

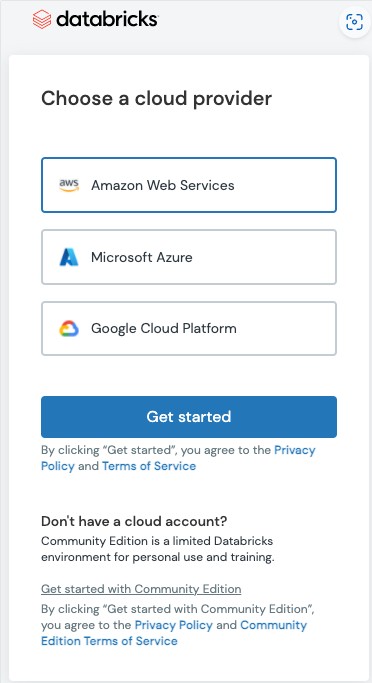
**Setup Guide [For Lab1]**

# Getting Started

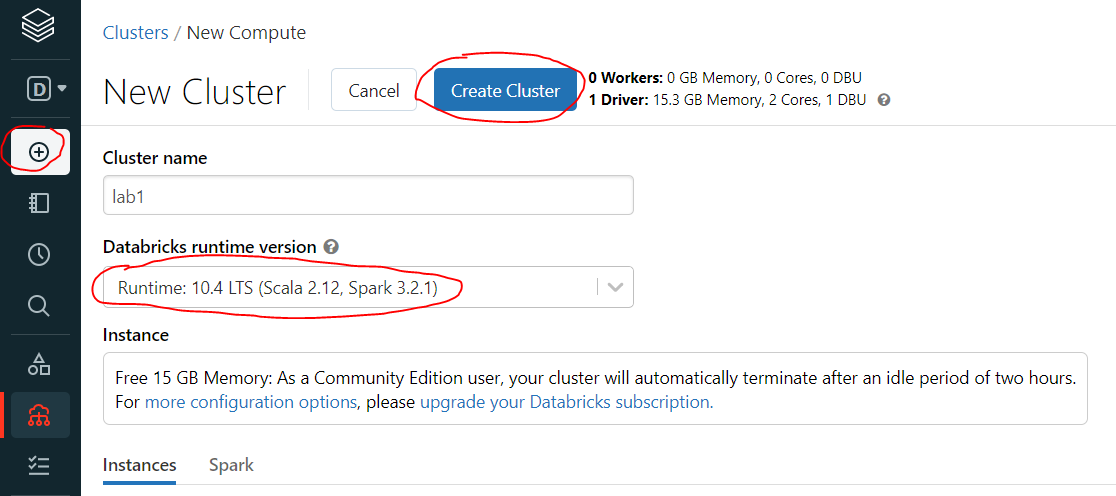
In this lab, we will use the Databricks platform to execute Spark/SQL tasks. Databricks has excellent [documentation](https://docs.databricks.com/index.html) and we defer to their guidance instead of reproducing it here. Follow these steps to get started:

1. Create a **Community Edition (**<https://community.cloud.databricks.com/>**)** account on Databricks. Do **NOT** select Databricks Platform - Free Trial; if you do, you will encounter many problems in the subsequent sections.

NOTE: select “Get started with community edition” on this page

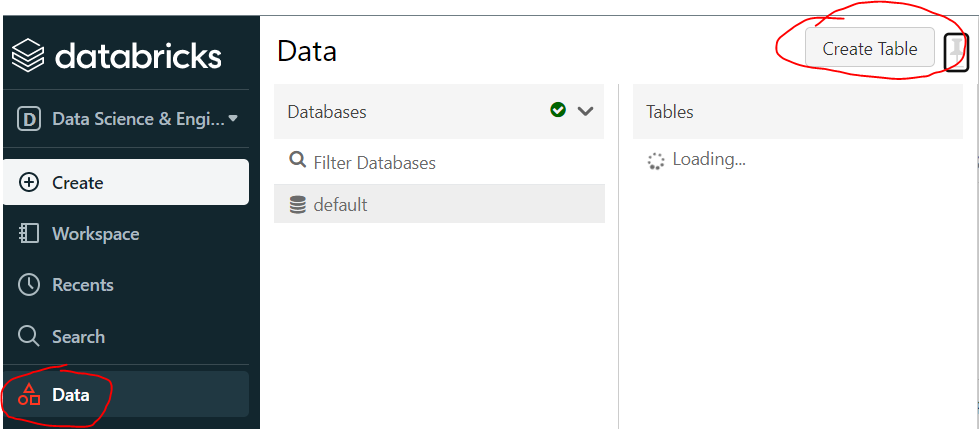


After setting up a **Community Edition** account Follow the [Quickstart Steps 1](https://docs.databricks.com/getting-started/quick-start.html#databricks-quickstart) to become familiar with the Databricks UI and to create a cluster. For this assignment, select the cluster Databricks Runtime version as ’10.4 LTS’. The cluster name can be defined by you.

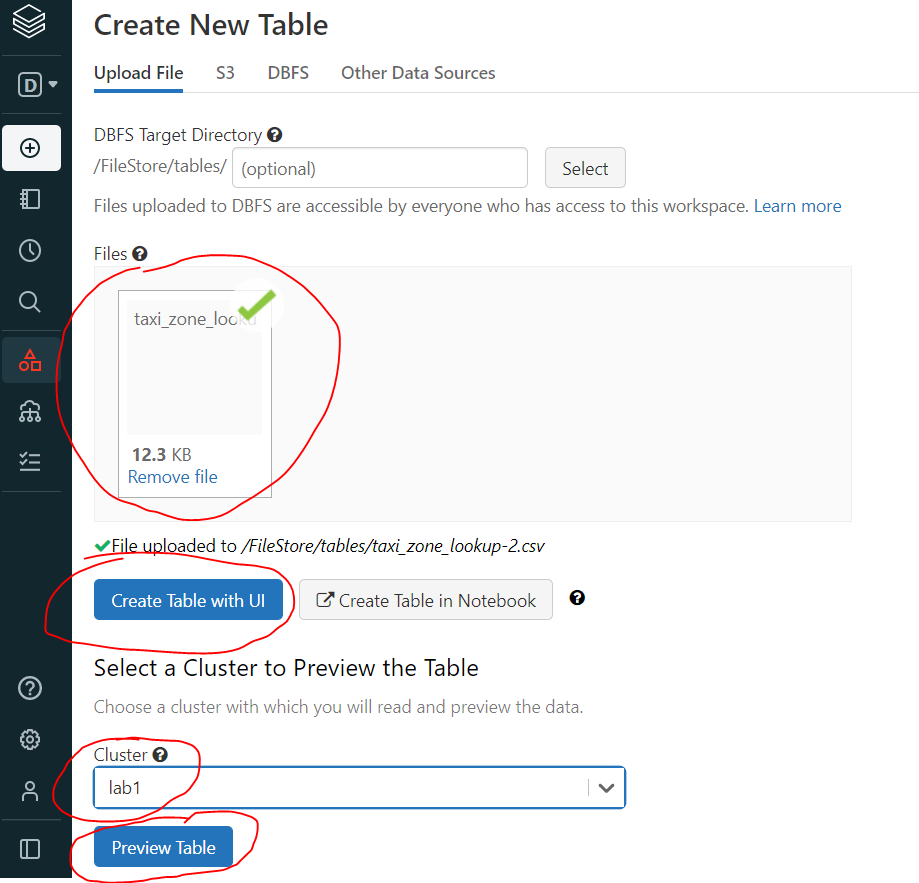
1. 

Note that your cluster will need to be re-created periodically. As a Community Edition user, your cluster will automatically terminate after an idle period of two hours.

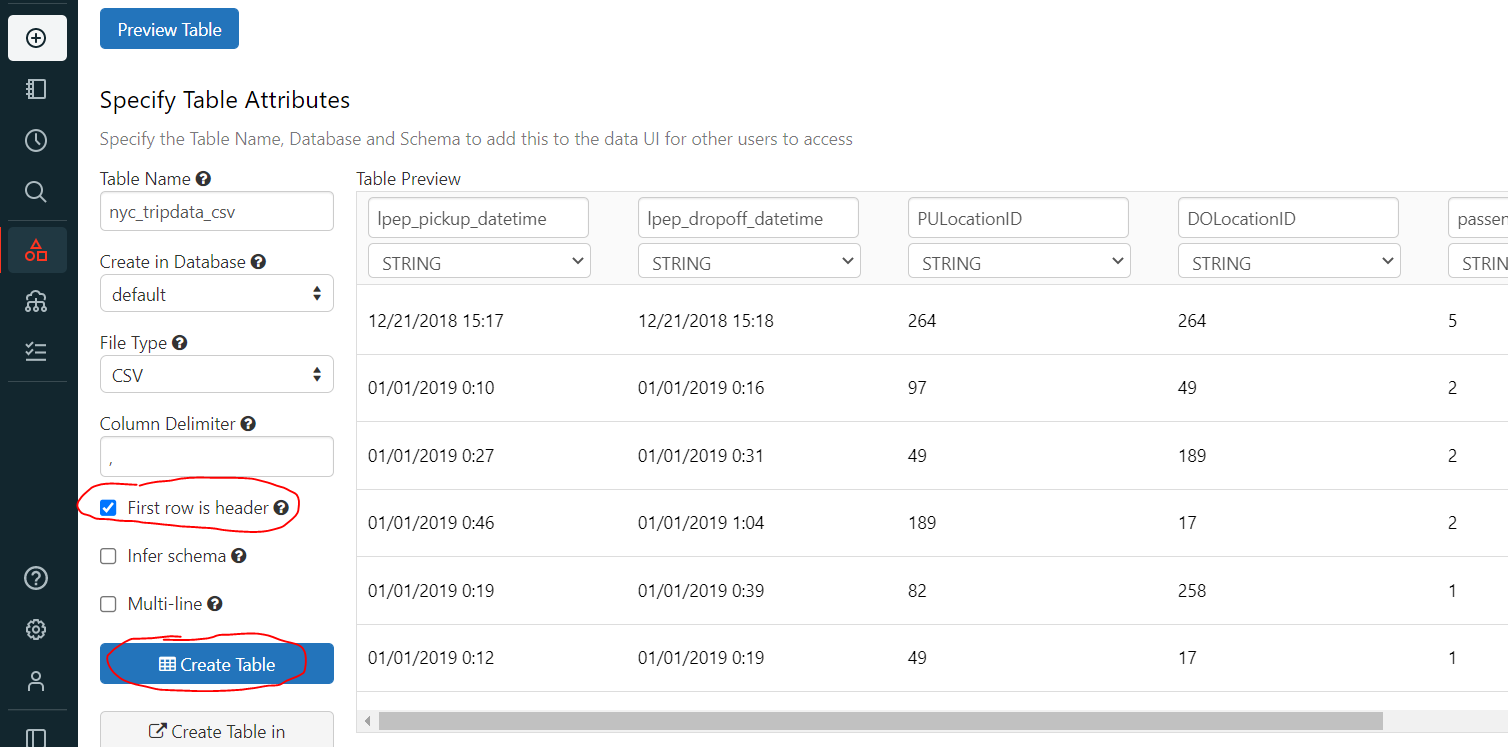
1. Import the data file, **taxi\_zone\_lookup.csv** into your workspace in the Data option using Data 🡪 Create Table 🡪 Upload File



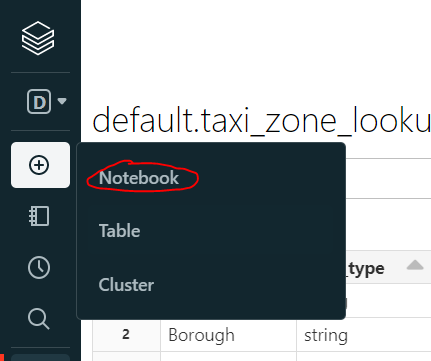
1. Click “Create Table with UI”, select the cluster that was created previously, click “Preview Table”



1. Select “First row is header” and “Create Table”.



1. Repeat the steps 3-5 to create table using **nyc-tripdata.csv.**
2. Create Notebook



1. Enter the Notebook name, select SQL as the language, select the cluster that was created previously. Either use the default name or define it by yourself.
2. Check the data using the following SQL code. You should see the top five rows of the two tables. You also can check the table name at Data 🡪 Tables. Note that the “+” above the output table is used to generate data visualization.

select \*

from nyc\_tripdata\_csv

limit 5

select \*

from taxi\_zone\_lookup\_csv

limit

-----------------

END