IOP Conf. Series: Materials Science and Engineering **263** (2017) 042052 doi:10.1088/1757-899X/263/4/042052

Comparative Study of Big data Analytics Tools: R and Tableau

Rajeswari C, Dyuti Basu, and Namita Maurya,

VIT University, Vellore- 632014, India
E-mail: rajeswari.c@vit.ac.in

Abstract. Big data is a huge collection of data from various sources. It can be of any type and tough to be interpreted and analyses hence we need some tool or mechanic that can easily analyses the data and give us some information out of it. Among various interesting tools R and tableau are the tools which deals with the big data analytics also it generates the output in visualization technique i.e., more understandable and presentable. In this paper we are comparing and contrasting the working of both the tools with some big dataset along with the importance and need of the tool in the field of big data analytics. This study gives the clear picture of growing data and the tools which can help more effectively, accurately and efficiently.

1 Introduction

Nowadays the data is been generated everywhere like facebook, twitter, gmail and industries. These data can be of any form audio video simple text etc. Understanding this data is very important as this is crucial and very important entity of an organization. The management of big data ensures a high level of data accuracy and accessibility for business intelligence. Big data is a collection of large and complex data which are difficult to be handled with traditional data processing application software. Analyzing the data sets can give a new business trends, prevent diseases, and combat crime and so on. Scientists, practitioners of medicine, business executives, advertising and governments come across difficulties with large data-sets like in internet search, urban informatics, finance, and business informatics. As many sources of data are getting generated, business managers at all levels ready to get data visualization software that let them to analyze it visually and take fast decisions. Currently, the most used tools for visualizations, data discovery are r and tableau. Tableau is one of the fastest upcoming business

intelligence (BI) tool. It is fast to deploy, easy to learn and very useful for a customer. Tableau has five main products facilitate to diverse needs for professionals and organizations. They are:

- tableau desktop: for individual use
- tableau server: collaboration for any organization
- tableau online: business intelligence in case of cloud
- tableau reader: used for reading files saved in tableau desktop.
- tableau public: for publishing interactive data online.

Tableau desktop has both a professional and personal edition. Tableau online is available to users in annual subscription, and enlarges to support thousands of users. Whereas, the "r" is statistical version of what is used to handle big data. R is a scripting language like a programming language hence it is a better tool. It integrates smoothly with the complex document publishing system or in other words it is easily possible that the statistical output and graphics generated by r can be merged with publication-quality documents. It is easy to operate and also an economical tool**.**

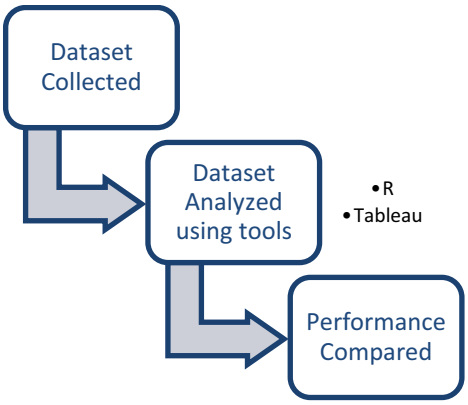
2 Methodologies

In this work we have followed a methodology. It includes three steps like:

2.1. Data collection: we are using three data sets in this work. The data set are collected from some websites. The description of the dataset is given in the section 5.

2.2. Analyzing the data using tools: this is the second step of our methodology. Here the datasets are analyzed using the tools r and tableau public.

2.3. Comparing the performance: this is the last and final step of the proposed methodology. Here the tools are compared on the basis of their performances. We have given the screen shots which was obtain at the time of analyzing the dataset.



3 Problem Statement

The aim of this work is to analyze the dataset using big data tools. This analysis will be done using tools like R and Tableau. A comparative study will be done on the basis of the performance of the tools.

4 Dataset Description

Three dataset have been used in this study. Those are:

a. ***Blood Transfusion Service Centre Data Set:*** This dataset is collected from the UCI Repository. This is a multivariate dataset. Number of instances (row) in this dataset is 748. It has 5 attributes named Recency, Frequency, Monetary, Time and whether he/she has donated blood. Recency says about months since last donation. Frequency gives the information about total number of donation. Monetary attribute holds the value of total blood donated in c.c. Time is the attribute which shows months since the first donation. The last attribute holds binary value stating whether a person has donated blood or not. The dataset does not contain any missing value. We have used the dataset in CSV (Comma Separated Value) for R and .xls format in Tableau.

b. ***Forest Fires Data Set:*** This dataset is also collected from UCI Repository. The characteristics of this dataset is multivariate which have 517 numbers of instances and 13 attributes. The attributes are: 1. X - x-axis spatial coordinate within the Montesinho park map; 2. Y - y-axis spatial coordinate within the Montesinho park map; 3. month - month of the year: 'jan' to 'dec' ; 4. day - day of the week: 'mon' to 'sun' ; 5. FFMC - FFMC index from the FWI system; 6. DMC - DMC index from the FWI system; 7. DC - DC index from the FWI system; 8. ISI - ISI index from the FWI system; 9. temp - temperature in Celsius degrees: 2.2 to 33.30 ; 10. RH - relative humidity in %; 11. wind - wind speed in km/h;12. rain - outside rain in mm/m2 ; 13. area - the burned area of the forest (in ha). The CSV format of this

dataset is used in R whereas the .xls format is used in Tableau. This dataset does not contain any missing value as well as the previous one.

c. Crime dataset: The crime dataset used in the present research work was downloaded from the Integrated Network for Societal Conflict Research (INSCR) website.

5 Implementations

5.1. Analysis using R:

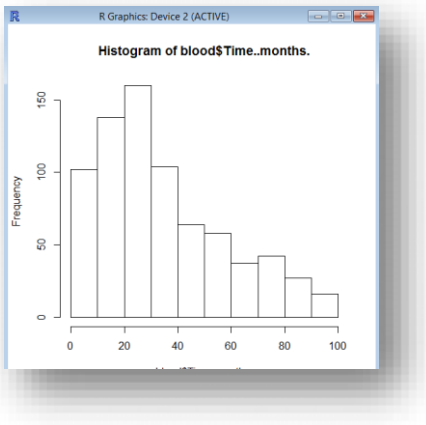


Fig. 1. Blood Trannsfution Service Center Dataset (Histogram)

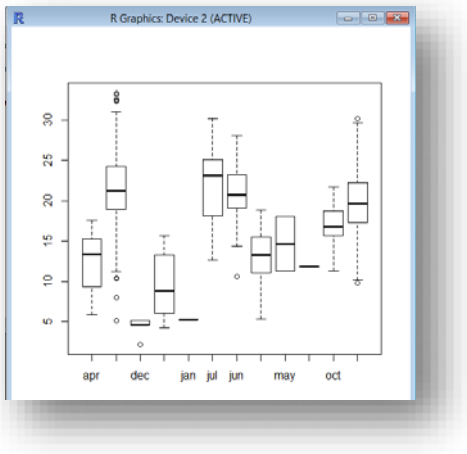


Fig. 2. Forest Fire Dataset (Scatter plot)

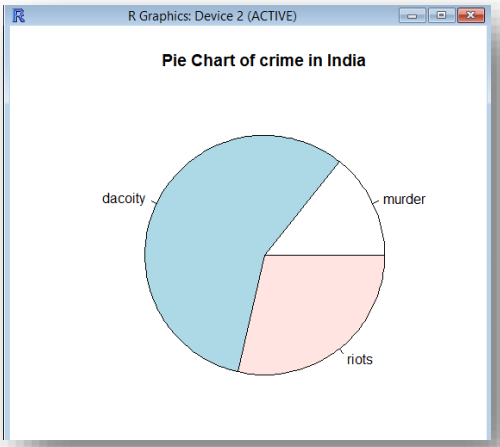


Fig. 3. Crime dataset (Pie chart)

5.2. Using Tableau:

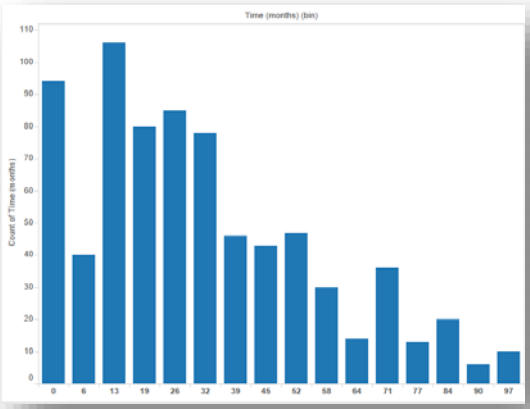


Fig. 4. Blood Transfusion Service center Dataset (Histogram)

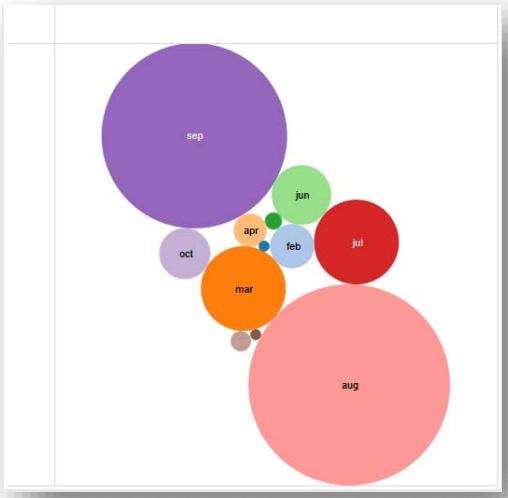


Fig. 5. Forest Fire dataset (Packed bubbles)

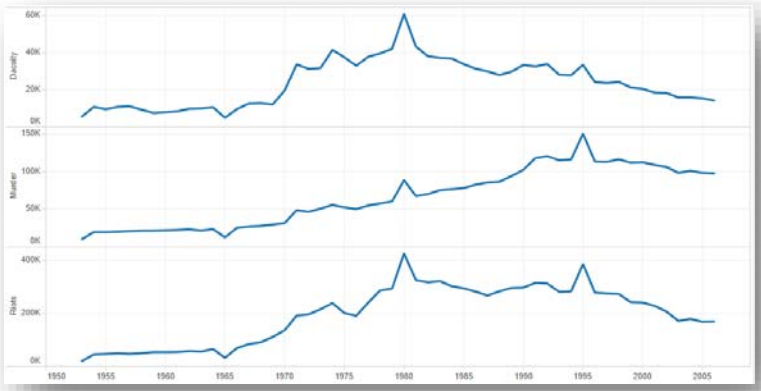


Fig. 6. Crime dataset (Lines)

6 Result:

The result shows us that the tools worked on the dataset. It has generated different kinds of pictorial representation of the analyzed data. While working with both the tools, we found some advantages and disadvantages of them. They are tabulated as follows:

Advantages:

R	Tableau
R is the most thorough measurable examination package available.	Tableau has an excellent user interface.
The graphical abilities of R are extraordinary, giving a completely programmable design dialect that outperforms most other measurable and graphical packages.	Its integration feature is also very attractive. I t can integrate with other big data platforms like Hadoop.
R is free and open source programming, permitting anybody to utilize and, imperatively, to alter it. R is authorized under the GNU General Public License, with a copyright assisted by The R Foundation for Statistical Computing.	Tableau supports in mobile devices . The report on the tableau dashboard is automatically optimized in mobile.
R has no license restrictions (other than ensuring our freedom to use it at our own discretion), and so we can run it anywhere and at any time, and even sell it under the conditions of the license.	It is low cost software which is also very easy to upgrade. It also consumes less memory space.

Disadvantages

R	Tableau
To use R one needs to learn it very well. Otherwise it cannot be used effectively.	Initial data processing is needed here in Tableau. And this should be done by professional kit expert.
All the packages used in R is not always give perfect result.	There is an argument that Tableau does not support all the statistical

	features.
Many R commands have very little memory management. So, when those commands perform their task they occupy the available memory very quickly.	As well as the other BI tools, financial reporting cannot be done using tableau. For this, financial analyst is needed.

From the above study of advantages and disadvantages, we can say that Tableau is better from the point of view of data analytics. Tableau is a tool that has more advantage than R. As we discussed earlier, it is a fast tool. It manufacturer provides maintenance yearly. It creates interesting and attractive charts and graphs really quick. Users of Tableau also have a chance to discuss the problems in the Tableau forum they face during operating the tool. Moreover Tableau public gives its user a chance to publish their dashboard.

7 Conclusion

Big data analytics using the tools is very effective, less time consuming and interesting. From our study we conclude that Tableau is the more efficient tool than R in big data analytics. The usefulness of Tableau in big data analytics can be measured by its performance, user friendly environment, and speed. However, there are more features of Tableau which was not taken into account in this study as the time was limited.

References

[1] Keim, Daniel A., et al. "Visual analytics: Scope and challenges", *Visual Data Mining*, Springer Berlin Heidelberg, 2008, 76-90.
[2] Fan, Wei, and Albert Bifet. "Mining big data: current status, and forecast to the future." *ACM SIGKDD Explorations Newsletter*, **14(**2), (2013), 1-5.
[3] Katal, Avita, Mohammad Wazid, and R. H. Goudar, "Big data: issues, challenges, tools and good practices", Contemporary Computing (IC3), *Sixth International Conference on IEEE*, 2013.
[4] Kalambe, Yogesh S., D. Pratiba, and Pritam Shah. "Big Data Mining Tools for Unstructured Data: A Review, *IJITR*, **3**, (2015), 2012-2017.
[5] Sanchita Patil, “Big Data Analytics using R”, *IRJET*, **3**, (2016)
[6] Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey. "Business intelligence and analytics: From big data to big impact." *MIS quarterly*, **36**, (4), (2012): 1165-1188.
[7] Kumar, Prashant, and Khushboo Pandey. "Big data and distributed data mining: an example of future networks." *International Journal of Advance Research and Innovation*, **(2),** (2013), 36-39.

[8] Bobade, Varsha B, "Survey Paper on Big Data and Hadoop." *International Research Journal of Engineering and Technology* (IRJET), 2016, 2395-0056.