CUSP-GX-5008.001: Big Data Management & Analysis

SPRING 2016

Homework 6 – Apache Spark

For this homework, we will be practicing Apache Spark with joining multiple data sets. This will be using a very similar approach as the Pig example (i.e. *find the top 10 most visited pages for each web category*) that we discussed in class a few lectures ago. However, we'll be using NYC open data sets instead of web clicks, in particular: the ***SAT Results*** and the ***NYC High School Directory*** data sets. Both are attached for your convenience.

DATA SET:

SAT_Results.csv
Source: https://nycopendata.socrata.com/Education/SAT-Results/f9bf-2cp4
Description: "The most recent school level results for New York City on the SAT. Results are available at the school level for the graduating seniors of 2012."

DOE_High_School_Directory_2014-2015.csv
Source: https://data.cityofnewyork.us/Education/DOE-High-School-Directory-2014-2015/n3p6-zve2
Description: "Directory of NYC High Schools."

Please note that each school is uniquely identified by an DBN code, which should be found on both data sets.

OBJECTIVE: (10 pts)

You are asked to compute the average SAT Math score of all high schools with 500 students or more, for each borough of the city. Meaning: what is the average SAT Math score of all high schools with 500 students or more in Manhattan, in Brooklyn, in Queens, in Bronx and in Staten Island.

You must use Apache Spark for this assignment. Both data sets must be loaded into RDDs, where all your manipulations must be applied on, though you are free to transform these RDDs into **Spark**'s DataFrame or SQL Context. The final result is expected to be a list of tuples borough names as the first elements, and the average scores as the second.

Note 1: since the SAT Results also provide the number of test takers along with the average scores, you should use this information in computing the exact average scores above.

Note 2: if a DBN in the SAT Results data set is not found in the High School Directory, you can safely ignore the test scores for that school.

Your submission: you can turn in either a PySpark notebook (as **NetID_hw6.ipynb**) if you do your homework on your local machines, or a Python application (as **NetID_hw6.py**) if you are using CUSP's or NYU HPC's cluster. For sanity check, please include the list of tuples in the body part of your submission.

EXTRA CREDITS 1: (5 pts) – ***to be due with (and as a part of) this assignment***
We would like to know how the Math scores vary across bus lines or subway lines serving the schools. Your task is to compute the average Math scores of all schools along each bus line and subway line. You can find the bus and subway lines serving each school in the High School Dictionary as **bus** and **subway** columns.

The expected results are two lists:

1. A list of key/value pairs: with **bus** line as keys, and the average Math scores as values.
2. A list of key/value pairs: with **subway** line as keys, and the average Math scores as values.

You are also required to use Spark for this extra credit.

Your submission: you can turn in either a PySpark notebook (as **NetID_hw6_extra1.ipynb**) if you do your homework on your local machines, or a Python application (as **NetID_hw6_extra1.py**) if you are using CUSP's or NYU HPC's cluster. For sanity check, please include the two lists in the body part of your submission.

EXTRA CREDITS 2: (5 pts) – ***to be due in two weeks as a separate entry on NYU Classes***
Complete the main objective of this assignment using only **Pig**. The expected result should be written to the standard output instead of storing to files.

Your submission: you must turn in your Pig script (as a text file **NetID_hw6_extra2.pig**), where we can just directly copy and paste your script into the HUE platform; and expected to get results on the standard output. Remember, unlike Extra Credit 1, there will be a separate entry for this extra credit on NYU Classes.