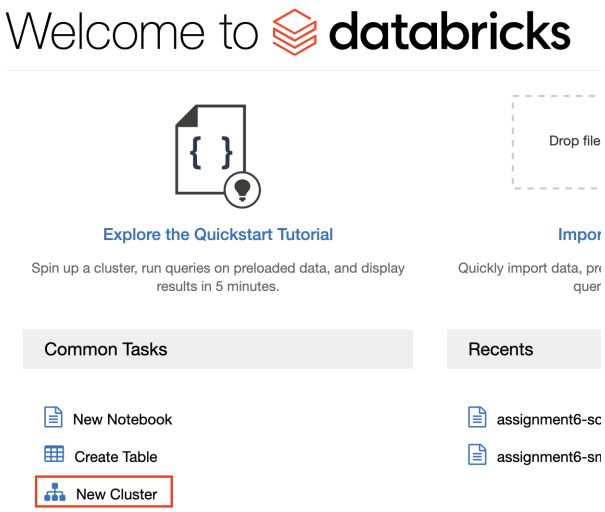
Homework Assignment #6 Setup

(Spark)

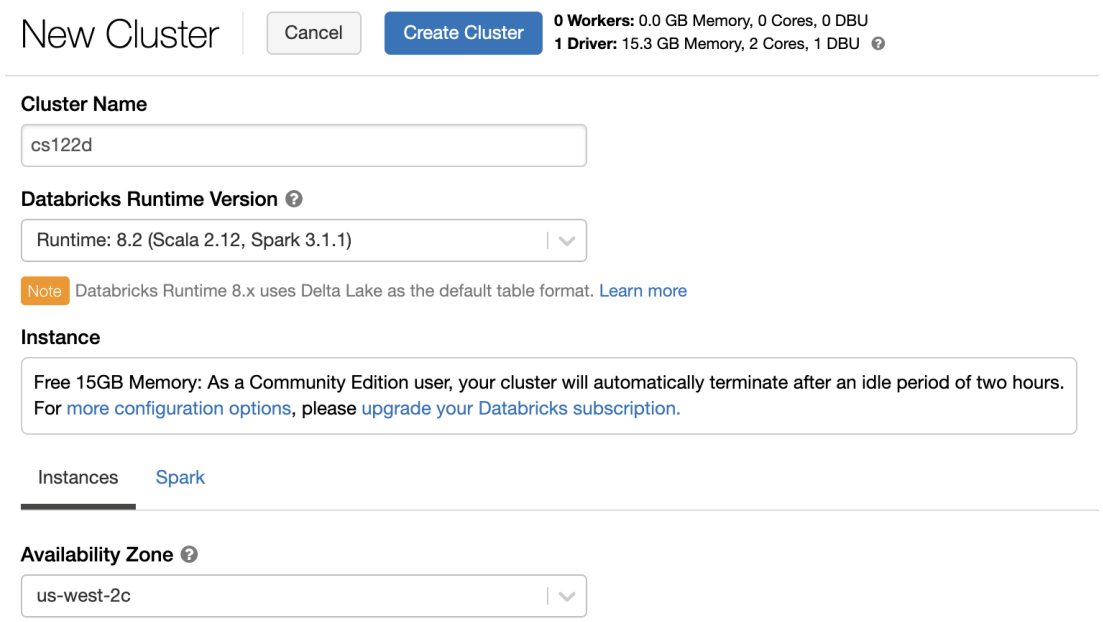
1. Create a Databricks account by following this link and be very careful to choose the ***community edition*** https://databricks.com/try-databricks.
2. Fill out the form. For the Company name you can type UCI, and for work email please provide your UCI email. You will receive an email back to confirm your email address and create a password.

Note: Clusters in Databricks self-terminate every 2 hours. You can either manually recreate the cluster using the steps 3-4 below, or you can let Databricks create a new cluster for you when running your notebook **(recommended)**. When running a cell on a detached notebook, it will ask you to automatically attach and launch a cluster.

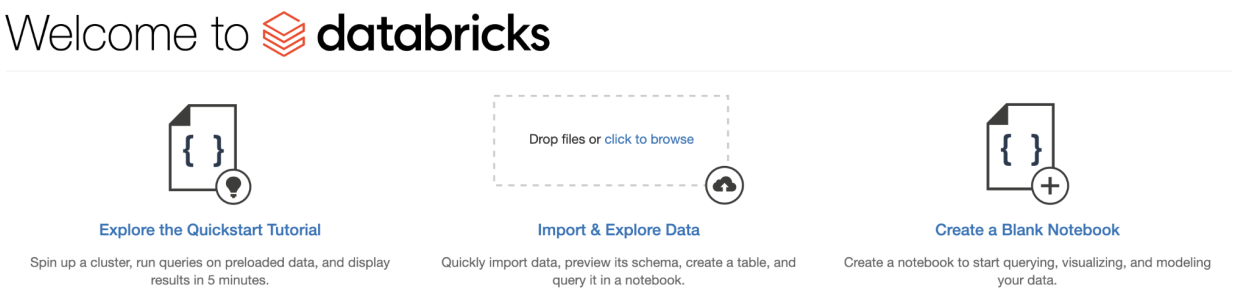
3. Click on the create cluster link when you wish to create a new cluster.



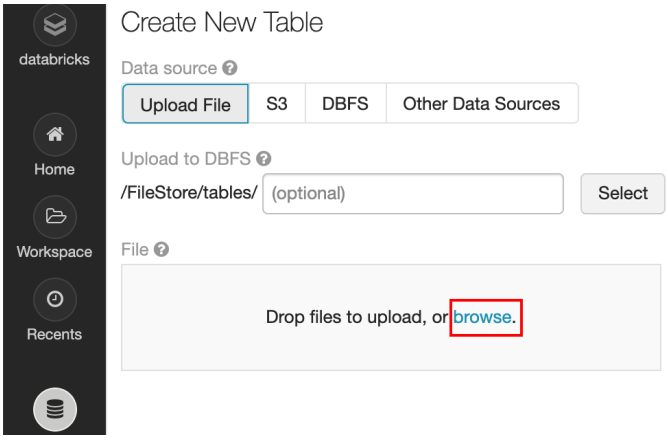
4. Pick any name for your cluster and make sure that you have the same settings as in the picture below and then create the cluster.



5. Download the json files from this Google Drive folder https://drive.google.com/drive/folders/1dOGaWH5wPfcTP72fw8mdjMMIW7zvSNXK?usp=sharing
6. Go to the main databricks page. Click the link for import & explore data.

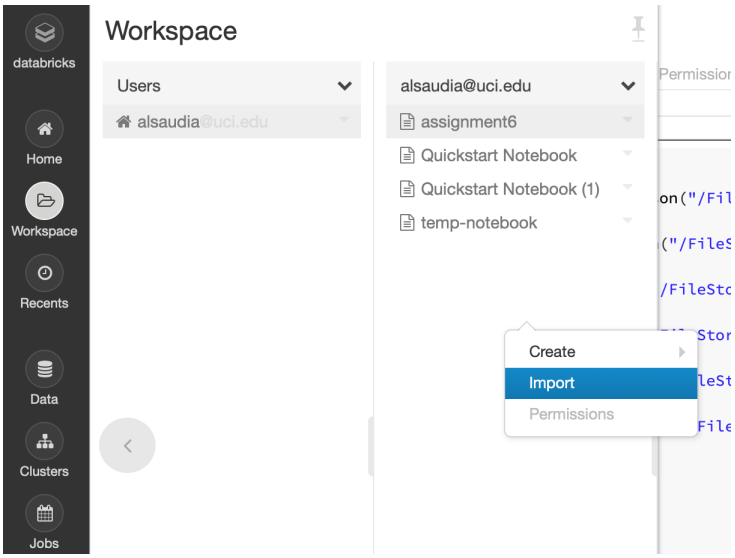


7. Click browse in order to select and import all of the json files.

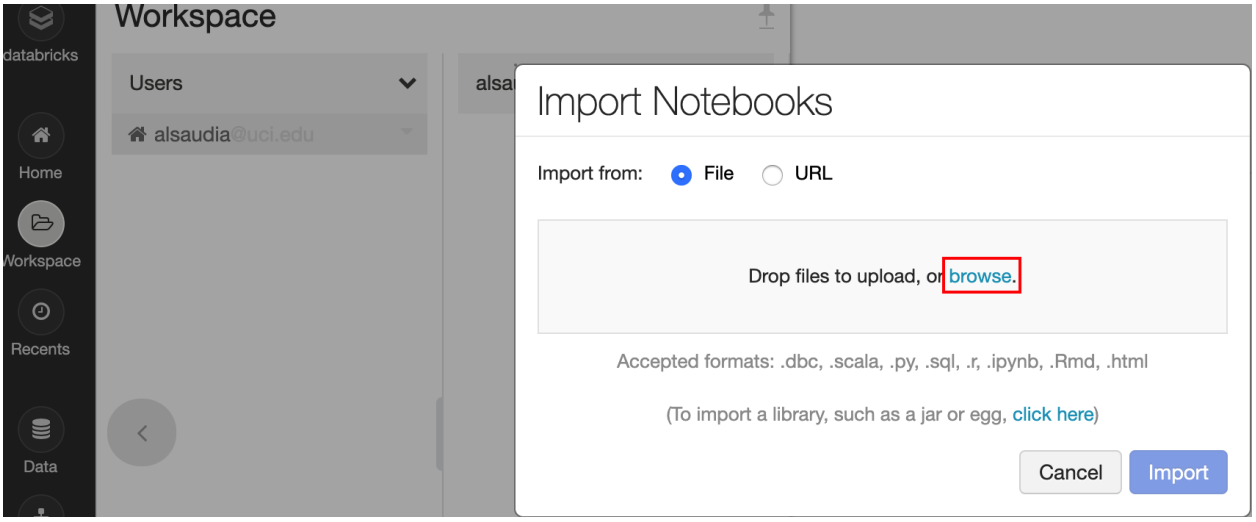


The files should now be stored in the cloud under the "/FileStore/tables" directory! You have also been provided with a notebook (ipynb) template file that contains some initial environment setup code that will load the files into dataframes that you can run queries on. Let's load that into Databricks as well by doing the following:

8. Download the hw6_template.ipynb file from the class website to your computer.
9. Click on workspace -> Users -> <user> -> right click and then import.



10. Click browse and choose the file hw6_template.ipynb and then click import.



11. After importing and opening the file, run the setup environment cell.

NOTE before running the cell: if the dropdown below says "Detached", make sure to attach the notebook to a cluster! (You can't compute without nodes to compute on…)

assignment6-smith_john-1234 (Python)

Detached | File | Edit | View: Standard | Permissions | Run All | Clear

Cmd 1

Assignment 6

Cmd 2

setup environment (must run first)

Cmd 3

```
%python
orders_df = spark.read.option("multiLine", True).json("/FileStore/tables/orders.json")
orders_df.cache() # Cache data for faster reuse
stores_df = spark.read.option("multiLine", True).json("/FileStore/tables/stores.json")
stores_df.cache() # Cache data for faster reuse
users_df = spark.read.option("multiLine", True).json("/FileStore/tables/users.json")
users_df.cache() # Cache data for faster reuse
own_df = spark.read.option("multiLine", True).json("/FileStore/tables/own.json")
own_df.cache() # Cache data for faster reuse
workfor_df = spark.read.option("multiLine", True).json("/FileStore/tables/workfor.json")
workfor_df.cache() # Cache data for faster reuse
products_df = spark.read.option("multiLine", True).json("/FileStore/tables/products.json")
products_df.cache() # Cache data for faster reuse
stockedby_df = spark.read.option("multiLine", True).json("/FileStore/tables/stockedby.json")
stockedby_df.cache() # Cache data for faster reuse
vehicles_df = spark.read.option("multiLine", True).json("/FileStore/tables/vehicles.json")
vehicles_df.cache() # Cache data for faster reuse

```

12. You are now ready to proceed with the assignment itself!