IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

A REVIEW ON BIG DATA ANALYTICS: TOOLS AND ITS APPLICATIONS

[1]T.Venkatesan,
Asst.Professor,
Department of Computer Applications,
A.V.C. College of Engineering,
Mayiladuthurai,Tamil Nadu,India
tvntvn09@gmail.com

Abstract —Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics. This useful information for companies or organizations with the help of gaining richer and deeper insights and getting an advantage over the competition. For this reason, big data implementations need to be analyzed and executed as accurately as possible. In this paper, the important characteristics, architecture related to Big Data management has been explored. The various big data analytic tools and techniques have also been discussed here in this work. It is concluded that Big Data Analytics is emerging and highly significant field of research these days. Big Data analytics can have variety of applications in different fields. The review of research carried out by various authors has also been discussed briefly. The motive of this review paper is to help the researchers who knew nothing about this field but want to explore this research area of Big Data Analytics.

Keywords—Big Data Analytics, Big Data Application, Apache Hadoop, Apache Drill, Project Storm

1. INTRODUCTION

The term "Big Data" has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems. They are data sets whose size is beyond the ability of commonly used software tools and storage systems to capture, store, manage, as well as process the data within a tolerable elapsed time [1]. Big data sizes are constantly increasing, currently ranging from a few dozen terabytes (TB) to many petabytes (PB) of data in a single data set. Consequently, some of the difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing[2].

There are various resources of Big Data For Example: Audio, Videos, and Post in Social Media, Various Database Tables, and Email Attachment etc. People uses twitter in diverse form and store 250 Million tweets Per Day. 4 Billion People watching YouTube per Day. Nowadays, Data produced in Zettabytes. Big data has many opportunities like financial services, Healthcare, Retail, Web/social, Manufacturing and Government [3].

Digital universe is flooded with huge amount of data generated by number of users worldwide. These data are of diverse in nature, come from various sources and in many forms. To keep with the desire to store and analyze ever larger volumes of complex data, relational databases vendors have delivered specialized analytical platforms that come in many shapes and sizes from software only to analytical services that run in third party hosted environments. In addition new technologies have emerged to address exploding volumes of complex data, including web traffic, social media content and machine generated data including sensor data, global positioning system data. New non-relational database vendors combine text indexing and natural language processing techniques with traditional database technologies to optimize ad-hoc queries against semi-structured data. Number of analytical platform are available in the market for analysis of complex structured and unstructured data, each of which is designed to handle specific type of data/workload .In this paper we will discuss three open source Big Data Analytics frameworks suitable for different types of workload.

This paper is organized as follows: Section 2 it will discuss Literature survey of Big Data, Section 3 it will discuss architecture and characteristics of Big Data, Section 4 will discusses open source Big Data processing tools. Section 5 contains applications of big data analytics in detail and Section 6 concludes the paper.

2. LITERATURE SURVEY

Over the last many years, there are many researchers has completed their work successfully on big data. Hundreds of articles have appeared in the general business press (For example Forbes, Fortune, Bloomberg, Business week, The Wall street journal, The Economist)[4]. National Institute of Standards and Technology [NIST] said that Big Data in which data volume, velocity and data representation ability to perform effective analysis using traditional relational approaches [5]. In March 2012, The Obama Administration announced that the US would invest 200 Million Dollars to launch a big data research plan [6].

An IDC Reports predicts that from 2005 to 2020, the global data volume will grow by a factor of 300, from 130 Exabyte's to 40,000 Exabyte's, representing a double growth every two years[9]. IBM estimates that everyday 2.5 quintillion bytes of data are created out of which 90% of the data in the world today has created in the last two years. It is observed that social networking sites like Facebook have 750 Million users, LinkedIn has 110 million users and Twitter has 250 million users [7].

3. BIG DATA ARCHITECTURE

Big Data architecture is premised on a skill set for developing reliable, scalable, completely automated data pipelines. That skill set requires profound knowledge of every layer in the stack, beginning with cluster design and spanning everything from Hadoop tuning to setting up the top chain responsible for processing the data. The following diagram shows the complexity of the stack, as well as how data pipeline engineering touches every part of it in Figure 1.

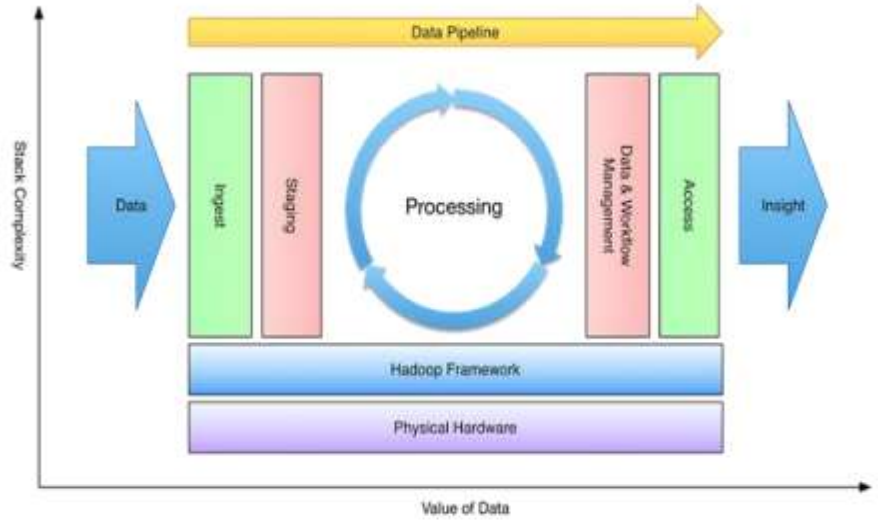

Fig.No. 1. Architecture of Big Data

The main detail here is that data pipelines take raw data and convert it into insight (or value). Along the way, the Big Data engineer has to make decisions about what happens to the data, how it is stored in the cluster, how access is granted internally, what tools to use to process the data, and eventually the manner of providing access to the outside world. The latter could be BI or other analytic tools, the former (for the processing) are likely tools such as Impala or Apache Spark. The people who design and/or implement such architecture I refer to as Big Data engineers.

Characteristics of Big Data

Big Data Big Data can be characterized by different aspects. The commonly used aspects are Volume, Velocity and Variety. Veracity and Value are also used to characterize Big Data. They are helpful lens through which we can understand the nature of Big Data and the platform available to exploit them.

Volume- As infrastructure becomes increasingly available and afford able, data generated by different sources is very huge in size; petabytes or zettabytes. This huge amount of data is called Big Data.

Velocity - The sheer velocity at which we are creating data is huge cause of Big Data. Digital universe expands from 130 million to 40 trillion in 8 years (2005-2013). The data generated from various sources range from Batch to Real time. So this high velocity data defines new term called"Big Data" .

Variety - The representation of data generated by various sources are diverse in nature; for example ecommerce web sites deal with structured data, web server logs deal with semi structured data and social websites deal with unstructured data like audio, video, images etc. Hence big data can be categorized into structured, unstructured and semistructured types and digital universe deals with combination of all.

Veracity - Duo to sheer velocity of some data we cannot spend time in cleans the data before using it. Compiling multisource data and use it for decision making for business requires mechanism that deals with imprecise data. Hence combination of precise, imprecise, accurate, data can be called big data.

Value - By processing huge volume, high velocity, variety and veracity of data, presents a new dimension for analyzing big data called "value". Collaborating different types of data, putting them all together in order to extract hidden knowledge for business and getting competitive advantage from it represents value of big data.

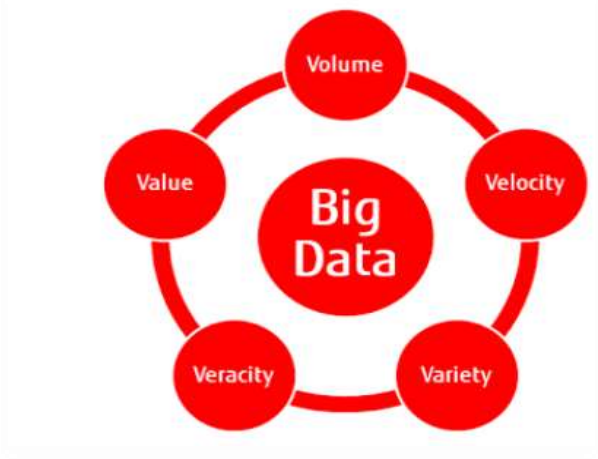

Fig. No. 2. Characteristics of Big Data

4. BIG DATA PROCESSING TOOLS

Large numbers of tools are available to process big data. In this section, we discuss some current techniques for analyzing big data with emphasis on three important emerging tools namely Map Reduce, Apache Spark, and Storm. Most of the available tools concentrate on batch processing, stream processing, and interactive analysis. Most batch processing tools are based on the Apache Hadoop infrastructure such as Mahout and Dryad. Stream data applications are mostly used for real time analytic. Some examples of large scale streaming platform are Strom and Splunk. The interactive analysis process allow users to directly interact in real time for their own analysis. For example Dremel and Apache Drill are the big data platforms that support interactive analysis. These tools help us in developing the big data projects. A fabulous list of big data tools and techniques is also discussed by much researchers [8][9].

Apache Hadoop and Map Reduce

The most established software platform for big data analysis is Apache Hadoop and Map reduce. It consists of hadoop kernel, map reduce, hadoop distributed file system (HDFS) and apache hive etc. Map reduce is a programming model for processing large datasets is based on divide and conquer method. The divide and conquer method is implemented in two steps such as Map step and Reduce Step. Hadoop works on two kinds of nodes such as master node and worker node. The master node divides the input into smaller sub problems and then distributes them to worker nodes in map step. Thereafter the master node combines the outputs for all the sub problems in reduce step. Moreover, Hadoop and MapReduce work as a powerful software framework for solving big data problems. It is also helpful in fault-tolerant storage and high throughput data processing.

Apache Mahout

Apache mahout aims to provide scalable and commercial machine learning techniques for large scale and intelligent data analysis applications. Core algorithms of mahout including clustering, classification, pattern mining, regression, dimensionalty reduction, evolutionary algorithms, and batch based collaborative filtering run on top of Hadoop platform through map reduce framework. The different companies those who have implemented scalable machine learning algorithms are Google, IBM, Amazon, Yahoo, Twitter, and facebook [10]

Apache Spark

Apache spark is an open source big data processing framework built for speed processing, and sophisticated analytics. It is easy to use and was originally developed in 2009 in UC Berkeleys AMPLab. It was open sourced in 2010 as an Apache project. Spark lets you quickly write applications in java, scala, or python. In addition to map reduce operations, it supports SQL queries, streaming data, machine learning, and graph data processing. Spark runs on top of existing hadoop distributed file system (HDFS) infrastructure to provide enhanced and additional functionality. Spark consists of components namely driver program, cluster manager and worker nodes. The driver program serves as the starting point of execution of an application on the spark cluster. The cluster manager allocates the resources and the worker nodes to do the data processing in the form of tasks.

Dryad

It is another popular programming model for implementing parallel and distributed programs for handling large context bases on dataflow graph. It consists of a cluster of computing nodes, and an user use the resources of a computer cluster to run their program in a distributed way. Indeed, a dryad user uses thousands of machines, each of them with multiple processors or cores. The major advantage is that users do not need to know anything about concurrent programming. A dryad application runs a computational directed graph that is composed of computational vertices and communication channels [11].

Storm

Storm is a distributed and fault tolerant real time computation system for processing large streaming data. It is specially designed for real time processing in contrasts with hadoop which is for batch processing. Additionally, it is also easy to set up and operate, scalable, fault-tolerant to provide competitive performances. The storm cluster is apparently similar to hadoop cluster. On storm cluster users run different topologies for different storm tasks whereas hadoop platform implements map reduce jobs for corresponding applications. There are number of differences between map reduce jobs and topologies. The basic difference is that map reduce job eventually finishes whereas a topology processes messages all the time, or until user terminate it. A storm cluster consists of two kinds of nodes such as master node and worker node. The master node and worker node implement two kinds of roles such as nimbus and supervisor respectively. The two roles have similar functions in accordance with jobtracker and tasktracker of map reduce framework. Nimbus is in charge of distributing code across the storm cluster, scheduling and assigning tasks to worker nodes, and monitoring the whole system. The supervisor complies tasks as assigned to them by nimbus

Apache Drill

Apache drill is another distributed system for interactive analysis of big data. It has more flexibility to support many types of query languages, data formats, and data sources. It is also specially designed to exploit nested data. Also it has an objective to scale up on 10,000 servers or more and reaches the capability to process patabytes of data and trillions of records in seconds. Drill use HDFS for storage and map reduce to perform batch analysis

Splunk

In recent years a lot of data are generated through machine from business industries. Splunk is a real-time and intelligent platform developed for exploiting machine generated big data. It combines the up-to-the-moment cloud technologies and big data. In turn it helps user to search, monitor, and analyze their machine generated data through web interface. The results are exhibited in an intuitive way such as graphs, reports, and alerts. Splunk is different from other stream processing tools. Its peculiarities include indexing structured, unstructured machine generated data, real-time searching, reporting analytical results, and dashboards.

5. BIG DATA APPLICATIONS

The applications of Big Data can be observed in various domains, as follows

Big data in retail

Competitions in the retail Industry is very fierce and retailers are continuously striving to achieve a competitive edge over others, and in order to thrive, it is important that retailers understand their customers really well. Having the awareness of the needs of the customers and how to optimally satisfy them will give the company a competitive edge. Also, by performing advanced analysis on their customer's data, retailers could fully understand their customers. The data of customers can be obtained via many resources including social media, loyalty programs and so forth. For retailers, all details of the customers are of value and having understanding of all of these minute details brings the retailers to their customers as close as possible. Consequently, the retailers could provide their customers with more personalized services and also forecast their future demands. Loyal customer can thus be established. Costco, Walmart, Walgreens, and Sears and Holdings are among the retailers that heavily utilize Big Data. Relevantly, the National Retail Federation estimated that about 30% of retail annual sales come from sales made in November and December [12]

Big data in healthcare

Big Data greatly facilitates the healthcare industry as this industry consistently has to deal with very large amount of data. Such amount of data has made it rather impossible for the healthcare practitioners to harness them. The use of Big Data can be regarded as lifesaving as it facilitates the practitioners and researchers in this industry to detect and cure diseases such as cancer. Also through Big Data and analytics, more personalized medications can be established, and more effective treatments can be provided to the patients. Furthermore, unique patterns of certain medicines can be identified, allowing the development of more cost-effective solutions [13].

Big data in education

Within the realm of education, data are generally important for future references. Hence, data are highly important in this domain. The use of Big Data greatly enhances the system of education, by specifically revitalizing the skills, both academic and non-academic ones. Also, the use of Big Data facilitates the evaluation of performances of students and teachers. Big Data has also been used in academic curriculum reformation in some leading universities. Equally, Big Data can be used in tracking the rate of dropout and then in determining the most appropriate measures to reduce it [14].

Big data in e-commerce

E-commerce has been regarded as a remarkable revolution in this era and it has become an integral part of life of people today. Hence, it is common for people to be thinking about E-commerce when they want to purchase something. In this regard, Amazon, Flipkart and Alibaba are among the most notable global E-commerce companies and the use of Big Data in these companies is extensive. Relevantly, Amazon as the world's biggest Ecommerce company is one of the leaders in Big Data and analytics. Meanwhile, Flipkart which is an Indian-based company, has one of the most vigorous data platforms in the country [15] . Within the domain of Big Data, the recommendation engine of Big Data is by far the most extraordinary applications of Big Data as it provides a 360- degree view of customers to the companies. This allows the companies to make appropriate recommendations to the customers, making the services more personalized. Indeed, the experiences of online shipping of people are completely redefined through Big Data.

Big data in media and entertainment

Media and Entertainment industry generally involves art and the use of Big Data is regarded as part of this art. Even though discrete from one another, the combination of art and science can generate remarkable outcomes especially in this industry. The general aim of this industry is to please customers and thus, it is crucial that this industry is able to consistently present new content to customers in order to retain them. In this regard, it is vital to have recommendation engine. Meanwhile, viewers today are inclined to choose the contents that they want, and generally, viewers prefer fairly new contents. Prior to the emergence of Big Data, companies would randomly broadcast their advertisement without performing any analysis first, and now with Big Data analytics, companies could determine the type of Ads to broadcast (i.e., those that would Attract customers) and the best broadcast time to achieve the maximum attention [15]

Big data in finance

Financial organizations greatly rely on data in their operations, and in fact, for such organizations, data are their second most vital commodity after money. Owing to such importance, financial organizations need to assure safety of their data which is a challenging task. Financial firms were in fact among the first adopters of Big Data and Analytics, and prior to that, these firms were already mastering the technical field. Relevantly, Digital banking and payments have been among the most trending buzzwords with Big Data as their important element. In financial firms, Big Data handles the major domains including algorithmic trading, fraud detection, risk analysis, and customer contentment. With Big Data, the financial system becomes fluent,improved, and empowered in making available superior services to the customers [16]

CONCLUSION

In this paper, it discussed the concept of Big Data Analytics in detail. Big Data has been discussed in detail along with its characteristics and architecture. The applications and importance of Big Data Analytics have been described in detail. It focused on the current scenario of Big Data Analytic Tools also. To conclude it can say that Big Data Analytics have both positive and negative impacts. Big Data Analytics is the latest emerging research area in the field of Big Data Processing. Big Data Analytics is beneficial for various fields of applications like in Education, Healthcare, Agriculture, Military, Banking, especially for Business Organizations . In Future, one can work on finding the solution for the various challenges that are creating problem on the path of Big Data Analytics. Further the work can be done on designing new techniques and algorithms for the processing of Big data. The collection of data from variety of sources can be improved by developing different new techniques. More focus should be given to the data security issues which have not considered in this work but it is very crucial issue that needs to be discussed. The techniques should be developed for the analysis of real time data instead of examining the static past data.

REFERENCES

1. Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F. B., & Babu, S. (2011, January). Starfish: A self-tuning system for big data analytics. In *Cidr* (Vol. 11, No. 2011, pp. 261-272).
2. Singh, D., & Reddy, C. K. (2015). A survey on platforms for big data analytics. *Journal of big data*, *2*, 1-20.
3. Peglar, R., & Isilon, E. M. C. (2012). Introduction to Analytics and Big Data-Hadoop. *Storage Networking Industry Association*.
4. Siddiqui, S., & Gupta, D. (2014). Big data process analytics: a survey. *Int J Emerg Res Manag Technol*, *3*(7), 117-23.
5. http://csrc.nist.gov/groups/SMA/Forum/document/June2012Presentation/f%25CSM_june2012_cooper_Neul.pdf
6. Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE access*, *2*, 652-687.
7. www.ebizmba .com/articles/social-networking-websites.
8. Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information sciences*, *275*, 314-347.
9. Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. *Journal of Big data*, *1*, 1-35.
10. Ingersoll, G. (2009). Introducing apache mahout. *IBM developerWorks Technical Library*.
11. Li, H., Fox, G., & Qiu, J. (2012, November). Performance model for parallel matrix multiplication with dryad: Dataflow graph runtime. In *2012 Second International Conference on Cloud and Green Computing* (pp. 675-683). IEEE.
12. Santoro, G., Fiano, F., Bertoldi, B., & Ciampi, F. (2018). Big data for business management in the retail industry. *Management Decision*, *57*(8), 1980-1992.
13. Baro, E., Degoul, S., Beuscart, R., & Chazard, E. (2015). Toward a literature-driven definition of big data in healthcare. *BioMed research international*, *2015*.
14. Williamson, B. (2017). Big data in education: The digital future of learning, policy and practice.
15. Akter, S., & Wamba, S. F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. *Electronic Markets*, *26*, 173-194.
16. Lippell, H. (2016). Big data in the media and entertainment sectors. *New Horizons for a data-driven economy: a Roadmap for usage and Exploitation of big data in Europe*, 245-259.
17. Begenau, J., Farboodi, M., & Veldkamp, L. (2018). Big data in finance and the growth of large firms. *Journal of Monetary Economics*, *97*, 71-87.