BIG DATA
DEVELOPMENT

ACADGILD

Session 3: YARN

Assignment 1

Big Data Hadoop and Spark Development

Assignment 1 – You must perform the given tasks.

Table of Content

Big Data Hadoop and Spark Development

Introduction

In this assignment, you need to perform the given tasks.

1. Objective

This assignment will help you to consolidate the concepts learnt in the session 2.

2. Prerequisites:

None

3. Associated Data Files

None

4. Problem Statement

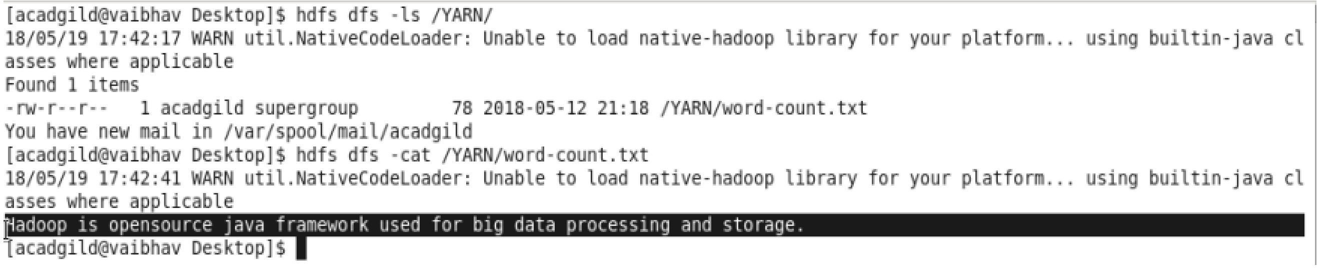
Task 1:

Execute **WordMedian** , **WordMean** , **WordStandardDeviation** programs using hadoop-mapreduce-examples-2.9.0.jar file present in your AcadGild VM.

Command used to execute the jar file in hdfs:
Hadoop jar <jar file name> <class to be executed/the fully qualified name of the package> <Input file path on which the job is executed> <unique output file path>

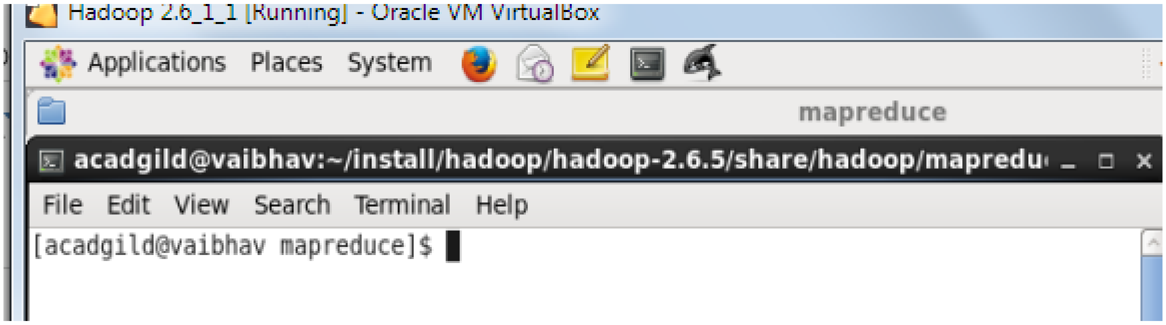
Word Median:

- The text file word-count.txt present in the YARN directiory in HDFS was used to perform Word median programs.
- The content of the file is shown in the following screenshot:



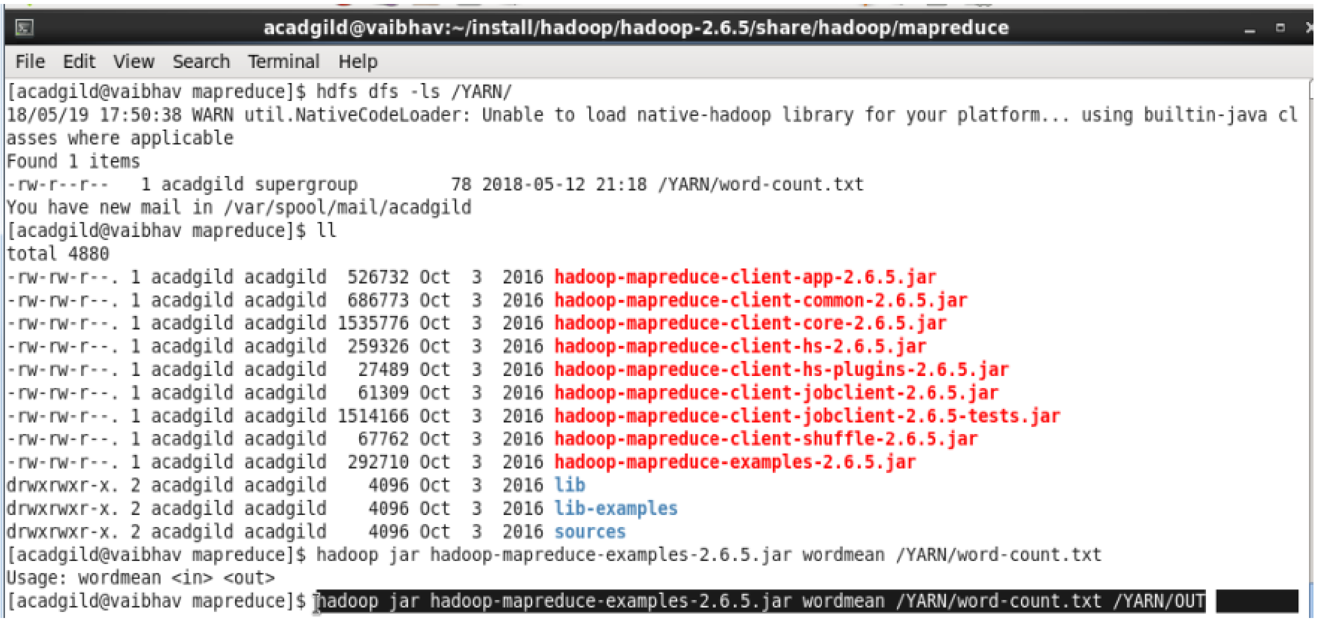
```
[acadgild@vaibhav Desktop]$ hdfs dfs -ls /YARN/
18/05/19 17:42:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 1 items
-rw-r--r--   1 acadgild supergroup         78 2018-05-12 21:18 /YARN/word-count.txt
You have new mail in /var/spool/mail/acadgild
[acadgild@vaibhav Desktop]$ hdfs dfs -cat /YARN/word-count.txt
18/05/19 17:42:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Hadoop is opensource java framework used for big data processing and storage.
[acadgild@vaibhav Desktop]$
```

- Now to execute the program that calculates the word median of the content of the file, we moved to the directory where the .jar with program was present file was present


Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
Applications Places System
mapreduce
acadgild@vaibhav:~/install/hadoop/hadoop-2.6.5/share/hadoop/mapredu
File Edit View Search Terminal Help

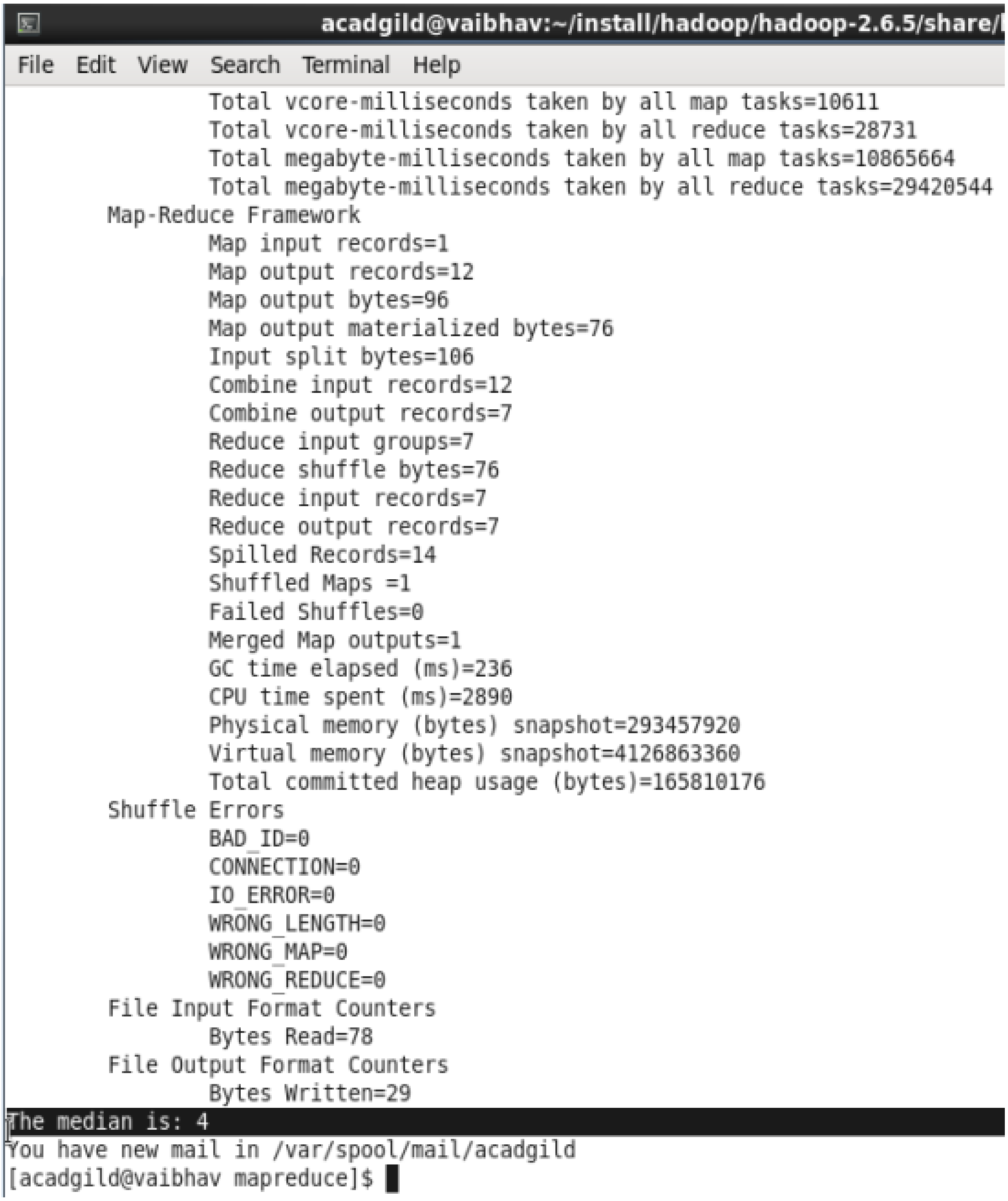
```
[acadgild@vaibhav mapreduce]$
```

- And then the program was executed using the following command:


acadgild@vaibhav:~/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce
File Edit View Search Terminal Help

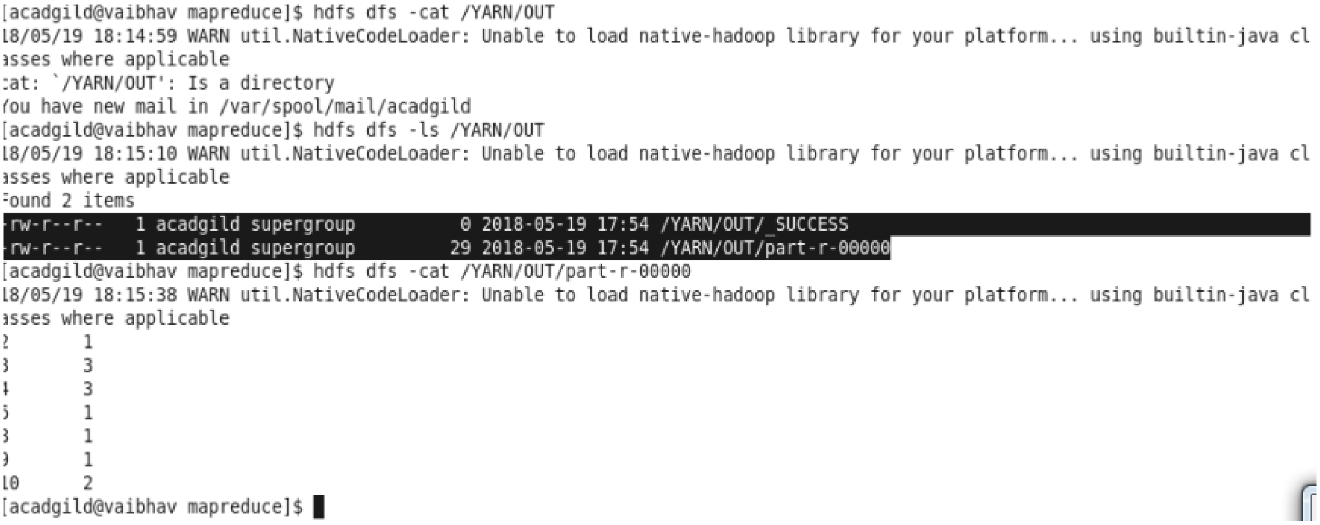
```
[acadgild@vaibhav mapreduce]$ hdfs dfs -ls /YARN/
18/05/19 17:50:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 1 items
-rw-r--r--   1 acadgild supergroup         78 2018-05-12 21:18 /YARN/word-count.txt
You have new mail in /var/spool/mail/acadgild
[acadgild@vaibhav mapreduce]$ ll
total 4880
-rw-rw-r--. 1 acadgild acadgild  526732 Oct  3  2016 hadoop-mapreduce-client-app-2.6.5.jar
-rw-rw-r--. 1 acadgild acadgild  686773 Oct  3  2016 hadoop-mapreduce-client-common-2.6.5.jar
-rw-rw-r--. 1 acadgild acadgild 1535776 Oct  3  2016 hadoop-mapreduce-client-core-2.6.5.jar
-rw-rw-r--. 1 acadgild acadgild  259326 Oct  3  2016 hadoop-mapreduce-client-hs-2.6.5.jar
-rw-rw-r--. 1 acadgild acadgild   27489 Oct  3  2016 hadoop-mapreduce-client-hs-plugins-2.6.5.jar
-rw-rw-r--. 1 acadgild acadgild   61309 Oct  3  2016 hadoop-mapreduce-client-jobclient-2.6.5.jar
-rw-rw-r--. 1 acadgild acadgild 1514166 Oct  3  2016 hadoop-mapreduce-client-jobclient-2.6.5-tests.jar
-rw-rw-r--. 1 acadgild acadgild   67762 Oct  3  2016 hadoop-mapreduce-client-shuffle-2.6.5.jar
-rw-rw-r--. 1 acadgild acadgild  292710 Oct  3  2016 hadoop-mapreduce-examples-2.6.5.jar
drwxrwxr-x. 2 acadgild acadgild    4096 Oct  3  2016 lib
drwxrwxr-x. 2 acadgild acadgild    4096 Oct  3  2016 lib-examples
drwxrwxr-x. 2 acadgild acadgild    4096 Oct  3  2016 sources
[acadgild@vaibhav mapreduce]$ hadoop jar hadoop-mapreduce-examples-2.6.5.jar wordmean /YARN/word-count.txt
Usage: wordmean <in> <out>
[acadgild@vaibhav mapreduce]$ hadoop jar hadoop-mapreduce-examples-2.6.5.jar wordmean /YARN/word-count.txt /YARN/OUT
```

- The following was the output of the job : The median is 4 for the content present in the text file.


acadgild@vaibhav:~/install/hadoop/hadoop-2.6.5/share/
File Edit View Search Terminal Help

```
                Total vcore-milliseconds taken by all map tasks=10611
                Total vcore-milliseconds taken by all reduce tasks=28731
                Total megabyte-milliseconds taken by all map tasks=10865664
                Total megabyte-milliseconds taken by all reduce tasks=29420544
        Map-Reduce Framework
                Map input records=1
                Map output records=12
                Map output bytes=96
                Map output materialized bytes=76
                Input split bytes=106
                Combine input records=12
                Combine output records=7
                Reduce input groups=7
                Reduce shuffle bytes=76
                Reduce input records=7
                Reduce output records=7
                Spilled Records=14
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=236
                CPU time spent (ms)=2890
                Physical memory (bytes) snapshot=293457920
                Virtual memory (bytes) snapshot=4126863360
                Total committed heap usage (bytes)=165810176
        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=78
        File Output Format Counters
                Bytes Written=29
The median is: 4
You have new mail in /var/spool/mail/acadgild
[acadgild@vaibhav mapreduce]$
```

- The following are the files present in the output folder :

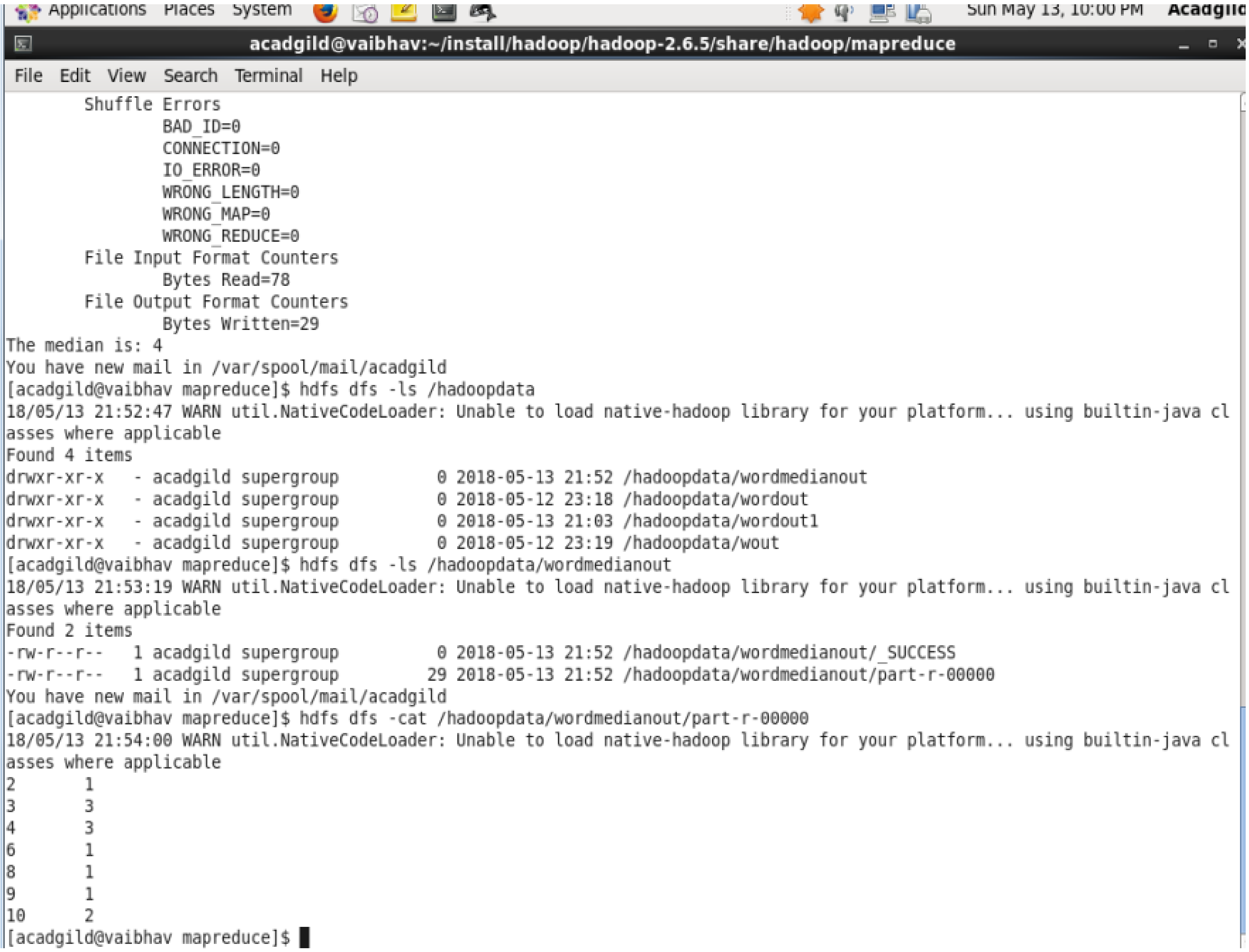


```
[acadgild@vaibhav mapreduce]$ hdfs dfs -cat /YARN/OUT
18/05/19 18:14:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
cat: `/YARN/OUT': Is a directory
You have new mail in /var/spool/mail/acadgild
[acadgild@vaibhav mapreduce]$ hdfs dfs -ls /YARN/OUT
18/05/19 18:15:10 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-05-19 17:54 /YARN/OUT/_SUCCESS
-rw-r--r--   1 acadgild supergroup         29 2018-05-19 17:54 /YARN/OUT/part-r-00000
[acadgild@vaibhav mapreduce]$ hdfs dfs -cat /YARN/OUT/part-r-00000
18/05/19 18:15:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
2       1
3       3
4       3
5       1
8       1
9       1
10      2
[acadgild@vaibhav mapreduce]$ 
```

WordMean

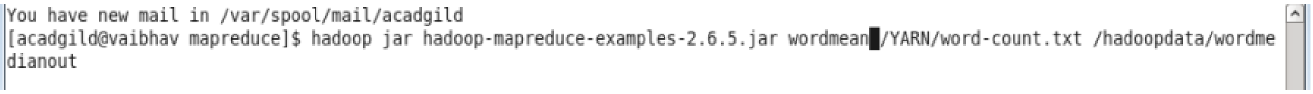
Step 1: Using Word Mean class for word-count.txt file :

- The following are the outputs for the mean of the text item present in the file :


Applications Places System
Sun May 13, 10:00 PM Acadgild
acadgild@vaibhav:~/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce
File Edit View Search Terminal Help

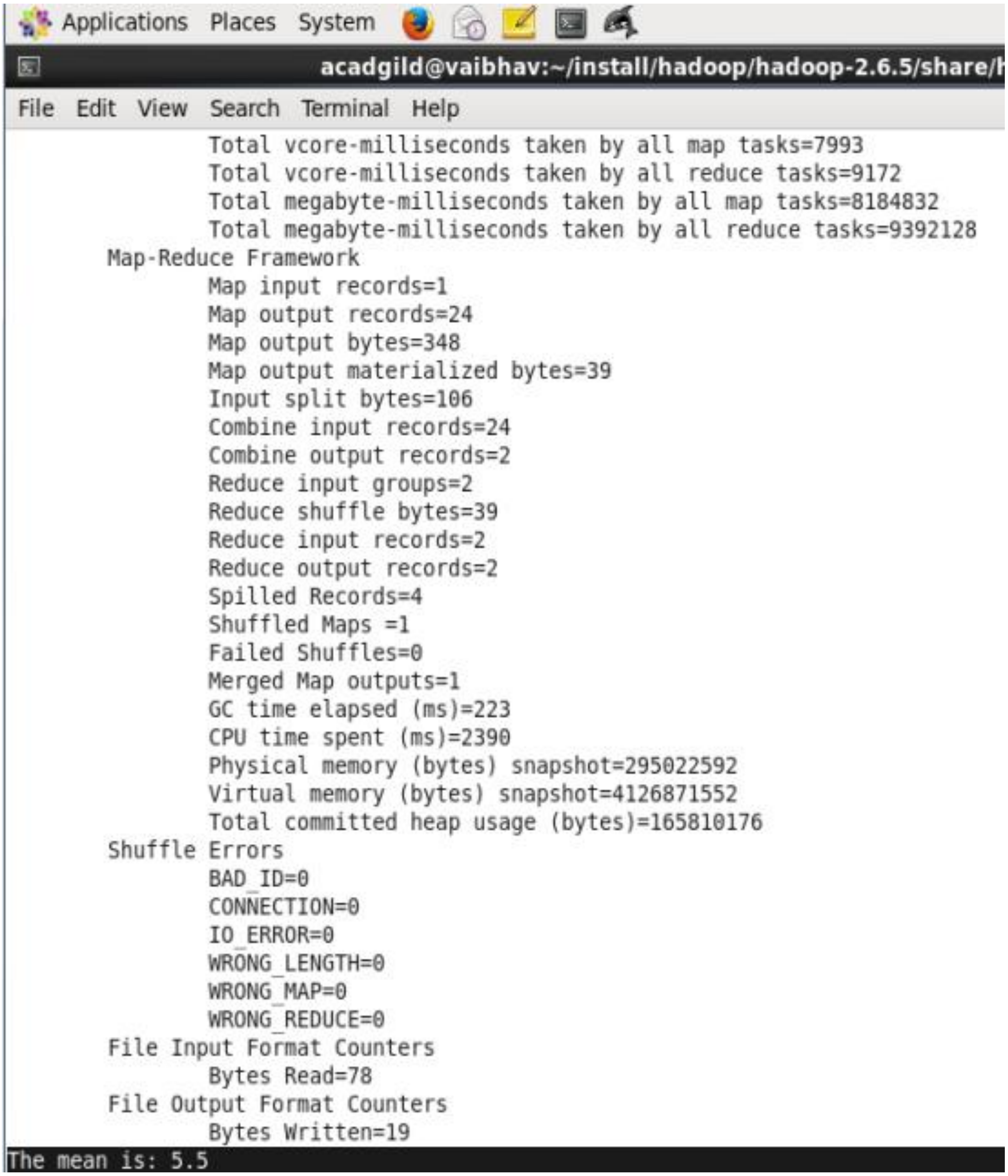
```
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=78
        File Output Format Counters
                Bytes Written=29
The median is: 4
You have new mail in /var/spool/mail/acadgild
[acadgild@vaibhav mapreduce]$ hdfs dfs -ls /hadoopdata
18/05/13 21:52:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 4 items
drwxr-xr-x   - acadgild supergroup          0 2018-05-13 21:52 /hadoopdata/wordmedianout
drwxr-xr-x   - acadgild supergroup          0 2018-05-12 23:18 /hadoopdata/wordout
drwxr-xr-x   - acadgild supergroup          0 2018-05-13 21:03 /hadoopdata/wordout1
drwxr-xr-x   - acadgild supergroup          0 2018-05-12 23:19 /hadoopdata/wout
[acadgild@vaibhav mapreduce]$ hdfs dfs -ls /hadoopdata/wordmedianout
18/05/13 21:53:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-05-13 21:52 /hadoopdata/wordmedianout/_SUCCESS
-rw-r--r--   1 acadgild supergroup         29 2018-05-13 21:52 /hadoopdata/wordmedianout/part-r-00000
You have new mail in /var/spool/mail/acadgild
[acadgild@vaibhav mapreduce]$ hdfs dfs -cat /hadoopdata/wordmedianout/part-r-00000
18/05/13 21:54:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
2       1
3       3
4       3
6       1
8       1
9       1
10      2
[acadgild@vaibhav mapreduce]$ █
```

Word Mean: the following command is used to initiate the process:



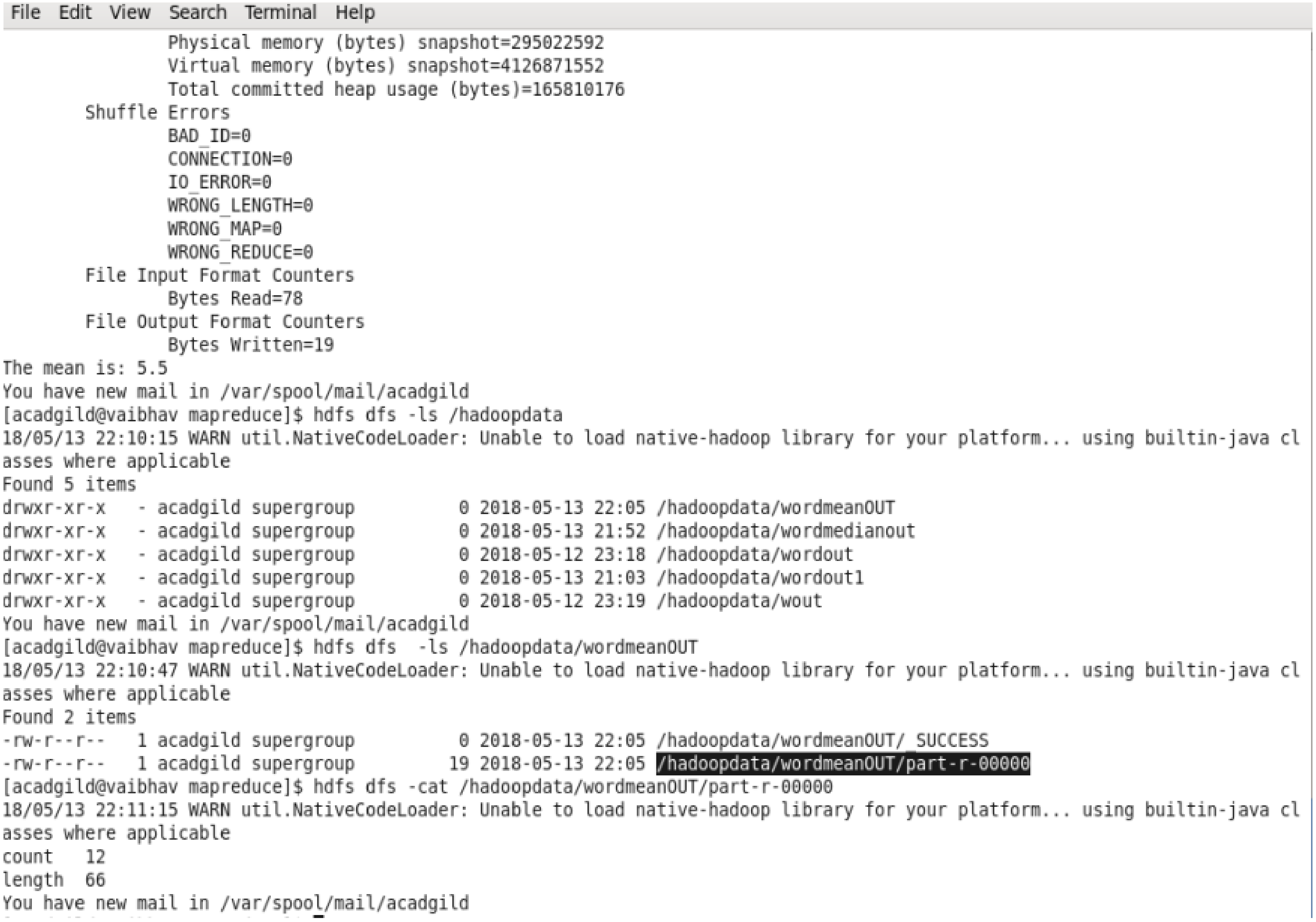
```
You have new mail in /var/spool/mail/acadgild
[acadgild@vaibhav mapreduce]$ hadoop jar hadoop-mapreduce-examples-2.6.5.jar wordmean█/YARN/word-count.txt /hadoopdata/wordme
dianout
```

The following results were observed:


Applications Places System
acadgild@vaibhav:~/install/hadoop/hadoop-2.6.5/share/h
File Edit View Search Terminal Help

```
                Total vcore-milliseconds taken by all map tasks=7993
                Total vcore-milliseconds taken by all reduce tasks=9172
                Total megabyte-milliseconds taken by all map tasks=8184832
                Total megabyte-milliseconds taken by all reduce tasks=9392128
        Map-Reduce Framework
                Map input records=1
                Map output records=24
                Map output bytes=348
                Map output materialized bytes=39
                Input split bytes=106
                Combine input records=24
                Combine output records=2
                Reduce input groups=2
                Reduce shuffle bytes=39
                Reduce input records=2
                Reduce output records=2
                Spilled Records=4
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=223
                CPU time spent (ms)=2390
                Physical memory (bytes) snapshot=295022592
                Virtual memory (bytes) snapshot=4126871552
                Total committed heap usage (bytes)=165810176
        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=78
        File Output Format Counters
                Bytes Written=19
The mean is: 5.5
```

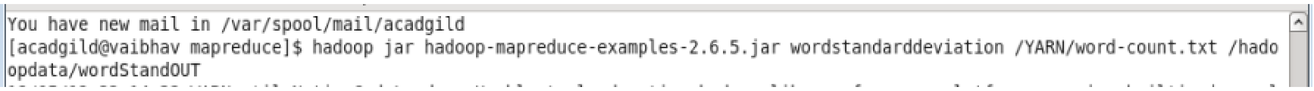
And there were two file generated after the job :



```
                Physical memory (bytes) snapshot=295022592
                Virtual memory (bytes) snapshot=4126871552
                Total committed heap usage (bytes)=165810176
        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=78
        File Output Format Counters
                Bytes Written=19
The mean is: 5.5
You have new mail in /var/spool/mail/acadgild
[acadgild@vaibhav mapreduce]$ hdfs dfs -ls /hadoopdata
18/05/13 22:10:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 5 items
drwxr-xr-x   - acadgild supergroup          0 2018-05-13 22:05 /hadoopdata/wordmeanOUT
drwxr-xr-x   - acadgild supergroup          0 2018-05-13 21:52 /hadoopdata/wordmedianout
drwxr-xr-x   - acadgild supergroup          0 2018-05-12 23:18 /hadoopdata/wordout
drwxr-xr-x   - acadgild supergroup          0 2018-05-13 21:03 /hadoopdata/wordout1
drwxr-xr-x   - acadgild supergroup          0 2018-05-12 23:19 /hadoopdata/wout
You have new mail in /var/spool/mail/acadgild
[acadgild@vaibhav mapreduce]$ hdfs dfs  -ls /hadoopdata/wordmeanOUT
18/05/13 22:10:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-05-13 22:05 /hadoopdata/wordmeanOUT/_SUCCESS
-rw-r--r--   1 acadgild supergroup         19 2018-05-13 22:05 /hadoopdata/wordmeanOUT/part-r-00000
[acadgild@vaibhav mapreduce]$ hdfs dfs -cat /hadoopdata/wordmeanOUT/part-r-00000
18/05/13 22:11:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
count   12
length  66
You have new mail in /var/spool/mail/acadgild
```

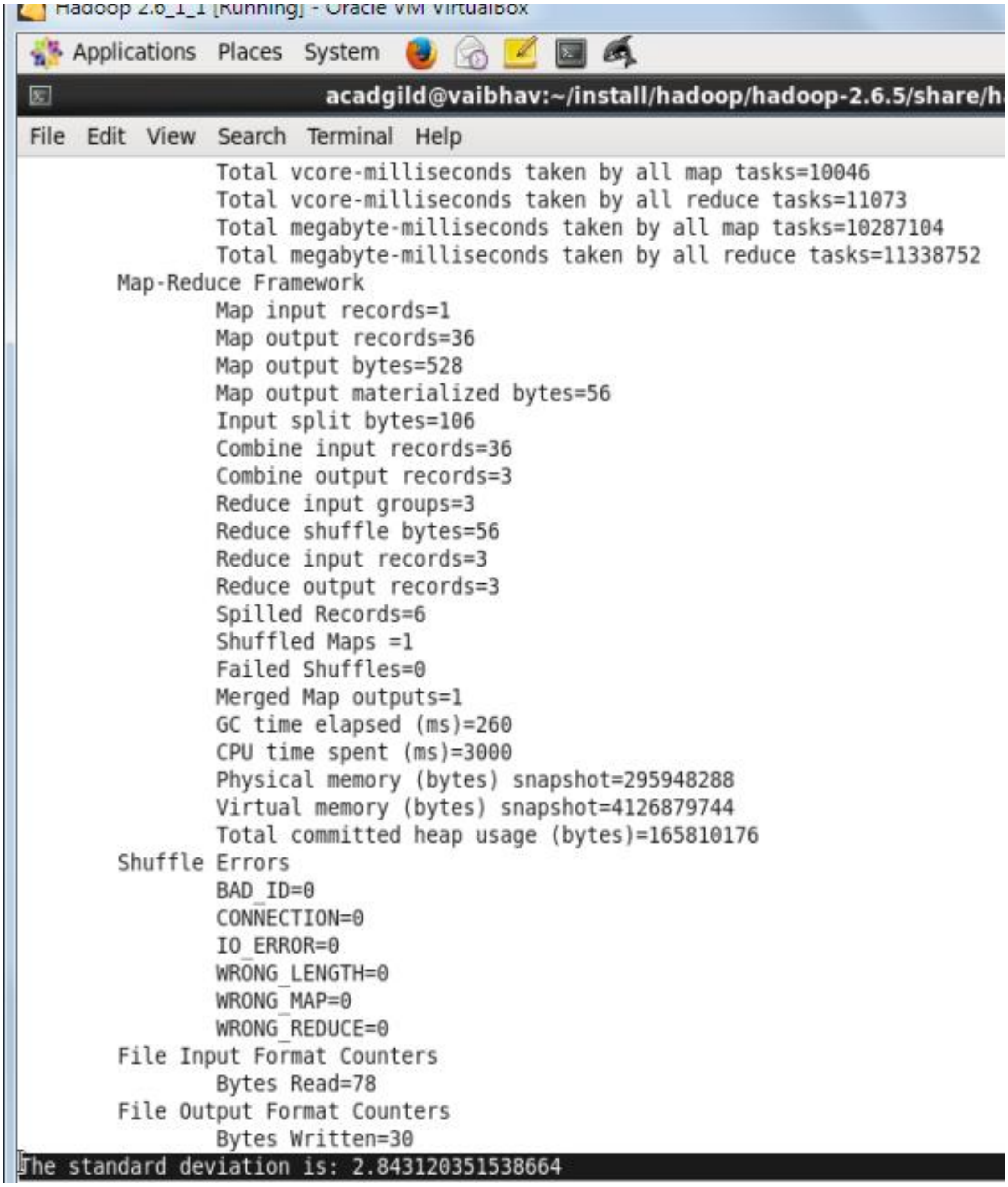
WordStandardDeviation:

Command used:



```
You have new mail in /var/spool/mail/acadgild
[acadgild@vaibhav mapreduce]$ hadoop jar hadoop-mapreduce-examples-2.6.5.jar wordstandarddeviation /YARN/word-count.txt /hado
opdata/wordStandOUT
```

The following is the standard deviation of the content present in the file:


Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
Applications Places System
acadgild@vaibhav:~/install/hadoop/hadoop-2.6.5/share/h
File Edit View Search Terminal Help

```
                Total vcore-milliseconds taken by all map tasks=10046
                Total vcore-milliseconds taken by all reduce tasks=11073
                Total megabyte-milliseconds taken by all map tasks=10287104
                Total megabyte-milliseconds taken by all reduce tasks=11338752
        Map-Reduce Framework
                Map input records=1
                Map output records=36
                Map output bytes=528
                Map output materialized bytes=56
                Input split bytes=106
                Combine input records=36
                Combine output records=3
                Reduce input groups=3
                Reduce shuffle bytes=56
                Reduce input records=3
                Reduce output records=3
                Spilled Records=6
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=260
                CPU time spent (ms)=3000
                Physical memory (bytes) snapshot=295948288
                Virtual memory (bytes) snapshot=4126879744
                Total committed heap usage (bytes)=165810176
        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=78
        File Output Format Counters
                Bytes Written=30
The standard deviation is: 2.843120351538664
```

5. Expected Output

Solution report with commands, explanation to commands and screenshot for output.

Report shall be   in PDF format. Submitted in GitHub.

6. Approximate Time to Complete Task

Big Data Hadoop and Spark Development

3hours