

Class 6: Homework

Big Data Application Development **Spring 2018**



Homework

Class 6

Analytics Project

1. Study for Midterm Exam.

Study for the midterm exam which will be given during regular class time.

The midterm exam covers chapter 1, 2, and 3 of the Learning Spark text.

It is recommended that you study the slides, your notes, book readings, and the homework to prepare for the exam.

Content from any paper readings that were assigned will not be on the exam.

Autonomic Computing slides will not be on the exam.

The content from Class 6 will not be on this exam.

2. Read Chapter 4 in Learning Spark.

3. Submit the first draft of your Team Project Proposal (TPP) - use the template provided.

This is your first draft, you will have an opportunity to refine this proposal, or even change it completely in the coming weeks. **All team members should upload the TPP.**

Honor Pledge: *Many of you are working with project partners with whom you must share your code and findings, and that is fine and expected. With this honor pledge you assure that no project deliverable, or part of a project deliverable, has been included in your respective projects without proper attribution to the original author/programmer.*

Homework

Class 6

3. Spark Homework

Provide a program that does the following:

Use Pair RDDs to Join Two Datasets ([Provide in NYU Classes Assignment the code you used.](#))

You will use the web server log file `2014-03-15.log` and the user account data in key-value Pair RDDs.

1. Start two terminal windows. In one window, start the Scala Spark Shell: `$ spark-shell` Use the other window for command line operations.
2. Copy the `accounts.zip` file to the VM, unzip it, and store it in HDFS to: `loudacre/accounts`

Example line:

```
178,2008-12-09 12:09:14.0,\N,Kimberly,Mulder,2383 Patton Lane,San Francisco,CA,
94114,4150916606,2014-03-18 13:29:47.0,2014-03-18 13:29:47.0
```

Schema:

```
AccountID (UserID), Date1, Date2, FirstName, LastName, Street, City, State,
Zip, Phone, Date3, Date4
```

You'll also need to use the weblog data from an earlier exercise. The weblog directory might already be in the `loudacre` directory.

(Note: Complete this assignment using the weblog dir that has just one log file - `2014-03-15.log`)

Homework

Class 6

3. Spark Homework (continued)

3. Count the number of requests from each user.

- a. You'll need to use the user ID field - it is the **third field** in each line of the weblogs data.

Hint: Create a Pair RDD and use the WordCount approach covered in the RBDA course.

Your data will look something like this:

```
(useridA, 1)
(useridB, 1)
(useridA, 1)
```

- b. Sum the values for each user ID.

Hint: Your RDD data will look something like this:

```
(useridA, 5)
(useridB, 7)
(useridC, 5)
```

- c. Determine how many users appear (visited) once, twice, three times, and so on.

Generate data in this format: frequency:user-count pairs

Hint: The data shown in b. above produces the following -

```
(5:2)
(7:1)
```

- d. Create an RDD where the user id is the key, and the value is the list of all the IP addresses that the user has connected from.

You will need the accounts data in order to only output ip addresses for user IDs that appear in the accounts files.

Ensure that the output only contains user IDs of actual customers.

Input example: (useridX, 20.1.34.55)

(useridX, 74.125.239.98)

becomes: (useridX, [20.1.34.55, 74.125.239.98])