ASSIGNMENT 5

Submitted By-

Yugaank Arun Sharma

CS 549

Date of Submission: 10/12/2017

Implement the PageRank algorithm to find the most popular pages in a set of Web pages. You will use your implementation of PageRank to find the 'most popular' pages in a dump of Wikipedia pages that you are provided with.

Input format: We will assume that the link graph is initially available as a list of vertices and their adjacency lists (a list of "friend" vertices for each vertex). Each entry is a line in a file of the form:

node-id: to-node1 to-node2

To complete this assignment-:

I had to fill in the missing code for different mappers and reducers classes like itermap etc.

To test the code, I first installed Hadoop on cent os on my local virtual box. I tested the application on my local Hadoop but the testing video on localhost couldn't be made because of the windows update that screwed my virtual box and the cent operating system.

To demonstrate the Hadoop map reduce, I tested the PageRank jar file on Amazon Elastic Map Reduce or Amazon EMR. I first tested the simple graph given to us in the specification and then tested the Wikipedia data. Although the Wikipedia testing takes a lot of time, I only showed the output of the simple graph.

I ran the composite task on amazon EMR after thoroughly testing the other commands. The composite function runs the init task, then it alternates between running the iter and diff tasks until convergence has occurred (diff <=30 in the case of the test data), at which point it runs the finish task and places the output into.

Since EMR doesn't support the Hadoop cache, a more scalable way to do this is to perform an equipartitioned join of the output page ranks and the page names, where the page name table includes node identifiers as keys.

The Wikipedia data showed a different result when tested against 5 reducers compared to 20 reducers. Since the run takes a lot of time I didn't bother to record the whole thing.

In the zip archive, there is another archive of source code, two testing videos and a output file of the simple graph.

This was the first time I used Hadoop, so the testing didn't go perfectly and had lots of issues especially when done locally. The local testing demonstration couldn't be done cause of latest windows update screwing virtual box and my cent operating system on it. Had no time to set it up again.