Research on Various Tools in Big Data

R.S. Karthiga, Senthil Kumar Janahan, U.V. Anbazhagu

Abstract--- The field big data has wonderful changes in the last few years. Data are collected very massive amount and cheaply through network devices such as mobile, camera, microphone, software logs etc. Information is coming various sources and also the nature of the information is different. It is very difficult to perform with very huge data sets and different nature by using traditional data application software techniques. For many field we need to take out the valuable information from enormous and noisy data sets. In this paper we analyses the characteristics among four different tools and the comparison value is very useful to determine the efficient analytic tool. Based on the value which made user to select the best tool for to advance the performance of big data in a easiest way.

1. INTRODUCTION

Big data is an enormous amount of information's which organised, unorganised and semi-organised data. In many fields of science such as chemistry, zoology, computer, physics, industry etc are generating enormous quantity of information which is organised and unorganised data. Traditional approaches are very hard to perform in massive data set. Big data analytics is the difficult method to inspect enormous data and different data sets to discover form, connection that could be unseen and that may deliver valued vision to the user. Big data platform is encompasses various tools and techniques into single packed result that help for managing and analysing data. These information runs in various platform and domain such as cloud computing, Apache spark and Mango DB.

2. BIG DATA ANALYTICS TOOLS

There are many big data analytic tools are used to explore the concealed data and perform processing of data. Some of the tools are made to be analysed are Apache spark, Sky tree, R programming tool, Mango DB and Rapid miner.

(i) *Apache spark* is developed at California University and after that spark source code was given to Apache software foundation. Spark and Resilient Distributed Dataset (RDD) were developed in 2012. It is processing of large dataset 100 times faster than Hadoop. API easily accesses the spark.

(ii) *Sky tree* is predictive analytic software. California based software company introduce and issues sky tree, machine learning platform for analytic purposes. Cloudera, Amazon EMR, Horton works and Map R qualified Sky tree is a Hadoop. In the market, it has proven that Sky tree is fastest ML software. In machine learning big data sky tree is an end to end organisation platform. Logic inside ML conclusion can be easily envision and recognize by sky tree. Sky tree records each and every data set, split the data, apply transformation, run algorithm and obtain the result for every data sets construct with sky tree.

(iii) *Rapid Miner* In 2001, at AI unit of technical university of Dortmund, Ralf Klinkenberg, Ingo Mierswa, Simon Fisher unit developed Rapid miner previously called YALE (YET ANOTHER LEARNING ENVIRONMENT). In 2007, YALE is changed into Rapid-I by the same company. In 2013, Rapid-I is rebranded into Rapid. It is also data science and predictive analytic software. It offers data mining and machine learning processes. For designing and implement analytical processes it offers graphical user interface.

(iv) *Mango DB*is developed by Mango DB Inc. Apart from traditional database method of loading data into columns and rows, Mango DB store enormous amount of data as individual document collection. It offers automatic scalability, great performance and great availability.

(v) *R – Programming Tool* is established by R core team. It provides platform for statistical computation of big data with high performance. It allows large scale of statistical examination and data visualization. It is essential tools for economics and analytic driven organization such as Facebook, Google and LinkedIn.

Big data analytic tools are used to analysis data such as examining, cleaning, transforming and modelling data to explore the useful information for achieving the backing of decision. It reduces the work of data scientists to get the best solution and enhance the business outcomes.

Table 1: Big Data Tool Characteristics

BIG DATA TOOL CHARACTERISTICS	DEVELOPER	PROGRAMMING LANGUAGE	CURRENT VERSION	COMMUNITY SUPPORT	OPERATING SYSTEM
APACHE SPARK	Apache software foundation	Java, Scala and Python	Spark 2.4	Distributed File System	Cross Platform
SKY TREE	Skytree INC.	SAS, R, Java and Python	Skytree 16.0	Relational database, HADOOP system, Flat file system	Cross Platform
RAPID MINER	Rapid Miner	Language independent	Rapid Miner 9.2	All Relational Database System	Cross Platform
MANGO DB	Mango DB Inc	Ruby, Java Script, Python	Mango DB 4.0	Grid File system	Cross Platform
R PROGRAMMING TOOL	R Core team	C, Fortan	R 3.5.2	Platform Independent	Cross Platform

Revised Version Manuscript Received on April 12, 2019.

R.S. Karthiga, Department of CSE, VISTAS, Pallavaram, Chennai, T.N, India. (e-mail: rskarthigacse@gmail.com)

Senthil Kumar Janahan, Department of CSE, VISTAS, Pallavaram, Chennai, T.N, India. (e-mail: skumar.se@velsuniv.ac.in)

U.V. Anbazhagu, Department of CSE, VISTAS, Pallavaram, Chennai,T.N, India. (e-mail: anbazhagu.se@velsuniv.ac.in)

3. BIG DATA ANALYTIC CHARACTERISTICS & RESULTS

Big data analytic is a group of processes that are associated with industry. Big data analytic tool is used to develop the business outcomes by decreasing the work of data scientists. There are various characteristics of big data analytic tool. Based on these characteristics we can easily analysis the performances and efficiency of the tool. They are,

(A) VOLUME

In big data there are large amount of information. Volume of data can be categorized as megabyte, kilobyte, terabyte, petabyte etc. Volume has not much problem when compare to other characteristics of V features. Every day each user create enormous amount of data. The major problem is determined by decreasing storage rate.

(B) VELOCITY

It denotes the users how fast the data to be generated. Data velocity is fundamental task for some organizations. Many social medias are did millions of photos uploaded and billions of searches on every day. It is similar as nuclear explosion. Big data assist the organization to hold this blast, receive the arriving stream of information and at the same time process it fast so that it does not generate blockage.

(C) VARIETY

Data is produced either by human beings or by machines. Received data is classified into various categorize. Variety is mention as structured data, or unstructured data. Structured data are such as image, text and videos. Unstructured data are such as audio, hand writing text, ECG reading and emails. Various unstructured data causes definite problems for storage, mining and analysing data.

(D) VARIABILITY

It describes the essential to acquire significant data under all probable situations. It mentions to found structure data in case of extreme unpredictability situations. Big data is variable due to gathering of data element resulting from multiple dissimilar data category and sources.

(E) VERACITY

All the above characteristics of big data analytic tool increases but veracity decreases. It is very similar to validity. It describes origin or consistency of the data sources, its circumstance and how significant it is to the analysis based on it. Based on the information of data veracity we can know the risk associated with analysis and help to take decision making based on those data sets.

(F) VALUE

Value of the big data is used to understand the consumer better, aiming them consequently, enhancing processes and improving machine or well performances. It is change a business to more competitiveness in world-wide stand. It suggests that big data bring big social value. There is pure connection between data and its visions.

(G) VISUALIZATION

It is very helpful to obtain correct vision of the data. Because of generating enormous of data in every second, it is essential to predict the data to trace the style, outlier, and pattern interrelate with it to obtain the correct decision. It is useful to remove the noise in your data to take the important value or pattern efficiently and faster.

(H) CONNECTIVITY

It is a very important feature of big data analytics. It is more essential to access the connectivity or how well the product can easily access the other system.

(I) COST

The costs of big data analytic tools are extremely variable. In business various analysis methods is used to enhance cost. Cost can influence on order, market stake, market diffusion, incomes, trades and turnover of essential business metrics. There are five steps influence big data analytic to enhance cost of your business. They are leveraging the power of data, be product specific, segment and separate, ensuring stock clearance while maximizing margin and reduce cost. Based on cost paid, a dealer will offer different properties, abilities or free from some limitations like examined data capacities.

(J) CONVERSION

Conversion characteristics of big data analytics is mainly used for converting one form of data into another form that able the user to run in any platform. Spark SQL tool in apache spark used to convert the data. The rapid miner converts nominal to numerical values. This conversion tool in big data analytic translates individual form of function and data into alternative form of data type which is needed by user and enables ease of use running program in any platform.

(K) SIMPLICITY

It enables very simple easy designing and implements the task very cosy to user while using the tool. In big data analytic data visualization, main key to achievement is simplicity and clarity.

(L) PRODUCTIVITY

Big data is very important in obtaining productivity and effectiveness improvement and finding new vision to drive invention. In big data analytic, manufactures can realize new data and recognize pattern that able them to develop method, rises supply chain efficiency and find the items that disturb production. Mean while production gain depends on maximizing the worth of resources; asset performance improvements can lead to huge productivity enhancement. Big data analytic also disclose needs, permitting manufacturer to boost production processes and generate another strategy to address drawback.

(M) METHODS

One of the tasks in big data is how to examine big data application performances in order to determine the important issue that disturb the quality of them.

Due to various reasons big data analytic can be very complex. To obtain the performance requirement of the organization it is essential that organization is planned and construct to meet these performance requirements.

(N) PLATFORM

Big data platform is a kind of information technology solution that associations the properties and skills of various big data application and services within a single solution. It is an enterprises class IT platform that allow the organization in developing, installing, operating and managing big data setup. Big data also maintain tradition improvement, demanding and incorporation with other structure. The main advantage behindhand of big data platform is to decrease the difficulty of various dealers into single organized solution.

Table 2: Big Data Characteristics

BIG DATA CHARACTERISTICS	VOLUME	VARIETY	VELOCITY
APACHE SPARK	Terabyte to peta byte	Structured, Unstructured data	100 times faster than Hadoop
SKYTREE	Massive amount of data	Structured, Unstructured data	10000 times faster than previous approaches
RAPIDMINER	Few hundred megabytes	Unstructured data	10x on average, 16x up to some of data
MANGO DB	Thousands of nodes, peta bytes	Unstructured data	It is faster than MYSQL.
R PROGRAMMING TOOL	Gigabytes	Structured, Unstructured data	It allow parallelization operation

Table 3: Big Data Characteristics

BIG DATA CHARACTERISTICS	SIMPLICITY	PERFORMANCE	SOFTWARE TYPE
APACHE SPARK	Easy to use for developer	Faster performance	Open source
SKYTREE	7.6 rating for easy for use	7.4 rating for performance	Open source
RAPIDMINER	Easy to use	Efficient performance	Open source
MANGO DB	Simple to install and implement	Efficient performance	Open source
R PROGRAMMING TOOL	Easily implement	Slow compared to other programming languages	Open source

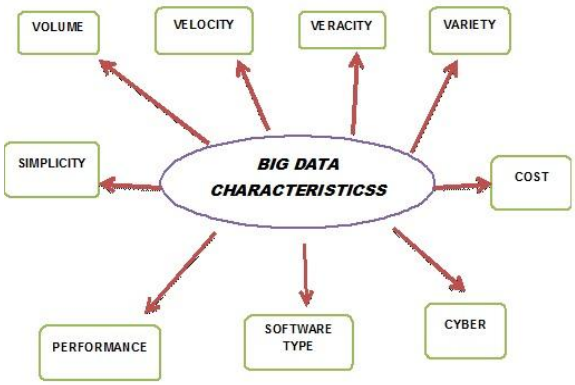

Fig. 1: Big Data Characteristics

We have discussed the characteristics of big data and its analytic tools which are used to promote the functionalities of big data. The numerous characteristics of several big data analytic tools are matched and identify which top tool the companies is used. This analysis done for each analytic tool based on the demand of customer and corporate, the project which has been done. The above tools are different categories based on the many characteristics. The tools are used to analyse complex data and also different area. All the areas use this tools which can saving the time, cost and improve the production of the organization.

4. CONCLUSION

Big data analytic play a central part in all the organization. By this current product can be changed with new advanced features. It is easy to use for both customer and service providers. Even in the complex data, the big data analytic tools produce more accurate result and improve the process. The user and data scientist can select any of the tools based on the comparison and also for future empowerment additional characteristics added to the tools invented. The supply and request of the tools are increasing by everyday among customers due to the simplicity and performances of the big data analytics tool.

REFERENCES

1. A COMPREHENSIVE SURVEY ON BIG DATA ANALYTICS TOOLS J. Vijayaraj, R. Saravanan, P. Victer Paul, R. Raju Department of Information Technology Sri Manakula Vinayagar Engineering College, Puducherry India.
2. Nawsher Khan,et.al,Big Data: Survey, Technologies, Opportunities, and Challenges, Hindawi Publishing Corporation, the Scientific World Journal, Volume 2014, Article ID 712826, 18 pages.
3. A Review paper on Big Data: Technologies, Tools and Trends Anurag Agrahari1, Prof D.T.V. Dharmaji Rao2 1. M.tech Student, Dept. Of Computer Sci & Engg, AITAM College, Tekkali, Srikakulam, Andhra Pradesh, India. 2. Professor, Dept of Computer Sci & Engg, AITAM College, Tekkali, Srikakulam, Andhra Pradesh, India
4. Comparative Study on Tools and Techniques of Big Data Analysis B.THILLAIESWARI M.S., M.Phil., B.Ed., Assistant Professor, Department of Computer Science, TBAK College for Women, Kilakarai
5. "Big Data for Development: Challenges and Opportunities", Global Pulse, May 2012 Yuri Demchenko ―The Big Data Architecture Framework (BDAF) Outcome of the Brainstorming Session at the University of Amsterdam 17 July 2013.
6. Joseph McKendrick, "Big Data, Big Challenges, Big Opportunities: 2012 IOUG Big Data Strategies Survey", IOUG, Sept 2012.
7. James Nunns, "10 of the most popular Big Data tools for developers", Last Accessed 24 November 2015.
8. Menon, S.P.; Hegde, N.P. "A survey of tools and applications in big data "Intelligent Systems and Control (ISCO), 2015 IEEE 9th International Conference, pp: I - 7.

9. Sofiya Mujawar, Aishwarya Joshi." Data Analytics Types, Tools and their Comparison" IIJARCE 2015Vol. 4, Issue 2, pp. 488-491.
10. Nigel Wallis, "Big Data in Canada: Challenging Complacency for Competitive Advantage", IDC, Dec 2012.
11. COMPARATIVE STUDY OF BIG DATA ANALYTICAL TOOLS J.VIJAYALAKSHMI Ph.D Research Scholar Department of Computer Science Alagappa University,Karaikudi. Email:viji.jayaprakash@gmail.com Dr. E. RAMARAJ Professor & HOD Department of Computer Science Alagappa University Karaikudi.