Names: Bob Skowron, Jason Walker
Keys: rskowron, jwalker
SVN: jwalker: https://svn.seas.wustl.edu/repositories/jwalker/cse427s_fl17/

1. a. The data format is a flat text file where each record represents a user making GET requests to some server. The information stored in each record is:

   * IP Address
   * Timestamp (date and time)
   * URL accessed
   * HTTP response code
   * Elapsed time of the request

   The test log has 5000 rows and the full data set has 4,477,843, so our sample is approximately .11% of the total.

   b. Mapper Input:
   (byteoffset, 10.223.157.186 - - [15/Jul/2009:21:24:17 -0700] "GET /assets/img/media.jpg HTTP/1.1" 200 110997)
   (byteoffset, 10.223.157.186 - - [15/Jul/2009:21:24:18 -0700] "GET /assets/img/pdf-icon.gif HTTP/1.1" 200 228)
   (byteoffset, 10.216.113.172 - - [16/Jul/2009:02:51:28 -0700] "GET / HTTP/1.1" 200 7616)
   (byteoffset, 10.216.113.172 - - [16/Jul/2009:02:51:29 -0700] "GET /assets/js/lowpro.js HTTP/1 .1" 200 10469)
   (byteoffset, 10.216.113.172 - - [16/Jul/2009:02:51:29 -0700] "GET /assets/css/reset.css HTTP/1.1" 200 1014)

   Mapper Output:
   (10.223.157.186, 1)
   (10.223.157.186, 1)
   (10.216.113.17, 1)
   (10.216.113.17, 1)
   (10.216.113.17, 1)

   Reducer Input:
   (10.223.157.186, [1, 1])
   (10.216.113.17, [1, 1, 1])

   Reducer Output:
   (10.223.157.186, 2)
   (10.216.113.17, 3)

   The reducer is summing over an array of ones, same as we saw with the word count examples.

   c. N/A

   d. i. Running the job locally, your inputs and outputs are local files and folders, respectively. When running on the cluster, the inputs and outputs are stored in hdfs. Print outputs, when run locally, are written out to the console for the end user to review. When run on the cluster, those outputs are not written out to the user. Running locally, there is no job management. The current job is what is run. When run on the cluster, hadoop (namely YARN) has to manage the resources on the cluster.

   ii. *hadoop jar logfileanalysis.jar stubs.ProcessLogs -fs=file:/// -jt=local /workspace/log_file_analysis/src/test_log_file output_test*

   iii. Used toolrunner. Indifferent between the two

   e. Based on the output, there are 10 unique IP addresses in the testlog file. Yes, every line contributed to a count. The sum of the counts in the output file sum to 5000, which is the original number of line items

   f. When running the job on an actual cluster we need to keep in mind....

   * Total number of IP Addresses: 333,923
   * 10.1.100.199: 35
   * 10.1.100.5: 1
   * 10.99.99.58: 21

   The results are globally sorted because the text sort done in the shuffle and sort step... blah blah blah

2.