Big Data 18CS322

Assignment 2

Page Rank Algorithm implementation with Map Reduce

1. Assignment Objectives and Outcomes

a. The objective of this assignment is for the students to run iterative processing with Map Reduce and learn how the Map Reduce algorithm works..

b. At the end of this assignment, the student will be able to write and debug Page Rank code on Map Reduce.

2. Ethical practices

Please submit original code only. You can discuss your approach with your friends but you must write original code. All solutions must be submitted through the portal. We will perform a plagiarism check on the code.

3. The Dataset:

a. For this assignment, we shall be using a web graph from Google released in 2002 of approximately 900,000 nodes.http://snap.stanford.edu/data/web-Google.html

b. Use this dataset for the tasks given below.

4. Software/Languages to be used:

a. Python (Version **3.6**) (Use Hadoop Streaming) or Java (Version **1.8** Only)

b. Please make sure to use Hadoop Version **3.2.0** only.

5. Marks:

a. 20 (Scaled down to 5)

b. Each Task is for 10 marks

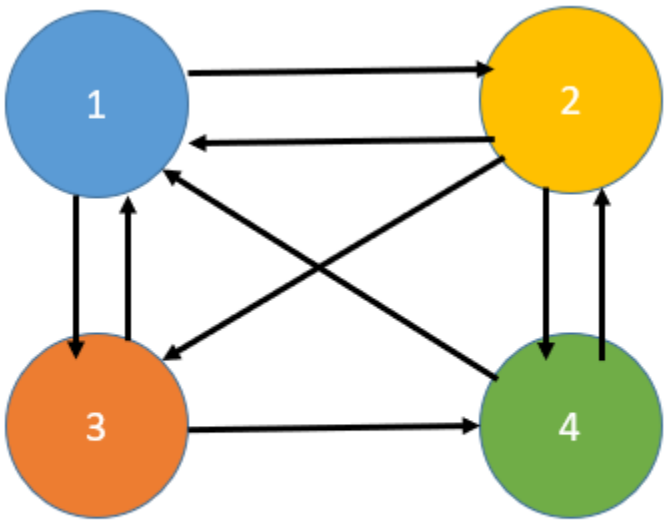
6. Submission Date:

a. To be announced on the forum

7. Tasks:

a. These following tasks are to be computed

i. First implement the page rank algorithm described in class using the MultiInputFile and demonstrate with the sample 4 node web graph described in the class given below. You have to represent the data in the form of a sparse matrix.



ii. Next convert the SNAP dataset to a sparse matrix representation and store in HDFS. Then run the page rank algorithm using Map Reduce. How many iterations does it require the algorithm to converge. Measure performance numbers of total time taken.

b. Write the output of the mapper and reducer for each question in the report.

c. Submit one page report based on template and answer the questions on the report.

8. Submission Link:

Will be shared with you on the portal at a later date.