Homework 2: (Deadline: 09/25)

1. Introduction

The data set we are using is an anonymized web server log from a public relations company whose clients were DVD distributors. The log file "access_log" can be downloaded from gitlab (https://gitlab.encs.vancouver.wsu.edu/xuechen.zhang/cs453-hw2-dataset.git), and it's currently compressed using Zip. So you'll need to decompress it and then put it in HDFS. If you take a look at the file, you'll see that each line represents a hit to the Web server. It includes the IP address which accessed the site, the date and time of the access, and the name of the page which was visited.

The logfile is in Common Log Format: %h - - %l %u "%r" %>s %b
(e.g., 10.223.157.186 – [15/Jul/2009:15:50:35 -0700] "GET /assets/js/lowpro.js HTTP/1.1" 200 10469)
where:

- %h is the IP address of the client
- %l is identity of the client, or "-" if it's unavailable
- %u is username of the client, or "-" if it's unavailable
- %t is the time that the server finished processing the request.
  The format is [day/month/year:hour:minute:second zone]
- %r is the request line from the client is given (in double quotes). It contains the method, path, query-string, and protocol or the request.
- %>s is the status code that the server sends back to the client. You will see mostly status codes 200 (OK - The request has succeeded), 304 (Not Modified) and 404 (Not Found). See more information on status codes in W3C.org
- %b is the size of the object returned to the client, in bytes. It will be "-" in case of status code 304.

2. Problems to solve

For each of the following problems, we would like you to write a MapReduce job to solve the problem and when you have done that you should be able to answer the question we are going to ask you.

- P1: Write a MapReduce program which will display the number of hits for each different file on the Web site and answer the question: *How many hits were made to the page: /assets/js/the-associates.js?*
- P2: Write a MapReduce program which determines the number of hits to the site made by each different IP Address and answer the question: *How many hits were made by the IP address: 10.99.99.186?*
- P3: Find the most popular file on the Web site. In other words, the file which had the most hits. Your Reducer should just write out the name of the file and number of hits into HDFS.

What You Need to Submit:
Submit your PDF file to Blackboard http://learn.wsu.edu before 11:59pm on the due date.