

Spring 2022 Setup Guide [For Q2]

Getting Started

A video tutorial has been created to help walk you through the steps in this document. You can view it <u>here</u>.

NOTE: This setup guide is for reference only and is not meant to serve as an exact stepby-step guide.

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

- Create a Community Edition (<u>https://community.cloud.databricks.com/</u>) account on Databricks. Do NOT select Databricks Platform - Free Trial; if you do, you will encounter many problems in the subsequent sections. More info: <u>https://docs.databricks.com/getting-started/try-databricks.html</u>
- After setting up a Community Edition account Follow the <u>Quickstart Steps 1-2</u> to become familiar with the Databricks UI and to create a cluster. For this assignment, select the cluster Databricks Runtime version as '6.4 (includes Apache Spark 2.4.5, Scala 2.11)' (without GPU). Grading will be done using this version of DBR. 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.

	Create Cluster
D-	New Cluster Cancel Create Cluster 0 Workers:0 GB Memory, 0 Cores, 0 DBU 1 Driver:15.3 GB Memory, 2 Cores, 1 DBU @
Ð	Cluster Name
	hw3q2
₽	Databricks Runtime Version @
\bigcirc	Runtime: 6.4 Extended Support (Scala 2.11, Spark 2.4.5)
Q	Instance
	Free 15 GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle period of two hours. For more configuration options, please upgrade your Databricks subscription.
æ	Instances Spark
×=	Availability Zone 😡
	us-west-2c

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. <u>https://docs.databricks.com/getting-started/quick-start.html</u>

3. Import the data files, i.e., nyc-tripdata.csv and taxi_zone_lookup.csv into your workspace in the Data option using Create Table. Record the path of your file when it is uploaded. This path will be used to read the file into your Scala Notebook. Note the path below is: '/FileStore/tables/nyc_tripdata-1.csv'. You will probably get a different file-path, which you must copy and use in your code.

⊜ d	atabricks	Data				Create Table
D Da	ata Science & E	Databases	•	✓ Ta	bles	
		Q Filter Data	Q Filter Databases		No Tables	
(+) Cr	eate	Sefault				
₽ ₩	orkspace					
C Re	ecents					
Q Se	earch					
<mark>∆</mark> Da	ata					
ኇ	ompute					
ŏ⊒ Jo	bs					
	Create New	Table				
	Create New	Table				
\$ •	Create New Data source @ Upload File	Table 3 DBFS	Other Data Sources	Partne	r Integrations	
	Create New Data source • Upload File S DBFS Target Direct	Table 3 DBFS ony O	Other Data Sources	Partne	r Integrations	
	Create New Data source O Upload File S DBFS Target Direct /FileStore/tables/	Table 3 DBFS ory • • optional)	Other Data Sources	Partne	r Integrations	
	Create New Data source • Upload File S DBFS Target Direct /FileStore/tables/ (Files uploaded to D	Table 3 DBFS ory • optional) BFS are access	Other Data Sources	Partne Select	r Integrations	space. Learn more
	Create New Data source O Upload File S DBFS Target Director /FileStore/tables/ (Files uploaded to D	Table 3 DBFS ory	Other Data Sources	Partne Select	r Integrations	space. Learn more
	Create New Data source O Upload File S DBFS Target Direct /FileStore/tables/ (Files uploaded to D Files O	Table 3 DBFS ory	Other Data Sources ible by everyone who	Partne Select	r Integrations	space. Learn more
()) (Create New Data source O Upload File S DBFS Target Direct /FileStore/tables/ ((Files uploaded to D Files O	Table 3 DBFS ory • • optional) BFS are access	Other Data Sources ible by everyone who or click to browse	Partne Select	r Integrations	space. Learn more
 Image: Second second	Create New Data source O Upload File S DBFS Target Direct /FileStore/tables/ ((Files uploaded to D Files O	Table 3 DBFS ory • optional) BFS are access	Other Data Sources ible by everyone who or click to browse	Partne Select	r Integrations	space. Learn more
 () 	Create New Data source O Upload File S DBFS Target Direct /FileStore/tables/ (Files uploaded to D Files O Drop	Table 3 DBFS ory O optional) BFS are access files to upload,	Other Data Sources ible by everyone who or click to browse	Partne Select	r Integrations	space. Learn more

	Create New Table					
₽₹	Data source O Upload File S3 DBFS Other Data Sources Partner Integrations					
Ð	DBFS Target Directory O /FileStore/tables/ (optional) Select					
₽	Files uploaded to DBFS are accessible by everyone who has access to this workspace. Learn more					
\bigcirc	Files O					
Q	nyc- tripdata.csv					
ጜ	33.9 MB					
<u>*=</u>	Remove me					
	✓File uploaded to /FileStore/tables/nyc_tripdata-1.csv					
	Create Table with UI					
▲ 徐 三	33.9 MB Remove file ✓ File uploaded to /FileStore/tables/nyc_tripdata-1.csv Create Table with UI Create Table with UI					

 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. <u>https://docs.databricks.com/user-guide/notebooks/notebook-manage.html#import-an-archive</u>

🗟 databricks	Workspace	Home
D Data Science & E	Workspace	* stagrations
	Shared	Create
Create	📽 Users	Import
Workspace		Permissions earn more
C Recents		Copy Link Address Sort
Q Search		
o⊟ Data		
G Compute		
<u>×</u> Jobs		

5. Attach your cluster to the imported notebook. <u>https://docs.databricks.com/user-guide/notebooks/notebook-manage.html#attach-a-notebook-to-a-cluster</u>

- 6. Review documentation on developing and using notebooks. <u>https://docs.databricks.com/user-guide/notebooks/notebook-use.html</u>
- 7. 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. <u>https://docs.databricks.com/user-guide/dbfs-databricks-file-system.html#access-dbfs-with-dbutils</u>.
- 8. Renaming the files.



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

	q2_gburdell3 (Scala)		
• •	<pre>hw3q2 Cmd 1 // STARTER CODE - DO NOT import org.apache.spark.s import org.apache.spark.s import spark.implicits import org.apache.spark.s</pre>	■ ✓ ▷ ✓ ■ ✓ ■ ✓ New Notebook Clone Rename Move Delete Upload Data	0 0.
Q 🐣	<pre>import org.apache.spark.sql. import org.apache.spark.sql. import org.apache.spark.sql. import spark.implicits import org.apache.spark.sql.</pre>	Export Publish Clear Revision History Change Default Language	DBC Archive Source File HTML

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