CS 8001 Big Data

HW #3: Web Server Log Analysis (10 points + 2 bonus points)

Spring 2016

(Due 2/10, Wednesday, midnight)

In this exercise, you will write Python code on Spark that analyzes a data set from NASA Kennedy Space Center WWW server in Florida, which is freely available at http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html. Use the first log of the data set in this homework, which was collected from 00:00:00 July 1, 1995 through 23:59:59 July 31, 1995, a total of 31 days. This exercise consists of 4 steps:

1. Apache Web Server Log file format
2. Sample Analyses on the Web Server Log File
3. Analyzing Web Server Log File
4. Exploring 404 Response Codes

See reference here.

Part I (10 points): Run Spark locally with one or more worker threads.

Submit electronically in Blackboard the following:

1) A brief description of your Spark environment, your code structure, and the execution process of the code.
2) Output of your code for the following analysis:
    a. Top Ten Error Endpoints. Create a sorted list containing top ten endpoints and the number of times that they were accessed with non-200 return code.
    b. Number of Unique Hosts.
    c. Number of Unique Daily Hosts
    d. Counting 404 Response Codes. How many 404 records are in the log?
    e. Listing the Top Twenty 404 Response Code Endpoints
3) Execution time of your code in seconds.
4) Your Python code with appropriate comments.

Part II (2 bonus points): Run Spark on a cluster in Amazon AWS.

Submit electronically in Blackboard the following:

1) A brief description of your cluster environment and the execution process of the code.
2) Execution time of your code in seconds.
3) Comparison of the execution results with those in Part I