CSC 555: Mining Big Data

Project, Phase 2 (due Friday June 8th)

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/ (we will use Scale4)

In your submission, please note what instance and what cluster you are using (you can reuse your existing cluster, but submit the screenshot of it). 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 without associated 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

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

In Hive and Hadoop streaming (but **not in Pig**), you must expand all numeric fields to a minimum of 3 digits (i.e. 0=>000, 2=>002, 28=>028). In Pig, you can keep the numbers as-is.

Part 2: Querying

Implement the following query:

```
select sum(lo_revenue), p_brand1
from lineorder, part, supplier
where lo_partkey = p_partkey
  and lo_suppkey = s_suppkey
  and p_category = 'MFGR#12'
  and s_region = 'EUROPE'
group by p_brand1;
```

using Hive, MapReduce with HadoopStreaming and Pig (i.e. 3 different solutions). In Hadoop streaming, this will require 2 different passes.

Part 3: Clustering

Create a new numeric file with 10,000 rows and 5 columns, separated by space – you can generate numeric data as you prefer, but submit the code that you have used.

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.

B. Using Hadoop streaming perform three iterations manually **using 9 centers** (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 with new centers can be given to the next pass.

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.