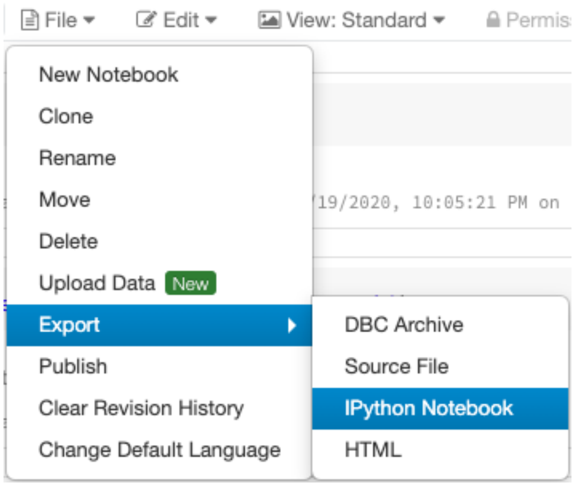
DS 610 Week 2 Assignment
Big Data Analytics

Due Date:

Please Note: As you will be working on Databricks console for this assignment, please submit the IPython Notebook [File -> Export -> IPython Notebook]. Please don't submit screenshots / word documents or any other format.



1.
Using databricks console and 'walmart_stock.csv', Create a table called *walmart_stock.* Note that the file has first row as the column names. Also, using your best judgment, assign the datatype to each column. Submit the screenshot of the resultant screen after you click on **Create Table** showing the 'Schema' and the 'Sample Data'.

2.
Explain the difference between SQLContext & SparkSession.

3.
Using **sqlContext**, create a DataFrame object holding the number of rows in the table created in #1. Show the content of the object on console.

4.
Using **collect()** & **asDict()**, grab & print only the number of rows in the table created in #1 on console.

5.
Using **sqlContext**, create a DataFrame object holding the first 10 rows in the table created in #1. Show the content of the object on console.

Created by Deepak Agrawal

6.
Using **sqlContext**, create a DataFrame object holding the **Date** for the maximum **Volume** in the table created in #1. Show the content of the object on console.

7.
Using **sqlContext**, create a DataFrame object holding the number of rows for which **Open** is less than **Close** in the table created in #1. Show the content of the object on console.

8.
Using **sqlContext**, create a DataFrame object holding the first 10 rows order by **High** in descending order in the table created in #1. Show the content of the object on console.

9.
Using **sqlContext**, create a DataFrame object holding the stock details for the month of January in the year 2015 in the table created in #1. Show the content of the object on console.

10.
Using **sqlContext**, create a DataFrame object holding the content of table created in #1 along with one more column called 'cost' where 'cost' is 'Very Expensive' if the **High** is greater than 85, 'Expensive' if the **High** is greater than 70 but less than or equal to 85, 'Not Expensive' otherwise. Show the complete content of the object on console [not only the first 20 rows].

Thank you.