Characteristics and Challenges of Big Data

Piyush Bhardwaj, Dr. Suruchi Gautam, Dr. Payal Pahwa, Neha Singh
Research Scholar, CSE Deptt. Uttarakhand Technical University, Dehradun
Associate Professor, CSE Deptt.,Delhi University, New Delhi
Professor,CSE Deptt., BPIT, New Delhi
Research Scholar, USICT, GGSIPU, New Delhi
piyushb88@gmail.com, suruchi.gautam@gmail.com,pahwa.payal@gmail.com, join7neha@gmail.com

Abstract: In today's digital-era, we are bowed down by the massive data that is generated at exponential rates. Technically, this massive data is referred to as Big Data. Simultaneously, the need to manage Big Data arises. Big Data, due to its high volume, velocity, veracity, value, variety, leads to various issues. In this paper, we talk about the various challenges faced because of the exorbitant amount of data. We not only face challenges in processing, but also in designing, analysing, storage, management, privacy and security issues.

Keywords— *Big Data, velocity, variety, veracity, verification, value, challenges*

*****

I. INTRODUCTION

If our ancestors were to return to earth they would be amazed to see the world revolutionized by the advancements in science and technology. There has been a transformation from manuscripts to file processing systems and now to more recent trends in data management for storing data. Since the data is being generated at exponential rates, the need of the hour is to devise methods to manage this huge data. According to Geoffrey Moore, without Big Data analytics, companies are blind and deaf wandering out on to the web like deer on a freeway. Typical sources of Big Data in current scenario are: Data Storage, docs, archives, social media, public web, business apps, media, sensor data, and machine log data [1]. According to "Human Face Of Big Data", during the first day of a baby's life, the amount of data generated by humanity is equivalent to 70 times the information contained the library of Congress [2].

The current statistics of Facebook are:

1.28 billion daily active users on average for March 2017
1.94 billion monthly active users as of March 31, 2017[3].

Statistics of YouTube:
More than 1 billion unique users visit YouTube each month Over 6 billion hours of video are watched each month on YouTube—that's almost an hour for every person on Earth 100 hours of video are uploaded to YouTube every minute. Millions of subscriptions happen each day. The number of people subscribing daily is up more than 3x since last year, and the number of daily subscriptions is up more than 4x since last year [4].

Twitter usage:
271 mil313 Million Monthly active users
500 million Tweets are sent per day [5].

This paper is worded as follows: Section II presents related work. In Section III we present various characteristics of Big Data. Section IV looks at the challenges due to big data. Section V concludes our work along with the future research and development.

II. RELATED WORK

In paper [6], the author moves forward with the challenges by beginning a collaborative research program into methodologies for Big Data analysis and design. Paper [7] presents the main issues along with the complete description of the tools and techniques associated with Big Data. In this article [8], an overview of big data's content, scope, samples, methods and challenges have been reviewed. A critical issue of privacy and security of Big Data is revisited. Paper [9]proposes the Scientific Data Infrastructure (SDI) architecture model. The authors have shown the models proposed can be implemented with the use of cloud based infrastructure services provisioning model. In paper [10] authors have analysed various social media sites like Flickr, Facebook, Locr and Google+. Based on this analysis they have discussed the privacy implications due to these sites.

III. CHARACTERISTICS OF BIG DATA

Big Data comes with a lot of promises – as a Dilbert Cartoon would have it, "It comes from everywhere. It knows all". Big Data requires a radical change from traditional data management to newer methods and techniques. It is characterized by its 5 main components referred as the 5 V's of Big Data [7] [8][9][11][12].

A. Velocity

Velocity in Big Data not only deals with the speed of the data coming from the sources but also it's processing in real-time, near real-time or in batch. For example, the data from Sensex

changes every second and is constantly moving to the database store which makes it difficult for the existing systems to perform analytics on the moving data.

B. Variety

Variety in data is the reason for big data being really big. Variety not only deals with complexity of big data but also information and semantics behind this data. Big data is generated from a number of sources, which may be structured, semi-structured or unstructured. It is the prominent obstacle in using large volumes of data as it not only includes traditional data warehouses but also semi structured form like graphs, charts, documents etc. This imposes requirements for new data storage designs to adapt to changing data formats and thus advanced analytic systems to increase processing speed.

C. Volume

The term Big data itself defines the torrent of data. The largest cardinalities of the most datasets: the numbers of distinct entities about which the observations are made are small as compared with the total number of observations. Hence the volume increases manifold. It is of the range of petabytes and exabytes but is supposed to increase in the range of zetabytes in the near future. As data volume increases, need to store and process such large volumes of data arise.

D. Veracity

The dictionary meaning of veracity is 'conformity to truth or a fact'. Thus this aspect deals with the authenticity and consistency of the data. Big Data veracity ensures that the data is protected from modification and unauthorized access. Hence the data becomes secure and its integrity is preserved.

E. Value

The data may be in any form but our main purpose is to extract the information from it. The objective is to derive the insights not the quantity of data. This closely relates value to volume and variety. The value of data is different for different people: for industrialists, value refers to obtaining maximum profit from their businesses whereas for IT professionals, the main concern is the technicalities of storage and processing.

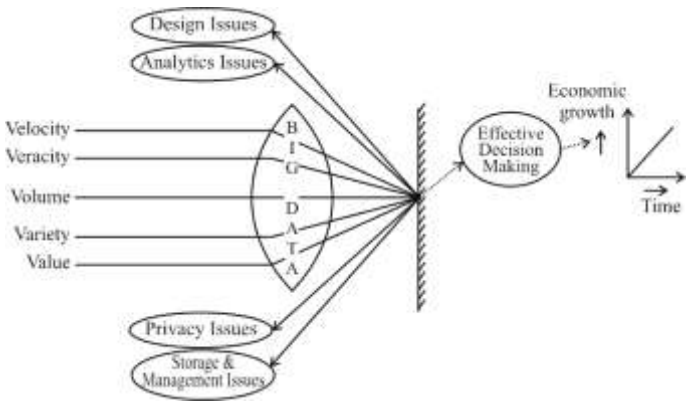

FIG. 1: CHARACTERISTICS AND CHALLENGES OF BIG DATA

IV. CHALLENGES DUE TO BIG DATA

A. Design challenges

At the rudimentary stages, due to the absence of well-defined tools to deal with the profuse amount of data, we face a lot of design challenges.

1. Fault tolerance: Fault tolerance [1] computing is extremely hard involving large volumes of varied data. Therefore designing of 100% reliable fault tolerable machines or software is simply next to impossible. Thus the need of the hour is to devise methods which reduce the probability of failure to an acceptable level.
2. Heterogeneous data: Unstructured data represents almost every kind of data collected through various public social media, recorded meetings, fax transfers, emails, audio and videos. It is completely raw and unorganised and therefore working with this type of data is onerous task and to an extent in convenient. Converting such data to structured data is a tedious job for the software designers and developers.

Reference [13] looks into the challenges faced in devising methods for Knowledge Discovery from this exorbitant amount of data in three principles:

a. KDD involves a wide range of analytic methods as distributed programming, pattern recognition, data mining, statistical and visual analysis and human-computer interaction. Therefore software architecture and designs for knowledge discovery in big data must support a variety of methods and analysis techniques rather than forcing users to use a limited set of tools.
b. In addition of data analysing methods, the system design should provide means of storing and processing of data at all stages. A single storage may suffice for small volumes but lack to fulfil its need for large amount of data and thus efficient processing cannot be carried out.
c. No analysis or processing is complete if the data obtained is not accessible or it cannot provide the required insight to the given data. The different approaches to accomplish this are using open standards, adoption of lightweight architectures and exposing results via web-based API's.

B. Analytic Challenges

The explosion of data today is changing our world as we are in the very nascent stage of the so-called "Internet of Things". Big data brings along with it some huge analytical challenges [7][9][13]. Data analysis is being conducted at a velocity that is increasingly approaching real time, i.e., there is growing potential of big data analytics. Some organizations are even carrying out predictive analysis for fraud detection. Big data analytics is very difficult but it helps in finding patterns and

relationships among lots of data. It is not always true as the information revealed by big data analytics isn't necessarily perfect. It may not help in rendering judgements. As big data is becoming a "buzzword", the skills required to analyse such large amounts of data are not sufficient. It requires advance skills for decision-making. The techniques like data fusing make data analytics more powerful but require de-identification and re-identification of data for privacy purpose. In an effort to process large sets of imperfect & incomplete data, a "Data to Knowledge to Action" initiative has been launched by DARPA, which focuses on tools & techniques needed to access, organize & discover from huge volumes of digital data. In Philadelphia, police are using software designed to predict which parolees are more likely to commit a crime after release from prison & thus should have greater supervision.

C. Storage and Processing Challenges

It is becoming very challenging and complex to produce knowledge from large datasets, which includes data from various sources including online activities, scientific activities, research institutions, large business enterprises etc. [13]. It is difficult for a single or a few machines to store, analyse and process this huge amount of data in given time limits, as it is not known in the beginning which data is important to be stored. To meet the storage and processing demands of big data, uploading it onto cloud is an option but it doesn't help in linking and correlating the data available, which is necessary to extract value and important information out of it, consequently hampering the effective decision making process. Effective knowledge discovery requires proper storage and organizational practices to use data to best advantage. Not all data can be easily modelled using relational techniques as data can be unstructured or semi-structured. Data stored in the form of graphs and hierarchical documents helps in discovering interesting relationships among massive datasets, thus helping visual analytics and easy processing of data. To model whole data, it needs to be scanned completely which surpasses the time tolerance. So building up the indexes from the very beginning while storing and collecting the data meets the time constraint and preserves the integrity and authenticity of data as well.

D. Privacy and Security Issues

The FTC's recent report on data brokers warned that, "collecting and storing large amounts of data not only increases the risk of a data breach or other unauthorized access but also increases the potential harm that could be caused."

A new report finds that 432 million online accounts in the US have been hacked this year, concerning about 110 million Americans. In the last year, 70 million Target customers, 33 million Adobe users, 4.6 million Snapchat users, and potentially all 148 million eBay users had their personal information exposed by database breaches. Earlier this month, the President's science advisors found little risk in the continued collection of personal data[14]

One of the major problems that a layman faces due to big data is privacy. Privacy and security means protecting our data and information from unauthorized access and modification. Privacy of data is no longer in the hands of the owner[8].

According to Chairwoman Ramirez, the First Commandment of data hygiene is: "Thou shall not collect and hold onto personal information unnecessary to an identified purpose." Similarly, Commissioner Julie Brill laments the fact that firms, "Without our knowledge or consent, can amass large amounts of private information about people to use for purposes we don't expect or understand."[15]

This issue not only deals with technical aspects but also has legal significance. The personal information of the users is accessed by the large companies and associations to add value to their businesses. This is done by creating insights in their lives which they are unaware of [16].

Another important upshot is social assortment that means an educated person could take out advantages out of big data analysis whereas an underprivileged can be identified and treated worse [7].

Big data used by law enforcement will increase chances of certain tagged people to suffer from adverse consequences without the ability to fight back or even having the knowledge that they are being discriminated [7].

V. Conclusion and Future Scope

Big Data was the watchword of year 2012. Even before one could really understand what it is, it began getting tossed up in huge doses in almost every analytical report. There is a plethora of data available and what the time demands, is a transition from the existing methods to the newly devised techniques. Our paper gave an insight about Big Data, its charateristics and the challenges faced due to Big Data. For the enterprises to add value to their businesses they should embark to flexible and multidisciplinary big data analytic methods. Pioneers such as Google have devised a tool that is popularly known as Hadoop, open source software that allows distributed processing of large data sets. Another tool developed for the same purposeis HPCC (high performance computing clusters) systems whose data model is defined by the user unlike Hadoop. But a lot is still to be done to bridge the gap between the myriad amount of data and their storage and management methods. Our future work will focus on bridging the gap between big data and store and management methods.

References

[1] Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce", 2012 nirma university international conference on engineering, nuicone-2012, 06-08december, 2012
[2] http://www.humanfaceofbigdata.com
[3] http://newsroom.fb.com/company-info/
[4] https://www.youtube.com/yt/press/statistics.html

[5] https://about.twitter.com/company

[6] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, "Big Data: Issues and Challenges Moving Forward", IEEE, 2013 46th Hawaii International Conference on System Sciences

[7] AvitaKatal, Mohammad Wazid, R H Goudar, "Big Data: Issues, Challenges, Tools and Good Practices", IEEE Conference 2013

[8] Seref SAGIROGLU and Duygu SINANC, "Big Data: A Review", IEEE Conference 2013

[9] Yuri Demchenko, Paola Grosso, Cees de Laat, Peter Membrey, "Addressing Big Data Issues in Scientific Data Infrastructure", IEEE Conference 2013

[10] Matthew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt,"Big Data Privacy Issues in Public Social Media", IEEE, 6th International Conference on Digital Ecosystems Technologies (DEST), 18-20 June 2012.

[11] Xin Luna Dong, Divesh Srivastava, "Big Data Integration", IEEE, ICDE Conference 2013.

[12] Du Zhang," Inconsistencies in Big Data", IEEE, Proc. 12th IEEE Int. Conf. on Cognitive Informatics & Cognitive Computing (ICCI*CC'13)

[13] Edmon Begoli, James Horey, "Design Principles for Effective Knowledge Discovery from Big Data", IEEE, 2012 Joint Working Conference on Software Architecture & 6th European Conference on Software Architecture

[14] http://epic.org/privacy/big-data/

[15] http://www.techpolicyinstitute.org/files/lenard_rubin_thebigdatarevolutionprivacyconsiderations.pdf

[16] M. Smith, C. Szongott, B. Henne and G. Voigt , "Big Data Privacy Issues in Public Social Media", Digital Ecosystems Technologies (DEST), 6th IEEE International Conference on, Campione d'Italia, June 2012