

# 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.

- (a) Source code (see WordCount.java)
- (b) screen shots of your execution process.
- (c) Runtime statistics in excel plots or tabular format.

Data files downloaded from [Gutenberg](#)

Ulysses by James Joyce (pg4300.txt) 1.50Mb

Metamorphosis by Franz Kafka (pg5200.txt) 138 kb

....

(b)

## General Set up

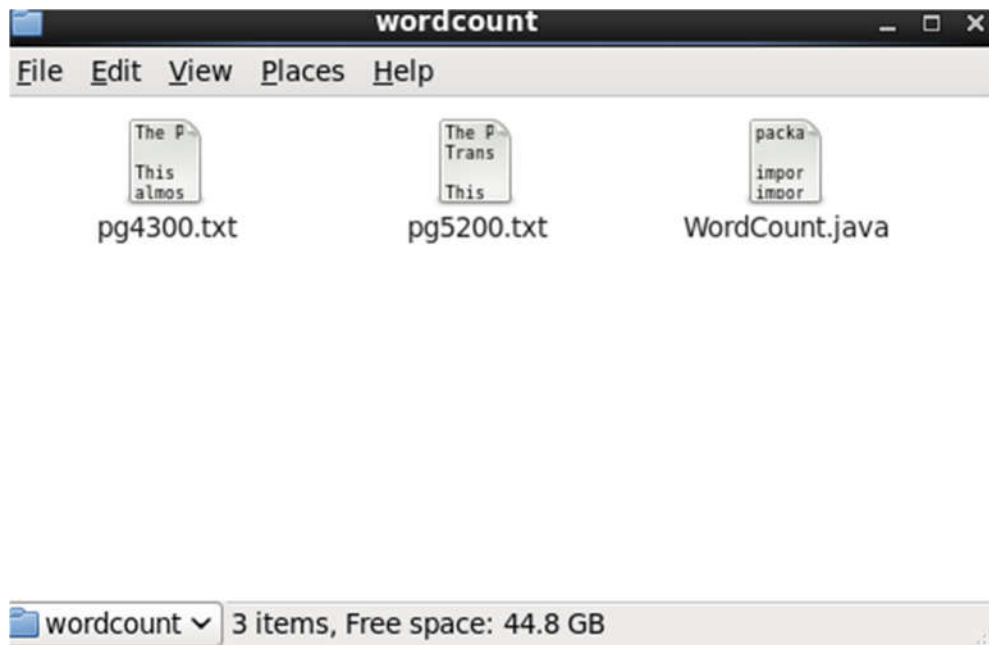


Fig. 1. Sample data files and Source code.

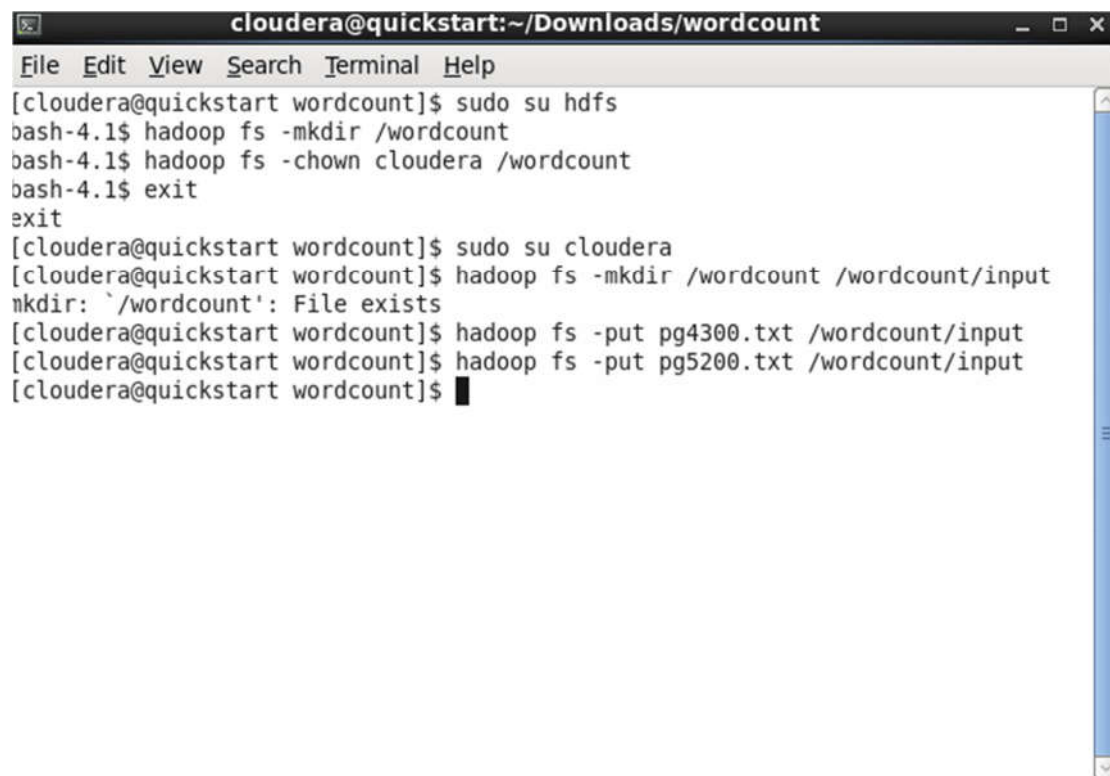



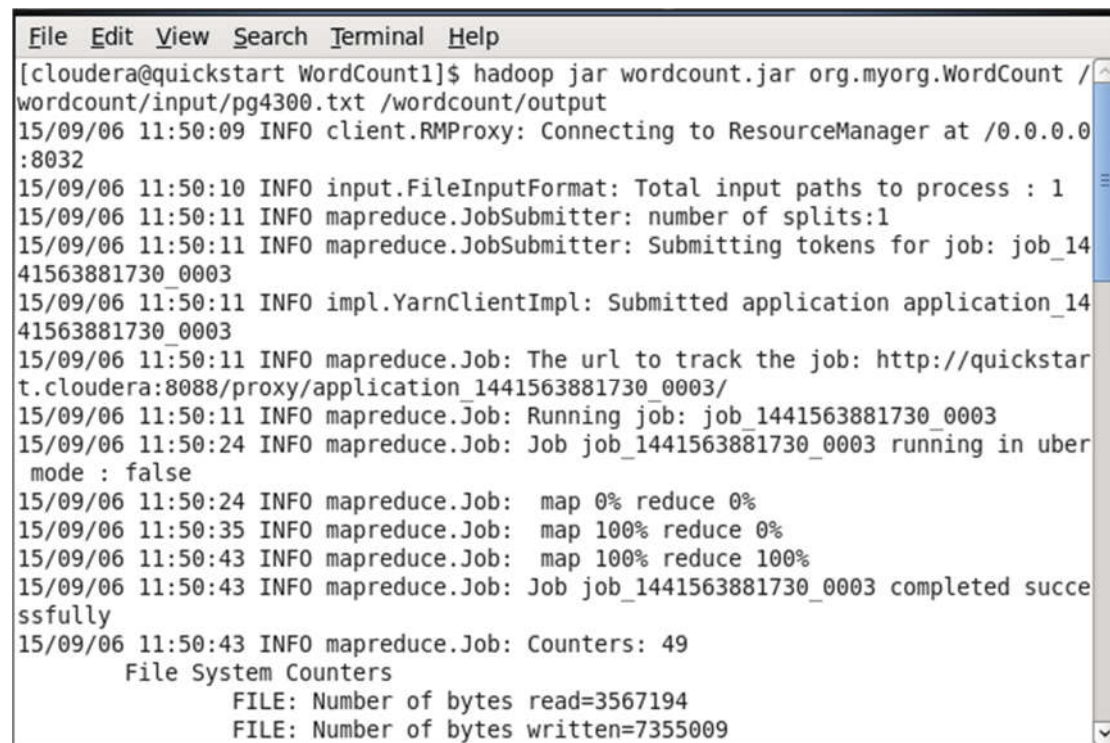
Fig. 2. Make directory /wordcount/input and put data files in it.



```
File Edit View Search Terminal Help
[cloudera@quickstart WordCount1]$ mkdir -p build
[cloudera@quickstart WordCount1]$ javac -cp /usr/lib/hadoop/*:/usr/lib/hadoop-mapreduce/* WordCount.java -d build -Xlint
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jaxb-api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/activation.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jsr173_1.0_api.jar": no such file or directory
warning: [path] bad path element "/usr/lib/hadoop-mapreduce/jaxb1-impl.jar": no such file or directory
4 warnings
[cloudera@quickstart WordCount1]$ jar -cvf wordcount.jar -C build/ .
added manifest
adding: org/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/WordCount.class(in = 1985) (out= 989)(deflated 50%)
adding: org/myorg/WordCount$Map.class(in = 2209) (out= 986)(deflated 55%)
adding: org/myorg/WordCount$Reduce.class(in = 1647) (out= 692)(deflated 57%)
[cloudera@quickstart WordCount1]$
```

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

## Run



```
File Edit View Search Terminal Help
[cloudera@quickstart WordCount1]$ hadoop jar wordcount.jar org.myorg.WordCount /wordcount/input/pg4300.txt /wordcount/output
15/09/06 11:50:09 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
15/09/06 11:50:10 INFO input.FileInputFormat: Total input paths to process : 1
15/09/06 11:50:11 INFO mapreduce.JobSubmitter: number of splits:1
15/09/06 11:50:11 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1441563881730_0003
15/09/06 11:50:11 INFO impl.YarnClientImpl: Submitted application application_1441563881730_0003
15/09/06 11:50:11 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1441563881730_0003/
15/09/06 11:50:11 INFO mapreduce.Job: Running job: job_1441563881730_0003
15/09/06 11:50:24 INFO mapreduce.Job: Job job_1441563881730_0003 running in uber mode : false
15/09/06 11:50:24 INFO mapreduce.Job: map 0% reduce 0%
15/09/06 11:50:35 INFO mapreduce.Job: map 100% reduce 0%
15/09/06 11:50:43 INFO mapreduce.Job: map 100% reduce 100%
15/09/06 11:50:43 INFO mapreduce.Job: Job job_1441563881730_0003 completed successfully
15/09/06 11:50:43 INFO mapreduce.Job: Counters: 49
File System Counters
FILE: Number of bytes read=3567194
FILE: Number of bytes written=7355009
```

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

```
File Edit View Search Terminal Help
15/09/06 11:50:24 INFO mapreduce.Job: map 0% reduce 0%
15/09/06 11:50:35 INFO mapreduce.Job: map 100% reduce 0%
15/09/06 11:50:43 INFO mapreduce.Job: map 100% reduce 100%
15/09/06 11:50:43 INFO mapreduce.Job: Job job_1441563881730_0003 completed successfully
15/09/06 11:50:43 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=3567194
    FILE: Number of bytes written=7355009
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=1573274
    HDFS: Number of bytes written=359516
    HDFS: Number of read operations=6
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Launched reduce tasks=1
    Data-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=8469
    Total time spent by all reduces in occupied slots (ms)=5163
    Total time spent by all map tasks (ms)=8469
```

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

```
File Edit View Search Terminal Help
  Job Counters
    Launched map tasks=1
    Launched reduce tasks=1
    Data-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=8469
    Total time spent by all reduces in occupied slots (ms)=5163
    Total time spent by all map tasks (ms)=8469
    Total time spent by all reduce tasks (ms)=5163
    Total vcore-seconds taken by all map tasks=8469
    Total vcore-seconds taken by all reduce tasks=5163
    Total megabyte-seconds taken by all map tasks=8672256
    Total megabyte-seconds taken by all reduce tasks=5286912
  Map-Reduce Framework
    Map input records=33055
    Map output records=329286
    Map output bytes=2908616
    Map output materialized bytes=3567194
    Input split bytes=123
    Combine input records=0
    Combine output records=0
    Reduce input groups=34660
    Reduce shuffle bytes=3567194
    Reduce input records=329286
    Reduce output records=34660
```


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

```

File Edit View Search Terminal Help
Reduce shuffle bytes=3567194
Reduce input records=329286
Reduce output records=34660
Spilled Records=658572
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=105
CPU time spent (ms)=4320
Physical memory (bytes) snapshot=478732288
Virtual memory (bytes) snapshot=3103940608
Total committed heap usage (bytes)=427294720
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=1573151
File Output Format Counters
  Bytes Written=359516
[cloudera@quickstart WordCount1]$

```

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



MapReduce Job job\_1441938454224\_0022

Logged in as: drisho

+ Application
+ Job
+ Overview
+ Configuration
+ Map tasks
+ Reduce tasks
+ Tools

Job Name: wordcount
User Name: cloudera
Queue: root.cloudera
State: SUCCEEDED
Uberized: false
Submitted: Thu Sep 10 20:51:09 PDT 2015
Started: Thu Sep 10 20:51:15 PDT 2015
Finished: Thu Sep 10 20:51:32 PDT 2015
Elapsed: 16sec
Diagnostics:
Average Map Time: 5sec
Average Shuffle Time: 1sec
Average Merge Time: 0sec
Average Reduce Time: 11sec

| ApplicationMaster | Attempt Number               | Start Time               | Node | Logs |
|-------------------|------------------------------|--------------------------|------|------|
| 1                 | Thu Sep 10 20:51:11 PDT 2015 | quickstart.cloudera:8042 |      | logs |

| Task Type | 1 | Total | 1 | Complete |
|-----------|---|-------|---|----------|
| Map       | 1 |       | 1 |          |
| Reduce    | 1 |       | 1 |          |

| Attempt Type | Failed | Killed | Successful |
|--------------|--------|--------|------------|
| Maps         | 0      | 0      | 1          |
| Reduces      | 0      | 0      | 1          |

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)



```

15/09/10 21:26:31 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1441938454224_0029/
15/09/10 21:26:31 INFO mapreduce.Job: Running job: job_1441938454224_0029
15/09/10 21:26:40 INFO mapreduce.Job: Job job_1441938454224_0029 running in uber
mode : false
15/09/10 21:26:40 INFO mapreduce.Job: map 0% reduce 0%
15/09/10 21:27:11 INFO mapreduce.Job: map 3% reduce 0%
15/09/10 21:27:18 INFO mapreduce.Job: map 7% reduce 0%
15/09/10 21:27:20 INFO mapreduce.Job: map 9% reduce 0%
15/09/10 21:27:21 INFO mapreduce.Job: map 10% reduce 0%
15/09/10 21:27:23 INFO mapreduce.Job: map 14% reduce 0%
15/09/10 21:27:24 INFO mapreduce.Job: map 16% reduce 0%
15/09/10 21:27:25 INFO mapreduce.Job: map 19% reduce 0%
15/09/10 21:27:32 INFO mapreduce.Job: map 24% reduce 0%
15/09/10 21:27:35 INFO mapreduce.Job: map 25% reduce 0%
15/09/10 21:27:38 INFO mapreduce.Job: map 29% reduce 0%
15/09/10 21:27:51 INFO mapreduce.Job: map 31% reduce 0%
15/09/10 21:27:54 INFO mapreduce.Job: map 32% reduce 0%
15/09/10 21:27:57 INFO mapreduce.Job: map 33% reduce 0%
15/09/10 21:27:58 INFO mapreduce.Job: map 40% reduce 0%
15/09/10 21:28:00 INFO mapreduce.Job: map 44% reduce 0%
15/09/10 21:28:01 INFO mapreduce.Job: map 48% reduce 0%
15/09/10 21:28:09 INFO mapreduce.Job: map 50% reduce 16%
15/09/10 21:28:12 INFO mapreduce.Job: map 51% reduce 16%
15/09/10 21:28:13 INFO mapreduce.Job: map 52% reduce 16%
15/09/10 21:28:15 INFO mapreduce.Job: map 52% reduce 17%
15/09/10 21:28:28 INFO mapreduce.Job: map 56% reduce 17%
15/09/10 21:28:29 INFO mapreduce.Job: map 62% reduce 17%
15/09/10 21:28:30 INFO mapreduce.Job: map 65% reduce 17%
15/09/10 21:28:31 INFO mapreduce.Job: map 68% reduce 17%
15/09/10 21:28:32 INFO mapreduce.Job: map 71% reduce 17%
15/09/10 21:28:35 INFO mapreduce.Job: map 71% reduce 24%
15/09/10 21:28:36 INFO mapreduce.Job: map 76% reduce 24%
15/09/10 21:28:38 INFO mapreduce.Job: map 76% reduce 25%
15/09/10 21:29:01 INFO mapreduce.Job: map 87% reduce 25%
15/09/10 21:29:02 INFO mapreduce.Job: map 90% reduce 25%
15/09/10 21:29:03 INFO mapreduce.Job: map 100% reduce 32%
15/09/10 21:29:06 INFO mapreduce.Job: map 100% reduce 68%
15/09/10 21:29:09 INFO mapreduce.Job: map 100% reduce 100%
15/09/10 21:29:10 INFO mapreduce.Job: Job job_1441938454224_0029 completed successfully
15/09/10 21:29:10 INFO mapreduce.Job: Counters: 49
    File System Counters
      FILE: Number of bytes read=36709354
      FILE: Number of bytes written=75846022
      FILE: Number of read operations=0
      FILE: Number of large read operations=0
      FILE: Number of write operations=0
      HDFS: Number of bytes read=16087593
      HDFS: Number of bytes written=759817
15/09/10 22:03:44 INFO mapreduce.Job: Running job: job_1441938454224_0033
15/09/10 22:03:53 INFO mapreduce.Job: Job job_1441938454224_0033 running in uber mode : false
15/09/10 22:03:53 INFO mapreduce.Job: map 0% reduce 0%
15/09/10 22:04:20 INFO mapreduce.Job: map 6% reduce 0%
15/09/10 22:04:27 INFO mapreduce.Job: map 12% reduce 0%
15/09/10 22:04:28 INFO mapreduce.Job: map 15% reduce 0%
15/09/10 22:04:29 INFO mapreduce.Job: map 18% reduce 0%
15/09/10 22:04:30 INFO mapreduce.Job: map 24% reduce 0%
15/09/10 22:04:32 INFO mapreduce.Job: map 33% reduce 0%
15/09/10 22:04:34 INFO mapreduce.Job: map 42% reduce 0%
15/09/10 22:04:35 INFO mapreduce.Job: map 55% reduce 0%
15/09/10 22:05:05 INFO mapreduce.Job: map 67% reduce 0%
15/09/10 22:05:06 INFO mapreduce.Job: map 67% reduce 18%
15/09/10 22:05:07 INFO mapreduce.Job: map 73% reduce 18%
15/09/10 22:05:08 INFO mapreduce.Job: map 85% reduce 18%
15/09/10 22:05:09 INFO mapreduce.Job: map 100% reduce 24%
15/09/10 22:05:12 INFO mapreduce.Job: map 100% reduce 71%
15/09/10 22:05:13 INFO mapreduce.Job: map 100% reduce 100%
15/09/10 22:05:14 INFO mapreduce.Job: Job job_1441938454224_0033 completed successfully
15/09/10 22:05:14 INFO mapreduce.Job: Counters: 49
    File System Counters
      FILE: Number of bytes read=18226009

```

Fig. 6. Map task and Reduce task completion.

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

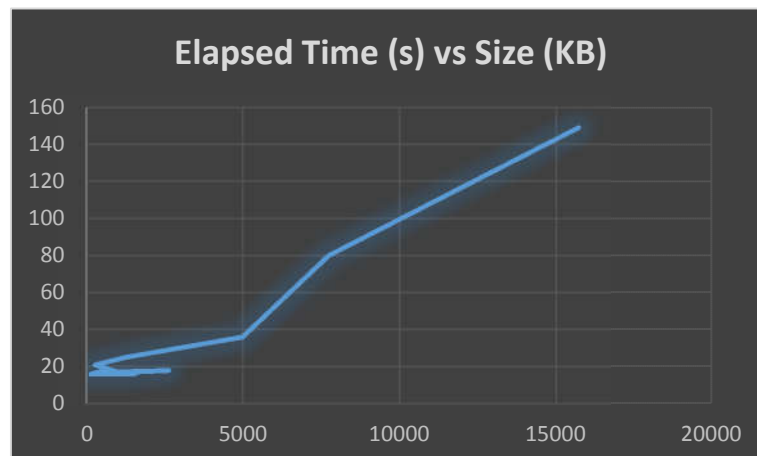


Fig. 7. Elapsed time vs Size.