IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 2, February 2020

Big Data - A Study on Challenges, Technologies and Tools

Mr. S. Antony Joseph Raj[1], Dr. P. Joseph Charles[2]

Assistant Professor, Departmentof Information Technology, St.Joseph's College, Trichy, India[1,2]

Abstract: With the growing IT infrastructures and technologies such as Internet of Things and Cloud Computing information generated is enormous in amount. As the data is generated in every nanosecond, probably faster than that the data exceeds the boundary often. In order to collect such an enormous amount of data and process it there are a number of techniques tied up under the term big data. As, such data is evolving form every industry namely medical, food, business, education, entertainment and etc., it is put into a severe treatment for the benefit of profits in various means called as analytics. Lot organizations adapt to analytics for processing and utilizing the customary information for various applications. This paper aims at providing a study on challenges in the big data analytics for the extensive usage of various techniques. The work also focuses on technologies involved in analytics based on the need and requirements. The result and analysis presented would help the researchers in choosing the technologies and also to identify the challenges in big data.

Keywords: Big data, Hadoop, Map Reduce, Apache Spark, Storm, Python, R, Tableau

I. INTRODUCTION

Before the evolution in internet and online applications, the system had to work with few structured data set of data. And slowly the giant called Information Technology gave its entry to every nook and corner of the world, which is the end of structured data analytics. As business firms entered the realm of big data, size of data went booming for terabytes into petabytes in a single set [1]. But for the increase rate of growth of unstructured data (emails, videos, audio files and etc) buzzword "Big Data" has been evolved.

With the emergence of IT as IoT and CC otherwise called as Internet of Things and Cloud Computing, the data created is mostly unstructured. This comes under the estimate of more than 95% of data generated in terabytes or even peta bytes, derived from social networks, sensor networks and other federated data with replications. Big data is the answer for handling such dissimilar data. It was predicted that by IDC that by 2020 the size of storage will be reaching 40 ZB as the information of the world doubles once in two years. The prediction was that by the year 2020 world would generate 50 times the larger amount of information. [2]

Big data is the term used for coining larger amount of data which are in various forms [3]. It can be characterized by three aspects: as volume (numerous data), Variety (cannot be categorized), and Velocity (generated quickly). It is also a term used for a large and complex data set, difficult for storing and processing using traditional DBM tools or processing applications. Here comes the necessity for learning the technologies that are existing, tools available and the challenges that persists in the real world.

II. BIG DATA: DIMENSIONS

'Big data' as a growing trend of computer science and has historic back ground. The drastic growth of data had made the traditional data analytics tools scarce and so the modern Business analytics tools and technologies have been evolved to extract the knowledge from the huge amount of perishable data.

The following are the major dimensions and characteristics of Big Data [4]. These V's are the major challenges in the Big Data.

1) **Volume**: Data is increasing in enormous rate. The size of Big Data is accumulated in larges terabyte to petabytes due to storage of data, live streaming, ICT's, product codes etc.
2) **Velocity**: Velocity indicates the speed at which the data generated and in turn responded. Big data is able to handle the incoming and outgoing data rapidly. Unprecedented rate of data has been created from sensors and smart phones, which are needed for real time analytics and evidence-based planning.
3) **Variety:** These data do not have a fixed structure. They are generated from various different sources and in different formats like audio, video, documents, logs etc. That data set can be structured or unstructured, public or private, shared or confidential, incomplete or complete etc.

4) **Veracity:** This represents the unreliability inherent in some sources of data. The customer sentiments in the social media are uncertain in nature even though the data contain the valuable information. Identifying and verifying inconsistent information is significant, to accomplish faithful study. Thus creating faith in big data is a big challenge to manage even more variety of data is available.
5) **Variability**: It can be added to the above 'V' which highlight on semantics, or the variability of meaning in language and communication protocols.
6) **Value:** This is the main concept behind the new technology called "Big Data", Oracle introduces Value and defines that big data are often characterized by relatively "low value density". That is, the data received in the original form usually has a low value relative to its volume. By analysing large Volume of data High Value can be obtained and in turn **Value Based Services** can be provided to the customers.

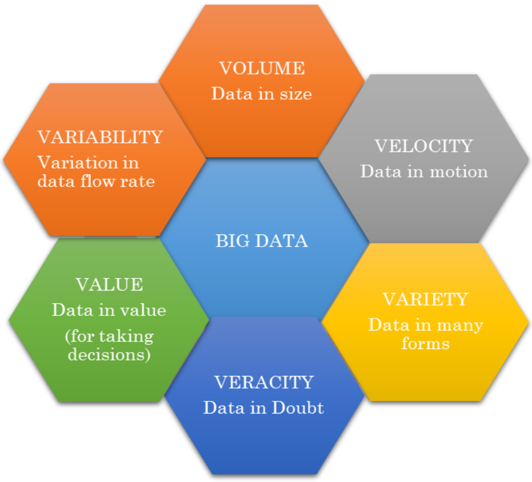

Fig 1: Big Data – Dimension's

Data Sources

Everything has become digitalized and technology has also enlarged the size of data. The data generated starting from astronomy to IoT, now becomes the source. Data can be retrieved from agriculture industry, medical care, etc of the smart cities. To analyse the different dissimilar data large numbers of tools are available. In this section, we discuss some current tools and techniques for analysing big data with emphasis on three important emerging tools namely Map Reduce, Apache Spark, and Storm.

III. ANALYSIS ON THE BIG DATA ASPECTS

There are a number of technologies present in analysis of huge volume of data. Also, a rich set of tools are existing for visualisation of results. Big data holding every industry in its hands is also having certain challenges yet to be studied. Here the work presents the major challenges, a comparison on existing technologies and a comparison on the existing visualization tools.

3.1 Major Challenges

On evolution of new technologies, threesome new challenges in all the aspects will also be evolved. Once it was functional challenges are in place, the next kin are the technical challenges. Here below we have discussed some big data technical challenges which are on the street of the research. [5, 6]

3.1.1. Data size

In current scenario, new technologies have been presented to reduce human burden and workload which enable us to store, query and analyse large data sets. Now it is difficult to use the complete data set because of volume (size) that enables us to new techniques.

3.1.2. Fault Tolerance

Developing 100% reliable systems on the go is not an easy task. Systems can be devised in such a way that the probability of failure must fall within the permitted threshold. Fault tolerance is a technical challenge in big data. When a process started it may involve with numerous network nodes & the whole computation process becomes cumbersome. Retaining check points and fixing the threshold level for process restart in case of failure, are greater concerns.

DOI 10.17148/IJARCCE.2020.9217

3.1.3. Data Heterogeneity

Big data deals with unstructured, semi-structured and structured data. Linking unstructured data with structured data, converting data from one form into another required form needs a lot of research.

3.1.4. Awareness in using the analytics tool

There is lack of understanding on how to use analytics to reduce the size, to improve the business value. This occurs because the objects which have to be modelled are huge, complex, and distributed. To overcome new modelling and simulation software are needed which should be simple, robust, distributed and parallel computing.

3.1.5. Content validation

As we get data from different resources, the data has to be validated. We have different types of data (text, audio, video, logs etc...) From different types of sources like are blogs, social sites etc.)And different types of content such as tweets, comments, articles, etc…. this vast amount of data has to be validated. It is very difficult and this becomes the major challenge in Big Data.

3.2. Technologies

The need of technologies in the analysis of the data generated by kith and kin, passers-by and native of the home are greater in ratio. As the data collected is simply stored or is on the flow, receiving it and working through its parameters, can also be called as analytics. Placing a boundary for the data collected, coming up with new decision with the present data are the few occasions where these technologies play their role.

3.2.1. Hadoop and Map Reduce

Apache Hadoop and Map reduce are the most established software platform for big data analysis. 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 Map Reduce work as a powerful software framework for solving big data problems. It is also helpful in fault-tolerant storage and high throughput data processing[4].

3.2.2. Apache Spark

Apache spark is an open source big data processing framework built for speed processing, and sophisticated analytics. The prime focus of spark includes resilient distributed datasets (RDD), which store data in-memory and provide fault tolerance without replication. It supports iterative computation, improves speed and resource utilization[4].

3.2.3. 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 [4].

3.3. Visualization Tools

As the analytics is done, it doesn’t enough to stop there, but still require something for putting it into a form as presentable output. Here comes the use of visualization tools. There are a number of visualization tools used in analytics, among which a few are reviewed here:

Table 1. Visualization Tools – Big Data Analytics

Tool	Type of Data	Interface	Advantages
Python	Heterogeneous	Command line interface	Packages and API’s
R	Heterogeneous		Provide many packages
GraphViz	Network		Powerful
Gephi	Network	Graphical User interface	Powerful, Easy to use
Cytoscape	Network		Powerful
Circos	Circular		Visualize chromosomes
Tableau	Tables and Maps		Google map style

Presentation of results in graphical form is essential to interpret the data, for which tools used to visualize the Data are shown in Table 1. As a language, R provides six thousand packages that put before the users a number of functions [7]. The presence of such packages confirms the need of R, which helps plot results using ggplot2 and other plot functions. For visualization of network specific data, Cytoscape[8] can provide rich outputs, Circos [9] is becoming a standard for the visualization of genomic chromosomes, Gephi [10] is an application helps creating hierarchical graphs and GraphViz is a command-line tool that can print network with large amount of nodes. Among all Python could stand first with rich set of its packages and api support. Also Tableau is also another choice as it is commercial as it can be employed in visualizing location data as Google map.

IV. CONCLUSION

As the importance of analytics is ever green in every industry, it will be necessary to study on the challenges and techniques existing. This paper therefore reflects such existing challenges and technologies of big data. The focusing is challenges it present various existing challenges such as volume of data, heterogeneity, validation, fault tolerance and awareness on tools. While presenting the technologies the major technologies noted are Hadoop, Spark and Storm. As form the analysis it was found that R could be efficient next to python with rich set of packages. And one among the commercial tools suggested would be Tableau. The comparative analysis presented could help researchers choose the best tools for the visualization of results.

REFERENCES

[1]. JunPing Wang, WenSheng Zhang, YouKang Shi, ShiHuiDuan, Jin Liu "Industrial Big Data Analytics: Challenges,Methodologies, and Applications", IEEE, 2018, pp 1 – 12.

[2]. NawsherKhan,IbrarYaqoob,IbrahimAbakerTargioHashem,ZakiraInayat,WaleedKamaleldinMahmoudAli,MuhammadAlam,MuhammadShiraz, andAbdullahGani1"Big Data: Survey, Technologies, Opportunities, and Challenges"The Scientific World Journal, Volume 2014, Article ID 712826, 18 pages http://dx.doi.org/10.1155/2014/712826

[3]. Jui-Chien Hsieh , Ai-Hsien Li and Chung-Chi Yang "Cloud, and Big Data Computing",Int.J.Environ.Res.Public Health 2013,10,6131-6153

[4]. P. Acharjya,Kauser Ahmed P," A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools" , (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016.

[5]. Dr.R.Saravanakumar and Dr.C.Nandini,"A Survey on the Concepts and Challenges of Big Data: Beyond the Hype",Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 5 (2017) pp. 875-884 © Research India Publications http://www.ripublication.com.

[6]. S. Justin Samuel, Koundinya RVP, KothaSashidhar and C.R. Bharathi, "A Survey on Big Data and its Research Challenges", ARPN Journal of Engineering and Applied Sciences, VOL. 10, NO. 8, ISSN 1819-6608.

[7]. Ihakaa, Gentleman, "R: a language for data analysis and graphics", J. Comput. Graph. Stat. 5 (1996) 299–314.

[8]. Shannon, Markiel, Ozier, Baliga, Wang, Ramage, Amin, Schwikowski, Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks", Genome Res. 13 (2003), 2498–2504

[9]. Krzywinski, Schein, Birol, Connors, Gascoyne, Horsman, Jones, Marra, "Circos: an information aesthetic for comparative genomics", Genome Res. 19 (2009) 1639–1645.

[10]. Bastian, Heymann, Jacomy, "Gephi: an open source software for exploring and manipulating networks", in: International AAAI Conference on Weblogs and Social Media, 2009.

DOI 10.17148/IJARCCE.2020.9217