Lab 6: Hadoop Streaming API with Python

1. Outline

In this lab, you will perform (1) A MapReduce job using Hadoop Streaming API using Python for counting the frequency of unique words in a document. (2) Generating density a density map with New York Taxi data (using the records of January, 2013).

2. Materials

The data and scripts are stored in: /gpfs\_scratch/geog479/lab6

3. Tasks

Task 1:

* Login to ROGER and copy data to your home directory:
  + >> ssh [NetID@roger-login.ncsa.illinois.edu](mailto:NetID@roger-login.ncsa.illinois.edu)
  + >> cp -r /gpfs\_scratch/geog479 ~/
* Login to cg-hm08
  + >> ssh cg-hm08
* copy data into HDFS
  + >> cd lab6/word\_count\_hadoop\_python
  + >>hdfs dfs -copyFromLocal const.txt
* Run the word count example
  + >> hadoop jar /usr/hdp/2.3.2.0-2602/hadoop-mapreduce/hadoop-streaming-2.7.1.2.3.2.0-2602.jar **-file** mapper.py **-mapper** mapper.py **-file** reducer.py **-reducer** reducer.py **-input** const.txt **-output** results.txt
* View the results
  + >> hdfs dfs -getmerge results.txt results.txt
  + use nano to view the file
* Now, view the details in **mapper.py** and **reducer.py** respectively
* Test the mapper and reducer code locally
* Test the mapper:
  + >> echo "This is a great day (yes, a great day), but we are sitting inside doing coding" | ~/lab6/word\_count\_hadoop\_python/mapper.py
* Test the reducer:
  + >> echo "This is a great day (yes, a great day), but we are sitting inside doing coding" | ~/lab6/word\_count\_hadoop\_python/mapper.py | sort -k1,1 | ~/lab6/word\_count\_hadoop\_python/reducer2.py
* Remove the data in HDFS and run modified script
  + >> hdfs dfs -rm -r results.txt const.txt

Task 2: Generating a density map of taxi pick-ups in New York during January, 2013.

* Get the data from ~/lab6/ny\_taxi: >> cd ~/lab6/ny\_taxi
* Load the data into HDFS
  + >> hdfs dfs -copyFromLocal ny\_taxi\_1.csv
* Run the script
  + >> cd straming\_py
  + >> ./program.sh 2013-01-01 2013-02-01 40.479636 40.930724 -74.402322 -73.630027 0.005 0.005 taxiImage\_yourname.asc
* Generate a TIFF image for the result
  + ./plotTaxi
* View the results
  + Using remote Firefox
  + scp the results to your local computer and view it with QGIS or ArcGIS
* **Understand the code!**
  + **What happened?**
  + **What are different and common between the word count example?**