CS585: Big Data Management
Spring-2015

Project 2

Total Points: 120

Release Date: 02/10/2015

Due Date: 02/24/2014 (11:59 PM)

Teams: Project to be done in teams of two.

Short Description

In this project, you will write map-reduce jobs in Pig high-level language as well as in streaming mode and run them on Hadoop system.

Detailed Description

You will use the datasets that you have created in Project 1, namely the "*Customers*" and "*Transactions*" datasets. Based on these datasets, answer the following queries.

Hint: You should first go over Pig and Hadoop streaming examples in these links and understand them before working out this project:

https://cwiki.apache.org/PIG/pigmix.html
http://pig.apache.org/docs/r0.10.0/perf.html#join-optimizations
http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/

1) Query 1 [20 Points]

Write a Pig query that reports for every customer, the number of transactions that each customer did and the total sum of these transactions. The output file should have one line for each customer containing:

CustomerID, NumTransactions, TotalSum

2) Query 2 [20 Points]

Write a Pig query that joins the Customers and Transactions datasets (based on the customer ID) and reports for each customer the following info:

CustomerID, Name, Salary, NumOf Transactions, TotalSum, MinItems

Where *NumOfTransactions* is the total number of transactions done by the customer, *TotalSum* is the sum of field "TransTotal" for that customer, and *MinItems* is the minimum number of items in transactions done by the customer.

3) Query 3 [20 Points]

Write a Pig query that reports for every country code, the number of customers having this code as well as the min and max of *TransTotal* fields for the transactions done by those customers. The output file should have one line for each country code containing:

CountryCode, NumberOfCustomers, MinTransTotal, MaxTransTotal

4) Query 4 [20 Points]

Repeat Query 2 but take advantage of the fact that the *Customers* dataset is very small compared to the *Transactions* dataset. In this case, you can use Pig feature of ***Replicated Joins***. Compare the query plans generated by Pig for Q2 and Q5 and the impact on performance. In your final report, show the query plans and performance of each of Q2 and Q5 queries.

5) Query 5 [20 Points]

Use Hadoop streaming (any language of your choice) to join the customers who have country code = 5 with the transaction dataset and report one line for each of these customers containing:

CustomerID, CustomerName, CountTransactions

Where CountTransactions is the number of transactions done by this customer.

6) Query 6 [20 Points]

Repeat Query 1 but using Hadoop streaming. Use only C code to answer this query. The output file should include one line for each customer containing:

CustomerID, NumTransactions, TotalSum

Compare the performance between Q1 and Q7 and include your observations and comments in the submitted report.

What to Submit

You will submit a single zip file containing the Pig queries, and the streaming programs needed to answer the queries above. Also include a .doc or .pdf report file containing any required documentation.

How to Submit

Use blackboard system to submit your files.