Behavioural analysis of the expenses of customers using MapReduce on Hadoop

Trupti Kulkarni

X15012948

MSc Cloud Computing, National College of Ireland
Mayor Street IFSC, Dublin 1, Ireland

trupti.kulkarni@student.ncirl.ie

***Abstract*— Public sector banks deal with large amount of data on a daily basis. To manage such large datasets, Bigdata analytics techniques like Mapreduce that run on Hadoop framework are used. The basic entities in public sector are customers and their transactions. The question of interest here is to analyse the behavioural patterns of the spending habits of customers and patterns of their transactions. These behaviours are encoded into data through certain medium that capture behaviour. Various data sources are then accessed, prepared, consolidated and analyzed. Ultimately, this gives rise to insights into the patterns of the expenses by the customers across a period of time.**

Keywords— **Hadoop; Mapreduce; Java; Pig; Hive;**

I. Introduction

Hadoop is present in all the vertical industries today for leveraging big data analytics so that organizations can gain competitive advantage. With increase in the amount of data processing, the traditional programming techniques turn out to be very time consuming, thereby affecting performance in a much broader way. A number of High Level Query Languages (HLQLs) have been constructed on top of the Hadoop MapReduce realization, primarily Pig, Hive, and JAQL. Finance sector generally handles petadata from transactions amassed on regular basis. Many of the banks have already shifted to Hadoop. Many of the transactions are done using credit and debit cards. In order to achieve significant customer satisfaction, it is necessary to have a clear picture of a Bank's customer expenses. This report analyses the pattern of expenditure by the users that are part of a particular bank for year 2014 and 2015. This analysis is based on the basic information of the customers and their transaction types.

II. Related Work

In 2009, analysis was done (McSweeney, 2009) on the behaviour of online frauds. This analysis suggests that there are various ways for online bank frauds, few of them being:

1. Stealing of a legitimate user’s banking authentication details by some form of identity theft. These details are utilized for fraudulent and criminal purposes for instance phishing and crimeware attacks.
2. By breaching the security standards followed by the banks, which allows criminals access to bank’s internal systems
3. By fraudulent activity by bank employees
4. Stealing of authentication details and performing fraud by the trusted members close to the legitimate customer

From this analysis it can be seen that:

1. Number of crimeware spreading URLs infecting PCs with password-stealing code rose 93 percent in 2008 to 6,500 sites, nearly double the previous high of November, 2007 - and an increase of 337% percent from the number detected in 2007. (McSweeney, 2009)

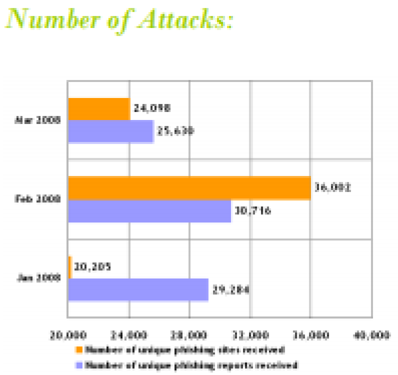


Figure 1: Number of Attacks between Jun 2008 and Mar 2009

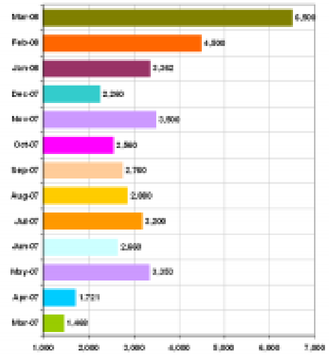
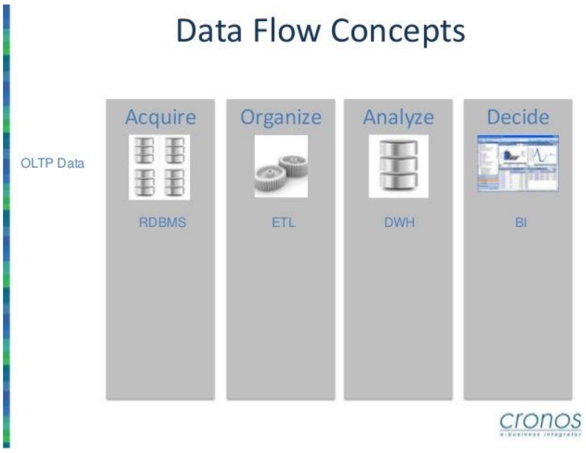


Figure 2: Monthly Attacks Recorded

III. METHODOLOGY:

This section describes about the methodologies used to analyse the banks transactions data to identify the expenditure of its customers. This report describes the behavioural pattern of the customer based on their transactions and their related information. At a higher level, transaction analysis is done using a core **Acquire-Organize-Analyse-Decide** cycle(McSweeney, 2009). These stages perform the following tasks:

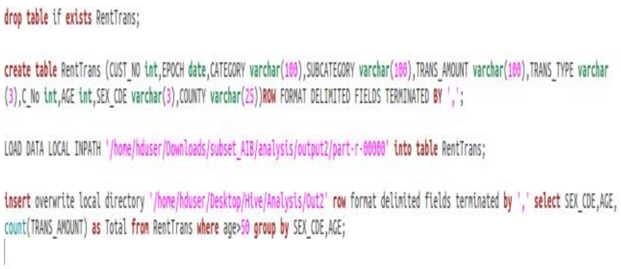


Collect: At this stage, the information is collected. It is necessary to understand the kind of datasets are needed to perform these analysis. As the analysis needs to be carried out on the banks' transaction information and customers' expenditure, related data is collected from a reliable source. Also, transactional data needs to be categorized based on different characteristics of the customer and patterns of the transactions. For this analysis, a dataset of around 850MB was collected which consists of 5 .csv formatted files. This information was split among 5 different files namely, transaction(8909689 records), rent(12789 records), income(1579 records), demographics(219111 records) and balances(76578 records).

Organize:
In order to analyse the data further, these files are divided into primary and secondary files as per the analysis being carried out. As this report analyses the expenditure of customers depending upon other factors. Transactions and Rent files are considered as the primary files, as these files contain majority of the data about the expenditure of the customers. As other 3 files namely, income, demographics and income contain the other supporting information about the transaction and the customer, these files are treated as secondary.
In order to perform proper analysis about the expenditure, a JOIN of rent file and transaction file was done since paying rent is also one kind of expenditure. Also, in order to have quality of the data, few inconsistencies were removed from the data such as records with null values in customer number field.
The original table format for the files had few inconsistencies and some data was missing. For instance, as mentioned earlier, rent is considered one type of expenditure, 2 columns namely, categories and sub-categories were missing in the original file.

Analyze:
Case study 1 :
As the rent varies from individual to individual , so in order to understand the amount spent by the senior citizens (age>50) analysis needs to be performed. Taking into consideration this question , individual has been categorised gender wise and age wise. Depending upon this analysis ,bank can offer various schemes to senior citizens to achieve better customer satisfaction.
In order to get the amount spent by the senior citizens over the period of two years, accounts of these customers are scanned and the total amount spent amount by them till date on rent is calculated. The combined transaction and rent file is joined(full outer join) with demographics file using pig script. The results are stored back to the local disk in csv format. This .csv format output file is given as output to either hive or imported on to database .The database table is then imported into hbase so as it can be accessed through Hadoop. Once the data is uploaded onto Hadoop , Java mapreduce program is written so as to carry out the required analysis.

Following the Hive script in order to calculate the total amount spent depending on the gender and age.



```
drop table if exists RentTrans;

create table RentTrans (CUST_NO int,EPOCH date,CATEGORY varchar(100),SUBCATEGORY varchar(100),TRANS_AMOUNT varchar(100),TRANS_TYPE varchar
(3),C_No int,AGE int,SEX_CDE varchar(3),COUNTY varchar(25))ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

LOAD DATA LOCAL INPATH '/home/hduser/Downloads/subset_AIB/analysis/output2/part-r-00000' into table RentTrans;

insert overwrite local directory '/home/hduser/Desktop/Hive/Analysis/Out2' row format delimited fields terminated by ',' select SEX_CDE,AGE,
count(TRANS_AMOUNT) as Total from RentTrans where age>50 group by SEX_CDE,AGE;
```

Case Study 2 :

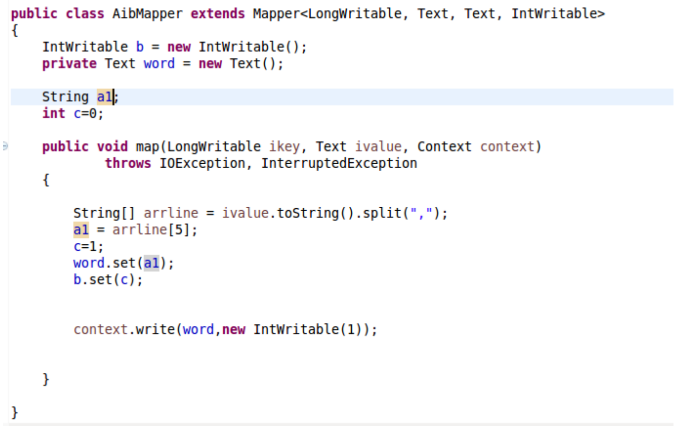
It has been observed that, customers perform transactions either by debit card or by credit card. In order to understand the usage of debit/credit card by customers depending upon the category analysis needs to be performed. Taking into account this requirement of the analysis , essential files has to be selected.

Viewing the requirements , the required information is available in transaction file.In order to carry out this experiment we have imported transaction.csv file into mysql relational database by using following mysql script .

load data local infile '/home/hduser/Downloads/subset_AIB/transaction.csv' into table Transaction fields terminated by ',' lines terminated by '\n';

Once the data is imported into mysql table , the table is uploaded onto Hbase usingfollowing command:
bin/swoop-import --connect jdbc:mysql://127.0.0.1/Test --username root --table Transaction --password Hadoop2015

Following is the Java Mapper code in order to perform the group by (trans_mode,category)::



```
public class AibMapper extends Mapper<LongWritable, Text, Text, IntWritable>
{
    IntWritable b = new IntWritable();
    private Text word = new Text();

    String a1;
    int c=0;

    public void map(LongWritable ikey, Text ivalue, Context context)
            throws IOException, InterruptedException
    {

        String[] arrline = ivalue.toString().split(",");
        a1 = arrline[5];
        c=1;
        word.set(a1);
        b.set(c);


        context.write(word,new IntWritable(1));


    }
}
```

Java Reducer code in order to calculate the total amount spent::

Case Study 3::

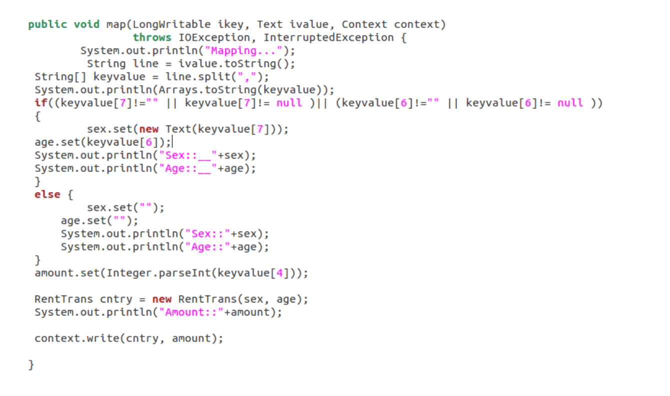
Transaction details of each individual differs from each other. Students' spending are varies from that of any professional. The pattern of outflow of money change depending open the age, gender and profession. In order to understand this pattern, analysis has to be carried out.

Depending upon the available information about the customers, analysis is carried out to calculate the total amount spent by categorizing age and the gender of the customer.

In order to perform this, information is imported from Transaction.csv to mysql and in turn is pushed to HBase so to make it available on Hadoop.

Java MapReduce program is written in order to perform the task.

Following is the Java Mapper code to get group by(sex,age)

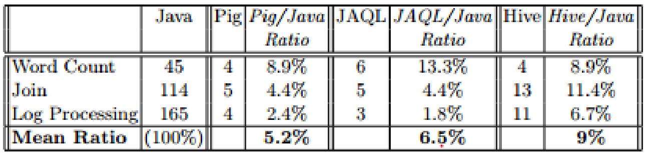


```
    public void map(LongWritable ikey, Text ivalue, Context context)
                throws IOException, InterruptedException {
            System.out.println("Mapping...");
             String line = ivalue.toString();
     String[] keyvalue = line.split(",");
     System.out.println(Arrays.toString(keyvalue));
     if((keyvalue[7]!="" || keyvalue[7]!= null )|| (keyvalue[6]!="" || keyvalue[6]!= null ))
     {
             sex.set(new Text(keyvalue[7]));
     age.set(keyvalue[6]);
     System.out.println("Sex::__"+sex);
     System.out.println("Age::__"+age);
     }
     else {
             sex.set("");
         age.set("");
         System.out.println("Sex::"+sex);
         System.out.println("Age::"+age);
     }
     amount.set(Integer.parseInt(keyvalue[4]));

     RentTrans cntry = new RentTrans(sex, age);
     System.out.println("Amount::"+amount);

     context.write(cntry, amount);

    }
```

Java Reducer code to calculate the total expense of customer grouped by (sex,age).

Decide:

After running the scripts using various language , the performance efficiency of each of the languages can be understandable. Hive was shown to achieve the quickest runtime for every benchmark, whilst Pig and JAQL produced largely similar results for the scalability experiments, with the exception of the join application, where JAQL achieved relatively poor runtime(Stewart, R.J., Trinder, P.W. and Loidl, H.W., 2011.). Both Hive and Pig have the mechanics to handle skewed key distribution for some SQL like functions, and these abilities were tested with the use of skewed data in the join application. The results highlight the success of these optimizations, as both languages outperformed Java when input size was above a certain threshold.(Stewart, R.J., Trinder, P.W. and Loidl, H.W., 2011.)



	Java	Pig	*Pig/Java Ratio*	JAQL	*JAQL/Java Ratio*	Hive	*Hive/Java Ratio*
Word Count	45	4	8.9%	6	13.3%	4	8.9%
Join	114	5	4.4%	5	4.4%	13	11.4%
Log Processing	165	4	2.4%	3	1.8%	11	6.7%
Mean Ratio	(100%)		**5.2%**		**6.5%**		**9%**

From the analysis done from the **case study 1**, it can be interpreted from the given data that the customers' usage of the debit card is twice as compared to that of the usage of the credit cards(Fig 1)

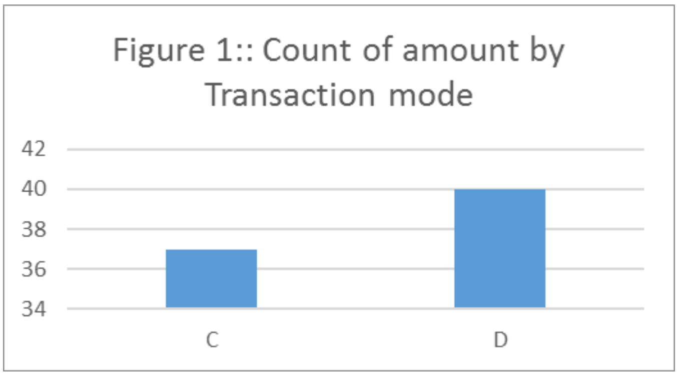


Figure 1:: Count of amount by Transaction mode

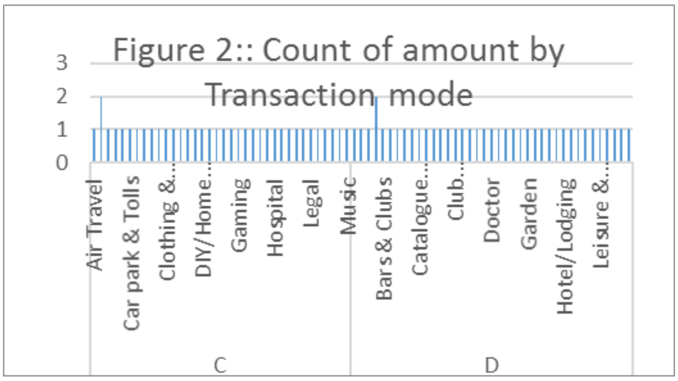


Figure 2:: Count of amount by Transaction mode

Also the transfer mode is further categorized specifying where customer has used debit and credit cards max(Fig 2).

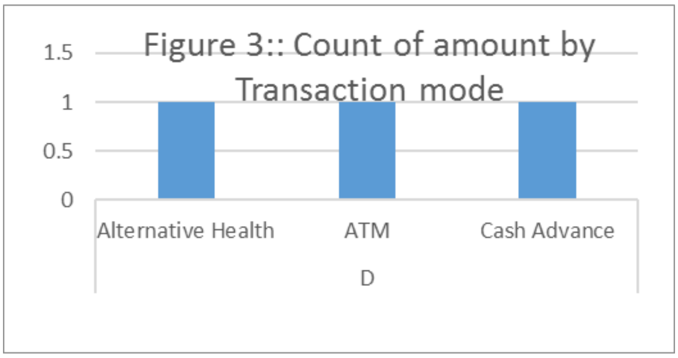


Figure 3:: Count of amount by Transaction mode

Also from Fig 3 it can be seen that there are some categories like "Alternative Health", "ATM" and "Cash Advance" that are performed only by debit cards.

Most of the population spend variable amount on rent , depending upon their occupation , age group , personal habits and gender. After analyzing the outcome for case study 2, the total amount of the rent paid by the senior citizens, it can be concluded that female spend very less amount on rent as compared to the male.

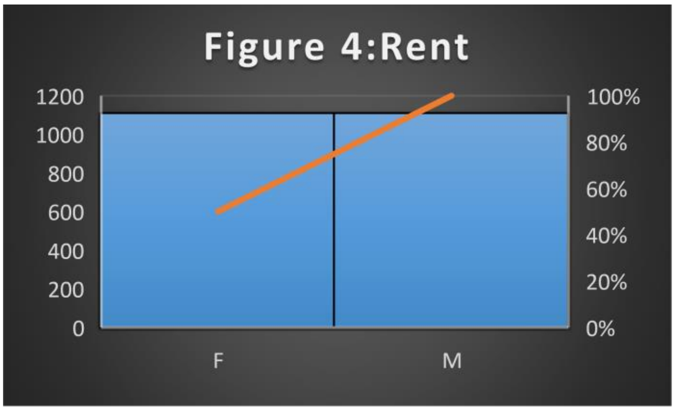


Figure 4:Rent

CONCLUSION.

While overviewing the expenses of the huge amount of population, expenses may differ depending upon miscellaneous factors. This report describes about the analysis of expense of customers depending upon the various factors like transaction mode, category, age and gender. Depending upon the results of all the analysis done in this report, particular bank can such extract data in order to build special schemes for their customers depending upon numerous factors like age and category of transaction in order to enhance the customer satisfaction.

BIBLOGRAPHY.

Stewart, R.J., Trinder, P.W. and Loidl, H.W., 2011. Comparing high lev
Processing Technologies(pp. 58-72). Springer Berlin Heidelberg.

Vallaey, M. (2014) Big data in public services. Available at: http://ww
services (Accessed: 21 April 2016).