CENTER FOR URBAN SCIENCE+PROGRESS

NEW YORK UNIVERSITY

CUSP-GX-5008.001: Big Data Management & Analysis

SPRING 2016

Homework 5 – Hadoop Streaming

The objectives of this homework is to get used to Hadoop environment, especially Hadoop Streaming. There is only one task in this homework that is to compute the median trip time at the minute interval using the CitiBike's data set that we have used in previous labs and assignments. The data set has already been uploaded onto HDFS at CUSP as well as at HPC. Below are their locations:

On CUSP HDFS	: /data/share/bdm/citibike.csv
On HPC HDFS	: /tmp/citibike.csv

Please note that both of these files have already had their headers removed to avoid issues in parsing their headers. Their stripped headers are included here for your convenience:

```
cartodb_id,the_geom,tripduration,starttime,stoptime,start_station_id,start_station_name,start_station_latitude,start_station_longitude,end_station_id,end_station_name,end_station_latitude,end_station_longitude,bikeid,usertype,birth_year,gender
```

For this assignment, we are only interested in the `tripduration` field.

Your task: please write one or many **mapper.py** and **reducer.py** files, where you can supply to the Hadoop Streaming framework (you may also use the run.sh script that was accompanied with Lab 5) to produce a median with one of the two above input file on CUSP or HPC cluster.

Your submission: you are to submit all files and the steps needed to run your Hadoop Streaming job(s) as required above.