CS8803 BDS / CS4365

Homework Assignment 1

(Programming Category)

Student Name: Enmao Diao

**Problem 1. Learning HDFS and Hadoop MapReduce.**

*Suitable for students who are the beginner of Hadoop MapReduce Platform*

Install HDFS and Hadoop MapReduce on your laptop, and run the word count map-reduce program, and report the runtime for two different sizes of datasets.

You may use excel file to generate your runtime statistics plot or organize the performance measurement data in a tabular format.

You are encouraged to learn by observing the runtime performance of Hadoop MapReduce program through different ways of programming the same problem and show their impact on the runtime performance of the MapReduce job.

Deliverable.

1. Source code (see WordCount.java)
2. screen shots of your execution process.
3. Runtime statistics in excel plots or tabular format.

Data files downloaded from [Gutenberg](http://www.gutenberg.org/wiki/Main_Page)

Ulysses by James Joyce (pg4300.txt) 1.50Mb

Metamorphosis by Franz Kafka (pg5200.txt) 138 kb

….

(b)

**General Set up**

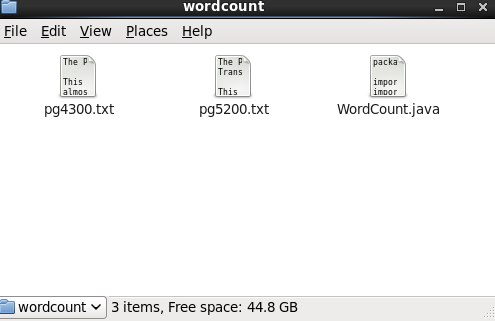
****

Fig. 1. Sample data files and Source code.

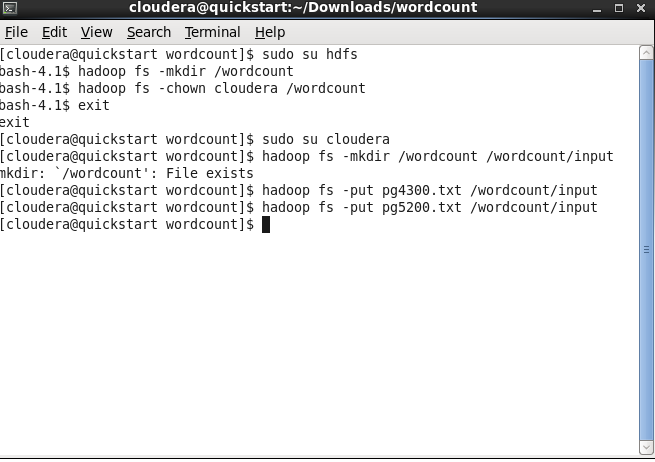
****

Fig. 2. Make directory /wordcount/input and put data files in it.

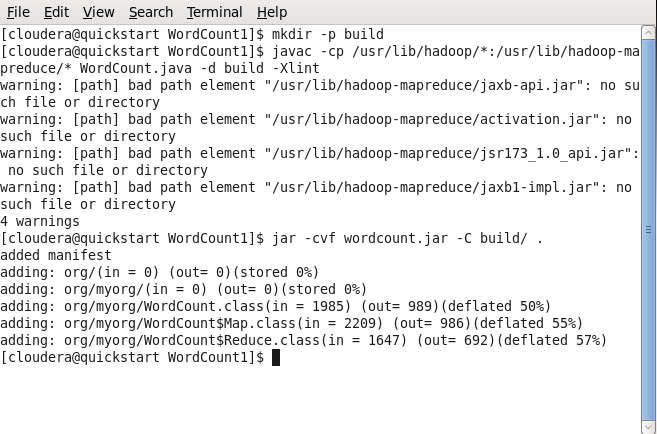
****

Fig. 3. Compile Source code and get jar file.

**Run**

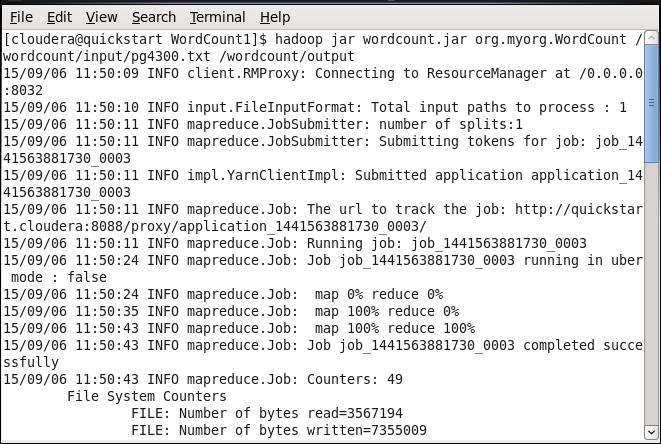
****

Fig. 4a. Start running on pg4300.txt.

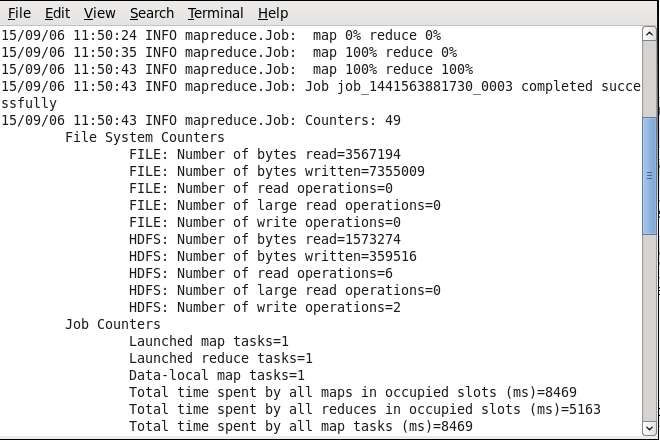
****

Fig. 4b. Start running on pg4300.txt.

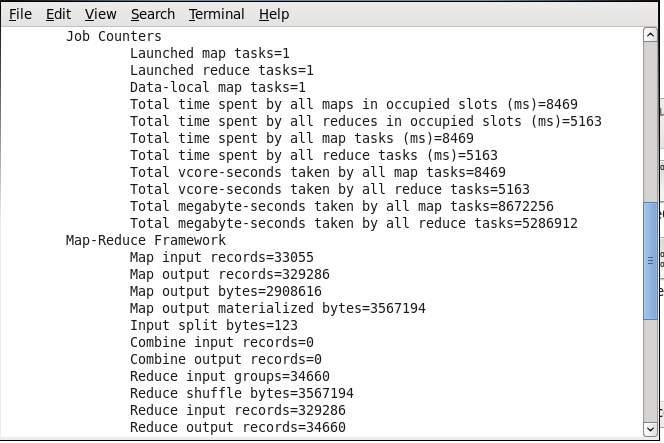
****

Fig. 4c. Start running on pg4300.txt.

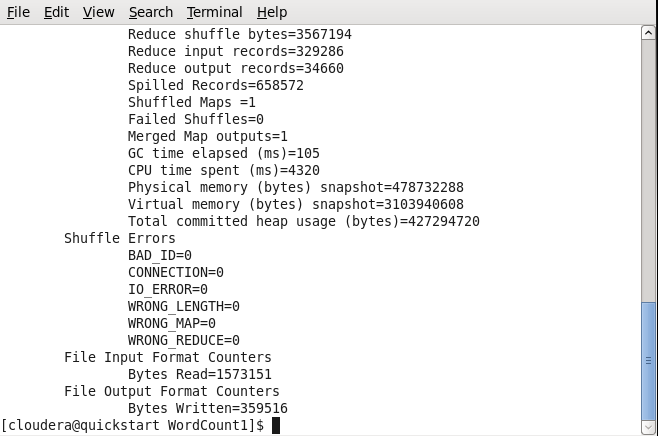
****

Fig. 4d. Start running on pg4300.txt.

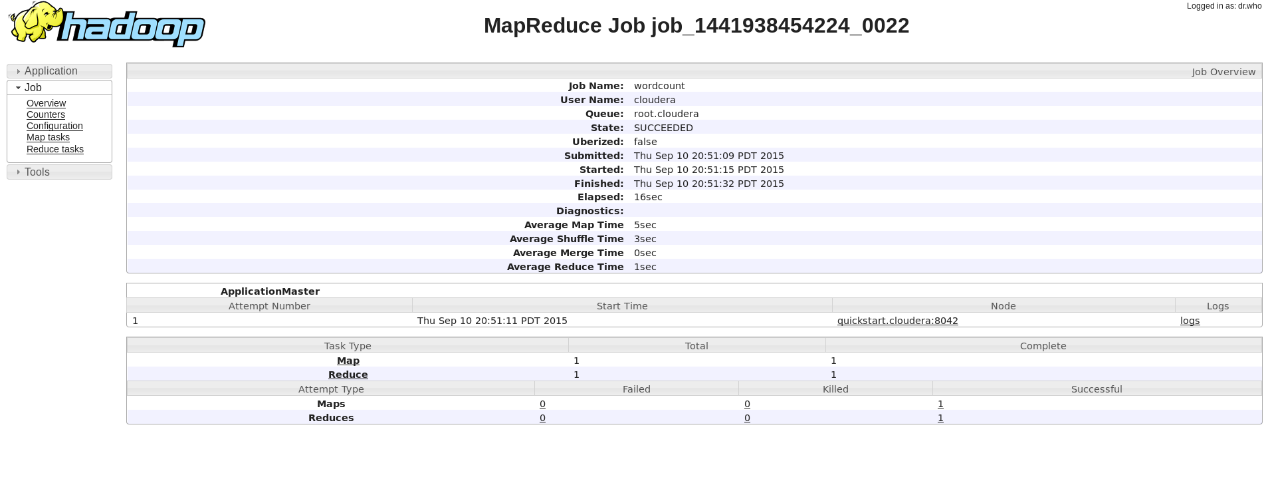
****

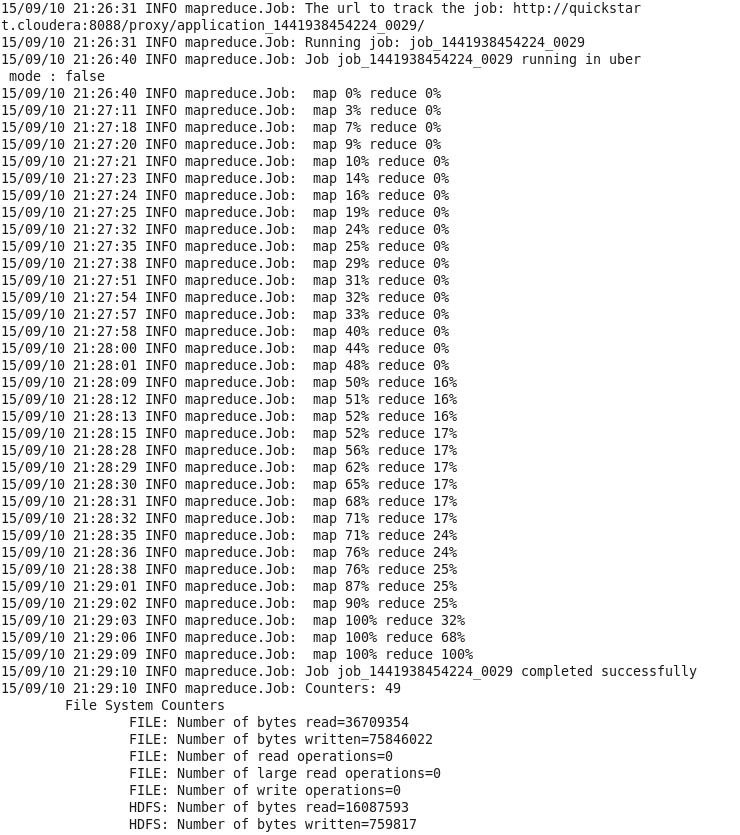
Fig. 5. Job Summary from ResourceManager at Port Localhost:8088.

(c)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Filename | Size  (KB) | Elapsed time (s) | Average Map Time (s) | Average Shuffle Time (s) | Average Merge Time (s) | Average Reduce Time (s) |
| pg23 | 243 | 16 | 5 | 5 | 0 | 1 |
| pg74 | 412 | 17 | 6 | 4 | 0 | 1 |
| pg76 | 596 | 17 | 6 | 5 | 0 | 1 |
| pg84 | 439 | 17 | 5 | 5 | 0 | 1 |
| pg98 | 775 | 16 | 5 | 4 | 0 | 1 |
| pg158 | 898 | 17 | 6 | 4 | 0 | 1 |
| pg174 | 452 | 16 | 4 | 4 | 0 | 1 |
| pg345 | 863 | 18 | 6 | 5 | 0 | 1 |
| pg844 | 140 | 16 | 5 | 4 | 0 | 1 |
| pg1184 | 2624 | 18 | 6 | 4 | 0 | 1 |
| pg1232 | 299 | 17 | 4 | 5 | 0 | 1 |
| pg1322 | 758 | 17 | 5 | 4 | 0 | 1 |
| pg1661 | 581 | 16 | 5 | 4 | 0 | 1 |
| pg2591 | 537 | 16 | 5 | 4 | 0 | 1 |
| pg2701 | 1228 | 17 | 5 | 4 | 0 | 1 |
| pg4300 | 1537 | 16 | 5 | 3 | 0 | 1 |
| pg5200 | 139 | 16 | 5 | 4 | 0 | 1 |
| pg6130 | 1174 | 17 | 6 | 4 | 0 | 1 |
| pg8800 | 627 | 18 | 5 | 5 | 0 | 1 |
| pg27827 | 352 | 17 | 5 | 5 | 0 | 1 |
| pg30254 | 1045 | 17 | 5 | 4 | 0 | 1 |
| Sum | 15719 | 352 | 109 | 90 | 0 | 21 |
| pg(844+5200) | 279 | 21 | 10 | 4 | 0 | 0 |
| pg(23+74+76) | 1251 | 25 | 14 | 4 | 0 | 1 |
| pg(2701+4300+6130+30254) | 4984 | 36 | 24 | 4 | 0 | 1 |
| pg(23to1232) | 7741 | 80 | 34 | 35 | 1 | 2 |
| All data | 15719 | 149 | 34 | 85 | 1 | 2 |
| All data in one file | 15719 | 30 | 14 | 4 | 1 | 3 |

Thought:

1. For each single file, run time does not vary much. If the file is large, its map time, shuffle time and reduce time slightly increase.
2. When multiple files run, map time increases significantly. If the number of files is extremely large (about 10), shuffle time also increases However, all the other run time remain stable. More files mean more map tasks but only one reduce task.
3. If Map task completes more than 50% and run time passes over a time threshold, the Reduce task will get started. (By observation)



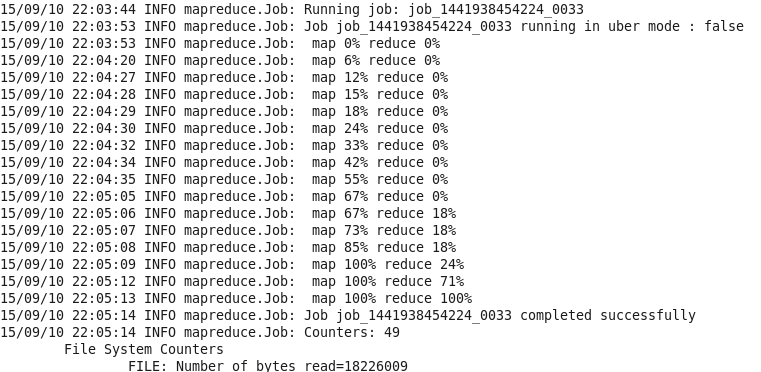


Fig. 6. Map task and Reduce task completion.

1. Size and the number of files affect run time, and I guess the number of files affect the most.

Fig. 7. Elapsed time vs Size.