Assignment 1 – Read read/write data from/to HDFS

Due date: Friday 4/16/2021 at 11:59PM Pacific Time

Remember – when in doubt, read the documentation first. It's always helpful to search for the class that you're trying to work with, e.g. pyspark.sql.DataFrame.

Resources:

- PySpark API: https://spark.apache.org/docs/latest/api/python/index.html
- Spark DataFrame: https://spark.apache.org/docs/latest/sql-programming-guide.html
- PySpark_SQL_Cheat_Sheet_Python.pdf uploaded on canvas

Tasks:

1. (From JupyterLab terminal) Copy data file(BookReviews_1M.txt) from local FS to HDFS
2. Start spark session
3. Read data from HDFS into Spark DataFrame
4. Print number of lines read in
5. Show first 20 lines using `pyspark.sql.DataFrame.show`
6. Stop spark session

Instructions for submission:

1. Save the Python notebook as Python file (.py) – File > Export Notebook as > Executable Script
2. Submit both the Python notebook(.ipynb) and Python file(.py) on Gradescope