CSC 555: Mining Big Data

Project, Phase 2 (due Friday, March 16th)

In this part of the project, you will various queries using Hive, Pig and Hadoop streaming. The schema is available below, but don't forget to apply the correct delimiter:
http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/SSBM1/SSBM_schema_hive.sql
The data is available at:
http://rasinsrv07.cstcis.cti.depaul.edu/CSC553/data/ (this is Scale4)

In your submission, 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 (pig, python and Hive). You should also submit the command lines you use and a screenshot of a completed run (just the last page, do not worry about capturing the whole output). An answer submission with screenshot/results but without the code will not receive credit.

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

Part 1: Data Transformation

Use Scale4 data to perform data processing, unless otherwise specified.

A. Transform lineorder.tbl table into a comma-separated file: Use Hive, MapReduce with HadoopStreaming and Pig (i.e. 3 different solutions)

B. Extract three of the numeric columns (lo_quantity, lo_linenumber, lo_revenue) for rows where lo_discount is between 6 and 8 into a space-separated text file (for K-Means clustering later). Use Hive, MapReduce with Hadoop Streaming, and Pig (3 different solutions)

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 and don't forget to time your results.

A. Run SSBM queries 2.1, 3.3 and 4.3 using Hive only (if you have issues running the queries, try placing lineorder table first in the FROM clause of the query)

B. Create a pre-join (i.e. a new data file) that corresponds to the following query below. You can think of it as a materialized view. What is the size of the new file? Use Hive and Pig (2 different solutions and be sure to report the file size for both).

SELECT lo_partkey, lo_suppkey, s_suppkey, d_year, lo_revenue

```
FROM lineorder, dwdate, lo_supplier
WHERE lo_orderdate = d_datekey and lo_suppkey = s_suppkey;
```

Part 3: Clustering

Using the file you have created in 1-B, run KMeans clustering using 11 clusters.

A. Using Mahout synthetic clustering as you have in a previous assignment on sample data. This entrails running the same clustering command, but substituting your own input data instead of the sample.

   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). If you **still** run out of memory in 3-A submit the screenshot of that change and you will get full credit for the question.

B. Using Hadoop streaming perform three iterations manually (initially with randomly chosen centers). This would require passing a text file with cluster centers using -file option, opening the centers.txt in the mapper with open('centers.txt', 'r') and assigning a key to each point based on which center is the closest to each particular point. Your reducer would then compute the new centers, and at that point the iteration is done and the output of the reducer can be given to the next pass.

   NOTE: Not attempting to answer this question will result in an additional grade penalty

Part 4: Performance

Compare the performance given following combinations.

A. All three of your solutions to Part-1A with

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

B. Both of your solutions for 2-B.

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

C. Summarize the results and cluster performance/scaling in at least a paragraph.

Extra Credit

Research and describe the most affordable way to build a 10-Petabyte drive. The drive should be built to own, not to rent (Dropbox or similar services doesn’t count, even if it does say “unlimited” storage).

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.