CS561: Advanced Topics In Database Systems
Spring-2012

Homework 3
Hadoop Project

Total Points: 90
Release Date: 02/16/2012
Due Date: 03/01/2012 (In class)

Description

In this homework, you will work on Hadoop system to upload data and execute MapReduce jobs on these data. The outcome from this homework is to learn how Hadoop works, how to write MapReduce jobs, and understand how these jobs run on Hadoop.

Hadoop Account and Setup

The instructions for setting up your account and testing your connection is found in document "AccessHadoopCluster" (available in blackboard and the website as well).

Note: It is your responsibility to test your account and make sure it is working as early as possible. If you faced any technical issues, let me know ASAP.

Project

You are asked to perform three activities in this project, (1) Create datasets, (2) upload the datasets into Hadoop, (3) Query the data by writing MapReduce Java code.

1-Createing Datasets [20 Points]

Write a java program that creates two datasets (two files), ***Customers*** and ***Transactions***. Each line in Customers file represents one customer, and each line in Transactions file represents one transaction. The attributed within each line are comma separated.

The ***Customers*** dataset should have the following attributes for each customer:

ID: unique sequential number (integer) from 1 to 20,000 (that is the file will have 20,000 line)
Name: random sequence of characters of length between 10 and 20 ***(do not include commas)***
Age: random number (integer) between 10 to 60
CountryCode: random number (integer) between 1 and 10
Salary: random number (float) between 100 and 10000

The ***Transactions*** dataset should have the following attributes for each transaction:

TransID: unique sequential number (integer) from 1 to 2,000,000 (the file have 2M transactions)
CustID: References one of the customer IDs, i.e., from 1 to 20,000 (on Avg. a customer has 100 trans.)
TransTotal: random number (float) between 10 and 1000
TransNumItems: random number (integer) between 1 and 10
TransDesc: random text of characters of length between 20 and 50 ***(do not include commas)***

Note: The column names will NOT be stored in the file. Only the values comma separated. Form the order of the columns, you will know each column represents what.

2-Uploading Data to Hadoop [10 Points]

Use hadoop file system commands (e.g., put) to upload the files you created to Hadoop cluster. Make sure you upload the data under your own HDFS directory "/user/<your-username>/"

Note: It is good to check your files at: http://cs-master.wpi.edu:50770/dfshealth.jsp
and see how the files are divided into blocks and each block is replicated.

3-Writing MapReduce Jobs [60 Points]

You will write Java program to query the data in Hadoop. Before writing your code you should perfectly understand the "WordCount" example in:

http://hadoop.apache.org/common/docs/r0.17.0/mapred_tutorial.html

You should have tried this code while testing your Hadoop account.

Notes:

- You should know whether each query is a Map-only job or Map-Reduce job, and write your code based on that.
- The output directory to which the query output will be written must be under your HDFS directory "/user/<user-name>/"
- Always delete the output directory before re-running your job. The cluster space is limited. So if you need to re-run the query multiple times (e.g., because you fix bugs), then first delete the old output directory. To delete files check Hadoop file system commands at:
  http://hadoop.apache.org/common/docs/r0.18.3/hdfs_shell.html
- You can always check the query output file from the HDFS website and see its content.

3.1) Query 1 [20 Points]

Write a job that reports the customers whose CountryCode between 2 and 6 (inclusive).

3.2) Query 2 [20 Points]

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

CustomerID, NumTransactions, TotalSum

3.2) Query 3 [20 Points]

Write a job 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 did by the customer, TotalSum is the sum of field "TransTotal" for that customer, MinItems is the minimum number of items in transactions did by the customer.

<u>Hint</u>: It is important two know how Hadoop reads and writes integers, floats, and text fields. Check IntWritable, FloatWritable, and Text classes to know which one to use and when.

What to Submit

You will submit a single zip file containing the Java programs for ***Creating Data Files*** and ***MapReduce Queries*****,** plus a document describing how to run these programs.

How to Submit

Use blackboard system to submit your files.