

Handed out: 02/21/2014

Due by 17:25PM on Friday, 02/28/2014

Problem 1. Install Ruby 1.8.7 and the elastic map reduce command interface tools. Demonstrate that you can run word count example we mentioned in class from the command line of your PC or Mac. The command looks like:

```
ruby elastic-mapreduce --create -stream \  
  --mapper s3://elasticmapreduce/samples/wordcount/wordSplitter.py \  
  --input s3://elasticmapreduce/samples/wordcount/input \  
  --output s3n://yourbucket \   # A path to your bucket on Amazon S3  
  --reducer aggregate
```

The output will look similar to:
Created job flow JobFlowID.

Find a way to tell EMR not to terminate the cluster once the job is done. We are doing this to make sure you have a running cluster so that you could examine HDFS. Download results to your PC or Mac and verify the content. Show a few lines to us. Use the facilities of Hadoop distributed file system shell to fetch the second input file 00002. We always look at 00001 and are quite curious what is in 00002. In your report, please display the first 20 and the last 20 lines of that file. Generate those lines using Hadoop dfs commands.

Problem 2. While you have the above cluster running examine the content of the public bucket s3://elasticmapreduce/samples. Try to figure out what is the volume of files in that bucket. Transfer those files to the HDFS directory on your cluster. Subsequently transfer those files to your local machine, PC or Mac. Please shut down your cluster.

Problem 3. Attached are two ruby scripts: max_temperature_map.rb and max_temperature_reduce.rb. Also are attached two sample file containing recordings of meteorological data from years 1901 and 1902. Your scripts will extract the year and the temperature in Celsius from the every line of those files and then determine what was the highest temperature for each year. Both values are buried in the lines. Years are spelled out as 1901, and 1902. Temperatures are presented as 100 time the actual temperature in Celsius. So, -6.11°C is written as N9-00611+.. Year is extracted as 4 digits starting from position 15 and the temperature as five digits starting at position 87. We are actually not that much interested into those data. We are just familiarizing ourselves with the Elastic MapReduce environment. Upload both scripts to a folder in one of your S3 buckets. Upload two data files to perhaps another bucket. Direct the output to a third bucket and logs to yet another. Run an Elastic Map Reduce job using the above mapper and reducer. As the type of the Application Step select Streaming program. Retrieve the results and the logs and submit. Capture the interaction with AWS console.

Please shut down your cluster.

SUBMISSION INSTRUCTIONS:

Your main submission should be an MS Word document containing your code, results produced by that code and brief textual descriptions of what you did and why. Typically, you just copy your code and results from the R console and paste them into the Word document. Start with this text of homework assignment as the template. Please add your code (R, Java, Ruby, Python) into a single. It is more convenient for us to open one or two files than a large number of files. If we recognize from your Word document what your code is doing and the results it is producing we will not run your code. If we have doubts we will run your code. In order to be able to do that it is convenient if your code is in a txt document. In special cases we might request more convenient formats.

Package your submission into an archive called E63_LastNameFirstNameHW04.zip. Naming your file properly is important. We download many files and if they are all named Assignment01.zip it becomes hard not to overwrite and lose them. Please do not use archiving tools like RAR or TAR which do not produce ZIP files. If you are using a Mac, please make sure that your files are READABLE to users of Windows. You are welcome to save your work as a PDF file, but please, always submit a Word document, as well. You can use Open source imitations of Microsoft Office as well.

Upload your ZIP archive to the course web site. Every assignment has its drop box. If you miss the deadline, please submit your solution into the 00_AnyHW_WayLate Drop Box. Those assignments will be graded as well. **We will chop 10% of your grade for every day you are late.** Your grade for every assignment will be entered as a comment next to your submission.

If you have issues with the formulation of the assignment or the software you are using, please FIRST go to the Discussion Forum on the class web site: <http://isites.harvard.edu/icb/icb.do?keyword=k102025> and check whether someone else raised the same issue and whether the answer is already there. If not, raise the issues yourself. A person from the class or a member of the teaching stuff will respond. The discussion forum is a very important tool. We all learn from the discussions on the forum.

If the issue is not address for a while, please send an inquiry to cscie63@fas.harvard.edu. A member of the teaching stuff will respond.

If you have issues with the class web site, please let us know right away. In the past, we experienced issues with the visibility of various folders, upload permissions and so on. We will try to resolve such issues as soon as we hear about them. For some issues we depend on the university support services and delays are possible.