CS595—Big Data Technologies

Assignment #7

Worth: 12 points (2 points for each problem)

Due

For this assignment you will be using your Hadoop environment including the pyspark CLI.

Some basic notes:

- We will again be using files generated by the program TestDataGen. But even though the files this program generates end is the '.txt' suffix, I want you to treat them as if they were '.csv' files. In fact, if you like, when you copy them to HDFS you can change their suffixes form '.txt' to '.csv'. But this is not necessary to complete the exercises.
- Also, don't forget that before starting pyspark enter the following into the command line of your maria_dev account to use Spark 2 capabilities:
    - export SPARK_MAJOR_VERSION=2

Exercise 1)

Step A

Use the TestDataGen program from previous assignments to generate new data files

Copy the files to HDFS.

Step B

Load the 'foodratings' file as a 'csv' file into a DataFrame called ex1_foodratings. When doing so specify a schema having fields of the following names and types:

Field Nampee	Field Type
name	String
food1	Integer
food1	Integer
food1	Integer
food1	Integer
placeid	Integer

As the results of this exercise provide the magic number, the code you execute and screen shots of the following commands:

foodratings.printSchema()

foodratings.head(5)

Exercise 2)

Load the ‘foodplaces’ file as a ‘csv’ file into a DataFrame called foodplaces. When doing so specify a schema having fields of the following names and types:

Field Nampee	Field Ty
placeid	integer
placename	string

As the results of this exercise provide the code you execute and screen shots of the following commands:

foodratings.printSchema()

foodratings.head(5)

Exercise 3)

Step A

Register the DataFrames created in exercise 1 and 2 as tables called “foodratingsT” and “foodplacesT”

Step B

Use a SQL query on the table “foodratingsT” to create a new DataFrame called foodratings_ex3 holding records which meet the following condition: food2 < 25 and food4 > 40

As the results of this step provide the code you execute and screen shots of the following commands:

foodratings.printSchema()

foodratings.head(5)

Step C

Use a SQL query on the table “foodplacesT” to create a new DataFrame called foodplaces_ex3 holding records which meet the following condition: placeid > 3

As the results of this step provide the code you execute and screen shots of the following commands:

foodratings.printSchema()

foodratings.head(5)

Exercise 4)

Use an operation (not a SQL query) on the DataFrame ‘foodratings’ create in exercise 1 to create a new DataFrame called foodratings_ex4 that includes only those records (rows) where the ‘name’ field is “Mel” and food3 < 25.

As the results of this step provide the code you execute and screen shots of the following commands:

foodratings.printSchema()

foodratings.head(5)

Exercise 5)

Use an operation (not a SQL query) on the DataFrame ‘foodratings’ create in exercise 1 to create a new DataFrame called foodratings_ex5 that includes only the columns (fields) ‘name’ and ‘placeid’

As the results of this step provide the code you execute and screen shots of the following commands:

foodratings.printSchema()

foodratings.head(5)

Exercise 6)

Use an operation on the DataFrame ‘to create a new DataFrame called ex6 which is the inner join, on placeid, of the DataFrames ‘foodratings; and ‘foodplaces’ created in exercises 1 and 2

As the results of this step provide the code you execute and screen shots of the following commands:

ex6.printSchema()

ex6.head(5)