

BIG DATA HADOOP AND SPARK DEVELOPMENT

ASSIGNMENT – 3

Table of Contents:

1. Introduction	1
2. Objective	1
3. Problem statement	1
4. Expected Output	
• Task 1 – Executing Median program	4
• Task 2 – Executing Mean program	9
• Task 3 – Executing Standard deviation program	13

BIG DATA HADOOP AND SPARK DEVELOPMENT

1. Introduction

In this assignment, the given task is performed and Output of the task is performed and Screenshots are attached.

2. Objective

This assignment consolidates the deeper understanding of the Session – 3 Introduction to YARN (Yet Another Resources Negotiator) and High level YARN components.

3. Problem Statement

- Task 1 - Execute WordMedian program,
- Task 2 -Execute WordMean program,
- Task 3 -Execute WordStandardDeviation programs using

[hadoop-mapreduce-examples-2.9.0.jar](#) file present in your AcadGild VM.

4. Expected Output

To perform the given tasks, first we need to copy a text file(file327.txt) with the size of 312Mb from local to Acadgild VM.

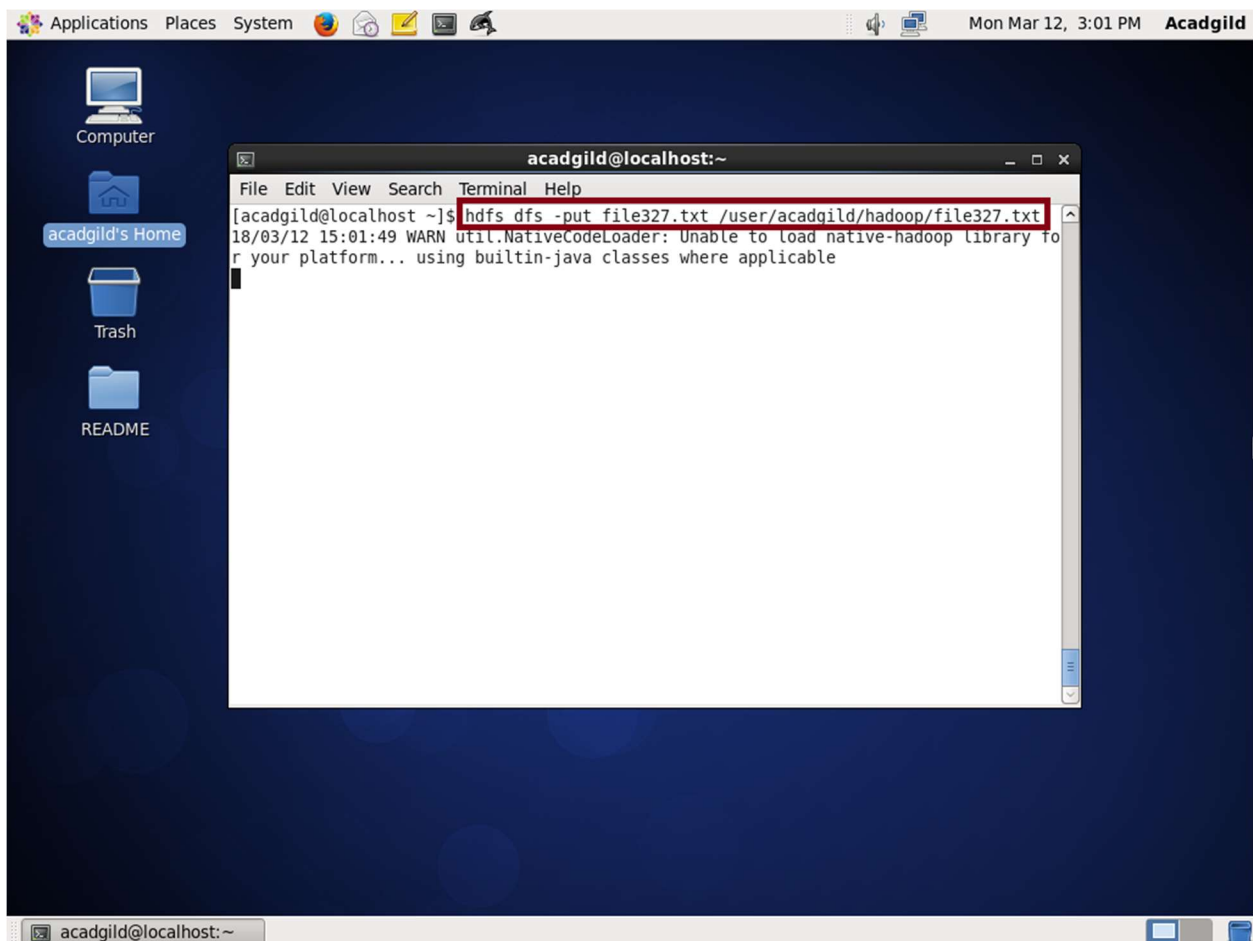
Preparing to perform tasks:

By copying a text file (file327.txt) with the size of 312Mb from local to Acadgild VM.

hdfs dfs -put file327.txt /user/acadgild/Hadoop/file327.txt

by using **-put** command the **file327.txt** is copied from local to **Acadgild VM**

The following screenshot show the process of copying **file327.txt** from local to **Acadgild VM**.



hadoop [--config conf dir]

jar <jar> run a jar file

by using the above syntax, we can run the jar file

- Task 1 – Executing Word Median program

Word Median:

A Map/Reduce program that counts the median length of the words in the input files.

`hadoop <hadoop jar file path> <program name> <path of the file name> <directory name where the output can be stored>`

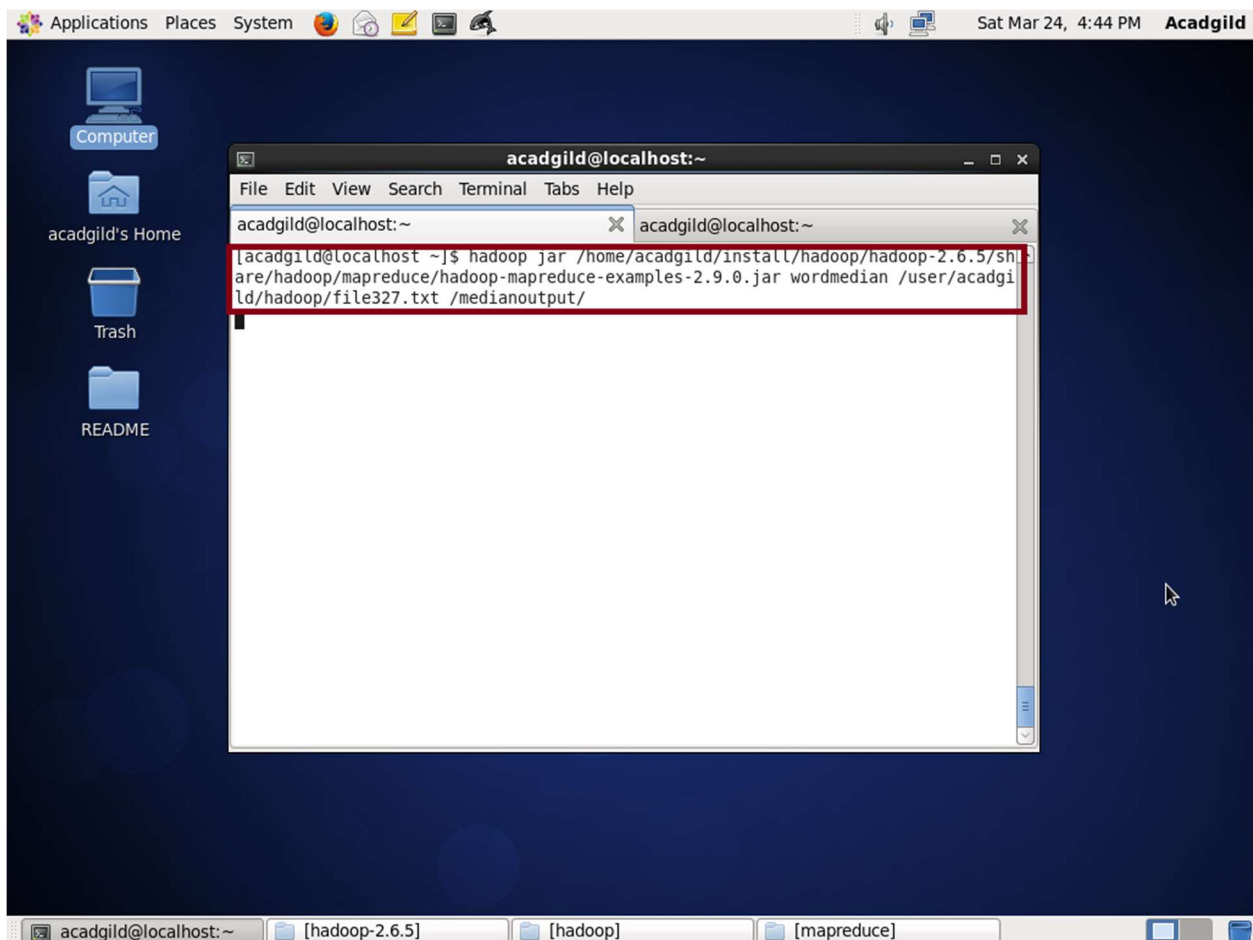
the above is the syntax for the jar file to run and output save in the directory.

To run wordmedian program

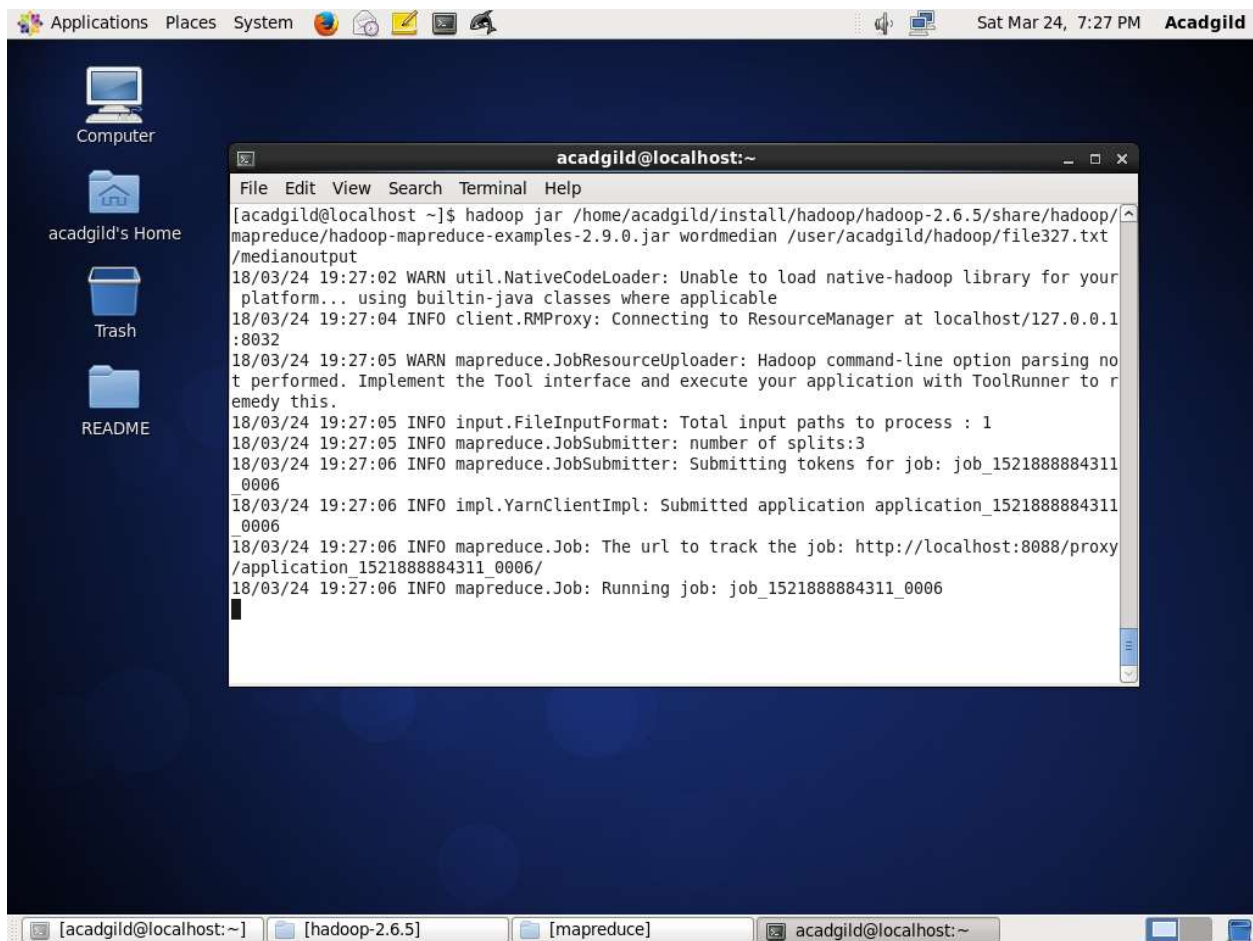
- jar file path is `/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.0.jar`
- program name is `wordmedian`
- File path is `/user/acadgild/hadoop/file327.txt`

The command which is used here is

`hadoop jar /home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.0.jar wordmedian /user/acadgild/hadoop/file327.txt /medianoutput`



Word median program is executed and the Map reduce process is performed.



The screenshot shows a Linux desktop with a dark blue background. On the left, there are icons for 'Computer', 'acagdild's Home', 'Trash', and 'README'. The top panel displays 'Applications', 'Places', 'System', and system status 'Sat Mar 24, 7:27 PM' and 'Acadgild'. A terminal window titled 'acadgild@localhost:~' is open, showing the command to run a Hadoop MapReduce job and its output logs. The logs indicate the job is running successfully.

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
[acadgild@localhost ~]$ hadoop jar /home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/  
mapreduce/hadoop-mapreduce-examples-2.9.0.jar wordmedian /user/acadgild/hadoop/file327.txt  
/medianoutput  
18/03/24 19:27:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your  
platform... using builtin-java classes where applicable  
18/03/24 19:27:04 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1  
:8032  
18/03/24 19:27:05 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing no  
t performed. Implement the Tool interface and execute your application with ToolRunner to r  
emedy this.  
18/03/24 19:27:05 INFO input.FileInputFormat: Total input paths to process : 1  
18/03/24 19:27:05 INFO mapreduce.JobSubmitter: number of splits:3  
18/03/24 19:27:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1521888884311  
_0006  
18/03/24 19:27:06 INFO impl.YarnClientImpl: Submitted application application_1521888884311  
_0006  
18/03/24 19:27:06 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy  
/application_1521888884311_0006/  
18/03/24 19:27:06 INFO mapreduce.Job: Running job: job_1521888884311_0006
```

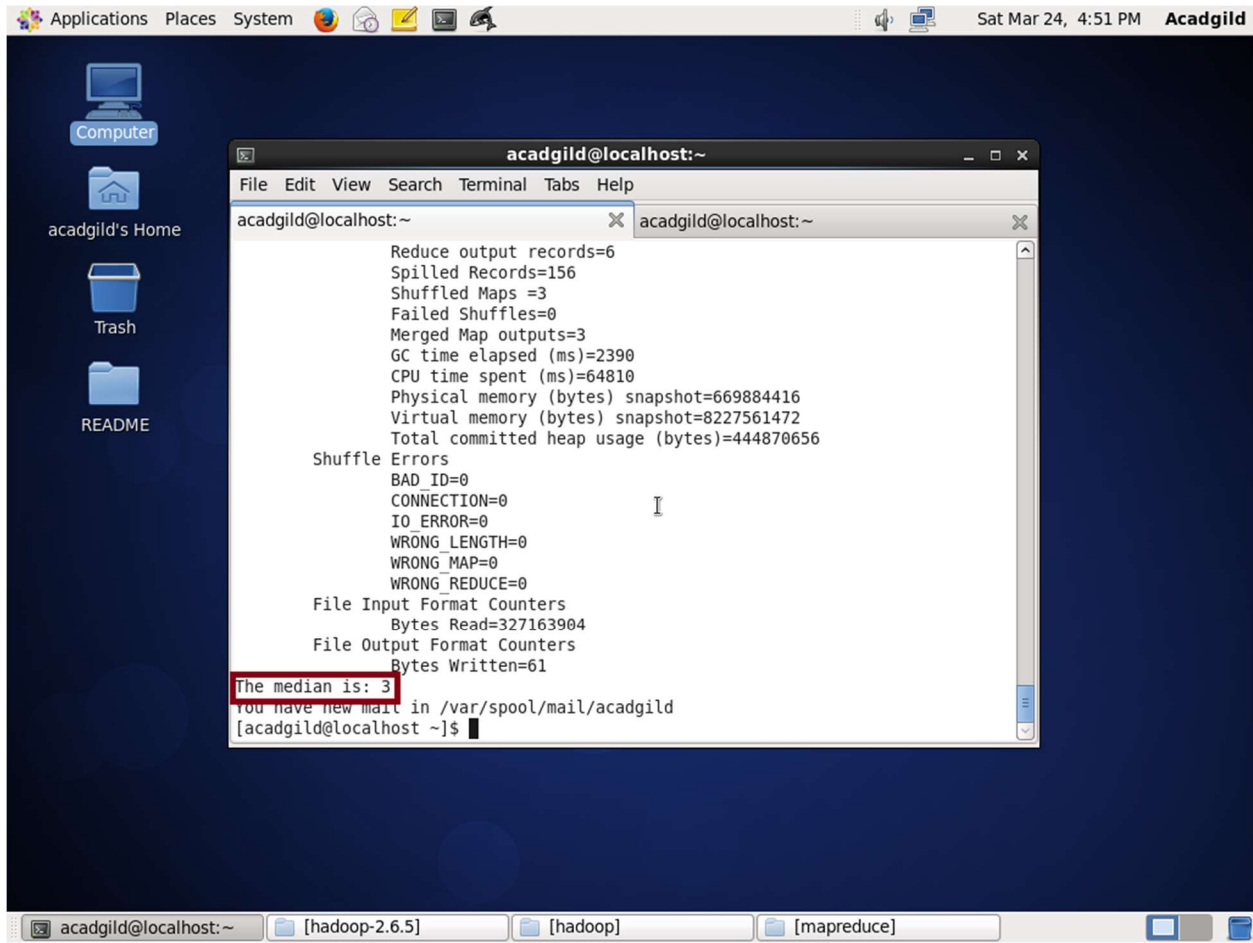
As in individual series (if number of observation is very low) first one must arrange all the observations in order. Then count(n) is the total number of observation in given data.

If n is odd then Median (M) = value of $((n + 1)/2)$ th item term.

If n is even then Median (M) = value of $[(n/2)$ th item term + $(n/2 + 1)$ th item term]/2

The Median is calculated as 3.

And the output is saved in **medianoutput** directory.

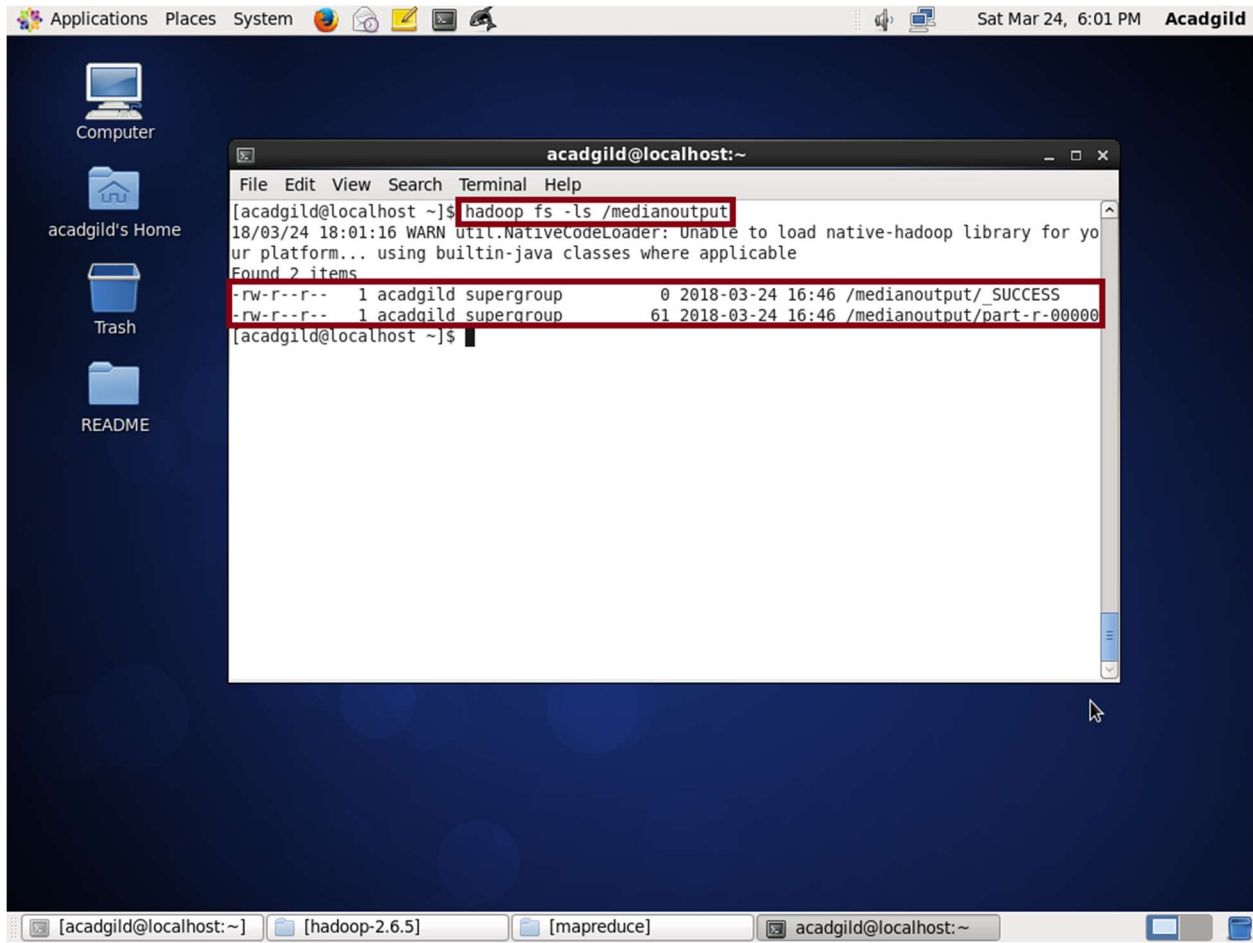


The screenshot shows a Linux desktop with a dark blue background. On the left sidebar, there are icons for 'Computer', 'acagdild's Home', 'Trash', and 'README'. The top panel displays 'Applications', 'Places', 'System', and system status 'Sat Mar 24, 4:51 PM' and 'Acadgild'. A terminal window titled 'acadgild@localhost:~' is open, showing the output of a Hadoop MapReduce job. The output includes statistics like 'Reduce output records=6', 'Spilled Records=156', and 'Shuffled Maps =3'. It also lists 'Shuffle Errors' which are all zero. At the bottom of the terminal output, 'The median is: 3' is highlighted with a red box. Below this, a message says 'You have new mail in /var/spool/mail/acadgild'. The terminal prompt is '[acadgild@localhost ~]\$'.

```
acadgild@localhost:~  
File Edit View Search Terminal Tabs Help  
acadgild@localhost:~  
    Reduce output records=6  
    Spilled Records=156  
    Shuffled Maps =3  
    Failed Shuffles=0  
    Merged Map outputs=3  
    GC time elapsed (ms)=2390  
    CPU time spent (ms)=64810  
    Physical memory (bytes) snapshot=669884416  
    Virtual memory (bytes) snapshot=8227561472  
    Total committed heap usage (bytes)=444870656  
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=327163904  
File Output Format Counters  
    Bytes Written=61  
The median is: 3  
You have new mail in /var/spool/mail/acadgild  
[acadgild@localhost ~]$
```

The output is saved in **medianoutput** directory, which is listed by following command,

hadoop fs -ls /medianoutput



The screenshot shows a Linux desktop environment with a dark blue background. On the left side, there are icons for 'Computer', 'acagild's Home', 'Trash', and 'README'. The top panel displays 'Applications', 'Places', 'System', and the date 'Sat Mar 24, 6:01 PM' along with the username 'Acagild'. A terminal window titled 'acagild@localhost:~' is open in the center. The terminal shows the command 'hadoop fs -ls /medianoutput' being executed. The output of the command is displayed below the command line, showing two items in the directory. The output is as follows:

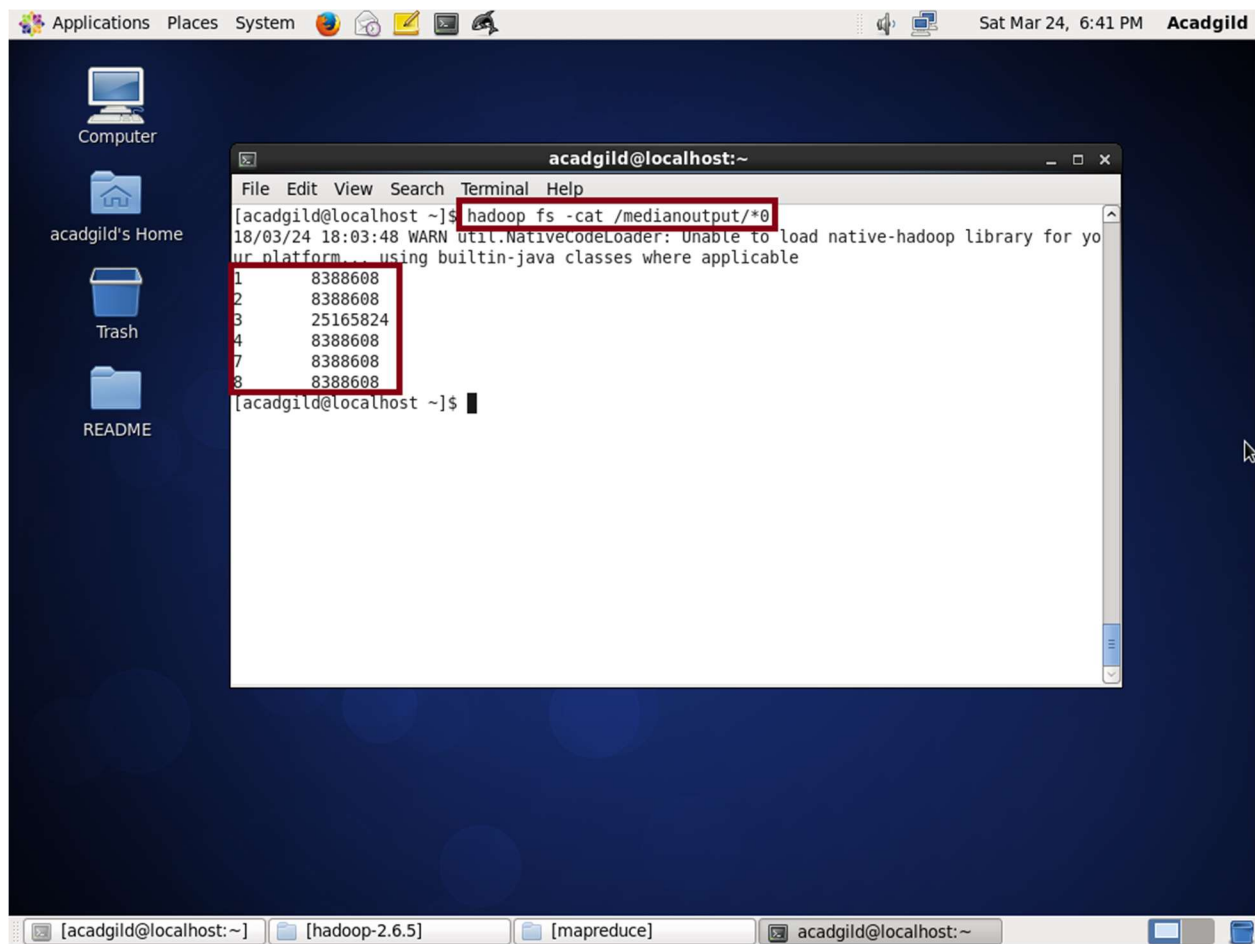
```
acagild@localhost:~$ hadoop fs -ls /medianoutput
18/03/24 18:01:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for yo
ur platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r-- 1 acagild supergroup 0 2018-03-24 16:46 /medianoutput/_SUCCESS
-rw-r--r-- 1 acagild supergroup 61 2018-03-24 16:46 /medianoutput/part-r-00000
acagild@localhost:~$
```

The data is viewed by using

```
hadoop fs -cat /medianoutput/part-r-00000
```

or

```
hadoop fs -cat /medianoutput/*0
```



The screenshot shows a Linux desktop environment with a dark blue background. On the left side, there are icons for 'Computer', 'acadgild's Home', 'Trash', and 'README'. The top of the window displays a menu bar with 'Applications', 'Places', and 'System', along with system status icons and the date 'Sat Mar 24, 6:41 PM' and the username 'Acadgild'. A terminal window titled 'acadgild@localhost:~' is open in the center. The terminal shows the command 'hadoop fs -cat /medianoutput/*0' being executed. The output consists of a warning message followed by a list of numbers: 1 8388608, 2 8388608, 3 25165824, 4 8388608, 7 8388608, and 8 8388608. The terminal window has a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. The bottom of the screen shows a taskbar with several open windows: '[acadgild@localhost: ~]', '[hadoop-2.6.5]', '[mapreduce]', and 'acadgild@localhost: ~'.

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
[acadgild@localhost ~]$ hadoop fs -cat /medianoutput/*0  
18/03/24 18:03:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
1      8388608  
2      8388608  
3     25165824  
4      8388608  
7      8388608  
8      8388608  
[acadgild@localhost ~]$
```


- Task 2 – Executing Mean Program

Word mean:

A Map/ Reduce program that counts the average length of the words in the input files.

hadoop <hadoop jar file path> <program name> <path of the file name> <directory name where the output can be stored>

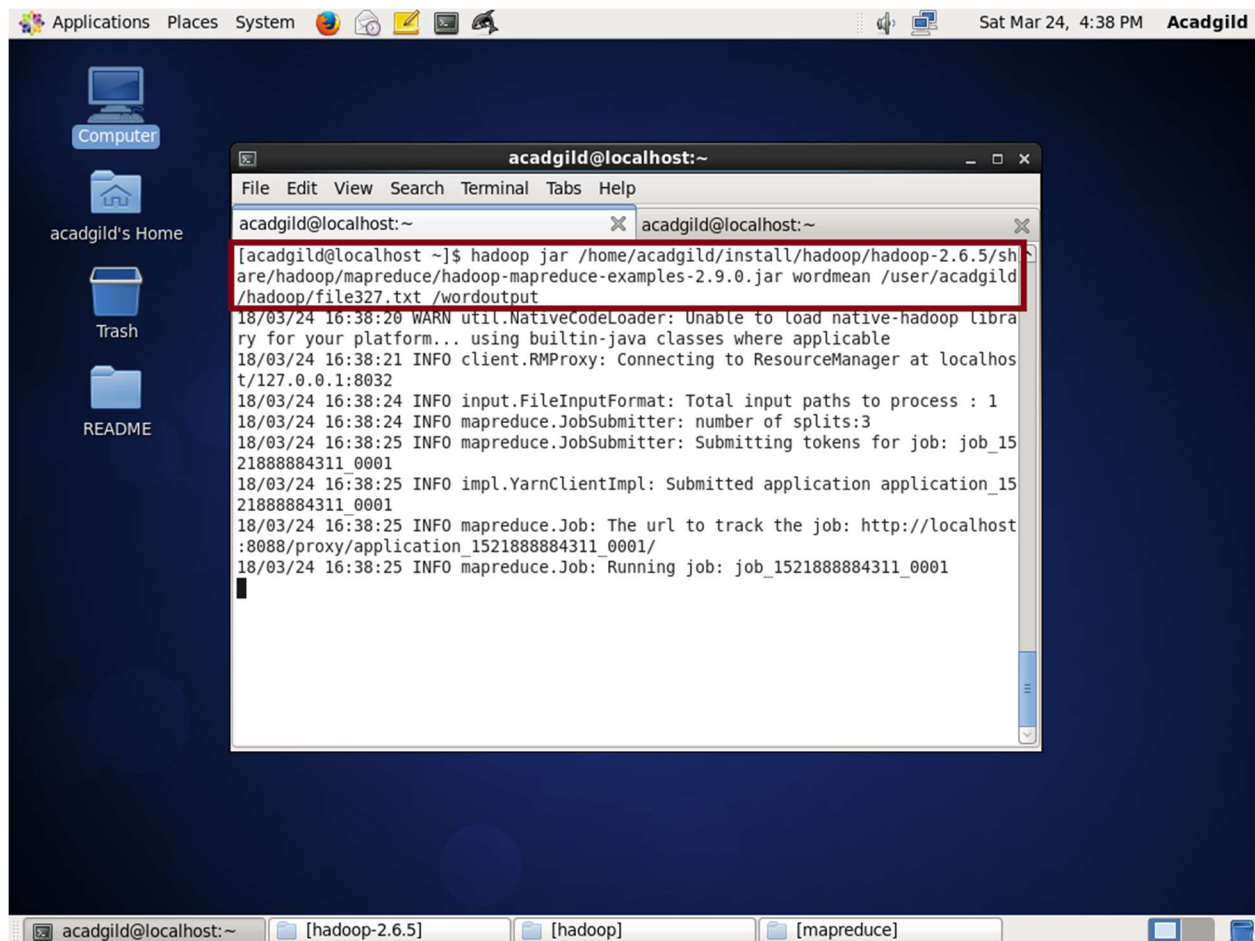
the above is the syntax for the jar file to run and output save in the directory.

To run wordmean program

- jar file path is [/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.0.jar](#)
- program name is [wordmean](#)
- File path is [/user/acadgild/hadoop/file327.txt](#)
- Output directory is [/wordoutput](#)

The following command is used to execute the wordmean program and save the output in the wordoutput directory.

`hadoop jar /home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.0.jar wordmean /user/acadgild/hadoop/file327.txt /wordoutput`



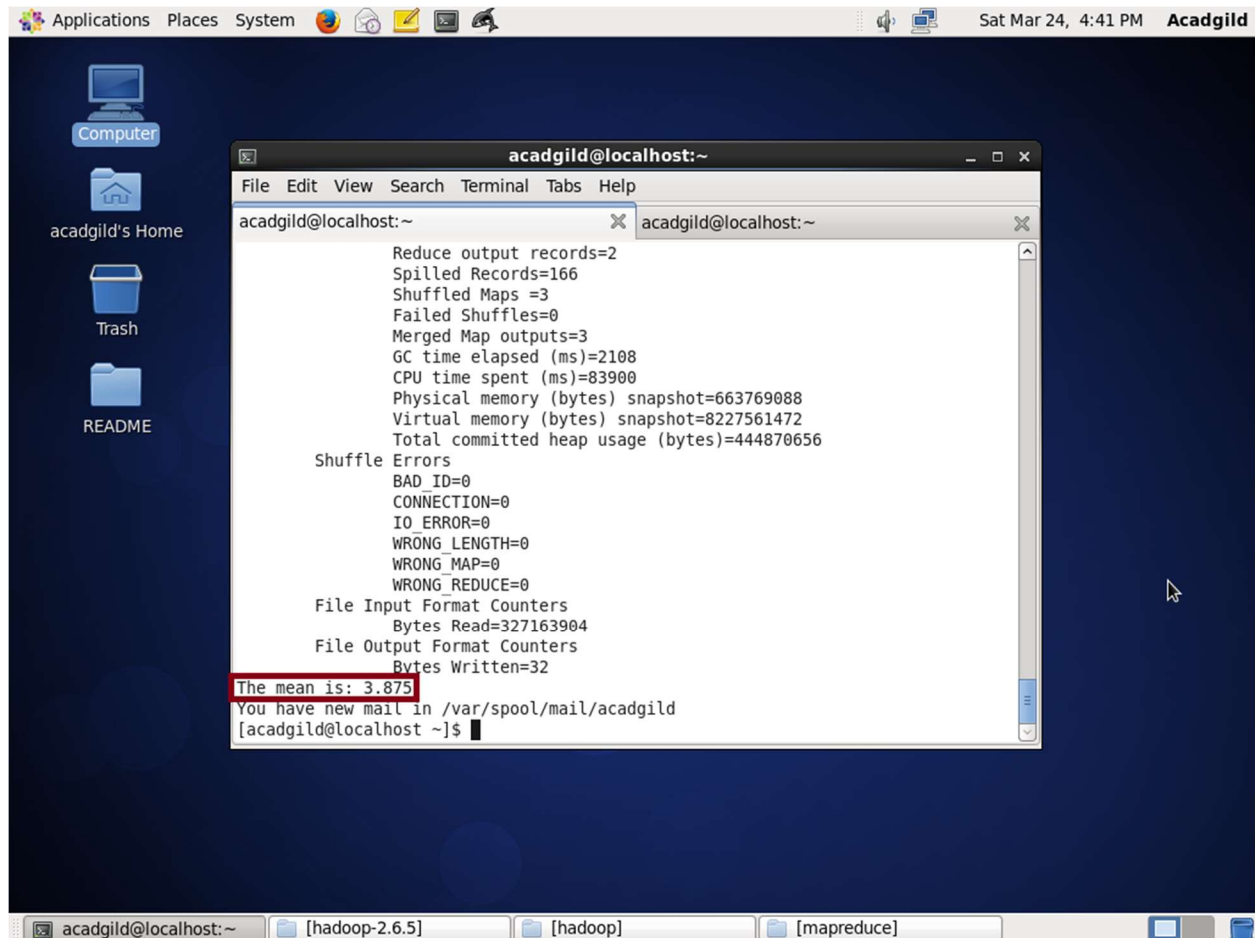
The screenshot shows a Linux desktop with a dark blue background. On the left sidebar, there are icons for 'Computer', 'acadgild's Home', 'Trash', and 'README'. The top panel displays 'Applications', 'Places', 'System', and the date 'Sat Mar 24, 4:38 PM' next to the username 'Acadgild'. A terminal window titled 'acadgild@localhost:~' is open in the center. The terminal shows the command: `hadoop jar /home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.0.jar wordmean /user/acadgild/hadoop/file327.txt /wordoutput`. The output of the command is as follows:

```
18/03/24 16:38:20 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/03/24 16:38:21 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/03/24 16:38:24 INFO input.FileInputFormat: Total input paths to process : 1
18/03/24 16:38:24 INFO mapreduce.JobSubmitter: number of splits:3
18/03/24 16:38:25 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1521888884311_0001
18/03/24 16:38:25 INFO impl.YarnClientImpl: Submitted application application_1521888884311_0001
18/03/24 16:38:25 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1521888884311_0001/
18/03/24 16:38:25 INFO mapreduce.Job: Running job: job_1521888884311_0001
```

Word Mean is calculated by

The word mean is the average of the number of input words in the text file calculated by a "central" value of a set of number of input words in the text file.

Map reduce process is performed and Mean is displayed.



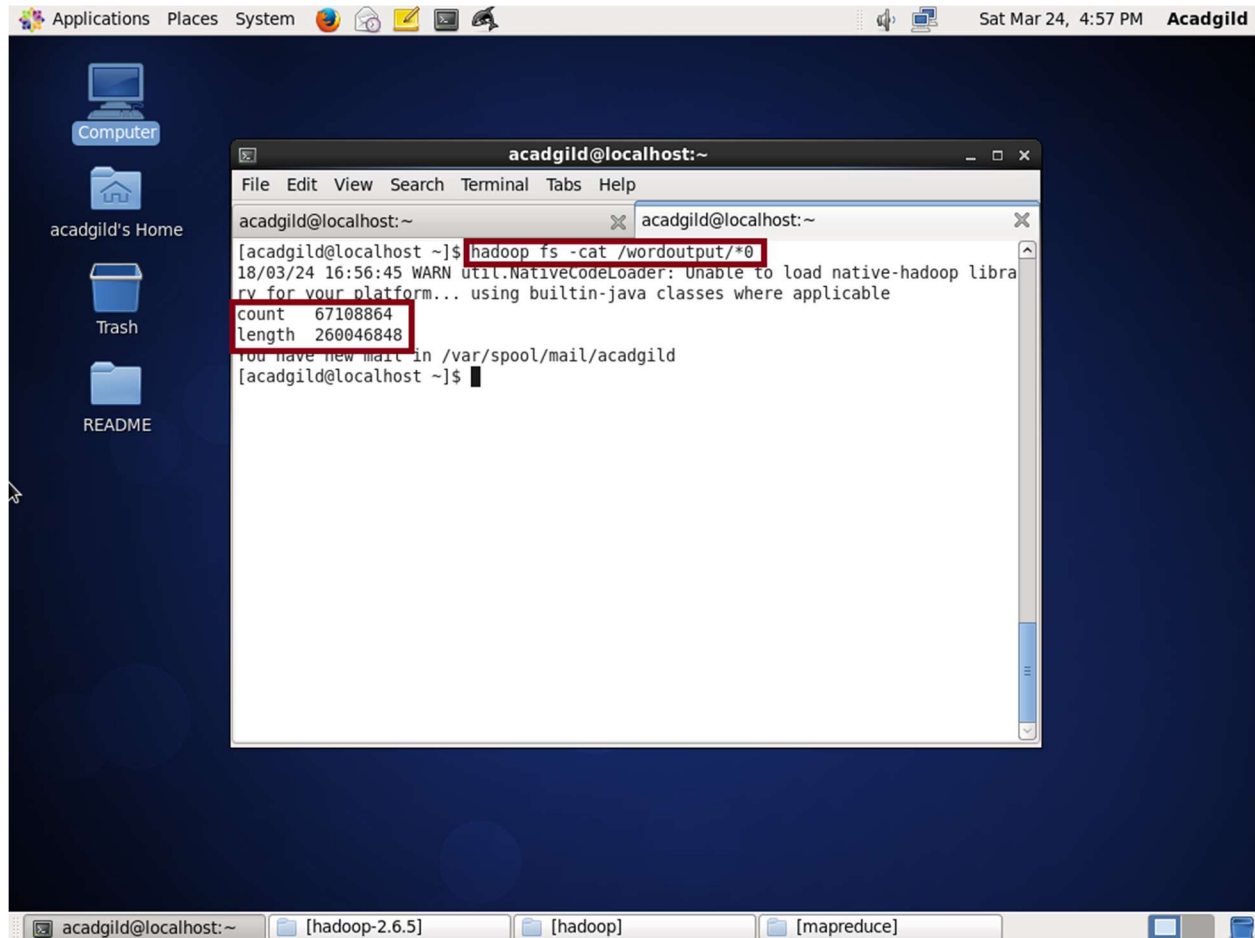
The screenshot shows a Linux desktop with a dark blue background. On the left sidebar, there are icons for 'Computer', 'acagild's Home', 'Trash', and 'README'. The top panel displays 'Applications', 'Places', 'System', and system status 'Sat Mar 24, 4:41 PM Acadgild'. A terminal window titled 'acadgild@localhost:~' is open, showing the output of a Hadoop MapReduce job. The output includes statistics for reduce output, GC time, CPU time, memory usage, and shuffle errors. The final line of the output, 'The mean is: 3.875', is highlighted with a red box. Below this, a system message indicates new mail in the user's inbox. The terminal window has a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', 'Tabs', and 'Help'. The bottom panel shows the terminal window and three open folders: '[hadoop-2.6.5]', '[hadoop]', and '[mapreduce]'.

```
acadgild@localhost:~  
File Edit View Search Terminal Tabs Help  
acadgild@localhost:~  
Reduce output records=2  
Spilled Records=166  
Shuffled Maps =3  
Failed Shuffles=0  
Merged Map outputs=3  
GC time elapsed (ms)=2108  
CPU time spent (ms)=83900  
Physical memory (bytes) snapshot=663769088  
Virtual memory (bytes) snapshot=8227561472  
Total committed heap usage (bytes)=444870656  
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=327163904  
File Output Format Counters  
Bytes Written=32  
The mean is: 3.875  
You have new mail in /var/spool/mail/acagild  
[acadgild@localhost ~]$
```

Mean is **3.875**

The output of the `wordmean` program is saved in the `wordoutput` directory and it can be displayed by following command,

```
hadoop fs -cat /wordoutput/*0
```



The screenshot shows a Linux desktop environment with a dark blue background. On the left side, there are icons for 'Computer', 'acagdild's Home', 'Trash', and 'README'. The top panel displays 'Applications', 'Places', 'System', and the date 'Sat Mar 24, 4:57 PM' along with the username 'Acadgild'. A terminal window titled 'acadgild@localhost:~' is open in the center. The terminal shows the command `hadoop fs -cat /wordoutput/*0` being executed. The output of the command is displayed below the command line, showing the count and length of the data. The output is as follows:

```
acadgild@localhost:~$ hadoop fs -cat /wordoutput/*0
18/03/24 16:56:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
count 67108864
length 260046848
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

- Task 3- Executing word standard deviation program

Word Standard Deviation:

A Map/Reduce program that counts the standard deviation of the length of the words in the input files.

hadoop <hadoop jar file path> <program name> <path of the file name> <directory name where the output can be stored>

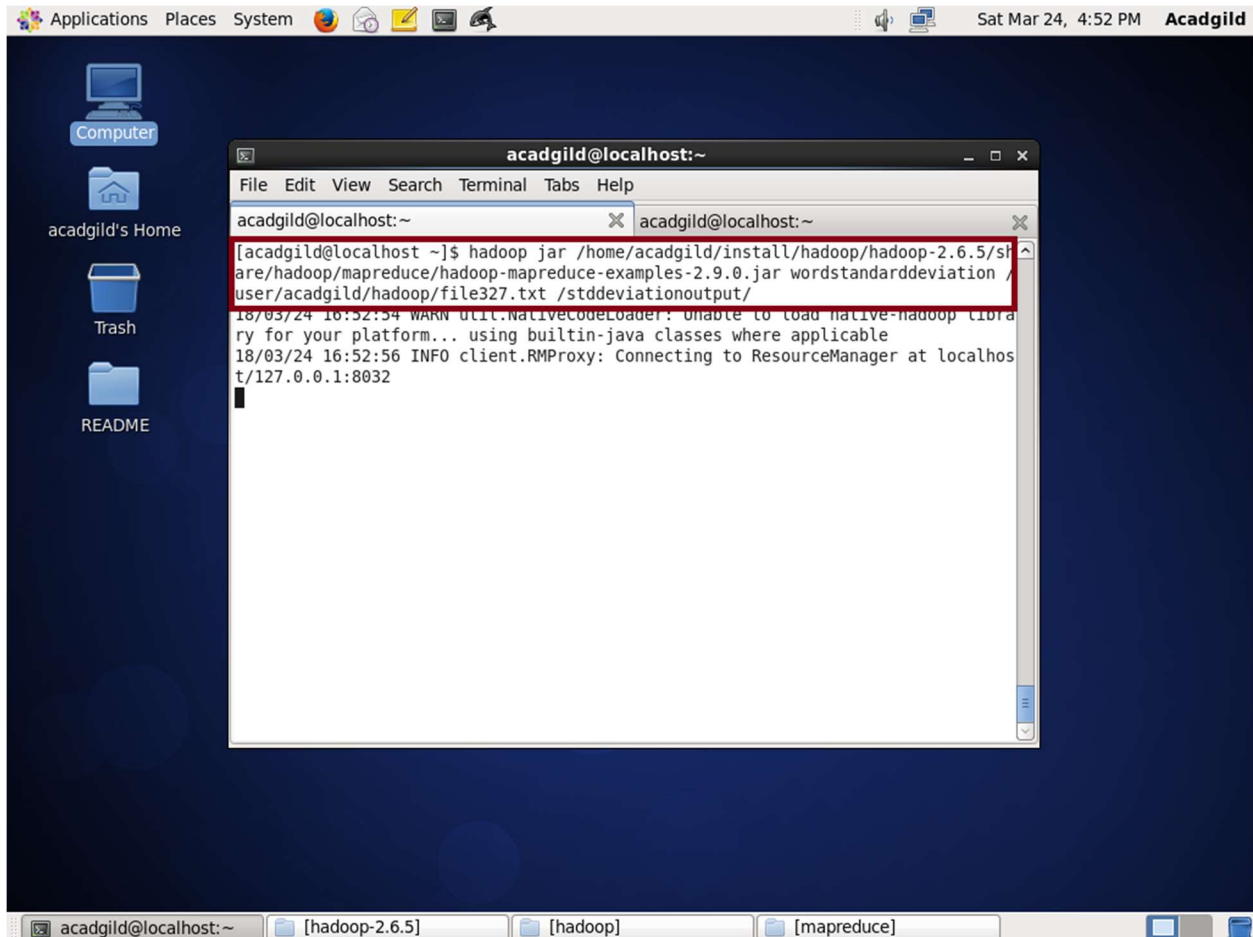
the above is the syntax for the jar file to run and output save in the directory.

To run wordstandarddeviation program

- jar file path is [/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.0.jar](#)
 - program name is [wordstandarddeviation](#)
 - File path is [/user/acadgild/hadoop/file327.txt](#)
- Output directory is [/stddeviationoutput](#)

The following command is used to execute the wordstandarddeviation program and save the output in the stddeviationoutput directory.

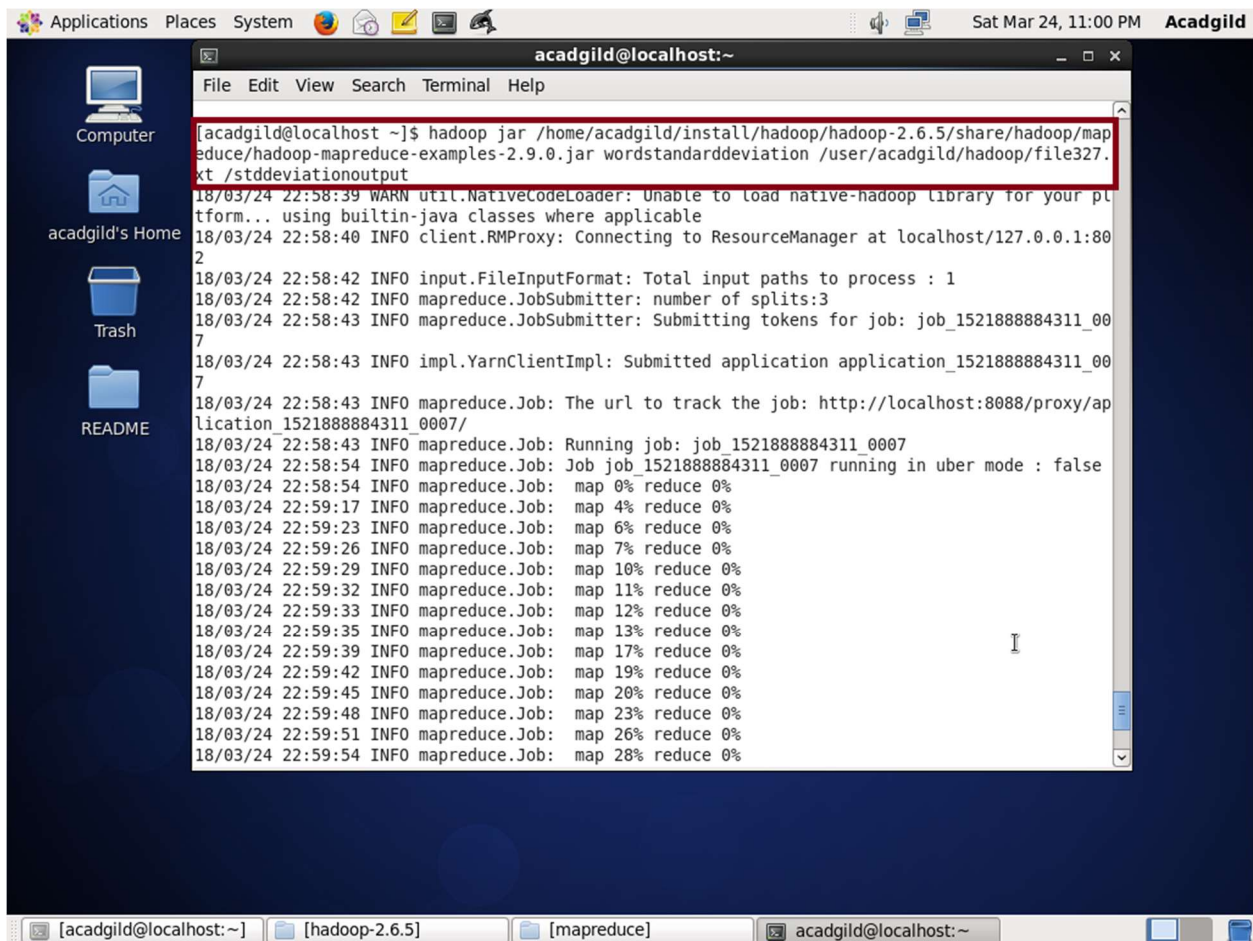
```
hadoop jar /home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.0.jar wordstandarddeviation /user/acadgild/hadoop/file327.txt /stddeviationoutput
```



The screenshot shows a Linux desktop environment with a dark blue background. On the left side, there is a sidebar with icons for 'Computer', 'acagild's Home', 'Trash', and 'README'. The top of the window features a menu bar with 'Applications', 'Places', and 'System', along with system status icons and the date 'Sat Mar 24, 4:52 PM' and the username 'Acadgild'. A terminal window titled 'acadgild@localhost:~' is open in the center. The terminal has a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', 'Tabs', and 'Help'. It shows a command prompt 'acadgild@localhost:~' followed by the command: `hadoop jar /home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.0.jar wordstandarddeviation /user/acadgild/hadoop/file327.txt /stddeviationoutput/`. The command is highlighted with a red box. Below the command, the terminal displays the following output: `18/03/24 16:52:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable` and `18/03/24 16:52:56 INFO client.RMPProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032`. The terminal window has a scrollbar on the right side. At the bottom of the desktop, there is a taskbar with icons for the terminal, '[hadoop-2.6.5]', '[hadoop]', and '[mapreduce]'. The terminal icon is highlighted.

The **Standard Deviation** is a measure of how spread out numbers are. It is the square root of the Variance, and the Variance is the average of the squared differences from the Mean.

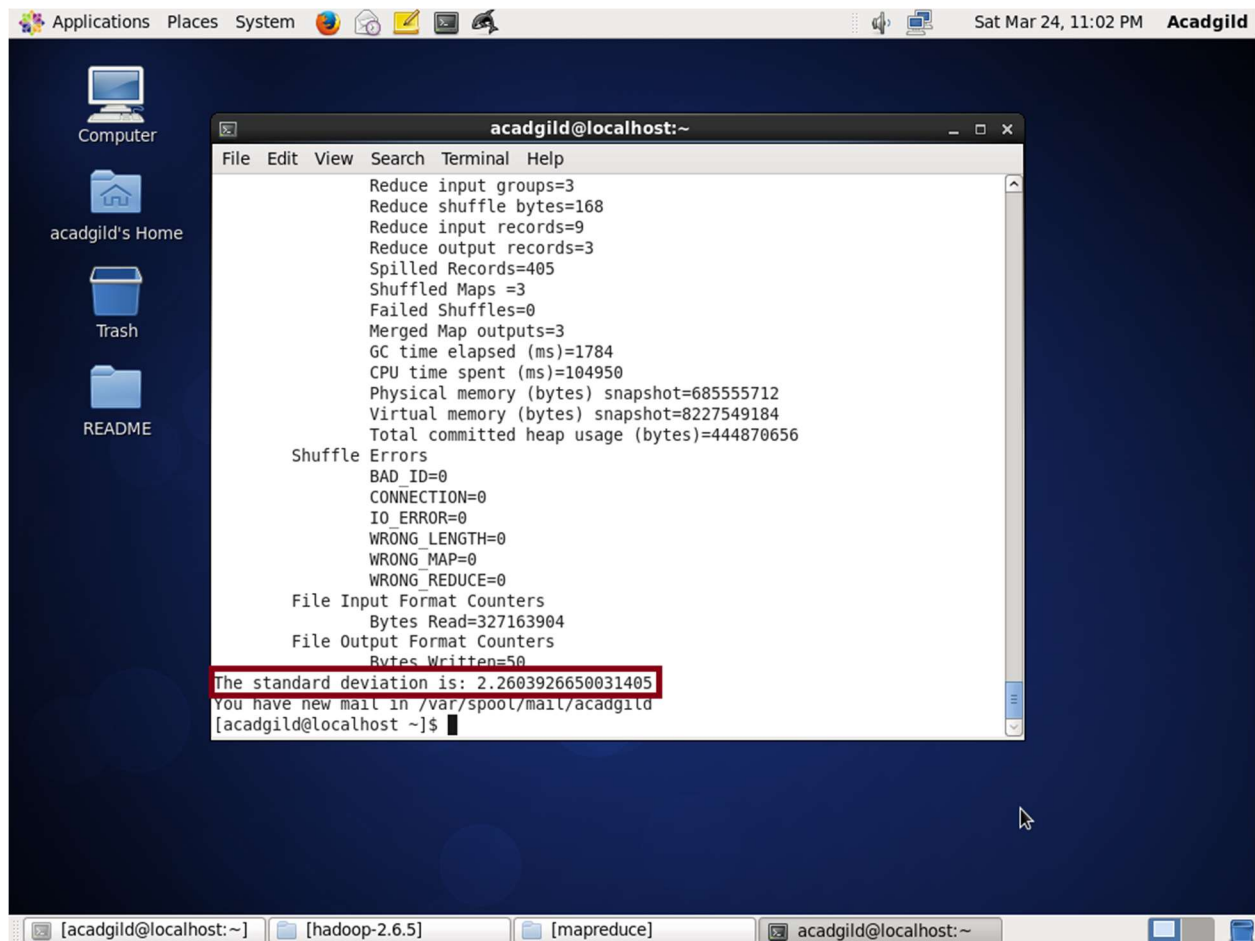
The map/reduce process is performed.



The screenshot shows a Linux desktop environment with a terminal window titled "acadgild@localhost:~". The terminal displays the command to run a Hadoop MapReduce job: `hadoop jar /home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.0.jar wordstandarddeviation /user/acadgild/hadoop/file327.txt /stddeviationoutput`. The output shows the job's progress, including warnings about native code loading, connection to the Resource Manager, and the submission of the application. The job is identified as `job_1521888884311_0007`. The progress is shown as a series of log entries: `map 0% reduce 0%`, `map 4% reduce 0%`, `map 6% reduce 0%`, `map 7% reduce 0%`, `map 10% reduce 0%`, `map 11% reduce 0%`, `map 12% reduce 0%`, `map 13% reduce 0%`, `map 17% reduce 0%`, `map 19% reduce 0%`, `map 20% reduce 0%`, `map 23% reduce 0%`, `map 26% reduce 0%`, and `map 28% reduce 0%`. The desktop background is dark blue, and the terminal window has a menu bar with "File", "Edit", "View", "Search", "Terminal", and "Help". The system tray at the bottom shows the time as "Sat Mar 24, 11:00 PM" and the user as "Acadgild".

```
acadgild@localhost:~$ hadoop jar /home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.0.jar wordstandarddeviation /user/acadgild/hadoop/file327.txt /stddeviationoutput
18/03/24 22:58:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/03/24 22:58:40 INFO client.RMPProxy: Connecting to ResourceManager at localhost/127.0.0.1:802
18/03/24 22:58:42 INFO input.FileInputFormat: Total input paths to process : 1
18/03/24 22:58:42 INFO mapreduce.JobSubmitter: number of splits:3
18/03/24 22:58:43 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1521888884311_0007
18/03/24 22:58:43 INFO impl.YarnClientImpl: Submitted application application_1521888884311_0007
18/03/24 22:58:43 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1521888884311_0007/
18/03/24 22:58:43 INFO mapreduce.Job: Running job: job_1521888884311_0007
18/03/24 22:58:54 INFO mapreduce.Job: Job job_1521888884311_0007 running in uber mode : false
18/03/24 22:59:17 INFO mapreduce.Job: map 4% reduce 0%
18/03/24 22:59:23 INFO mapreduce.Job: map 6% reduce 0%
18/03/24 22:59:26 INFO mapreduce.Job: map 7% reduce 0%
18/03/24 22:59:29 INFO mapreduce.Job: map 10% reduce 0%
18/03/24 22:59:32 INFO mapreduce.Job: map 11% reduce 0%
18/03/24 22:59:33 INFO mapreduce.Job: map 12% reduce 0%
18/03/24 22:59:35 INFO mapreduce.Job: map 13% reduce 0%
18/03/24 22:59:39 INFO mapreduce.Job: map 17% reduce 0%
18/03/24 22:59:42 INFO mapreduce.Job: map 19% reduce 0%
18/03/24 22:59:45 INFO mapreduce.Job: map 20% reduce 0%
18/03/24 22:59:48 INFO mapreduce.Job: map 23% reduce 0%
18/03/24 22:59:51 INFO mapreduce.Job: map 26% reduce 0%
18/03/24 22:59:54 INFO mapreduce.Job: map 28% reduce 0%
```

The output is displayed as Standard deviation as 2.2603926650031405



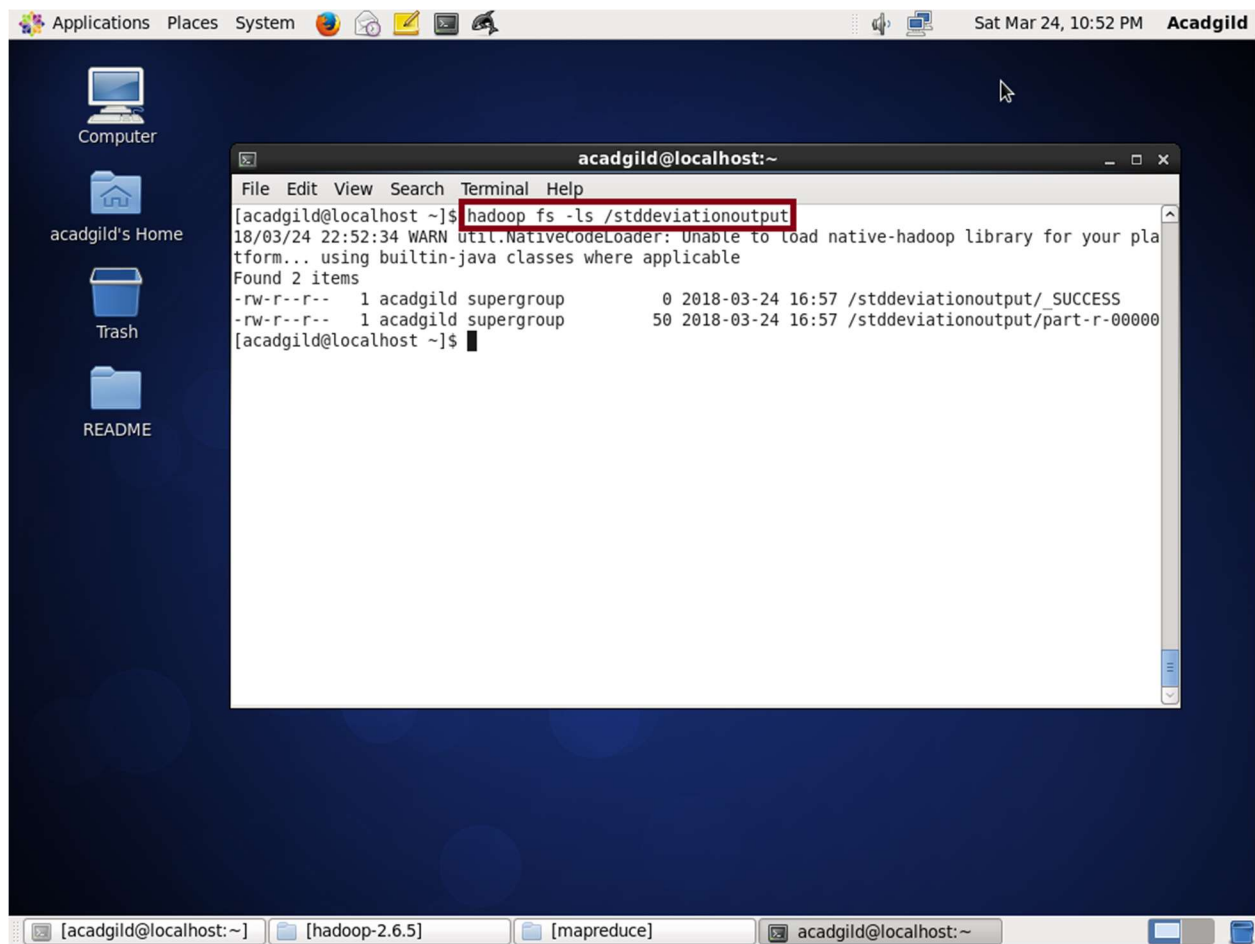
The screenshot shows a Linux desktop with a dark blue background. On the left sidebar, there are icons for 'Computer', 'acadgild's Home', 'Trash', and 'README'. The top panel displays 'Applications', 'Places', 'System', and the date 'Sat Mar 24, 11:02 PM' next to the username 'Acadgild'. A terminal window titled 'acadgild@localhost:~' is open in the center, showing the output of a Hadoop MapReduce job. The output includes various statistics such as 'Reduce input groups=3', 'Reduce shuffle bytes=168', 'GC time elapsed (ms)=1784', and 'CPU time spent (ms)=104950'. A red box highlights the line 'The standard deviation is: 2.2603926650031405'. Below this, it says 'You have new mail in /var/spool/mail/acadgild' and the prompt '[acadgild@localhost ~]\$'.

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
Reduce input groups=3  
Reduce shuffle bytes=168  
Reduce input records=9  
Reduce output records=3  
Spilled Records=405  
Shuffled Maps =3  
Failed Shuffles=0  
Merged Map outputs=3  
GC time elapsed (ms)=1784  
CPU time spent (ms)=104950  
Physical memory (bytes) snapshot=685555712  
Virtual memory (bytes) snapshot=8227549184  
Total committed heap usage (bytes)=444870656  
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=327163904  
File Output Format Counters  
Bytes Written=50  
The standard deviation is: 2.2603926650031405  
You have new mail in /var/spool/mail/acadgild  
[acadgild@localhost ~]$
```


The output is saved in the `stddeviationoutput` directory.

It is listed by using following command.

`hadoop fs -ls /stddeviationoutput`



The screenshot shows a Linux desktop environment with a dark blue background. On the left side, there are icons for 'Computer', 'acagdild's Home', 'Trash', and 'README'. The top panel displays 'Applications', 'Places', 'System', and the date 'Sat Mar 24, 10:52 PM' along with the username 'Acadgild'. A terminal window titled 'acadgild@localhost:~' is open in the center. The terminal shows the command `hadoop fs -ls /stddeviationoutput` being executed. The output indicates a warning about the native-hadoop library and lists two files in the directory: `_SUCCESS` and `part-r-00000`. The terminal window has a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. The bottom panel shows several open windows: '[acadgild@localhost: ~]', '[hadoop-2.6.5]', '[mapreduce]', and 'acadgild@localhost: ~'.

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
[acadgild@localhost ~]$ hadoop fs -ls /stddeviationoutput  
18/03/24 22:52:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pla  
tform... using builtin-java classes where applicable  
Found 2 items  
-rw-r--r--  1 acadgild supergroup          0 2018-03-24 16:57 /stddeviationoutput/_SUCCESS  
-rw-r--r--  1 acadgild supergroup       50 2018-03-24 16:57 /stddeviationoutput/part-r-00000  
[acadgild@localhost ~]$
```

By using following command, the output data which is saved are displayed.

hadoop fs -cat /stddeviationoutput/*0

