**Name:** V Venkata Sri Prasad

**Batch:** Data Engineering

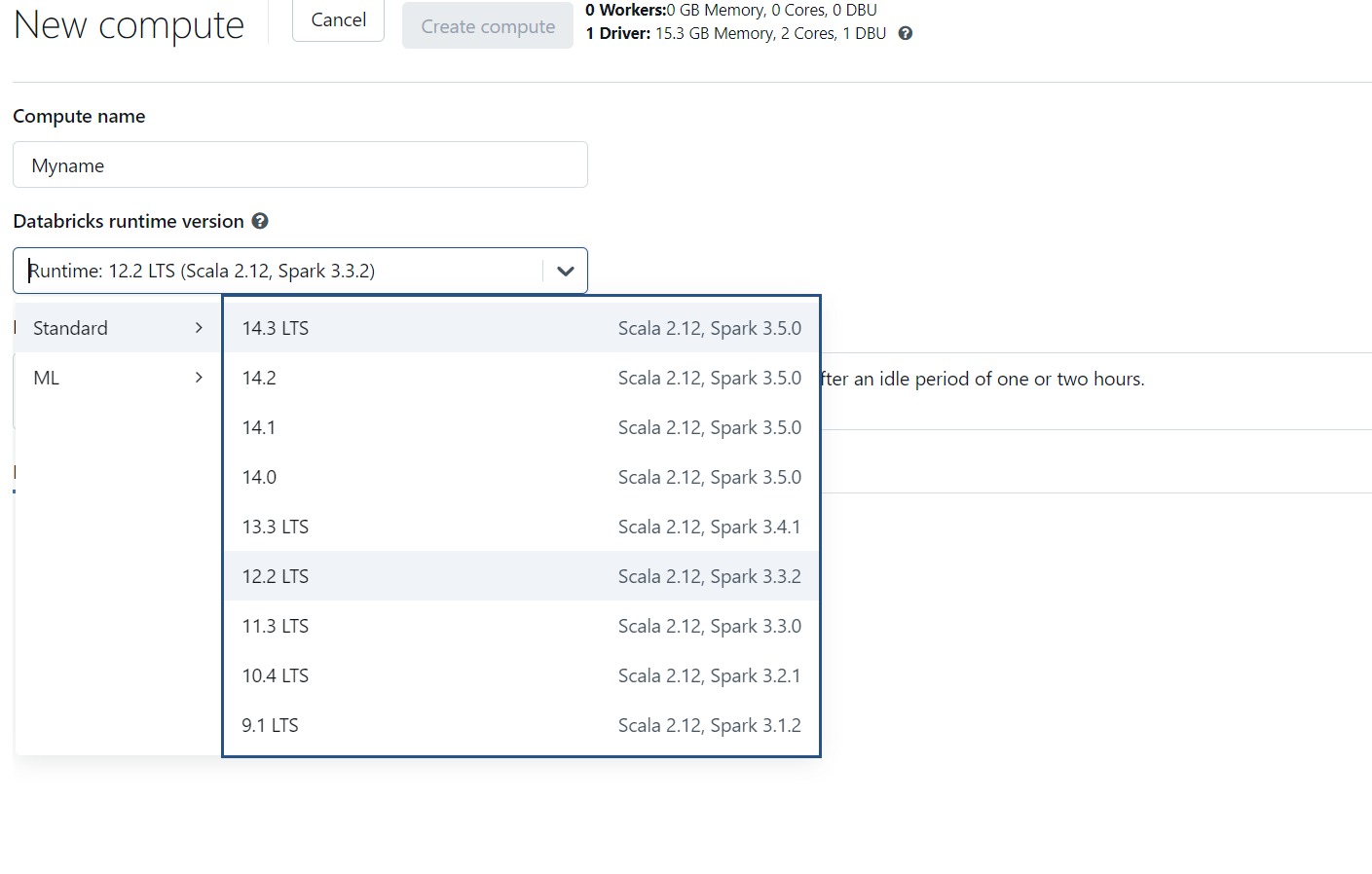
**Date:** 5/02/2024 – (Day 14)

Topics:

1. Introduction to PySpark and setting up the environment
2. Hands-on Exercise: Writing basic PySpark programs
3. Hands-on Exercise: Analyzing a sample dataset using PySpark"
4. "RDDs and transformations in PySpark
5. Hands-on Exercise: Transforming data with PySpark RDDs"

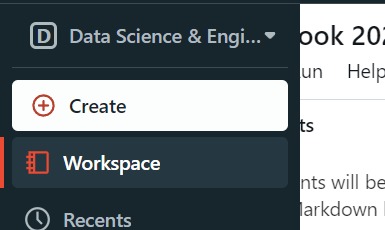
**Creating Cluster**

* Click on create cluster and give the cluster name and runtime version

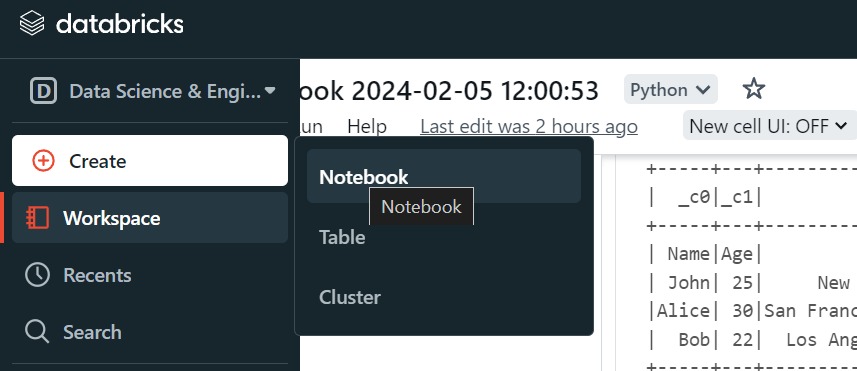


**Creating Notebook**

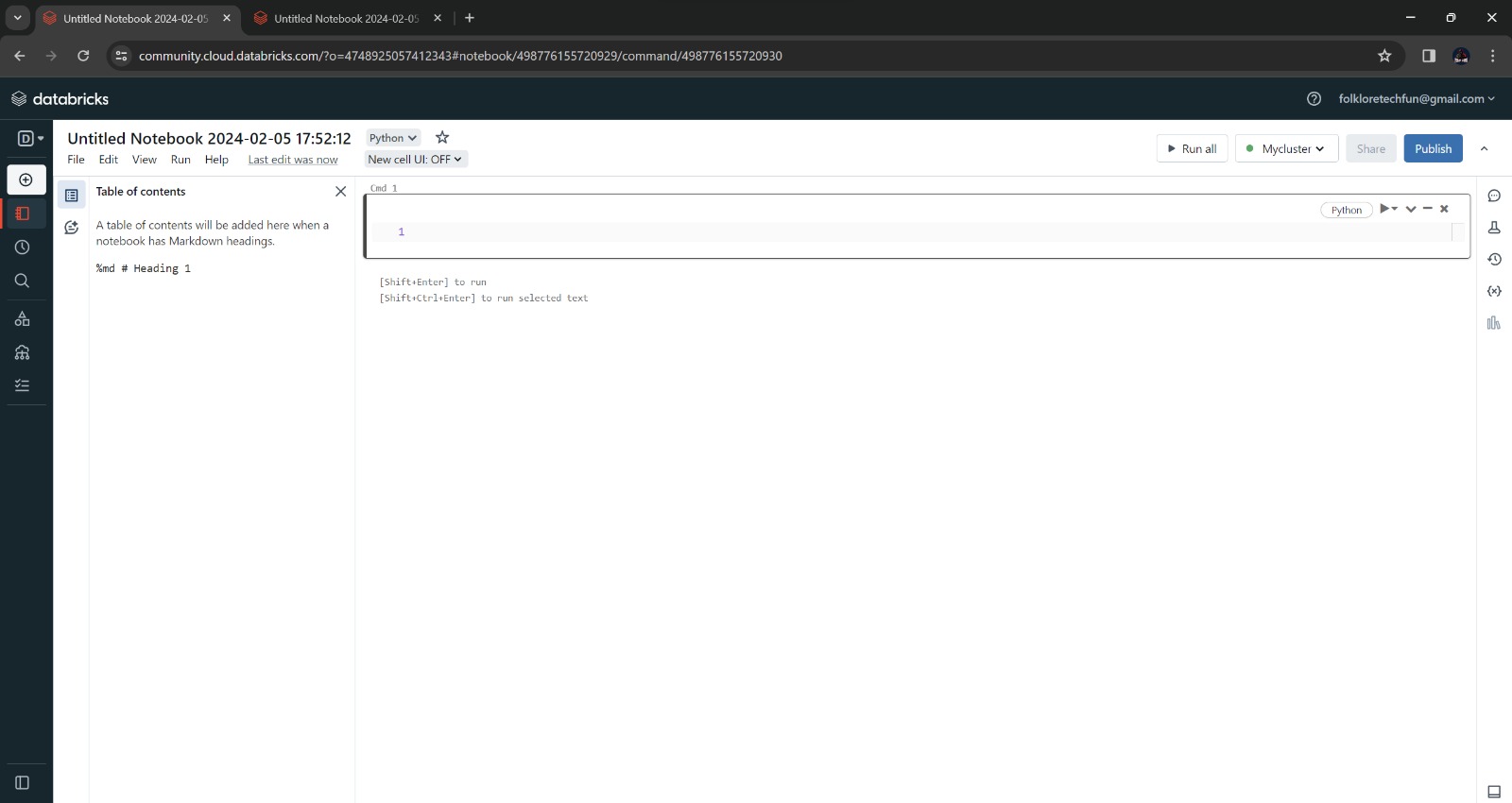
* Click on create button



* Click on Notebook

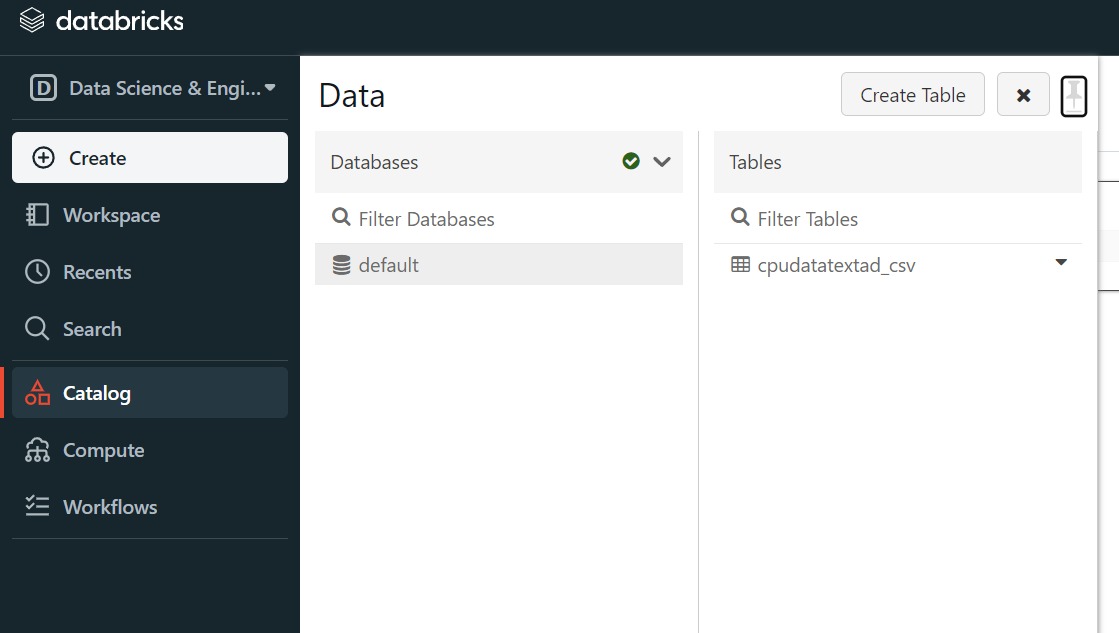


* Select cluster

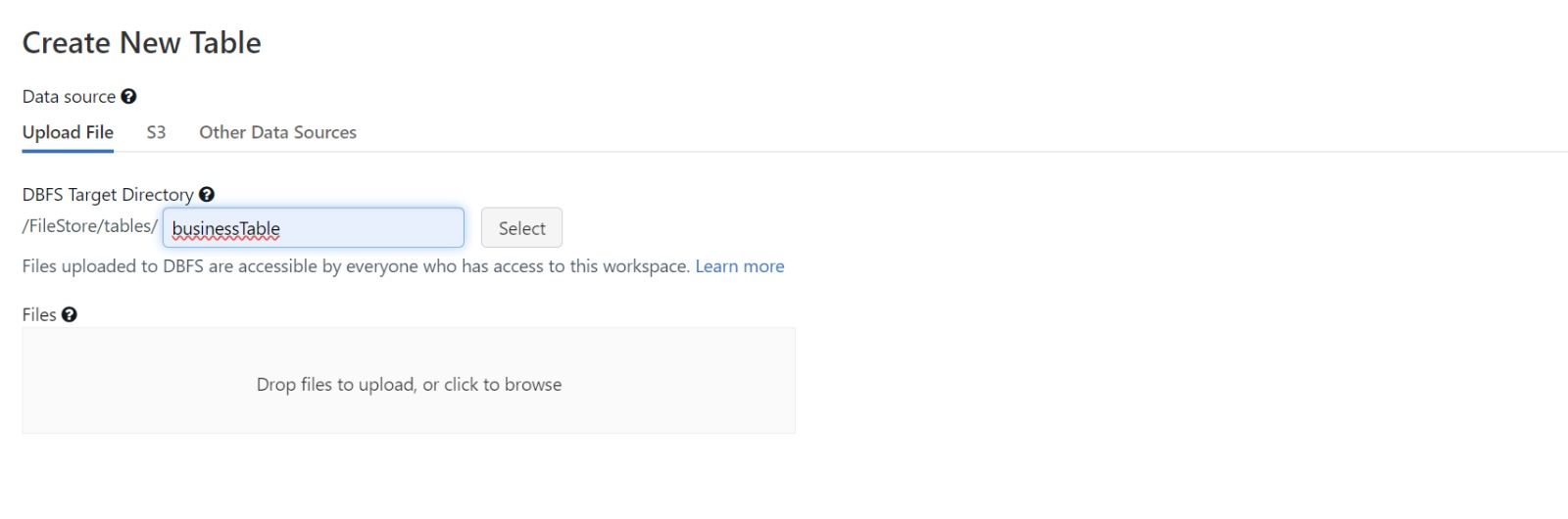


**Creating Table**

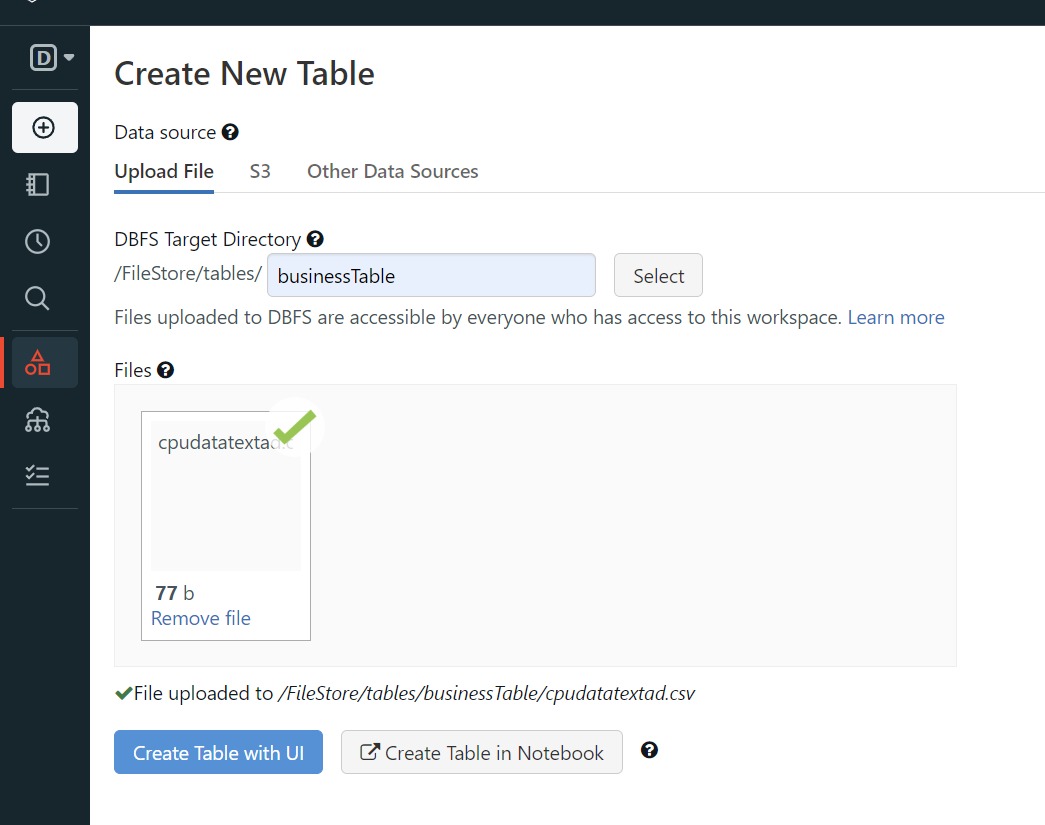
* Go to catalog session and click on create table



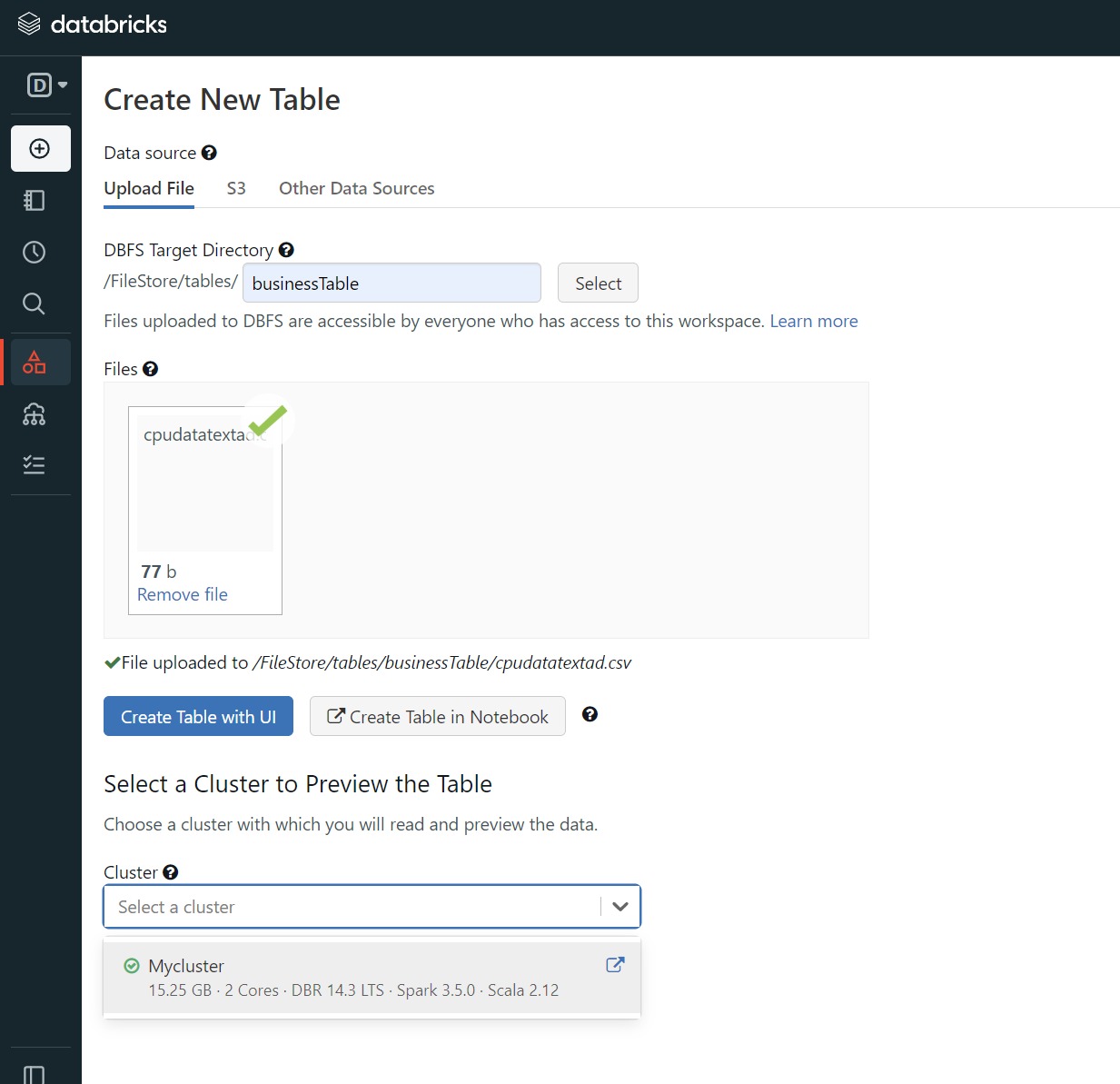
* Write the table and drag the folder to the files bar



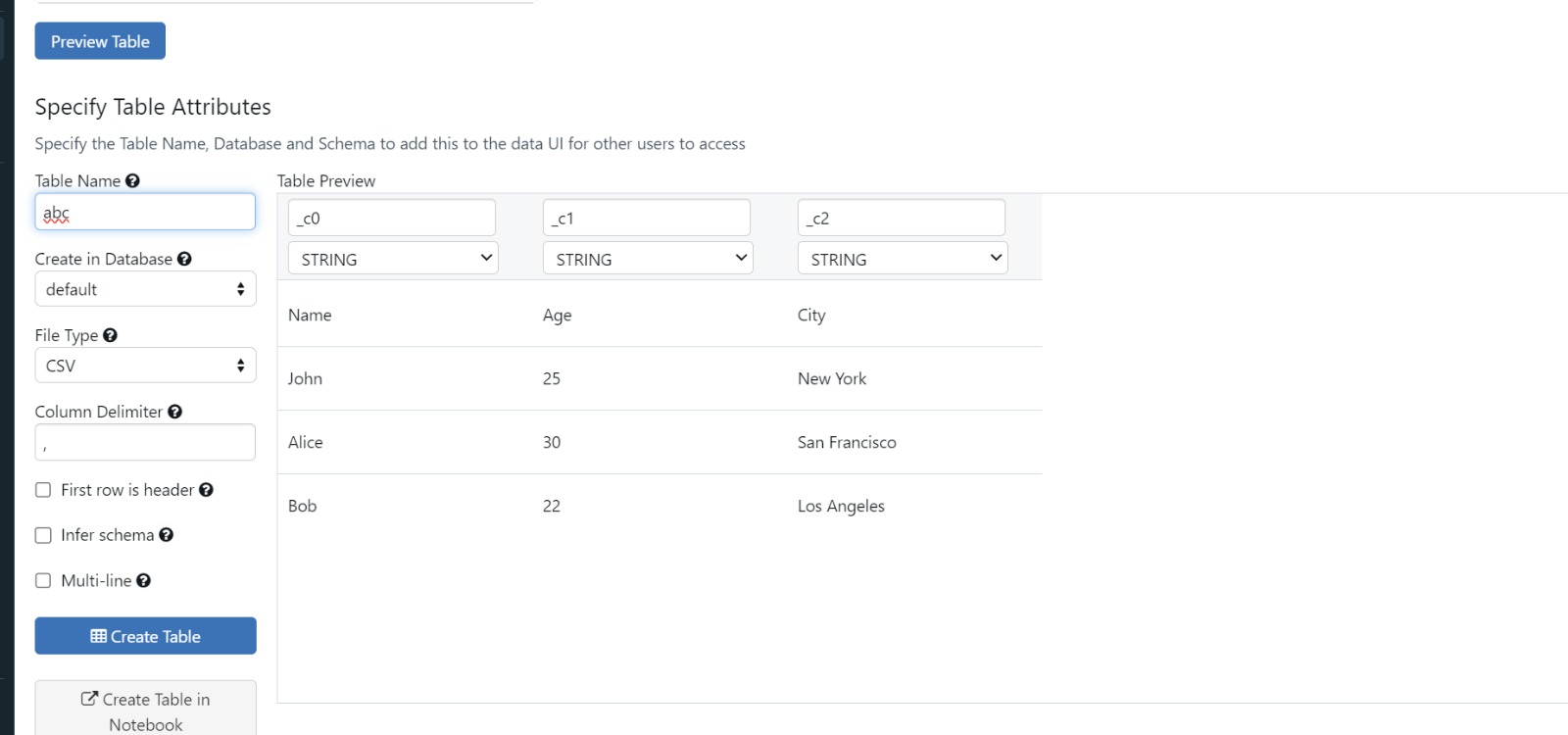
* Click on Create Table with UI



* Select the which you have created



* Click on Create table and your table is created



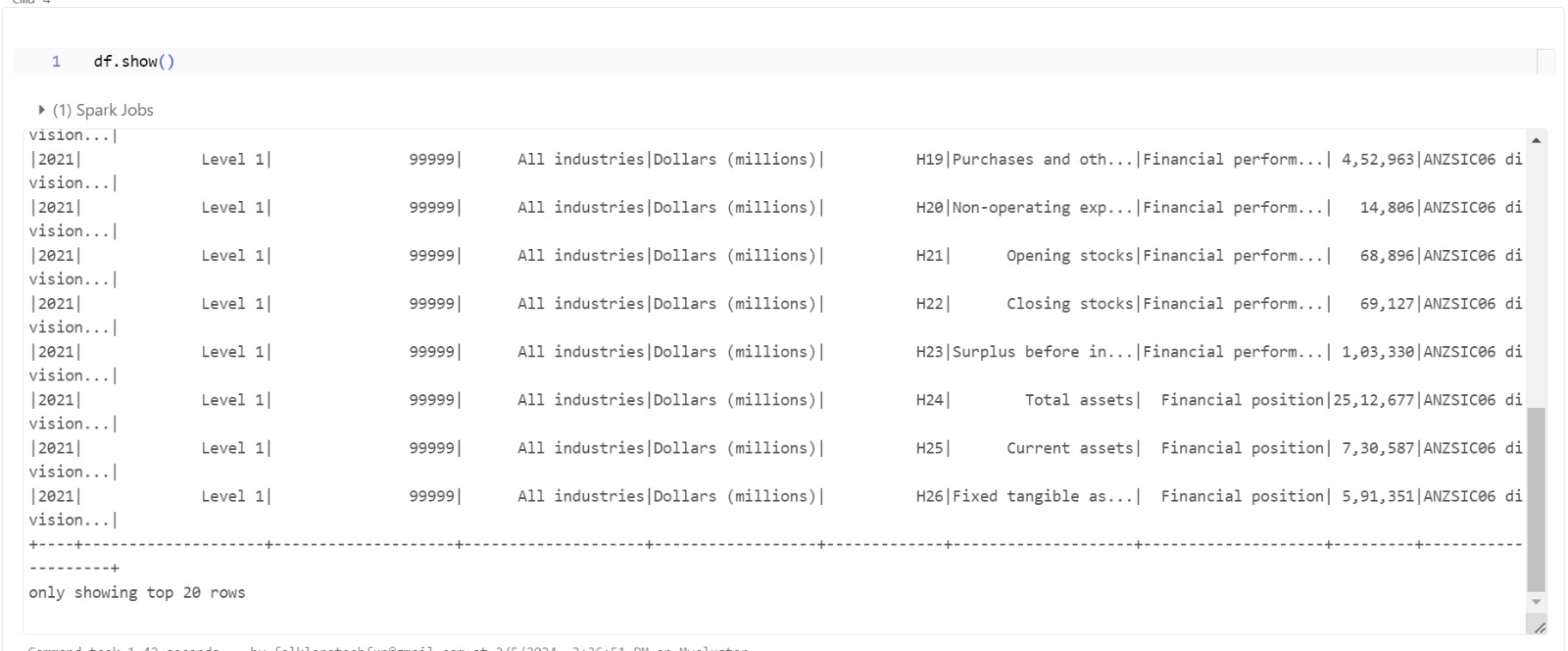
**Spark Session creation** :



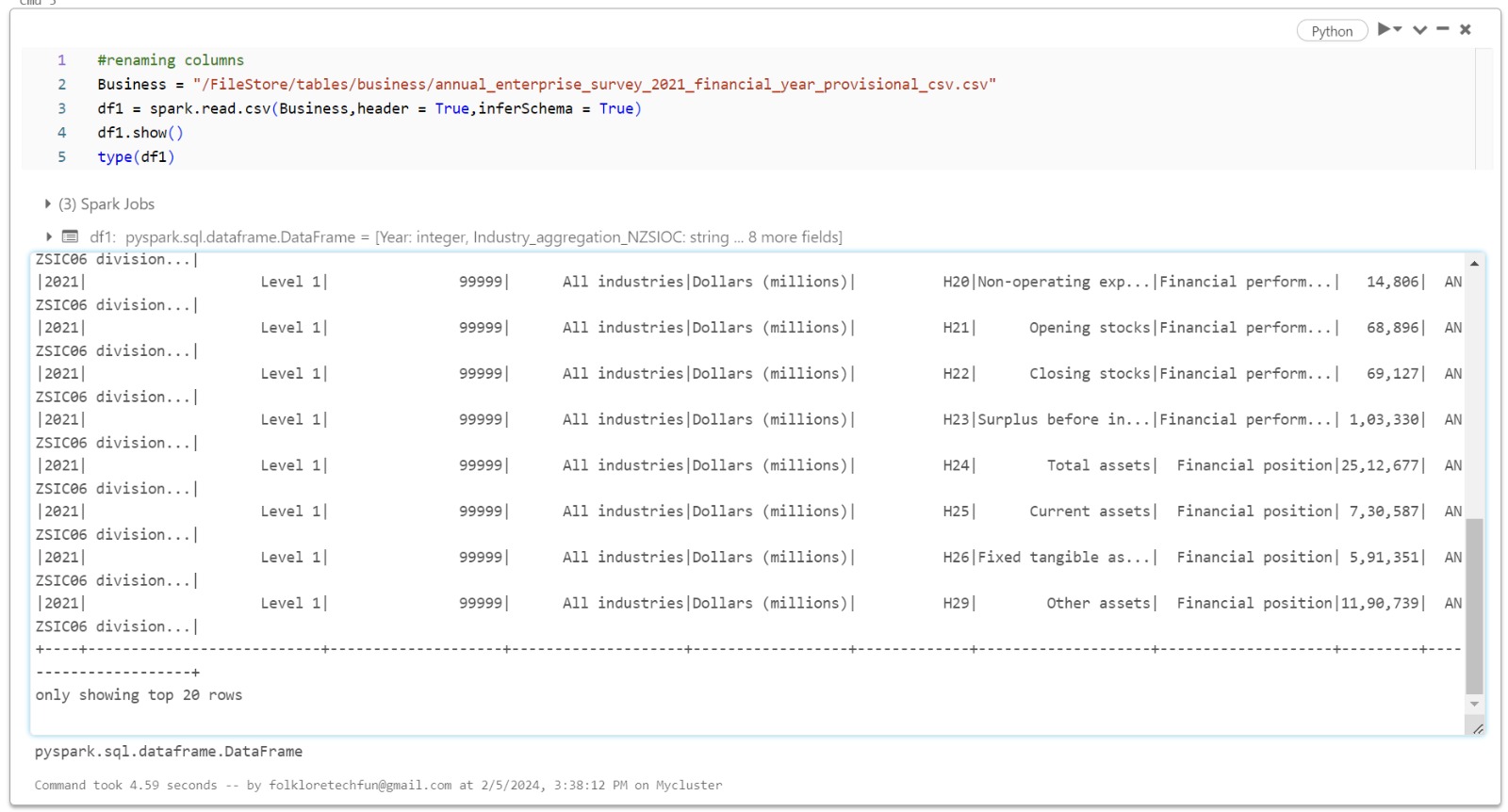
Reading CSV File



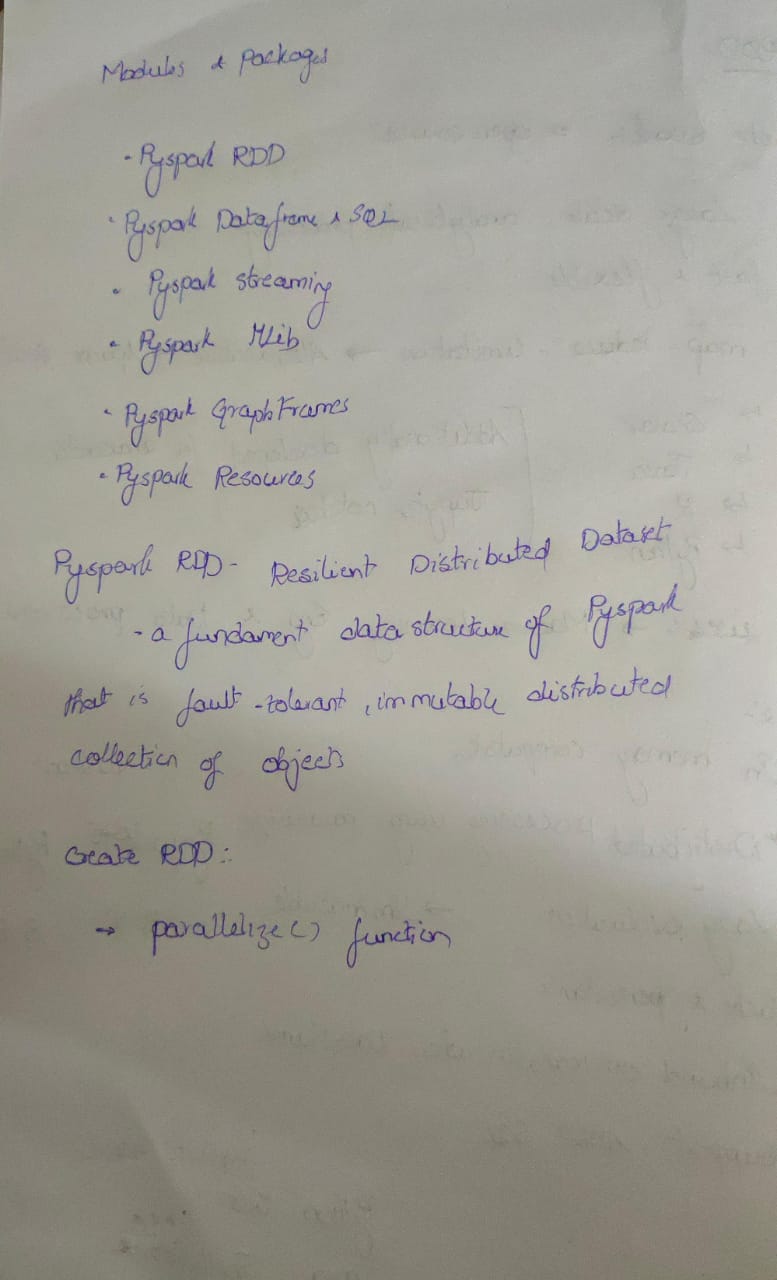
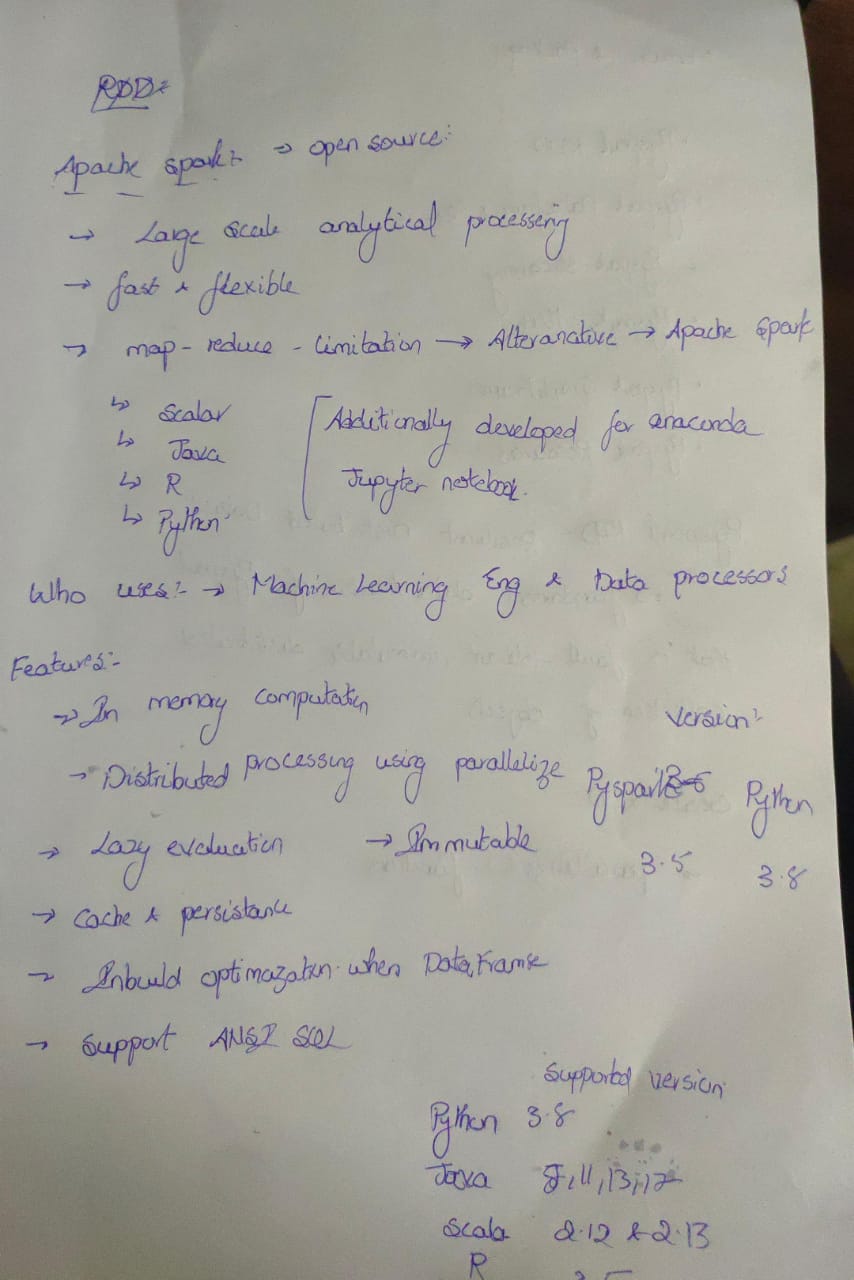
Showing CSV file data



**Giving properties like header and inferschema**



**Printing Schema**

**Day notes: **