NIIT

SAN-JAN DATA ANALYST AND TECH

BIG DATA FOR BETTER SOLUTION

Version *<1.0>*

<22/11/2016>

Author: Janakiraman

Table of Contents

Project Definition

1.1 what is BIG DATA

Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analysed for insights that lead to better decisions and strategic business moves. Big data is changing the way people within organizations work together. It is creating a culture in which business and IT leaders must join forces to realize value from all data. Insights from big data can enable all employees to make better decisions—deepening customer engagement, optimizing operations, preventing threats and fraud, and capitalizing on new sources of revenue. But escalating demand for insights requires a fundamentally new approach to architecture, tools and practices.

1.2 Background

We are provide accurate data solution to the industries who`s generating huge amount data, using various tools and techniques under Big Data technology.

1.3 Business Case

Huge amount of data being generated by everything around us at all times. Every digital process and social media exchange produces it. Industries struggling with handle this amount of data. So we made it as a business to give accurate data solution.

1.4 Project objectives and desired outcomes

San-Jan Data Analyst and Tech is a MNC (Multi National Company) for data analysis and to provide technology solution to the major industries. We uses various scenarios in Big Data to provide desired outcomes.

Here our current project fully based on E-Commerce, there are tons of transaction process, log files and feedback files created. All Industries struggling to handle these huge volume data generated by the customers. This project specifically for provide clean data solution about E-Commerce.

1.5 Project outputs

Our project outputs come with clear documentation along with input and output snapshots, so a non-technical person also can understand the results.

1.6 Project scope

Our E-Commerce project scope to provide clear outcomes as per the client`s requirement.

1.7 Project budget

We offer less charge depend with volume of data, and for this particular E-Commerce project our offer is very flexible charge for clients.

E-Commerce

2.1 What is E-Commerce?

E-commerce (electronic commerce or EC) is the buying and selling of goods and services, or the transmitting of funds or data, over an electronic network, primarily the internet. These business transactions occur either as business-to-business, business-to-consumer, consumer-to-consumer or consumer-to-business.

2.2 Database for Transaction details

Here is the transaction database for sample.

Tid	Date	Cust id	Amount	Category	Product	City	State	Payment mode
00000001	06-26-2015	4007024	040.33	Exercise & Fitness	Cardio Machine Accessories	Clarksville	Tennessee	credit
00000002	06-26-2015	4007023	238	Gymnastics	Gymnastics Rings	Des Moines	Iowa	credit
00000005	06-26-2015	4007028	1265	Outdoor Recreation	Camping & Backpacking & Hiking	Chicago	Illinois	credit

2.3 Database for Customer details

Here is the transaction database for sample.

Custid	Fname	Lname	Age	Profession
4000001	Sunitha	Devi	22	Teacher
4000002	Arul	Selvan	25	mentor
4000003	Vinod	Kumar	33	manager

2.4 List of use cases

Test case 1:

==========

To get all the details from the transaction amount detail that is greater than with a specific amount which the user wants.

Output file:

```
00049986        156.38
00049991        191.29
00049994        177.22
00049996        163.81
00049998        180.41
00049999        168.49
hduser@ubuntu64server:~$
```

Test case 2:

==========

To count all the transaction where amount is between 150 and 500.

I/P data:

```
hduser@ubuntu64server:~$ hadoop fs -ls /johnamtbw
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2016-11-21 13:17 /johnamtbw/_SUCCESS
-rw-r--r--   1 hduser supergroup          5 2016-11-21 13:17 /johnamtbw/part-r-00000
hduser@ubuntu64server:~$ hadoop fs -cat /johnamtbw/p*
5068
hduser@ubuntu64server:~$
```

O/p data;

```
hduser@ubuntu64server:~$ hadoop fs -cat /johncount/p*^C
hduser@ubuntu64server:~$ hadoop jar amtbw.jar /johnin/txns-large.dat /johnamtbw
Enter the lower limit
150
Enter the upper limit
170
16/11/21 13:16:55 INFO client.RMProxy: Connecting to ResourceManager at /192.168.56.123:8032
16/11/21 13:16:57 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed.
```

Test case 3:

==========

Calculate the total sum and total count of all the transaction for each user id.

I/P data:

```
hduser@ubuntu64server:~$ hadoop jar sum.jar /johnin/txns-large.dat /johnsum
Enter the User Id
4004613
16/11/21 13:24:42 INFO client.RMProxy: Connecting to ResourceManager at /192.168.56.123:8032
16/11/21 13:24:44 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed.
```

O/P file

```
hduser@ubuntu64server:~$ hadoop fs -ls /johnsum
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2016-11-21 13:25 /johnsum/_SUCCESS
-rw-r--r--   1 hduser supergroup         63 2016-11-21 13:25 /johnsum/part-r-00000
hduser@ubuntu64server:~$ hadoop fs -cat /johnamtbw/p*^C
hduser@ubuntu64server:~$ hadoop fs -cat /johnsum/p*
4004613 Sum : 800.05   Count : 9   Average : 88.89444444444445
hduser@ubuntu64server:~$
```

Test case 4:

Calculate total sales amt for each Month.

I/p file:

```
hduser@ubuntu64server:~$ hadoop jar ttlsl.jar /johnin/txns-large.dat /johnttlsl
Enter the Months
05
16/11/21 13:37:24 INFO client.RMProxy: Connecting to ResourceManager at /192.168.56.123:8032
16/11/21 13:37:26 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed.
```

O/p file :

```
hduser@ubuntu64server:~$ hadoop fs -ls /johnttlsl/p*
-rw-r--r--   1 hduser supergroup         22 2016-11-21 13:38 /johnttlsl/part-r-00000
hduser@ubuntu64server:~$ hadoop fs -cat /johnttlsl/p*
05      432627.58000000013
hduser@ubuntu64server:~$
```

Test case 5:

==========

Divide the file into 12 files, each file containing each month of data. For eg. file 1 should contain data of january txn, file 2 should contain data of feb txn.

I/p file:

```
hduser@ubuntu64server:~$ hadoop fs -ls /johnmonthofdata/p*
-rw-r--r--   1 hduser supergroup     432933 2016-11-21 13:51 /johnmonthofdata/part-r-00000
-rw-r--r--   1 hduser supergroup     389153 2016-11-21 13:51 /johnmonthofdata/part-r-00001
-rw-r--r--   1 hduser supergroup     442575 2016-11-21 13:51 /johnmonthofdata/part-r-00002
-rw-r--r--   1 hduser supergroup     422696 2016-11-21 13:51 /johnmonthofdata/part-r-00003
-rw-r--r--   1 hduser supergroup     426463 2016-11-21 13:51 /johnmonthofdata/part-r-00004
-rw-r--r--   1 hduser supergroup     422470 2016-11-21 13:51 /johnmonthofdata/part-r-00005
-rw-r--r--   1 hduser supergroup     430830 2016-11-21 13:52 /johnmonthofdata/part-r-00006
-rw-r--r--   1 hduser supergroup     429555 2016-11-21 13:52 /johnmonthofdata/part-r-00007
-rw-r--r--   1 hduser supergroup     422035 2016-11-21 13:52 /johnmonthofdata/part-r-00008
-rw-r--r--   1 hduser supergroup     427267 2016-11-21 13:52 /johnmonthofdata/part-r-00009
-rw-r--r--   1 hduser supergroup     409774 2016-11-21 13:52 /johnmonthofdata/part-r-00010
-rw-r--r--   1 hduser supergroup     424714 2016-11-21 13:52 /johnmonthofdata/part-r-00011
hduser@ubuntu64server:~$
```

O/p file:

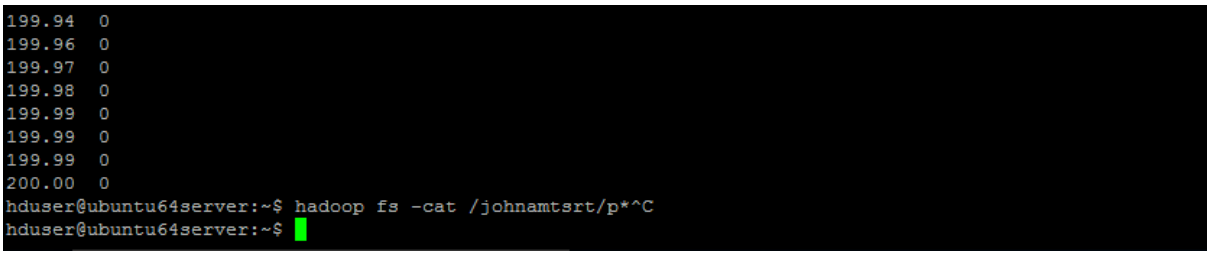
```
6       00004827        06-26-2015      4008117         092.66  Gymnastics      Pommel Horses   El Paso         Texas   credit
6       00043676        06-02-2015      4007212         034.62  Combat Sports   Martial Arts    Boston  Massachusetts   cash
6       00040361        06-16-2015      4008509         157.99  Outdoor Play Equipment  Playhouses      Tampa   Florida         credit
6       00033323        06-29-2015      4003604         155.52  Exercise & Fitness      Cardio Machines         Columbia        South Carolina  credit
6       00008611        06-10-2015      4003102         018.02  Water Sports    Life Jackets    Long Beach      California      credit
6       00038036        06-22-2015      4001694         108.44  Team Sports     Cricket         Milwaukee       Wisconsin       credit
6       00026380        06-11-2015      4006188         187.19  Gymnastics      Vaulting Horses         Midland         Texas   credit
6       00017388        06-23-2015      4007484         036.71  Exercise & Fitness      Jump Ropes      Stamford        Connecticut     cash
6       00021249        06-05-2015      4008624         141.27  Outdoor Recreation      Riding Scooters         Midland         Texas   credit
6       00029186        06-01-2015      4009240         091.88  Outdoor Play Equipment  Water Tables    Charlotte       North Carolina  credit
6       00045234        06-08-2015      4006223         159.87  Exercise & Fitness      Free Weight Bars        Omaha   Nebraska        credit
6       00043675        06-01-2015      4005590         011.22  Jumping         Trampolines     Portland        Oregon  cash
6       00016906        06-04-2015      4005417         061.17  Winter Sports   Downhill Skiing         Washington      District of Columbia    credit
6       00029182        06-23-2015      4005664         144.61  Gymnastics      Springboards    New York        New York        credit
6       00029181        06-13-2015      4000964         184.74  Games   Board Games     San Antonio     Texas   credit
6       00024472        06-11-2015      4007815         142.86  Outdoor Recreation      Fishing         Brownsville     Texas   credit
6       00018115        06-17-2015      4000792         198.25  Jumping         Trampolines     Pasadena        Texas   credit
6       00045236        06-03-2015      4000447         125.59  Exercise & Fitness      Weightlifting Machine Accessories       Philadelphia    Pennsylvania    credit
6       00033325        06-30-2015      4005382         028.95  Racquet Sports  Racquetball     Kansas City     Kansas  credit
6       00000598        06-28-2015      4007452         062.34  Team Sports     Cricket         Denton  Texas   credit
6       00007593        06-05-2015      4006795         078.51  Team Sports     Cheerleading    Oklahoma City   Oklahoma        credit
6       00001743        06-13-2015      4006354         173.72  Gymnastics      Gymnastics Rings        Scottsdale      Arizona         credit
6       00029173        06-06-2015      4007332         037.89  Outdoor Recreation      Camping & Backpacking & Hiking  Oakland         California      cash
6       00029172        06-02-2015      4007044         110.54  Outdoor Recreation      Riding Scooters         Indianapolis    Indiana         credit
6       00037182        06-14-2015      4003597         097.44  Gymnastics      Vaulting Horses         Detroit         Michigan        credit
6       00024474        06-23-2015      4008215         180.41  Water Sports    Water Polo      Memphis         Tennessee       credit
```

Test case 6:

==========

Sorting all the transaction amount in ascending order.

O/p file:



Conclusion

With these different scenarios we can give different dimensional solution in accurate with a huge dataset. This type of data solution will help Industries to maintain their customer and transaction details as well. We are very trustable MNC for data solution and we can proudly say, we are the best data analyser and tech solution provider for other MNC`s and start-ups companies who needs data solution.