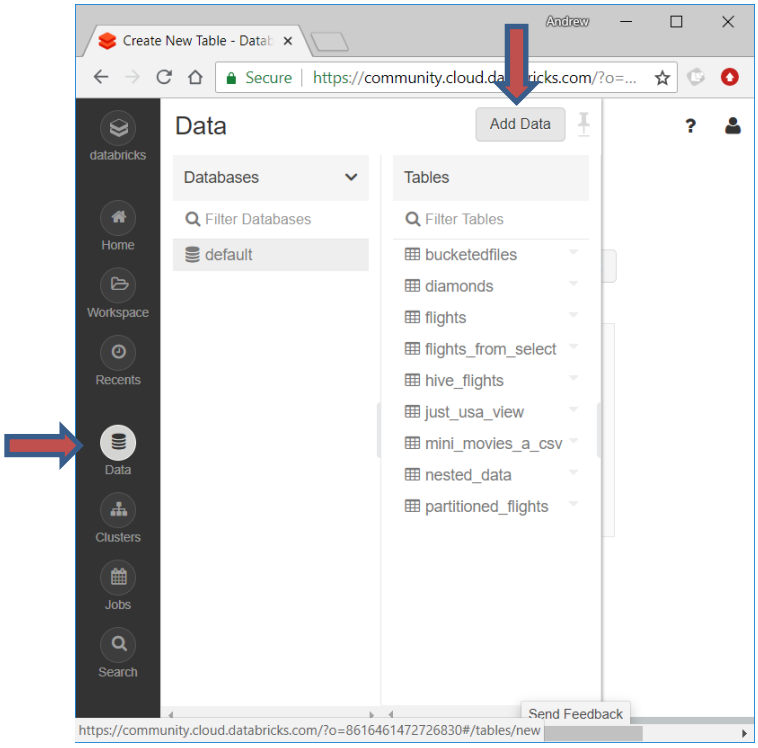
CSS 735: HW 2

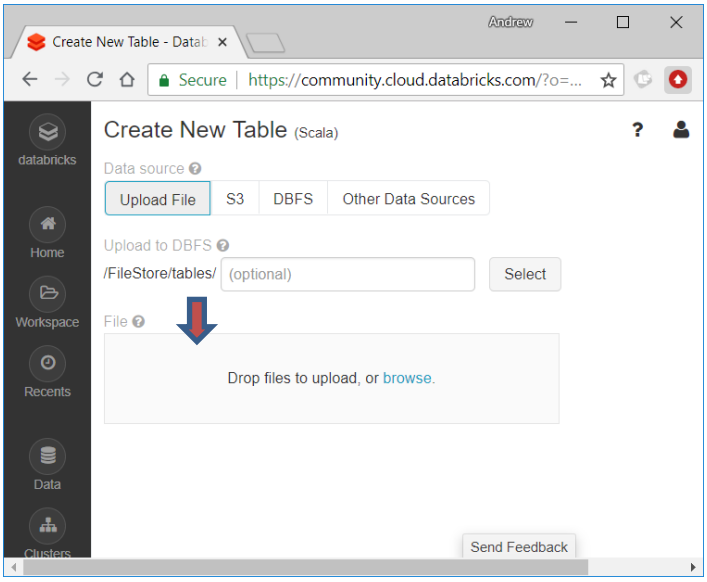
Due Date: *Oct 13 (F) 11:59 PM* ***Overdue submissions*****:** subject to penalties

Download the file **movies.csv** from Blackboard. Login to your Databricks Community Edition account. Follow these instructions to upload the file movies.csv to your Databricks account:

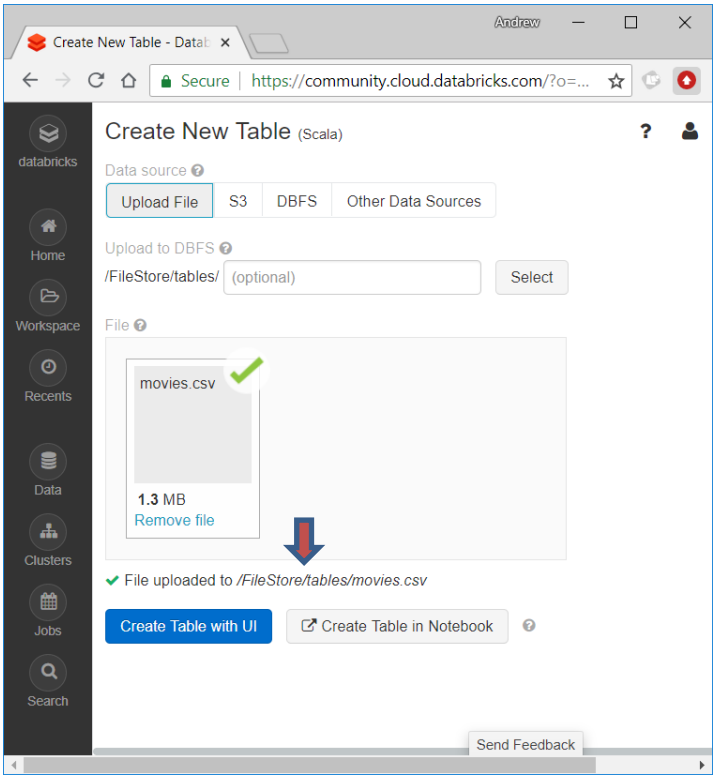
1) Create a cluster
2) Click on the **Data** icon in the vertical navigation bar on the left side of the page
3) Click on the **Add Data** button located near the top-right corner of the Data worksheet.



4) At the bottom of the landing page, you will see an area labeled **File** (see image next page). You can either **drag and drop** files into this area or click the **browse** link to browse and select a file from your computer to upload.



5) You should see a green check mark indicating that your file has been uploaded successfully.
6) Pay attention to the location where your file has been uploaded because this will be the path that you will need to use in your Spark code to load the file contents into a DataFrame.



7) Now, you are ready to create a Notebook for this homework assignment. The file that you will eventually upload to Blackboard will be named hw2.scala.
8) Use this command to load the dataset

```
val df = spark.read.option("header", "true")
  .option("inferSchema","true")
  .csv("/FileStore/tables/movies.csv")
```

9) Execute the following commands to ensure that the data has been uploaded successfully.

```
df.show(false)
df.count()
// res6: Long = 31394
```

10) Use the following command to register the DataFrame as an SQL table:

```
df.createOrReplaceTempView("movies_table")
```

11) **[20points]** Write an SQL-based Spark code to compute the number of movies produced in each year. The output should have two columns for year and count. The output should be ordered in ascending order by year. It is alright that Spark will only display the first 20 rows.
12) **[20 points]** Write DataFrame-based Spark code to do the same thing as in the previous item.
13) **[20 points]** Write an SQL-based Spark code to find the five top most actors who acted in the most number of movies. Schema of the output must be (actor, number_of_movies), or in other words, rename the column with the count as number_of_movies.
14) **[20 points]** Write DataFrame-based Spark code to do the same thing as in the previous item. Make sure that schema of the output is (actor, number_of_movies).
15) **[20 points]** Write DataFrame-based Spark code to find the title and year for every movie that Tom Hanks acted in (the name is stored as Hanks, Tom in the csv file). Make sure that the output is sorted in ascending order by year. Notice that schema of the output must be (title, year). This means that actor name should not be part of the output.

Deliverables

- All files to be submitted should be placed in a single directory and **zipped** together into a single file identified by your name and student ID (**YourFamiliyName_StudentID.zip**) for uploading to Blackboard.
- Submit your **notebook** and the corresponding **HTML** file named in this format: Asg2_YourFamiliyName_StudentID.**scala**, Asg2_YourFamiliyName_StudentID.**html**

- **Ensure that your notebook can run from start to finish without crashing** (e.g. click Run=>Run all and double check the output of each cell before submitting)
  To download, click File=>Export=>Source file, and File=>Export=>HTML
- To prove that your code works correctly, you will also submit screenshot of the output of your program in a file with an extension of **.png** or **.pdf**.

Policies and procedures

- Assignments must be completed **individually and independently,** and **NOT** in a group.
- Use comments to document and explain your code where needed.
- Assignments are to be handed in by the **due date**; otherwise, **penalty per day** will be incurred: 1 day late 10% penalty; 2 days late 20% penalty; after 2 days no credit will be given for a late assignment. Homework that is habitually late will not be accepted.
- Any incomplete work that is submitted will be considered for partial marks.
- If exceptional circumstances have occurred that have kept you from submitting your assignment on time, you should contact the course instructor as soon as possible. If you are ill, a valid **document**, e.g., a doctor note, must be provided.