CS561: Advanced Topics In Database Systems
Spring-2016

Assignment 2

Total Points: **120**

Release Date: **02/12/2016**

Due Date: **02/22/2016 (11:59PM)**

Teams: **Project to be done in teams of two**.

Short Description

In this project, you will write map-reduce jobs in high-level languages 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.

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]

Write a Pig query that reports the customer names who have the least number of transactions. Your output should be the customer names, and the number of transactions.

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]

Write an RHadoop script to do the following:

1) Create a map-reduce job that aggregates the records based on the CountryCode, i.e., For each country code, we need the count of customers.
2) Plot the output where country codes on the x-axis, and the count on the y-axis
3) Sort the output descending based on the count, and re-plot the chart.

What to Submit

You will submit a single zip file containing the Pig queries, the streaming programs, and the RHadoop code 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.