CSC 555: Mining Big Data

Project, Phase 2 (due Tuesday, November 22nd)

The schema is available at:
http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/SSBM1/SSBM_schema_hive.sql

The data is available at (note that the data is |-separated, not comma separated):
http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/SSBM1/ (this is Scale1)
http://rasinsrv07.cstcis.cti.depaul.edu/CSC553/data/ (this is Scale4)
http://rasinsrv07.cstcis.cti.depaul.edu/CSC553/data/Scale14/ (This is Scale14)

Please note what instance and what cluster you are using (you can reuse your existing cluster for most of the questions).

Please be sure to submit all code. You should also submit the command line you use and a screenshot of a completed run (just the last page of output, do not worry about capturing the whole thing). You can use time command to record time of execution of anything you run.

I highly recommend creating a small sample input to test your code first (e.g., by running **head lineorder.tbl > lineorder.tbl.sample** and testing your code with it, you can use head -n 100 to get 100 lines).

Part 1: Data Transformation

Using Scale4 data perform the following data processing. For each part below you should use two of the three choices (Hive, Pig, Hadoop Streaming) to produce the same result.

A. Transform lineorder.tbl table into a csv (comma-separated file)

B. Transform dwdate.tbl table into a csv (comma-separated file)

C. Extract all of the numeric columns from lineorder.tbl into a space-separated file (for K-Means clustering later)

D. Create a pre-join (i.e. a new data file) that corresponds to (all the columns needed for Q2.1 that you ran before). What is the size of the new file?

```
SELECT lo_partkey, lo_suppkey, p_category, s_region, d_year, p_brand1, lo_revenue
FROM lineorder, dwdate
WHERE lo_orderdate = d_datekey;
```

Do not forget to use two different tools for each of the steps above. (Hint: Pig is likely to be easier than Hive, especially for parts A-C).

Part 2: Querying

All queries from SSBM benchmark are available here:

http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/SSBM1/SSBM_queries_all.sql

Using Scale4 data perform the following data processing. For each part below you should use two of the three choices (Hive, Pig, Hadoop Streaming) to produce the same result.

A. Run SSBM query 3.1

B. Run SSBM Query 2.1 using what you have created in 1-D (instead of lineorder and dwdate tables). You would need to rewrite the query accordingly (change the FROM clause to use your new table instead of lineorder and dwdate).

There are only two queries here, but please do not forget to use two methods (Hive is the simplest, of course – but you need a second method).

Part 3: Clustering

Using the file you have created in 1-C, run KMeans clustering using 10 clusters.

A. Using Mahout as you have in a previous assignment (by default it takes input from testdata in /home/ec2-user)

B. Using Hadoop streaming perform two iterations (manually) with randomly chosen input centers. (This would require passing a text file with cluster centers using **-file centers.txt** option, opening the centers.txt in the mapper with open('centers.txt', 'r') and choosing a key to each point based on which center is the closest to each particular point. Reducer simply needs to compute the new center of all of the points it receives under the same key). You will then collect the new centers and repeat the iteration again. You only need to perform 2 iterations and you can do that manually.

NOTE: if you get a java.lang.OutOfMemoryError error, you will need to reconfigure Hadoop to supply the java virtual machine with more memory. You can do this by editing the mapred-site.xml (Mapper should not need much RAM):

```
 <property>
   <name> mapreduce.reduce.java.opts</name>
   <value>-Xmx1024m</value>
 </property>
```

The amount of memory can be tweaked (you can go higher, but keep in mind how much physical memory your machine has).

Do not forget to restart Hadoop after any configuration file change.

Part 4: Performance

Compare the performance given following combinations. If you already ran that combination before (e.g., Query 2.1 in Hive and single-node from previous assignment or 2-B from this work), it is sufficient to copy the runtime for comparison.

A. Query 2.1 in Hive with

   a. Scale1: single node and cluster of at least 4 nodes

   b. Scale4: single node and a cluster of at least 4 nodes

   c. Scale14: cluster of at least 4 nodes

B. Both of your solutions for 2-B.

   a. Scale4: single node and a cluster of at least 4 nodes

   b. Scale14: cluster of at least 4 nodes

Submit a single document containing your written answers. Be sure that this document contains your name and "CSC 555 Project Phase 2" at the top.