CSP554—Big Data Technologies

Assignment #5 (Modules 05)

Worth: 15 points

The general theme of this week's assignment is to write Pig commands and query scripts to perform various tasks.

I have included the code to demo use of Pig as one of the files—pigdemo.zip—associated with this assignment. There is also a file of instructions on how to set up and use the demo code—pigdemoreadme.txt. There are useful bits to study and reuse, so have a look now.

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

Exercise 1) 2 points

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

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 referred to as '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) 2 points

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, food_ratings_subset, back to HDFS (perhaps as the file /user/hadoop/fr_subset)

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) 2 points

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) 2 points

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) 2 points

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) 2 points

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.

Exercise 7) (3 points) Identify the one correct answer for each the following questions. These questions are similar to the ones you might find on the mid-term covering Pig. Each is worth ½ point.

I. Which keyword is used to select a certain number of rows from a relation when forming a new relation?

Answer: ____

Choices:

A. LIMIT
B. DISTINCT
C. UNIQUE
D. SAMPLE

II. Which keyword returns only unique rows for a relation when forming a new relation?

Choices:

Answer: ____

A. SAMPLE
B. FILTER
C. DISTINCT
D. SPLIT

III. Assume you have an HDFS file with a large number of records similar to the examples below

- Mel, 1, 2, 3
- Jill, 3, 4, 5

Which of the following would NOT be a correct pig schema for such a file?

Choices:

Answer: ____

A. (f1: CHARARRY, f2: INT, f3: INT, f4: INT)
B. (f1: STRING, f2: INT, f3: INT, f4: INT)
C. (f1, f2, f3, f4)
D. (f1: BYTEARRAY, f2: INT, f3: BYTEARRAY, f4: INT)

IV. Which one of the following statements would create a relation (relB) with two columns from a relation (relA) with 4 columns? Assume the pig schema for relA is as follows:

(f1: INT, f2, f3, f4: FLOAT)

Answer: ____

Choices:

A. relB = GROUP relA GENERATE f1, f3;
B. relB = FOREACH relA GENERATE $0, f3;
C. relB = FOREACH relA GENERATE f1, f5;
D. relB = FOREACH relA SELECT f1, f3;

V. Pig Latin is a _____ language. Select the best choice to fill in the blank.

Choices:

A. functional
B. data flow
C. procedural
D. declarative

VI. Given a relation (relA) with 4 columns and pig schema as follows: (f1: INT, f2, f3, f4: FLOAT) which one statement will create a relation (relB) having records all of whose first field is less than 20

Answer: ____

Choices:

A. relB = FILTER relA by $0 < 20
B. relB = GROUP relA by f1 < 20
C. relB = FILTER relA by $1 < 20
D. relB = FOREACH relA GENERATE f1 < 20