Report

Class Number: CS6240 SEC02

HW Number: 3

Name: Gautam Vashisht

Design Discussion

In preprocessing, Program is using the map reduce class to parse the input file. Parser is same as given in hw3 with a minor update of handling the error related to & in input file. Mapper is emitting the edges and parent node. Reducer is used to emit the parent nodes with edges and the dangling nodes which are not present as parent node in input file.

Program is using solution2 approach to calculate the dangling ranks in the previous reducer phase. It is handled using global counter dangling. It is better approach as we don't need extra map reduce job to count dangling node share. Only trade-off is using a global counter but as program is using few global counters. So it's not an issue.

For top100, Program is using an additional map reduce job to emit top 100 pageranks using TreeMap data structure(pageRank as key and pageName as value).

Data transferred in each iteration of pageRank –

Iteration	Mapper to Reducer	Reducer to HDFS
1	4067357658	1412417183
2	4148424554	1445859373
3	4250483288	1479433500
4	4352300754	1513027530
5	4454308535	1546631324
6	4554491484	1578712929
7	4652040595	1610783884
8	4750747683	1642877169
9	4849345643	1673425243

10	4946642184	1705501886

Data transferred is given in bytes and nearly same per iteration but it is increasing a bit per iteration. This small increase might be because of change in some precision of data points or so. As it should be same because number of records entering and leaving in each iteraton is same.

Performance Comparison

	Pre- Processing Time	Iteration Time	Top 100
6m4.large Machines	36 minutes	32 minutes	Failed after 2 minute
11m4.large Machine	22 minutes	20 minutes	Failed after 1 minute

Ouput of timing is as expected, it ran pretty faster for 11 machines as data is distributed over more number of machines. Therfore, it has processed more faster.

Speed up = 6m4 time/ 11m4 time = 70/43 = 1.62

My program is failing on AWS for predicting top 100 pageRanks using treeMap. It is running well on local machine for sample file but not working fine on AWS. I have attached the syslogs file for both the runs on AWS.