International Journal of Engineering & Technology, 7 (4.5) (2018) 608-611

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Minimizing the cost by effective utilization of resources which yields better profits for government organizations in a less interval of time

Ch. Nanda Krishna [1] *, M. Ramesh [1], Dr. M. Suneetha [2]

[1] Assistant Professor, Dept. of Information Technology, VRSEC
[2] Professor & Head, Dept. of Information Technology, VRSEC
**Corresponding author E-mail: nkcherukuri@gmail.com*

Abstract

Minimizing the cost and utilization of resources in an efficient manner helps any organization in its progress that is growth of the organization. This paper gives an optimized solution for reducing the cost of resources in government services of Transportation. If the data is structured can use any one of the Relational Database Management Systems (RDBMS), if the data is unstructured or semi-structured we can use hadoop for processing the data and analyze the data which is processed by using R. Strategic decision making can be made by the end results that we get from the visualizations of R-Language in-order to get better insights of the business and better decisions can be taken for better growth of the organization.

Keywords: *Optimized Solution; Structured Data; Semi-Structured and Unstructured Data; Visualizations and Decision Making*

1. Introduction

Many organizations deal with the real time data. The data is pulled into any one of the RDBMS, which is used to store data that is flowing from source. In the process of loading data if needed pre-processing is done and the required data is loaded for further analysis[4]. Upon which reports can be generated according to the need of the business users using R-Programming[7]. The end user of the business could understand the trends in business by comparisons of the reports from one week to the other or one month with the other or one quarter with another and changes can be made in business.

1.1. History in R

Before it is called as R the programming language is called S programming language which is combined with the lexical semantic. The creators of R were Robert Gentleman and Ross Ihaka whose names start with R so finally the name R-Programming came into existence.
R-Programming is an open-source language which is mainly for the purpose of statistical computing of Data and also the graphics of R, which is mainly used in the areas of machine learning and data mining by the data miners and statisticians for better visualizations which leads to better insights into the business and also improves decision making[10]. With R Programming we can implement linear, non-linear modellings, time-series analysis, clustering and classification of data. With-in the R-programming
Number of varieties of file formats of data can be dumped and analysed, and also the structured, semi-structured and un-structured data can also be dumped into R by the use of relational and non-relational database management systems. R has many predefined or built-in functions which reduces the lines of code that the programmers need to write.

1.2. Features of R

R covers wide range of areas like research, statistical applications in economics, medicine, business and sciences. R provides modellings like linear and non-linear, clustering, classification and regression analysis. R is the best tool. in the current industry for statistical modelling of data. The visualisations from R give us better insights into the business growth.
Data Analytics:
In order to draw conclusions of the client requirements analytics is the process which gives the information needed from huge repositories. The output of the analytics i.e information plays a key role in decision making.
Importance of Analytics:
The solution that analytics provides to the end-customer are cost reduction, faster, better decision making, new products and services.
Applications of Analytics:
Healthcare providers, education, communications, media and entertainment, banking and securities, government sectors, insurance companies, transportation, retail and wholesale trade are the areas where analytics can be done to get optimized solutions.
Benefits of R Programming:
R is an open-source programming language used for statistical analysis of data. The packages in R Programming reduce the lines of code in-context to the programmer. The graphical user interface of R is well designed and easy to implement.
Big Data Analysis using R:
The data which is in unstructured, semi-structured and structured data can be processed by using R. We can access the data from hive using R.
Types of Analytics:

Copyright © 2018 Ch. Nanda Krishna et. al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Different types of analytics can be implemented by using R-Programming. They are Prescriptive, predictive, diagnostic and descriptive [6].
Prescriptive Analytics:
The outcome of prescriptive analytics is that what type of actions can be taken. From this analytics, rules and recommendations for the next steps can be framed.
Diagnostic Analytics:
The outcome of diagnostic analytics is that by looking at the past performance and giving a solution why it happened likewise. Example for this type of analytics is dashboard analytics.
Descriptive Analytics:
The outcome of this analytics is what is happening now based on the incoming data; to mine the data we can use real-time dashboard data.
Predictive Analytics:
The outcome of predictive analytics is that what might happen in the future be able to be predicted.
In the next session i.e existing system the methodology which is in use is storing data in flat files or excel files and processing it which may not give the optimized solution.

2. Existing system

The system that is in existence is usage of Microsoft Excel or flat files for analysing the data, which takes lots of time for getting the needed information as well as making the decision takes lot of time because the data that is generated everyday is huge[4]. We need to process semi-structured as well as unstructured data at-times which is not possible to handle by the File Processing Systems. The methodologies that are implemented over here are static, each and every time as there is change in data dimensions a new methodology need to be taken into consideration which takes lot of time practically. In order to overcome the problems in the existing system, the suggested techniques are implementation of the concept on R programming with mysql relational database management system in the proposed system.



Fig. 2.1: Data Flow in Microsoft Excel.

3. Proposed system

In the proposed system it is suggested to use mysql for storage of huge repositories of data that is generated every day on top of it R-Programming is used for analysis of data and also the data can be viewed statistically which also help in decision making[4]. In this paper the implementation is done using Diagnostic Analytics Approach for better understanding of the previous data and trying to give a solution which is optimized from the current business perspectives. The type of technique that we are using in this paper implementation is Diagnostic Analysis of data that is by taking the past data into consideration and giving the solution for the future growth of the business [6].

3.1. Dataset

Is a collection of data which is usually presented in the tabular-format. The datasets that we are working on is in the form of Microsoft excel sheets, which consists of data for every month and source etc.

3.2. Processing

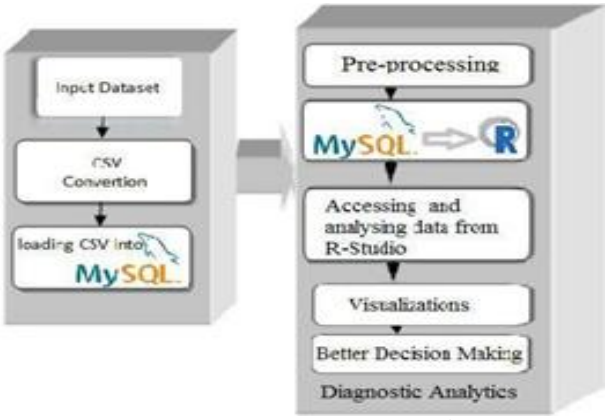


Fig. 3.2: Steps Involved in Processing of Data.

3.3. Dataset

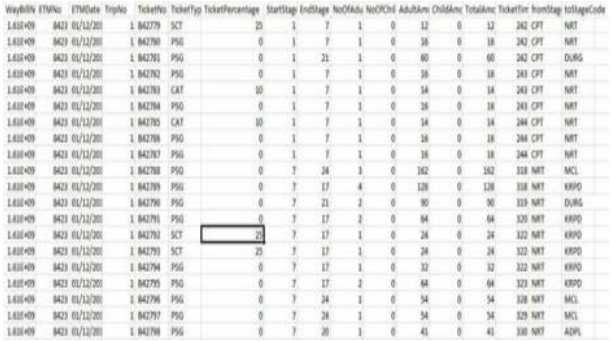


Fig. 3.3: The Excel File Dataset.

The excel file consists of many attributes (π). The useful attributes in the dataset were considered by attribute subset selection includes TicketNo, Tickettype, etmdate, servicetype, etmnumber and tripno etc. were identified as the major fields on top of which decisions can be made. The dataset consists of 50,000 records(σ) per one route as there are many such routes that the passengers travel for their lively need.

3.4. Excel file

The excel file with the required tuples and attributes that need to be analysed.

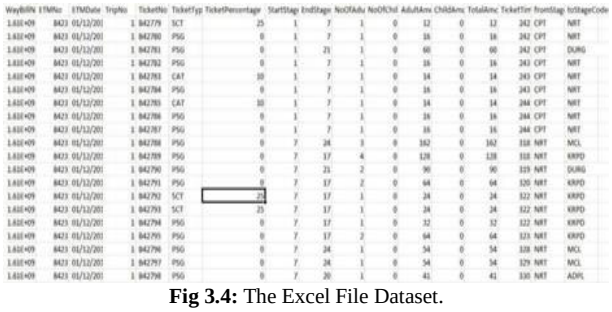


Fig 3.4: The Excel File Dataset.

Converting into .csv (comma separated values) file:
Converting the excel file into .csv file and then loading it into mysql database.

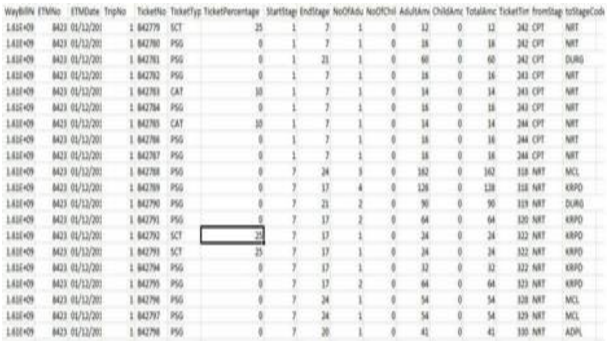


Fig. 3.5: The Excel File Dataset.

3.5. Loading data into my SQL database

There are many ways for loading data into mysql, they are using "load data inpath" or by import and export wizard etc choose the best among them which suits your requirements, such that the .csv file need to dumped into mysql relational database management system[2].

Pre-processing:

Pre-processing of data plays a key role in deciding what type of data is needed for analysis on top of which visualizations can be done for better understanding of perspectives of business growth by an end-client or customers [4]. Pre-processing of the data needs to be done as per the requirements of the business users. Among different pre-processing techniques methods like normalization, replacing missing values if needed and many other methods like cleansing of data, applying different types of transformations on the data can be implemented on the streaming data that comes in through the excel file from the end-client or customer.

Connecting mysql with R-Studio:

In-order to establish connection between R-Programming and mysql relational database management systems, using the concepts of libraries, connection managers, database connectivity

that is username, password, databasename and hostname are required.

```
Install·the·package·and·load·the·package.¶
¶
install.packages("RMySQL")¶
¶
library(RMySQL)¶
¶
Connecting·to·MySQL:¶
¶
mydb=dbConnect(MySQL(),user='username',·pass-¶
word='password',·dbname='database name',·host='hostname')¶
```

Fig. 3.6: install & load Package.

Accessing data from R-Studio, which is in mysql:

```
Tables and Columns in database:
dbListTables(mydb)
This command displays the list of tables in the
connected database.
dbListFields(mydb,'table_name')
This command displays the columns in the particular table.
Running the required queries as per requirements:
By using a built-in function dbSendQuery we can run
the queries.
dbSendQuery(mydb,'select * from tablename') Retrieving
the data from mysql database:
Rt=dbSendQuery(mydb,"select * from tablename where
condition")
Usage of fetch function to view the data:
Data=fetch(Rt,n=-1)
```

Fig. 3.7: Accessing Data in My Sql through R Analyzing the Data.

The data is analyzed as per the requirements of the client or business user on top of it analytics using algorithms like attribute subset selection is applied on the dataset in-order to decide the required attributes. The best way to analyze the data is through visualizations which can be done in R Programming [7]. There are many techniques in which the data can be visualized in R-Programming, the best suited technique that suits the requirements of the end-client or customer is taken into consideration.

4. Results & observations

The observations noticed are for a particular ticket type the number of tickets sold. From these statistics we can come to a decision that which ticket type is sold more or less and can further improve the growth of organization

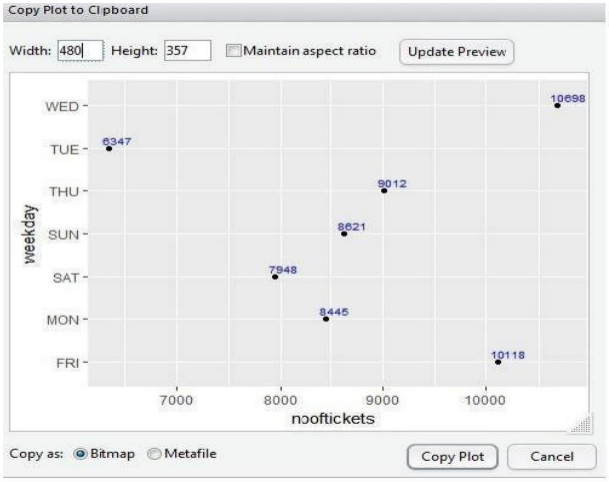


Fig. 4.1: Type of Tickets Sold vs Day of Week.

The observations noticed are the number of tickets sold on total per weekday on the total dataset. A decision can be taken such that the number of vehicles needs to be increased on particular day of a week in-order to increases the revenue.

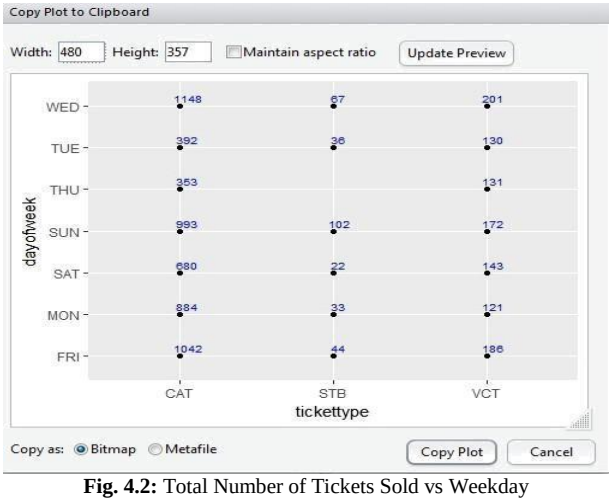


Fig. 4.2: Total Number of Tickets Sold vs Weekday

Describes the total number of tickets sold on each and every Monday of a month by this observation we can take a decision that is whether to increase the services of decrease the services via that particular route

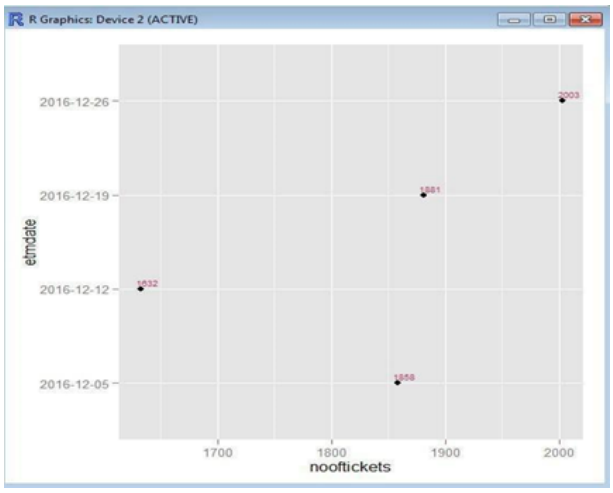


Fig. 4.3: Number of Tickets Sold on Monday is In a Month.

5. Conclusions & future work

The problem with Microsoft Excel and MySQL is that data is represented in tabular format only. For statistical representations R-Programming is the best suited mechanism from which we can even access the data that is in any of the RDBMS or any other flat file sources. As the data is presented in different statistical formats business users can easily visualize their business and can change their business needs in-order to get revenues. They can visualize their business either day-by-day or monthly or quarterly or annually and compare their growth and revenue at different intervals of time. In Statistical formats the growth of organization can be achieved by good decision making. As the data is
increasing day- by day the storage of the data plays a major role, data even includes structured, semi-structured and unstructured data issues need to be handled so concepts like hadoop with hive ecosystem or piglatin need to be used and also better graphical or pictorial representations ensure better insights into the business for better decision making.

References

[1] SmarandaBelciug, Florin Gorunescu, "Journal of Biomedical Informatics 53 (2015) 261-269.
[2] Thomas Rahlf. Data Visualisation with R. Springer International Publishing, New York, 2017. ISBN 978-3-319-49750-1.
[3] D. Yu Izai, Data storage basNed on distributed file systems with data replication, 2010, IEEE, ISBN978-966-335-401-9.
[4] Ch. Nanda krishna,dr m.suneetha, business intelligence solutions for processing huge data to the business user's using dashboards, international conference on signalprocessing,communication, power and embedded system(scopes), 2016, IEEE Explore Digital Library, 16980535.
[5] Kamleshkumarpandey, An Analytical and ComparativeStudy of Various Data Preprocessing Method in DataMining International Journal of Emerging Technology andAdvanced Engineering, ISSN 2250-2459, and ISO 9001:2008Certified Journal, Volume 4, Issue 10, October 2014.
[6] Sunghae Jun, an Efficient Connection between Statistical Software and Database Management System, IJCSBI.ORG, ISSN 694-2108 Vol. 8, 2013.
[7] P.VidyaSagar, Dr.N.Geethanjali, "An Improved Parallel Activity scheduling algorithm for large datasets", International Journal of Engineering Research and Applications, Vol. 5, Issue 7, pp.23-29, 2015.
[8] Daniel Keim, Big-Data Visualization, IEEE Computer Graphics and Applications, ISSN: 0272-1716, 2013.
[9] Nageswara Rao Moparthi,Dr. N.Geethanjali " Design and implementation of hybrid phase based ensemble technique for defect discovery using SDLC software metrics " An International Conference by IEEE, AEEICB16(978-1-4673-9745-2) PP. 269-276, ©2016 IEEE.