CSC 555: Mining Big Data

Project, Phase 1 (due Wednesday, October 26th)

In this part of the project, you will perform data warehousing queries using Hive, Pig and Hadoop streaming. The Hive schema is available at:
http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/SSBM1/SSBM_schema_hive.sql

It is based on SSBM schema benchmark which, in turn, is derived from industry standard TPCH benchmark. I have modified it to work in Hive (from SQL to HiveQL). This is Scale1, or the smallest unit of SSBM benchmark data – lineorder is the largest at about 0.6GB.
The tables are available at (note that the data is |-separated, not comma separated):
http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/SSBM1/dwdate.tbl
http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/SSBM1/lineorder.tbl
http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/SSBM1/part.tbl
http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/SSBM1/supplier.tbl
http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/SSBM1/customer.tbl
Please be sure to note what instance you are using (it should definitely be bigger than nano or micro, I recommend using medium – but at least small). Please be sure to submit all code (pig and python).

Part 1: Hive

Run the following eight (0.1, 0.2, 0.3, 1.1, 1.2, 1.3 and 2.1, 2.2) queries in Hive and record the time they take to execute:
http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/SSBM1/SSBM_queries.sql

Part 2: Pig

Convert and load the data into Pig, implementing queries 0.1, 0.2, 0.3, and 1.1. This should only require loading and processing 2 out of 5 tables (lineorder and dwdate). Check the disk storage, if your disk usage is over 90% Pig may hang without an error or a warning.

One easy way to time Pig is as follows: put your sequence of pig commands into a text file and then run, from command line in pig directory (e.g., [ec2-user@ip-172-31-6-39 pig-0.15.0]$), **bin/pig -f pig_script.pig** (which will inform you how long the pig script took to run).

Part 3: Hadoop Streaming

Implement query **0.3** using Hadoop streaming with python (you may implement this part in Java if you prefer).

You would need to create a mapper and reducer python code, which can be based on posted examples. In order to be able to run my examples, you need to copy the Hadoop streaming jar (**cp ./hadoop-2.6.4/share/hadoop/tools/lib/hadoop-streaming-2.6.4.jar .**) to current directory where you are running.

> Submit a single document containing your written answers. Be sure that this document contains your name and "CSC 555 Project Phase 1" at the top.