Third Homework Assignment (20% of grade)

Due on Monday, May 11 at 5 pm.

Deliver the following in folder /home/public/assignment (you should have write permissions; after you copy every file, execute: chmod 400 file_name)

1. Your java source code file: name the file lastname1 _exercise.java
2. One output file from a reducer: name the file lastname1_ exercise.txt

You need to implement the k-means clustering algorithm.

Assume that:

1) The number of clusters is given in advance.
2) The input files all reside in one folder. Each record corresponds to a data observation. Each column corresponds to a coordinate of the underlying vector. The coordinates are separated by comma.
3) The output must be a set of centroids.

As an external 'tool' you are free to use the tool of your choice, but all mapreduce routines must be written in java. You have to:

- Write a map reduce job that will execute a single iteration of k-means
- External script that will call this map reduce job many times. The script must take the output of the previous iteration, use it as input to map reduce.
- You will have to use the distributed cache concept. Without it it is impossible to do the assignment correctly.

Test data is available in /home/public/course/clustering/clustering.txt

File clustering.txt is a small test problem consisting of 600 records. The more interesting data set is the other one.

The US Government released medicare data. The second data set is this data set. It has 10 million records and it is 1 TB (so don't forget to copy it directly from the local folder to Hadoop – bypassing copying to your local directory; if you want to peek at the data, use 'more' or 'less'). There is an accompanying pdf file that describes the data. I noticed that there are several numerical fields and thus clustering is possible. I challenge you to come up with relevant clustering problems from this data set.

You can read more about the dataset at http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html (Don't download the data again, it is pretty large.)

Do not extract only relevant columns for clustering and then cluster in Hadoop (e.g., by using 'cut'). Be a real hadooper and read in mapreduce the entire data set, and let the mapper process only the relevant fields – you can use an extra preprocessing mapreduce job to filter only the relevant columns.

If someone comes up with interesting results, we'll send them to newspapers (they have been beating to death this data set – I haven't read about clustering studies, but more easy statistics).

In addition to the deliverables specified at the beginning of the document, for the medicare data set, you have to provide a short accompanying document (not more than 1 page). The document must outline:

1) The features selected and the reasoning
2) Insights from clustering (needless to say that you are welcome to use Tableau, R, d3 to produce breathtaking visualizations).