Project Two

The main steps in the implementation of Page Rank Algorithm using Hadoop are as follows:-

1. Generating <Key,Value> Pairs:- Nodes and unique IDs or Webpage names can be considered to be keys. Identification of dangling nodes is done through the values present in the adjacency matrix. The calculated value for each page is the combination of the current page rank value and the entire outward link information for that page. This information can be differentiated with the help of any separator in our case '#'. A <key,value> represents a lot of relevant information like the source and the target URLs, page rank values, edges between the nodes and also the dangling nodes. For dangling nodes we distribute its rank to all the other nodes equally.

2. Level 1 (Create Graph)
   To obtain key value pairs form adjacency matrix.
   Map Input: The Adjacency Matrix representing the complete Web graph.
   Map output: Add a column "initial Page Rank Value" and insert a "#" which separates the page rank value and the out links.
   Reducer: Take the input from the mapper and write it to the HDFS since the first level mapper runs on the local File System.
3. Level 2 (Page Rank implementation)
   Calculate the page rank values for each node iteratively
   Map: Take in the # separated input from the Level 1 MapReduce and calculate the page rank values of each node.
   Reduce: Take the input from the Map task and sum all the page rank values for each node iteratively.
4. MapReduce Level 3 (Format results)
   The sole purpose of this MapReduce job is to clean the results and write them in a presentable way.
   Cleanup:
   Map: Since we are only interested in the WebPages and their page rank values this function just removes the target URL list, so basically it removes everything after the Page Rank values including the #.
   Reduce: This further combines the result from the map and write the result as <pagerank,url>.