Hadoop

Online Transaction Analysis

Sandeep Kumar Bhandoria

Objective

Given with the data, my objective was to analyse the data to produce relevant information for which decision has to be made.

The data i have is the transaction details of online shopping website.

Use case 1

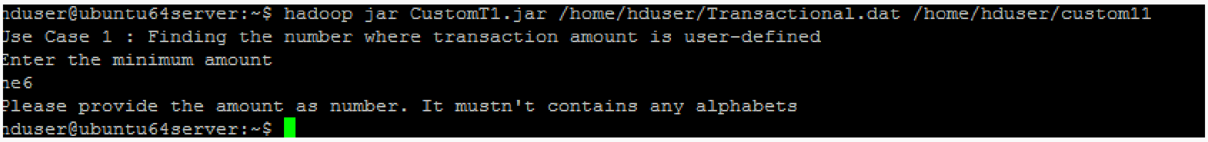
We are given the task to find the all the transaction done where amount spent is greater than some user-defined values.

In the case we are expecting input from user for the value of amount on which we have to decide the transaction.

We create a driver class where we put our logic in there to get the input from user. Basically we have used a Scanner class.

For Validation of we have checked the user input before processing it for further.

If the user is passing String in place of number he/she will be displayed a message showing an error message to provide valid input and start the job again.

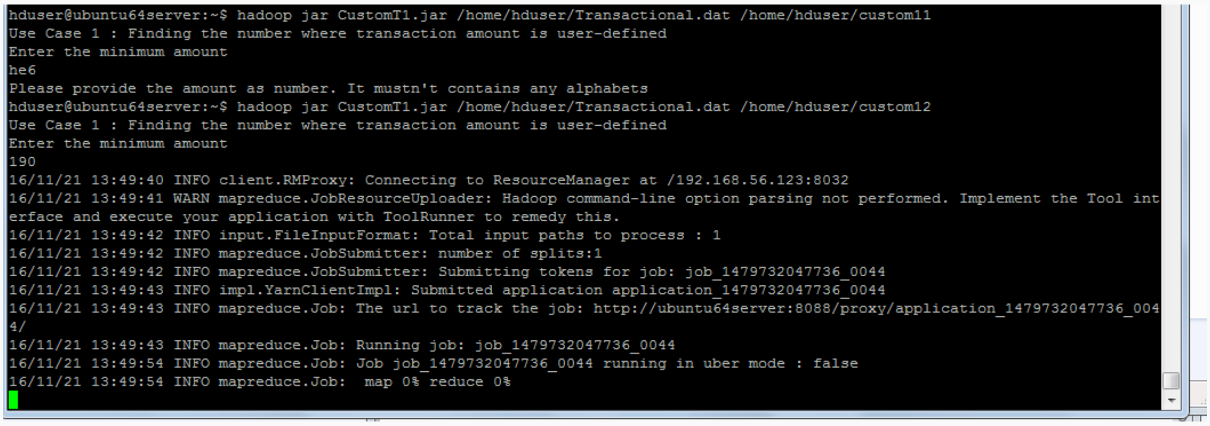


```
nduser@ubuntu64server:~$ hadoop jar CustomT1.jar /home/hduser/Transactional.dat /home/hduser/custom11
Jse Case 1 : Finding the number where transaction amount is user-defined
Enter the minimum amount
ne6
Please provide the amount as number. It mustn't contains any alphabets
nduser@ubuntu64server:~$
```

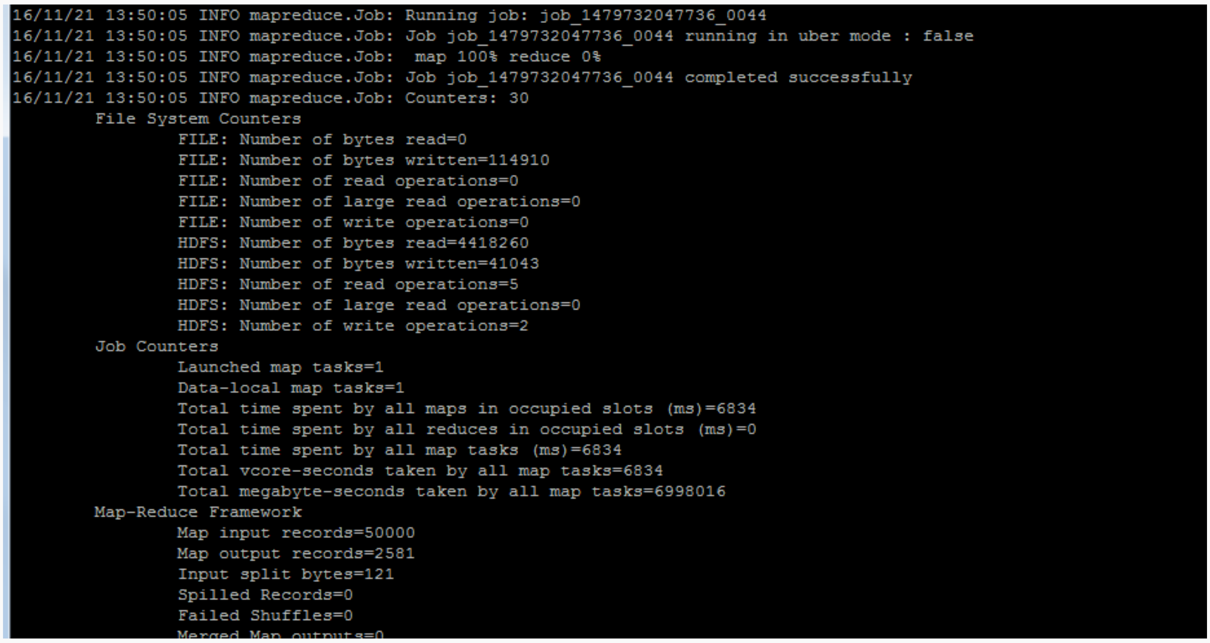
Since the program requires all the transaction we create transaction as a key and amount as value to check whether to or not to display the transaction id.

For This Separate MyKey and MyVal class are created.

We have used out own InputFormat class for our Customized RecordReader who help as to get the right key /value from a line.



```
hduser@ubuntu64server:~$ hadoop jar CustomT1.jar /home/hduser/Transactional.dat /home/hduser/custom11
Use Case 1 : Finding the number where transaction amount is user-defined
Enter the minimum amount
he6
Please provide the amount as number. It mustn't contains any alphabets
hduser@ubuntu64server:~$ hadoop jar CustomT1.jar /home/hduser/Transactional.dat /home/hduser/custom12
Use Case 1 : Finding the number where transaction amount is user-defined
Enter the minimum amount
190
16/11/21 13:49:40 INFO client.RMProxy: Connecting to ResourceManager at /192.168.56.123:8032
16/11/21 13:49:41 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool int
erface and execute your application with ToolRunner to remedy this.
16/11/21 13:49:42 INFO input.FileInputFormat: Total input paths to process : 1
16/11/21 13:49:42 INFO mapreduce.JobSubmitter: number of splits:1
16/11/21 13:49:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1479732047736_0044
16/11/21 13:49:43 INFO impl.YarnClientImpl: Submitted application application_1479732047736_0044
16/11/21 13:49:43 INFO mapreduce.Job: The url to track the job: http://ubuntu64server:8088/proxy/application_1479732047736_004
4/
16/11/21 13:49:43 INFO mapreduce.Job: Running job: job_1479732047736_0044
16/11/21 13:49:54 INFO mapreduce.Job: Job job_1479732047736_0044 running in uber mode : false
16/11/21 13:49:54 INFO mapreduce.Job:  map 0% reduce 0%
```



```
16/11/21 13:50:05 INFO mapreduce.Job: Running job: job_1479732047736_0044
16/11/21 13:50:05 INFO mapreduce.Job: Job job_1479732047736_0044 running in uber mode : false
16/11/21 13:50:05 INFO mapreduce.Job:  map 100% reduce 0%
16/11/21 13:50:05 INFO mapreduce.Job: Job job_1479732047736_0044 completed successfully
16/11/21 13:50:05 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=114910
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=4418260
                HDFS: Number of bytes written=41043
                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)=6834
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=6834
                Total vcore-seconds taken by all map tasks=6834
                Total megabyte-seconds taken by all map tasks=6998016
        Map-Reduce Framework
                Map input records=50000
                Map output records=2581
                Input split bytes=121
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
```

Note: There is no reducer in this program and for this we also set number of reducers to 0.

Job.setNumReduceTasks(2);

Below commands wait for the main method to close before completing the job.

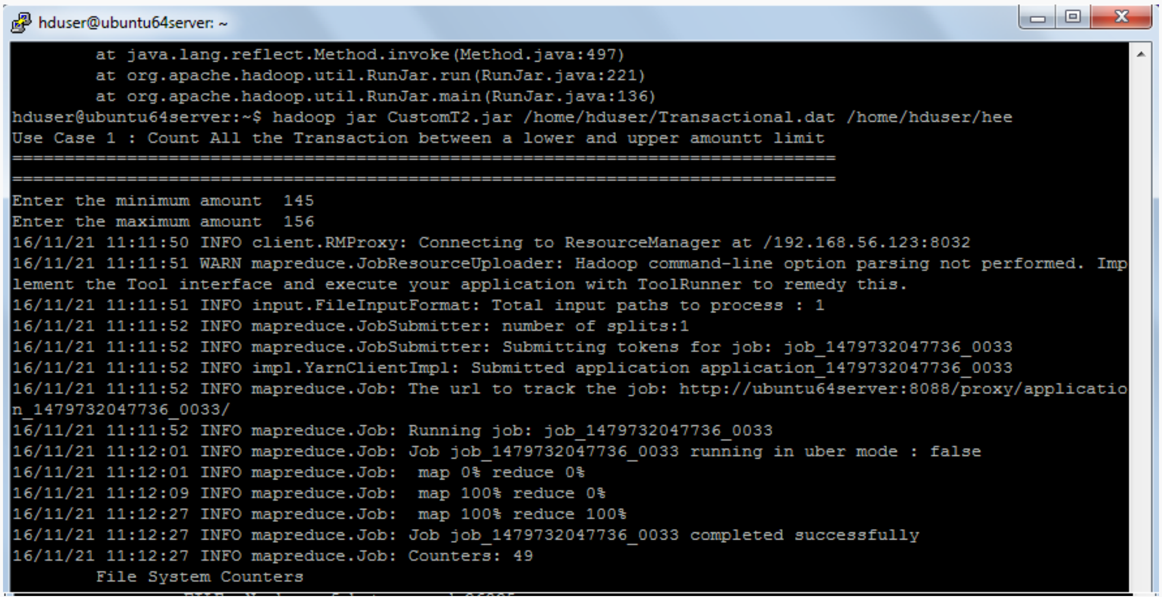
System.exit(job.waitForCompletion(true)?0:1);

Use case 2

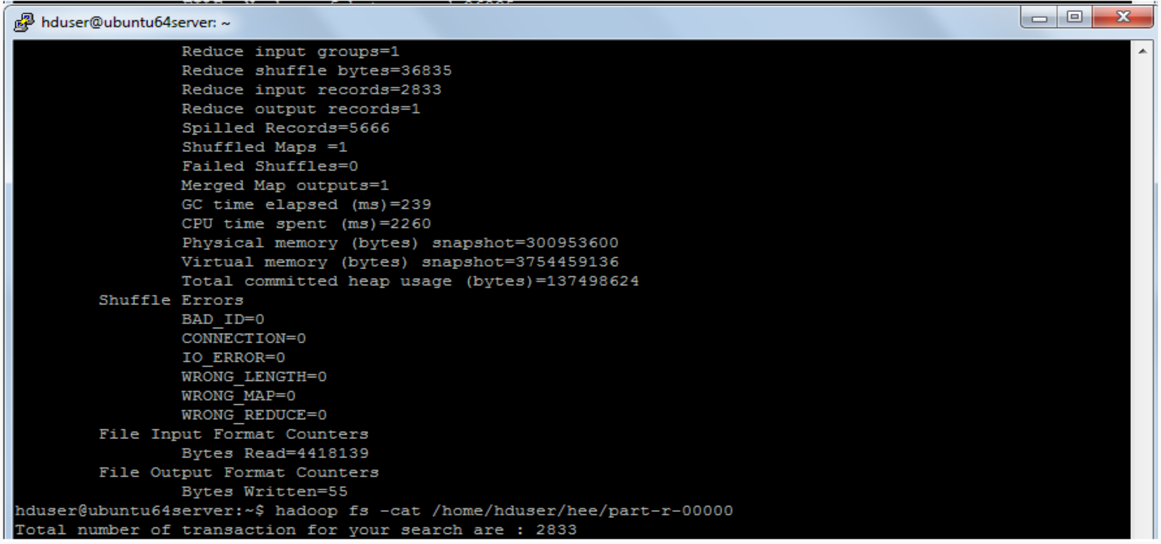
An analysis is required to find the number of products sold for a specific range of prices.

Our program contains 7 java files.

The program takes user input for lower and higher limit of price in MyDriver.class



```
        at java.lang.reflect.Method.invoke(Method.java:497)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
hduser@ubuntu64server:~$ hadoop jar CustomT2.jar /home/hduser/Transactional.dat /home/hduser/hee
Use Case 1 : Count All the Transaction between a lower and upper amountt limit
==============================================================================
==============================================================================
Enter the minimum amount  145
Enter the maximum amount  156
16/11/21 11:11:50 INFO client.RMProxy: Connecting to ResourceManager at /192.168.56.123:8032
16/11/21 11:11:51 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Imp
lement the Tool interface and execute your application with ToolRunner to remedy this.
16/11/21 11:11:51 INFO input.FileInputFormat: Total input paths to process : 1
16/11/21 11:11:52 INFO mapreduce.JobSubmitter: number of splits:1
16/11/21 11:11:52 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1479732047736_0033
16/11/21 11:11:52 INFO impl.YarnClientImpl: Submitted application application_1479732047736_0033
16/11/21 11:11:52 INFO mapreduce.Job: The url to track the job: http://ubuntu64server:8088/proxy/applicatio
n_1479732047736_0033/
16/11/21 11:11:52 INFO mapreduce.Job: Running job: job_1479732047736_0033
16/11/21 11:12:01 INFO mapreduce.Job: Job job_1479732047736_0033 running in uber mode : false
16/11/21 11:12:01 INFO mapreduce.Job:  map 0% reduce 0%
16/11/21 11:12:09 INFO mapreduce.Job:  map 100% reduce 0%
16/11/21 11:12:27 INFO mapreduce.Job:  map 100% reduce 100%
16/11/21 11:12:27 INFO mapreduce.Job: Job job_1479732047736_0033 completed successfully
16/11/21 11:12:27 INFO mapreduce.Job: Counters: 49
        File System Counters
```

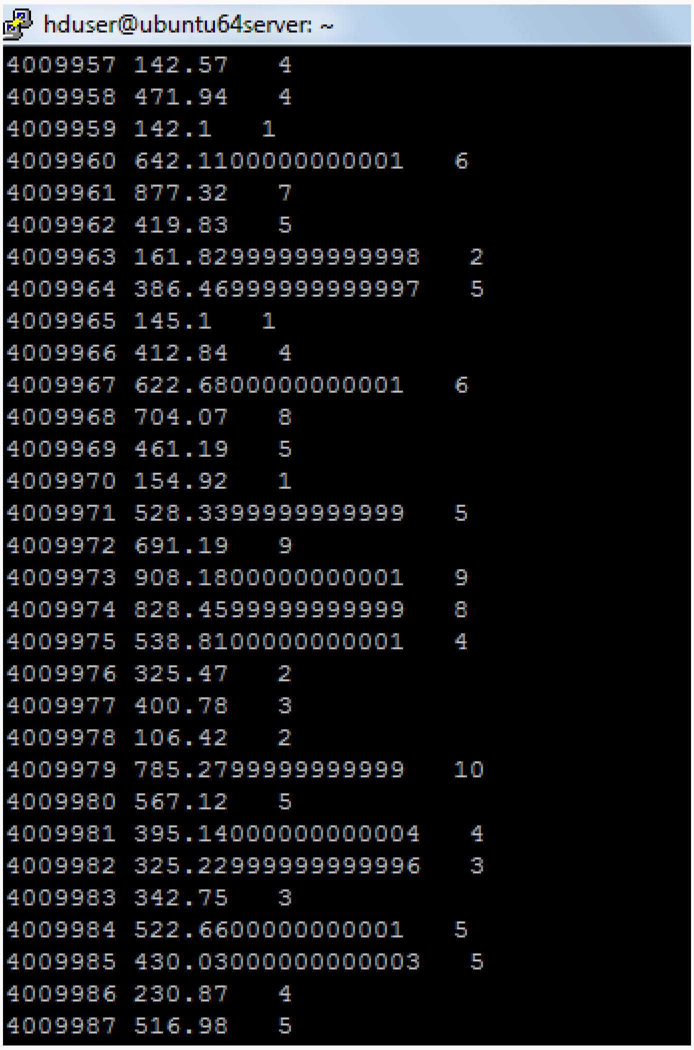


```
                Reduce input groups=1
                Reduce shuffle bytes=36835
                Reduce input records=2833
                Reduce output records=1
                Spilled Records=5666
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=239
                CPU time spent (ms)=2260
                Physical memory (bytes) snapshot=300953600
                Virtual memory (bytes) snapshot=3754459136
                Total committed heap usage (bytes)=137498624
        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=4418139
        File Output Format Counters
                Bytes Written=55
hduser@ubuntu64server:~$ hadoop fs -cat /home/hduser/hee/part-r-00000
Total number of transaction for your search are : 2833
```

Use case 3

The Online Web site is planning for a scheme to give offers to customers based on their past number of transaction and total purchase they did.

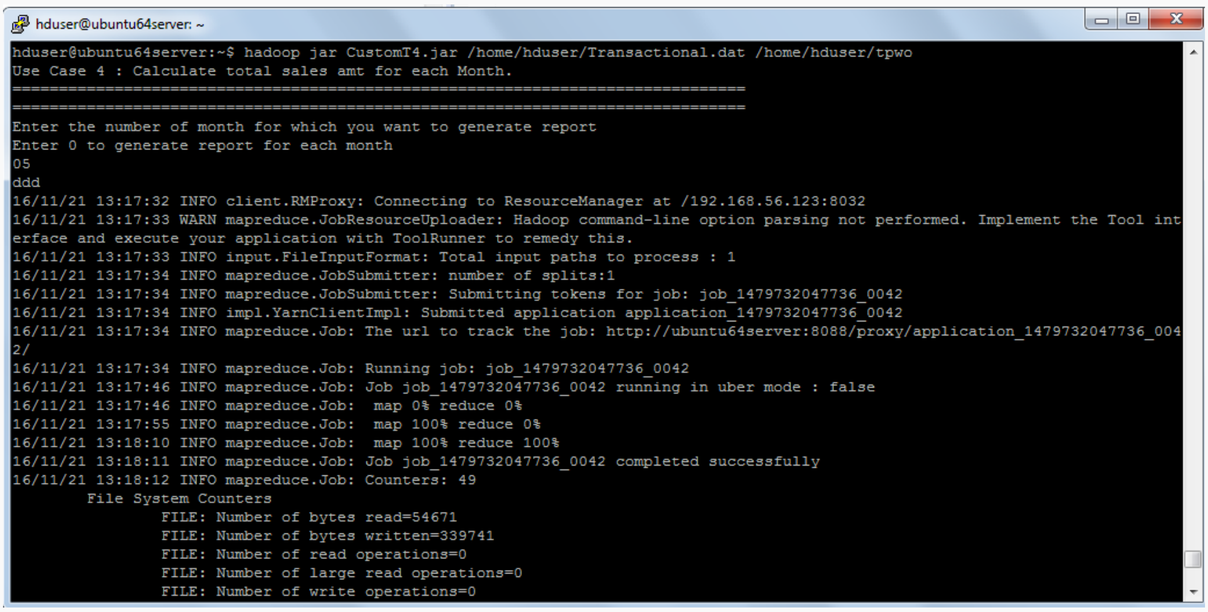
For this an analysis is required to prepare a report for each user.



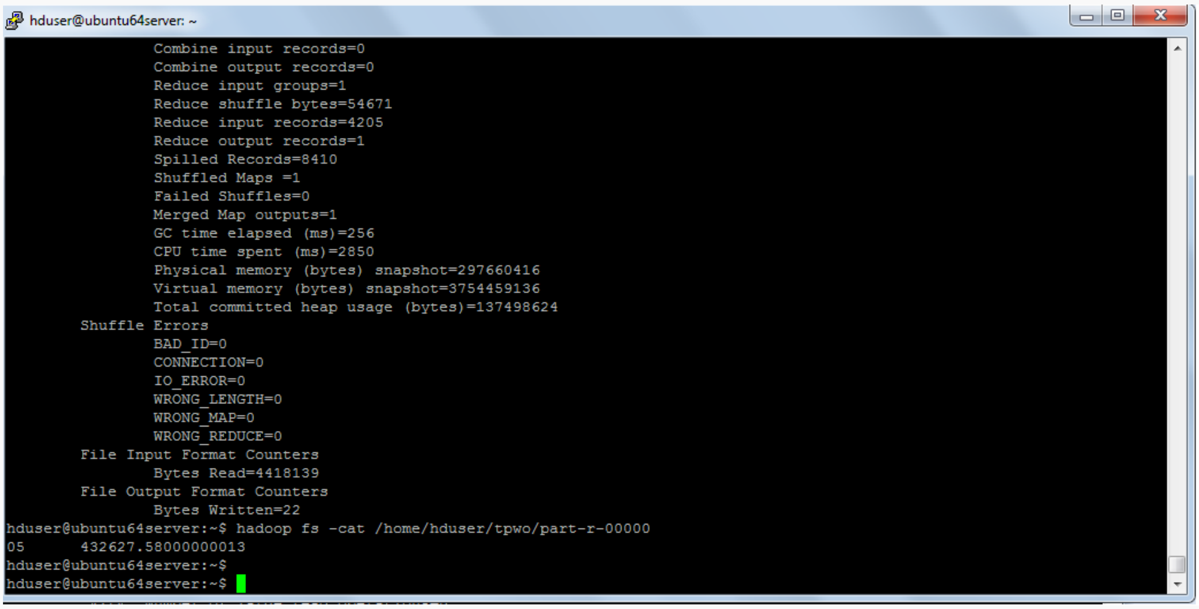
```
4009957 142.57   4
4009958 471.94   4
4009959 142.1   1
4009960 642.1100000000001   6
4009961 877.32   7
4009962 419.83   5
4009963 161.82999999999998   2
4009964 386.46999999999997   5
4009965 145.1   1
4009966 412.84   4
4009967 622.6800000000001   6
4009968 704.07   8
4009969 461.19   5
4009970 154.92   1
4009971 528.3399999999999   5
4009972 691.19   9
4009973 908.1800000000001   9
4009974 828.4599999999999   8
4009975 538.8100000000001   4
4009976 325.47   2
4009977 400.78   3
4009978 106.42   2
4009979 785.2799999999999   10
4009980 567.12   5
4009981 395.14000000000004   4
4009982 325.22999999999996   3
4009983 342.75   3
4009984 522.6600000000001   5
4009985 430.03000000000003   5
4009986 230.87   4
4009987 516.98   5
```

Use case 4

At the end of every year your company wants to do an analysis to know in which month people usually comes for shopping.



```
hduser@ubuntu64server:~$ hadoop jar CustomT4.jar /home/hduser/Transactional.dat /home/hduser/tpwo
Use Case 4 : Calculate total sales amt for each Month.
================================================================================
================================================================================
Enter the number of month for which you want to generate report
Enter 0 to generate report for each month
05
ddd
16/11/21 13:17:32 INFO client.RMProxy: Connecting to ResourceManager at /192.168.56.123:8032
16/11/21 13:17:33 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool int
erface and execute your application with ToolRunner to remedy this.
16/11/21 13:17:33 INFO input.FileInputFormat: Total input paths to process : 1
16/11/21 13:17:34 INFO mapreduce.JobSubmitter: number of splits:1
16/11/21 13:17:34 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1479732047736_0042
16/11/21 13:17:34 INFO impl.YarnClientImpl: Submitted application application_1479732047736_0042
16/11/21 13:17:34 INFO mapreduce.Job: The url to track the job: http://ubuntu64server:8088/proxy/application_1479732047736_004
2/
16/11/21 13:17:34 INFO mapreduce.Job: Running job: job_1479732047736_0042
16/11/21 13:17:46 INFO mapreduce.Job: Job job_1479732047736_0042 running in uber mode : false
16/11/21 13:17:46 INFO mapreduce.Job:  map 0% reduce 0%
16/11/21 13:17:55 INFO mapreduce.Job:  map 100% reduce 0%
16/11/21 13:18:10 INFO mapreduce.Job:  map 100% reduce 100%
16/11/21 13:18:11 INFO mapreduce.Job: Job job_1479732047736_0042 completed successfully
16/11/21 13:18:12 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=54671
                FILE: Number of bytes written=339741
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
```

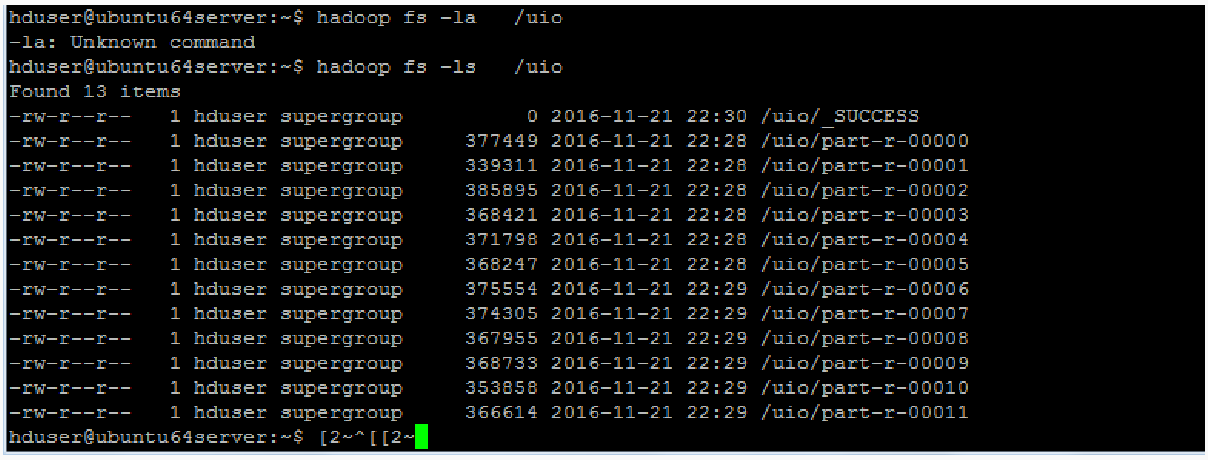


```
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=54671
                Reduce input records=4205
                Reduce output records=1
                Spilled Records=8410
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=256
                CPU time spent (ms)=2850
                Physical memory (bytes) snapshot=297660416
                Virtual memory (bytes) snapshot=3754459136
                Total committed heap usage (bytes)=137498624
        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=4418139
        File Output Format Counters
                Bytes Written=22
hduser@ubuntu64server:~$ hadoop fs -cat /home/hduser/tpwo/part-r-00000
05      432627.58000000013
hduser@ubuntu64server:~$
hduser@ubuntu64server:~$
```

Use Case 5

Being a Hadoop Developer and Admin, You may need to partition your final data to make further processing easy.

How have been asked to work on you company data to divide all the transaction based on the month and store each transaction according to the months. So 12 files are created for this one for each month.

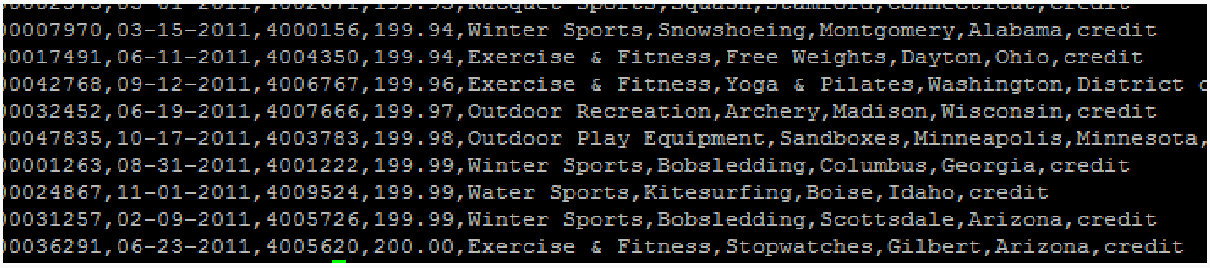


```
hduser@ubuntu64server:~$ hadoop fs -la   /uio
-la: Unknown command
hduser@ubuntu64server:~$ hadoop fs -ls   /uio
Found 13 items
-rw-r--r--   1 hduser supergroup          0 2016-11-21 22:30 /uio/_SUCCESS
-rw-r--r--   1 hduser supergroup     377449 2016-11-21 22:28 /uio/part-r-00000
-rw-r--r--   1 hduser supergroup     339311 2016-11-21 22:28 /uio/part-r-00001
-rw-r--r--   1 hduser supergroup     385895 2016-11-21 22:28 /uio/part-r-00002
-rw-r--r--   1 hduser supergroup     368421 2016-11-21 22:28 /uio/part-r-00003
-rw-r--r--   1 hduser supergroup     371798 2016-11-21 22:28 /uio/part-r-00004
-rw-r--r--   1 hduser supergroup     368247 2016-11-21 22:28 /uio/part-r-00005
-rw-r--r--   1 hduser supergroup     375554 2016-11-21 22:29 /uio/part-r-00006
-rw-r--r--   1 hduser supergroup     374305 2016-11-21 22:29 /uio/part-r-00007
-rw-r--r--   1 hduser supergroup     367955 2016-11-21 22:29 /uio/part-r-00008
-rw-r--r--   1 hduser supergroup     368733 2016-11-21 22:29 /uio/part-r-00009
-rw-r--r--   1 hduser supergroup     353858 2016-11-21 22:29 /uio/part-r-00010
-rw-r--r--   1 hduser supergroup     366614 2016-11-21 22:29 /uio/part-r-00011
hduser@ubuntu64server:~$ [2~^[[2~
```

We get 12 files one for each month. Each file has transaction for that particular month.

Use case 6

Sort the whole file on basis of amount spend.



```
)0007970,03-15-2011,4000156,199.94,Winter Sports,Snowshoeing,Montgomery,Alabama,credit
)0017491,06-11-2011,4004350,199.94,Exercise & Fitness,Free Weights,Dayton,Ohio,credit
)0042768,09-12-2011,4006767,199.96,Exercise & Fitness,Yoga & Pilates,Washington,District c
)0032452,06-19-2011,4007666,199.97,Outdoor Recreation,Archery,Madison,Wisconsin,credit
)0047835,10-17-2011,4003783,199.98,Outdoor Play Equipment,Sandboxes,Minneapolis,Minnesota,
)0001263,08-31-2011,4001222,199.99,Winter Sports,Bobsledding,Columbus,Georgia,credit
)0024867,11-01-2011,4009524,199.99,Water Sports,Kitesurfing,Boise,Idaho,credit
)0031257,02-09-2011,4005726,199.99,Winter Sports,Bobsledding,Scottsdale,Arizona,credit
)0036291,06-23-2011,4005620,200.00,Exercise & Fitness,Stopwatches,Gilbert,Arizona,credit
```

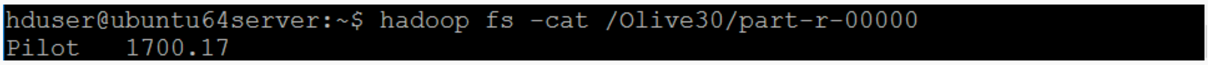
We have the transaction file and this file will be sorted based on the amounts available in each transaction.

Use Case 7

Top Profession who spent most amounts on our products.

Company wants to target the particular area where

People are more interested in their products so we have analysed the top profession.



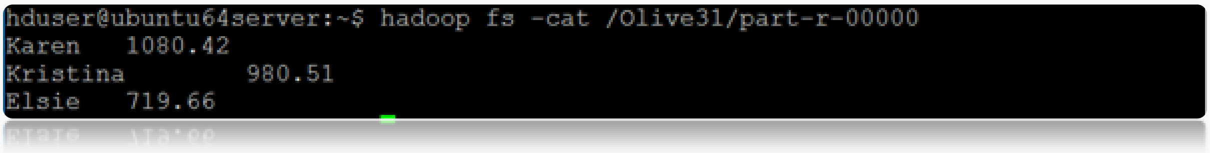
```
hduser@ubuntu64server:~$ hadoop fs -cat /Olive30/part-r-00000
Pilot   1700.17
```

It seems customers who are pilot are doing more transactions.

Use Case 8

Analyze Top 3 customers to give additional rewards.

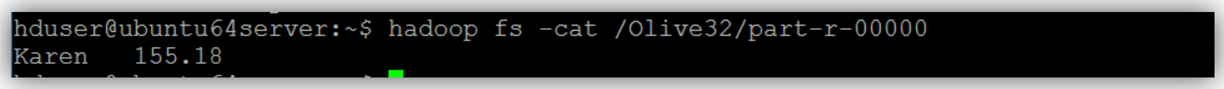
Our online shopping website wants to give rewards to some top 3 customers.



```
hduser@ubuntu64server:~$ hadoop fs -cat /Olive31/part-r-00000
Karen   1080.42
Kristina        980.51
Elsie   719.66
```

Use Case 9

Top customer who spent the most in month of July.



```
hduser@ubuntu64server:~$ hadoop fs -cat /Olive32/part-r-00000
Karen   155.18
```

Karen is the top customer who spent the most for online shipping.