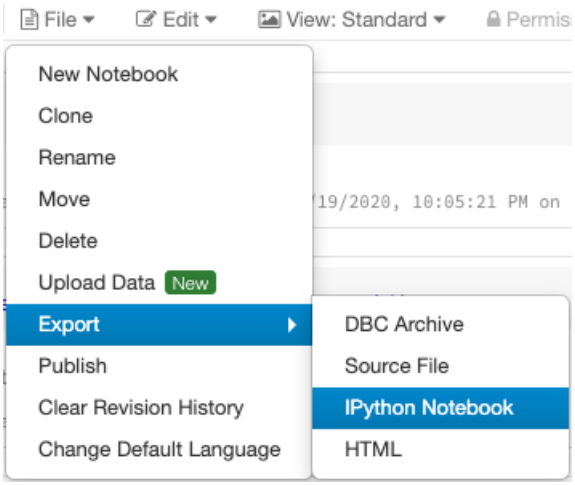
DS 610 Week 3 Assignment
Big Data Analytics

Due Date:

Please Note: As you will be working on Databricks console for this assignment, please submit the IPython Notebook [File-> Export-> IPython Notebook]. Use Markdown. Submissions in the form of screenshots / word documents or in any other format will **NOT** be evaluated.



1.
Compare & contrast Apache Spark with MapReduce.

2.
What are the important components of Apache Spark ecosystem?

3.
Why the Transformation is lazy in Apache Spark?

4.
Create a sample file called *marks.json* with the keys as ***name*** and ***marks***. Enter marks for 5 students. Explicitly define ***name*** as *StringType* and ***marks*** as *FloatType*. Show the initial & final output of *printSchema*().

5.
Explain *withColumn()* and *withColumnRenamed()* with the help of sample data and PySpark code.

6.
Using *SparkSession*, select & print only the ***marks*** column from *marks.json*.

7.
Using *SparkSession*, collect & print only the name of fourth student from *marks.json*.

8.
For all the names in *marks.json*, append 'LNU' (**L**ast **N**ame **U**nknown) to the ***name*** field and print the output.
Sample Output:
<first_name_1> LNU
<first_name_2> LNU
<first_name_3> LNU
<first_name_4> LNU
<first_name_5> LNU

9.
Explain the difference between *show()* and *collect()* with the help of sample data and PySpark code.

10.
For *marks.json*, create a new column called *scaled_marks* defined as 1.2 times of original marks and print the output.

Thank you.