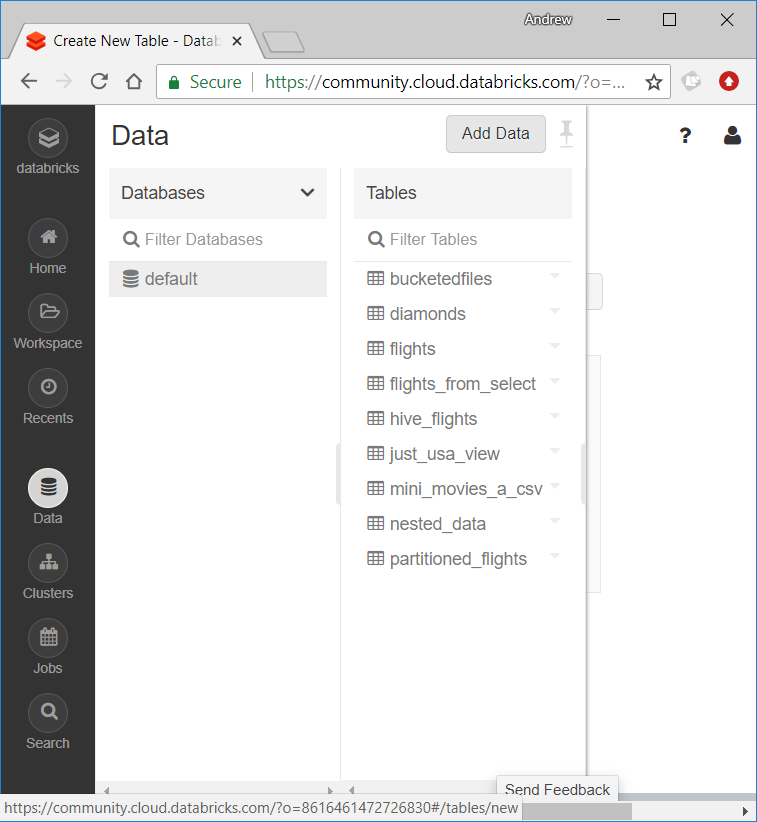
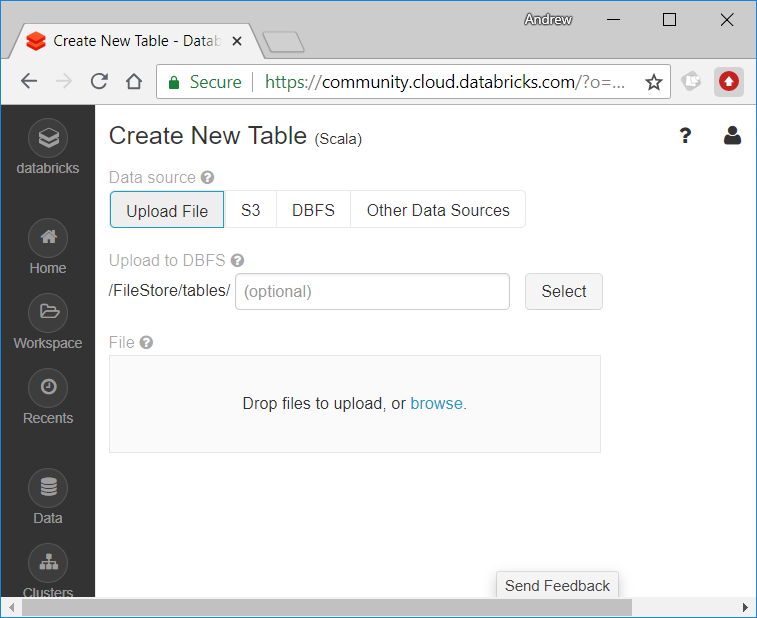
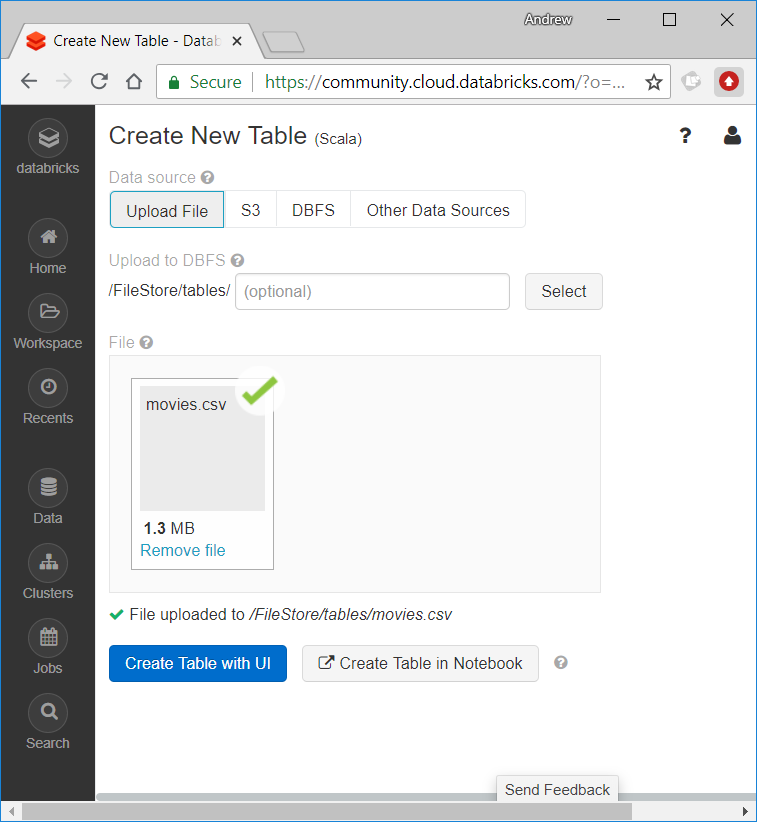
CSC 735 – HW 2  
Due Date: Thursday, October 25, 2018 at 9:30 am

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.



1. 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.   
     
     
   
2. You should see a green check mark indicating that your file has been uploaded successfully.
3. 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.   
     
   
4. 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.
5. Use this command to load the dataset  
     
   val df = spark.read.option("header", "true")   
    .option("inferSchema","true")  
    .csv("/FileStore/tables/movies.csv")
6. Execute the following commands to ensure that the data has been uploaded successfully.  
     
   df.show(false)  
   df.count()   
   // res6: Long = 31394
7. Use the following command to register the DataFrame as an SQL table:  
     
   df.createOrReplaceTempView("movies\_table")
8. **[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.
9. **[20 points]** Write DataFrame-based Spark code to do the same thing as in the previous item.
10. **[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.
11. **[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).
12. **[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.
13. **[Extra Credit 10 points]** Think of a nontrivial data analytics task that you can perform on the movies.csv dataset which you can solve using only material that we have studied by October 9, 2018. State the task and provide an SQL-based and DataFrame-based solutions for it. The number of points that you will receive will be based on how interesting and nontrivial your task is. Please do not e-mail the instructor to ask if a certain idea is good. I won’t reply to such e-mails. As a graduate student, you should be able to judge.

**What to turn in**: upload to Blackboard a file named **hw2.scala**. This file will contain your solution for items 11, 12, 13, 14, 15 and maybe 16 (if you did the extra credit). Please use a comment before each item so that it is easy to identify the solutions for the different questions. Also use comments to document and explain your code where needed. Make sure to **also turn in a stapled hard copy of your solution in class on the due date**. Write your name at the top of your file.