Assignment4

Chengze Li

1.Describe briefly how each step of your program is transforming the data. Be precise, e.g., by showing the structure of the input and output as a table. (10 points)

Steps	input	output	shuffle
preprocess	bz file lines	keyValueRdd(pagename, linkPages)	no
construct	RDD(pagename, linkPages)	RDD(pageName, node(weight, linkPages))	no
iterative process	RDD(pageName, node(weight, linkPages))	RDD(pageName, Any(linkPages, received weight))	yes
	RDD(pageName, Any(linkPages, received weight))	RDD(pageName, node(new weight, linkPages))	no
get final result	RDD(pageName, node(new weight, linkPages))	sort the input and get the first 100 records	yes

in my program, there are 11 stages, the preprocess and construct steps don't have shuffle, in iterative process step, there are ten shuffles, so there are 10 stages, in final step there is a shuffle, so totally there are 11 stages

Performance Comparison

6 machines:

Hadoop:

```
(1) preprocess:
GC time elapsed (ms)=180622 CPU time spent (ms)=17770690
(2) pagerank:
GC time elapsed (ms)=48575 CPU time spent (ms)=1094370
GC time elapsed (ms)=48250 CPU time spent (ms)=1104930
GC time elapsed (ms)=49497 CPU time spent (ms)=1105890
GC time elapsed (ms)=46347 CPU time spent (ms)=1102920
GC time elapsed (ms)=47865 CPU time spent (ms)=1104040
GC time elapsed (ms)=45966 CPU time spent (ms)=1095500
GC time elapsed (ms)=47407 CPU time spent (ms)=1101210
GC time elapsed (ms)=46583 CPU time spent (ms)=1096120
GC time elapsed (ms)=47808 CPU time spent (ms)=1096080
(3)top-k
GC time elapsed (ms)=34286 CPU time spent (ms)=202310
```

Spark:

```
INFO total process run time: 5752 seconds
2017-11-03T03:44:11.963Z INFO Step created jobs:
2017-11-03T03:44:11.963Z INFO Step succeeded with exitCode 0 and took 5752
seconds
```

11 machines

Hadoop:

```
(1) preprocessing:
GC time elapsed (ms)=174987
CPU time spent (ms)=17016360
(2) pagerank (10 iterations)
GC time elapsed (ms)=31971 CPU time spent (ms)=979900
GC time elapsed (ms)=33387 CPU time spent (ms)=981740
GC time elapsed (ms)=31344 CPU time spent (ms)=978160
GC time elapsed (ms)=31605 CPU time spent (ms)=971010
GC time elapsed (ms)=31693 CPU time spent (ms)=979200
GC time elapsed (ms)=30939 CPU time spent (ms)=987780
GC time elapsed (ms)=32061 CPU time spent (ms)=972490
GC time elapsed (ms)=33050 CPU time spent (ms)=975310
GC time elapsed (ms)=31335 CPU time spent (ms)=992060
GC time elapsed (ms)=31971 CPU time spent (ms)=979900
(3) top-k
GC time elapsed (ms)=19455 CPU time spent (ms)=166600
```

Spark:

```
INFO total process run time: 2956 seconds
2017-11-03T02:58:46.333Z INFO Step created jobs:
2017-11-03T02:58:46.333Z INFO Step succeeded with exitCode 0 and took 2956
seconds
```

. Discuss which system is faster and briefly explain what could be the main

. reason for this performance difference.

Obviously the spark is faster than hadoop, because spark use in-memory model whereas hadoop retrieve the data from disk. we can notice that the speedup effects is more obvious for 11 machines of spark test, I think the reason is with more memory, the more fast the program can run.