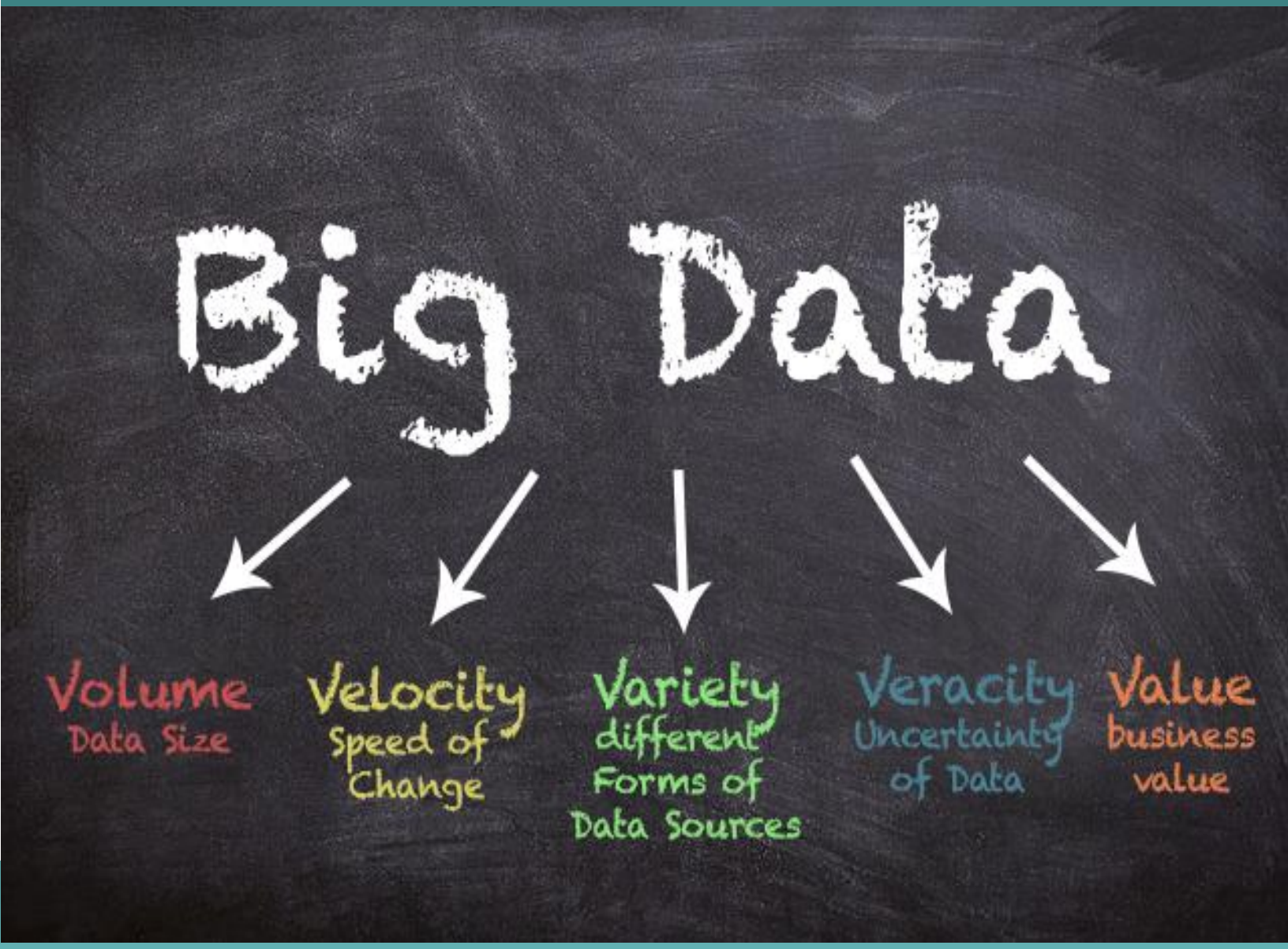
07/12/2018 | By: Ajit K Prasad



Big Data Engineering with Hadoop & Spark

Assignment on MapReduce Introduction

Acadgild

Session 4: Assignment 4.1

This assignment is aimed at consolidating the concepts that was learned during the MapReduce Introduction session of the course.

Problem Statement

- We have a dataset of sales of different TV sets across different locations in file ***"television.txt"***.
- Records look like:
  Samsung|Optima|14|Madhya Pradesh|132401|14200
- The fields are arranged like:
  Company Name|Product Name|Size_in_inches|State|Pin Code|Price
- There are some invalid records which contain 'NA' in either Company Name or Product Name.

Task 1:

- Write a Map Reduce program to filter out the invalid records. Map only job will fit for this context.

Solution:

- A project ***"TelevisionNARecordFilter"*** was created using Eclipse which had a class ***"NATVRF.java"***. This allowed to process the records by tokenizing each recording using **"|"** as separator and filtered records, excluding all invalid records which had either NA in company or product field. This was a "map" only program
- Once the code was developed, the project was exported to jar file ***"NATVRF.jar"***
- The steps to execute the job are shared as screenshots below along with the commands used
- The source code/class file ***"NATVRF.java"*** is shared along with the assignment report on ***GitHub*** in a compressed file
- Copy the file ***"television.txt"*** to HDFS folder
  hadoop fs -put television.txt /hadoopdata/assignmentjobs/television.txt

```
[acadgild@localhost ~]$
[acadgild@localhost ~]$ hadoop fs -put television.txt /hadoopdata/assignmentjobs/television.txt
18/07/12 12:19:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -ls /hadoopdata/assignmentjobs/
18/07/12 12:19:50 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
Found 1 items
-rw-r--r--   1 acadgild supergroup        733 2018-07-12 12:19 /hadoopdata/assignmentjobs/television.txt
[acadgild@localhost ~]$
```

– Check the content of the file ***"television.txt"***
hadoop fs -cat /hadoopdata/assignmentjobs/television.txt

```
[acadgild@localhost ~]$ hadoop fs -cat /hadoopdata/assignmentjobs/television.txt
18/07/12 12:33:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar Pradesh|232401|16200
Akai|Decent|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Zen|Super|14|Maharashtra|619082|9200
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar Pradesh|232401|16200
Onida|Decent|14|Uttar Pradesh|232401|16200
Onida|NA|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Zen|Super|14|Maharashtra|619082|9200
Samsung|Optima|14|Madhya Pradesh|132401|14200
NA|Lucid|18|Uttar Pradesh|232401|16200
Samsung|Decent|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Samsung|Super|14|Maharashtra|619082|9200
Samsung|Super|14|Maharashtra|619082|9200
Samsung|Super|14|Maharashtra|619082|9200You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

– Run the Map program using the command below
hadoop jar NATVRF.jar NATVRF /hadoopdata/assignmentjobs/television.txt /hadoopdata/assignmentjobs/NATVRFOutput

```
[acadgild@localhost ~]$
[acadgild@localhost ~]$ hadoop jar NATVRF.jar NATVRF /hadoopdata/assignmentjobs/television.txt /hadoopdata/assignmentjobs/NATVRFOut
put
18/07/12 12:30:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
18/07/12 12:30:09 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/07/12 12:30:14 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interfac
e and execute your application with ToolRunner to remedy this.
18/07/12 12:30:16 INFO input.FileInputFormat: Total input paths to process : 1
18/07/12 12:30:16 INFO mapreduce.JobSubmitter: number of splits:1
18/07/12 12:30:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1531377372055_0001
18/07/12 12:30:20 INFO impl.YarnClientImpl: Submitted application application_1531377372055_0001
18/07/12 12:30:21 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1531377372055_0001/
18/07/12 12:30:21 INFO mapreduce.Job: Running job: job_1531377372055_0001
18/07/12 12:31:11 INFO mapreduce.Job: Job job_1531377372055_0001 running in uber mode : false
18/07/12 12:31:11 INFO mapreduce.Job:  map 0% reduce 0%
18/07/12 12:31:40 INFO mapreduce.Job:  map 100% reduce 0%
18/07/12 12:31:42 INFO mapreduce.Job: Job job_1531377372055_0001 completed successfully
18/07/12 12:31:43 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=107021
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=860
                HDFS: Number of bytes written=646
                HDFS: Number of read operations=5
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=23763
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=23763
                Total vcore-milliseconds taken by all map tasks=23763
                Total megabyte-milliseconds taken by all map tasks=24333312
        Map-Reduce Framework
```

```
18/07/12 12:31:42 INFO mapreduce.Job: Job job_1531377372055_0001 completed successfully
18/07/12 12:31:43 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=107021
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=860
                HDFS: Number of bytes written=646
                HDFS: Number of read operations=5
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=23763
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=23763
                Total vcore-milliseconds taken by all map tasks=23763
                Total megabyte-milliseconds taken by all map tasks=24333312
        Map-Reduce Framework
                Map input records=18
                Map output records=16
                Input split bytes=127
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=313
                CPU time spent (ms)=2400
                Physical memory (bytes) snapshot=107552768
                Virtual memory (bytes) snapshot=2060169216
                Total committed heap usage (bytes)=62980096
        File Input Format Counters
                Bytes Read=733
        File Output Format Counters
                Bytes Written=646
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

– Finally, check the filtered output
hadoop fs -cat /hadoopdata/assignmentjobs/NATVRFOutput/part-m-00000

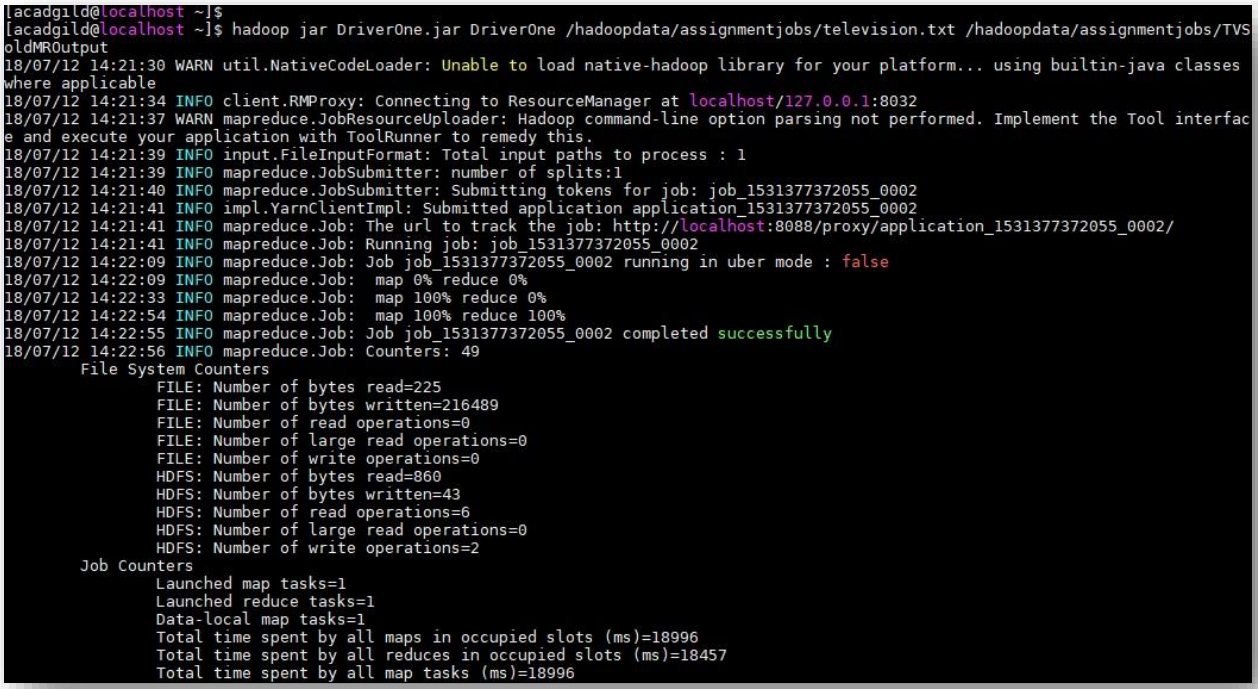
```
[acadgild@localhost ~]$ hadoop fs -cat /hadoopdata/assignmentjobs/NATVRFOutput/part-m-00000
18/07/12 12:35:20 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar Pradesh|232401|16200
Akai|Decent|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Zen|Super|14|Maharashtra|619082|9200
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar Pradesh|232401|16200
Onida|Decent|14|Uttar Pradesh|232401|16200
Lava|Attention|20|Assam|454601|24200
Zen|Super|14|Maharashtra|619082|9200
Samsung|Optima|14|Madhya Pradesh|132401|14200
Samsung|Decent|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Samsung|Super|14|Maharashtra|619082|9200
Samsung|Super|14|Maharashtra|619082|9200
Samsung|Super|14|Maharashtra|619082|9200
[acadgild@localhost ~]$
```

Task 2:

- Write a Map Reduce program to calculate the total units sold for each Company.

Solution:

- Since it was a "map-reduce" program, a project ***"TVSoldMR"*** was created using Eclipse which had 3 different classes, a main class ***"DriverOne.java"***, a mapper class ***"MyMapperOne.java"*** and a reducer class ***"MyReducerOne.java"***
- Once the code was developed, the project was exported to jar file ***"DriverOne.jar"***
- The steps to execute the job are shared as screenshots below along with the commands used
- All source code/class files are shared along with the assignment report on ***GitHub*** in a compressed file
- The file ***"television.txt"*** was already present on HDFS folder
- Run the MapReduce program using the command below
  hadoop jar DriverOne.jar DriverOne /hadoopdata/assignmentjobs/television.txt /hadoopdata/assignmentjobs/TVSoldMROutput



```
                Total time spent by all reduce tasks (ms)=18457
                Total vcore-milliseconds taken by all map tasks=18996
                Total vcore-milliseconds taken by all reduce tasks=18457
                Total megabyte-milliseconds taken by all map tasks=19451904
                Total megabyte-milliseconds taken by all reduce tasks=18899968
        Map-Reduce Framework
                Map input records=18
                Map output records=18
                Map output bytes=183
                Map output materialized bytes=225
                Input split bytes=127
                Combine input records=0
                Combine output records=0
                Reduce input groups=6
                Reduce shuffle bytes=225
                Reduce input records=18
                Reduce output records=6
                Spilled Records=36
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=537
                CPU time spent (ms)=4640
                Physical memory (bytes) snapshot=321081344
                Virtual memory (bytes) snapshot=4118192128
                Total committed heap usage (bytes)=222429184
        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=733
        File Output Format Counters
                Bytes Written=43
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

– Finally, check the MapReduce output using the command *hadoop fs -cat /hadoopdata/assignmentjobs/TVSoldMROutput/part-r-00000*

```
[acadgild@localhost ~]$ hadoop fs -ls /hadoopdata/assignmentjobs/TVSoldMROutput/
18/07/12 14:32:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-07-12 14:22 /hadoopdata/assignmentjobs/TVSoldMROutput/_SUCCESS
-rw-r--r--   1 acadgild supergroup         43 2018-07-12 14:22 /hadoopdata/assignmentjobs/TVSoldMROutput/part-r-00000
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -cat /hadoopdata/assignmentjobs/TVSoldMROutput/part-r-00000
18/07/12 14:32:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
Akai    1
Lava    3
NA      1
Onida   4
Samsung 7
Zen     2
[acadgild@localhost ~]$
```

Task 3:

- Write a Map Reduce program to calculate the total units sold in each state for **Onida** company.

Solution:

- Since it was a "map-reduce" program, a project ***"OnidaSUMR"*** was created using Eclipse which had 3 different classes, a main class ***"OnidaSUMR.java"***, a mapper class ***"OnidaSUMapper.java"*** and a reducer class ***"OnidaSUReducer.java"***
- Once the code was developed, the project was exported to jar file ***"OnidaSUMR.jar"***
- The steps to execute the job are shared as screenshots below along with the commands used
- All source code/class files are shared along with the assignment report on ***GitHub*** in a compressed file
- The file ***"television.txt"*** was already present on HDFS folder
- Run the MapReduce program using the command below
  hadoop jar OnidaSUMR.jar /hadoopdata/assignmentjobs/television.txt /hadoopdata/assignmentjobs/OnidaSUMROutput

```
[acadgild@localhost ~]$ hadoop jar OnidaSUMR.jar /hadoopdata/assignmentjobs/television.txt /hadoopdata/assignmentjobs/OnidaSUMROutp
ut
18/07/12 16:17:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
18/07/12 16:17:48 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/07/12 16:17:55 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interfac
e and execute your application with ToolRunner to remedy this.
18/07/12 16:17:57 INFO input.FileInputFormat: Total input paths to process : 1
18/07/12 16:17:57 INFO mapreduce.JobSubmitter: number of splits:1
18/07/12 16:17:58 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1531377372055_0003
18/07/12 16:17:59 INFO impl.YarnClientImpl: Submitted application application_1531377372055_0003
18/07/12 16:18:00 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1531377372055_0003/
18/07/12 16:18:00 INFO mapreduce.Job: Running job: job_1531377372055_0003
18/07/12 16:18:47 INFO mapreduce.Job: Job job_1531377372055_0003 running in uber mode : false
18/07/12 16:18:47 INFO mapreduce.Job:  map 0% reduce 0%
18/07/12 16:19:20 INFO mapreduce.Job:  map 100% reduce 0%
18/07/12 16:19:53 INFO mapreduce.Job:  map 100% reduce 100%
18/07/12 16:19:54 INFO mapreduce.Job: Job job_1531377372055_0003 completed successfully
18/07/12 16:19:55 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=26
                FILE: Number of bytes written=215005
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=860
                HDFS: Number of bytes written=16
                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)=27480
                Total time spent by all reduces in occupied slots (ms)=29612
                Total time spent by all map tasks (ms)=27480
                Total time spent by all reduce tasks (ms)=29612
                Total vcore-milliseconds taken by all map tasks=27480
```

```
                Total time spent by all reduce tasks (ms)=29612
                Total vcore-milliseconds taken by all map tasks=27480
                Total vcore-milliseconds taken by all reduce tasks=29612
                Total megabyte-milliseconds taken by all map tasks=28139520
                Total megabyte-milliseconds taken by all reduce tasks=30322688
        Map-Reduce Framework
                Map input records=18
                Map output records=3
                Map output bytes=54
                Map output materialized bytes=26
                Input split bytes=127
                Combine input records=3
                Combine output records=1
                Reduce input groups=1
                Reduce shuffle bytes=26
                Reduce input records=1
                Reduce output records=1
                Spilled Records=2
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=768
                CPU time spent (ms)=6890
                Physical memory (bytes) snapshot=320638976
                Virtual memory (bytes) snapshot=4121538560
                Total committed heap usage (bytes)=222429184
        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=733
        File Output Format Counters
                Bytes Written=16
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

– Finally, check the output using the following command. Fields which contained ***"NA"*** have been filtered out and hence the output
hadoop fs -cat /hadoopdata/assignmentjobs/OnidaSUMROutput/part-r-00000

```
[acadgild@localhost ~]$
[acadgild@localhost ~]$ hadoop fs -cat /hadoopdata/assignmentjobs/OnidaSUMROutput/part-r-00000
18/07/12 16:21:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
Uttar Pradesh   3
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```