HW4 – Hadoop data Analysis

Due Date : April 23rd

On the Hadoop cluster, I have uploaded a file named “nyc-traffic.csv” at the following HDFS location:

- /tmp/nyc.data

This data was collected from the City of New York’s data website, and contains all reports of vehicular incidents in New York City over a period of time. The file is roughly 175MB in size, and contains over 900,000 records.

There are a considerable number of fields, including columns with a common format that describe up to 5 vehicles that contributed to the particular incident.

Using the Hadoop streaming API (the one we demonstrated in class using the Python scripts, but you may use any similar script that can be invoked in a similar manner using STDIN and STDOUT), build mapper and reducer scripts that analyze the data and yield summary counts for each vehicle that describe the total count, per vehicle type, that the vehicle type was involved in an incident. If the same type of vehicle was involved more than once in an incident, count the vehicle twice for the purpose of the summary statistic.

Please refer to https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-VehicleCollisions/h9gi-nx95 for more information about the dataset.

The discussion we had in class should come in handy:

https://piazza.com/class_profile/get_resource/jqx71md1irpi8/jt6bq54ho6j3mb

http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/

You can also use JAVA map reduce API if you prefer to. (No extra credits though).

As always, use GitHub to host your code and add myself as a contributor to the repository

Submit, to blackboard:

- GitHub URL
- Text of the summary statistics output
- Command line(s) you used on the Hadoop server to execute your program and the corresponding log.

Github repository should include the following:

● Mapper & reducer scripts (Python or whatever you preferred)

● Any shell script you used to execute the above

● README.md that explains how, after I clone the repository into CSCloud, I can run your analysis

A successful mapreduce command looks like below

hadoop jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar -file /home/$USER/py_mapper.py -mapper /home/$USER/py_mapper.py -file /home/$USER/py_reducer.py -reducer /home/$USER/py_reducer.py -input /tmp/4300-0.txt -output /tmp/hadoopstreamingoutput/