Big Data Application Development - Class 9 Homework

New York University

Spring 2019

Homework

A. Spark Project

1. Develop code using Spark to characterize (profile) the data in each column you plan to use from the data source you are responsible for. Use Spark or Spark SQL for data profiling. The result of this will be the data type for each column, and ideally could include information about the range of values for each column you'll use and the maximum string lengths you expect (where applicable).

   You may need to write code to clean and/or format (ETL-Extract, Transform, and Load) your data after profiling it (you should submit this too). For example, you might want to drop some columns, you might want to normalize data in a column (e.g. you may want to change all references to NYC in a 'City' column to 'New York City' instead of NYC/nyc/NYCity/NYC), you might want to detect badly formatted rows that might be missing important data.

   Submit this code in NYU Classes - this is an individual assignment - only upload your own code.

2. Submit the schema (column names and their data types) for the dataset you're responsible for. You can just type the schema information into a schema.txt file.

   Submit this in NYU Classes - this is an individual assignment - only upload results for your data source.

B. Spark Homework

3. Complete and upload solutions to assignments #1 and #2 described on the following slides.

C. Readings

1. Complete the readings assigned to date.
2. Read Chapter 5: 71-81 (Chapter 5: Loading and Saving Your Data)
3. Read Chapter 7: 117-134 (Chapter 7: Running on a Cluster)

Homework

Spark Assignment #1:

Join Web Log Data with Account Data (provide the commands in NYU Classes)

Store the accounts.zip data to directory loudacre/accounts in HDFS. Use the small web log file you already stored to HDFS: 2014-03-15.log. Review the accounts file: the first field in each line is the user ID, which corresponds to the user ID in the web server logs. The other fields include account details such as creation date, first and last name and so on.

a. Join the accounts data with the weblog data to produce a dataset keyed by user ID which contains the user account information and the number of website hits for that user. Here are the steps:

1. Use the accounts data to Create an RDD named **`userData`** consisting of key/value-array pairs: (userid,[values,…])

(*userid1*,[*userid1*,2008-11-24 10:04:08,\N,Cheryl,West,4905 Olive Street,San Francisco,CA,…])
...

2. Join the **`userData`** RDD with the set of user-id/hit-count pairs calculated in the previous homework assignment to generate:

(*userid1*,([*userid1*,2008-11-24 10:04:08,\N,Cheryl,West,4905 Olive Street,San Francisco,CA,…],**4**))
...

3. Display the user ID, hit count, first name (3rd value), and last name (4th value) for the first 5 users, e.g.:

userid1 4 Cheryl West
userid2 8 Elizabeth Kerns
userid3 1 Melissa Roman
...

Homework

Spark Assignment #2:

Use keyBy, mapValues, and sort (provide the commands in NYU Classes)

a. Challenge 1: Use keyBy to create an RDD of account data with the postal code (9th field in the accounts CSV file) as the key.
Tip: Save this RDD for use in the next challenge

b. Challenge 2: Create a pair RDD with postal code as the key and a list of names (Last Name,First Name) in that postal code as the value.
Hint: First name and last name are the 4th and 5th fields respectively
Try using the mapValues operation

c. Challenge 3: Sort the data by postal code, then for the first five postal codes, display the code and list the names in that postal zone, e.g.

```
--- 85003
Jenkins,Thad
Rick,Edward
Lindsay,Ivy
…

--- 85004
Morris,Eric
Reiser,Hazel
Gregg,Alicia
Preston,Elizabeth
…
```