CSP554—Big Data Technologies

Assignment #7

Worth: 12 points (2 points for each problem)

Readings:

The mid-term will assume you have read at least the high points of the below. In "Free Books and Chapters"

- Make sure to read/skim "Spark - The Definitive Guide (Excerpts)"
- Make sure to read/skim "Spark - Python API - SQL & DataFrames"
- Also take a look at the two Spark Cheat Sheets

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 (comma separated) '.csv' files.
- In fact, if you like, when you copy them to HDFS you can change their suffixes from '.txt' to '.csv'. But this is not necessary to complete the exercises.

Demos

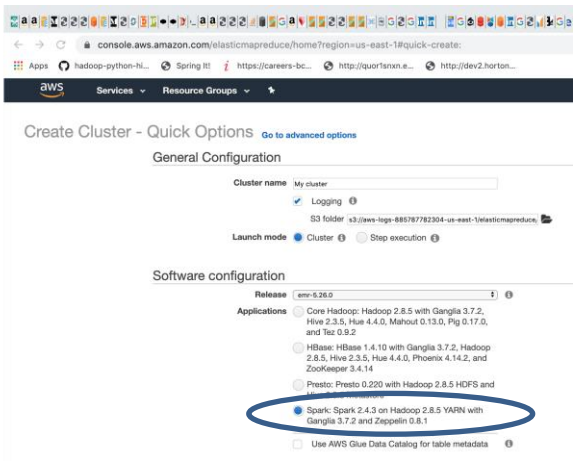
To sets of demo files have been included in this assignment. It is a good idea to look at and execute them before trying the following exercises.

1. One set of demo files provide examples of the use of RDDs. The instructions are in pyspark.txt and the demo files themselves are in pyspark.zip.
2. Another set of demo files provide examples of the use of DataFrames. The instructions are in dfdemo.txt and the demo files themselves are in sparkdf.zip.

Exercise 1)

Step A

Start up a Hadoop cluster as previously, but instead of choosing the "Core Hadoop" configuration chose the "Spark" configuration (see below), otherwise proceed as before.



Step B

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

Copy the files to the directory "/user/hadoop" in HDFS

Step C

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

Field Name	Field Type
name	String
food1	Integer
food2	Integer
food3	Integer
food4	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.show(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 Type
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.show(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_ex3a holding records which meet the following condition: food2 < 25 and food4 > 40. Remember, when defining conditions in your code use maximum parentheses.

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

foodratings_ex3a.printSchema()

foodratings_ex3a.show(5)

Step C

Use a SQL query on the table “foodplacesT” to create a new DataFrame called foodplaces_ex3b 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:

foodplaces_ex3b.printSchema()

foodplaces_ex3b.show(5)

Exercise 4)

Use a transformation (not a SparkSQL query) on the DataFrame ‘foodratings’ created 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_ex4.printSchema()

foodratings_ex4.show(5)

Exercise 5)

Use a transformation (not a SparkSQL query) on the DataFrame ‘foodratings’ created 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_ex5.printSchema()

foodratings_ex5.show(5)

Exercise 6)

Use a transformation (not a SparkSQL query) 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.show(5)