**Databricks Setup Guide [For Q2]**

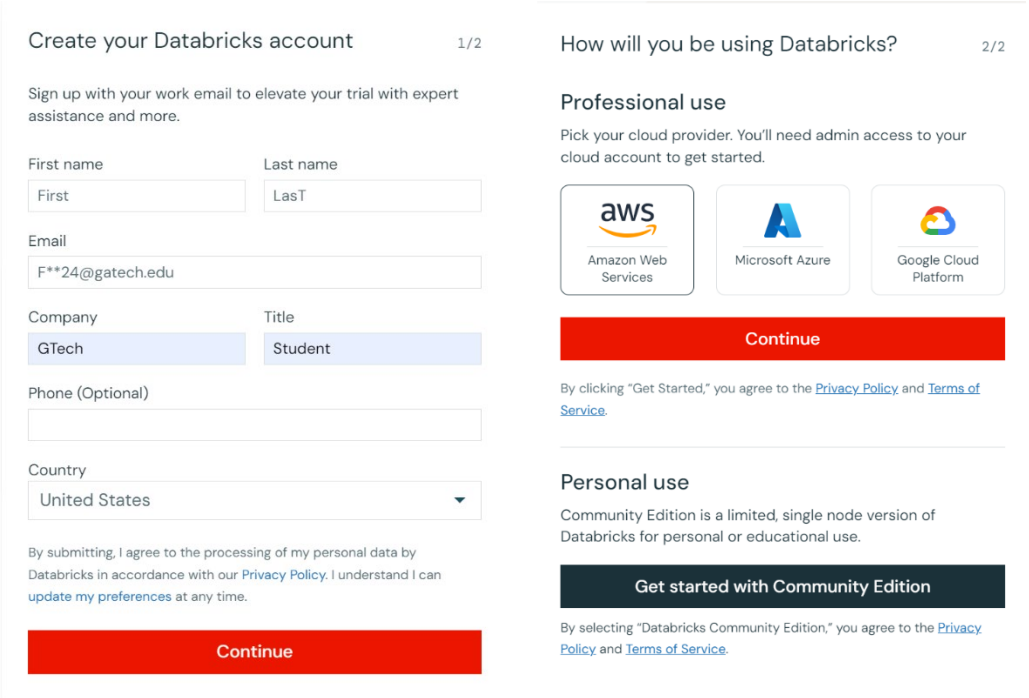
## Getting Started

For Q2, we will use the Databricks platform to execute Spark/Scala tasks. Databricks has excellent documentation and we defer to their guidance instead of reproducing it here. Follow these steps to get started:

1. Create a **Community Edition** account on Databricks. Do **NOT** select Databricks Platform - Free Trial; if you do, you will encounter many problems in the subsequent sections.

    a. Go to https://www.databricks.com/try-databricks#account to create an account

    b. b. Enter your info, see sample above, select **"Continue"**, then select "**Get started with Community Edition**" on the next page

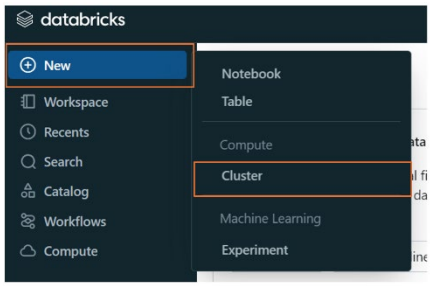



2. You should receive an email confirmation and account verification. After set-up Use https://community.cloud.databricks.com/ to login your account.

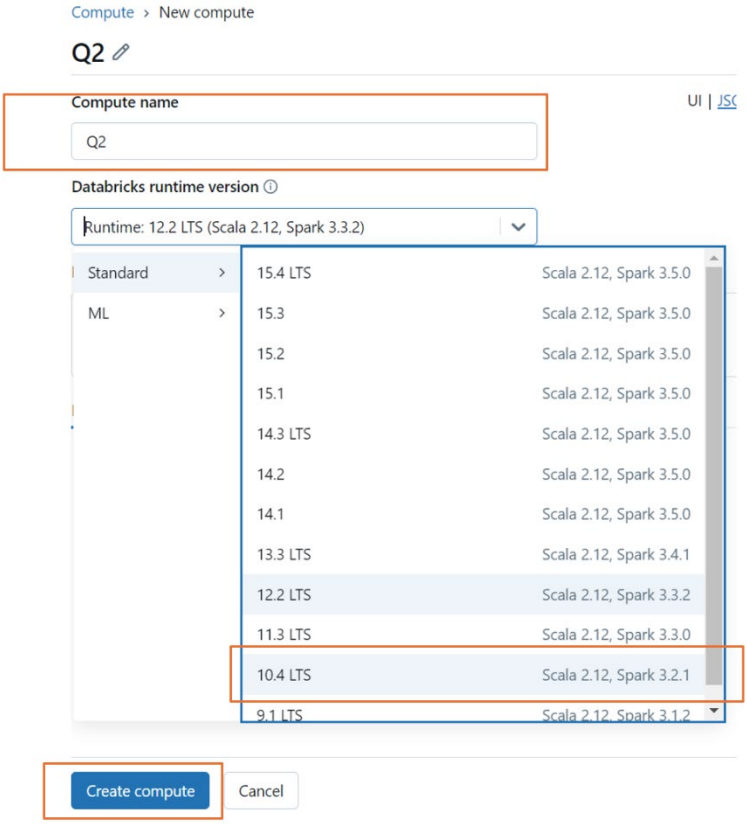
3. After setting up a **Community Edition** account Follow the Quickstart Steps 1-2 to become familiar with the Databricks UI and to create a cluster.

## Create Cluster

4. On the Databricks site, select **"+" –> New –> Cluster:**

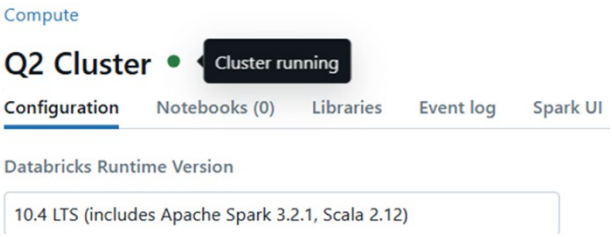



5. Type in a compute name, from the run time version drop-down menu, select the cluster Databricks Runtime (DBR) version as '10.4 LTS'. Grading will be done using this version of DBR. Then select **Create Compute**




The Python version does not matter. You do not need to set the "Availability Zone" and can leave it at default value. Give this step a few moments to initialize the cluster – you

will see a green dot by the cluster name on the next screen when cluster is ready(hovering over the cluster should provide the status, such as “cluster running”).




Note that your cluster will need to be re-created periodically.  As a Community Edition user, the cluster will automatically terminate after an idle period of 1-2 hours.
https://docs.databricks.com/getting-started/quick-start.html

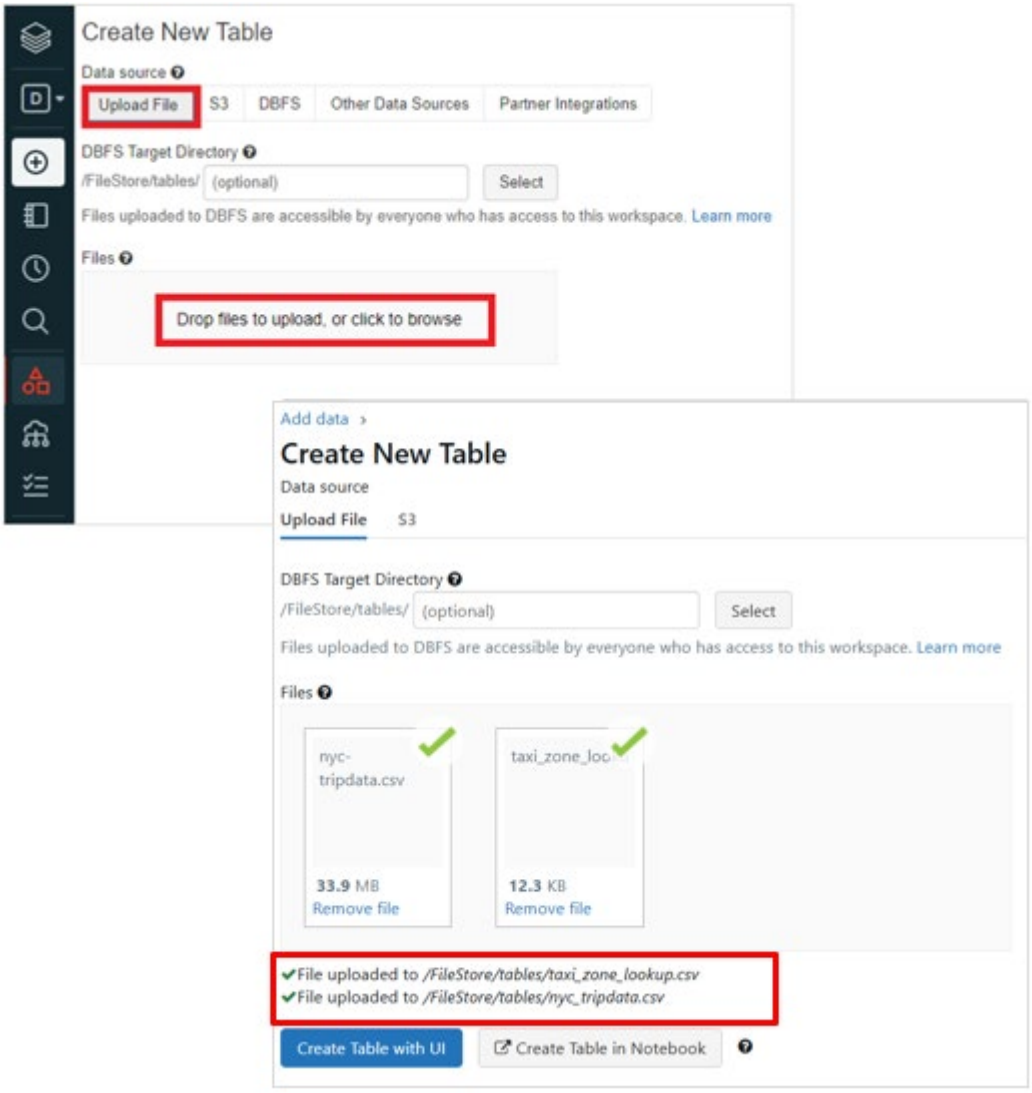
## Import Data

6. Import the data files, i.e., **nyc-tripdata.csv and taxi_zone_lookup.csv** into your workspace using **New –> Table**. Select **Upload File –> Drop Files to upload or click to browse**. Drag and drop or browse for files on your computer. Record the path of your files when uploaded. This path will be used to read the file into your Scala Notebook. Note the path below is:

   "/FileStore/tables/taxi_zone_lookup.csv",

   "/FileStore/tables/nyc_tripdata.csv'.

   **You will probably get a different file-path**, for you to copy and use in your code.

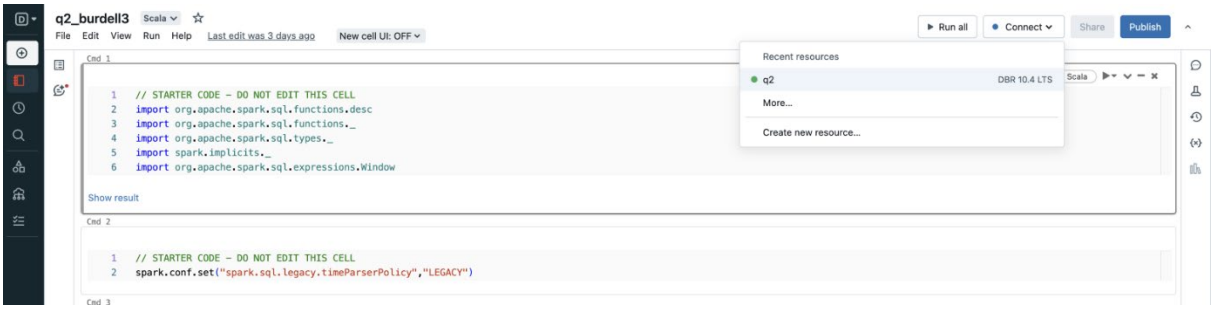


## Import Notebook and Attach cluster

7. There are several options to import the Q2.dbc notebook - this is a template notebook containing Scala code that you can use for Q2. Here are two options:
    a. Select **New** -> **Notebook** to create an untitled blank notebook. From within that new notebook, go to File, Import, select **Import**. Import the template Scala notebook, **q2.dbc** from **hw3-skeleton/q2** into your workspace. This is a template notebook containing Scala code that you can use for Q2.

    b. **Select Workspace –>** create a folder in the workspace, then select the ellipsis (three dots) next to the **"Users"** to pull up the menu, select **Import**. Import the template Scala notebook, **q2.dbc** from **hw3-skeleton/q2** into your workspace. This is a template notebook containing Scala code that you can use for Q2. https://docs.databricks.com/en/notebooks/notebook-export-import.html#import-a-notebook
    c. To access imported files, s**elect Workspace –>** you should see a list of available files, select the Q2 file to open the notebook

8. At the top of the notebook, select **Connect** to attach your cluster to the imported notebook. https://docs.databricks.com/en/notebooks/notebook-ui.html#attach-a-notebook-to-a-cluster

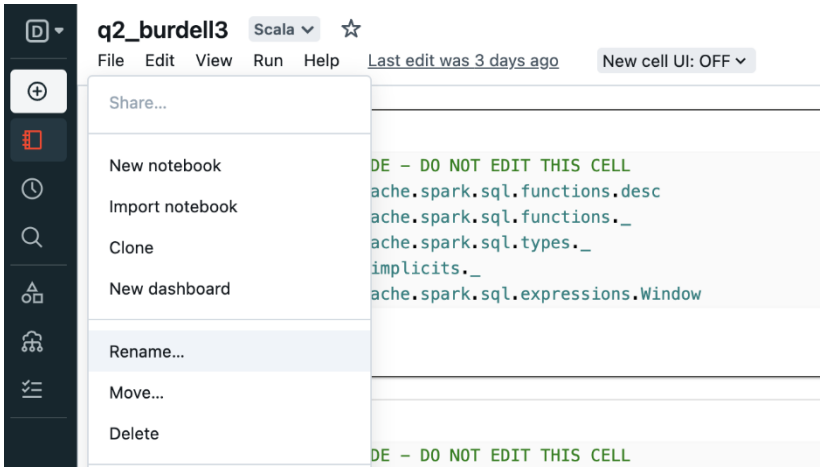



9. Review documentation on developing and using notebooks. https://docs.databricks.com/en/notebooks/notebooks-code.html

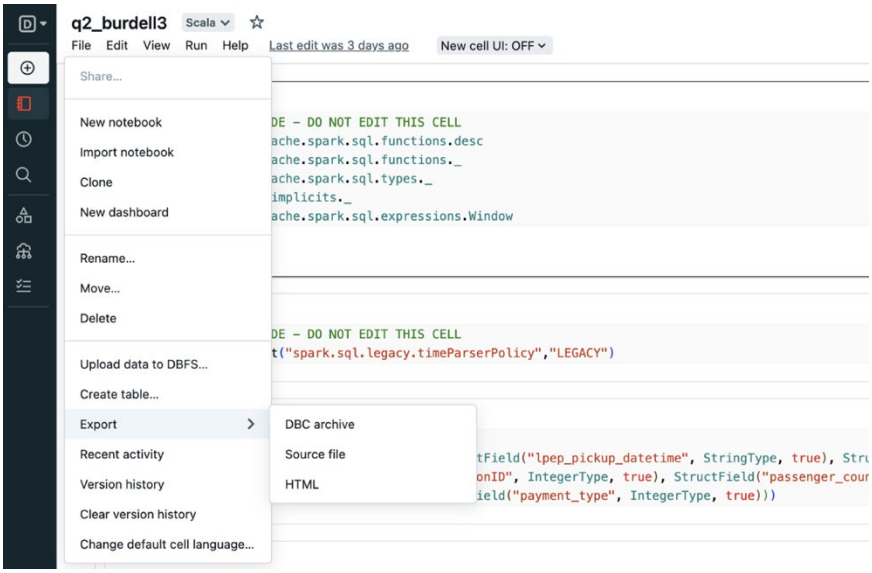
10. Within your notebook, you can access the uploaded data file **nyc-tripdata.csv and taxi_zone_lookup.csv** by using the Databricks file system utilities. Code snippet to read in the data is already provided in the dbc file present in the hw-3 skeleton. https://docs.databricks.com/user-guide/dbfs-databricks-file-system.html#access-dbfs-with-dbutils.

## Renaming and Exporting Files

11. Renaming the files.




12. Creating an exportable archive: Export your solution as **<filename>.dbc**. (see HW instructions on what this file should be named). https://docs.databricks.com/en/notebooks/notebook-export-import.html#export-notebooks: In this case, you will select, File -> Export -> DBC Archive.




Create an exportable source file: Export your solution as **<filename>.scala** (see HW instructions on what this file should be named). https://docs.databricks.com/en/notebooks/notebook-export-import.html#export-notebooks: In this case, you will select, File -> Export -> Source File.