Lab 5: Taming Big Geospatial Data with Hadoop

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) A MapReduce job using Apache Pig to extract Twitter data of Chicago from the data covering the entire North America.

2. Materials

The data and scripts are stored in: /gpfs\_scratch/geog479/lab5

3. Tasks

* Login to cg-hm08, which is the master node of the Hadoop cluster, and make sure your home directory in HDFS has already been created
  + ssh [NetID@roger-login.ncsa.illinois.edu](mailto:NetID@roger-login.ncsa.illinois.edu)
  + ssh cg-hm08
  + hdfs dfs -ls /user/
* copy data into HDFS
  + hdfs dfs -copyFromLocal file\_in\_local\_directory [PATH\_IN\_HDFS]
* 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 [PATH\_IN\_HDFS] PATH\_IN\_LOCAL\_DIRECTORY
  + 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 "foo foo quux labs foo bar quux" | [PATH]/word\_count\_hadoop\_python/mapper.py

* Test the reducer:

echo "foo foo quux labs foo bar quux" | [PATH]/word\_count\_hadoop\_python/mapper.py | sort -k1,1 | [PATH]/word\_count\_hadoop\_python/reducer2.py

* Remove the data in HDFS and run modified script
  + hdfs dfs -rm -r PATH\_IN\_HDFS
* Continue to Apache Pig
* Provided bounding box of Chicago: lower left (-88.707599, 41.201577) and upper right (-87.524535, 42.495775).
* pig **-f** name\_of\_pig\_script **-param** input=name\_of\_file\_in\_HDFS
* Be creative: e.g., keep/drop Twitter message content, switch to other region