Comparative study of tools for Big Data Analytics: An Analytical Study

Sanjib Kumar Sahu
Dept. of computer Science, Utkal University, Odisha(India)
sahu_sanjib@rediffmail.com

M Mary Jacintha
School of Information technology CDAC, Noida(India)
maryjacintha03@gmail.com

Amit Prakash Singh
University School of Information and Communication Technology, India
amit@ipu.ac.in

Abstract—**Data sets grow rapidly in different forms due to digitalization. When the data or information sets which are too large and are complex in nature in which traditional data processing techniques are not able to deal with those complex data, then that data is called Big data. Researchers, scientists, business organizations, government agencies, advertising agencies, medical researchers often come across more difficulty in dealing with data for any decision making. The data available for research has to be processed by using various techniques of data analytics which is called Big Data Analytics. These techniques helps in getting benefits in dealing with massive volume of either unstructured, structured or semi-structured data content that is fast changing nature, also not possible to process using conventional database techniques. This paper discusses the major utilization of big data analytics by comparing different tools available for big data validation. Furthermore, this paper discusses the case study conducted to overcome the big data challenges and needs.**

Key Words----Big Data Analytics, Data Sets, Challenges, big data validation, unstructured data, multiple techniques, semi-structured data.

I. INTRODUCTION (BIG DATA)

The ever increasing use of internet, sensors, and heavy machines at a very high rate with sheer volume, velocity, variety and veracity, the data is termed as Big Data. Data is everywhere, in every industry, in the form of numbers, images, videos, and text. As data continues to grow, it becomes difficult for computing systems to manage big data due to immense speed and volume at which it is generated. As the data is enormous and complex, the data is stored in a distribute architecture file system. Analyzing the complex data is a risky and time consuming task as it contains big distributed file systems, which should be fault tolerant, flexible, and scalable. The process of capturing or collecting big data is known as 'datafication'. Big data is datafied so that it can be used productively. Big Data cannot be made useful by simply organizing it, rather the data's usefulness lies in determining what we can do with it.

Big Data is being engendered by almost everything around us, like social network, government, healthcare, education at an alarming volume with high velocity and variety. To pull out meaningful assessment from this enormous data, it is necessary to do best possible processing control, analytical potential and skills.

As big data contains data which is big in size, the selection of right data within the larger data set to analyze the whole data should be appropriate. Predictive analytics and data mining solutions for the enterprises are currently available from a number of companies, like Predictive analytics Suit, IBM SPSS Statistics, Microsoft Dynamics CRM Analytics Foundation[1].Software on big data platforms and big data analytics focus on giving efficient analysis on data on enormously big datasets. The industries like banking, automobiles, healthcare, telecom, government, transportation and travel will have major impact through Big Data analytics(BDA). Its perspectives are to provide assistanceto number of industries to extract data into high-quality information for in-depth approach into their organizations status[2].

Global phenomena of using Big Data to gain business value and competitive advantage will only continue to grow as will the opportunities associated with it. As per MGI and McKinsey's Business Technology Office research, the utilization of enormous Data is most likely to become a key basis of competition for individual firms for success and growth and strengthening consumer surplus, production growth and innovation. The means of using Big Data is by selecting appropriate Big Data analytics platform and the tools which is very critical for an organization. Although there has been a numerous tools are available, we made an attempt to highlight some of the tools for BDA.

The rest of this paper is organized as follows. In Section II, we give an overview and basic understanding of Big Data Analytics tools. In Section III, we discuss the comparison of various software tools in analyzing big data. In Section IV, we present some related work on big data validation tools along with an analytical study is presented. Finally, Section V concludes the paper by providing some future perspective to overcome the global challenges.

II. UNDERSTANDING OF BIG DATA ANALYTICS TOOLS

The scalability of increase in volume, velocity and variety of data in an organization will be benefited by selecting appropriate big data technologies. Appropriate selection of tool will be the basis of global competition result in optimum investment in big data analytics, production growth and strengthening consumer surplus.Big Data Analytics tool made

ISBN:978-1-5090-6471-7/17/$31.00 ©2017 IEEE

the entire data management cycle technically and economically feasible from collection and storing of larger datasets to analyze the data inorder to provide new and valuable insights.

A. Process of Big Data Analytics Tool

The process of Big Data Analytics tool involves the data flow from collection of data from the larger dataset to provide valuable information for decision making.

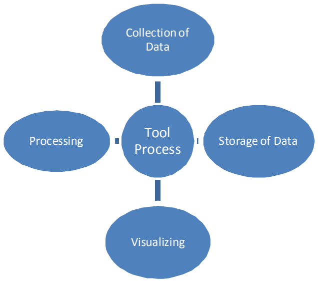


Fig. 1. Big Data Analytics Tool Process

Collection of data-. The structured, unstructured and semi-structured data available from three major sources: a) social data which is the data generated from twitter, facebook, linken and google. b) machine data which includes data generated from enterprise resource planning, global positioning system, weblogs. c) transactional data which includes data generated from amazon, walmart, ebay.

Storage of data-Securing enormous data before and after the analysis process needs a platform with secure, scalable, and durable As per the requirement of an organization, the presence parallel file systems which scales to billions of files and terabytes of capacity allowslarger data sets to link together across locations and interrogated.

Processing-The structured, unstructured and semi-structured data is transformed to a valuable format by means of categorizing, summarizing, matching-up and performing advanced functions and algorithms. The processed datasets are then used for business intelligence processing and also for data visualization using appropriate tools.

Visualizing-. To get accurate and valuable insights of the data, the availability of new in-memory uses data visualization for better way to analyze the data more quickly than ever before.

B. Selection of Big Data Analytics tool

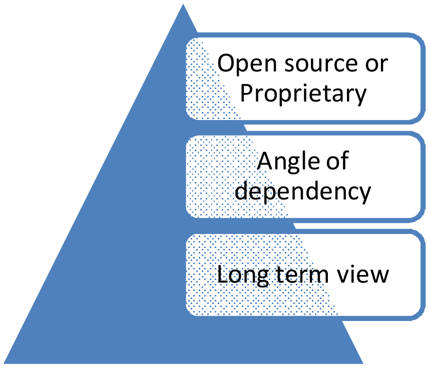


Fig. 2. Pyramid approach of selection of tool

The pyramid approach of selection of tool for BDA is the best practices for the selection of appropriate tools. As open source tools and technology is held themselves to payoffs on price and vendor lock-in, it is preferable than the proprietary tool. While selecting the tool, the angle of inter-tool dependency has to be tested as big data will require complex integration. A big data project is an ongoing long-term project which requires a balanced architectural view of the tool.

III. SOFTWARE TOOLS IN ANALYSING BIG DATA

There are numerous Big Data analytics tools are available with different vendors. All the tools are very promising by saving the time taken from data retrieval to data visualization; same money and provide uncover never-before-seen business insights. Previous studies indicate that there is no specific or valid tool for particular organization. Here the tools are compared in terms of the Big Data Analytics tool process: Collection of data, storage of data, procession of data and data visualization.

The big data technologies available to assess the data include Hadoop, Cloudera, MongoDB, Tableau, silk, CartoDB. [3]. The importance of big data analysis was highlighted through six fields such as structured data analysis, text data analysis, web data analysis, multimedia data analysis, network data analysis and mobile data analysis. The enhancement of framework of the Hadoop-MapReduce framework was observed[4] it was discussed about the comprehensive data processing mechanisms which was implemented using big data. He emphasized that the entire life cycle of the big data such as Data Collection, Data Storage, Data Processing, and Data Visualization is supported by big data analytics[5]. The comparison of Big Data tools along with data process cycle is discussed in Table number 1.

TABLE I. A COMPARISON OF BIG DATA ANALYTICS TOOLS ON THE BASIS OF DATA PROCESS

Process	Tools	Usage	Data Extraction ways
Data Collection	Import.io	A powerful tool to extract data from WebPages. It quickly create an API(Application Programming Interface) to a webpage. The API defines what to extract from a page and which page to be converted to data and run those as queries through the API.	Single URL Bulk Extract URLs from another API
Data Storage	Hadoop	An open-source software tool for distributed storage of very large datasets on computer clusters. It helps in scale the data up and down without hardware failures. Large amount of data can be stored for different kind of data, able to handle data without limit, simultaneouswork or tasks.	MapReduce- a parallel processing software framework. YARN- Yet Another Resource Negotiator provides resource management for the processes running HDFS- Hadoop Distributed File System, a Java based scalable system
	cloudera	It is a recent platform for data management and analysis within limit or in cloud.	It builds big data applications on Apache Hadoop with the most recent open source tools. It increases the skill for the formulation of best alternate management strategy.
	MongoDB	Start-up approach to databases. It manages data which changes frequently and the unstructured or semi-structured data.	For delivering a single view across multiple systems, MongoDBhelps in storing data. It helps for mobile apps, product catalogs, real-time personalization, content management and applications delivering.
	Talend	An open source product focusing on their Master Data Management (MDM). It simplifies real-time data integration for superior analytics and the real-time use cases that are driving business innovation	Talend real-time Big Data platform delivers high-scale, in-memory fast data processing to turn more data into business decisions, in real time at scale. It helps in self service data preparation.
Data Processing	Qubole	It is cloud based Hadoop platform which works with both structured and unstructured data. It makes things easierto speeds-up and scales big data analytics workloads in opposition tothe data stored on AWS, Google, or Azure clouds. It takeshassle out from communication setback.	It has a user interface which allows user to analyze the data sets in the absence of Hadoop system. It allows the user to integrate and consolidate the data available from various sources simultaneously.
	BigML	It makes things easierto machine learning and put forward a dominant Machine Learning examination with anuser-friendly interface to import the data available and to take decision from the results. The models are used for doing predictive analytics.	It is used to import data, take decision making and forecasting. It provides models for predictive analysis.
	Statwing	A new level data analysis is done through Statwing. It provides simple to complex visual analysis.	It has blog post on NFL data which is trouble-free to use. This can actually be started with Statwing in less than 5 minutes.
Data Visualization	Tableau	A data visualization tool with primary focus on business intelligence.	It is used to generate maps, bar charts, scatter plots and more without the need for programming. It recently released a web connector that allows linking a database or API to give the ability to get live data in visualization.
	Silk	It is also an data visualization tool much better than Tableau and more user friendly than Tableau.	It helps in bringing data to life by creating maps which is interactive and charts without any extra programming. Silk also allows collaborating on visualization with as many users.
	CartoDB	It specializes in making maps.	It is used to visualize location data without coding. CartoDB can manage a myriad of data files and types; it has sample dataset which plays around while the system is getting the hang of it.

IV. APPLICATION CASE STUDY

The visualization of two sets of data using selected tools was conducted with MBA(Information Technology) and MCA students at CDAC Noida. The Big Data Analytics tools for Visualization of data such as, Tableau, Silk and CartoDB were used for analysis. This study was conducted to see the variation in using the tools in two sets of data. Only visualization tools were given to students so as to differentiate the performance visually. The parameters taken for comparison are: Quality- whether the tool is bug free or it contains errors which are not worth to analyse; Efficiency- whether the CPU time utilization, availing memory capacity and storage disk space occupancy are optimum; connectivity-accessibility of the tool in connection with internet and intranet; Documentation and performance- the results provided are well understood and easy to take decisions. Table 2 depicts the comparison of two sets of data in two programmes(MBA(IT) and MCA).

TABLE 2 COMPARISON OF PARAMETERS APPLIED TOVISUALIZATION TOOLS

Parameters	Tableau		Silk		CartoDB	
	MBA	MCA	MBA	MCA	MBA	MCA
Quality	Yes	Yes	Yes	No	Yes	Yes
Efficiency	Yes	No	Yes	Yes	Yes	No
Connectivity	No	Yes	Yes	Yes	Yes	Yes
Documentation & Performance	Yes	Yes	Yes	Yes	Yes	Yes

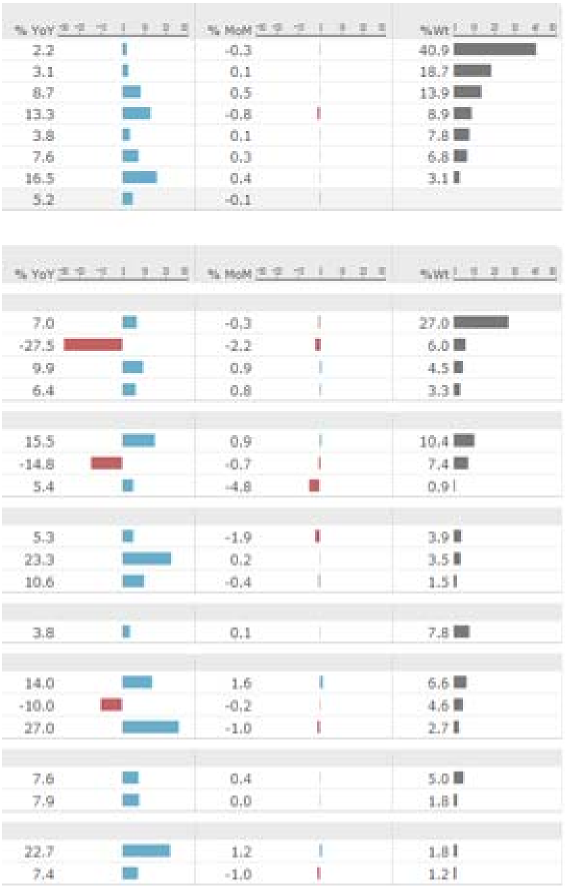
The case study was taken as pilot study to know about the degree of complexity the visualization tool can accommodate. All the three tools used for testing are visualization tool and there is no much of variation in their performance. The visualization tool depicts complex visuals in one place. The table 3 depicts the expenditure for two months September and October.

TABLE 3 MONTHLY EXPENDITURE FOR TWO MONTHS

Major Groups	Sep'16	Oct'16
Food	2532	2832
Hospitality	2259	2279
Household goods	925	1019
pharma and watches	600	652
departmental stores	482	509
clothing	421	511
recreation	190	210
total retail	7409	8012
Group	Sep'16	Oct'16
Food		
Supper Market	1675	1752
Grocery Stores	374	452
Soft and hot drinks	282	332
other food	202	242
Hospitality and other services		
Hotels and clubs	647	747
restaurants	456	556
other services	56	76
pharma and watches		
Retailing	243	253
Cosmetic	216	241
Watch &Jewellery	93	109
Departmental Stores		
Stores	482	490
Household goods		
furniture	457	551
domestic ware	282	323
other domestic items	169	179
Clothing		
Clothing	308	343
other things	123	153
Recreation		
Stationary	113	113
Recreation items	77	97

The data at Table 3 are analysed by using visualization tool and it provides vaious information in one place. The visual at Table 4 gives various analysis about the data.

TABLE 4 ANALYSIS OF DATA USING VISUALIZATION TOOL



The results at Table 4 provides the following information:

1. Comparison in percentage for two months
2. Bar chart showing the difference in expenditures
3. Line graph showing the variation

So, Visualization tool can able to provide multiple analysis at one place for the same data.

V. Conclusion

This paper explores various Big Data analytics tools in terms of Big Data Process. It proposes a framework for the selection of tool as each stage in the data process should use appropriate tool for that stage for optimum utilization of CPU time, cost and accuracy. The case study conducted proved that the efficiency in tool selection and utilization will lead to efficient management of data and decision making. It is expected that future research may be carried out in analyzing industry-wise Big Data Analytics tool which is very much in need as there are numerous tools are available: open source and private.

Refernces

[1] Imanuel, 2015. 43 bigdata platforms and bigdata analytics software. Retrieved on November 16,2015 from http://www.predictiveanalyticstoday.com/bigdata-platforms-bigdata-analytics-software/

[2] Minelli, M., Chambers, M., & Dhiraj, A. (2013). Big data, big analytics: emerging business intelligence and analytic trends for today's businesses. John Wiley & Sons.

[3] Chen,M.,Mao,S.,&Liu,Y.(2014).Big Data: A Survey. Mobile Networks and Applications, 19(2). 171-209 doi:10.100/s11036-013-0489-0.

[4] Zhao, L., Sakr, S., Liu, A., & Bouguettaya, A. (2014). Big Data Processing Systems. In Cloud Data Management (pp. 135-176). Heidelberg: Springer.

[5] Brandon, G. (2015). Guide to big data analytics: platforms, software, companies tools, solutions and hadoop. Retrieved on November 16, 2015 from http://cloudnewsdaily.com/big-data-analytics/

[6] Che, D., Safran,M., & Peng, A (2013). From Big Data to Big Data Mining: Challenges, Issues, and Opportunities, In B.Hong et al. (Eds.), Database systems for Advanced Applications (pp. 1-15), Heidelberg: Springer.

[7] Demchenka,Y.,Grosso, P.,Laat, C., & Membrey, P.(2013). Addressing Big Data Issues in Scientific Data Infrastructure. In 2013 International Conference on Collaboration technologies and Systems(pp.48-55). New Yokr:IEEE.

[8] Elgendy,N.,&Elragal,A.(2014).Big Data Analystics: A Literature Review Paper. In P.Perner(Ed.), ICDM 2014. LNAI, Vol8557(pp.214-227).Heidelberg:Springer.

[9] Henschen, D. (2014). 16 top big data analytics platforms. Retrieved on November 15,2015 from http://www.informationweek.com/big-data/big-data-analytics/16-top-big-data-analytics-platforms/d/d-id/1113609

[10] Kaisier, S., Armour, F.,Espinosa,JA., & Money, W(2013). Big Data: Issues and Challenges Moving Forward: In 46th Hawaii International Conference on System Sciences(pp.995-1004). New York:IEEE.

[11] Roggero. H (2015). Sample pricing comparison. Retrieved on November,16, 2015 from http://geekswithblogs.net/hroggero/archive/2015/08/12/sample-pricing-comparison-2-amazon-aws-and-microsoft-azure.aspx.