

Data Analysis on Socio-Economic and Well-being Fitness

Objective:

The project is based on Data Analysis for Economy, Seasonality, Product selling, financial exchange, Well-being Fitness and Sports interest. This project would help in analyzing socio-economic status, their wellbeing fitness on various factors like age group and period of time, etc.

Given Data:

Sports & Fitness Equipment sales details and Customer details

Trans id	Date	Cust id	Amount	Category	Item name	City	State	Mode of payment

Cust id	First Name	Last Name	Age	Profession

Analysis on following factors:

1. Economic Status
2. Seasonality
3. Financial Exchanges
4. Well-being Fitness
5. Sports Awareness

1. Economic Status:

Economic Analysis is a powerful way of thinking about the business world. The phenomenal growth of e-commerce is increased access to the internet. Financial Services organizations are using data mined from customer interactions to slice and dice their users into finely tuned segments. It also tells us how much individual, single organization or government is using e-commerce.

For Example: Analyzing Economic status for each age group and each profession on citizenship i.e. finding out the sales happened for person with age groups and profession.

We have achieved the same in our project in various Tasks mentioned below.

Task1: Finding all the transaction where Amount>160.

Task2: Counting all the transaction where amount is between 175 and 200.

Task3: Calculating the total sum, total count, average of all the transaction for each user id.

2. Seasonality

We know that retail sales move with the seasons. Seasonality is a characteristic of a time series in which the data experiences regular and predictable changes that recur every calendar year. Any predictable change or pattern in a time series that recurs or repeats over a one-year period can be said to be seasonal. Retail sales in general always rise in the fourth quarter, but e-commerce sales do so even more intensely.

For Example: Finding the period of time with maximum sales.

We have achieved the same in our project in various Tasks mentioned below.

Task 4: Calculating total sales amount for each Month.

Task 5: Dividing the file into 12 files, where each file contains each month of data. For ex: file 1 should contain data of January transactions, file 2 should contain data of February transactions and soon.

Task 9: Finding the user who has spent the max amount in July month.

3. Financial Exchanges:

E-Commerce or Electronics Commerce sites use cash and electronic payment where electronic payment refers to paperless monetary transactions. Electronic payment has revolutionized the business processing by various factors. This enables these financial institutions to create increasingly relevant and sophisticated offers. Being user friendly and less time consuming than manual processing, helps business organization to expand its market reach / expansion.

For Example: Finding out the persons who made their purchases on various modes of payment by cash, credit card, debit card, Internet Banking NEFT.

We have achieved the same in our project in various Tasks mentioned below.

Task 6: Sorting the whole file on the basis of amt.

Task 7: Finding the name of top 3 spenders.

Task 8: Finding the name of user who has spent the maximum amount.

Future Enhancement: Finding the number of persons who made their purchase based on mode of payment.

4. Well-being Fitness:

We have to focus on Healthcare - both in terms of fitness and diseases outbreak.

Healthcare organizations are turning to data analysis to improve their care delivery and the cost of care. This services to this population to help fight the obesity.

EX: identifying No of persons being physically fit i.e. finding their interest in sports, gyms, yoga exc.

Additional & Future Enhancement:

We can implement the same in our project in various Tasks likes finding the total number of persons based on category of purchase. Grouping the data based on Exercise & Fitness and Gymnastics.

5. Sports Awareness:

Sports and games form an essential part of human resource Development. Sport development is a national priority, as it promotes active Lifestyle, child and youth

development, social inclusiveness, Employment opportunities, peace and development, and above all a Sense of belongingness and national pride.

For example: Finding the number of person who is interest in various sports.

Additional & Future Enhancement:

We can implement the same in our project in various Tasks likes finding the total number of persons who were interest in various sports, we can group the persons in based on their interest. Also we can find the maximum persons who were interested in particular sports.

Technology used:

Apache Hadoop:

Apache Hadoop an open-source software framework used for distributed storage and processing of very large data sets. It consists of computer clusters built from commodity hardware.

Characteristics of Hadoop:

- Reliable
- Flexible
- Economical
- Scalable

Map reduce:

Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

Apache Pig:

Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

Features:

- Ease of programming
- Optimization opportunities
- Extensibility.

Apache Hive:

The Apache Hive data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL. Structure can be projected onto data already in storage

HARDWARE & SOFTWARE REQUIREMENTS:

Operating System	:	LINUX
RAM	:	8GB
System Type	:	64 bit OS
Development Tool	:	Eclipse
Language Used	:	MapReduce Java, Hive, Pig

TASK1: To Find all the transaction on given amount.

INPUT: MAPREDUCE : Data from user:Enter the amount

Data Validation:Yes

```
hduser@ubuntu64server:~$ hadoop jar p1.jar /nov22/trans.dat /f2
Enter your choice:--
1-->Find userid whose purchase amount should be greater than given value
2-->To Count all transaction between the given amount
3-->Calculate sum and count of transaction of each user id
4-->Calculate total sales amt for each Month
5-->To place total sales of each month in different files
6-->Sort the whole file on the basis of amt.
7-->Find the name of top 3 spenders
8-->Find the name of user who has spend the maximum amount
9-->Find the user who has spend the max amount in July month
1
Enter the Amount---
199
```

1.OUTPUT(MAPREDUCE JAVA)

```
hduser@ubuntu64server: ~
user id-->4000698---Amount--> 199.17
user id-->4001145---Amount--> 199.21
user id-->4003204---Amount--> 199.37
user id-->4009181---Amount--> 199.81
user id-->4001178---Amount--> 199.69
user id-->4006367---Amount--> 199.11
user id-->4006767---Amount--> 199.96
user id-->4003016---Amount--> 199.15
user id-->4003093---Amount--> 199.05
user id-->4001051---Amount--> 199.29
user id-->4009311---Amount--> 199.43
user id-->4005503---Amount--> 199.74
user id-->4008144---Amount--> 199.13
user id-->4008188---Amount--> 199.84
user id-->4008525---Amount--> 199.02
user id-->4001490---Amount--> 199.6
user id-->4008334---Amount--> 199.05
user id-->4002736---Amount--> 199.34
user id-->4009382---Amount--> 199.31
user id-->4006933---Amount--> 199.21
user id-->4009465---Amount--> 199.16
user id-->4009693---Amount--> 199.22
user id-->4002624---Amount--> 199.68
user id-->4003035---Amount--> 199.26
user id-->4000521---Amount--> 199.49
user id-->4000695---Amount--> 199.13
user id-->4002060---Amount--> 199.54
user id-->4007135---Amount--> 199.41
user id-->4008313---Amount--> 199.32
user id-->4007299---Amount--> 199.69
user id-->4004125---Amount--> 199.65
user id-->4000519---Amount--> 199.9
user id-->4009239---Amount--> 199.09
user id-->4003783---Amount--> 199.98
user id-->4006138---Amount--> 199.91
user id-->4002136---Amount--> 199.18
user id-->4003050---Amount--> 199.89
user id-->4007911---Amount--> 199.32
user id-->4005705---Amount--> 199.81
user id-->4002687---Amount--> 199.06
user id-->4009559---Amount--> 199.63
user id-->4008746---Amount--> 199.42
user id-->4006098---Amount--> 199.49
user id-->4003750---Amount--> 199.93
user id-->4007255---Amount--> 199.16
user id-->4005076---Amount--> 199.92
user id-->4000454---Amount--> 199.35
user id-->4000978---Amount--> 199.39
user id-->4004318---Amount--> 199.07
hduser@ubuntu64server:~$
```

1.OUTPUT(HIVE)

```
00048177      08-03-2015      4002687 199.06  Jumping Jumping Stilts  Scottsda
le      Arizona credit
00048324      04-26-2015      4009559 199.63  Indoor Games      Darts      Springfi
eld      Illinois      credit
00048673      02-15-2015      4008746 199.42  Jumping Bungee Jumping  Eugene O
region credit
00049043      02-26-2015      4006098 199.49  Water Sports      Kitesurfing  J
ersey City      New Jersey      credit
00049336      05-17-2015      4003750 199.93  Gymnastics      Balance Beams  C
olumbia Missouri      credit
00049472      07-17-2015      4007255 199.16  Exercise & Fitness      Cardio M
achine Accessories      Pasadena      Texas      credit
00049851      04-20-2015      4005076 199.92  Team Sports      Basketball      S
tamford Connecticut      credit
00049939      09-07-2015      4000454 199.35  Games      Portable Electronic Game
s      Newark      New Jersey      credit
00049955      09-13-2015      4000978 199.39  Exercise & Fitness      Jump Rop
es      Birmingham      Alabama      credit
00049980      03-13-2015      4004318 199.07  Outdoor Recreation      Track &
Field      Omaha      Nebraska      credit
Time taken: 33.587 seconds
hive>
```

1.OUTPUT(PIG)

```
cloudera@localhost:~
File Edit View Search Terminal Help
[cloudera@localhost ~]$ pig /home/cloudera/Desktop/mydata/ol
```

```
a = load '/user/cloudera/trans.dat' using PigStorage(',');
b = foreach a generate $2,$3;
c = filter b by $1>=199;
dump c;
```

```
cloudera@localhost:~
File Edit View Search Terminal Help
(4002060,199.54)
(4007135,199.41)
(4008313,199.32)
(4007299,199.69)
(4004125,199.65)
(4000519,199.90)
(4009239,199.09)
(4003783,199.98)
(4006138,199.91)
(4002136,199.18)
(4003050,199.89)
(4007911,199.32)
(4005705,199.81)
(4002687,199.06)
(4009559,199.63)
(4008746,199.42)
(4006098,199.49)
(4003750,199.93)
(4007255,199.16)
(4005076,199.92)
(4000454,199.35)
(4000978,199.39)
(4004318,199.07)
[cloudera@localhost ~]$
```

TASK2: To Count all the transaction where amount is between Given values

2INPUT:MAPREDUCE:DATA from user Enter Max and Min amount:

Data Validation: Yes

```
hduser@ubuntu64server:~$ hadoop jar p1.jar /nov22/trans.dat /f2
Enter your choice:--
1-->Find userid whose purchase amount should be greater than given value
2-->To Count all transaction between the given amount
3-->Calculate sum and count of transaction of each user id
4-->Calculate total sales amt for each Month
5-->To place total sales of each month in different files
6-->Sort the whole file on the basis of amt.
7-->Find the name of top 3 spenders
8-->Find the name of user who has spend the maximum amount
9-->Find the user who has spend the max amount in July month
2
Enter the Min amount
190
Enter the Max amount
199
16/11/24 00:30:37 INFO client.RMProxy: Connecting to ResourceManager at /192.168
```

2OUTPUT:MAPREDUCE

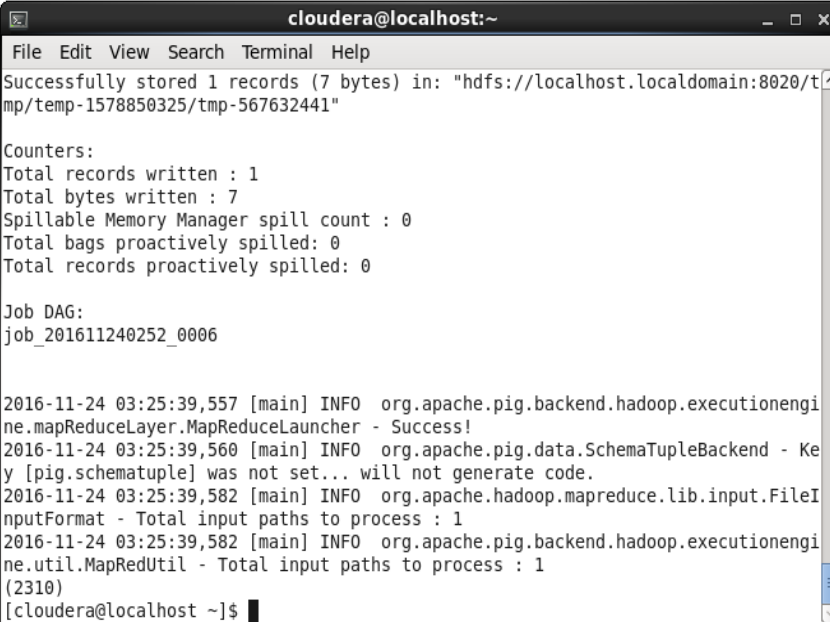
```
hduser@ubuntu64server: ~
Reduce output records=1
Spilled Records=4620
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=240
CPU time spent (ms)=2880
Physical memory (bytes) snapshot=313249792
Virtual memory (bytes) snapshot=3754459136
Total committed heap usage (bytes)=186126336
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=50
hduser@ubuntu64server:~$ hadoop fs -cat /f2/p*
Total transaction between those amount is--> 2310
hduser@ubuntu64server:~$
```


2 OUTPUT:HIVE

```
ite: 5 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 770 msec
OK
2310
Time taken: 25.386 seconds
hive> █
```

2 OUTPUT :PIG

```
A = load '/user/cloudera/trans.dat' using PigStorage(',') as (tid, d, uid, amt : double , cat, prod,city,state,pt);
B = foreach A generate uid, amt;
C = filter B by ($1>190 and $1<199);
D = foreach C generate 1 as one;
E = group D by one;
F = foreach E generate COUNT(D.one);
dump F;
```



```
Successfully stored 1 records (7 bytes) in: "hdfs://localhost.localdomain:8020/tmp/temp-1578850325/tmp-567632441"

Counters:
Total records written : 1
Total bytes written : 7
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201611240252_0006

2016-11-24 03:25:39,557 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2016-11-24 03:25:39,560 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-11-24 03:25:39,582 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2016-11-24 03:25:39,582 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(2310)
[cloudera@localhost ~]$ █
```

TASK3: Calculate the total sum, count and average of all the transaction for given user id.

INPUT :MAPREDUCE :DATA from user : Enter the userid

Data Validation: Yes

```
hduser@ubuntu64server: ~  
WRONG_LENGTH=0  
WRONG_MAP=0  
WRONG_REDUCE=0  
File Input Format Counters  
Bytes Read=4418139  
File Output Format Counters  
Bytes Written=50  
hduser@ubuntu64server:~$ hadoop fs -cat /f2/p*  
Total transaction between those amount is--> 2310  
hduser@ubuntu64server:~$  
hduser@ubuntu64server:~$ hadoop jar p1.jar /nov22/trans.dat /f2  
Enter your choice:--  
1-->Find userid whose purchase amount should be greater than given value  
2-->To Count all transaction between the given amount  
3-->Calculate sum and count of transaction of each user id  
4-->Calculate total sales amt for each Month  
5-->To place total sales of each month in different files  
6-->Sort the whole file on the basis of amt.  
7-->Find the name of top 3 spenders  
8-->Find the name of user who has spend the maximum amount  
9-->Find the user who has spend the max amount in July month  
3  
Enter the user id  
4009465
```

3OUTPUT:Mapreduce

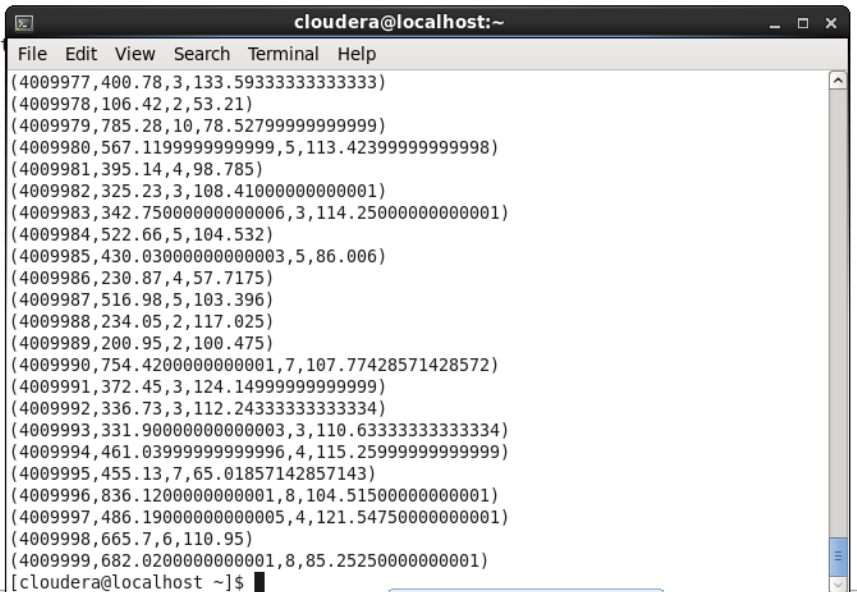
```
hduser@ubuntu64server: ~  
Reduce output records=1  
Spilled Records=6  
Shuffled Maps =1  
Failed Shuffles=0  
Merged Map outputs=1  
GC time elapsed (ms)=228  
CPU time spent (ms)=2340  
Physical memory (bytes) snapshot=312946688  
Virtual memory (bytes) snapshot=3754459136  
Total committed heap usage (bytes)=186126336  
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=74  
hduser@ubuntu64server:~$ hadoop fs -cat /f2/p*  
user id-->4009465--sum--> 513.02-->count 3--average--> 171.00666666666666  
hduser@ubuntu64server:~$
```

3OUTPUT:HIVE

```
hive> Select sum(amt),count(uid),avg(amt),uid from trans group by uid having uid
=4009465;
OK
513.02 3 171.00666666666666 4009465
Time taken: 25.95 seconds
hive>
```

3OUTPUT:PIG

```
A = load '/user/cloudera/trans.dat' using PigStorage(',') as (tid, d, uid, amt : double , cat, prod,city,state,pt);
B = foreach A generate uid, amt;
C = group B by uid;
D = foreach C generate group,SUM(B.am
dump D;
```

A terminal window titled 'cloudera@localhost:~' with a menu bar (File, Edit, View, Search, Terminal, Help). The window displays the output of a Pig script, showing a list of 20 rows of data. Each row contains a uid, an amount, and a group identifier. The output is as follows:
(4009977,400.78,3,133.59333333333333)
(4009978,106.42,2,53.21)
(4009979,785.28,10,78.52799999999999)
(4009980,567.11999999999999,5,113.42399999999998)
(4009981,395.14,4,98.785)
(4009982,325.23,3,108.41000000000001)
(4009983,342.75000000000006,3,114.25000000000001)
(4009984,522.66,5,104.532)
(4009985,430.03000000000003,5,86.006)
(4009986,230.87,4,57.7175)
(4009987,516.98,5,103.396)
(4009988,234.05,2,117.025)
(4009989,200.95,2,100.475)
(4009990,754.4200000000001,7,107.77428571428572)
(4009991,372.45,3,124.14999999999999)
(4009992,336.73,3,112.24333333333334)
(4009993,331.90000000000003,3,110.63333333333334)
(4009994,461.03999999999996,4,115.25999999999999)
(4009995,455.13,7,65.01857142857143)
(4009996,836.1200000000001,8,104.51500000000001)
(4009997,486.19000000000005,4,121.54750000000001)
(4009998,665.7,6,110.95)
(4009999,682.0200000000001,8,85.25250000000001)
[cloudera@localhost ~]\$

TASK4: Calculate total sales amt for each Month.

User Data: Enter the month

Data Validation: Yes

INPUT: MAPREDUCE

```

hduser@ubuntu64server:~$ hadoop jar p1.jar /nov22/trans.dat /f10
Enter your choice:--
1-->Find userid whose purchase amount should be greater than given value
2-->To Count all transaction between the given amount
3-->Calculate sum and count of transaction of each user id
4-->Calculate total sales amt for each Month
5-->To place total sales of each month in different files
6-->Sort the whole file on the basis of amt.
7-->Find the name of top 3 spenders
8-->Find the name of user who has spend the maximum amount
9-->Find the user who has spend the max amount in July month
4
Enter the Month
4

```

4OUTPUT:MAPREDUCE

```

Bytes Written=22
hduser@ubuntu64server:~$ hadoop fs -cat /f10/p*
04      420695.24000000075
hduser@ubuntu64server:~$
hduser@ubuntu64server:~$

```

4OUTPUT:PIG

```

pigtasks x o1 x
a = load '/user/cloudera/trans.dat' using PigStorage(',') as (tid,tdate:chararray,uid,amt:double,cat,acc,city,state,pay);
b = foreach a generate SUBSTRING(tdate,0,2) as mon, a;
c = group b by mon;
d = foreach c generate group, SUM(b.amt) as sum;
Dump d;

```

```

cloudera@localhost:~
File Edit View Search Terminal Help
job_201611240252_0013

2016-11-24 03:49:23,042 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2016-11-24 03:49:23,045 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-11-24 03:49:23,065 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2016-11-24 03:49:23,065 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(01,438165.7599999988)
(02,395262.3699999991)
(03,444664.2399999998)
(04,420695.2400000012)
(05,432627.5799999984)
(06,421074.55000000197)
(07,439560.8000000005)
(08,434255.01000000205)
(09,429321.6299999997)
(10,424856.28000000014)
(11,408846.34999999864)
(12,421490.7299999994)
[cloudera@localhost ~]$

```

TASK5: Dividing the file into 12 files, each file containing each month of data.

5INPUT: MAPREDUCE : Data from user: Enter the month:

Data Validation: NA

```
hduser@ubuntu64server:~$ hadoop jar p1.jar /nov22/trans.dat /f11
Enter your choice:--
1-->Find userid whose purchase amount should be greater than given value
2-->To Count all transaction between the given amount
3-->Calculate sum and count of transaction of each user id
4-->Calculate total sales amt for each Month
5-->To place total sales of each month in different files
6-->Sort the whole file on the basis of amt.
7-->Find the name of top 3 spenders
8-->Find the name of user who has spend the maximum amount
9-->Find the user who has spend the max amount in July month
5
```

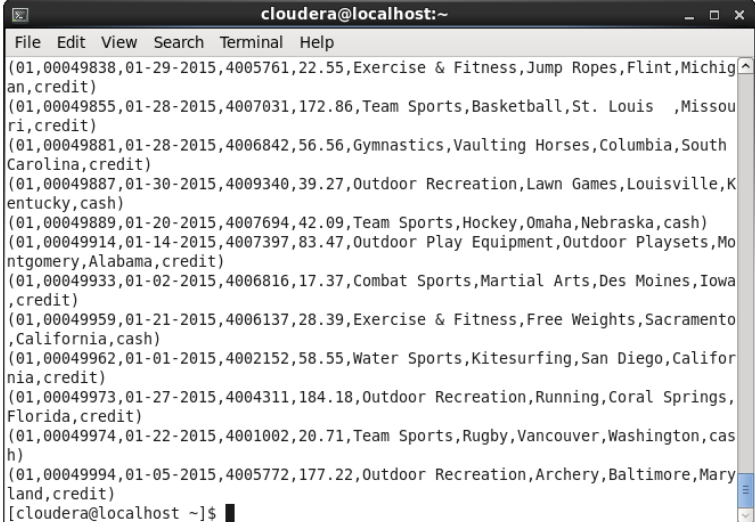
5OUTPUT:MAPREDUCE

```
hduser@ubuntu64server:~$ hadoop fs -ls /f11
Found 13 items
-rw-r--r-- 1 hduser supergroup 0 2016-11-24 01:00 /f11/_SUCCESS
-rw-r--r-- 1 hduser supergroup 432933 2016-11-24 00:59 /f11/part-r-00000
-rw-r--r-- 1 hduser supergroup 389153 2016-11-24 00:59 /f11/part-r-00001
-rw-r--r-- 1 hduser supergroup 442575 2016-11-24 00:59 /f11/part-r-00002
-rw-r--r-- 1 hduser supergroup 422696 2016-11-24 00:59 /f11/part-r-00003
-rw-r--r-- 1 hduser supergroup 426463 2016-11-24 00:59 /f11/part-r-00004
-rw-r--r-- 1 hduser supergroup 422470 2016-11-24 00:59 /f11/part-r-00005
-rw-r--r-- 1 hduser supergroup 430830 2016-11-24 00:59 /f11/part-r-00006
-rw-r--r-- 1 hduser supergroup 429555 2016-11-24 01:00 /f11/part-r-00007
-rw-r--r-- 1 hduser supergroup 422035 2016-11-24 01:00 /f11/part-r-00008
-rw-r--r-- 1 hduser supergroup 427267 2016-11-24 01:00 /f11/part-r-00009
-rw-r--r-- 1 hduser supergroup 409774 2016-11-24 01:00 /f11/part-r-00010
-rw-r--r-- 1 hduser supergroup 424714 2016-11-24 01:00 /f11/part-r-00011
hduser@ubuntu64server:~$
```

```
ennis Salt Lake City Utah credit
5 00046731 05-16-2011 4002373 128.91 Water Sports S
wimming Anaheim California credit
5 00025007 05-02-2011 4003900 120.62 Exercise & Fitne
ss Stopwatches Lexington Kentucky credit
5 00009184 05-03-2011 4009617 134.06 Team Sports C
urling Durham North Carolina credit
5 00046734 05-14-2011 4000255 189.73 Exercise & Fitne
ss Weightlifting Machines Boston Massachusetts credit
5 00015614 05-23-2011 4001576 184.97 Winter Sports S
nowmobiling Columbia South Carolina credit
5 00017570 05-13-2011 4007333 033.80 Gymnastics B
alance Beams Irving Texas cash
5 00001152 05-19-2011 4005034 045.61 Outdoor Recreati
on Equestrian Portland Oregon cash
5 00012862 05-31-2011 4009007 110.53 Exercise & Fitne
ss Weightlifting Machines Everett Washington credit
5 00020746 05-05-2011 4000389 054.42 Outdoor Recreati
on Ice Climbing Jersey City New Jersey credit
5 00009190 05-14-2011 4004470 099.20 Outdoor Play Equ
ipment Slides Paterson New Jersey credit
5 00016279 05-03-2011 4005730 134.61 Games Mahjong
Scottsdale Arizona credit
hduser@ubuntu64server:~$
```

SOUTPUT: PIG

```
a = load '/user/cloudera/trans.dat' using PigStorage(',') as (tid,tdate:chararray,uid,amt:double,cat,acc,city,state,pay);
b = foreach a generate SUBSTRING(tdate,0,2) as month;
c = filter b by month=='01';
dump c;
```



```
(01,00049838,01-29-2015,4005761,22.55,Exercise & Fitness,Jump Ropes,Flint,Michigan,credit)
(01,00049855,01-28-2015,4007031,172.86,Team Sports,Basketball,St. Louis ,Missouri,credit)
(01,00049881,01-28-2015,4006842,56.56,Gymnastics,Vaulting Horses,Columbia,South Carolina,credit)
(01,00049887,01-30-2015,4009340,39.27,Outdoor Recreation,Lawn Games,Louisville,Kentucky,cash)
(01,00049889,01-20-2015,4007694,42.09,Team Sports,Hockey,Omaha,Nebraska,cash)
(01,00049914,01-14-2015,4007397,83.47,Outdoor Play Equipment,Outdoor Playsets,Montgomery,Alabama,credit)
(01,00049933,01-02-2015,4006816,17.37,Combat Sports,Martial Arts,Des Moines,Iowa,credit)
(01,00049959,01-21-2015,4006137,28.39,Exercise & Fitness,Free Weights,Sacramento ,California,cash)
(01,00049962,01-01-2015,4002152,58.55,Water Sports,Kitesurfing,San Diego,California,credit)
(01,00049973,01-27-2015,4004311,184.18,Outdoor Recreation,Running,Coral Springs,Florida,credit)
(01,00049974,01-22-2015,4001002,20.71,Team Sports,Rugby,Vancouver,Washington,cash)
(01,00049994,01-05-2015,4005772,177.22,Outdoor Recreation,Archery,Baltimore,Maryland,credit)
[cloudera@localhost ~]$
```

TASK6: Sort the whole file on the basis of amt.

DATA Validation:NA

6INPUT: MAPREDUCE

```
hduser@ubuntu64server:~$ hadoop jar p1.jar /nov22/trans.dat /f12
Enter your choice:--
1-->Find userid whose purchase amount should be greater than given value
2-->To Count all transaction between the given amount
3-->Calculate sum and count of transaction of each user id
4-->Calculate total sales amt for each Month
5-->To place total sales of each month in different files
6-->Sort the whole file on the basis of amt.
7-->Find the name of top 3 spenders
8-->Find the name of user who has spend the maximum amount
9-->Find the user who has spend the max amount in July month
6
```

6 OUTPUT: MAPREDUCE

```
199.93 00049336,05-17-2011,4003750,199.93,Gymnastics,Balance Beams,Columbia,Mis  
souri,credit  
199.93 00002373,05-01-2011,4002671,199.93,Racquet Sports,Squash,Stamford,Connec  
ticut,credit  
199.94 00007970,03-15-2011,4000156,199.94,Winter Sports,Snowshoeing,Montgomery,  
Alabama,credit  
199.94 00017491,06-11-2011,4004350,199.94,Exercise & Fitness,Free Weights,Dayto  
n,Ohio,credit  
199.96 00042768,09-12-2011,4006767,199.96,Exercise & Fitness,Yoga & Pilates,Was  
hington,District of Columbia,credit  
199.97 00032452,06-19-2011,4007666,199.97,Outdoor Recreation,Archery,Madison,Wi  
sconsin,credit  
199.98 00047835,10-17-2011,4003783,199.98,Outdoor Play Equipment,Sandboxes,Minn  
neapolis,Minnesota,credit  
199.99 00001263,08-31-2011,4001222,199.99,Winter Sports,Bobsledding,Columbus,Ge  
orgia,credit  
199.99 00024867,11-01-2011,4009524,199.99,Water Sports,Kitesurfing,Boise,Idaho,  
credit  
199.99 00031257,02-09-2011,4005726,199.99,Winter Sports,Bobsledding,Scottsdale,  
Arizona,credit  
200.0 00036291,06-23-2011,4005620,200.00,Exercise & Fitness,Stopwatches,Gilber  
t,Arizona,credit  
hduser@ubuntu64server:~$
```

6 OUTPUT: PIG

```
a = load '/user/cloudera/trans.dat' using PigStorage(',') as (tid,tdate:chararray,uid,amt:double,cat,acc,city,state,pay);  
b = foreach a generate tid,tdate,uid,amt,cat,acc,city;  
c = order b by amt;  
dump c;
```

```
cloudera@localhost:~  
File Edit View Search Terminal Help  
(00019383,08-25-2015,4004377,199.92,Gymnastics,Gymnastics Protective Gear,Orange  
,California,credit)  
(00049336,05-17-2015,4003750,199.93,Gymnastics,Balance Beams,Columbia,Missouri,c  
redit)  
(00002373,05-01-2015,4002671,199.93,Racquet Sports,Squash,Stamford,Connecticut,c  
redit)  
(00017491,06-11-2015,4004350,199.94,Exercise & Fitness,Free Weights,Dayton,Ohio,  
credit)  
(00007970,03-15-2015,4000156,199.94,Winter Sports,Snowshoeing,Montgomery,Alabama  
,credit)  
(00042768,09-12-2015,4006767,199.96,Exercise & Fitness,Yoga & Pilates,Washington  
,District of Columbia,credit)  
(00032452,06-19-2015,4007666,199.97,Outdoor Recreation,Archery,Madison,Wisconsin  
,credit)  
(00047835,10-17-2015,4003783,199.98,Outdoor Play Equipment,Sandboxes,Minneapolis  
,Minnesota,credit)  
(00024867,11-01-2015,4009524,199.99,Water Sports,Kitesurfing,Boise,Idaho,credit)  
(00031257,02-09-2015,4005726,199.99,Winter Sports,Bobsledding,Scottsdale,Arizona  
,credit)  
(00001263,08-31-2015,4001222,199.99,Winter Sports,Bobsledding,Columbus,Georgia,c  
redit)  
(00036291,06-23-2015,4005620,200.0,Exercise & Fitness,Stopwatches,Gilbert,Arizon  
a,credit)  
[cloudera@localhost ~]$
```


TASK7: Find the name of top 3 spenders.

Data Validation: NA

7INPUT:MAPREDUCE

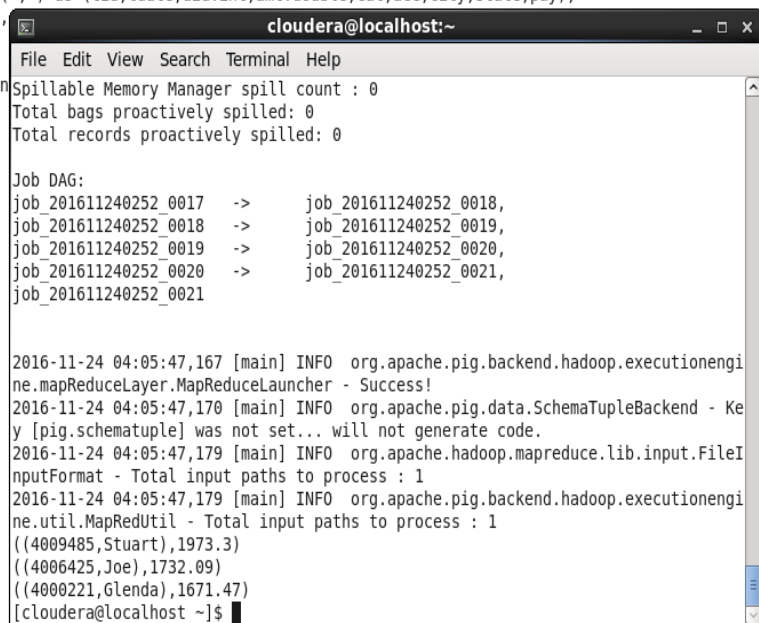
```
at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
hduser@ubuntu64server:~$
hduser@ubuntu64server:~$ hadoop jar p1.jar /nov22/cus.dat /nov22/trans.dat /f13
Enter your choice:--
1-->Find userid whose purchase amount should be greater than given value
2-->To Count all transaction between the given amount
3-->Calculate sum and count of transaction of each user id
4-->Calculate total sales amt for each Month
5-->To place total sales of each month in different files
6-->Sort the whole file on the basis of amt.
7-->Find the name of top 3 spenders
8-->Find the name of user who has spend the maximum amount
9-->Find the user who has spend the max amount in July month
7
```

7OUTPUT:MAPREDUCE

```
hduser@ubuntu64server:~$
hduser@ubuntu64server:~$ hadoop fs -cat /f39/p*
Ted      16991.869999999995
Calvin   16891.920000000006
Gretchen 16762.39
hduser@ubuntu64server:~$
```

7OUTPUT:PIG

```
a = load '/user/cloudera/trans.dat' using PigStorage(',') as (tid,tdate,uid:int,amt:double,cat,acc,city,state,pay);
b = load '/user/cloudera/cus.dat' using PigStorage(',') as (uid:int,fname:chararray,lname,age,prof);
c = join a by uid,b by uid;
d = foreach c generate $2 as uid, $3 as amt,$10 as fname;
e = group d by (uid,fname);
f = foreach e generate group, SUM(d.amt) as Total;
g = order f by Total DESC;
h = limit g 3;
dump h;
```



The screenshot shows a terminal window titled 'cloudera@localhost:~'. It displays the output of a Pig script execution. The top part shows status messages: 'Spillable Memory Manager spill count: 0', 'Total bags proactively spilled: 0', and 'Total records proactively spilled: 0'. Below this is the 'Job DAG' section, which lists five jobs and their dependencies. The bottom part of the screenshot shows log messages from the Pig execution engine, including 'Success!' and information about input paths and schema tuples.

```
cloudera@localhost:~$
File Edit View Search Terminal Help
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201611240252_0017 -> job_201611240252_0018,
job_201611240252_0018 -> job_201611240252_0019,
job_201611240252_0019 -> job_201611240252_0020,
job_201611240252_0020 -> job_201611240252_0021,
job_201611240252_0021

2016-11-24 04:05:47,167 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2016-11-24 04:05:47,170 [main] INFO org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2016-11-24 04:05:47,179 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2016-11-24 04:05:47,179 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
((4009485,Stuart),1973.3)
((4006425,Joe),1732.09)
((4000221,Glenda),1671.47)
[cloudera@localhost ~]$
```


TASK8: To Find the name of user who has spend the maximum amount.

Data Validation: NA

8INPUT: Map reduce

```
hduser@ubuntu64server:~$ hadoop jar p1.jar /nov22/trans.dat /f37
Enter your choice:--
1-->Find userid whose purchase amount should be greater than given value
2-->To Count all transaction between the given amount
3-->Calculate sum and count of transaction of each user id
4-->Calculate total sales amt for each Month
5-->To place total sales of each month in different files
6-->Sort the whole file on the basis of amt.
7-->Find the name of top 3 spenders
8-->Find the name of user who has spend the maximum amount
9-->Find the user who has spend the max amount in July month
8
```

8OUTPUT: MAP reduce

```
hduser@ubuntu64server:~$ hadoop fs -cat /f42/p*
Ted      16991.869999999995
hduser@ubuntu64server:~$
```

TASK9: To find the user who has spent the max amount in July month

9INPUT: Map Reduce

```
hduser@ubuntu64server:~$ hadoop jar p1.jar /nov22/trans.dat /f37
Enter your choice:--
1-->Find userid whose purchase amount should be greater than given value
2-->To Count all transaction between the given amount
3-->Calculate sum and count of transaction of each user id
4-->Calculate total sales amt for each Month
5-->To place total sales of each month in different files
6-->Sort the whole file on the basis of amt.
7-->Find the name of top 3 spenders
8-->Find the name of user who has spend the maximum amount
9-->Find the user who has spend the max amount in July month
9
```

9 OUTPUT: Map Reduce

```
hduser@ubuntu64server:~$ hadoop fs -cat /f37/p*
Toni      2082.44
hduser@ubuntu64server:~$
```

Additional:

Task1: Finding the total count based on category of items

Time taken: 33.764 seconds

```
hive> select 'Category-->',cate,'Total Count-->',count(*) from trans group by cate;
```

Output:Hive

```
OK
Category-->    Air Sports      Total Count-->  960
Category-->    Combat Sports   Total Count--> 1630
Category-->    Dancing Total Count-->  414
Category-->    Exercise & Fitness Total Count--> 7394
Category-->    Games      Total Count--> 3666
Category-->    Gymnastics    Total Count--> 3196
Category-->    Indoor Games  Total Count--> 2799
Category-->    Jumping Total Count-->  2015
Category-->    Outdoor Play Equipment Total Count--> 2910
Category-->    Outdoor Recreation Total Count--> 8383
Category-->    Puzzles Total Count-->  612
Category-->    Racquet Sports Total Count--> 1611
Category-->    Team Sports    Total Count--> 6010
Category-->    Water Sports   Total Count--> 5219
Category-->    Winter Sports  Total Count--> 3181
Time taken: 25.286 seconds
hive>
```

Details of person who bought items under the category gym and exercise:

```
hive>
> select * from trans where cate='Gymnastics' or cate='Exercise & Fitness';
Total MapReduce jobs = 1
```

Hive:Output:

00049839	07-20-2015	4006048	138.55	Exercise & Fitness	Gym Mats	Lowell	Massachusetts	credit	
00049840	08-17-2015	4001167	148.1	Exercise & Fitness	Exercise Balls	Miami	Florida	credit	
00049847	06-09-2015	4007535	105.81	Exercise & Fitness	Cardio Machine Accessories		Portland		Oregon credit
00049874	04-10-2015	4002731	101.75	Exercise & Fitness	Medicine Balls	Columbia		South Carolina	credit
00049876	10-04-2015	4001459	175.82	Exercise & Fitness	Abdominal Equipment		Philadelphia		Pennsylvania credit
00049878	03-22-2015	4004231	95.46	Exercise & Fitness	Weightlifting Machine Accessories		Plano		Texas credit
00049881	01-28-2015	4006842	56.56	Gymnastics	Vaulting Horses	Columbia		South Carolina	credit
00049883	03-10-2015	4003041	165.54	Gymnastics	Gymnastics Rings	Baltimore		Maryland	credit
00049884	08-09-2015	4007271	136.24	Exercise & Fitness	Weightlifting Belts		Jackson		Mississippi credit
00049886	08-15-2015	4008962	22.37	Exercise & Fitness	Exercise Bands	Boise		Idaho	cash
00049888	04-19-2015	4008433	124.88	Gymnastics	Vaulting Horses	Jersey City		New Jersey	credit
00049892	10-04-2015	4000927	196.09	Gymnastics	Gymnastics Mats	Eugene		Oregon	credit
00049895	08-06-2015	4008982	159.37	Exercise & Fitness	Weight Benches	San Diego		California	credit
00049900	06-27-2015	4000158	148.96	Exercise & Fitness	Exercise Balls	Louisville		Kentucky	credit
00049908	05-16-2015	4006924	177.9	Gymnastics	Gymnastics Rings		Coral Springs		Florida credit
00049913	05-08-2015	4002281	48.14	Exercise & Fitness	Exercise Balls	Boston		Massachusetts	credit
00049915	10-07-2015	4002030	141.12	Exercise & Fitness	Free Weights	Salem		Oregon	credit
00049921	11-16-2015	4000172	198.33	Gymnastics	Gymnastics Rings	Boise		Idaho	credit
00049943	02-26-2015	4008385	163.0	Exercise & Fitness	Weightlifting Machine Accessories			Gresham	Oregon credit
00049945	09-02-2015	4006965	117.9	Gymnastics	Vaulting Horses	Louisville		Kentucky	
00049950	09-30-2015	4007935	137.99	Exercise & Fitness	Weightlifting Belts		Huntsville		Alabama credit
00049955	09-13-2015	4000978	199.39	Exercise & Fitness	Jump Ropes	Birmingham		Alabama	credit
00049958	02-14-2015	4009456	37.49	Gymnastics	Gymnastics Rings		Sunnyvale		California cash
00049959	01-21-2015	4006137	28.39	Exercise & Fitness	Free Weights	Sacramento		California	cash
00049960	06-08-2015	4004939	198.32	Gymnastics	Springboards	Cincinnati		Ohio	credit
00049967	12-28-2015	4005452	101.34	Exercise & Fitness	Abdominal Equipment		Centennial		Colorado credit
00049968	06-16-2015	4002061	175.61	Gymnastics	Gymnastics Protective Gear		Lexington		Kentucky credit
00049990	08-10-2015	4002940	144.91	Exercise & Fitness	Medicine Balls	Minneapolis		Minnesota	credit
00049991	04-25-2015	4003685	191.29	Gymnastics	Pommel Horses	Santa Ana		California	credit
00049998	10-23-2015	4007843	180.41	Gymnastics	Vaulting Horses	Berkeley		California	credit

Time taken: 24.47 seconds

Task: Details of person in Sports:

```
hive> select * from trans where cate like '%Sports%';
```

Total MapReduce jobs = 1

00049918	03-05-2015	4006138	64.81	Team Sports	Rugby Durham	North Carolina	credit		
00049922	12-26-2015	4002856	23.88	Water Sports	Bodyboarding	Columbus	Ohio	credit	
00049923	03-18-2015	4003487	53.07	Water Sports	Windsurfing	Reno Nevada	credit		
00049924	09-02-2015	4005005	169.85	Water Sports	Bodyboarding	Des Moines	Iowa	credit	
00049925	06-18-2015	4000742	13.97	Team Sports	Basketball	New Orleans	Louisiana		credit
00049926	05-30-2015	4000041	178.06	Water Sports	Swimming	Madison Wisconsin		credit	
00049933	01-02-2015	4006816	17.37	Combat Sports	Martial Arts	Des Moines	Iowa	credit	
00049937	06-11-2015	4008000	122.83	Team Sports	Baseball	Columbus	Georgia	credit	
00049938	03-13-2015	4003301	136.85	Water Sports	Towed Water Sports	Louisville	Kentucky		credit
00049942	09-21-2015	4008179	158.69	Water Sports	Boating Springfield	Illinois		credit	
00049946	02-24-2015	4003737	38.15	Team Sports	Indoor Volleyball	Denver	Colorado		cash
00049947	08-29-2015	4002944	58.52	Water Sports	Surfing St. Petersburg	Florida	credit		
00049948	02-22-2015	4005290	74.77	Winter Sports	Luge Santa Ana	California		credit	
00049949	03-03-2015	4007112	189.15	Team Sports	Field Hockey	Boise	Idaho	credit	
00049951	02-10-2015	4006798	79.95	Combat Sports	Boxing Montgomery	Alabama	credit		
00049956	07-28-2015	4001371	187.72	Winter Sports	Snowboarding	Orange	California	credit	
00049962	01-01-2015	4002152	58.55	Water Sports	Kitesurfing	San Diego	California		credit
00049963	05-27-2015	4003448	42.88	Winter Sports	Bobsledding	Eugene	Oregon	cash	
00049969	10-19-2015	4002286	121.81	Winter Sports	Cross-Country Skiing	Coral Springs	Florida	credit	
00049972	05-21-2015	4008556	10.26	Water Sports	Life Jackets	Irving	Texas	cash	
00049974	01-22-2015	4001002	20.71	Team Sports	Rugby Vancouver	Washington		cash	
00049979	06-21-2015	4009736	66.19	Winter Sports	Snowshoeing	Cincinnati	Ohio	credit	
00049982	02-12-2015	4007202	129.43	Team Sports	Field Hockey	Dayton	Ohio	credit	
00049983	06-13-2015	4000024	7.44	Team Sports	Rugby Santa Ana	California		cash	
00049986	08-29-2015	4008092	156.38	Winter Sports	Sledding	Salem	Oregon	credit	
00049988	08-13-2015	4007405	65.54	Winter Sports	Cross-Country Skiing	Oklahoma City	Oklahoma		credit
00049992	10-31-2015	4002441	139.78	Water Sports	Water Tubing	Lexington	Kentucky	credit	
00049993	06-02-2015	4007367	50.32	Team Sports	Field Hockey	Stamford	Connecticut	credit	
00049997	05-03-2015	4003954	35.85	Racquet Sports	Squash New Orleans	Louisiana		cash	
00049999	12-14-2015	4001406	168.49	Team Sports	Team Handball	Rockford	Illinois	credit	

Time taken: 22.134 seconds
hive> █

Task: Total number of persons interested in Sports:

FAILED: SemanticException [Error 10025]: Line 1:27 Expression not in GROUP BY key cate

```
hive> select 'Category name-->', cate,'Total-->',count(*) from trans where cate like '%Sports%' group by cate;
```

Total MapReduce jobs = 1

```
OK
Category name-->      Air Sports      Total-->      960
Category name-->      Combat Sports   Total-->      1630
Category name-->      Racquet Sports   Total-->      1611
Category name-->      Team Sports      Total-->      6010
Category name-->      Water Sports     Total-->      5219
Category name-->      Winter Sports    Total-->      3181
Time taken: 26.354 seconds
```

Task: : Total number of persons interested in Yoga:

Time taken: 27.000 seconds

```
hive> select 'product name-->',pro,'Total Count-->', count(*) from trans where pro like '%Yoga%' group by pro;
```

```
OK
product name--> Yoga & Pilates  Total Count-->  444
```

Conclusion:

Data Analysis in E-Commerce can enhance various factors in terms of sales, transaction, communication with the customers. The project analyzed the status of Economy, Seasonality, product selling, well-being Fitness, Financial exchange, Sports interest on factors like each age group and various period.