2nd Programing Assignment
CS562 – Advanced Database Applications

Due: April 30, 2020 at 11:59PM using Gradescope

In this assignment you will use Hadoop MapReduce and Pig to implement a simple graph algorithm. For the assignment, you need to install Hadoop and Pig on you computer and then run some of the examples that come together with the distribution. After getting familiar with the systems, you have to implement an algorithm to compute the number of triangles per node in a graph that will be provided to you.

1) Install Hadoop on your computer (Single Node Installation): You will find the general directions to download and install Hadoop here:
https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html

Here are the directions for installing Hadoop on Windows:
http://wiki.apache.org/hadoop/Hadoop2OnWindows

Now, you can run some examples from the Hadoop distribution. For example, the following operations copy the unpacked conf directory to use as input and then find and displays every match of the given regular expression. Output is written to the given output directory.

```
 $ mkdir input
$ cp conf/*.xml input
$ bin/hadoop jar hadoop-examples-*.jar grep input output 'dfs[a-z.]+'
$ cat output/*
```

2) Install Apache Pig using the instructions from here:

https://pig.apache.org/docs/r0.16.0/start.html

Run some examples on Local Mode using:

```
$ pig –x local
```

Now you can do the rest of the assignment.

3) Text processing: Run the WordCount example on some text documents. Try to understand how it works.

4) Graph Algorithm: Implement in Map Reduce the naïve and the improved Triangle Counting algorithms that we discussed in class to find the number of triangles incident in each node in the graph. The improved algorithm is the Algorithm 3 in the triangles paper.

Run the algorithms on the file *graph.txt* that is provided to you in the link below. Also, count the total number of triangles in the graph. Report the total number of triangles in the graph.

https://www.cs.bu.edu/~gkollios/cs562s20/datasets/Graph/

Notice that the graphs provided (a small graph for testing and a larger graph for running the experiments) are undirected. That means that if there is an edge between i and j, there should be an edge between j and i. Every line represents an edge between two vertices.

5) Implement the triangle counting algorithm in Pig. Here you need to use joins to generate triplets that have three nodes that have edges with each other. Report again the total number of triangles in the graph.

6) Write a report on how you implemented the algorithms and the results.

Code Submission.
Put all your code and the report in a .gzip or tar.gz file and submit it through Gradescope.