The City College of New York | Department of Computer Science

CSC 445: Big Data Management and Analysis
FALL 2020

Lab 5 – Apache Spark

For this lab, we will be practicing Apache Spark with joining multiple data sets. We'll be using two NYC open data sets: the ***SAT Results*** and the ***NYC High School Directory*** data sets. Both can be downloaded from the links below, or from NYU Classes for your convenience.

DATA SET: *Please note that each school is uniquely identified by an DBN code, which should be found on both data sets.*
SAT_Results.csv (also available on Blackboard)
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 (also available on Blackboard)
Source: https://data.cityofnewyork.us/Education/DOE-High-School-Directory-2014-2015/n3p6-zve2
Description: "Directory of NYC High Schools."

TASK 1:

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. 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.

TASK 2:

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.