HU Extension Assignment 07 E185 Big Data Analytics

Handed out: 03/30/2013 Due by 5:30PM on Saturday, 04/05/2013

The National Bureau of Economic Research (www.nber.org) offers an interesting set of econometric and sociological datasets. From the page http://data.nber.org/patents/, please download file `acite75_99.zip` which contains ASCII, comma separated values describing citation data for US patents between years 1975 and 1999. The first column contains ordered list of CITING patents and the second column contains CITED patents. For the initial analysis you might want to extract a small section of cite75_99.txt file you will obtain when you decompress the archive.
My recommendation is that you open an interactive Pig session on Amazon's Elastic Map Reduce and work on the Master node. Record Pig commands as you develop them so that you could interrupt your AWS sessions. When you shut a grunt session all of your data structures and commands are lost. Keep the text file cite75_99.txt you extract from the above archive in one of your S3 buckets.

Problem 1) Using Pig create a file that contains an ordered list of patents where every patent is followed by the patents that cite it. The list that we want to create looks like the following:

```
1000026   4043055
1000033   4190903, 4975983
1000043   4091523
1000044   4082383, 4055371
1000045   4290571
1000046   5918892, 5525001
```

The above table tells us, for example, that patent `1000033` is cited by patents: `4190903` and `4975983`. In a way you are inverting the original data. You can organize your output in any data structure you like. A nested data structure would look nice, though. Could you write one or more SQL commands that would accomplish the same objective.

As you are working, please describe for us all the intermediary relations (data structures) and transformations you are using.

Problem 2) Once you are satisfied with the output of your Pig process, place all of your code into a Pig script and demonstrate that you can run that script from the command prompt where the source file or source bucket and the output file or the output bucket are specified as input parameters to the script.

Problem 3) Create another Pig script that transform results of the previous program and produce a data set which will tell us how many times a patent has been cited. For example, the small section of cited patens list displayed above in problem1 will turn into something like:

```
1000026   1
```

```
1000033   2
1000043   1
1000044   2
1000045   1
1000046   2
```

This new list is telling us that patent `1000026` is cited once, patent `1000033` is cited twice, and so on. Store results of this program in your file system. Please record and present in your solution all the transformations that you are using and the structure of all relations (data structures) used along the way. Take a small (20 rows) section of the result and display in your solution.

Problem 4) Create a Pig script that will read the result of the previous problem and create a "histogram" of number of citations. We want to know how many patents are not cited at all, how many patents are cited once, cited twice and so on. Create one ordered list where the number of patents with the largest number of citations is on the top and another where the number of patents with 0 or 1 (the smallest number of) citations is on the top of the list. Store both results on your file system.
Could you write one or more SQL commands that would accomplish the same objectives. For us, make a small extract from the top of both files.

As usual, please capture all the steps of your implementation, with comments indicating what is it you are accomplishing with every step, in an MS Word document.
Please place all files you want to submit in a folder named: HW07. Compress that folder into an archive named E185_LastNameFirstNameHW07. ZIP. Upload the archive to the course drop box on the class web site. Please send comments and questions to cscie185@fas.harvard.edu