HU Extension **Assignment 03** E-185 Big Data Analytics

Handed out: 02/23/2013 Due by 17:25PM on Friday, 03/01/2013

Those of you who work with Windows operating system should download Cygwin. When you do that please make sure that you include OpenSSH tools. Cygwin download presents you with a list of Unix (Linux) packages. Under Net, look for openssh and then toggle the selection to install. OpenSSH will provide you with both SSH and SCP (secure copy) tools. Cygwin will provide you with Linux environment on your PC. Mac and Linux users have all those tools already installed.

Please register for Amazon AWS. You TA-s have deposited a TOKEN worth $200 into the comment for your Assignment 02. Go to My Account -> Payment Method. At the bottom of the screen you will see “Redeem an AWS Credit Coupon”. Do it.
Be wise with your time at Amazon AWS. If you are careful and never leave instances and buckets to hang out needlessly, those $200 will carry you through the course.

Problem 1. Create a bucket. Enable versioning. Upload a single text document. A few minutes later modify the document a tiny bit. Upload modified document. Demonstrate that you can retrieve both versions of the document.

Problem 2. Create a folder for your dog’s Web site. This is a very simple Web site and has only two HTML pages but you can navigate from one to the other. Demonstrate that you can serve your Web site from an AWS S3 bucket.

Problem 3. Install Ruby 1.8.7 and the elastic map reduce command interface tools. Demonstrate that you can run word count example we did 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 own on Amazon S3
--reducer aggregate
```

The output will look similar to:

```
Created job flow JobFlowID.
```

Download results to your PC or Mac and verify the content. Show a few lines to us. Use facilities of hadoop distributed file system shell to fetch for use 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.

Problem 4. Attached are two ruby scripts: max_temperature_map.rb and max_temperature_reduce.rb. Also are attached two sample file containing recordings of metheorogical 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^{o}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 flow as your own application. As the Job Type select Streaming. Retrieve the results and the logs and submit. Capture the interaction with AWS console.

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 past 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 E185_LastNameFirstNameHW02.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=k93720 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 cscie185@fas.harvard.edu. A member of the teaching stuff will respond.

If we respond to your inquiry to class email address or to the inquiry to any email address of the teaching stuff, PLEASE DO NOT RESPOND WITH A “THANK YOU NOTE”. This is not a joke. We will take 2% of your grade for that week’s assignment for every “thank you note”. We know that you are both polite and thankful for the effort of the volunteers who make up the teaching stuff. We all have limited time. Please let us use it for something else rather than opening and closing emails.

We will apply the same penalty to any trivial email. Please do not complain when you lose a few points on your assignment.

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.