Lab: Using Counters and a Map-Only Job

In this lab you will create a Map-only MapReduce job.

Your application will process a web server's access log to count the number of times gifs, jpegs, and other resources have been retrieved. Your job will report three figures: number of gif requests, number of jpeg requests, and number of other requests.

Instructions

1. Input Data

   Unpack the access_log.gz file and upload it to HDFS. You can unpack the file with the following command:

   ```
   gunzip access_log.gz
   ```

2. Source code

   You will find the source files in the `src/stubs` directory. You will also find a `src/hints` directory that contains a copy of the source files modified to make the assignment easier to complete, and a `src/solution` directory that contains a copy of the source files modified to complete the assignment.

3. Complete the `ImageCounter` and `ImageCounterMapper` classes in the `src/stubs` directory to count the number of occurrences of files of type GIF, JPEG, and "other". Use counters to return the counts from the map task.

4. Assemble and test your solution.

Hints

1. You should use a Map-only MapReduce job, by setting the number of Reducers to 0 in the driver code.

2. You should test your code against the smaller version of the access log before testing against the full log file.

3. Use a counter group such as `ImageCounter`, with names `gif`, `jpeg`, and `other`.

4. In your driver code, retrieve the values of the counters after the job has completed and report them using `System.out.println`.

5. The output folder on HDFS will contain Mapper output files which are empty, because the Mappers did not write any data.