

Homework Assignment #5 Setup

Part I

(Spark on Databricks Community Version)

1. Create a Databricks Community Version account by following this link (<https://databricks.com/try-databricks>) and complete the following steps:
 - a. Fill out the registration form in the first dialog in order to get to the next account setup dialog. For the Company name you can type UCI, and for work email please provide your UCI email.

How will you be using Databricks? 2/2

Professional use

Pick your cloud provider. You'll need admin access to your cloud account to get started.

Amazon Web Services

Microsoft Azure

Google Cloud Platform

Enjoy \$400 in credits during your 14-day AWS trial. Trial ends when credits expire.

By clicking "Continue," you agree to Databricks' [Terms of Service](#).

Continue

Personal use

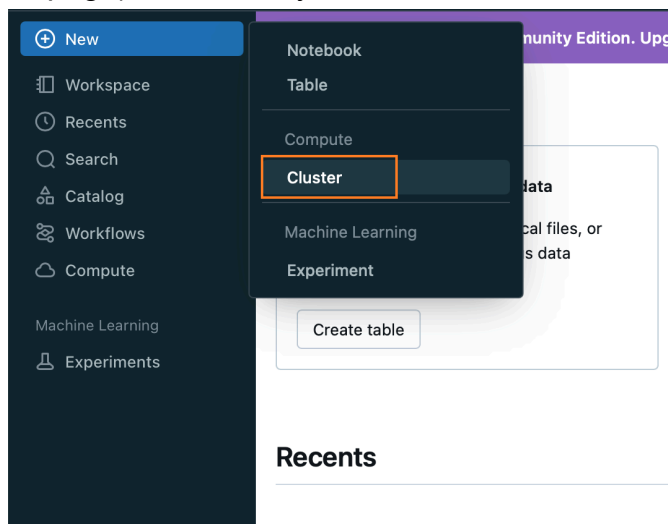
Community Edition is a limited, single node version of Databricks for personal or educational use. **Click here!**

By clicking "Get started with Community Edition," you agree to Databricks' [Terms of Service](#).

Get started with Community Edition

- b. Once you click "Continue" and see the account setup dialog as shown in the above screenshot, be very careful to choose the **Community Edition** for your account. Do NOT click the "Continue" button in this dialog as that requires you to have an existing cloud provider, which we will not do until Part II.
 - c. You will receive an email back to confirm your email address and create a password.

2. Click on the create cluster link (under “Create->Cluster” in the left panel of your account dashboard page) whenever¹ you wish to create a new cluster:



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

Compute > New compute

224P

Compute name

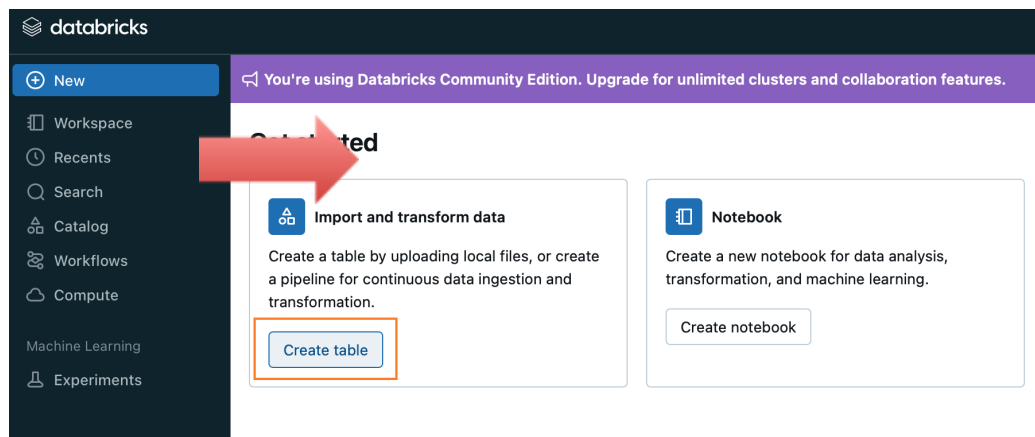
224P

Databricks runtime version ⓘ

Runtime: 15.4 LTS (Scala 2.12, Spark 3.5.0) | v

4. Download both the ZotMusic HW5_sample and the HW5_full json files from Canvas.
5. Go to the main dashboard page. Click **Create table**.

¹ Unfortunately you will have to do this multiple times, once each time you start working on the homework after having taken a break for a while, as the life expectancy of a Databricks CE cluster is quite short.



6. Click browse in order to select and import the two datasets in the appropriate folders:
 - a. Put the json files in the sample dataset under **zotmusic_sample**
 - b. Put the json files in the full dataset under **zotmusic_full**

[Add data](#) >

Create New Table

Data source ⓘ

Upload File S3 DBFS**Make sure to input the subdirectory name!**

DBFS Target Directory ⓘ

/FileStore/tables/

Files uploaded to DBFS are accessible by everyone who has access to this workspace. [Learn more](#)

Files ⓘ

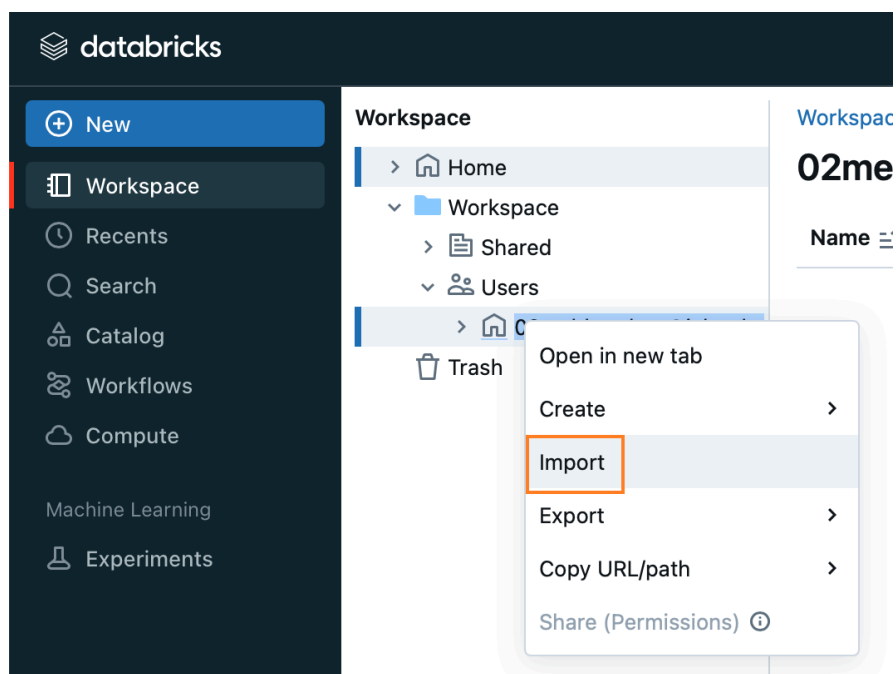
File Name	Size	Action
Users.json	0.1 MB	Remove file
Upvotes.json	6.9 MB	Remove file
Records.json	1.3 MB	Remove file
Reviews.json	2.2 MB	Remove file
Sessions.json	19.2 MB	Remove file

- ✓ File uploaded to /FileStore/tables/zotmusic_sample/Users.json
- ✓ File uploaded to /FileStore/tables/zotmusic_sample/Records.json
- ✓ File uploaded to /FileStore/tables/zotmusic_sample/Upvotes.json
- ✓ File uploaded to /FileStore/tables/zotmusic_sample/Reviews.json
- ✓ File uploaded to /FileStore/tables/zotmusic_sample/Sessions.json

[Create Table with UI](#)[Create Table in Notebook](#)

The files should now be stored in the cloud under the "/FileStore/tables/zotmusic_sample" and "/FileStore/tables/zotmusic_full" directories, respectively. You are also being provided with a notebook (ipynb) template file that contains some initial environment setup code that will load the files into Spark Dataframes that you can then run queries on. Let's load the provided notebook into Databricks as well by executing the following two steps.

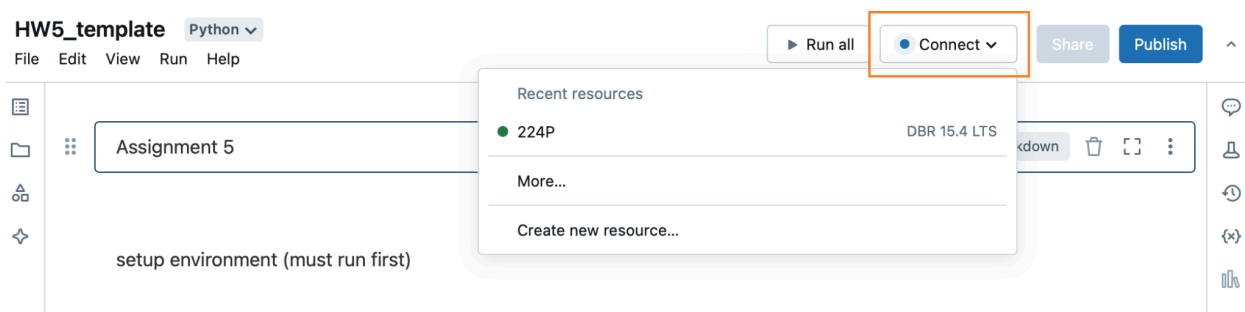
7. Download the *HW5_template.ipynb* file from Canvas to your computer.
8. In Databricks, click on Workspace -> Users -> <user> -> right click and then import, as follows:



9. Click browse and choose the file *HW5_template.ipynb*, or drag and drop the file, and then click import.

10. After importing and opening the HW6 notebook file, run its setup environment cell.

NOTE: Before running the cell, if the dropdown below says “Connect” (like the screenshot below), make sure to first attach the notebook to a cluster! (You can’t compute without nodes to compute on... 😊)

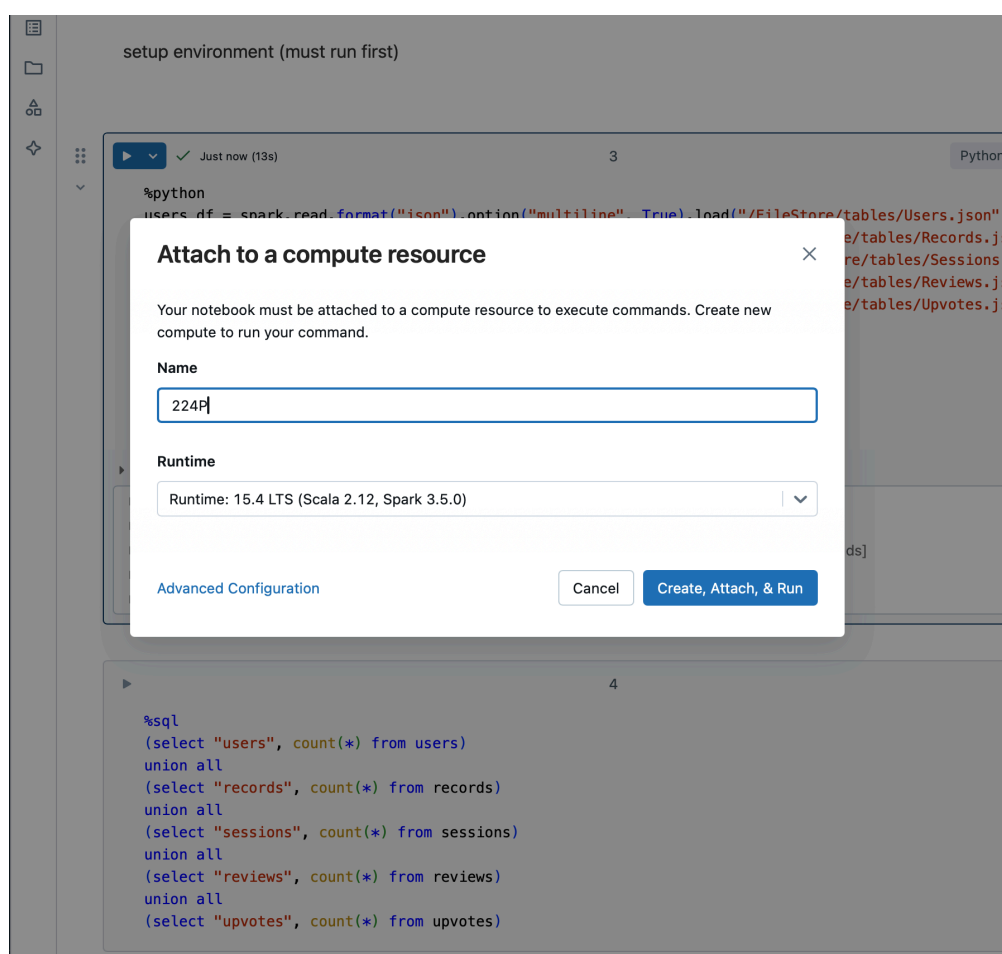


11. If the counting cell succeeds, you are ready to proceed with the assignment itself!

Important Note: Clusters in Databricks Community Edition will self-terminate after being idle for 2 hours, which means you will not be able to run your notebook on that same cluster again.

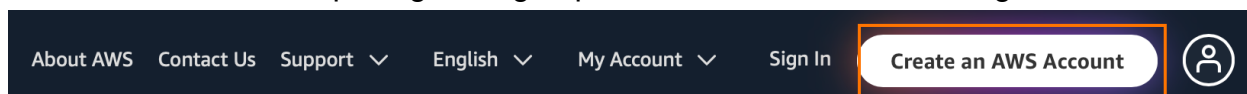
When that happens, you should first **delete** your old cluster, as Community Edition only allows 1 cluster to be created at a time. (Don't worry, the data and notebook will still be there.)

After the deletion, you can either manually recreate the cluster using steps 2-3 and attach the notebook to the new cluster, or you can ask Databricks to create a new cluster for you when running your notebook (**recommended**). When trying to run a cell on a detached notebook, it will ask if you want it to automatically attach and launch a cluster, as shown below:



Part II*(Spark on Databricks Premium Version + AWS)*

1. Go to <https://aws.amazon.com/> and create an AWS account if you do not already have one. When prompted for "AWS account name", you are free to use any name, e.g., "cs224p". You will go through several steps and enter your information. A few tips:
 - a. Keep all the credentials, e.g., email, root username, root user password.
 - b. When going through Step 2 (Contact Information), enter Personal - for your own projects.
 - c. As AWS is not free of charge, you will be asked to enter your billing information at Step 3. Don't worry, you will not be charged now, and will only be billed for a few dollars in order to complete this assignment.
 - d. In the last step, "Select a support plan", choose "Basic support - Free"
 - e. After completing the sign up, click "Go to the AWS management console."

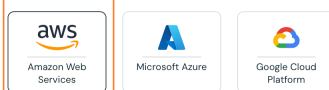


2. In another browser tab/window, create a Databricks AWS Version account by following this link (<https://databricks.com/try-databricks>) and complete the following steps:
 - a. Fill out the registration form in the first dialog in order to get to the next account setup dialog. For the Company name you can type UCI, and for work email please provide your UCI email.

How will you be using Databricks? 2/2

Professional use

Pick your cloud provider. You'll need admin access to your cloud account to get started.



Enjoy \$400 in credits during your 14-day AWS trial. Trial ends when credits expire.

By clicking "Continue," you agree to Databricks' [Terms of Service](#).

Continue**Personal use**

Community Edition is a limited, single node version of Databricks for personal or educational use.

By clicking "Get started with Community Edition," you agree to Databricks' [Terms of Service](#).

Get started with Community Edition

- b. Once you click "Continue" and see the account setup dialog as shown in the above screenshot, choose "Amazon Web Services" and click "Continue".
 - c. You will receive an email back for "Welcome to Databricks!" Click "Log in to get started", and you will need to enter an email verification code. After that you will enter your premium Databricks account.
3. Enter the workspace info as shown in the following screenshot.



Let's set up your first workspace

Enter a workspace name, select your AWS region, then start the quickstart. We'll send you to your AWS Console where a prepopulated template will create a new IAM role and S3 bucket and deploy the Databricks workspace.

If your company already has a contract with Databricks, talk to your company contact before you create your workspace to ensure any negotiated discounts are applied. For troubleshooting help, email us at onboarding-help@databricks.com.

Workspace name

Human-readable name for your workspace

AWS region

Oregon (us-west-2)

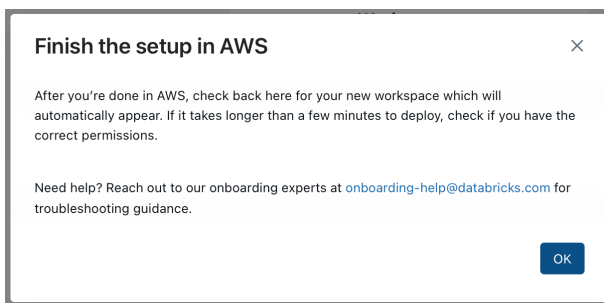
Start quickstart 

By clicking "Start quickstart" you agree to Databricks [terms and conditions](#) and you begin your 14-day free trial. After the trial ends you will be charged at the [list rates](#).

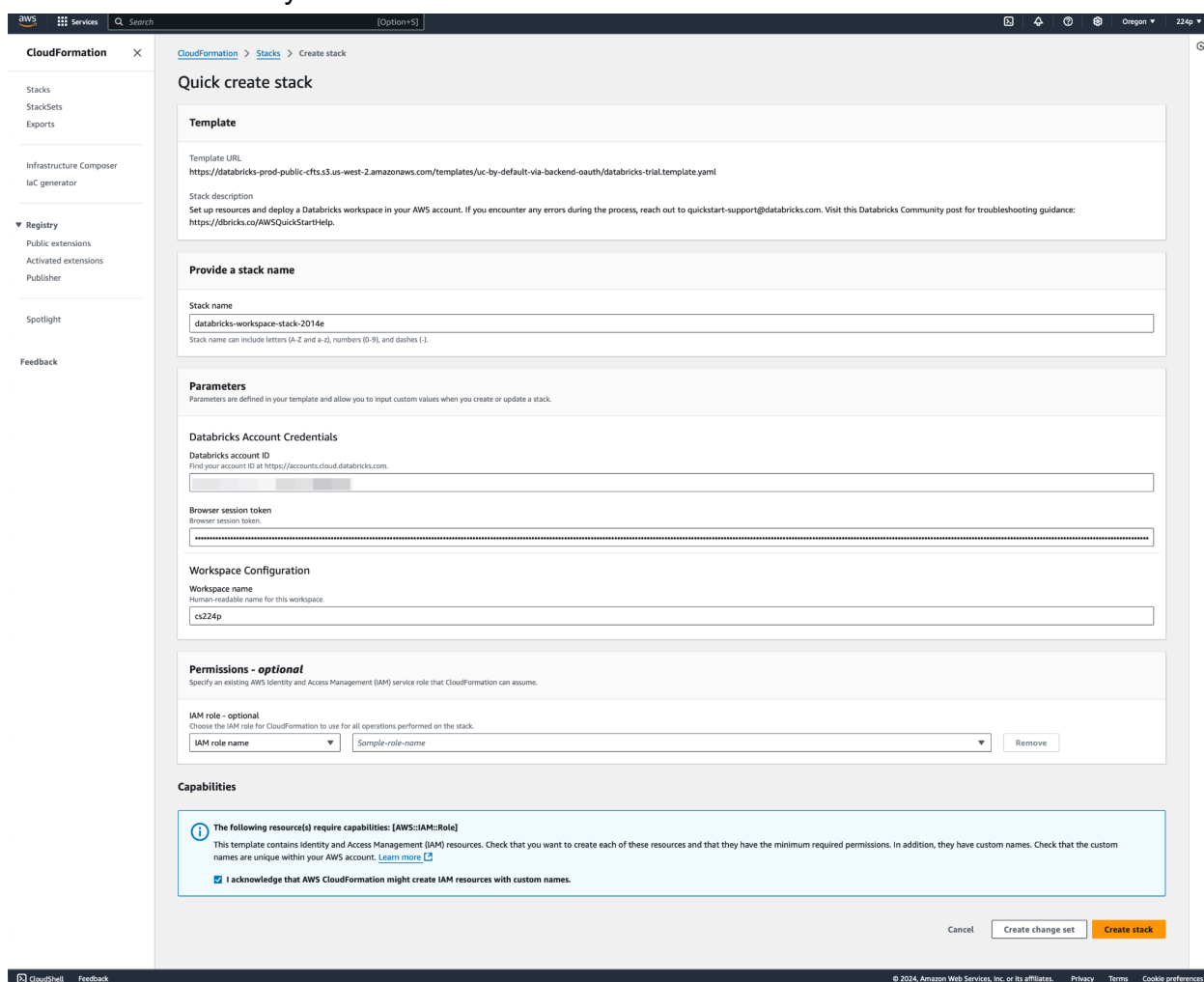
AWS will charge you for AWS resources used during and after the free trial (for example, instance costs).

Your workspace will be created on the Premium plan or [change to Enterprise plan](#).

4. After you click "Start quickstart", you will be redirected to an account management page and be prompted with the following dialog. **Another pop-up browser window/tab should appear.** If not, check if the pop-up has been disabled or blocked by your browser. Make sure the pop-up window goes through, and that you have already logged on to your AWS account.



5. In the pop-up window/tab, you will see the following. Click the acknowledge button and proceed to Create stack. Now Databricks will access your AWS resources on your behalf.

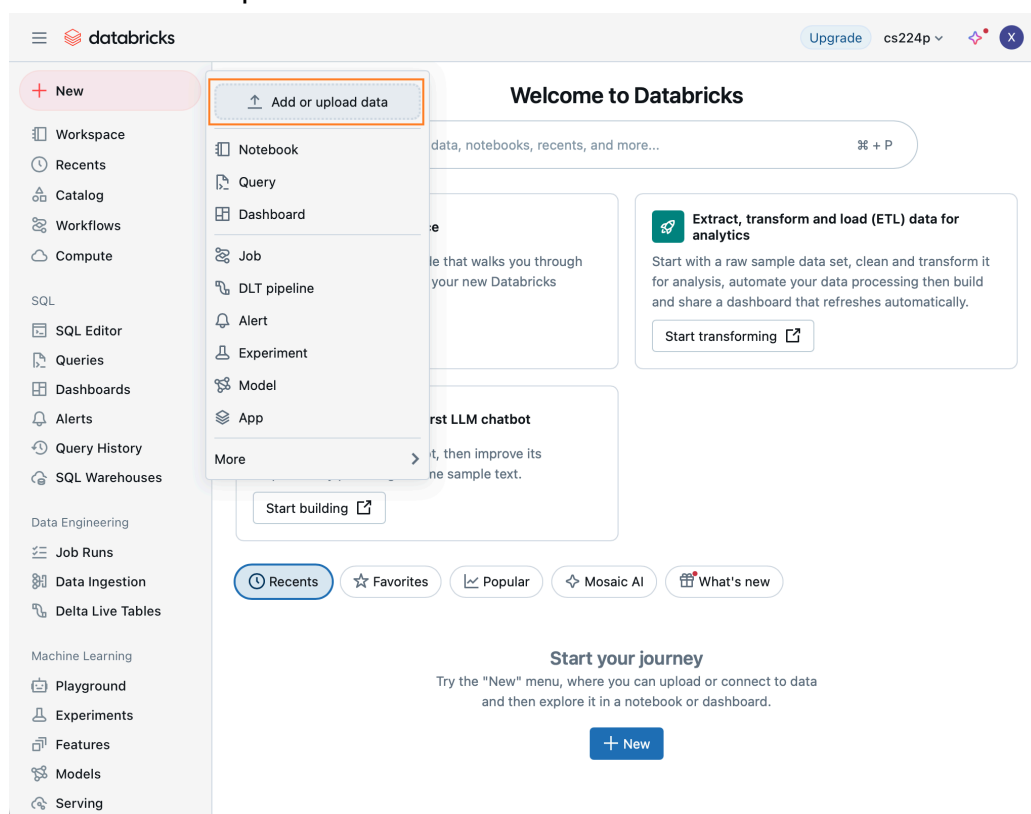


6. The databricks workspace stack will take a few minutes to finish. You will see CREATE_IN_PROGRESS status in AWS, and in your Databricks account console, you will see your new workspace being in "Provisioning" status.

7. After the workspace is ready (you will also receive an email indicating that), refresh your Databricks account console. You will see your new workspace as "Running". Click "Open" to be redirected to this workspace.



8. In the new tab for this workspace, you will see a new but familiar UI (looks like the one from the Community Version but with many more new features). Click "+New->Add or upload data".



9. Go to the bottom and click "Upload files to DBFS". Again, enter "" in the Target Directory, and upload the files from the full dataset. It will likely take a while for all 5 files to be uploaded. Patiently wait for all of them to be uploaded.

[Add data >](#)**DBFS**Upload File

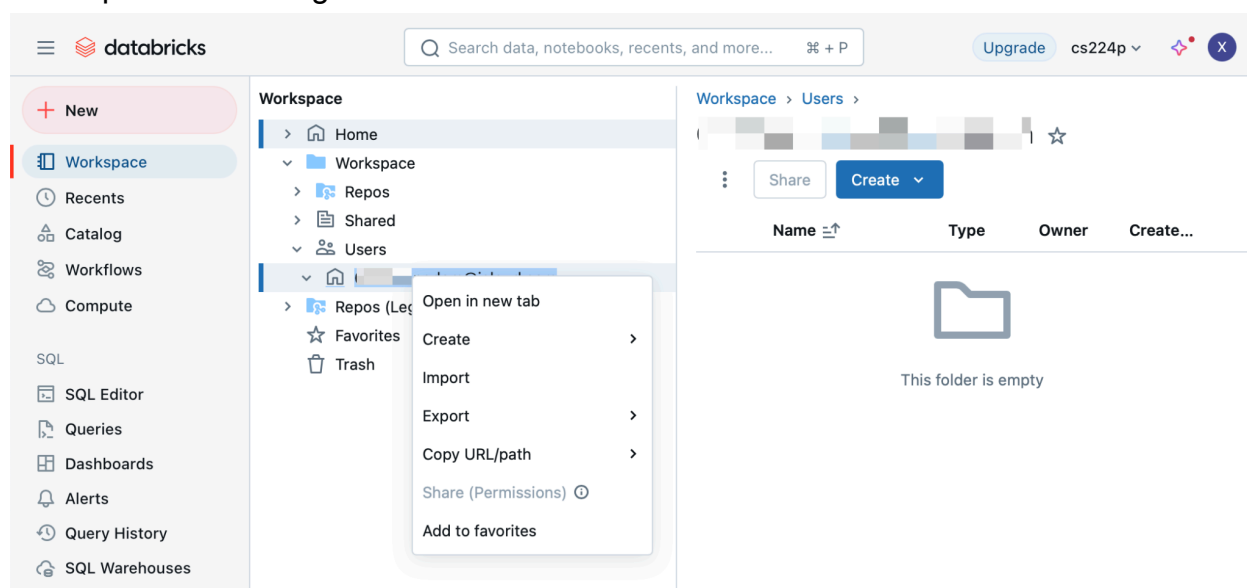
DBFS Target Directory ?

/FileStore/tables/ Files uploaded to DBFS are accessible by everyone who has access to this workspace. [Learn more](#)

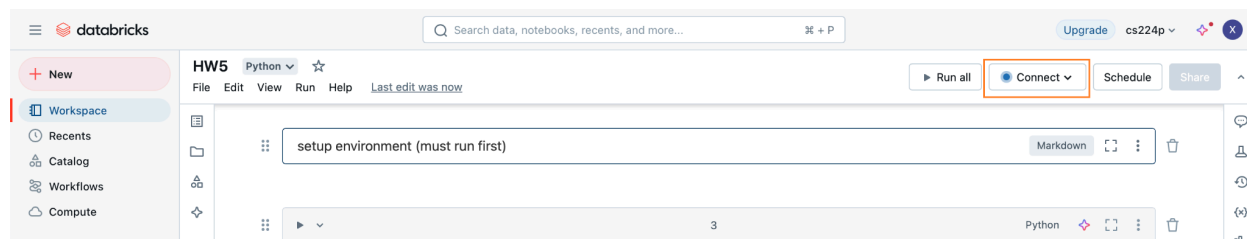
Files ?

Drop files to upload, or click to browse

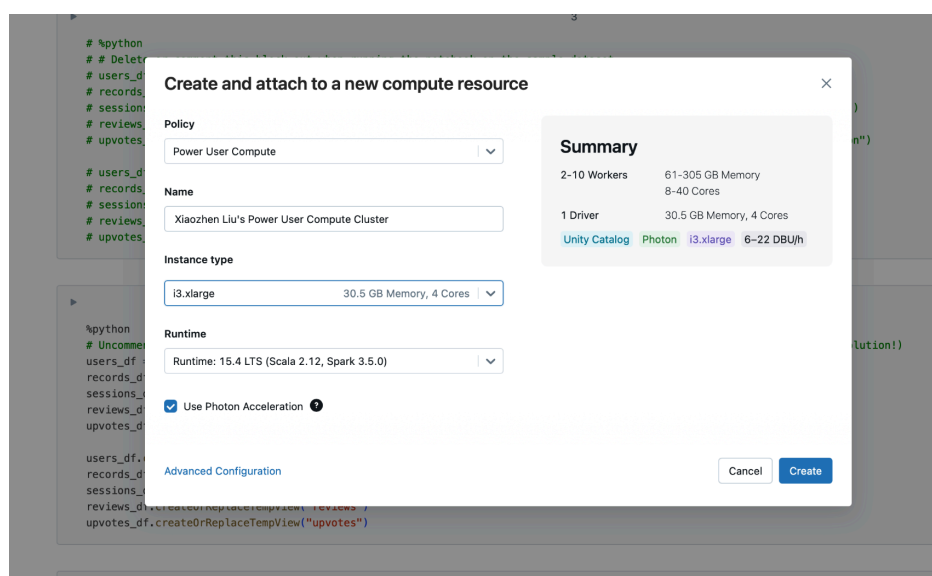
10. Go back to the dashboard, and from "Workspace->Users-><Username>", right click, and use "Import" to upload the notebook you have finished from the first part of the assignment.



11. Click your uploaded notebook, and click "Connect".



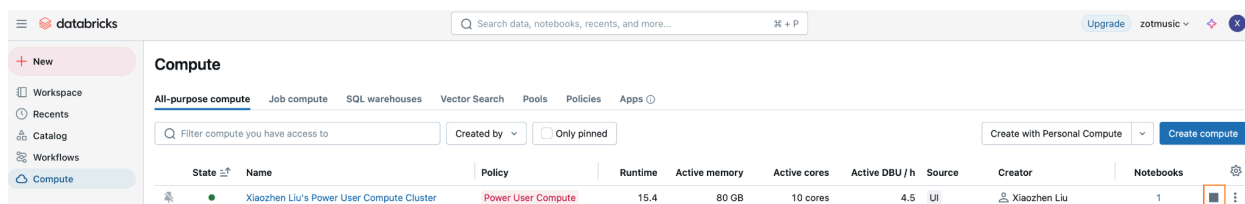
12. In the dialog, click "Create new resource", and follow these configurations (make sure to select **Power User Compute** to allow multiple worker nodes; for other configurations like number of workers and instance type, you are free to explore other options). Click "Create" to let Databricks utilize your AWS account to create a new compute cluster.



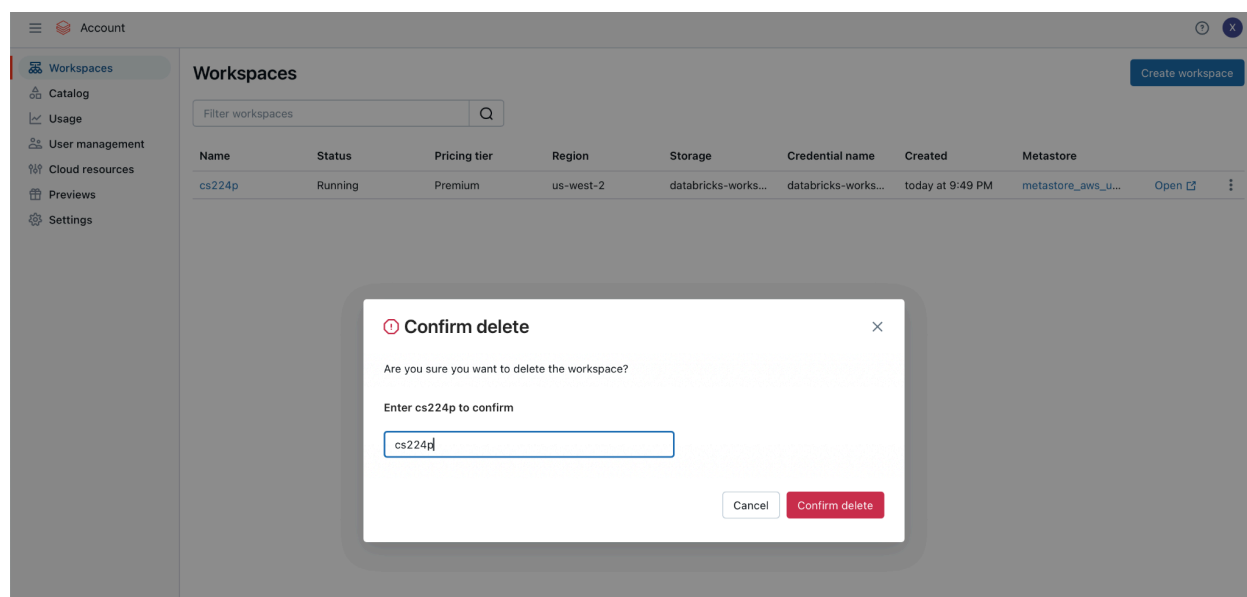
13. Wait a few minutes for the compute cluster to finish creation. After that, click "Run all" in the notebook and observe the execution process, and answer part II of HW5 accordingly. Pay special attention to the execution time of each cell like this:



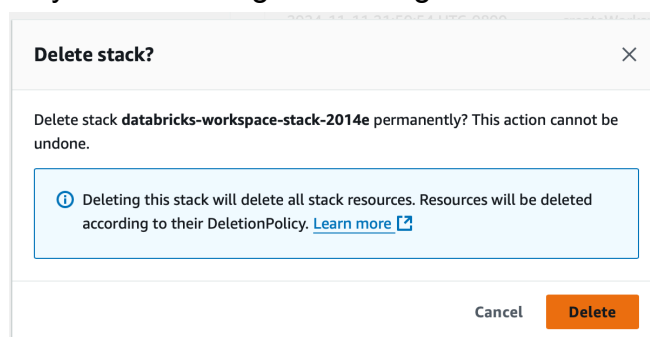
14. After you have finished everything and are certain you will not run any executions on the notebook again, go to Compute and terminate the cluster.



15. To ensure the workspace does not incur any more AWS cost, go back to the Databricks account console and delete the workspace.



16. Also, in your AWS CloudFormation console, delete the Stack created by Databricks. Now you will no longer be charged in AWS.



17. To view the details of your billing history, in AWS, go to account name -> Billing and Cost management (note: it usually takes 24 hours for costs to be updated)

