Program Discussion

Steps	Input	Output
Parsing	File of Lines	Pair RDD of pages and adjacency List
PageRank	Pair RDD of pages and adjacency List	Pair RDD of pages and page Rank
TopK	Pair RDD of pages and page Rank	Pair RDD of pages and page Rank

Steps	Shuffling
Parsing	Narrow Shuffling(map, filter)
PageRank	Wide Shuffling(reduceByKey)
TopK	Wide Shuffling (repartition)

Performance Comparison

Spark Job

Machines	ParseJob(in s)	PageRankJob(in s)	TopKJob(in s)
6 m4.Large	1085.152914136	546.363581093	15.435811096
11 m4.Large	638.526131646	370.991811585	9.175283556

Hadoop Job

Machines	ParseJob(in s)	PageRankJob(in s)	TopKJob(in s)
6 m4.Large	1720.723	1995.712	591.565
11 m4.Large	1313.246	1143.019	220.731

As expected from the above results, the Spark Job runs faster than Hadoop jobs.

Reasons:

1. The data processing is done in-memory, for Spark jobs while for Hadoop jobs, the data is written to the HDFS.
2. Spark launches tasks much faster. MapReduce starts a new JVM for each task, which can take seconds with loading JARs, etc.