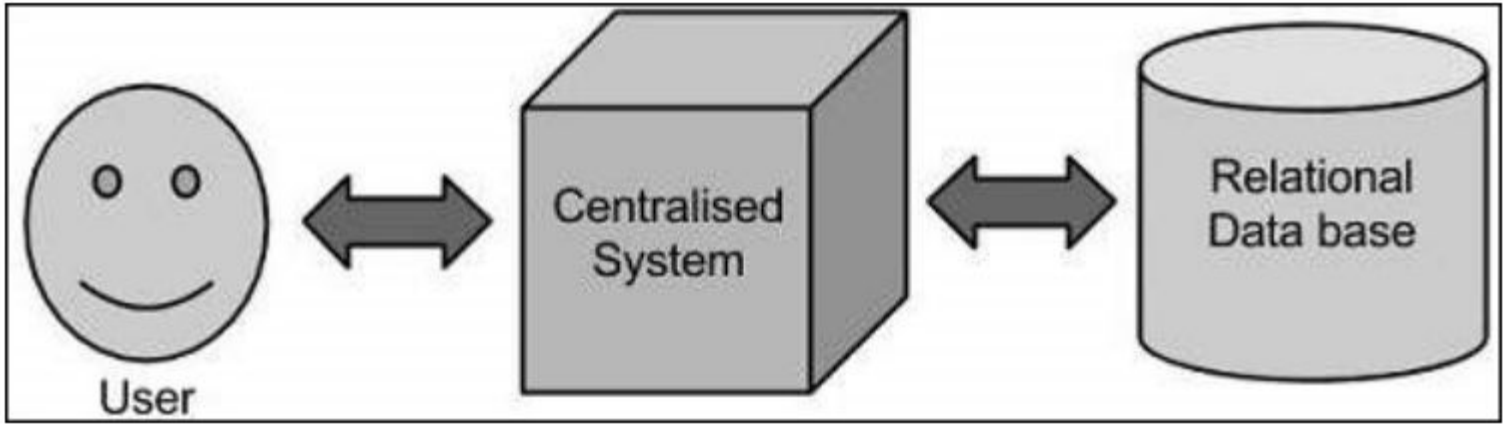
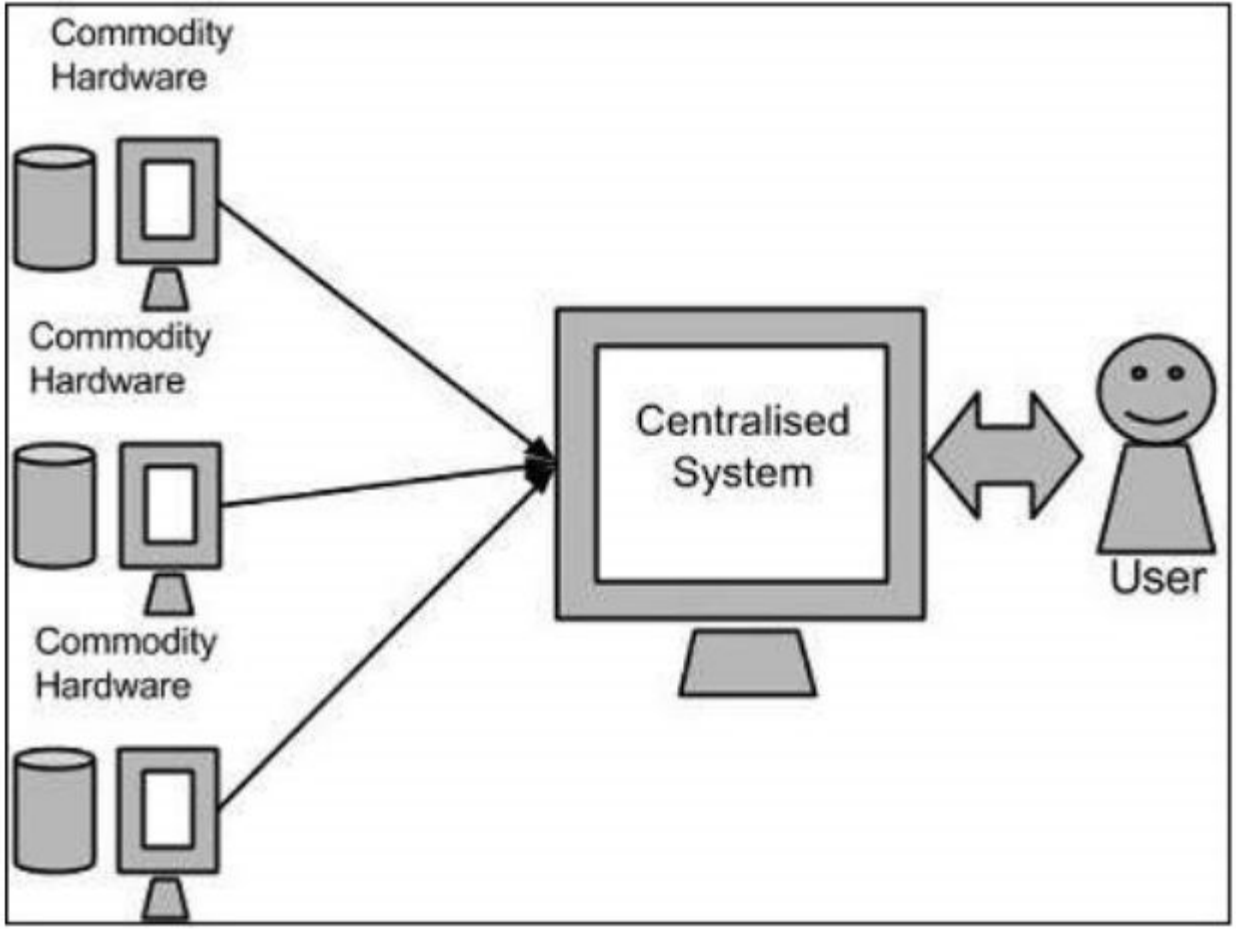
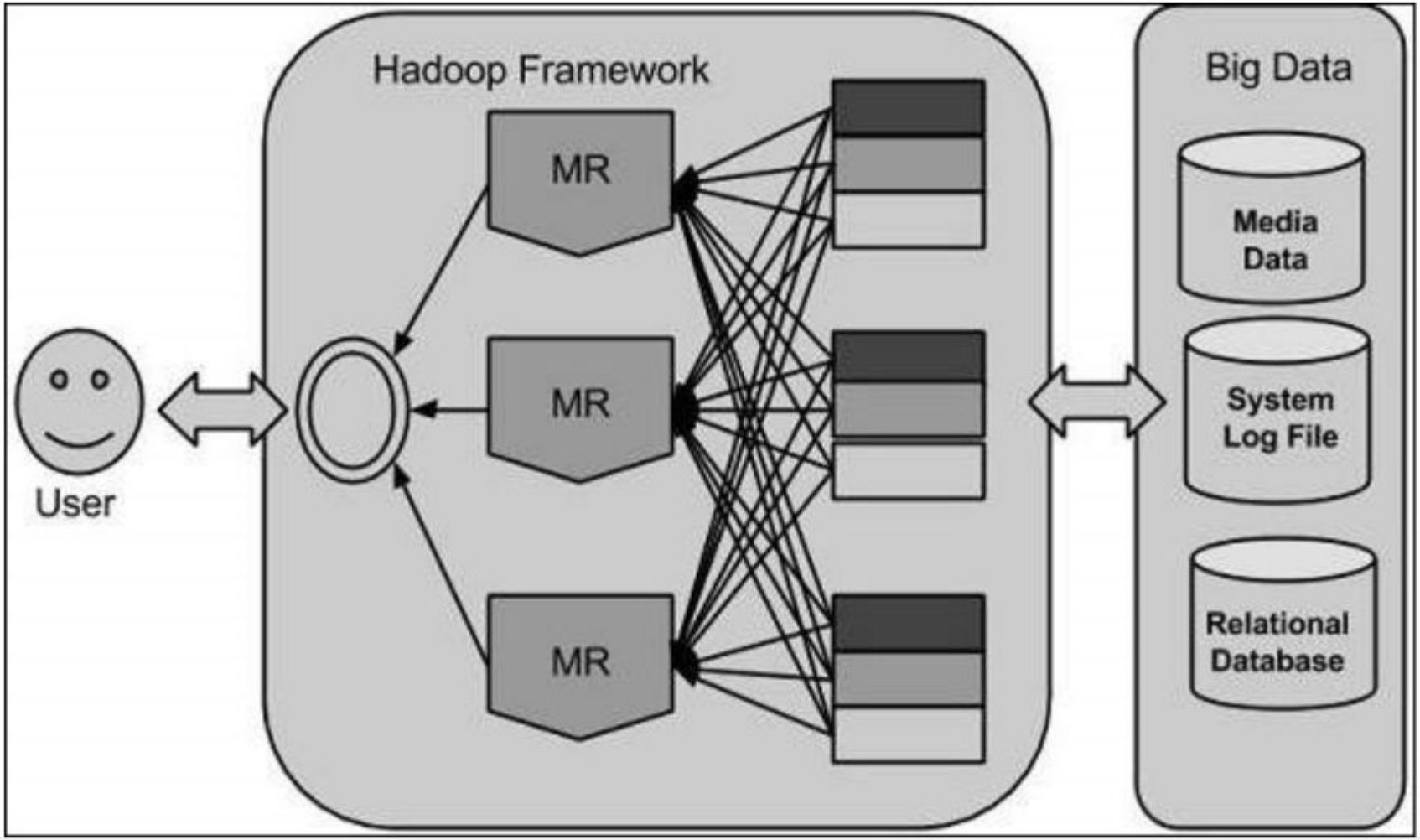
Big Data Challenges

The major challenges associated with big data are as follows –

- Capturing data
- Curation
- Storage
- Searching
- Sharing
- Transfer
- Analysis
- Presentation


User
Centralised System
Relational Data base


Commodity
Hardware
Commodity
Hardware
Commodity
Hardware
Centralised
System
User


Hadoop Framework
MR
MR
MR
User
Big Data
Media Data
System Log File
Relational Database

DOWNLOAD

https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-3.3.5/hadoop-3.3.5.tar.gz

COMMUNITY-LED DEVELOPMENT "THE APACHE WAY"

Projects ▾ People ▾ Community ▾ License ▾ Sponsors ▾

We suggest the following site for your download:

https://dlcdn.apache.org/hadoop/common/hadoop-3.3.5/hadoop-3.3.5.tar.gz

Alternate download locations are suggested below.

It is essential that you verify the integrity of the downloaded file using the PGP signature (`.asc` file) or a hash (`.md5` or `.sha*` file).

HTTP

https://dlcdn.apache.org/hadoop/common/hadoop-3.3.5/hadoop-3.3.5.tar.gz

BACKUP SITE

https://dlcdn.apache.org/hadoop/common/hadoop-3.3.5/hadoop-3.3.5.tar.gz

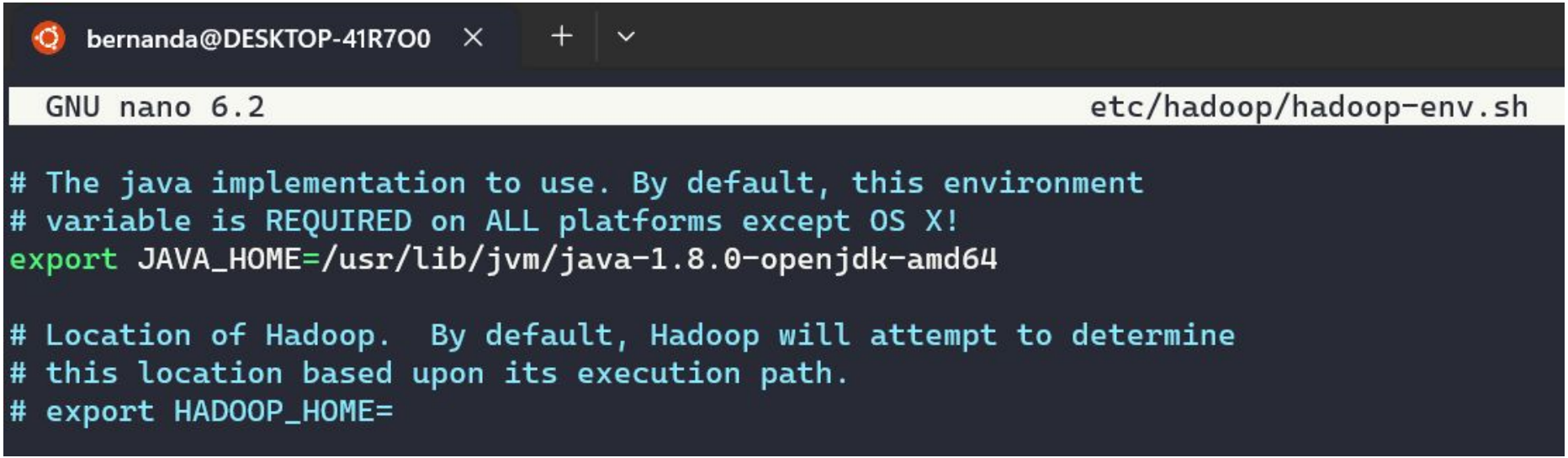
prerequisite

sudo apt install openjdk-8-jre-headless
sudo apt-get install openjdk-8-jdk

+ SSH

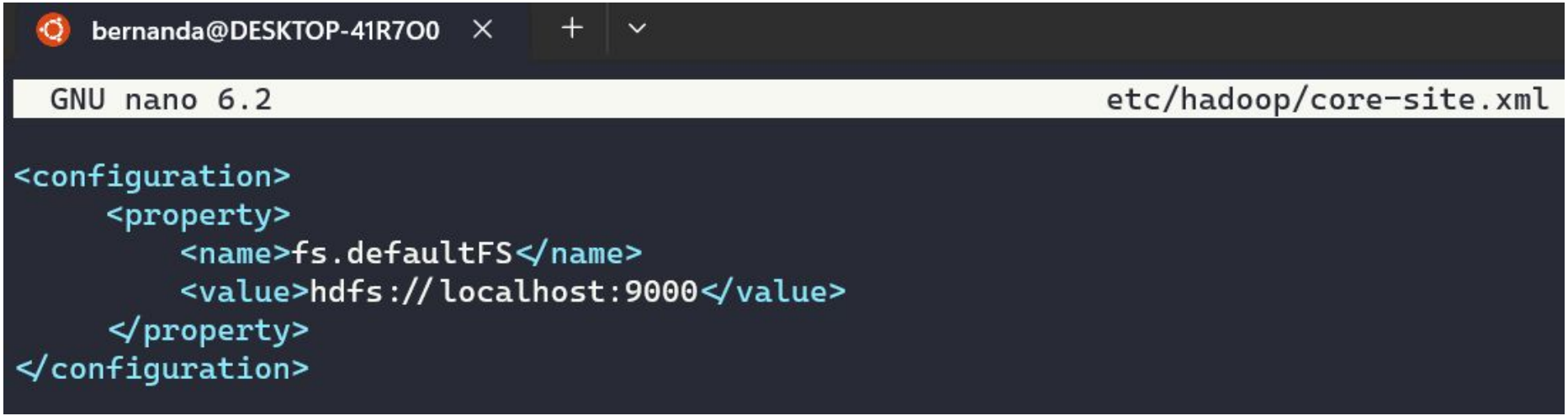
CONFIGURATION

etc/hadoop/hadoop-env.sh



CONFIGURATION

etc/hadoop/core-site.xml



CONFIGURATION

etc/hadoop/hdfs-site.xml

```
<configuration>
  <property>
      <name>dfs.namenode.name.dir</name>
      <value>/home/bernanda/hadoop/dfs/namenode</value>
  </property>
  <property>
      <name>dfs.datanode.data.dir</name>
      <value>/home/bernanda/hadoop/dfs/datanode</value>
  </property>
  <property>
      <name>dfs.replication</name>
      <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.secondary.http-address</name>
    <value>DESKTOP-41R7O0N:50090</value>
  </property>
</configuration>
```

CONFIGURATION

etc/hadoop/mapred-site.xml

```
<configuration>
    <property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
    </property>
    <property>
        <name>mapreduce.application.classpath</name>
        <value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/lib/*</value>
    </property>
</configuration>
```

```
    </property>
    <property>
        <name>mapred.child.java.opts</name>
        <value>-Xmx4096m</value>
    </property>
</configuration>
```

CONFIGURATION

etc/hadoop/yarn-site.xml

```
<configuration>
    <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
    </property>
   <property>
        <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
        <value>org.apache.hadoop.mapred.ShuffleHandler</value>
   </property>
    <property>
        <name>yarn.nodemanager.env-whitelist</name>
        <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_H>
    </property>
</configuration>
```

START THE SERVICE

```
bernanda@DESKTOP-41R7O0N:~/hadoop/hadoop-3.3.5$ ./bin/hdfs namenode -format
2023-05-10 23:21:30,956 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = DESKTOP-41R7O0N.localdomain/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.3.5
STARTUP_MSG:   classpath = /home/bernanda/hadoop/hadoop-3.3.5/etc/hadoop:/home
share/hadoop/common/lib/netty-transport-native-kqueue-4.1.77.Final-osx-aarch_6
```

START THE SERVICE

```
bernanda@DESKTOP-41R7O0N:~/hadoop/hadoop-3.3.5$ ./sbin/start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as bernanda in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [DESKTOP-41R7O0N]
Starting resourcemanager
Starting nodemanagers
bernanda@DESKTOP-41R7O0N:~/hadoop/hadoop-3.3.5$ jps
4192 ResourceManager
3572 NameNode
3972 SecondaryNameNode
3720 DataNode
4333 NodeManager
4719 Jps
bernanda@DESKTOP-41R7O0N:~/hadoop/hadoop-3.3.5$
```

Search

penerapan apa
yang cocok untuk
menggunakan
hadoop / map
reducer?

WORD COUNTING

```
hdfs dfs -mkdir -p /word_count/input
```

```
hdfs dfs -put
```

```
hadoop com.sun.tools.javac.Main WordCount.java
```

```
jar cf wc.jar WordCount*.class
```

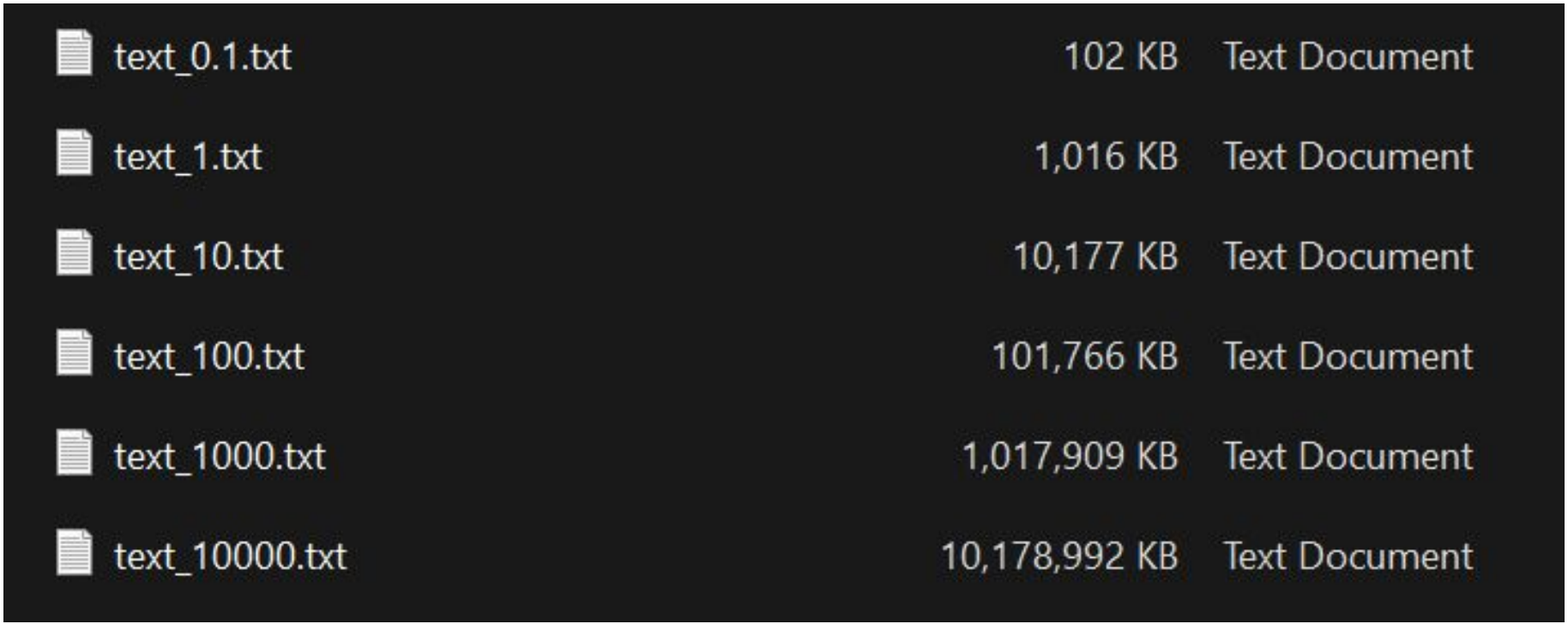
```
hadoop jar wc.jar WordCount /word_count/input /word_count/output
```

```
bernanda@DESKTOP-41R7O0N:~$ hadoop fs -ls /word_counting/output/
Found 2 items
-rw-r--r--   1 bernanda supergroup          0 2023-05-10 23:37 /word_counting/output/_SUCCESS
-rw-r--r--   1 bernanda supergroup       1467 2023-05-10 23:37 /word_counting/output/part-r-00000
bernanda@DESKTOP-41R7O0N:~$ hadoop fs -cat /word_counting/output/part-r-00000
"A      1
"Have   1
"How    1
"I      5
"Look   1
"My     1
"No!    1
"Now!   1
"Put    1
"This   1
"Why    1
"Why,   1
"With   1
ALL     1
All     1
And     11
As      1
BUMP!   2
But     4
By      1
Cat     5
Did     1
Dr.     1
For     1
Hat     2
Hat!    1
Hat.    2
Have    2
He      3
How     1
I       13
It      1
```

Hadoop (single node) MapRed
vs
Java TreeMap

Test Machine :

- UBUNTU 20.04
- RAM 16 Gigs
- 4 CORE

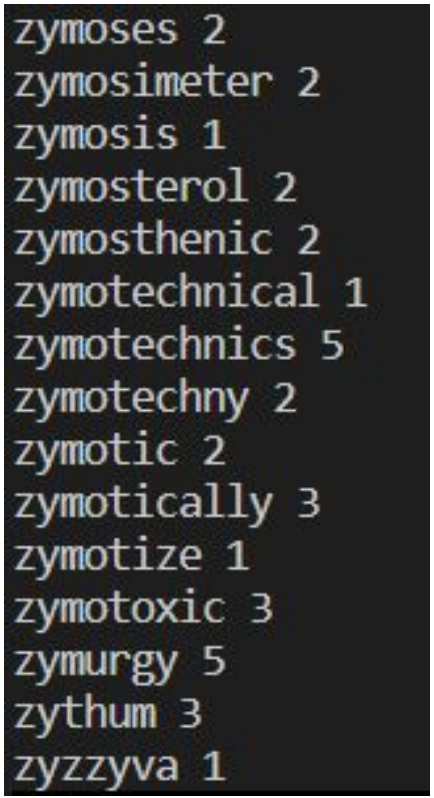

text_0.1.txt 102 KB Text Document
text_1.txt 1,016 KB Text Document
text_10.txt 10,177 KB Text Document
text_100.txt 101,766 KB Text Document
text_1000.txt 1,017,909 KB Text Document
text_10000.txt 10,178,992 KB Text Document

```python
def load_word_pool(file_path):
    with open(file_path, 'r') as file:
        word_pool = file.read().split()
    return word_pool


def generate_random_text(word_pool, size_mb):
    words_per_mb = 100000
    total_words = int(size_mb * words_per_mb)

    list_of_words = random.choices(word_pool, k=total_words)
    random_text = ' '.join(list_of_words)
    return random_text


def save_text_to_file(text, file_path):
    with open(file_path, 'w') as file:
        file.write(text)
```

Both program has the same result

HADOOP

=

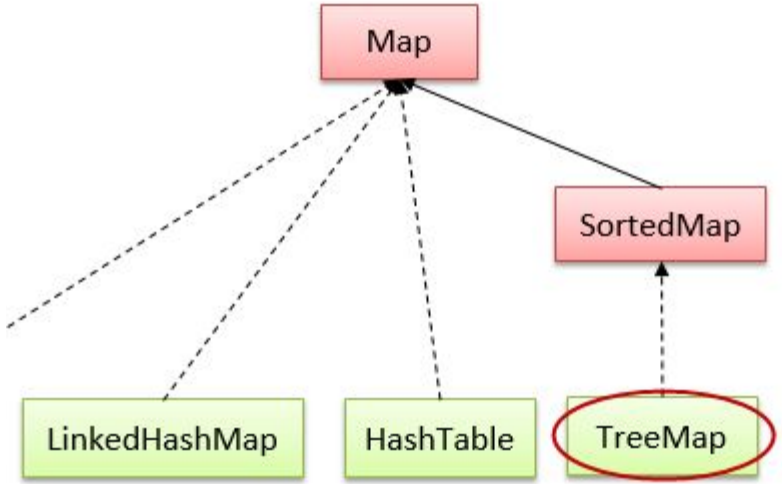


JAVA TREEMAP

JAVA TREEMAP

O(log n)

```java
public class WordCounter {
    Run | Debug
    public static void main(String[] args) {
        // Check if the file name argument is provided
        if (args.length < 1) {
            System.out.println("Usage : java WordCounter <file_name>");
            return;
        }

        String fileName = args[0];
        Map<String, Integer> wordCounts = new TreeMap<>();

        try {
            File file = new File(fileName);
            Scanner scanner = new Scanner(file);

            while (scanner.hasNext()) {
                String word = scanner.next();
                if (word.length() > 0) {
                    int count = wordCounts.getOrDefault(word, 0);
                    wordCounts.put(word, count + 1);
                }
            }
            scanner.close();
        } catch (FileNotFoundException e) {
            System.out.println("File not found: " + fileName);
            return;
        }
    }
}
```



```
bernanda@bernanda-pc:~/word_count/jv_treemap$ ./run.sh 20 text0.1.txt
------------------------
Average running time 88 ms
------------------------
bernanda@bernanda-pc:~/word_count/jv_treemap$ ./run.sh 10 text1.0.txt
------------------------
Average running time 244 ms
------------------------
bernanda@bernanda-pc:~/word_count/jv_treemap$ ./run.sh 10 text10.0.txt
------------------------
Average running time 1433 ms
------------------------
bernanda@bernanda-pc:~/word_count/jv_treemap$ ./run.sh 10 text20.0.txt
------------------------
Average running time 2694 ms
------------------------
bernanda@bernanda-pc:~/word_count/jv_treemap$ ./run.sh 10 text40.0.txt
------------------------
Average running time 5297 ms
------------------------
bernanda@bernanda-pc:~/word_count/jv_treemap$ ./run.sh 10 text60.0.txt
------------------------
Average running time 8013 ms
------------------------
bernanda@bernanda-pc:~/word_count/jv_treemap$ ./run.sh 10 text80.0.txt
------------------------
Average running time 10973 ms
------------------------
bernanda@bernanda-pc:~/word_count/jv_treemap$ ./run.sh 10 text100.0.txt
------------------------
Average running time 13477 ms
------------------------
bernanda@bernanda-pc:~/word_count/jv_treemap$ ./run.sh 10 text200.0.txt
------------------------
Average running time 26765 ms
------------------------
```

Hadoop (uber false)

0.1MB : 7.097sec

```
2023-06-08 21:03:26,975 INFO mapreduce.Job: Job job_1686232850366_0004 running in uber mode : false
2023-06-08 21:03:26,977 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 21:03:30,028 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 21:03:34,057 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:03:34,072 INFO mapreduce.Job: Job job_1686232850366_0004 completed successfully
```

1MB : 7.089sec

```
2023-06-08 21:07:52,823 INFO mapreduce.Job: Job job_1686232850366_0005 running in uber mode : false
2023-06-08 21:07:52,825 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 21:07:55,867 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 21:07:59,897 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:07:59,912 INFO mapreduce.Job: Job job_1686232850366_0005 completed successfully
```

10MB : 9.095sec

```
2023-06-08 21:10:15,726 INFO mapreduce.Job: Job job_1686232850366_0008 running in uber mode : false
2023-06-08 21:10:15,728 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 21:10:20,781 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 21:10:24,806 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:10:24,821 INFO mapreduce.Job: Job job_1686232850366_0008 completed successfully
```

20MB : 10.103sec

```
2023-06-08 21:14:27,121 INFO mapreduce.Job: Job job_1686232850366_0010 running in uber mode : false
2023-06-08 21:14:27,122 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 21:14:33,180 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 21:14:37,209 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:14:37,224 INFO mapreduce.Job: Job job_1686232850366_0010 completed successfully
```

40MB : 13.106sec

```
2023-06-08 21:16:01,307 INFO mapreduce.Job: Job job_1686232850366_0012 running in uber mode : false
2023-06-08 21:16:01,309 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 21:16:10,379 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 21:16:14,403 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:16:14,413 INFO mapreduce.Job: Job job_1686232850366_0012 completed successfully
```

60MB : 16.122sec

```
2023-06-08 21:17:08,481 INFO mapreduce.Job: Job job_1686232850366_0013 running in uber mode : false
2023-06-08 21:17:08,482 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 21:17:19,557 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 21:17:24,590 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:17:24,603 INFO mapreduce.Job: Job job_1686232850366_0013 completed successfully
```

80MB : 18.123sec

```
2023-06-08 21:19:05,625 INFO mapreduce.Job: Job job_1686232850366_0015 running in uber mode : false
2023-06-08 21:19:05,626 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 21:19:19,716 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 21:19:23,737 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:19:23,748 INFO mapreduce.Job: Job job_1686232850366_0015 completed successfully
```

100MB : 21.135sec

```
2023-06-08 21:20:28,614 INFO mapreduce.Job: Job job_1686232850366_0016 running in uber mode : false
2023-06-08 21:20:28,615 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 21:20:43,687 INFO mapreduce.Job:  map 67% reduce 0%
2023-06-08 21:20:45,715 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 21:20:49,736 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:20:49,749 INFO mapreduce.Job: Job job_1686232850366_0016 completed successfully
```

200MB : 36.200sec

```
2023-06-08 21:27:15,239 INFO mapreduce.Job: Job job_1686232850366_0021 running in uber mode : false
2023-06-08 21:27:15,240 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 21:27:30,334 INFO mapreduce.Job:  map 33% reduce 0%
2023-06-08 21:27:46,403 INFO mapreduce.Job:  map 50% reduce 0%
2023-06-08 21:27:48,413 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 21:27:51,428 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:27:51,439 INFO mapreduce.Job: Job job_1686232850366_0021 completed successfully
```

Hadoop (uber true)

```
<property>
    <name>mapreduce.job.ubertask.enable</name>
    <value>true</value>
</property>
```

reduce the overhead of inter-process communication between the AM (AppMaster) and RM (ResourceManager)

0.1MB : 1.043sec

```
2023-06-08 21:34:03,583 INFO mapreduce.Job: Job job_1686232850366_0022 running in uber mode : true
2023-06-08 21:34:03,584 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:34:04,626 INFO mapreduce.Job: Job job_1686232850366_0022 completed successfully
```

1MB : 2.055sec

```
2023-06-08 21:35:40,761 INFO mapreduce.Job: Job job_1686232850366_0024 running in uber mode : true
2023-06-08 21:35:40,762 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 21:35:42,804 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:35:42,816 INFO mapreduce.Job: Job job_1686232850366_0024 completed successfully
```

10MB : 3.057sec

```
2023-06-08 21:37:01,089 INFO mapreduce.Job: Job job_1686232850366_0025 running in uber mode : true
2023-06-08 21:37:01,090 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 21:37:04,131 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:37:04,146 INFO mapreduce.Job: Job job_1686232850366_0025 completed successfully
```

20MB : 5.060sec

```
2023-06-08 21:38:36,274 INFO mapreduce.Job: Job job_1686232850366_0028 running in uber mode : true
2023-06-08 21:38:36,274 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 21:38:39,303 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 21:38:41,319 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:38:41,334 INFO mapreduce.Job: Job job_1686232850366_0028 completed successfully
```

40MB : 7.064sec

```
2023-06-08 21:39:09,248 INFO mapreduce.Job: Job job_1686232850366_0029 running in uber mode : true
2023-06-08 21:39:09,249 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 21:39:15,301 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 21:39:16,305 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:39:16,312 INFO mapreduce.Job: Job job_1686232850366_0029 completed successfully
```

60MB : 10.061sec

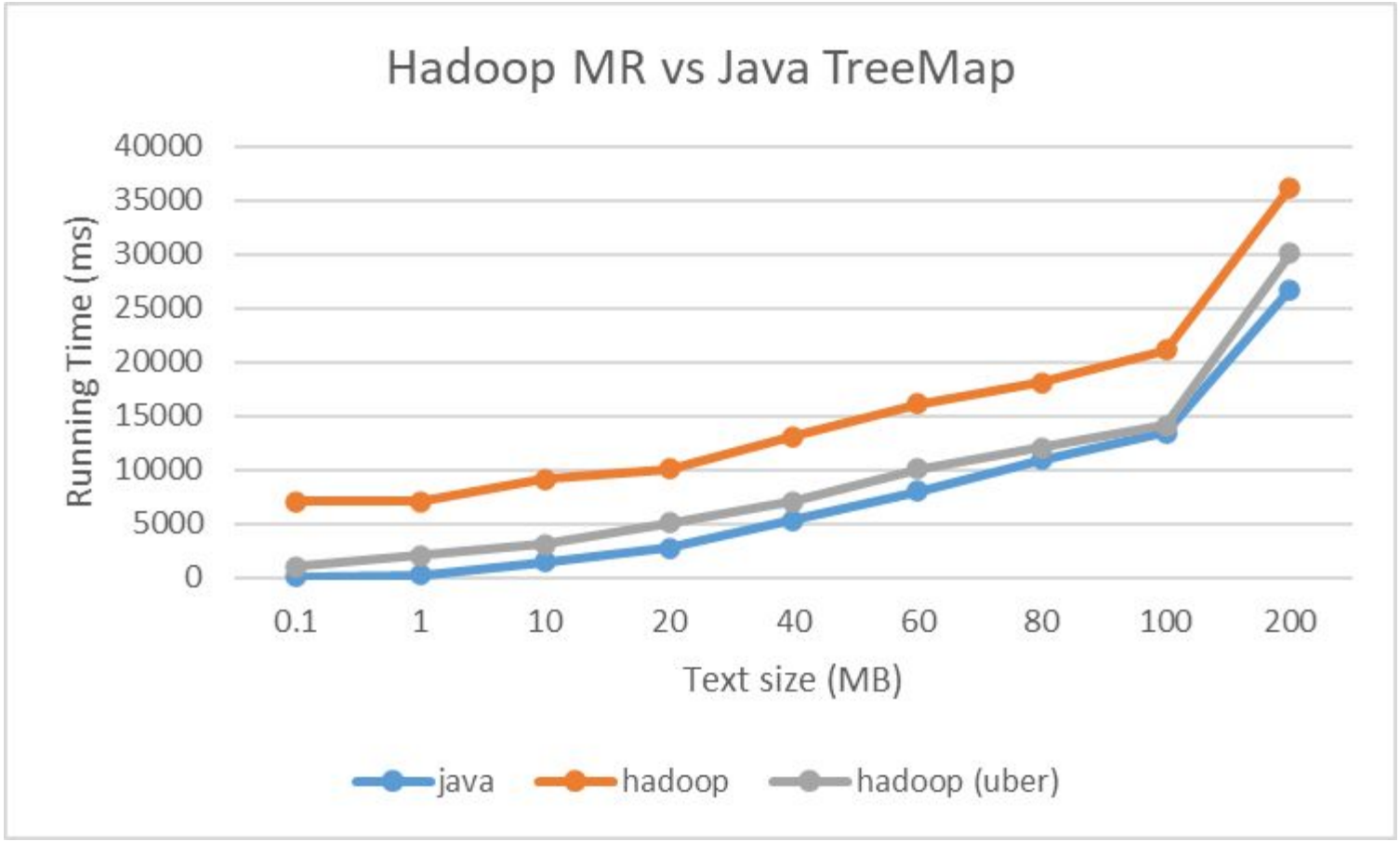
```
2023-06-08 21:41:08,641 INFO mapreduce.Job: Job job_1686232850366_0030 running in uber mode : true
2023-06-08 21:41:08,641 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 21:41:16,686 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 21:41:18,702 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:41:18,717 INFO mapreduce.Job: Job job_1686232850366_0030 completed successfully
```

80MB : 12.082sec

```
2023-06-08 21:42:24,293 INFO mapreduce.Job: Job job_1686232850366_0031 running in uber mode : true
2023-06-08 21:42:24,294 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 21:42:35,353 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 21:42:36,362 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:42:36,375 INFO mapreduce.Job: Job job_1686232850366_0031 completed successfully
```

100MB : 14.095sec

```
2023-06-08 21:43:46,841 INFO mapreduce.Job: Job job_1686232850366_0032 running in uber mode : true
2023-06-08 21:43:46,842 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 21:43:58,895 INFO mapreduce.Job:  map 67% reduce 0%
2023-06-08 21:43:59,919 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 21:44:00,927 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 21:44:00,936 INFO mapreduce.Job: Job job_1686232850366_0032 completed successfully
```


Hadoop MR vs Java TreeMap
Running Time (ms)
40000
35000
30000
25000
20000
15000
10000
5000
0
0.1
1
10
20
40
60
80
100
200
Text size (MB)
java
hadoop
hadoop (uber)

WHY?

- Hadoop has the **overhead** of setting up and managing a cluster.
- **Single node**, as all components run on a single machine, the **parallelism** and data locality advantages of distributed processing are **not fully utilized**.
- Java TreeMap is more **simple** and optimized for this type of problem. (only map and counting)

```
real    0m15.360s
user    0m4.343s
sys     0m0.288s
```

```
real    0m9.531s
user    0m4.761s
sys     0m0.434s
```

```
real    0m18.761s
user    0m5.455s
sys     0m0.529s
```

```
real    0m10.814s
user    0m5.762s
sys     0m0.372s
```

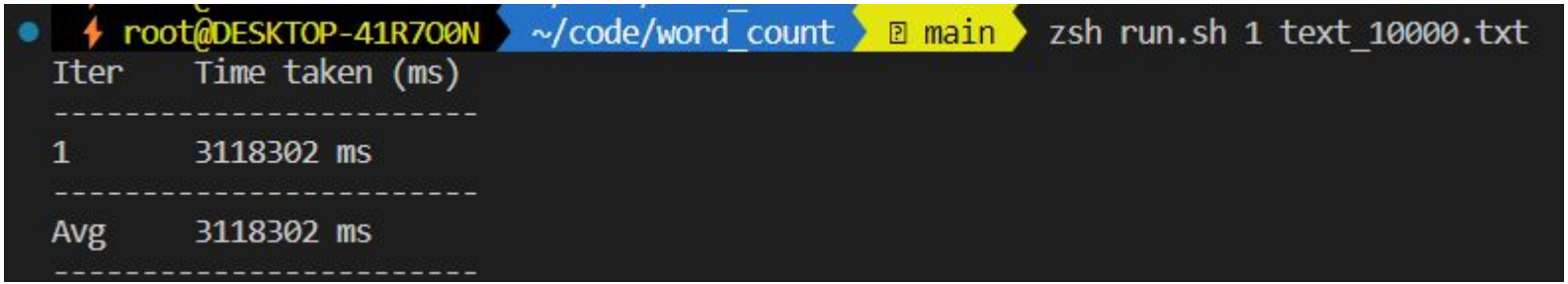
```
real    0m19.726s
user    0m5.045s
sys     0m0.458s
```

```
real    0m16.774s
user    0m4.857s
sys     0m0.205s
```

```
real    0m38.855s
user    0m5.291s
sys     0m0.426s
```

```
real    0m24.200s
user    0m5.641s
sys     0m0.507s
```

```
root@DESKTOP-41R7O0N ~/code/word_count main zsh run.sh 10 text_100.txt
Iter    Time taken (ms)
------------------------
1       14765 ms
2       15527 ms
3       15424 ms
4       16157 ms
5       16126 ms
6       17082 ms
7       17534 ms
8       16297 ms
9       16766 ms
10      17274 ms
------------------------
Avg     16295 ms
------------------------
root@DESKTOP-41R7O0N ~/code/word_count main zsh run.sh 4 text_1000.txt
Iter    Time taken (ms)
------------------------
1       152737 ms
2       156070 ms
3       218215 ms
4       171322 ms
------------------------
Avg     174586 ms
------------------------
```


root@DESKTOP-41R700N ~/code/word_count main zsh run.sh 1 text_10000.txt
Iter Time taken (ms)
1 3118302 ms
Avg 3118302 ms