Class 6: Homework

Big Data Application Development

Summer 2017

Homework

Class 6

Analytics Project

1. Submit the first draft of your Team Project Proposal (TPP) - use the template provided. All team members should upload the TPP.
2. Research your project. Each team member should upload a summary for one distinct paper. Try using GoogleScholar to find papers. Please do not choose papers on a Hadoop technology or other tool. Instead, choose papers related to your project thesis - the paper does not have to be in the same domain as your project. The paper you select might describe a technique or processing that might be useful in your project.

   Each team member should write a summary of a scholarly paper related to the team analytics project.

   Coordinate with your teammates to ensure each member reads a different paper. Share what you learn with your teammates. Please write a short, one paragraph summary for the paper and upload to NYU Classes. (In a future homework, you will add these summaries to the 'Related Work' section of your project paper.)

   Note: The MapReduce, HDFS, and other papers already assigned cannot be used for this assignment.

   Some places to look for papers:

   ACM KDD Conference: http://www.kdd.org/kdd2016/

   Google Scholar: https://scholar.google.com/

   ACM DL: http://dl.acm.org/

3. Bring your data into your VM and/or into the NYU Hadoop cluster, Dumbo.

   Start collecting your data - use the NYU Hadoop cluster if you will have lots of data (Big Data). Post a note on the forum if you have trouble finding data or loading it.

4. Upload to NYU Classes a text file containing the schema (field names and data types) for your dataset.

   It may be a subset of the dataset you downloaded. Each team member should upload a schema for their dataset.

Readings

1. Please read in the class text, "Learning Spark," pages: 36-46, 47-51 (Aggregations part that talks about reduceByKey too), 57 - all of page 60). Skip over the Python and Java sections.

Homework

Class 6

Spark Homework

1. 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 and the user account data in key-value Pair RDDs.

A. Start two terminal windows. In one window, start the Scala Spark Shell: $ spark-shell Use the other window for command line operations.

B. Copy the accounts.zip file to the VM, unzip it, and store it in HDFS to: loudacre/accounts

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`)

C. Using map-reduce, count the number of requests from each user.

1. Use map to create a Pair RDD with the user ID as the key, and the integer 1 as the value. (The user ID is the third field in each line.)

Your data will look something like this:

(useridA, 1)

(useridB, 1)

(useridA, 1)

...

2. Use reduce to sum the values for each user ID. Your RDD data will be similar to:

(useridA, 5)

(useridB, 7)

(useridC, 2)

...

3. Use countByKey to determine how many users visited once, twice, three times, and so on.

Use map to reverse the key and value. Use the countByKey to return a Map of frequency:user-count pairs.

4. 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.

(IP address is the first field in each request line.) Hint: Map to (userid, ipaddress) and then groupByKey.

(userid, 20.1.34.55)

...

becomes:

(userid, [20.1.34.55, 74.125.239.98])

…