CSE 512 Assignment 4

Maximum Points Possible – 5

The required task is to write a map-reduce program that will perform equijoin.

- The code should be in Java (use **Java 1.8.x**) using Hadoop Framework (use **Hadoop 2.7.x**).
- The code would take two inputs, one would be the HDFS location of the file on which the equijoin should be performed and other would be the HDFS location of the file, where the output should be stored.

Format of the Input File: -
Table1Name, Table1JoinColumn, Table1Other Column1, Table1OtherColumn2, ……..
Table2Name, Table2JoinColumn, Table2Other Column1, Table2OtherColumn2, ……..

Format of the Output File: -
If Table1JoinColumn value is equal to Table2JoinColumn value, simply append both line side by side for Output. If Table1JoinColumn value does not match any value of Table2JoinColumn, simply remove them for the output file. You should not include two joins contains same row (No duplicate joins in output file).

Note: -
Table1JoinColumn and Table2JoinColumn would both be Integer or Real or Double or Float, basically Numeric.

Example Input : -
R, 2, Don, Larson, Newark, 555-3221
S, 1, 33000, 10000, part1
S, 2, 18000, 2000, part1
S, 2, 20000, 1800, part1
R, 3, Sal, Maglite, Nutley, 555-6905
S, 3, 24000, 5000, part1
S, 4, 22000, 7000, part1
R, 4, Bob, Turley, Passaic, 555-8908
Example Output: -
R, 2, Don, Larson, Newark, 555-3221, S, 2, 18000, 2000, part1
R, 2, Don, Larson, Newark, 555-3221, S, 2, 20000, 1800, part1
R, 3, Sal, Maglite, Nutley, 555-6905, S, 3, 24000, 5000, part1
S, 4, 22000, 7000, part1, R, 4, Bob, Turley, Passaic, 555-8908

Another correct answer is:
R, 2, Don, Larson, Newark, 555-3221, S, 2, 18000, 2000, part1
R, 2, Don, Larson, Newark, 555-3221, S, 2, 20000, 1800, part1
R, 3, Sal, Maglite, Nutley, 555-6905, S, 3, 24000, 5000, part1
R, 4, Bob, Turley, Passaic, 555-8908, S, 4, 22000, 7000, part1

So it means that whether R is before S is not required in the result. But you cannot have both
S, 4, 22000, 7000, part1, R, 4, Bob, Turley, Passaic, 555-8908
and
R, 4, Bob, Turley, Passaic, 555-8908, S, 4, 22000, 7000, part1
in the output.

You cannot assume that the table are R and S all the time. They can be other two tables. Number of tables in the input are exactly 2.

Submission Instructions: -
Put all your submission files in a folderUpload a zip file named Assignment4.zip (.rar is not allowed and will cause point loss), which will contain three files, equijoin.java, equijoin.jar and a ReadMe.txt. (The name has to be exactly the same as required and it is case sensitive) I will use equijoin.jar to run the assignment. ReadMe.txt should contain the approach you used for doing this work, basically how your mapper, reducer and driver is working in short.

This is how I am going to run your submission: -
sudo -u <username> <path_of_hadoop> jar <name_of_jar> <class_with_main_function> <HDFSinputFile> <HDFSoutputFile>

Example: -
sudo -u hduser /usr/local/hadoop/bin/hadoop jar equijoin.jar
equijoin hdfs://localhost:54310/input/sample.txt
hdfs://localhost:54310/output

Instructions for Assignment: -
Please follow these instructions closely **else Marks will be deducted.**

1. Please make sure you follow the submission instructions carefully and **do not miss any files**.
2. Please make sure to run the jar before submitting and make sure there is no compilation or runtime error.
3. Make sure your jar can be run from arbitrary location.
4. For any case of doubt in the assignment, PLEASE USE Discussion Boards, Individual mails would not be entertained.
5. Also, it is an individual's responsibilities to clarify his/her doubts, so read and use Discussion Board extensively.