IJASRM

International Journal of Advanced Scientific Research and Management, Volume 3 Issue 9, Sept 2018

www.ijasrm.com

ISSN 2455-6378

A comparative study on three big data analytics tools

Ilangovan Pandian[1], Meena Krishnamoorthy[2], Meenalaxmi[3], Abinaya[4]

[1] Department of Computer Science and Engineering,
Periyar Maniammai Institute of Science & Engineering,
Thanjavur, Tamilnadu, India.

[2] Department of Software Engineering
Periyar Maniammai Institute of Science & Engineering,
Thanjavur, Tamilnadu, India.

[3,4] Department of Computer Science and Engineering,
Periyar Maniammai Institute of Science & Engineering,
Thanjavur, Tamilnadu, India.

Abstract

In the developing technological world, data in all the fields is increasing every day. All these continuous flow of data from different sources are captured, stored and processed for future analysis or to mine knowledgeable information based on these stored data. Due to the availability of massive volume of data, it is difficult to process and do an effective analytics on these data. Different sources of data and different formats of data namely unstructured data, semi structured data and structured data became difficult to process and pool useful information together using traditional techniques. The recent technology which could be the solution to the highly increasing of data is big data. Such big data could be the best solution for performing analytics and prediction of the data which is captured and stored. This paper tries to review and compare three big data analytics tools namely: Tableau public, R Programming and KNIME. Comparison is made using three tools based on their Mode of software, data sources, Real time analytics, Owner, Scalability.

Keywords: *unstructured data, semi structured data, structured data, mode of software, data sources, owner, scalability, real time analytics.*

1. Introduction

In the recent years data keeps on increasing day by day and we are living in the data world. In our daily life, usage of electronic gadgets, social media data, sensor data and many real time sources data is getting increased every minute. By 2025 the data volume may increase upto to 35 Zeta bytes. The data

collected from different sources remains a challenging task to monitor, capture, stored, process and analytics by the conventional methods. To solve this problem, the only solution is "big data analytics". Big data analytics can perform an effective analytics and predict informative decisions in various business applications (Katal et al., 2013). The buzz word big data was first introduced to the information and communication technology world by Roger Magoulas in 2005. The term big data is said to be the massive volume of data. The four characteristics of big data remains: volume, velocity, variety and veracity. This is also called as 4V of big data analytics.

Volume: It is the important feature of big data. It is said to be the massive volume of data. Twitter, Face book other social medias generate 25 TB of data every day. Some other applications can generate more than TB of data per hour. Managing this huge volume of data is becomes a challenging task.

Velocity: It means the speed of the data generated from various sources, and all these data are stored, processed and retrieved for future use, velocity is not related with speed of the data alone it also depends on the flow of the data. Data from various sensor devices flows continuously and it is stored in the database. This traditional system is facing a challenging task to analyze these data.

Variety: Variety it means data may be classified into different categories It means unstructured data, semi structured data and structured data. Today the usage of smart phones and other electronic gadgets embedded with sensors is increased as well as different varieties of data also flow from these devices is also increasing gradually.

Veracity: veracity refers to find knowledgeable information based on the huge storage of data. It may help to find a better decision on much business application.

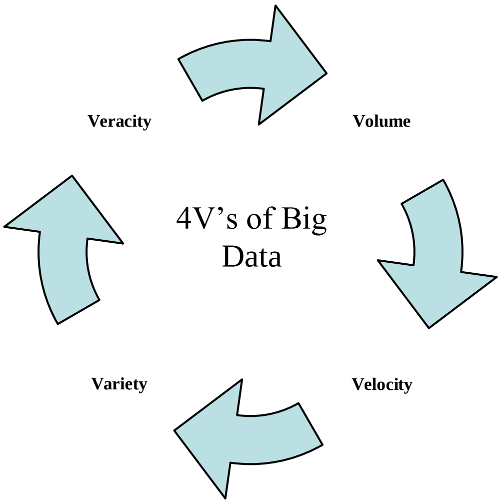

Fig.1 4 V's of Big Data

2. Data Analytics

Data analytics (DA) is the art of analyzing the data to mine useful information based on the stored data. It is used in many organizations, industries and others to mine information based on the stored data. In unprocessed stage, this massive volume of raw data captured doesn't contain any useful information. But, by the method of applying right data analytics tools we can mine an actionable insight for future betterments (Prasad et. al., 2016). There are four types of data analytics, they are: Prescriptive analytics, Predictive analytics, Diagnostic analytics, and Descriptive analytics.

Prescriptive analytics: This prediction analytics says what kind of action has to be taken based on the stored data. In this type of analytics, it suggests what should be done next.

Predictive analytics: In this predictive analytics, it is a forecast analytics which predicts what is going to happen next. This is a useful analytics which suggests you what is going to happen; a prediction is done based on the cleaned data stored in the data base.

Diagnostic analytics: This diagnostic analytics, it is based on the past performance to examine what next and what purpose it has to be practiced. This type of analytics is often made with a diagnostic decision.

Descriptive analytics: This descriptive analytics, is a kind of knowledgeable information which can be mined based on the stored data. In this type of analytics there is a lot of chance to mine useful information's.

In this next section we are going to compare the three data analytics tools namely Tableau public, R Programming, and KNIME. These tools are compared based on their Mode of software, data sources, Real time analytics, Owner, Scalability. This comparison is a general study of these three tools.

3. Comparison of three Big Data Analytics Tools

Tableau: It is one of the business analytics tool for analyzing the data visually. In Tableau, data can be viewed in the form of graphs and charts. The important feature of tableau is the discovery of data and its analytics gives useful information in few seconds. We have the option of combing multiple database and analyzed data are presented visually in different formats (Huddar et. al, 2013). It can be easy to understand by a common man. Any complex business problems can be analyzed and presented easily also. The analyzed data can also be shared with others.

Tableau can be an important application in many sectors like industries, organization and business analytics area. Some of the unique features of this tableau are velocity, self sustainable, visual analytics of data, combining different data, compatibility, and common data storage.

Velocity analysis: Continuous flowing of data can be stored and analyzed shortly also the information can be mined from those analyzed data in few seconds.

Self sustainable: Tableau doesn't require any special software for installing. It can be easily installed and analytics can be done.

Visual analytics of data: In Tableau data can be analyzed in the format of tables and graphs and these data can be presented visually in different color formats.

Combining different data: unstructured data and semi structured data from real time sources can be easily handled and these data collected from different sources are accepted and analyzed (Bobade, et.al., 2016).

Compatibility: Tableau can be easily accepted by any kind of devices where the real time data flows from different sources. It can be easily blended with any kind of hardware and software easily.

Common data storage: Data can be stored in a centralized server where all the updating, deletion, can be easily managed

R Programming: R is said to one of the statistical analysis tool and the reports are presented visually by means of graphical presentation. R has a separate computer programming style. R programming is free and open source software, which is compatible for all operating systems. R is named by the first two letters of the author namely Robert gentle man and Ross Ihaka in Bell labs. R can be integrated with all other programming languages easily. R requires some time for the user to learn the programming language. R can also be a comfortable zone for all the programmers. Some of the features of R Programming are:

- R is simple programming language which has the concepts like how all other programming language.
- Effectively handling of data and multiple storage capacity could be possible in R programming.
- R has the capacity of integrating all the other data analytics tool easily.
- Graphical presentation and visual analytics of the data could be easily presented by R programming.

KNIME: KNIME stands for Konstanz Information Miner it is been developed at the university of Konstanz, Germany. KNIME is considered to be the advanced analytic tool. KNIME is a free and open source platform for data analytics. In KNIME machine learning and data mining concepts can be integrated easily. KNIME uses a separate graphical user interface and also a set of nodes which integrates different set of sources which are ETL (Extraction, Transformation and Loading) for presenting the data visually in an understandable manner. In KNIME analytics platform it provides every feature easily to the end user. KNIME is a leading analytical platform which mines the knowledgeable information hidden in the stored data

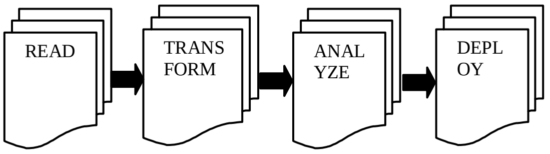

Fig. 2 Flow of KNIME

Some of the analytical features of KNIME are powerful, scalable, data integration, tool integration and visual.

Table 1: Comparison of three big data tools

Tools	Mode of software	Data sources	Real-Time Analytics	Owner	Scalability
Tableau	Open and Free	Structured / Semi-structured / unstructured	Yes	Chris Stolte	Yes
R Programming	Open and Free	Structured / Semi-structured / unstructured	Yes	Robert gentle man and Ross ihaka	Yes
KNIME	Open and Free	Structured / Semi-structured / unstructured	Yes	Konstanz Information Miner	Yes

Powerful: Native nodes, community collaborations and tool integration make KNIME more powerful. KNIME is used by many users for a long time so it is considered to be the most trustful data analytics tool.

Scalable: Streaming of big data, real time data and new technologies can be easily integrated by KNIME with the existing infrastructure hence, it is said to be most scalable.

Data Integration: Text files, databases, documents, images and hadoop based files can also be easily integrated.

Tool Integration: Many recent tools like legacy scripting, code can be easily integrated. Code reusing concepts can also be implemented in KNIME.

Visual: Results are presented graphically in an understandable manner.

The main objective of this comparison between these three tools is not to criticize, but to tell its basic usage and to create awareness of these tools in various fields.

4. Conclusion

In this paper big data, data analytics and comparison between three data analytics tools are discussed. Also discussed the challenges faced by the continuous flow of unstructured, semi structured, structured data in big data analytics and the 4 Vs of big data. The main objectives of this comparison are not to highlight which is the best but to create an awareness of these tools in various fields. Based on these analysis it is stated that some of the advantages and features of these three data analytics tools. These three data analytics tools can be used in various applications efficiently and it can give information in many business sectors. In future, a detailed study can be made to compare various other data analytics tools.

Reference

[1] Avita Katal., Mohammad Wazid., R H Goudar., Big Data: Issues, Challenges, Tools and Good Practices, IEEE , PP 404 -409, (2013).

[2] Bakshi Rohit Prasad., Sonali Agarwal., Comparative Study of Big Data Computing and Storage Tools: A Review, International Journal of Database Theory and Application, Vol.9, No.1, pp.45-66, (2016).

[3] Gartner, Big Data Definition, http://www.gartner.com/it-glossary/big-data/

[4] Mahesh G Huddar., Manjula M Ramannavar., A Survey on Big Data Analytical Tools, International Journal of Latest Trends in Engineering and Technology (IJLTET), Special Issue - IDEAS, ISSN: 2278-621X, PP 85-91, (2013).

[5] Varsha B. Bobade, Survey paper on Big Data and Hadoop, International Research Journal of Engineering and Technology, Volume 03, Issue: 01, pp.861-863, (2016).