Big Data Application Development - Class 7 Homework

New York University

Summer 2017

Homework

Class 7

A. Spark Project

1. Draw *initial* diagrams using PowerPoint, Visio, etc. to describe your project. Include a diagram of the software architecture, the data flows, and anything else you think is important to show. This is a first draft, you will refine it in the coming weeks. All team members should upload the diagrams.

2. Create a list of tasks for your project - you must use the TaskList.xlsx template in Resources. Assign team members to tasks, and assign a due date to each task. Try to identify milestones – that will help you know if you are on or off track. All team members should upload the schedule.

3. Develop Spark code using Scala or Python to ETL (clean/format) your data sources as needed (there should be at least one data source per team member). Submit this code.
   **Any problems with access to data sources need to be resolved quickly or you will fall behind.
   Submit this code in NYU Classes - this is an individual assignment - only upload your own code.

4. Start designing and developing your application.

B. Spark Homework

1. Complete and upload the Spark homework assigned last week.
2. Complete and upload solutions to assignments #1 and #2 described on the following slides.

C. Readings

1. Complete the readings assigned last week.
2. TDG pages: 117-134 (Chapter 7: Running on a Cluster)
3. TDG pages: 141-154 (Chapter 8: Tuning and Dubugging Spark)
4. TDG pages: 71-81 (Chapter 5: Loading and Saving Your Data)

Homework

3. 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

4. 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
…
```