Assignment 4.1

Problem Statement

We have a dataset of sales of different TV sets across different locations.

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.

Task 2:

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

Task 3:

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

company.

Program Structure

The java program consists of following classes

1. MR Driver class – (TvExample.java)
2. TvMapper.java (mapper class used for task1)
3. TvCalcMapper.java (mapper class used for task2)
4. TvCalcMapperOnida.java (mapper class used for task3)
5. TvReducer.java (common reducer class for all 3 tasks)

MR Driver Class (TvExample)

This class the main class of the program. It expects user 3 types of command line inputs for the 3 tasks described in the problem statement. Based on the user input, it initializes the job and sets up mapper and reducer classes for the job. It also sets the input/output type for the job etc.

The program expects following command line arguments

TvExample <option> <Input File> <Output Dir>
User can give one of the 3 options

1. Filter – This option creates a MR job for task1 where the invalid records are filtered. No reducers are needed for this
2. CalcTotalUnitsPerComany – This option calculates total products sold by each company
3. CalTotalUnitsForOnida – This option calculates total products sold in each state for the 'Onida' company

TvMapper Class

This class is used for filtering the invalid records having NA in either company and product name. It extends the MR Mapper class for this purpose.

The <map> method is called for each line of the file television.txt. Sample line from the file is below

Samsung|Optima|14|Madhya Pradesh|132401|14200

In the map method each line is passed as key which then split in words with "|" delimiter. The result is stored in an array and items at offset 0 (company) and 1 (Product) are compared with NA. If there is a match, the map method returns. If NA is not found then it uses context.write() method to output the same line. This way only valid records are the output of the mapper

TvCalcMapper Class

This class is used for calculating the number of product sold by each company which extends MR Mapper class. The <map> method is called for each line of the file television.txt. First filter the input line to make sure there is no NA field present in the input line. This is described in the description of TvMapper class. The output of the mapper is <K,V> where key is the company name and value=1. We use context.write() method to output the company name and value=1.

TvCalcMapperOnida Class

This class is used for calculating the number of product sold "Onida" company for every state. The <map> method is called for each line of the file television.txt. We first filter the input line to the map method to make sure that we process only records with company name as Onida. We parse the state name in a variable which becomes the key for the mapper output.

The output of the mapper is <K,V> where key is the state name and value=1. We use context.write() method to output the state name and value=1.

TvReducer Class

The reducer class is used only for Task2 and 3 of the assignment. Its only job is to calculate sum of the values (integers) passed to the reducer against each key and output the resulting value against that key using context.write()

Program execution logs

```
                          • MobaXterm 10.4 •
                (SSH client, X-server and networking tools)

→ SSH session to acadgild@192.168.56.2
   • SSH compression : v
   • SSH-browser     : v
   • X11-forwarding  : v  (remote display is forwarded through SSH)
   • DISPLAY         : v  (automatically set on remote server)

→ For more info, ctrl+click on help or visit our website
```

```
Last login: Thu Jul  5 12:14:13 2018 from 192.168.56.1
[acadgild.mmisra ~]$ hadoop fs -ls /
18/07/05 12:15:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
Found 6 items
drwxr-xr-x   - acadgild supergroup          0 2018-07-04 13:21 /files
drwxr-xr-x   - acadgild supergroup          0 2018-06-24 10:50 /mohit
drwxr-xr-x   - acadgild supergroup          0 2018-07-01 10:01 /sqoopout
drwxr-xr-x   - acadgild supergroup          0 2018-07-01 11:21 /sqoopoutbyid
drwx------   - acadgild supergroup          0 2018-06-24 11:27 /tmp
drwxr-xr-x   - acadgild supergroup          0 2018-07-03 14:12 /user
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
```

```
# Contents of the input file
[acadgild.mmisra ~]$ cat television.txt
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|9200
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$ hadoop jar TvExample.jar
Tv Example
Valid options are:
TvExample filter <input path> <output path>
TvExample CalcTotalUnitsPerCompany <input path> <output path>
TvExample CalcTotalUnitsForOnida <input path> <output path>
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
```

```
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
```

#Option to run Task1 of the assignment

```
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$ hadoop jar TvExample.jar filter /files/television.txt /out1
Tv Example
18/07/05 12:17:10 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
18/07/05 12:17:12 INFO client.RMProxy: Connecting to ResourceManager at
localhost/127.0.0.1:8032
18/07/05 12:17:13 WARN mapreduce.JobResourceUploader: Hadoop command-line option
parsing not performed. Implement the Tool interface and execute your application with
ToolRunner to remedy this.
18/07/05 12:17:14 INFO input.FileInputFormat: Total input paths to process : 1
18/07/05 12:17:15 INFO mapreduce.JobSubmitter: number of splits:1
18/07/05 12:17:15 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1530772907907_0001
18/07/05 12:17:17 INFO impl.YarnClientImpl: Submitted application
application_1530772907907_0001
18/07/05 12:17:17 INFO mapreduce.Job: The url to track the job:
http://localhost:8088/proxy/application_1530772907907_0001/
18/07/05 12:17:17 INFO mapreduce.Job: Running job: job_1530772907907_0001
18/07/05 12:17:31 INFO mapreduce.Job: Job job_1530772907907_0001 running in uber mode
: false
18/07/05 12:17:31 INFO mapreduce.Job:  map 0% reduce 0%
18/07/05 12:17:38 INFO mapreduce.Job:  map 100% reduce 0%
18/07/05 12:17:39 INFO mapreduce.Job: Job job_1530772907907_0001 completed
successfully
18/07/05 12:17:39 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=107342
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=842
                HDFS: Number of bytes written=678
                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)=4909
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=4909
                Total vcore-milliseconds taken by all map tasks=4909
                Total megabyte-milliseconds taken by all map tasks=5026816
        Map-Reduce Framework
                Map input records=18
                Map output records=16
                Input split bytes=107
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=87
                CPU time spent (ms)=650
                Physical memory (bytes) snapshot=158007296
```

```
                Virtual memory (bytes) snapshot=2082119680
                Total committed heap usage (bytes)=114819072
        File Input Format Counters
                Bytes Read=735
        File Output Format Counters
                Bytes Written=678
Job success
Tv Example Success
[acadgild.mmisra ~]$ hadoop fs -ls /out1
18/07/05 12:17:47 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 2018-07-05 12:17 /out1/_SUCCESS
-rw-r--r--   1 acadgild supergroup        678 2018-07-05 12:17 /out1/part-m-00000
You have new mail in /var/spool/mail/acadgild
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$ hadoop fs -cat /out1/part-m-00000
18/07/05 12:17:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable

#Output - no lines with NA
Samsung|Optima|14|Madhya Pradesh|132401|14200   1
Onida|Lucid|18|Uttar Pradesh|232401|16200       1
Akai|Decent|16|Kerala|922401|12200      1
Lava|Attention|20|Assam|454601|24200    1
Zen|Super|14|Maharashtra|619082|9200    1
Samsung|Optima|14|Madhya Pradesh|132401|14200   1
Onida|Lucid|18|Uttar Pradesh|232401|16200       1
Onida|Decent|14|Uttar Pradesh|232401|16200      1
Lava|Attention|20|Assam|454601|24200    1
Zen|Super|14|Maharashtra|619082|9200    1
Samsung|Optima|14|Madhya Pradesh|132401|14200   1
Samsung|Decent|16|Kerala|922401|12200   1
Lava|Attention|20|Assam|454601|24200    1
Samsung|Super|14|Maharashtra|619082|9200        1
Samsung|Super|14|Maharashtra|619082|9200        1
Samsung|Super|14|Maharashtra|619082|9200        1
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$ hadoop jar TvExample.jar
Tv Example
Valid options are:
TvExample filter <input path> <output path>
TvExample CalcTotalUnitsPerCompany <input path> <output path>
TvExample CalcTotalUnitsForOnida <input path> <output path>
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$

#Option to run Task2 of the assignment
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$ hadoop jar TvExample.jar CalcTotalUnitsPerCompany
/files/television.txt /out2
Tv Example
18/07/05 12:18:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
18/07/05 12:18:24 INFO client.RMProxy: Connecting to ResourceManager at
localhost/127.0.0.1:8032
18/07/05 12:18:25 WARN mapreduce.JobResourceUploader: Hadoop command-line option
parsing not performed. Implement the Tool interface and execute your application with
```

```
ToolRunner to remedy this.
18/07/05 12:18:25 INFO input.FileInputFormat: Total input paths to process : 1
18/07/05 12:18:25 INFO mapreduce.JobSubmitter: number of splits:1
18/07/05 12:18:25 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1530772907907_0002
18/07/05 12:18:26 INFO impl.YarnClientImpl: Submitted application
application_1530772907907_0002
18/07/05 12:18:26 INFO mapreduce.Job: The url to track the job:
http://localhost:8088/proxy/application_1530772907907_0002/
18/07/05 12:18:26 INFO mapreduce.Job: Running job: job_1530772907907_0002
18/07/05 12:18:33 INFO mapreduce.Job: Job job_1530772907907_0002 running in uber mode
: false
18/07/05 12:18:33 INFO mapreduce.Job:  map 0% reduce 0%
18/07/05 12:18:38 INFO mapreduce.Job:  map 100% reduce 0%
18/07/05 12:18:44 INFO mapreduce.Job:  map 100% reduce 100%
18/07/05 12:18:44 INFO mapreduce.Job: Job job_1530772907907_0002 completed
successfully
18/07/05 12:18:44 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=204
                FILE: Number of bytes written=215663
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=842
                HDFS: Number of bytes written=38
                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)=3443
                Total time spent by all reduces in occupied slots (ms)=3229
                Total time spent by all map tasks (ms)=3443
                Total time spent by all reduce tasks (ms)=3229
                Total vcore-milliseconds taken by all map tasks=3443
                Total vcore-milliseconds taken by all reduce tasks=3229
                Total megabyte-milliseconds taken by all map tasks=3525632
                Total megabyte-milliseconds taken by all reduce tasks=3306496
        Map-Reduce Framework
                Map input records=18
                Map output records=16
                Map output bytes=166
                Map output materialized bytes=204
                Input split bytes=107
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=204
                Reduce input records=16
                Reduce output records=5
                Spilled Records=32
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=151
                CPU time spent (ms)=1530
                Physical memory (bytes) snapshot=424235008
                Virtual memory (bytes) snapshot=4167303168
                Total committed heap usage (bytes)=312999936
        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=735
        File Output Format Counters
                Bytes Written=38
Job success
Tv Example Success
[acadgild.mmisra ~]$ hadoop fs -ls /out2
18/07/05 12:18:51 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 2018-07-05 12:18 /out2/_SUCCESS
-rw-r--r--   1 acadgild supergroup         38 2018-07-05 12:18 /out2/part-r-00000
You have new mail in /var/spool/mail/acadgild
[acadgild.mmisra ~]$ hadoop fs -cat /out1/part-r-00000
18/07/05 12:18:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
cat: `/out1/part-r-00000': No such file or directory
[acadgild.mmisra ~]$ hadoop fs -cat /out2/part-r-00000
18/07/05 12:19:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
```

#Output – Number of products sold by each company

```
Akai    1
Lava    3
Onida   3
Samsung 7
Zen     2
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$ hadoop jar TvExample.jar
Tv Example
Valid options are:
TvExample filter <input path> <output path>
TvExample CalcTotalUnitsPerCompany <input path> <output path>
TvExample CalcTotalUnitsForOnida <input path> <output path>
[acadgild.mmisra ~]$
```

#Option to run Task3 of the assignment

```
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$ hadoop jar TvExample.jar CalcTotalUnitsForOnida
/files/television.txt /out3
Tv Example
18/07/05 12:19:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
18/07/05 12:19:33 INFO client.RMProxy: Connecting to ResourceManager at
localhost/127.0.0.1:8032
18/07/05 12:19:34 WARN mapreduce.JobResourceUploader: Hadoop command-line option
parsing not performed. Implement the Tool interface and execute your application with
ToolRunner to remedy this.
18/07/05 12:19:35 INFO input.FileInputFormat: Total input paths to process : 1
18/07/05 12:19:35 INFO mapreduce.JobSubmitter: number of splits:1
18/07/05 12:19:36 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1530772907907_0003
18/07/05 12:19:36 INFO impl.YarnClientImpl: Submitted application
```

```
application_1530772907907_0003
18/07/05 12:19:36 INFO mapreduce.Job: The url to track the job:
http://localhost:8088/proxy/application_1530772907907_0003/
18/07/05 12:19:36 INFO mapreduce.Job: Running job: job_1530772907907_0003
18/07/05 12:19:43 INFO mapreduce.Job: Job job_1530772907907_0003 running in uber mode
: false
18/07/05 12:19:43 INFO mapreduce.Job:  map 0% reduce 0%
18/07/05 12:19:48 INFO mapreduce.Job:  map 100% reduce 0%
18/07/05 12:19:53 INFO mapreduce.Job:  map 100% reduce 100%
18/07/05 12:19:53 INFO mapreduce.Job: Job job_1530772907907_0003 completed
successfully
18/07/05 12:19:53 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=66
                FILE: Number of bytes written=215397
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=842
                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)=2745
                Total time spent by all reduces in occupied slots (ms)=3087
                Total time spent by all map tasks (ms)=2745
                Total time spent by all reduce tasks (ms)=3087
                Total vcore-milliseconds taken by all map tasks=2745
                Total vcore-milliseconds taken by all reduce tasks=3087
                Total megabyte-milliseconds taken by all map tasks=2810880
                Total megabyte-milliseconds taken by all reduce tasks=3161088
        Map-Reduce Framework
                Map input records=18
                Map output records=3
                Map output bytes=54
                Map output materialized bytes=66
                Input split bytes=107
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=66
                Reduce input records=3
                Reduce output records=1
                Spilled Records=6
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=136
                CPU time spent (ms)=1380
                Physical memory (bytes) snapshot=419835904
                Virtual memory (bytes) snapshot=4163325952
                Total committed heap usage (bytes)=310902784
        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=735
        File Output Format Counters
                Bytes Written=16
Job success
Tv Example Success
You have new mail in /var/spool/mail/acadgild
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$ hadoop fs -ls /out3
18/07/05 12:20:00 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 2018-07-05 12:19 /out3/_SUCCESS
-rw-r--r--   1 acadgild supergroup         16 2018-07-05 12:19 /out3/part-r-00000
[acadgild.mmisra ~]$
```

#Output – Number of products sold in each state by Onida

```
[acadgild.mmisra ~]$ hadoop fs -cat /out3/part-r-00000
18/07/05 12:20:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
Uttar Pradesh   3
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
```