CS 6240: Assignment 4

Xiao Wang

GitHub: <https://github.ccs.neu.edu/xiaowang/CS6240-MapReduce>

1. Program Discussion

**Q:** Describe briefly how each step of your program is transforming the data. Be precise: show the schema of each RDD or DataSet (i.e., the “table” header) and briefly state what type of information each record (i.e., “row”) stores.

**A:** (also commented in source code)  
 graph: [(pageName, [link])]  
 scores: [(pageName, score)]  
 contribs: [(pageName, outScore)]

**Q:** For each step, state if the dependency is narrow (no shuffling) or wide (shuffling). How many stages does your Spark have?

**A:** Stage 1: Preprocessing: narrow  
 Stage 2: Iteration: wide  
 Stage 3: Ranking: wide

1. Performance

|  |  |  |
| --- | --- | --- |
| time (s) | 5 workers | 10 workers |
| MapReduce | 4036 | 1990 |
| Spark | 41563 | 21236 |

**Q:** Discuss which system is faster and briefly explain what could be the main reason for this performance difference.

**A:** The MapReduce system is faster. I think the main reason is that during the iteration stage my Spark program use a join operation to get the full graph information with pagerank which requires a lot of time to complete.