CS595—Big Data Technologies

Assignment #5 (Modules 05)

Worth: 12 points (2 points for each problem) + 2 points extra credit

Due by the start of the next class period

Assignments should be uploaded via the Blackboard portal

It is ok to ask for hints from me to help solve the problems below. I will try to be helpful without giving away the answers.

Note: There may be short quiz questions about readings, assignments or articles (except extra credit) in the class period when they are due.

Read from (TW)

- Chapter 19

For this assignment you will be using your Hadoop environments

The general theme of this week's assignment is to write Pig commands and queries programs to perform various tasks

Recall that the files generated by TestDataGen have comma separated fields.

Exercise 1)

Create new versions of the foodratings and foodplaces files by using TestDataGen (as described in assignment #4) and copy them to HDFS

Write and execute a sequence of pig latin statements that loads the foodratings file as a relation. Call the relation 'food_ratings'. The load command should associate a schema with this relation where the first attribute is referred to as 'name' and is of type chararray, the next attributes are referred to as 'f1' through 'f4' and are of type int, and the last field is refereed to a 'placeid' and is also of type int.

Execute the describe command on this relation.

Provide the magic number, the load command you wrote and the output of the describe command as the result of this exercise.

Exercise 2)

Now create another relation with two fields of the initial (food_ratings) relation: ‘name’ and ‘f4’. Call this relation ‘food_ratings_subset’.

Store this last relation back to HDFS.

Also write 6 records of this relation out to the console.

Submit the pig latin statements you used and the six records printed out to the console as the result of this exercise.

Exercise 3)

Now create another relation using the initial (food_ratings) relation. Call this relation ‘food_ratings_profile’. The new relation should only have one record. This record should hold the minimum, maximum and average values for the attributes ‘f2’ and ‘f3’. (So this one record will have 6 fileds).

Write the record of this relation out to the console.

Submit the pig latin statements you used and the record printed out to the console as the result of this exercise.

Exercise 4)

Now create yet another relation from the initial (food_ratings) relation. This new relation should only include tuples (records) where f1 < 20 and f3 > 5. Call this relation ‘food_ratings_filtered’.

Write 6 records of this relation out to the console.

Submit the pig latin statements you used and the six records printed out to the console as the result of this exercise.

Exercise 5)

Using the initial (food_ratings) relation, write and execute a sequence of pig latin statements that creates another relation, call it ‘food_ratings_2percent’, holding a random selection of 2% of the records in the initial relation.

Write 10 of the records out to the console.

Submit the pig latin statements and the records printed out to the console.

Exercise 6)

Write and execute a sequence of pig latin statements that loads the foodplaces file as a relation. Call the relation ‘food_places’. The load command should associate a schema with this relation where the first attribute is referred to as ‘placeid’ and is of type int and the second attribute is referred to as ‘placename’ and is of type chararray.

Execute the describe command on this relation.

Now perform a join between the initial place_ratings relation and the food_places relation on the placeid attributes to create a new relation called ‘food_ratings_w_place_names’. This new relation should have all the attributes (columns) of both relations. The new relation will allow us to work with place ratings and place names together.

Write 6 records of this relation out to the console.

Submit the pig latin statements you used and the six records printed out to the console as the result of this exercise.

Extra Credit (2 points)

Read the article about Spark available on the blackboard in the ‘Articles’ section:

“Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing”

Provide a half-page summary and include some of your own comments.