**BIG DATA ENGINEERING FOR HADOOP & SPARK TRAINING**

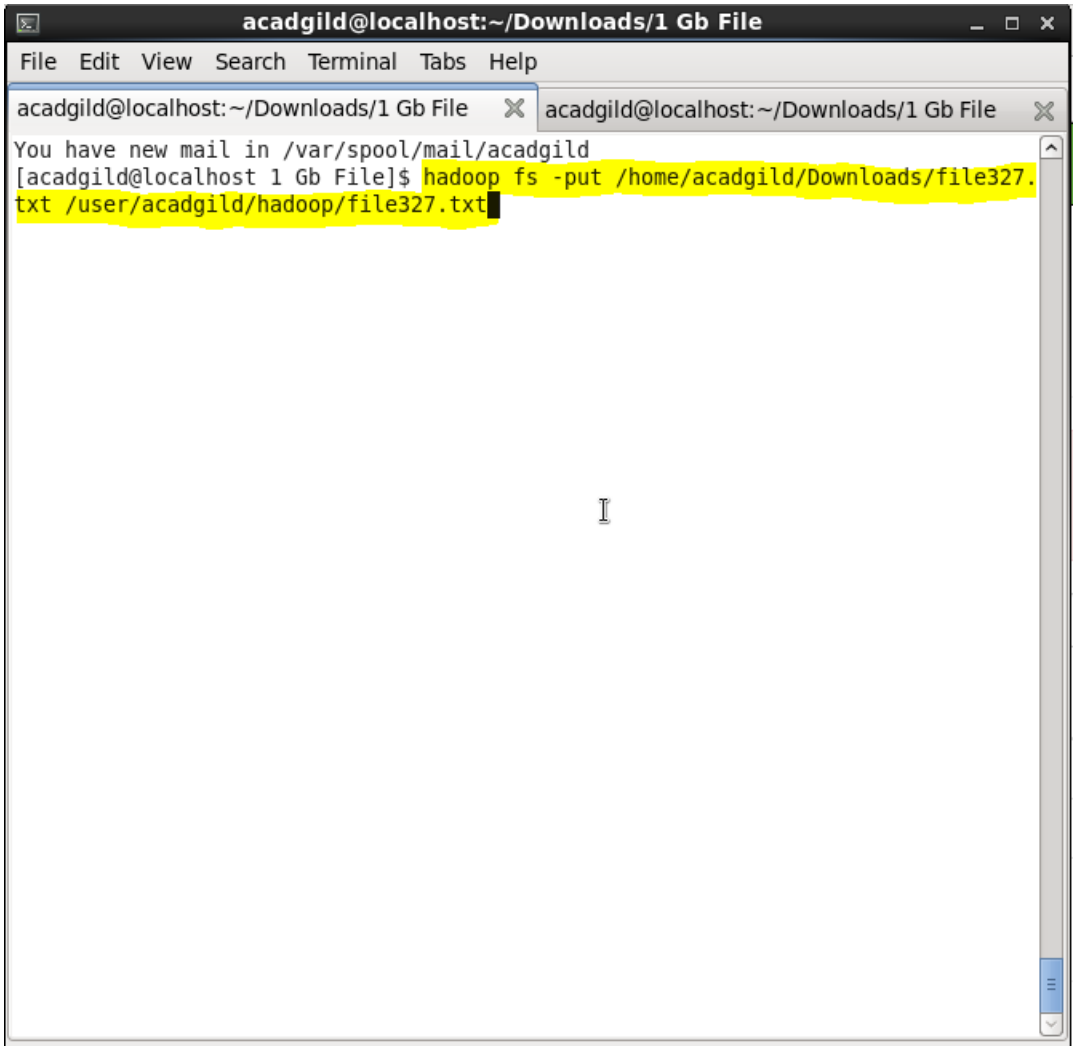
**ASSIGNMENT 3**

**Task 1:**

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

**Sol :** First copy a file to HDFS (hadoop file system) which should be more than 300 MB.

Using the commnad **"hadoop fs –put <path to file in local dir> <path of file in hdfs >"**



In hadoop home directory  we have a jar file which consists of sample java programs that can be used for performing mapreduce on sample data.

Command to check the jar file programs is

**“hadoop jar <HADOOP_HOME_DIR>/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.5.jar”**

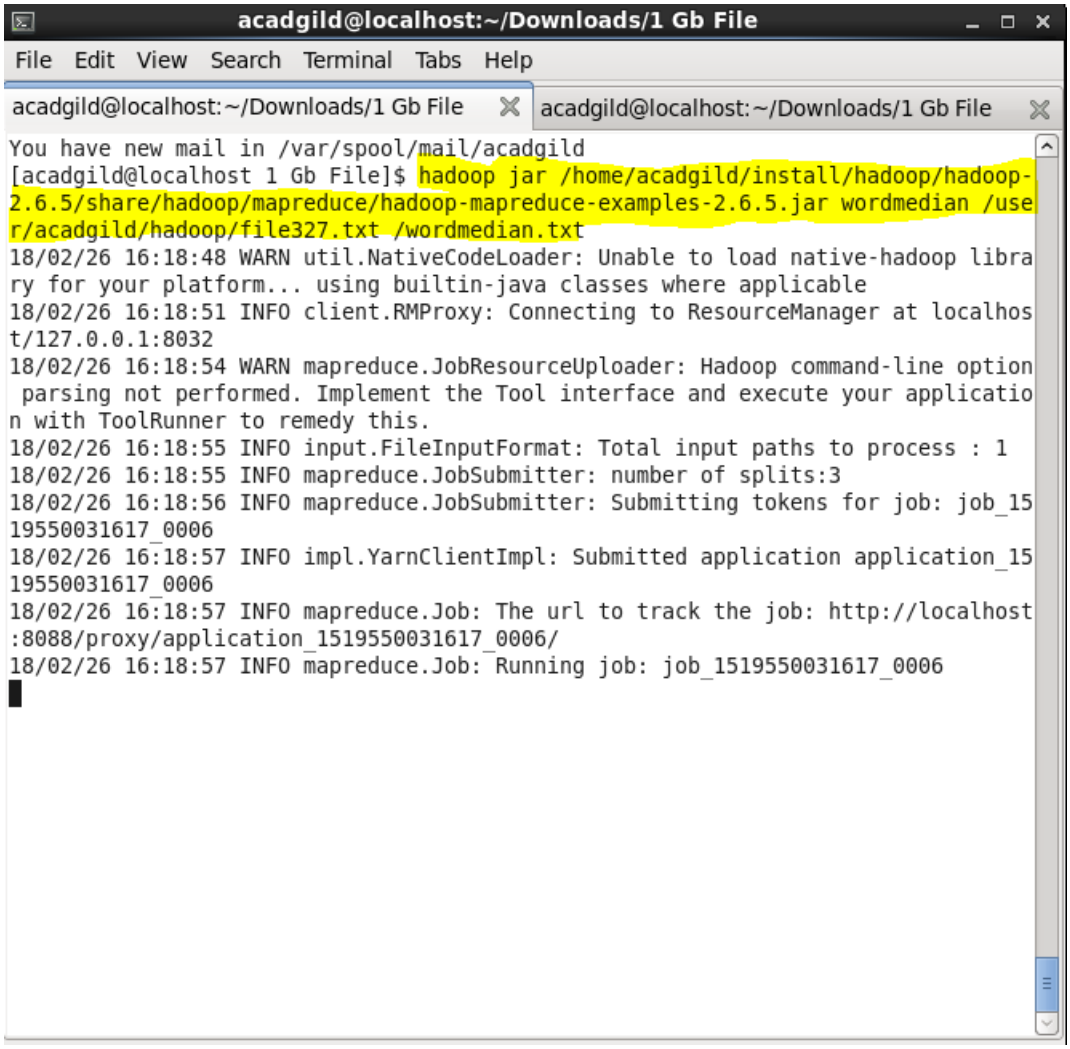
The above command lists all the programs that are available in the jar file.

**Note: file327.txt is a file that I used for demo which will be around 315MB and that is stored /user/acadgild/hadoop directory.**

**Execute wordmedian on file327.txt**

Command used to execute wordmedian program on file327.txt is

**"hadoop jar <HADOOP_HOME_DIR>/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.5.jar wordmedian /user/acadgild/hadoop/file327.txt /wordmedian.txt"**




The **hadoop jar** streaming utility enables you to create and run MapReduce jobs with any executable or script as the mapper and/or the reducer.

Here hadoop jar will execute wordmedian program and the input to the file will be file327.txt and the output of the program will be stored in /wordmedian.txt folder.

The **median** is the middle number in a group of numbers which are in Ascending order.

The Output of the above command will be

```
acadgild@localhost:~/Downloads/1 Gb File
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~/Downloads/1 Gb File    acadgild@localhost:~/Downloads/1 Gb File

                Map output bytes=536870912
                Map output materialized bytes=198
                Input split bytes=357
                Combine input records=67108984
                Combine output records=138
                Reduce input groups=6
                Reduce shuffle bytes=198
                Reduce input records=18
                Reduce output records=6
                Spilled Records=156
                Shuffled Maps =3
                Failed Shuffles=0
                Merged Map outputs=3
                GC time elapsed (ms)=4407
                CPU time spent (ms)=85140
                Physical memory (bytes) snapshot=678318080
                Virtual memory (bytes) snapshot=8227274752
                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 1 Gb File]$ clear
```

The Output of the java program is stored in /wordmedian.txt/part-r-00000 file which looks as

```
acadgild@localhost:~/Downloads/1 Gb File
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~/Downloads/1 Gb File    acadgild@localhost:~/Downloads/1 Gb File
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost 1 Gb File]$ hadoop fs -cat /wordmedian.txt/part-r-00000
18/02/26 16:28:38 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
1       8388608
2       8388608
3       25165824
4       8388608
7       8388608
8       8388608
[acadgild@localhost 1 Gb File]$
```

**Execute wordmean on file327.txt**

Command used to execute wordmean program on file327.txt is

**"hadoop jar <HADOOP_HOME_DIR>/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.5.jar wordmean /user/acadgild/hadoop/file327.txt /wordmean.txt"**

```
acadgild@localhost:~/Downloads/1 Gb File
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~/Downloads/1 Gb File   acadgild@localhost:~/Downloads/1 Gb File
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost 1 Gb File]$ hadoop jar /home/acadgild/install/hadoop/hadoop-
2.6.5/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.5.jar wordmean /user/
acadgild/hadoop/file327.txt /wordmean.txt
18/02/26 16:31:15 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
18/02/26 16:31:17 INFO client.RMProxy: Connecting to ResourceManager at localhos
t/127.0.0.1:8032
18/02/26 16:31:21 INFO input.FileInputFormat: Total input paths to process : 1
18/02/26 16:31:21 INFO mapreduce.JobSubmitter: number of splits:3
18/02/26 16:31:22 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_15
19550031617_0007
18/02/26 16:31:23 INFO impl.YarnClientImpl: Submitted application application_15
19550031617_0007
18/02/26 16:31:23 INFO mapreduce.Job: The url to track the job: http://localhost
:8088/proxy/application_1519550031617_0007/
18/02/26 16:31:23 INFO mapreduce.Job: Running job: job_1519550031617_0007
18/02/26 16:31:41 INFO mapreduce.Job: Job job_1519550031617_0007 running in uber
 mode : false
18/02/26 16:31:41 INFO mapreduce.Job:  map 0% reduce 0%
```

The **hadoop jar** streaming utility enables you to create and run MapReduce jobs with any executable or script as the mapper and/or the reducer.

Here hadoop jar will execute wordmean program and the input to the file will be file327.txt and the output of the program will be stored in /wordmean.txt folder.

The **mean** is calculated as sum of all the observations divided by total no of observations.

Mean = x1+x2+x3…….xi/N

Where N is the total no of observations and x1,x2 x3,….xi are the observations

The Output of the above command will be

```
acadgild@localhost:~/Downloads/1 Gb File
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~/Downloads/1 Gb File    acadgild@localhost:~/Downloads/1 Gb File
                Map output bytes=1946157056
                Map output materialized bytes=117
                Input split bytes=357
                Combine input records=134217830
                Combine output records=108
                Reduce input groups=2
                Reduce shuffle bytes=117
                Reduce input records=6
                Reduce output records=2
                Spilled Records=166
                Shuffled Maps =3
                Failed Shuffles=0
                Merged Map outputs=3
                GC time elapsed (ms)=3821
                CPU time spent (ms)=115360
                Physical memory (bytes) snapshot=676962304
                Virtual memory (bytes) snapshot=8227274752
                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/acadgild
[acadgild@localhost 1 Gb File]$ clear
```

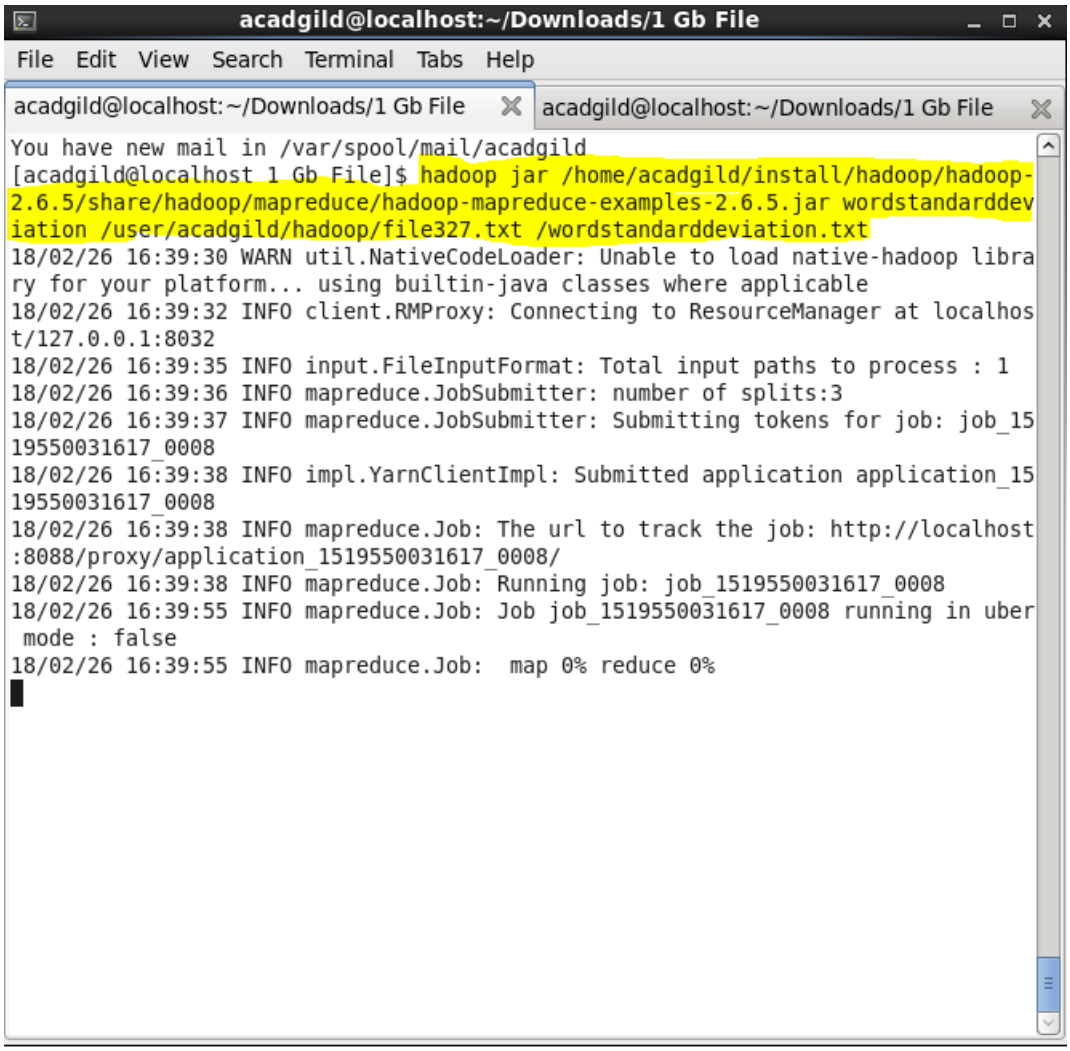
The Output of the java program is stored in /wordmean.txt/part-r-00000 file which looks as

```
acadgild@localhost:~/Downloads/1 Gb File
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~/Downloads/1 Gb File   acadgild@localhost:~/Downloads/1 Gb File
[acadgild@localhost 1 Gb File]$ hadoop fs -cat /wordmean.txt/part-r-00000
18/02/26 16:37:04 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
count   67108864
length  260046848
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost 1 Gb File]$
```

**Execute wordstandarddeviation on file327.txt**

Command used to execute wordstandarddeviation program on file327.txt is

**"hadoop jar <HADOOP_HOME_DIR>/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.5.jar wordstandarddeviation /user/acadgild/hadoop/file327.txt /wordstandarddeviation.txt"**




The **hadoop jar** streaming utility enables you to create and run MapReduce jobs with any executable or script as the mapper and/or the reducer.

Here hadoop jar will execute wordstandarddeviation program and the input to the file will be file327.txt and the output of the program will be stored in /wordstandarddeviation.txt folder.

The **Standard Deviation** is calculated by using the formula

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

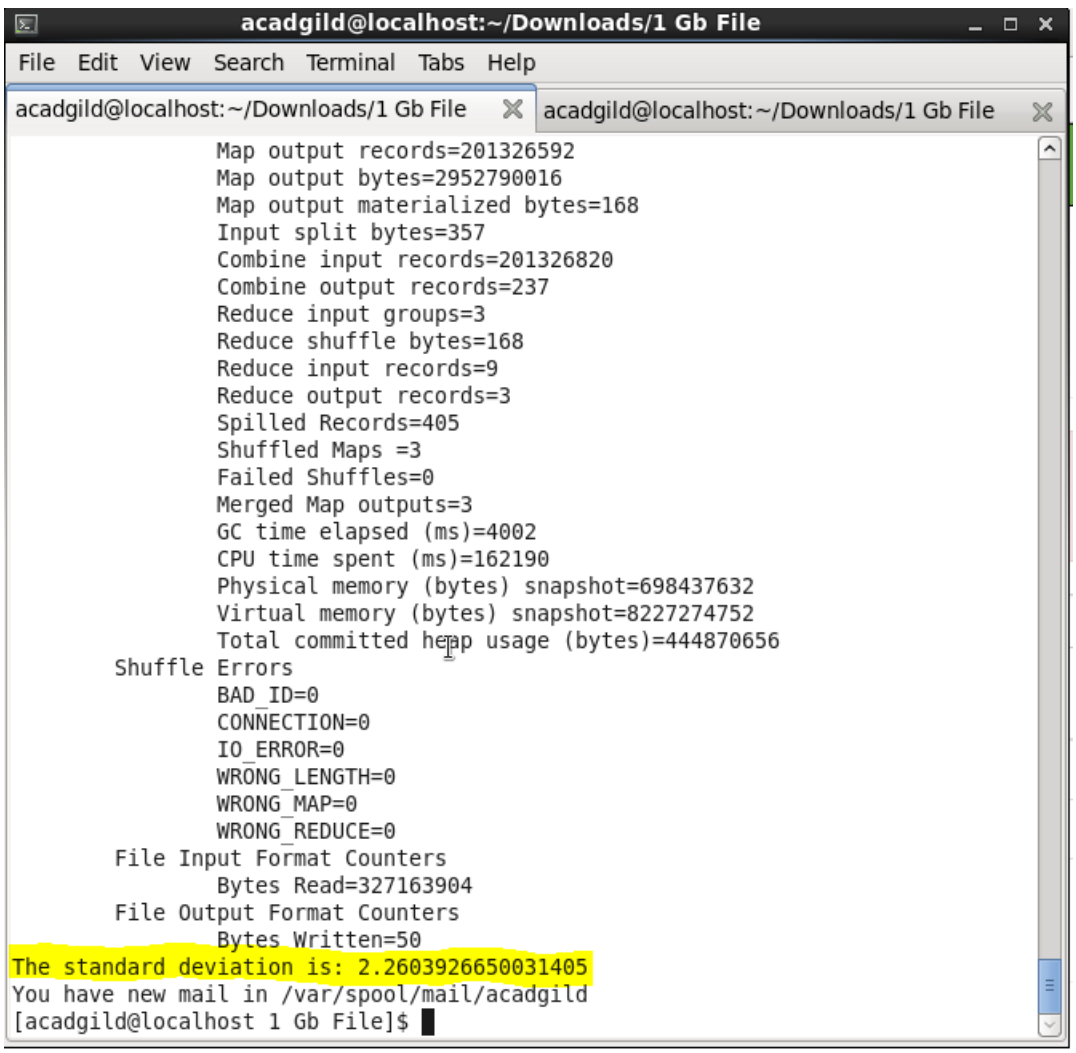
S - >Standard Deviation

N-> Total no of Observations

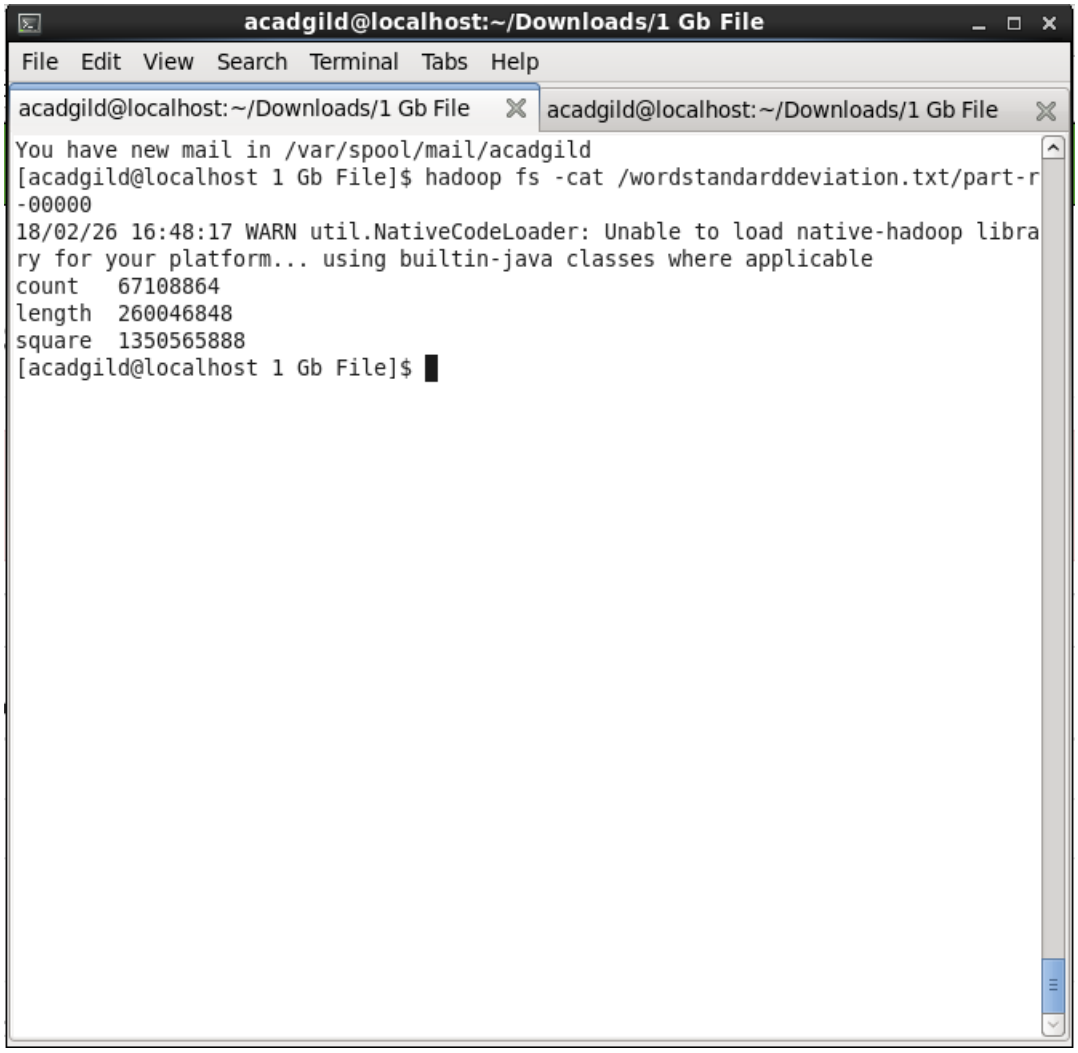
xi-> observation

$\overline{X}$ -> Mean

The Output of the above command will be



The Output of the java program is stored in /wordstandarddeviation.txt/part-r-00000 file which looks as



```
acadgild@localhost:~/Downloads/1 Gb File
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~/Downloads/1 Gb File   acadgild@localhost:~/Downloads/1 Gb File
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost 1 Gb File]$ hadoop fs -cat /wordstandarddeviation.txt/part-r
-00000
18/02/26 16:48:17 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
count   67108864
length  260046848
square  1350565888
[acadgild@localhost 1 Gb File]$
```

To check the process that are running while Mapreduce is going on on file327.txt we can open a new tab and use jps command to check the deamons or processes that are executed background.