iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VIII **Month of publication:** August 2017

DOI: http://doi.org/10.22214/ijraset.2017.8118

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VIII, August 2017- Available at www.ijraset.com

Challenges Involved in Big Data Processing & Methods to Solve Big Data Processing Problems

Diksha Sharma[1], Gagan Pabby[2], Neeraj Kumar[3]
[1,2]*B.Tech* , [3]*Assistant Professor, Electronics and Communication Engineering Department, Chitkara University*

Abstract: ***The term "Big Data" refer to the gigantic bulkiness of data which cannot be dealt with by conventional data-handling techniques. Big Data is a new conception, and in this article we are going to intricate it in a clear fashion. It commences with the conception of the subject in itself along with its properties and the two general approaches of dealing with it. The widespread study further goes on to explain the applications of Big Data in all various aspects of economy and being. The deployment of Big Data Analytics after integrating it with digital capabilities to secure business growth and its apparition to make it intelligible to the technically apprenticed business analyzers has been discussed deeply. Also the challenge that hinders the growth of Big Data Analytics is explained in the paper. A brief description about "Hadoop" & Machine learning is also given in the article.***
Keywords: Big data, Hadoop, Machine learning

I. INTRODUCTION

Big data refers to data sets or combinations of data sets whose size complication and rate of expansion make them hard to be processed & analyzed by traditional technologies such as relational databases and desktop information within the time necessary to make them useful. While the size used to decide whether a particular data set is considered big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from terabytes to multiple petabytes(1). Big data challenges include capture, storage, analysis, data curation, search, sharing, transfer, visualization, querying, and updating and information privacy. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. "There is slight uncertainty that the volume of data now available is certainly large, but that's not the most appropriate attribute of this novel data ecosystem. Analysis of data sets can find new correlations to "spot business trends, prevent diseases, and combat crime and so on. Business executives, medical practitioners, repeatedly face difficulties with huge data-sets in Internet search, urban informatics, and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics, complex physics simulations, biology and environmental research. Data sets raise rapidly because they are gradually gathered by cheap and numerous information-sensing IOT devices such as mobile devices, aerial (remote sensing), software logs, cameras, microphones, RFID readers and wireless sensor networks [1, 2]. Big data can be explained by 3V's namely volume, variety & velocity [3, 4].

II. DATA CLASSIFICATION

Data can be classified as either primary & secondary & Qualitative &Quantitative data. Primary data means original data that has been collected specially for the purpose in mind. It means someone collected the data from the original source first hand. Data collected this way is called primary data. The community who collect primary data can be an approved society, examiner, or they might be just somebody with a clipboard. Those who gather primary data may have knowledge of the study and may be motivated to make the study a success.Secondary data is data that has been unruffled for another purpose. It means that solitary purpose's Primary Data is other purpose's Secondary Data. Secondary data is data that is being reused. Qualitative data is a inflexible measurement uttered not in terms of statistics, but relatively by means of a natural language explanation. In figures, it is repeatedly used interchangeably with "definite" data. Although there might have categories, the categories might have a structure to them. When there is not a natural ordering of the categories, it is known as nominal categories. When the categories might be prearranged, these are called ordinal variables. Categorical variables that judge size (small, medium, large, etc.) are ordinal variables. Note that the distance between these categories is not something we can measure. Quantitative data is a arithmetic quantity articulated not by means of a natural language explanation, but relatively in terms of numbers. However, not all numbers are continuous & measurable

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VIII, August 2017- Available at www.ijraset.com

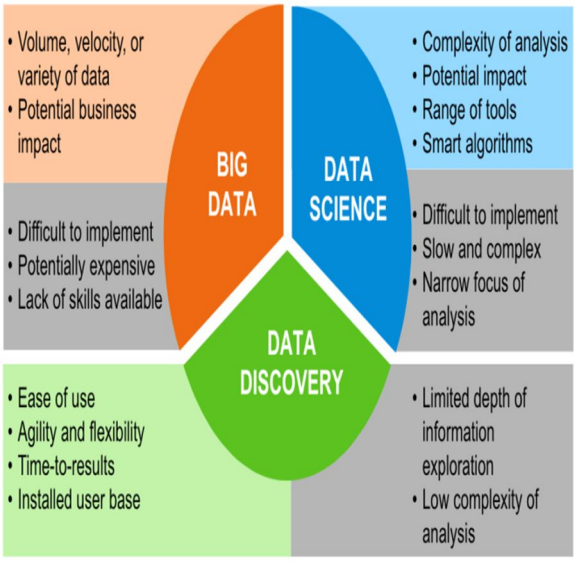


Fig 2. Big Data

Quantitative data always are associated with a scale measure. Probably the most common scale type is the ratio-scale. Observations of this type are on a scale that has a meaningful zero value but also have an equidistant measure (i.e., the difference between 10 and 20 is the same as the difference between 100 and 110). For example, a 10 year-old girl is twice as old as a 5 year-old girl. Since you can measure zero years, time is a ratio-scale variable. Money is another common ratio-scale quantitative measure. Observations that you count are usually ratio-scale (e.g., number of widgets).

III. PROBLEMS IN BIG DATA PROCESSING

A. Heterogeneity and Incompleteness

When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured

B. Scale of Course

The first thing anyone thinks of with Big Data is its size. After all, the word “big” is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore’s law, to provide us with the resources needed to cope with increasing volumes of data. But, there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static[5].

C. Timeliness

The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data. Rather, there is an acquisition rate challenge

D. Privacy

The privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done. For other data, regulations, particularly in the US, are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VIII, August 2017- Available at www.ijraset.com

IV. HOW TO SOLVE PROBLEM OF BIG DATA PROCESSING USING HADOOP

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's MapReduce that is a software framework where an application break down into various parts. The Current Appache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper. HDFS and MapReduce are explained in following points.

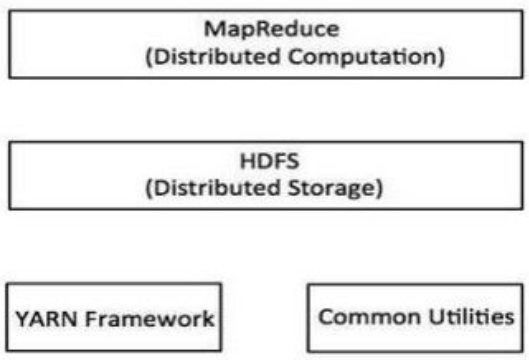


Fig 2. HDFS block diagram

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job. The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. It is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications having large datasets. Apart from the above-mentioned two core components, Hadoop framework also includes the following two modules: Hadoop Common: These are Java libraries and utilities required by other Hadoop modules. Hadoop YARN: This is a framework for job scheduling and cluster resource management. How Does Hadoop Work? It is quite expensive to build bigger servers with heavy configurations that handle large scale processing, but as an alternative, you can tie together many commodity computers with single-CPU, as a single functional distributed system and practically, the clustered machines can read the dataset in parallel and provide a much higher throughput. Moreover, it is cheaper than one high-end server. So this is the first motivational factor behind using Hadoop that it runs across clustered and low-cost machines. Hadoop runs code across a cluster of computers. This process includes the following core tasks that Hadoop performs: Data is initially divided into directories and files. Files are divided into uniform sized blocks of 128M and 64M (preferably 128M). These files are then distributed across various cluster nodes for further processing. HDFS, being on top of the local file system, supervises the processing Blocks are replicated for handling hardware failure.

V. CONCLUSION

The article illustrates the conception of Big Data alongside with 3 Vs, Volume, Velocity and variety of Big Data. The article also highlights problems of Big Data processing .These technical challenges must be addressed for efficient and fast processing of Big Data. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These

technical challenges are common across a large variety of application domains, and therefore not costeffective to address in the context of one domain alone. The paper describes Hadoop which is an open source software used for processing of Big Data.

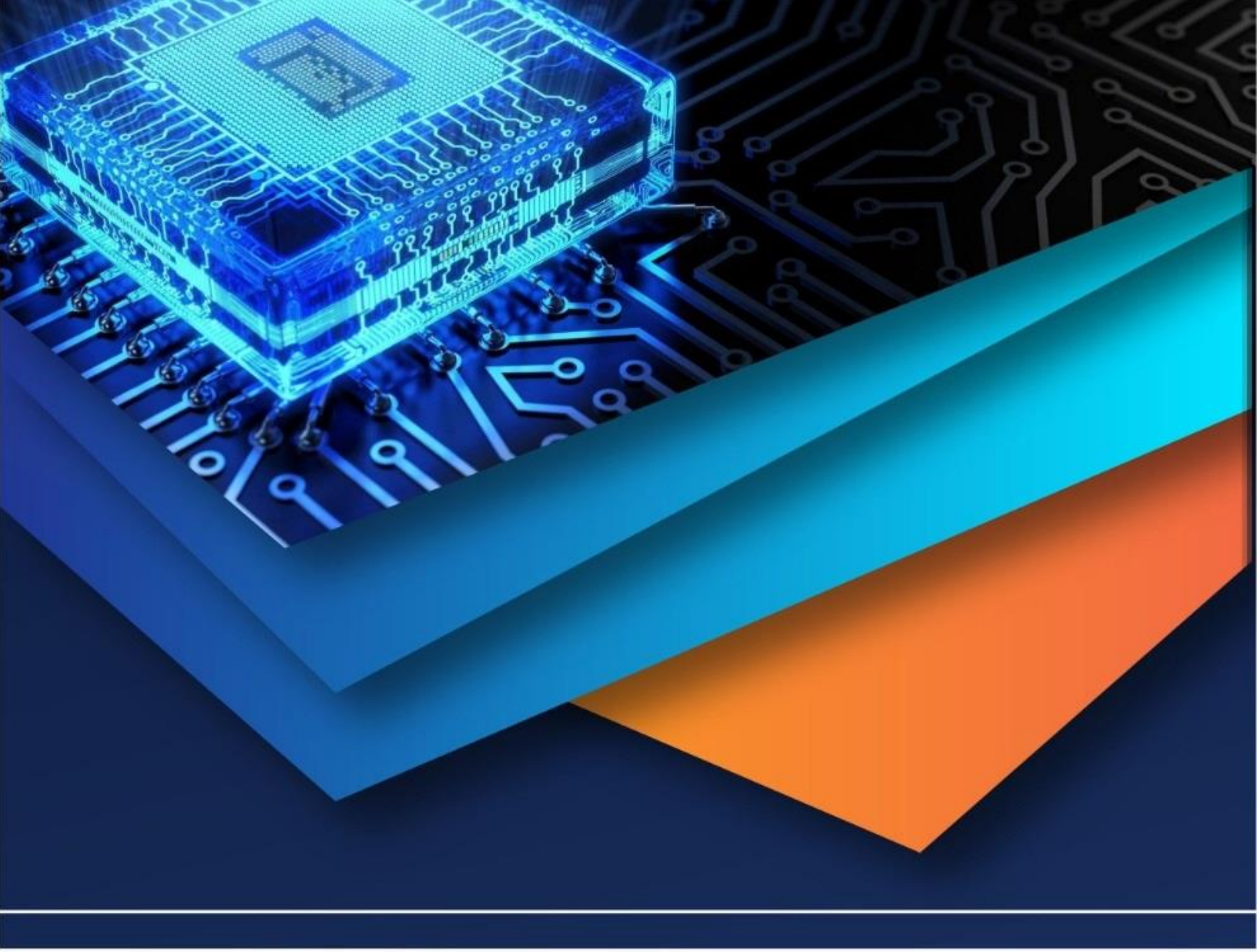
REFERENCES

[1] Jump up Hellerstein, Joe (9 November 2008). "Parallel Programming in the Age of BigData".Gigaom-Blog.
[2] Jump up Segaran, Toby; Hammerbacher, Jeff (2009). Beautiful Data: The Stories Behind Elegant Data Solutions. O'Reilly Media. p. 257. ISBN 978-0-596-15711-
[3] Beyer, Mark. "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data". Gartner. Archived from the original on 10 July 2011. Retrieved 13 July 2011.
[4] Chen, C.P. and Zhang, C.Y., 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 275, pp.314-347
[5] Bhosale, H.S. and Gadekar, D.P., 2014. A Review Paper on Big Data and Hadoop. International Journal of Scientific and Research Publications, 4(10), p.1.

Ms. Diksha completed her B.Tech from Chitkara University, Himachal Pradesh in the stream of Electronics and Communication Engineering. She is now planning to pursue Masters in science from abroad.

Mr. Gagan Pabby completed his B.Tech from Chitkara University, Himachal Pradesh in the stream of Electronics and Communication Engineering. He is now planning to pursue Masters in science from abroad.

Mr. Neeraj Kumar is presently working as Assistant Professor in Electronics and Communication Engineering Department at Chitkara University, Himachal Pradesh, India. He has more than 6 years of teaching experience. His area of interest is digital image processing, digital signal processing.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY