**Secondary Sort**

For this problem you are to write a MapReduce program that performs a secondary sorting on a log file which stores the access history to a webpage. Every line of the input dataset contains the IP\_address that accessed the web page, the time stamp when the web page was accessed and the http request that was submitted by the IP address. Here is an example line from the input file:

10.216.113.172 - - [16/Jul/2009:02:51:29 -0700] "GET /assets/js/lowpro.js HTTP/1.1" 200 10469

The http request line consists of a Method token (such as : GET , POST, PUT, etc.), a request URI (such as: /assets/js/lowpro.js ) and a HTTP code (such as: HTTP/1.1)

Your MapReduce program must read this input file and produce as output **the latest access time for each request URI in the input file**. For example, if the input file contains the following records:

10.125.118.102 - - [08/Sep/2010:10:20:15 -0700] "GET /assets/img/banner/ten-years-banner-white.jpg HTTP/1.1" 200 18077

10.125.118.102 - - [08/Nov/2010:11:25:15 -0700] "GET /assets/img/banner/ten-years-banner-white.jpg HTTP/1.1"200 17903

10.125.118.102 - - [08/Apr/2010:10:20:17 -0700] "GET /images/filmpics/0000/1143/197-the-greatest-stills-1-2362x1575\_thumb.jpg HTTP/1.1" 200 35392

10.125.118.102 - - [08/Apr/2010:10:20:17 -0700] "GET /images/filmpics/0000/1143/197-the-greatest-stills-1-2362x1575\_thumb.jpg HTTP/1.1" 200 30065

10.125.118.102 - - [08/Apr/2010:10:20:17 -0700] "GET /images/filmpics/0000/1143/197-the-greatest-stills-1-2362x1575\_thumb.jpg HTTP/1.1" 200 27939

10.125.118.102 - - [08/Apr/2010:10:20:17 -0700] "GET /images/filmpics/0000/1143/197-the-greatest-stills-1-2362x1575\_thumb.jpg HTTP/1.1" 200 76418

10.125.118.102 - - [08/Apr/2010:10:20:18 -0700] "GET /images/filmpics/0000/1143/197-the-greatest-stills-1-2362x1575\_thumb.jpg HTTP/1.1" 200 35695

10.125.118.102 - - [08/Apr/2010:10:20:18 -0700] "GET /images/filmpics/0000/1143/197-the- greatest-stills-8-2362x1575\_thumb.jpg HTTP/1.1" 200 42179

10.125.118.102 - - [08/Apr/2010:10:20:19 -0700] "GET /favicon.ico HTTP/1.1" 200 1406

Then the output file should contain the following records:

/assets/img/banner/ten-years-banner-white.jpg [08/Nov/2010:11:25:15 -0700]

/images/filmpics/0000/1157/197-the-greatest-stills-8-2362x1575\_thumb.jpg [08/Apr/2010:10:20:18 -0700] /favicon.ico [08/Apr/2010:10:20:19 -0700]

The input file is relatively small. It is only 0.5 GB. You can download it from the following URL:

https://s3.amazonaws.com/class-data-set/access\_log

I have also prepared a smaller sample of the input file which you can download from here:

https://s3.amazonaws.com/class-data-set/access\_log\_test .