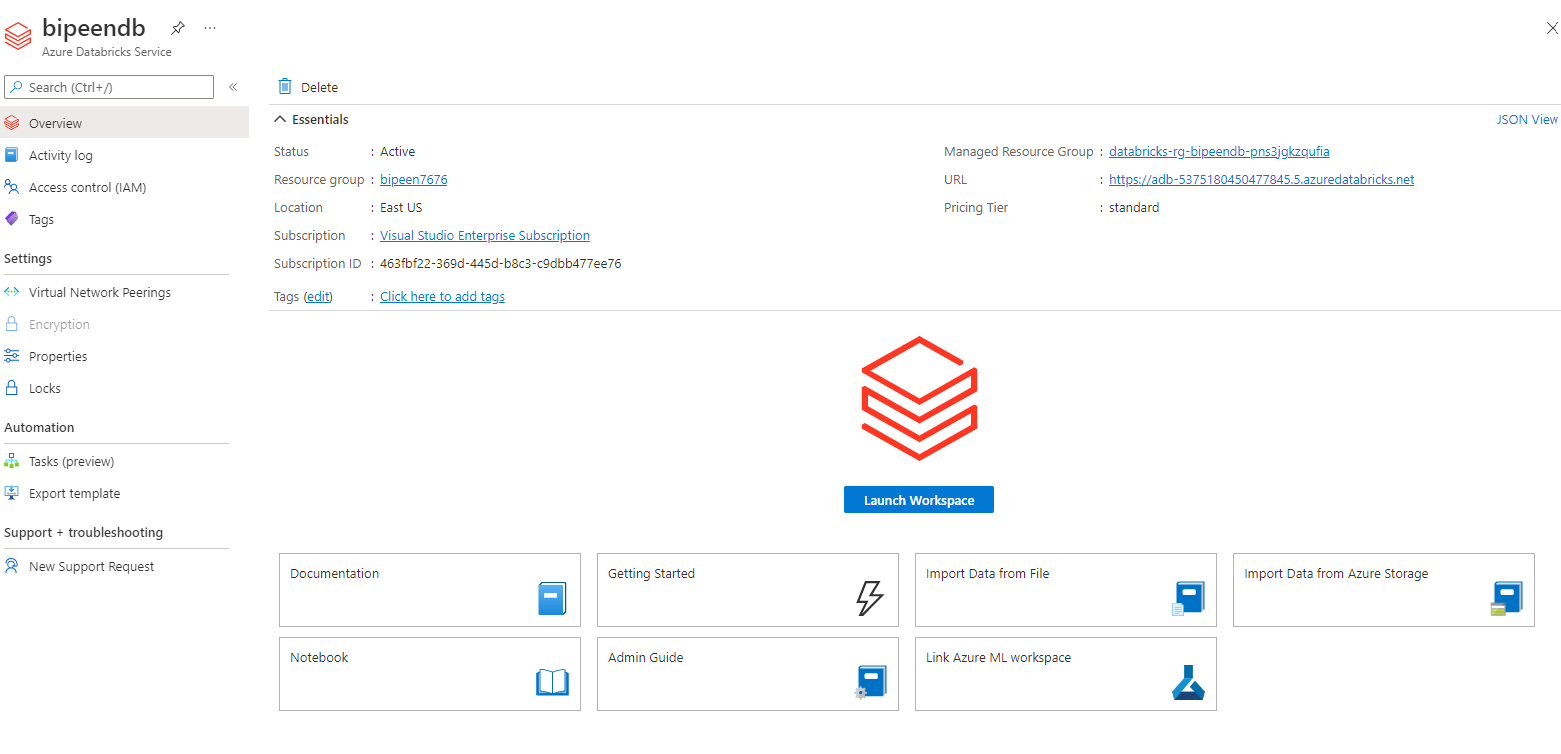


1. Click **Review and create**

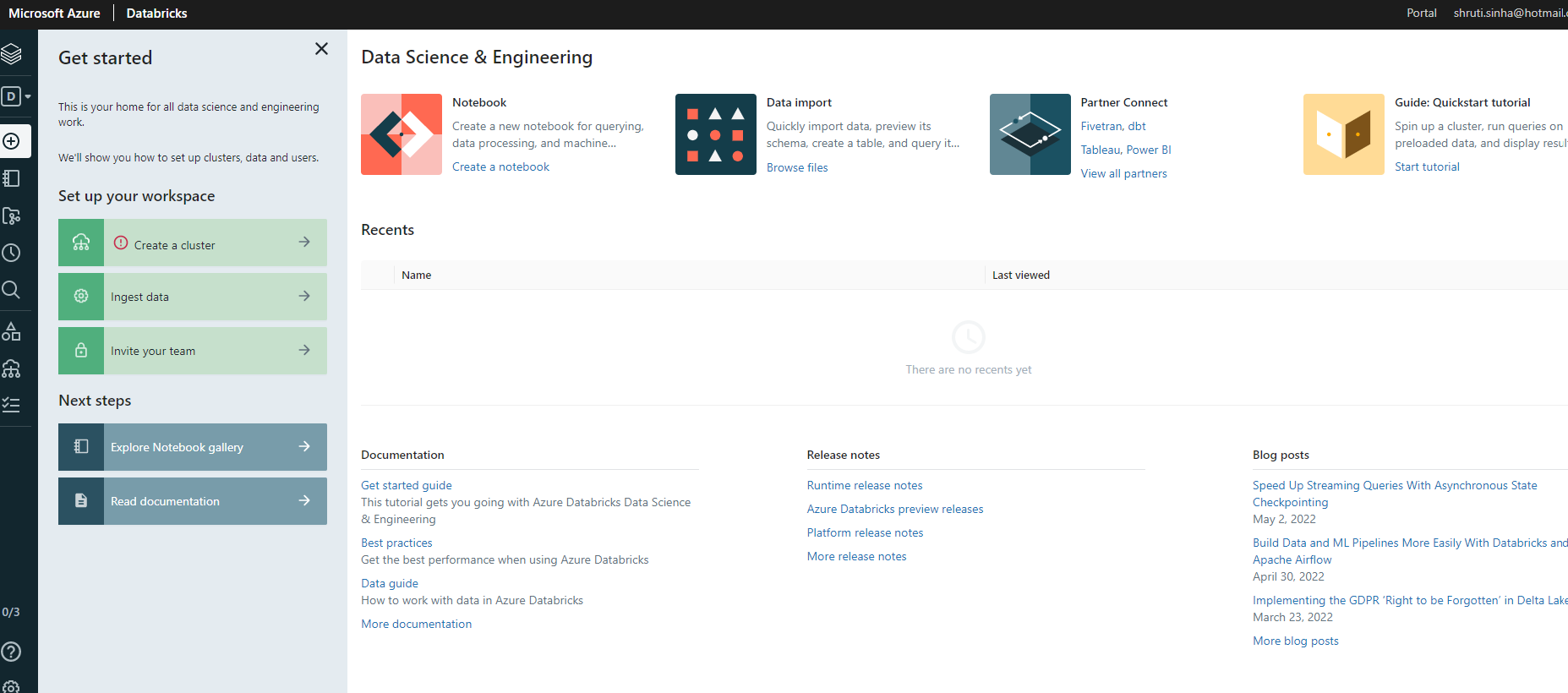


# Lab 2: Configure Databricks Cluster with Apace Spark

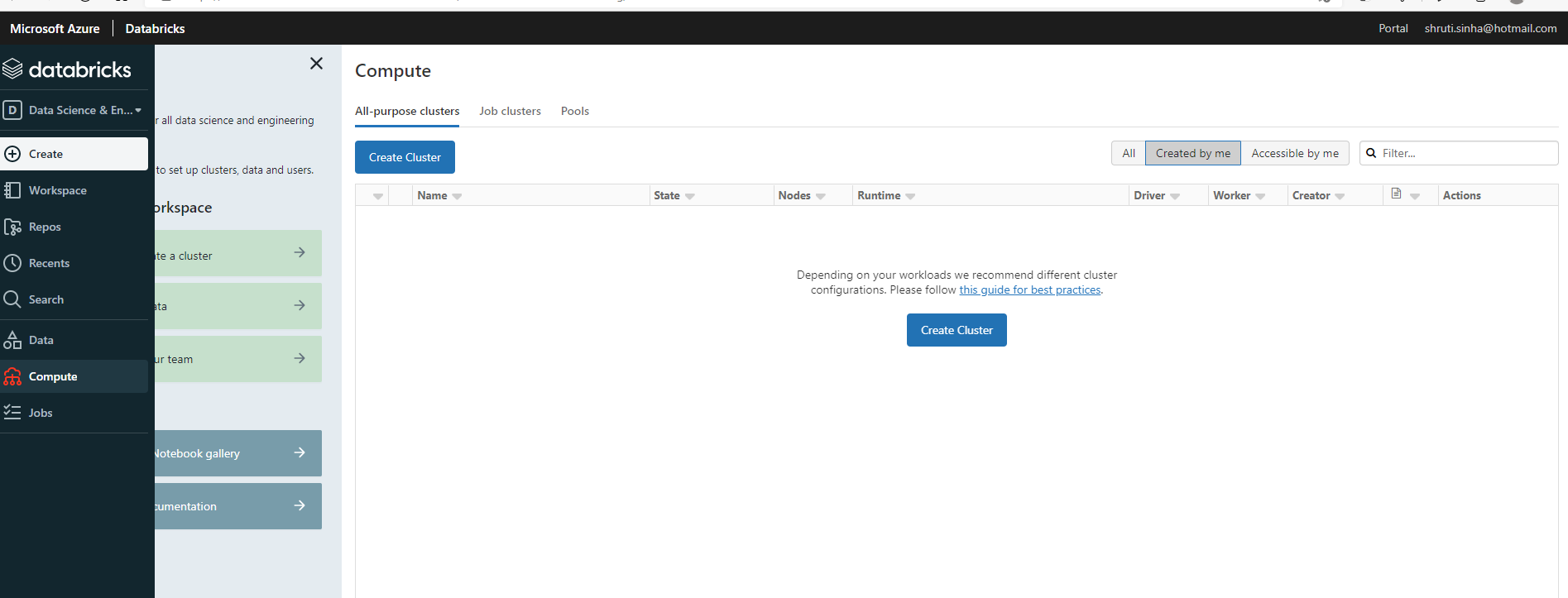
* Click your workspace to open it.



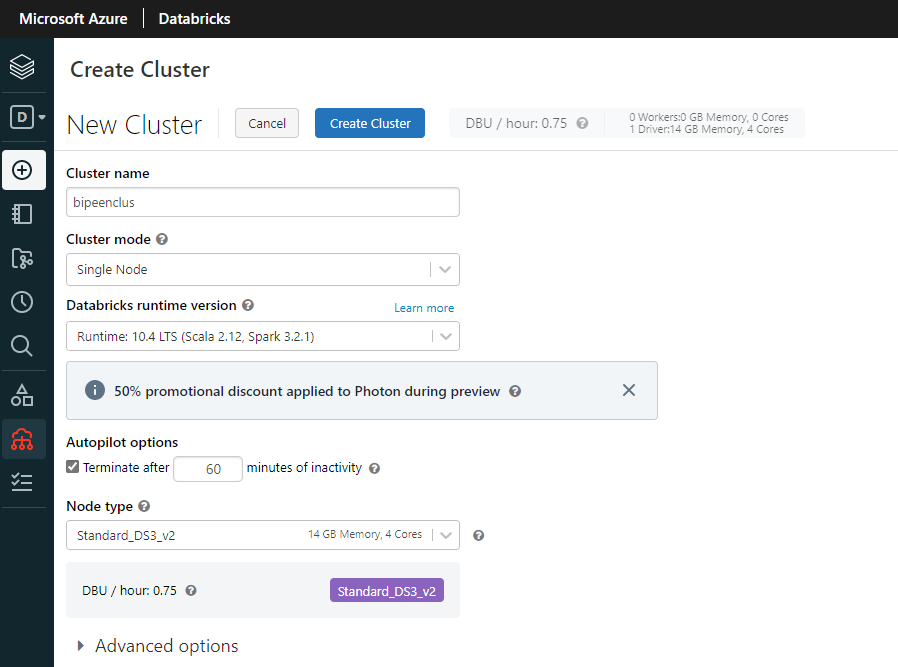
* Click Launch workspace to open **Databricks studio**



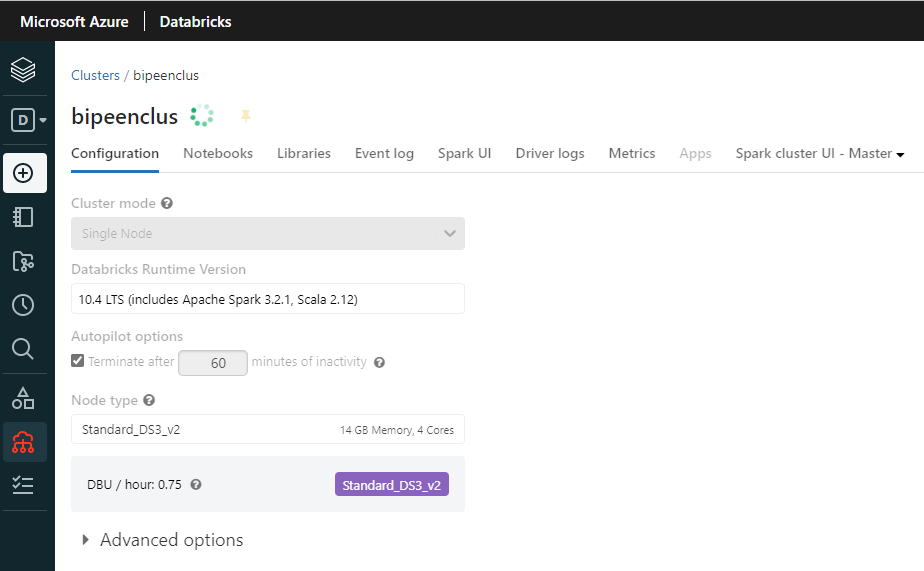
* In the Databricks Studio, Click **Compute** from left hand side menu



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

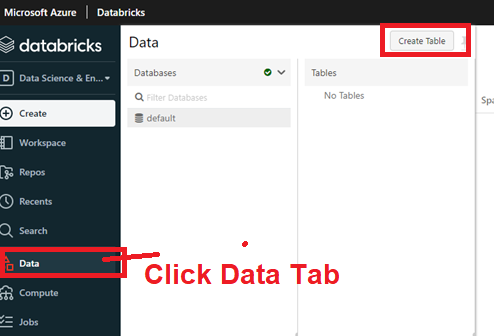


* Click **Create Cluster**. It will take 5-10 Minute

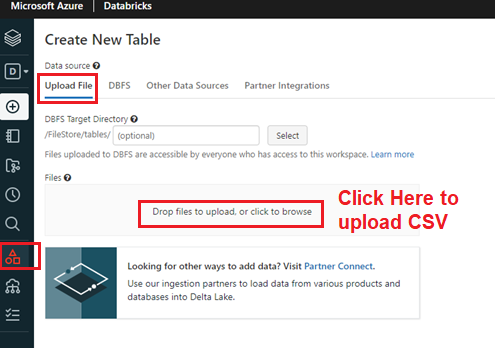


# Lab 3: Data Ingestion using Delta lake Table

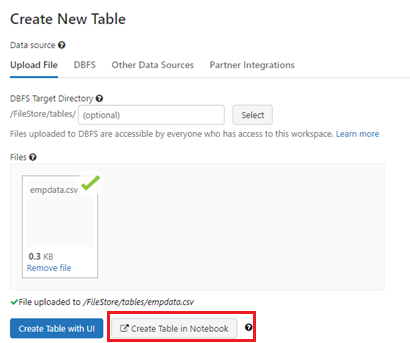
* Click **Data** menu from Left side Pane and Click **Create Table**



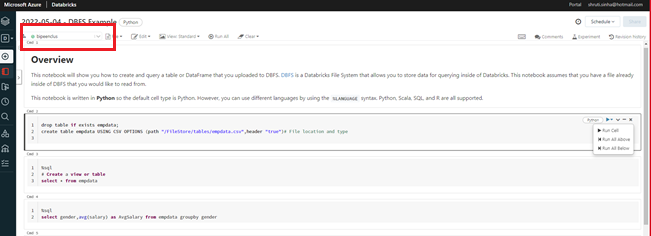
* **Upload** your CSV File



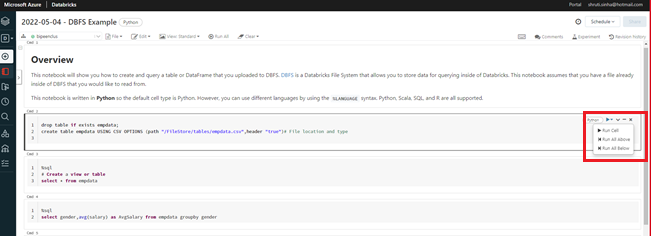
* Once your CSV Uploaded click “**Create Table in Notebook**”



* 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**

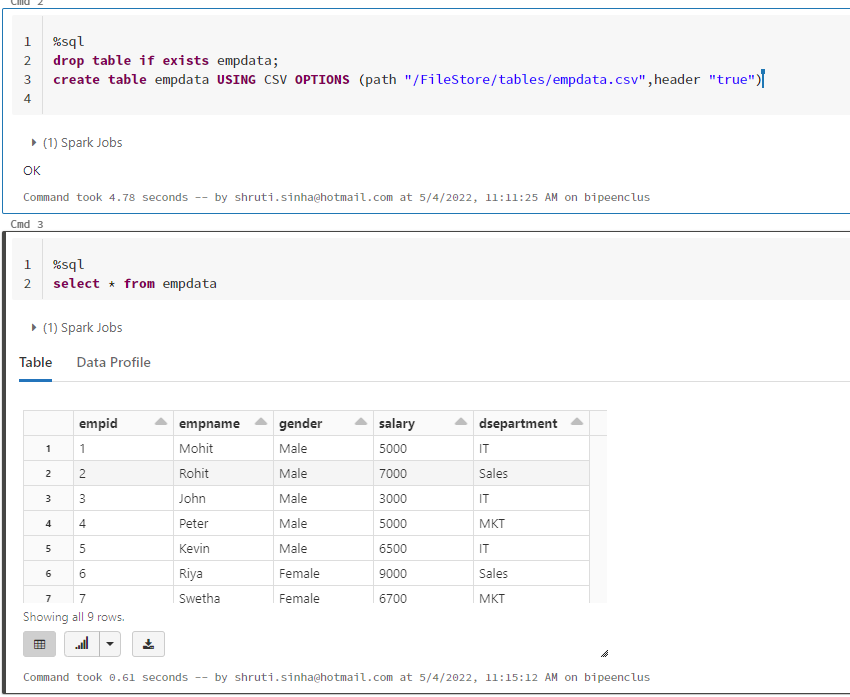
**drop table if exists empdata;**

**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'**

