Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 7, Issue. 12, December 2018, pg.1 – 5

A Review toward Powers of Big Data and its Challenges, Tools and Techniques

Priyanka Tyagi
Priya21.tyagi@gmail.com
Lalita Devi Institute of Management and Science, Mandi Road New Delhi, India

Dr. Pranav Mishra
mishrpranav@gmail.com
Lalita Devi Institute of Management and Science, Mandi Road New Delhi, India

Abstract— Big Data applies to data that can't be prepared or broke down utilizing customary procedures or devices. For decades, organizations have been settling on business choices in light of value-based information put away into relational databases. In any case, there is a potential fortune of non-conventional information got from different sources, for example, online networking, messages, emailing, online overviews, online shopping and so forth that can be dug for helpful information. With the diminishing in the expense of storage capacity and calculation, it has ended up workable for ventures to utilize this information to profit. However, the fast growth rate of such large data generates numerous challenges, such as data inconsistency and incompleteness, scalability, timeliness, and security. This paper provides a brief introduction to the Big data technology and its importance in the contemporary world and addresses various challenges and issues that need to be emphasized to present the full influence of big data.
Keywords: Big Data, characteristics, definition, challenges, Hadoop, MapReduce

I. INTRODUCTION

Big Data has gained much attention from the last few years in the IT industry. As we can witness billions of people are connected to internet worldwide, generating large amount of data at the rapid rate. The generation of this large amount of engenders various challenges. Along with Big Data's huge benefits to many organizations, the challenges and issues should also be brought into light. A forecast from International Data Corporation (IDC), the Big Data technology and services market represents a fast-growing multibillion-dollar worldwide opportunity. In fact, a recent IDC forecast shows that the Big Data technology and services market will grow at a 26.4% compound annual growth rate to $41.5 billion through 2018, or about six times the growth rate of the overall information technology market. Additionally, by 2020 IDC believes that line of business buyers will help drive analytics

beyond its historical sweet spot of relational (performance management) to the double-digit growth rates of real-time intelligence and exploration/discovery of the unstructured worlds.

II. DEFINITION OF BIG DATA

At present, the industry does not have a bound together meaning of Big Data. It has been characterized in varying courses as takes after by different gatherings: As per McKinsey, "Enormous Data alludes to datasets whose size are past the capacity of regular database programming apparatuses to catch, store, oversee and break down". IDC characterizes Big Data advancements as another era of advances and models intended to concentrate esteem financially from substantial volumes of a wide assortment of information by empowering high speed catch, revelation and investigation. As indicated by O'Reilly, "Enormous Data will be information that surpasses the handling limit of ordinary database frameworks. The information is too huge, moves too quickly, or does not fit the structures of existing database designs. To pick up quality from this information, there must be an option approach to process it." As indicated by Wikipedia, "Huge Data for the most part incorporates datasets with sizes past the capacity of generally utilized programming apparatuses to catch, clergyman, oversee, and handle the information inside a decent slipped by time".

As indicated by Gartner, "Huge Data is high volume, high speed, and/or high assortment data resources that require new types of handling to empower improved basic leadership, knowledge revelation, and procedure enhancement". In the nutshell, efficacy of Big Data is that it is used to describe massive volumes of unstructured and structured data that are so large that it is very difficult to process this data using traditional databases and software technologies.

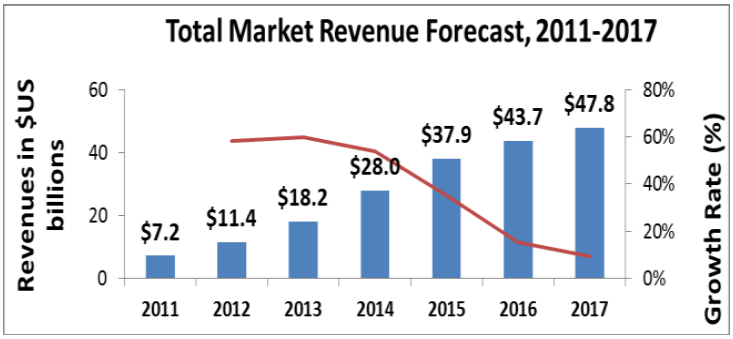

Figure 1: Growth rate of Big Data from 2011-2017

III. CHARACTERISTICS AND BENEFITS OF BIG DATA

The McKinsey Global Institute estimates that data volume is growing 40-50% per year, and will grow 44x between 2009 and 2020 [4]. However, volume of data is not the only characteristic that matters. In fact, Big Data has four main characteristics: Volume, Velocity, Variety, and Value commonly referred to as "4V," referencing the huge amount of data volume, fast processing speed, various data types, and low-value density.

Enormous Data coordinates both organized and unstructured information. The examination of information should be possible continuously or near ongoing, following up on full datasets instead of condensed components. The fundamental expense of the foundation to control the examination of information has fallen drastically, making it monetary to mine the data. Like customary investigation, it can likewise bolster interior business choices. The advancements and ideas driving Big Data permit associations to accomplish an assortment of goals. At the point when Big Data is refined and dissected.

In Big Data, the product bundles give a rich arrangement of instruments and choices where an individual could delineate whole information scene over the organization, in this manner permitting the person to investigate the dangers he/she confronts inside. This is considered as one of the primary favorable circumstances as Big Data keeps the information safe. With this an individual can have the capacity to identify the possibly touchy data that is not secured in a suitable way and ensures it is put away as indicated by the administrative necessities. A percentage of the ranges where Big Data is entirely valuable are expressed underneath.

IV. PROPERTIES OF BIG DATA

Sr. No.	Properties	Description
1.	Volume	Many factors contribute towards increasing Volume streaming data , live streaming data and data collected from sensors etc.,
2.	Variety	Data comes in all types of formats-from traditional databases ,text documents, emails, video, audio, transactions etc.,
3.	Velocity	This means how fast the data is being produced and how fast the data needs to be processed to meet the demand.
4.	Variability	Along with the Velocity, the data flows can be highly inconsistent with periodic peaks.
5.	Complexity	Complexity of the data also needs to be considered when the data is coming from multiple sources. The data must be linked, matched, cleansed and transformed into required formats before actual processing.

Challenge	Impact	Risk
Uncertainty of the market landscape	Difficulty in choosing technology components Vendor lock-in	Committing to failing product or failing vendor
Big data talent gap	Steep learning curve Extended time for design, development, and implementation	Delayed time to value
Big data loading	Increased cycle time for analytical platform data population	Inability to actualize the program due to unmanageable data latencies
Synchronization	Data that is inconsistent or out of date	Flawed decisions based on flawed data
Big data accessibility	Increased complexity in syndicating data to end-user discovery tools	Inability to appropriately satisfy the growing community of data consumers

Figure 2. Challenges, its impact and risk involved in Big data

V. BIG DATA CHALLENGES

Big data due to its various properties like volume, velocity, variety, variability, value and complexity put forward many challenges. The Figure 3 shows various challenges in big data. Figure 2 list some of the challenges in Big data along with its impact and risks involved.

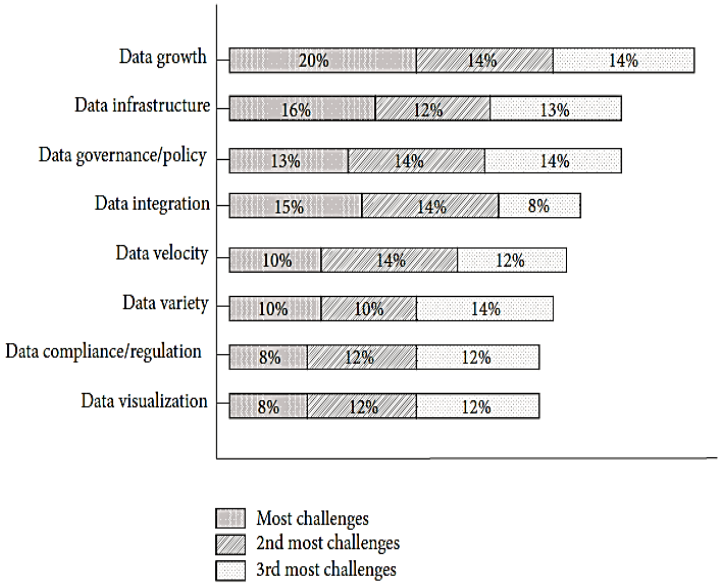

Figure 3. Challenges in Big data

VI. TECHNIQUES FOR BIG DATA HANDLING

There are many techniques available for data management. That includes Google BigTable, Simple DB, Not Only SQL (NoSQL), Data Stream Management System (DSMS), MemcacheDB, and Voldemort [7]. But these traditional approaches are only applicable to traditional data and not Big data as it cannot be stored on a single machine. The Big Data handling techniques and tools include Hadoop, MapReduce, and Big Table. Out of these, Hadoop is one of the most widely used technologies.

Hadoop

Hadoop is an Apache open source framework which is written in java. High volumes of data, in any structure, are processed by Hadoop. Hadoop allows distributed storage and distributed processing for very large data sets. The main components of Hadoop are:

1. Hadoop distributed file system (HDFS)
2. MapReduce

HDFS (Storage layer):- Hadoop has a distributed File System called HDFS, which stands for Hadoop Distributed File System. It is a File System used for storing very large files with streaming data access patterns, running on clusters on commodity hardware. There are two types of nodes in HDFS cluster, namely namenode and datanodes. The name node manages the file system namespace, maintains the file system tree and the metadata for all the files and directories in the tree. The datanode stores and retrieve blocks as per the instructions of clients or the namenode. The data retrieved is reported back to the namenode with lists of blocks that they are storing. Without the namenode it is not possible to access the file. So it becomes very important to make name node resilient to failure.

MapReduce (Processing/Computation layer):- It is a programing paradigm which is meant for managing applications on multiple distributed servers. In MapReduce divide and conquer method is used to break the large complex data into small units and process them. It reads the data from HDFS in an optimal way. However, it can read the data from other places too; including mounted local file systems, the web, and databases. It divides the computations between different computers (servers, or nodes). It is also fault-tolerant. If some of nodes fail, Hadoop knows how to continue with the computation, by re-assigning the incomplete

work to another node and cleaning up after the node that could not complete its task. It also knows how to combine the results of the computation in one place. The other core components in Hadoop architecture includes Hadoop YARN, it is a framework for job scheduling and cluster resource management. The other component is the cluster which is the set of host machines (nodes).

VII. CONCLUSION

There is most likely Big Data is the hot outskirts of today's data innovation advancement. The measure of information at present produced by the different exercises of the general public has never been so enormous, and is being created at a continually expanding speed. Through better investigation of the substantial volumes of data that are getting to be accessible, there is the potential for making speedier advances in a several disciplines and enhancing the gainfulness and accomplishment of numerous enterprises. Finally, so as to completely profit from Big Data, the above expressed difficulties should be taken care. As there are huge volumes of data that are produced every day, so such large size of data it becomes very challenging to achieve effective processing using the existing traditional techniques Big data is data that exceeds the processing capacity of conventional database systems. In this paper fundamental concepts about Big Data are presented. These concepts include Big Data characteristics, challenges and techniques for handling big data.

REFERENCES

[1]. "The Emerging Big Returns on Big Data", A TCS 2013 Global Trend Study.
[2]. Elena Geanina Ularu, Florina Camelia Puican, Anca Apostu, Manole Velicanu, "Perspectives on Big Data and Big Data Analytics", Database Systems Journal, Volume 3, No. 4, 2012.
[3]. Bernice M Purcell, "Big Data Using Cloud Computing", OC13030
[4]. "Big Data for the Enterprise", An Oracle White Paper, June 2013.
[5]. Chris Eaton, Dirk Deroos, Tom Deutsch, George Lapis, Paul Zikopoulos, "Understanding Big Data".
[6]. "Challenges and Opportunities with Big Data", A Community White Paper Developed by Leading Researchers Across United States, 2012.
[7]. https://www.idc.com/prodserv/4Pillars/bigdata
[8]. www.Wikibon.org
[9]. A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices." Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.]
[10]. Golfarelli, M., & Rizzi, S. (2009). Data warehouse design: modern principles and methodologies. Columbus: McGraw-Hill
[11]. Almeida, F., and Calistru, C, "The Main Challenges and Issues of Big Data Management", International Journal of Research Studies in Computing, 2(1), 2013, pp. 11-20.
[12]. https://www.progress.com

Author's Bibliography:

Priyanka Tyagi was born in sonipat, in 1984. She is currently working as an Assistant Professor in Lalita Devi Institute of management & science, Mandi. She received the M.Tech degree in computer science from Banasthali University, Rajasthan in 2009 and B.Tech degree in Computer Science from Hindu College of Engineering, Sonipat in 2006 and. Her research experience includes 4 years in the Academics.

