INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 8, Issue 2 - V8I2-1268)
Available online at: https://www.ijariit.com

Analytical study of tools for Big data analytics

Shubham Milind Lokhande
lokhandes1404@gmail.com
Patkar Varde College, Mumbai, Maharashtra

ABSTRACT

Due to digitalization there is a rapid growth of data sets into different forms. When there are instances of a large and complex amount of data and the traditional data processing techniques are failing to deal with these types of complex data, this data is knownas Big Data. For any decision making in sectors like advertising, government agencies, medical researches and such wecome across big data. The given data for the research has to be processed using various techniques of data analytics which is known as Big Data Analytics. These techniques are used on structured or unstructured available data which arenot possible to be processed using traditional methods. In this paper I'll be discussing some of the major utilization of big data analytics by comparing different big data tools used for validating the data. I'll also be discussingsome of the big data challenges and needs.

Keywords *– Data Sets, Big Data Validation, Big Data Analytics, Structured And Unstructured Data, Semi-Structured Data.*

1. INTRODUCTION

Day by day the use of internet, sensors and heavy machines, all requires stored data which is known as big data because of the volume of the data available. Data is everywhere in the form of numbers, videos, images and texts. As the data is growing rapidly it becomes very difficult for computing systems to manage this big data due to the immense speed and volume at which is been generated. As a result the complex nature of the data is also increasing.Analyzing this complex data is very time consuming and there is a lot of risk of losing data as it is very big. The process of collecting this big data is called as 'datafication'. For using this big data in a productive way it is 'datafied'. Simply organizing this big data does not help, we need to determine what to do with the data and how we can use it effectively.

Because of the social network, healthcare, education and the government with their huge volumeof data with high velocity and variety, big data has been engendered [1]. For making meaningful use of this data and assessing it correctly, it is necessary for having the best possible processing controls and analytical potential. Selecting the right data withing the larger data set available to analyze the whole data is very important. There are multiple companies whichare providing predictive analytics and data mining solutions for various enterprises. Software used by thebig data platforms on the data focus on giving efficientanalysis on the datasets available. Globally Big Data is used to gain business value and competitive advantages which are going to grow more eventually [2]. Selecting the appropriate Big Data analytics platform and tools is very critical for any organization.In this paper some of these tools will be highlighted.

2. BIG DATA ANALYTICS TOOLS

The increase in the volume, variety of the data and the velocity in an organization will be usefulby selecting appropriate big data technologies. Selection of appropriate tools will be the basis of the result with optimum investment in big data analytics, the growth of the production and strengthening the consumer surplus. Big data analytics tools have made the entire data management cycle feasible from collection and storing larger datasets to analyze the given data sets in order to provide new and valuable insights.

There is a certain process in which big data analytics tools are used in for efficient data flow from collectionof the data to the end result. They are as follows:

Collection of data: It can be structures, unstructured or semi structured. This data is mainly acquired from three major sources.

1) Social data generated by social websites like Twitter or Facebook. 2) Machine data which is generated from enterprise resources planning, GPS or weblogs. 3) Transactional data which is generated by e-commerce websites like amazon, Walmart or Ebay.

Storing the data: Storing and securing this huge amount of data before and after the analysis and the process is done needs a platform with security, scalability and durability. According to the organizations needs, the file system must have terabytes of capacities which will allow larger data sets to link together across locations and could also beinteracted.

Processing: Certain techniques are used for categorizing,

summarizing, matching and performing advanced functions and algorithms for transforming structured, unstructured or semi-structured data into valuable information. This information is then used bybusiness intelligence for processing.

Visualizing: Data visualization is basically used asa better way of analyzing the given data more quickly and efficiently. Visualization results into a more accurate data and gives valuable insights which help in data correction.

After the data has been processed, selection of the Big Data Analytics tool is the major step in the approach. A pyramid approach is used for the selectionof the BDA tool as shown in Fig. 1.

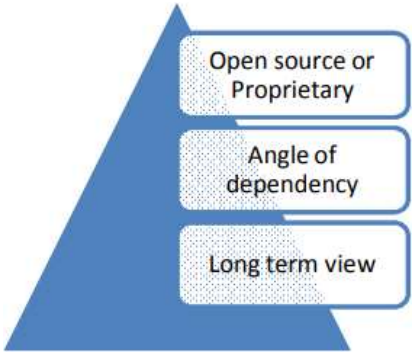

Fig 1: Pyramid approach

The pyramid approach of selection of tools for BDA isone of the best practices for selecting the appropriate tools. Open-source tools and technologies held themselves to payoffs on price and vendor lock-in, it is more preferred over the proprietary tool. When the tool is been selected, the angle of inter-tool dependency is been tested as big data will require

complex integration. A big data project requires abalanced architectural view of the selected tool.

3. TOOLS USED FOR ANALYSING BIGDATA

There are various number of Big Data analytics tools that are available with different vendors. All of these tools save the time taken for data retrieval to the data visualization, also saving money and providing business insights. In this paper I'll be comparing different tools in terms of Big Data Analytics tool process, such as collection of data, storage of data, processing the data and the visualization.

Some of the big data technologies that are currently available to asses the data includes the software like Hadoop, MongoDB. Tableu, Cloudera [3]. There are six different fields which highlight the importance of Big Data Analysis such as structured data analysis, web data analysis, mobile data analysis,network data analysis, multimedia data analysis and text data analysis. Below I'll be discussing the comparison of Big data tools along with it's dataprocess life cycle.

For Data Collection

A tool called Import.io is used for this purpose. It is a powerful tool which is used to extract data from Webpages. It can also quickly make an Application Programming Interface (API) to a webpage. This API is used to define what toextract from the page and which page has to be converted to a data set and run these as queries through the Application. The ways in which this application extracts data is by Single URL, Bulk extract and URLsfrom another API.

Data Storage – For data storage there are four of the tools that I'll be comparing in this paper.

1. **Hadoop** – This is an open-source software tool used for distributed storage of very large datasets on the computer clusters. It also helps in scaling the dataup and down without any hardware complications or failures. Hadoop helps in storing different kinds of data and handle the data with simultaneous work or tasks. MapReduce, YARN and HDFS is used in Hadoop for extracting data.
2. **Cloudera** – This is a recent platform for data management and analysis based on a cloud. It is used for building data applications on Hadoop with the most recent open-source tools [4]. It also increased the
skill for formulation of best alternate management strategies.
3. **MongoDB** – This is a start-up approach to the databases, it basically manages unstructured or semi- structured data which frequently changes [5].
4. **Talend** – It is an open-source product whichfocuses on their master data management. It is used tosimplify real-time data integration for superior analytics and the real time use cases which are drivingbusiness innovation [6]. For data extraction Talend uses in-memory fast data processing to turn more data into business decisions and scale them in real-time.

Data Processing

For data processing there are 2tools that I have compared in this paper.

1. **BigML** – This tool makes the machine learning easier and puts forward a dominant machine learning examination with an user-friendly interface. This toolis usually used for predictive analysis[7]. The data transfer is in a way of importing the data and taking decisions accordingly.
2. **Qubole** – This is a cloud based Hadoop platform which is used for processing structured as well as unstructured data. It is used for speeding-up and scaling big data analytics workloads in opposition to that of the data stored on google or azure clouds. Qubole's user interface allows users to analyze the given data set in the absence of the Hadoop system[8].

Data visualization – For the visualization of datathere are 2 tools that I've compared, they are as follows

1. **Silk** – This tool is a much better data visualizationtool than Tableau and it is more user-friendly. It is used for creating maps which are interactive and charts without any extra programming or coding. It alsoallows collaborations of different users[9].
2. **CartoDB** – CartoDB is mainly specializes inmaking maps. It is basically used to visualize location data without any programming. It can also manage myriad of data files and types [10]

4. CONCLUSION

This paper compares various Big Data analytics toolsin terms of the big data processing of the given data and explores more on which tool is more appropriate in regards with the organization. It gives a frameworkfor selection of the tools which should be of appropriate use for the data process for resulting with optimum utilization of the time, cost and accuracy.

There are multiple Big data analytics tools mentionedin this paper which are both open-source and private.

5. ACKNOWLEDGEMENTS

I would like to thank all the staff members of Patkar Varde College, teachers, my colleagues and classmates for their valuable support in writing this research paper.

6. REFERENCES

[1] "Big data analytics and software",

http://www.predictiveanalyticstoday.com/bigdata-platforms bigdata-analytics-software/

[2] "Big data, big analytics: emerging business intelligence and analytics trends for todays businesses", Minelli, M. Chambers & Dhiraj.

[3] "A survey of mobile networks and application", Chen, M, Mao, S & Liu. doi:10.100/s11036-013- 0489-0

[4] 'Big data processing systems. In cloud data management", Zhao, L., Sakr, S, Liu % Bouguettaya,springer.com

[5] MongoDB for startups, https://www.mongodb.com/startups

[6] "Talend Big Data Basics", https://www.talend.com/academy/calendar/big-data-basics/

[7] https://bigml.com/about/

[8] "Qubole review : self-service big data analytics",Martin Heller, https://www.infoworld.com/article/3449896/qubole- review-self-service-big-data-analytics.html

[9] "Silk- A link discovery framework for the web of data" byJulius Volz, Christian Bizer, Martin Gaedke, Georgi Kobilarov.

[10] https://carto.com/blog/what-exactly-is-big-data/

