© 2016 IJSRSET | Volume 2 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Big Data : A Review of Challenges, Tools and Techniques

Anureet Kaur

Department of Computer Science and Applications, Khalsa College, Amritsar, Punjab, India

ABSTRACT

Big Data is the large amount of data that cannot be processed by making use of traditional methods of data processing. Due to widespread usage of many computing devices such as smartphones, laptops, wearable computing devices; the data processing over the internet has exceeded more than the modern computers can handle. Due to this high growth rate, the term Big Data is envisaged. 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. This paper addresses various challenges and issues that need to be emphasized to present the full influence of big data. The tools used in Big data technology are also discussed in detail. This paper also discusses the characteristics of Big data and the platform used in Big Data i.e. Hadoop.

Keywords: Big Data, 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. [1].

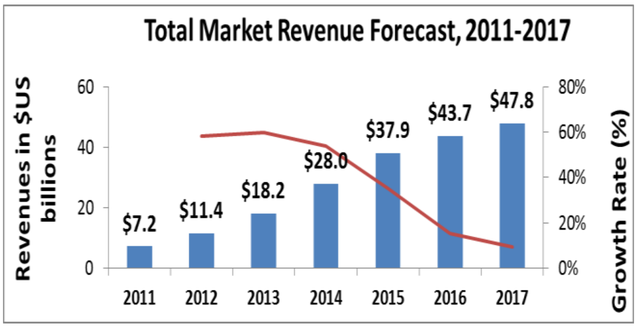
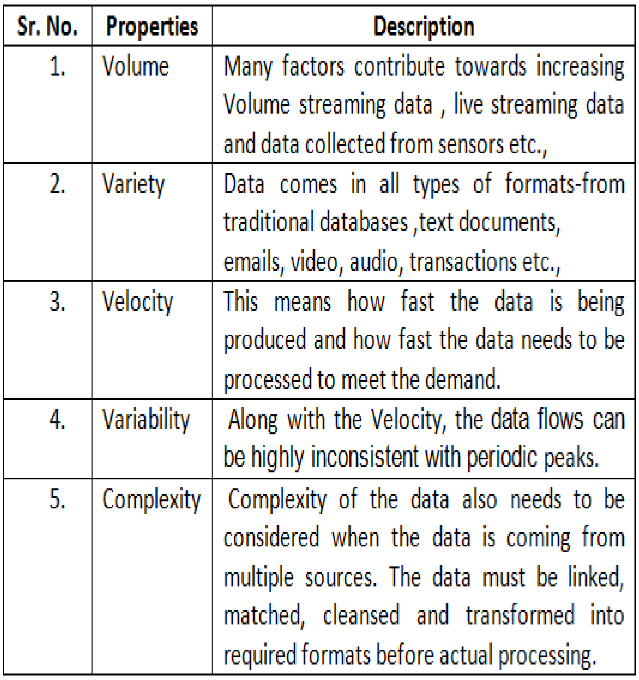


Figure 1: Growth rate of Big Data from 2011-2017[2]

II. BIG DATA OVERVIEW

Big Data is a compendium of big datasets that cannot be processed using traditional computing techniques. It is not a technique that can be worked on its own or in isolation; rather it involves many areas of business and technology. The properties of signify Big Data are volume, Variety, Velocity, Variability and Complexity as shown in figure 2[3]



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.

Figure 2. Properties of Big Data

Big data involves the data produced by different devices and applications. Some of the sources of Big Data are shown in the figure.

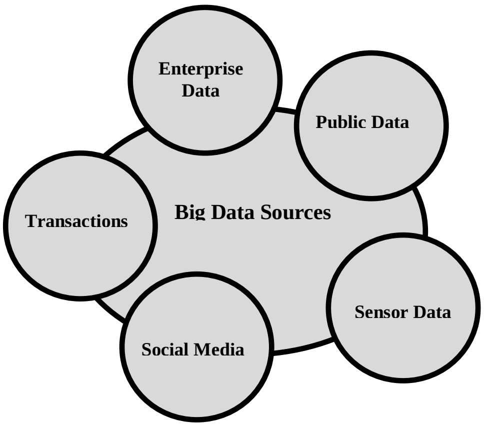


Figure 3. Some Sources of Big Data

III. PHASES IN BIG DATA PROCESSING

Before processing Big data it must be recorded from various data generating sources. After recording, it must be filtered and compressed. Only the relevant data should be recorded by means of filters that discard useless information. In order to facilitate this work specialized tools are used such as ETL. ETL tools represent the means in which data actually gets loaded into the warehouse. The figure 3 demonstrates different stages in the process.

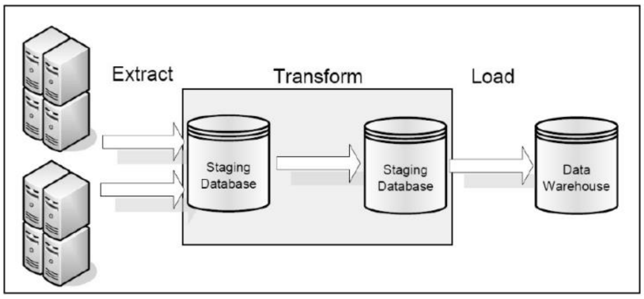


Figure 4. ETL process [4]

Table 1: Various phases in ETL [5]

Sr. No	Phase	Description of the Phase
1.	Extraction	In this phase relevant information is extracted. To make this phase efficient, only the data source that has been changed since recent last ETL process is considered.
2.	Transformation	Data is transformed through various phases[10] The phases are 1. Data analysis; 2. Definition of transformation workflow and mapping rules; 3. Verification; 4. Transformation; and 5. Backflow of cleaned data.
3.	Loading	At the last, after the data is in the required format, it is then loaded into the data warehouse.

IV. BIG DATA CHALLENGES

Big data due to its various properties like volume, velocity, variety, variability, value and complexity put forward many challenges. The Figure 5 shows various challenges in big data [12]. Figure 6 list some of the challenges in Big data along with its impact and risks involved. [6]

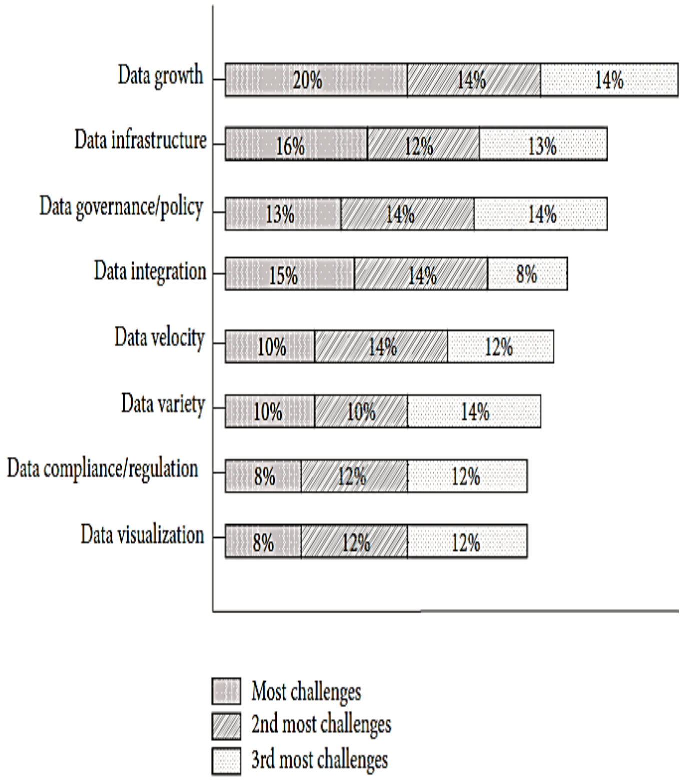
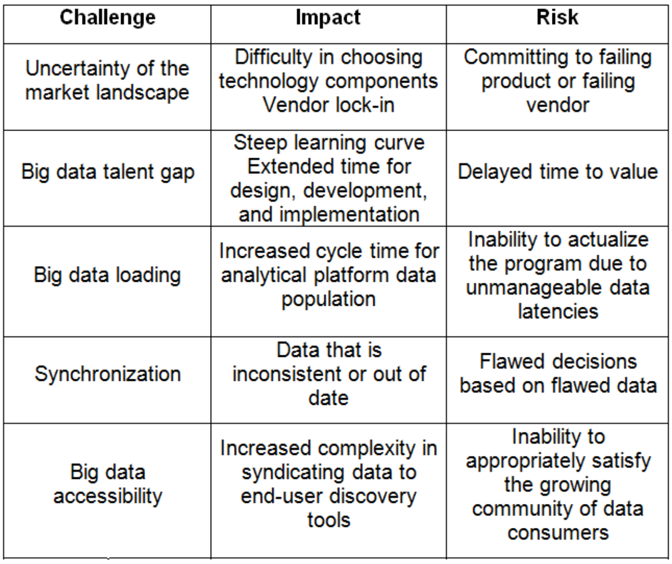


Figure 5. Challenges in Big data



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 6. Challenges, its impact and risk involved in Big data

V. 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

The architecture of Hadoop is shown in the figure 7. Hadoop has three layers. The two major layers are MapReduce and HDFS.

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 .[8] 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. [11]

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. [9]. 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).

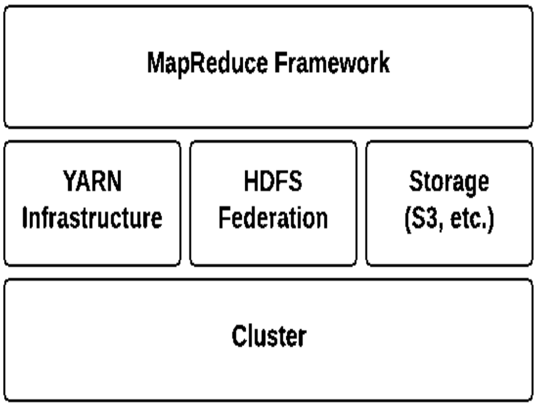


Figure 7. Hadoop Architecture

VI. CONCLUSION

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.

VII. REFERENCES

[1] https://www.idc.com/prodserv/4Pillars/bigdata

[2] www.Wikibon.org

[3] 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.]

[4] Golfarelli, M., & Rizzi, S. (2009). Data warehouse design: modern principles and methodologies. Columbus: McGraw-Hill

[5] 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.

[6] https://www.progress.com

[7] M. Chen, S. Mao, and Y. Liu, “Big data: a survey,” *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014

[8] Apache Hadoop (2013). HDFS Architecture Guide [Online]. Available: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.ht

[9] Amrit pal, Pinki Aggrawal, Kunal Jain, Sanjay Aggrawal *“A Performance Analysis of MapReduce Task with Large Number of Files Dataset in Big Data using Hadoop”* Forth International Conference on Communication Systems and Network Technologies, 2014.

[10] Rahm, E., & Hai Do, H. (2000). Data cleaning: problems and current approaches. Bulletin of the Technical Committee on Data Engineering, 23(4), 3-13.),

[11] Apache Hadoop (2013). HDFS Architecture Guide [Online]. Available: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.ht

[12] Intel, “Big Data Analaytics,”2012, http://www.intel.com/content/dam/www/public/us/en/documents/reports/data-insightspeer-research-report.pdf