Real-Time Analytics in Ecommerce Transaction

Project Author : Niyaz Ahmed

Date : 22/11/2016

Objective : To analyze the data of the E-Commerce Transaction and Customer Details so that the E-Commerce Company can Analyze them and apply business methodology so that it helps them in future growth in terms of Sales, Customer Satisfaction and many more.

Test Data :

1) Ecommerce Transaction Data (txns-large.dat)

TransID	Date	UserId	Amount	Cat	Machine	Sate	City	Payment

2) Customer Details (custs-large.dat)

UserID	First Name	Last Name	Age	Position

Use Case 1 : Constraint Based Amount Scenario

- To find the product based on the user search or product the user has purchased.
- Whenever user purchases a product of a particular price or within range of amount than at time the user will provide with similar type of product within the same range.
- **Task** included are Task 1 : Find all the transaction where amount >160 and Task 2 : Count all the transaction where amount is between 175 to 200.
- **Data Validation :** Yes.

  Constraint : User Input Can be Only Numbers.

OUTPUT

```
Last login: Mon Nov 21 04:31:42 2016 from 192.168.56.1
hduser@ubuntu64server:~$ hadoop jar Ctask1.jar /17Sep/txns-large.dat/ ct1
Enter the amount
170
```

```
hduser@ubuntu64server:~$ hadoop fs -ls /ct1
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2016-11-21 05:27 /ct1/_SUCCESS
-rw-r--r--   1 hduser supergroup      82866 2016-11-21 05:27 /ct1/part-m-00000
hduser@ubuntu64server:~$ hadoop fs -cat /ct1/p*
```

```
00046831        191.38
00046832        198.49
00046836        199.41
00046842        189.37
00046846        184.89
00046849        186.0
00046860        196.9
```

```
hduser@ubuntu64server:~$ hadoop jar Ctask2.jar /17Sep/txns-large.dat/ ct2
Enter the lower limit
120
Enter the upper limit
140
```

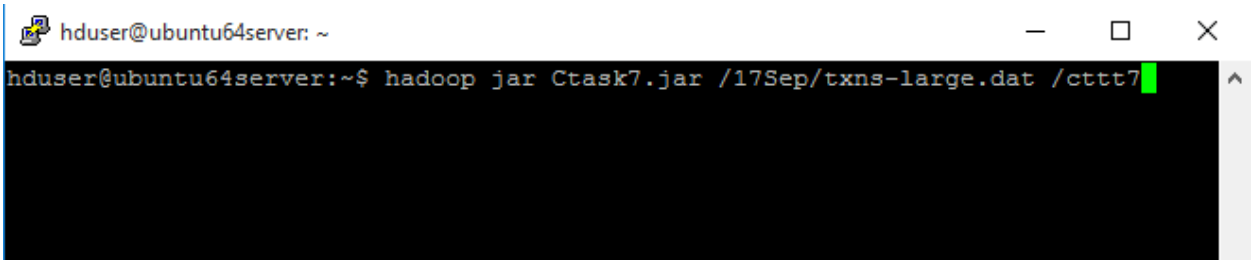
```
hduser@ubuntu64server:~$ hadoop fs -ls /ct2
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2016-11-21 05:44 /ct2/_SUCCESS
-rw-r--r--   1 hduser supergroup          5 2016-11-21 05:44 /ct2/part-r-00000
hduser@ubuntu64server:~$ hadoop fs -cat /ct2/p*
6524
hduser@ubuntu64server:~$
```

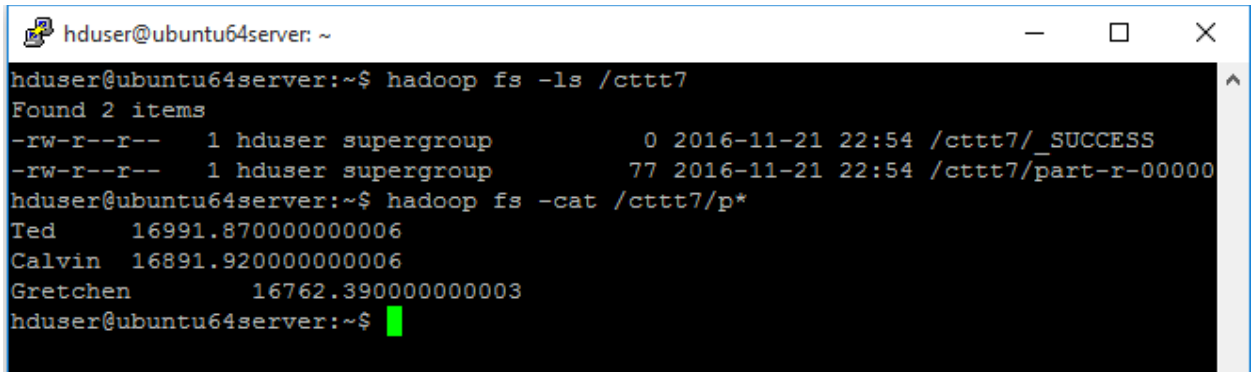
Use Case 2 : Top Three Spender or in a particular month

- During a e-shopping festival many people shop on a particular day so based on amount spend by the customer, the top three customer who will spend maximum money on shopping will be provided with gift hamper.
- The same scenario can be applied when the shopping festival is for the entire month.
- **Tasks** included are Task 7 : Find the name of top 3 spenders

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

OUTPUT





Use Case 3 : Revenue made via transaction

- This Use Case is use by the E-Commerce Company to check the revenue they have generated in a single day, month or any month or a complete year.
- Revenue sorted via them can be categorized on their needs.
- **Tasks** included are Tasks 4 : Calculate total sales amount for each Month and Tasks 5 : Divide the file into each month for all 12 months.
- **Data Validation :** Yes.

  Constraint : User Input Can be Only Numbers for Months

OUTPUT

```
hduser@ubuntu64server:~$ hadoop jar Ctask4.jar /17Sep/txns-large.dat/ ct4
Enter the Months
11
```

```
hduser@ubuntu64server:~$ hadoop fs -ls /ct44
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2016-11-21 09:03 /ct44/_SUCCESS
-rw-r--r--   1 hduser supergroup         22 2016-11-21 09:03 /ct44/part-r-00000
hduser@ubuntu64server:~$ hadoop fs -cat /ct44/p*
12      421490.72999999876
hduser@ubuntu64server:~$
```

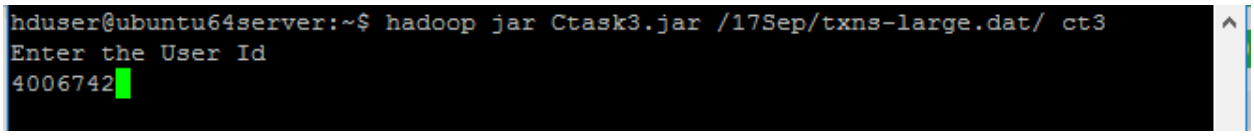
```
hduser@ubuntu64server:~$ hadoop fs -cat /ct55/part-r-00002
03      444664.23999999935
hduser@ubuntu64server:~$ hadoop fs -cat /ct55/p*
01      438165.7600000004
02      395262.3700000005
03      444664.23999999935
04      420695.239999999
05      432627.57999999967
06      421074.54999999976
07      439560.8000000002
08      434255.01000000106
09      429321.63000000105
10      424856.2800000009
11      408846.3499999989
12      421490.7299999997
hduser@ubuntu64server:~$
```

Use Case 4 : Customer Dilemma

- Normally people who do shopping online face some situation regarding the product, delivery or any other problem.
- So in that scenario the customer calls the Customer Care and the people there addresses the problem. For which they ask for Customer Id No so that all the data regarding the user can be fetched and problem regarding the issues can resolved.
- **Data Validation :** Yes.

  Constraint : For Customer Care People Input Can be Only Numbers

OUTPUT



Use Case 5 : Product Promotion and Customer Targeting

- We can do a complete analysis of customer regarding their shopping behavioral pattern
- We can calculated their total sum of money they spent on a particular E-Commerce site, their average transaction and number of visits.
- Depending upon their shopping behavioral pattern we can analyze them and target a particular customer by any new product launch in the market.
- **Task** included is Task 3 to find the total sum, count and average of customers.

OUTPUT :

```
A = load '/user/cloudera/txns-large.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.amt),COUNT(B.amt),AVG(B.amt);
dump D;
```

```
(4009979,785.28,10,78.52799999999999)
(4009980,567.1199999999999,5,113.42399999999998)
(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)
```

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**.**

Java MapReduce Program: 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.

Apache Hive: Apache Hive is data warehouse infrastructure built on top of Apache Hadoop for providing data summarization, ad-hoc query, and analysis of large datasets. It provides a mechanism to project structure onto the data in Hadoop and to query that data using a SQL-like language called HiveQL (HQL).

SOFTWARE Used:

1) Virtual Box

2) Eclipse

3) Ubuntu Terminal (for MapReduce)

4) Cloudera OS (for HIVE)

SYSTEM REQUIREMENT:

- Minimum 50 Gb of HardDrive Space.
- Minimum 4 Gb RAM.
- Next Generation Processor Chips like Intel I3 and so on.

CONCLUSION : Thus a Real Time Analysis of E-Commerce Transaction Data happens which helps in maintaining the needs as per requirements.