Big Data Hadoop & Spark Training

Assignment4:MapReduce Programs

RASHMI.K

Task 1: Write a Map Reduce program to filter out the invalid records. Map only job will fit for this context.

Below is the input file TVDataset.csv to perform task1.

Company	Product	Size in inches	state	Pincode	price
Samsung	NA	10	Karnataka	345354	1093453
LG	LED	10	Andhra Pradesh	213434	1043454
Onida	LED	10	Odisha	534533	1083453
NA	LED	10	Madhya Pradesh	565443	1042346
Haier	LED	10	Gujurat	456764	1023456
NA	LED	10	Delhi	345656	1010354
Sharp	LED	10	Haryana	345456	990234
Akai	LED	10	Jammu	67658	985534
Mitashi	NA	10	Kashmir	75634	995345
Samsung	Plasma	13	Karnataka	345354	1264534
NA	Plasma	13	Andhra Pradesh	213434	1234566
Onida	Plasma	13	Odisha	534533	1345745
Sansui	Plasma	13	Madhya Pradesh	565443	1436865
Haier	NA	13	Gujurat	456764	1457794
NA	Plasma	13	Delhi	345656	1576357
Sharp	Plasma	13	Haryana	345456	924365
Akai	Plasma	13	Jammu	67658	923456
Mitashi	Plasma	13	Kashmir	75634	972442
Samsung	Ultra slim	22	Karnataka	345354	1353464
NA	Ultra slim	22	Andhra Pradesh	213434	1345736
Onida	Ultra slim	22	Odisha	534533	1686544
Sansui	Ultra slim	22	Madhya Pradesh	565443	1535786
Haier	Ultra slim	22	Gujurat	456764	1474876
Philips	NA	22	Delhi	345656	1346787
Sharp	Ultra slim	22	Haryana	345456	1497865
Akai	Ultra slim	22	Jammu	67658	1346543
Mitashi	Ultra slim	22	Kashmir	75634	1572654
Samsung	LCD	31	Karnataka	345354	943534
LG	NA	31	Andhra Pradesh	213434	923454
Onida	LCD	31	Odisha	534533	993453
Sansui	LCD	31	Madhya Pradesh	565443	992234
Haier	LCD	31	Gujurat	456764	943524
Philips	LCD	31	Delhi	345656	832452
Sharp	LCD	31	Haryana	345456	884325

Akai	LCD	31	Jammu	67658	893453
Mitashi	LCD	31	Kashmir	75634	973455
Samsung	Flat	26	Karnataka	345354	1034553
LG	Flat	26	Andhra Pradesh	213434	1935324
Onida	Flat	26	Odisha	534533	1345534
Sansui	Flat	26	Madhya Pradesh	565443	1345554
Haier	Flat	26	Gujurat	456764	934756
Philips	Flat	26	Delhi	345656	934622
Sharp	Flat	26	Haryana	345456	993453
Akai	Flat	26	Jammu	67658	982345
Mitashi	Flat	26	Kashmir	75634	972345

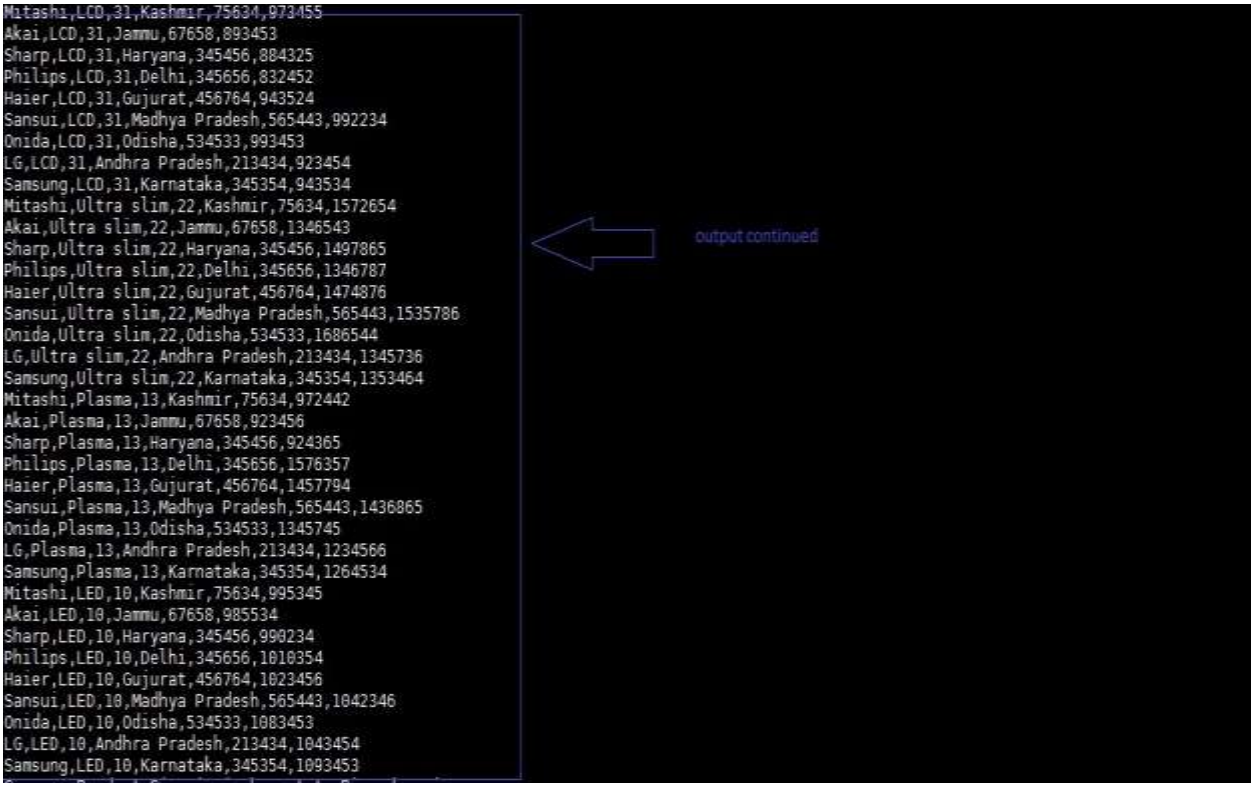
Moved the input data set TVDataset.csv from local to HDFS and executing the jar file Task1TV to eliminate the invalid records with NA as highlighted above and print valid records which does not have invalid records in both Company Name and Product Name.

```
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost projects]$ hadoop jar Task1TV.jar /hadoopdata/TVDataset.csv /TV1Task1outfinal
18/03/06 16:29:57 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
18/03/06 16:30:00 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/03/06 16:30:02 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and
execute your application with ToolRunner to remedy this.
18/03/06 16:30:03 INFO input.FileInputFormat: Total input paths to process : 1
18/03/06 16:30:03 INFO mapreduce.JobSubmitter: number of splits:1
18/03/06 16:30:04 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1520331066586_0014
18/03/06 16:30:04 INFO impl.YarnClientImpl: Submitted application application_1520331066586_0014
18/03/06 16:30:04 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1520331066586_0014/
18/03/06 16:30:04 INFO mapreduce.Job: Running job: job_1520331066586_0014
18/03/06 16:30:23 INFO mapreduce.Job: Job job_1520331066586_0014 running in uber mode : false
18/03/06 16:30:23 INFO mapreduce.Job:  map 0% reduce 0%
18/03/06 16:30:35 INFO mapreduce.Job:  map 100% reduce 0%
18/03/06 16:30:49 INFO mapreduce.Job:  map 100% reduce 100%
18/03/06 16:30:50 INFO mapreduce.Job: Job job_1520331066586_0014 completed successfully
18/03/06 16:30:50 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=1908
                FILE: Number of bytes written=218307
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1967
                HDFS: Number of bytes written=1810
                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
```

```
                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)=9358
                Total time spent by all reduces in occupied slots (ms)=10478
                Total time spent by all map tasks (ms)=9358
                Total time spent by all reduce tasks (ms)=10478
                Total vcore-milliseconds taken by all map tasks=9358
                Total vcore-milliseconds taken by all reduce tasks=10478
                Total megabyte-milliseconds taken by all map tasks=9582592
                Total megabyte-milliseconds taken by all reduce tasks=10729472
        Map-Reduce Framework
                Map input records=46
                Map output records=46
                Map output bytes=1810
                Map output materialized bytes=1988
                Input split bytes=111
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=1988
                Reduce input records=46
                Reduce output records=46
                Spilled Records=92
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=296
                CPU time spent (ms)=2780
                Physical memory (bytes) snapshot=298590208
                Virtual memory (bytes) snapshot=4118188032
                Total committed heap usage (bytes)=170004480
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
```

```
        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=1856
        File Output Format Counters
                Bytes Written=1810
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost projects]$ hadoop fs -cat /TV1Task1outfinal/part-r-00000
18/03/06 16:31:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
Mitashi,Flat,26,Kashmir,75634,972345
Akai,Flat,26,Jammu,67658,982345
Sharp,Flat,26,Haryana,345456,993453
Philips,Flat,26,Delhi,345656,934622
Haier,Flat,26,Gujurat,456764,934756
Sansui,Flat,26,Madhya Pradesh,565443,1345554
Onida,Flat,26,Odisha,534533,1345534
LG,Flat,26,Andhra Pradesh,213434,1935324
Samsung,Flat,26,Karnataka,345354,1834553
Mitashi,LCD,31,Kashmir,75634,973455
Akai,LCD,31,Jammu,67658,893453
Sharp,LCD,31,Haryana,345456,884325
Philips,LCD,31,Delhi,345656,832452
Haier,LCD,31,Gujurat,456764,943524
Sansui,LCD,31,Madhya Pradesh,565443,992234
Onida,LCD,31,Odisha,534533,993453
LG,LCD,31,Andhra Pradesh,213434,923454
Samsung,LCD,31,Karnataka,345354,943534
Mitashi,Ultra slim,22,Kashmir,75634,1572654
Akai,Ultra slim,22,Jammu,67658,1346543
Sharp,Ultra slim,22,Haryana,345456,1497865
Philips,Ultra slim,22,Delhi,345656,1346787
Haier,Ultra slim,22,Gujurat,456764,1474876
```

Output after eliminating "invalid records with NA"



Input File for Task2 & Task 3

Samsung	LED	10	Karnataka	345354	1093453	344
LG	LED	10	Andhra Pradesh	213434	1043454	3455
Onida	LED	10	Odisha	534533	1083453	3245
Sansui	LED	10	Madhya Pradesh	565443	1042346	47244
Haier	LED	10	Gujurat	456764	1023456	3243
Philips	LED	10	Delhi	345656	1010354	43524
Sharp	LED	10	Haryana	345456	990234	63454
Akai	LED	10	Jammu	67658	985534	23455
Mitashi	LED	10	Kashmir	75634	995345	234555
Samsung	Plasma	13	Karnataka	345354	1264534	4553
LG	Plasma	13	Andhra Pradesh	213434	1234566	344
Onida	Plasma	13	Odisha	534533	1345745	4372
Sansui	Plasma	13	Madhya Pradesh	565443	1436865	4377
Haier	Plasma	13	Gujurat	456764	1457794	7432
Philips	Plasma	13	Delhi	345656	1576357	2437
Sharp	Plasma	13	Haryana	345456	924365	7843
Akai	Plasma	13	Jammu	67658	923456	247
Mitashi	Plasma	13	Kashmir	75634	972442	84234
Samsung	Ultra slim	22	Karnataka	345354	1353464	24786
LG	Ultra slim	22	Andhra Pradesh	213434	1345736	2421
Onida	Ultra slim	22	Odisha	534533	1686544	34778

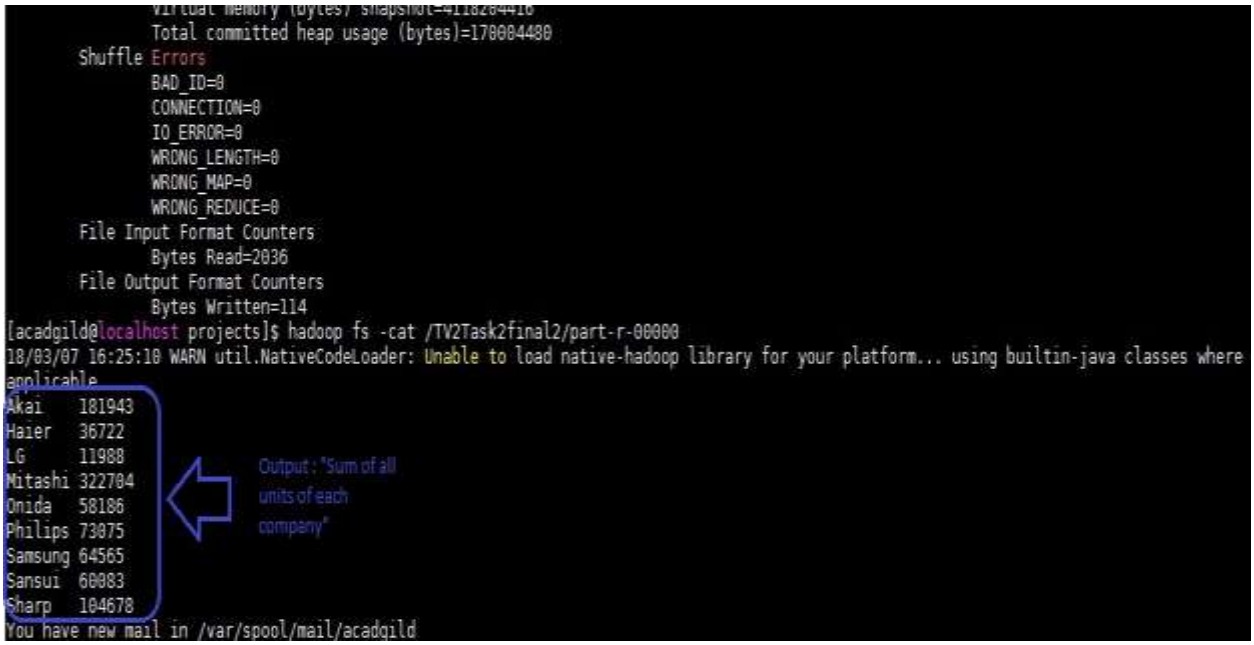
Sansui	Ultra slim	22	Madhya Pradesh	565443	1535786	7444
Haier	Ultra slim	22	Gujurat	456764	1474876	2347
Philips	Ultra slim	22	Delhi	345656	1346787	24634
Sharp	Ultra slim	22	Haryana	345456	1497865	2442
Akai	Ultra slim	22	Jammu	67658	1346543	23443
Mitashi	Ultra slim	22	Kashmir	75634	1572654	2347
Samsung	LCD	31	Karnataka	345354	943534	34134
LG	LCD	31	Andhra Pradesh	213434	923454	2345
Onida	LCD	31	Odisha	534533	993453	2344
Sansui	LCD	31	Madhya Pradesh	565443	992234	234
Haier	LCD	31	Gujurat	456764	943524	23466
Philips	LCD	31	Delhi	345656	832452	2346
Sharp	LCD	31	Haryana	345456	884325	23455
Akai	LCD	31	Jammu	67658	893453	134555
Mitashi	LCD	31	Kashmir	75634	973455	1434
Samsung	Flat	26	Karnataka	345354	1034553	748
LG	Flat	26	Andhra Pradesh	213434	1935324	3423
Onida	Flat	26	Odisha	534533	1345534	13447
Sansui	Flat	26	Madhya Pradesh	565443	1345554	784
Haier	Flat	26	Gujurat	456764	934756	234
Philips	Flat	26	Delhi	345656	934622	134
Sharp	Flat	26	Haryana	345456	993453	7484
Akai	Flat	26	Jammu	67658	982345	243
Mitashi	Flat	26	Kashmir	75634	972345	134

Task 2: Write a Map Reduce program to calculate the total units sold for each Company.

Moved the input file newTV21.csv from local to HDFS and executing jar to find total units sold for each company.

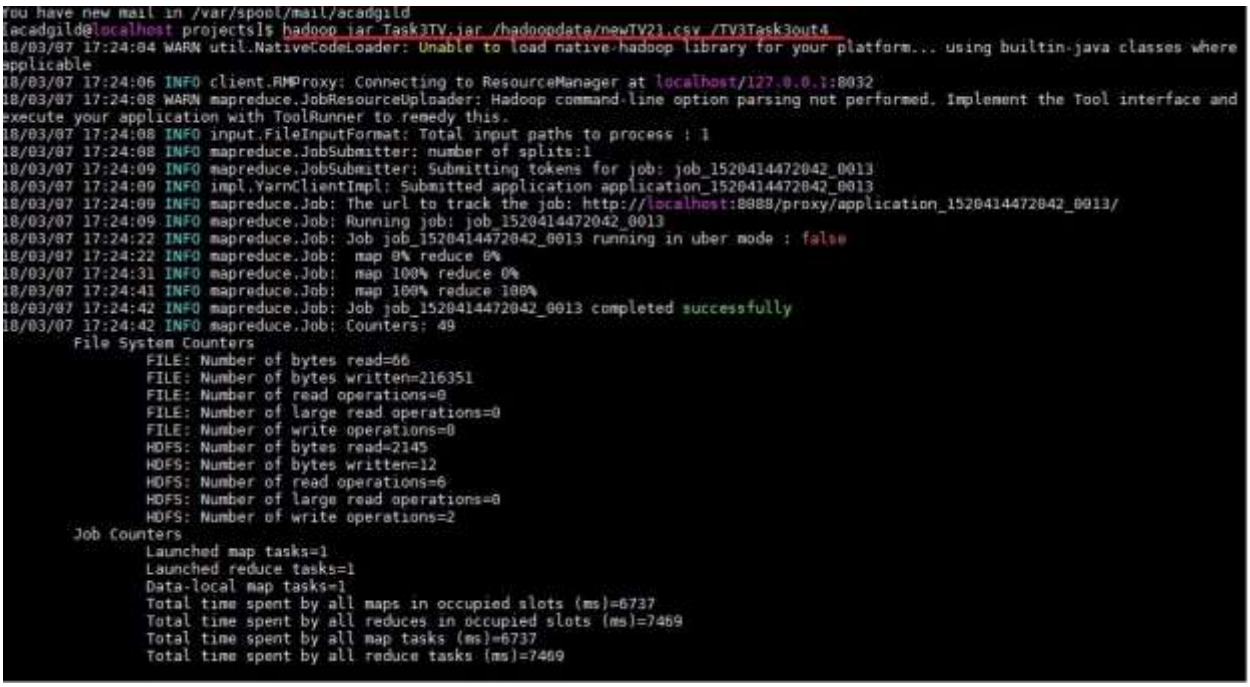
```
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost projects]$ hadoop jar Task2TV.jar /hadoopdata/newTV21.csv /TV2Task2final2
18/03/07 16:24:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
18/03/07 16:24:25 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/03/07 16:24:26 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and
execute your application with ToolRunner to remedy this.
18/03/07 16:24:27 INFO input.FileInputFormat: Total input paths to process : 1
18/03/07 16:24:27 INFO mapreduce.JobSubmitter: number of splits:1
18/03/07 16:24:27 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1520414472042_0008
18/03/07 16:24:28 INFO impl.YarnClientImpl: Submitted application application_1520414472042_0008
18/03/07 16:24:28 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1520414472042_0008/
18/03/07 16:24:28 INFO mapreduce.Job: Running job: job_1520414472042_0008
18/03/07 16:24:41 INFO mapreduce.Job: Job job_1520414472042_0008 running in uber mode : false
18/03/07 16:24:41 INFO mapreduce.Job:  map 0% reduce 0%
18/03/07 16:24:50 INFO mapreduce.Job:  map 100% reduce 0%
18/03/07 16:24:59 INFO mapreduce.Job:  map 100% reduce 100%
18/03/07 16:25:00 INFO mapreduce.Job: Job job_1520414472042_0008 completed successfully
18/03/07 16:25:00 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=561
                FILE: Number of bytes written=217095
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2145
                HDFS: Number of bytes written=114
                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)=6396
                Total time spent by all reduces in occupied slots (ms)=7172
                Total time spent by all map tasks (ms)=6396
                Total time spent by all reduce tasks (ms)=7172
```

```
                Total time spent by all map tasks (ms)=6396
                Total time spent by all reduce tasks (ms)=7172
                Total vcore-milliseconds taken by all map tasks=6396
                Total vcore-milliseconds taken by all reduce tasks=7172
                Total megabyte-milliseconds taken by all map tasks=6549504
                Total megabyte-milliseconds taken by all reduce tasks=7344128
        Map-Reduce Framework
                Map input records=45
                Map output records=45
                Map output bytes=465
                Map output materialized bytes=561
                Input split bytes=109
                Combine input records=0
                Combine output records=0
                Reduce input groups=9
                Reduce shuffle bytes=561
                Reduce input records=45
                Reduce output records=9
                Spilled Records=90
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=210
                CPU time spent (ms)=1950
                Physical memory (bytes) snapshot=381125632
                Virtual memory (bytes) snapshot=4118204416
                Total committed heap usage (bytes)=170004480
        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=2036
        File Output Format Counters
                Bytes Written=114
```



Task 3: Write a Map Reduce program to calculate the total units sold in each state for Onida company.

Moved the input file newTV21.csv from local to HDFS and executing jar to find total units sold by Onida company.



```
        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)=6737
                Total time spent by all reduces in occupied slots (ms)=7469
                Total time spent by all map tasks (ms)=6737
                Total time spent by all reduce tasks (ms)=7469
                Total vcore-milliseconds taken by all map tasks=6737
                Total vcore-milliseconds taken by all reduce tasks=7469
                Total megabyte-milliseconds taken by all map tasks=6898688
                Total megabyte-milliseconds taken by all reduce tasks=7648256
        Map-Reduce Framework
                Map input records=45
                Map output records=5
                Map output bytes=50
                Map output materialized bytes=66
                Input split bytes=109
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=66
                Reduce input records=5
                Reduce output records=1
                Spilled Records=10
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=269
                CPU time spent (ms)=2020
                Physical memory (bytes) snapshot=298225664
                Virtual memory (bytes) snapshot=4118188032
                Total committed heap usage (bytes)=170004480
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
```

```
                Total committed heap usage (bytes)=170004480
        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=2036
        File Output Format Counters
                Bytes Written=12
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost projects]$ hadoop fs -cat /TV3Task3out/part-r-00000
18/03/07 17:25:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
cat: `/TV3Task3out/part-r-00000': No such file or directory
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost projects]$ hadoop fs -cat /TV3Task3out4/part-r-00000
18/03/07 17:25:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
Onida   58186
[acadgild@localhost projects]$
```

Output:"Sum of all units of Onida company