JCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Big Data: A New Way to Look At World

Kamal Kumar Ranga[1*], **C.K. Nagpal**[2]

[1] Department of Computer Engineering, YMCA University of Science and Technology, Faridabad, Haryana, India
[2] Department of Computer Engineering, YMCA University of Science and Technology, Faridabad, Haryana, India

Available online at: www.ijcseonline.org

Accepted: 23/Jul/2018, Published: 31/Jul/2018

Abstract- Big Data is a term which now a days everyone is aware of. This has changed view of every normal human or technocrat towards world. Technocrats sees this as a world full of challenges yet more of opportunities to explore this world more vividly. The voluminous data evolving these days with exponential rate is one of the biggest concern of technological world, yet bigger challenge is to explore and extract useful information out of this world and exploit is to the fullest to get maximum benefit. This paper ca be taken as reference paper or a tutorial for the one who want to gain in depth basic knowledge of big data and pursue research further.

Keywords: Big Data, Varacity, Variety, Value, Volume, Map Reduce, Hadoop

I. INTRODUCTION

BIG DATA:
Definition: BIG DATA is a term evolving since last decade, which describes any voluminous amount of structured, semi-structured and unstructured data coming from enormous sources such as web, medical, business, scientific organizations, software logs, cameras, microphones, RFID's readers, and WSN and social media etc.

With BIG DATA the world's technological per-capita capacity to store information has doubled every 40 months since 1980's and as of 2012, every day over 2.5 EB (2.5×10^{18}) of data were created, which has grown exponentially up to many order now [1]. The main focus point is that data should have potential to be mined for useful information. Although big data doesn't refer to any specific quantity or size of data but the term is often used when we speak about PB and EB of data.

II. CHARACTERISTICS

Big data can be characterized by 7Vs:

1. **VOLUME:** The massive amounts of data generated/second is termed as volume. Also, over 90% of world's data is generated in last 2 years. This volume of data every day is posing an immediate challenge.

For Example**:**
- Daily upload on Facebook crosses 100's TB data.
- Over 75 MN events are analyzed by Akami per day to target online ads.
- Wallmart has over 1 MN transactions every hour.

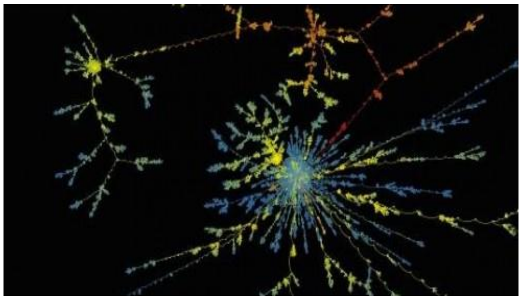

Fig 1: 2.5 Lakh Facebook comments Visualization.

2. VELOCITY: It refers to exponentially increasing speed at which data is created, processed, stored and analyzed by RDB's. Processing data at exceptionally high speed in real-time is an area of particular interest, which allows companies to do things like display personalized ads on webpages based on your recent search, views and purchase history.

For example:
- Stock Exchange of NY captures over 1 TB information daily. It reacts and analyze business data fast enough. Although, with increasing data this still troubling businesses where speed at peak periods being inconsistent.
- In 1999, Wal-Mart's data warehouse stored over 1K TB of data and in 2012, it had access to over 2.5 PB of data.
- We upload over 100 hrs of video on YouTube and send over 200 MN mails & posts 0.3 MN tweets.[2]

3. VARIETY: Various forms of data that we collect and use is called variety. As data is generated in wide variety of formats categorized as unstructured, structured and semi-structured.
The explosion of data over social media further making it more challenging.

Gone are the days when company's data was neatly organized in table for analysis. Now, over 90% data is 'unstructured' which is in all shapes, sizes and formats like sensors and geospatial data, photos, videos and even tweets etc.[3].

4. VARIABILITY: Data whose meaning keeps changing is referred as Variability. With each passing year amount of generated data increases exponentially. Thus, we now know much about what defines big data, So, IBM introduced fourth V, Veracity, as outlined here.

Brian Hopkins has cited "Supercomputer WATSON a prime example of this. In a game show Jeopardy, Watson had to dissect an answer into its meaning and to figure out what the right question was". The context changes the meaning of words greatly.

Fig 2: The Watson supercomputer.

For Example: Say a company was trying to gauge sentiment towards a cafe using these 'tweets':

"Delicious muesli from the @kamalcafe: what a great way to start the day!"

"Greatly disappointed that my Kamal Cafe have stopped stocking BLTs."

"Had to wait in line for 45 minutes at the Kamal Cafe today. Great, well there's my lunchbreak gone…"

Evidently, "great" here is donot signifies positive emotion, rather companies should develop dedicated programs that 'understand' context of word in which it is said to decode its precise meaning. Challenging yet it possible; e.g.: Bloomberg, has introduced program that correctly analyzed social media buzz about companies last year [4].

5. VERACITY: It refers to uncertainty of data which is due to incompleteness & inconsistency of data that leads to another challenge. Although everyone knows that if data is not accurate it is virtually worthless which is true particularly in cases that involves supplying data to unsupervised machine learning algorithm and automated decision-making systems.

Cloudera's Sr. Director said "Let's say that, in theory, you have customer behavior data and want to predict purchase intent. In practice you have log files in 4 formats and from 6 systems, with noise and errors. These have to be copied, translated and unified." Owens' US counterpart, Josh Wills, said "We are Data Scientist but our job revolves so much around the cleaning up of messy data that we are more a Data janitor." [5].

6. VISUALISATION: Visualization refers how we represent data that is understandable, readable and accessible. Visualizations can have good number of variables/parameters apart from traditional x/y used for standard charts, and that lets you present information and findings clearer, is yet another challenge for Big Data. This problem has let various new visualization packages appear every now and then in market.

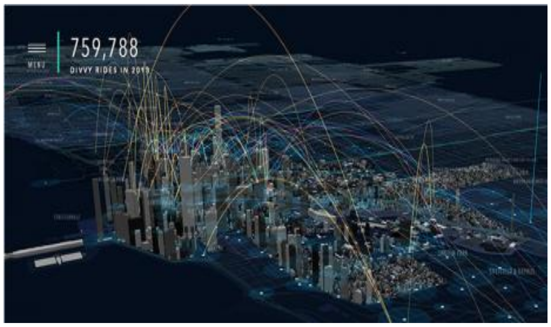

Fig 3*: A visualization of bike rides across Chicago.*

7. VALUE: Value acronyms importance of data which can only gained by thorough and effective mining and analysis. Huge amount of data can be used to generate business opportunities. Valuable business information can be extracted from numerous forms of data that can help in improving supplychains & programme planning and to measure performance, track sales so as to enhance on-demand business. Big data strategies enables better analysis of business data that leads to profitable growth [6] [7].

III. ANALYSIS

The analysis of Big Data is often done by analyzing large data sets in real-time by storing large data sets over distributed clusters and then combine and coordinate and process this data. There exist various tools for this like Hadoop for storing and MapReduce for processing this data.

As loading Big Data onto traditional RDB's requires too much time and money, new approaches have emerged that relies very less on data schema and data quality. Rather, raw data with extended metadata is gathered then processed by applying complex machine learning and AI algorithms to find repeatable patterns [8].

With this another term has evolved called small data which meant to describe data whose format & volume can be used for self analytics. A commonly quoted axiom is that "big data is for machines; small data is for people." [9].

IV. BIG DATA TECHNOLOGIES

Big corporations are constantly focused on creating and acquiring unstructured data, and experimenting with cloud computing and associated technologies using MapReduce and Hadoop for high-speed data analysis and across to cluster of computers. Google and Amazon are the companies whose technological use has raised modern business status. Also there are few other companies whose successes represents lode star for Big Data [10]. This has led open source to dominate much of Big Data that makes application development managers to find varied solutions to Big Data.

Hadoop: Hadoop is Apache Software Foundation project and is Java-based, free programming framework that supports processing of large data sets in distributed environment. Hadoop permits scalable, reliable distributed computing to process huge datasets across cluster of computers using simple programming models through its library. It is designed such that it can scale up from 1 to 1000's of machines, provided each machine has local computation and storage. Further, High availability is offered by library itself. The library is designed such a way that it detects & handles failure at application layer which delivers highly-available service on top of cluster of computers [11].

MapReduce: MapReduce is a software framework using which programs are written that can process very huge (massive) amount of unstructured data in parallel across distributed cluster of 1 or 1000's of computers in reliable, fault-tolerant manner. It splits input data into small independent chunks which are further processed by map() parallelly. The framework then sorts outputs of the mappers, which is further input to reduce() [12].

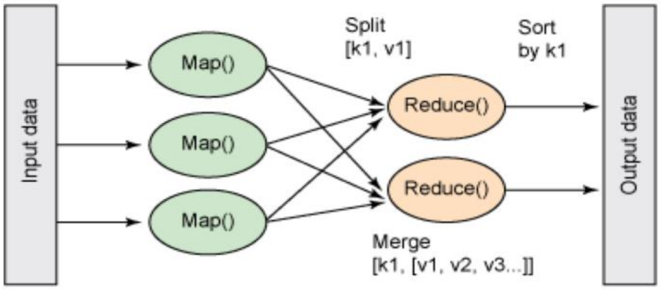

Fig 4: MAPREDUCE processing within the Hadoop cluster.

These days Apache Hive data warehousing component having HiveQL a query language that translates SQL-like queries into MapReduce jobs automatically, is also gaining popularity.

V. MAJOR CHALLENGES

To realize the full potential of Big Data, various technical challenges needs to be addressed such as:

- **Speed:** Merely finding and analyzing data is not sufficient but finding it quickly is new demand of today's hypercompetitive business environment. Although, visualization helps is quick and better decisions making, but the bigger challenge here is to go through huge volumes of data while accessing the level of details needed, at extremely high speed [13].
- **Understanding the data:** Understanding data and getting it into right shape requires lot of efforts. This helps in better visualization for better analysis. For example, for social media data, you must need user in general sense, also without knowing context of data, visualization tools are less likely to be of value.
- **Data Quality:** If data is not accurate and on time then there may be a danger of loss, harm, or failure. This challenge becomes more complex and pronounced when the volumes of information involved is very high.
- **Displaying meaningful results**: With variety and enormous information it is very difficult to present this information in clearly. For example, suppose we want to plot about a billion rows of retail data, which need to be compared, here it becomes very difficult for user find data points.
- **Query Processing, and Analysis**: Big Data's Query methods must be capable to deal with heterogeneous, noisy, dynamic, untrustable data that can be characterized by complex relationships. The redundant relationships between data offers an opportunity for thorough analysis of data so as to improve data trustworthiness, which requires scalable mining algorithms and powerful computing infrastructures for data analysis and query processing.

VI. APPLICATIONS

1. Government :

Adoption and use of Big Data in government processes is very beneficial in terms of cost, productivity and innovation. Although, data analysis may sometimes require multiple departments work collectively to create new & innovative processes that deliver desired outcome. For example:

India

- BJP's bumper win was propelled by big data analysis in Indian General Election 2014.[14]
- Numerous techniques on policy augmentation and the response of Indian electorate is analyzed by Indian Govt. using Big Data only.

United States of America

- Obama's win in re-election campaign was backed by Big Data only. [15]

- US Government owns six of ten most powerful supercomputers in world.[16]

2. Manufacturing

TCS study of Global Trends suggests, Big Data's greatest benefits in manufacturing is improvements in planning of product quality and supply. Further, Big Data in manufacturing provides an infrastructure for transparency that unveils uncertainties like inconsistent component, performance and availability. Yet another approach, that requires huge amount of data along with advanced prediction tools called predictive manufacturing that helps getting useful information by systematic processing [17].

3. Technology

- EBay has 2 data warehouses of capacity 7.5 PB & 40PB dedicated for searching data, providing recommendations to customers and merchandising.
- Millions of back-end operations are handled by Amazon every day. Further it also handles over half millions of queries from third-party sellers. It used Linux-based technology including world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB.[18]

4. Retail

Over a million transactions are handled every hour at Walmart, which estimates to a size of over 2.5 PB, that equals 167 times of information contained in all books at US Library of Congress.

5. Research

Originally, human genome decoded in over 10 years while now it's less than 1 day. Likewise the cost of various technologies got cheaper with introduction of more complex and active technologies which led further interests of scientists to explore more in various domains to find better solution to ease human life in all aspects [19].

6. Education

Education is one of prime area need immediate eye to analyze real facts that can lead to excellent growth of our young generation as Big Data has potential to revolutionize education as well. Big Data can be applied to analyze the better contents, instruction manual and also monitor growth and performance of each individual student at any level [20]. This data so collected can be effectively used to design most effective approaches for education.

7. Medical

It is believed that use of Big Data can improve the quality and cost of healthcare services. Using Big Data more personalized and preventive care can be given to patients. Further, complex algorithms can be designed to combat critical illness and analyzing the success rate in advance to save a patient life.

8. Employment

A considerable demand in information system specialists is witnessed in companies like Microsoft, Oracle, SAP, EMC, HP and Dell. These companies has spent billions of dollars on data management and analytics. In 2016 this industry worth more than $2000 bn and growing at an exponential rate [21].

9. Serves Humanity

Recently an IIT surveyed in nearby villages that there is flood in some areas during rainy season while others suffering drought. They analyzed this data and implemented a system to divert extra water to drought places. This has led the even distribution of water that eased lives of the villagers to great extent.

VII. CONCLUSION

This papers gives detailed information what Big Data is, how it is analysed, what technologies are useful for real analysis of Big Data and what are major challenges which big data is facing today. We have also discussed various research and real time application of Big Data. This paper can be used as reference or base paper for the persons, scholars or technocrats to get in depth knowledge about Big Data and its future.

REFERENCES

[1]. IBM Bringing BIG DATA to enterprise: http://www-01.ibm.com/software/in/data/bigdata/

[2]. Jacobs, A. (6 July 2009). "The Pathologies of Big Data". ACMQueue.

[3]. Magoulas, Roger; Lorica, Ben (February 2009). "Introduction to Big Data". Release 2.0 (Sebastopol CA: O'Reilly Media) (11).

[4]. Avnet Advantage: The Blog http://www.ats.avnet.com/na/en-us/news/Pages/The-5-Vs-of-Big-Data.aspx

[5]. Dataconomy: Understanding Big Data : http://dataconomy.com/seven-vs-big-data/

[6]. Data Alchamists: http://dataalchemists.com.au/2013/05/5-vs-big-data/)

[7]. Donna Burbank, Enterprise Architects: the 5 V's of Big Data http://enterprisearchitects.com/the-5v-s-of-big-data/

[8]. Billings S.A. "Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains". Wiley, 2013

[9]. Bertolucci, Jeff "Hadoop: From Experiment To Leading Big Data Platform", "Information Week", 2013. Retrieved on 14 November 2013.

[10]. Bertolucci, Jeff "Hadoop: From Experiment To Leading Big Data Platform", "Information Week", 2013. Retrieved on 14 November 2013.

[11]. Apache Hadoop: http://hadoop.apache.org/

[12]. Big Data Technologies: http://bigdata.impetus.com/technologies.

[13]. Big Data Challanges: http://www.sas.com/resources/asset/five-big-data-challenges-article.pdf.

[14]. Drowning in numbers -- Digital data will flood the planet—and help us understand it better. The Economist, Nov 18, 2011. http://www.economist.com/blogs/dailychart/2011/11/big-data-0

[15]. Using Data for Systemic Financial Risk Management. Mark Flood, H V Jagadish, Albert Kyle, Frank Olken, and Louiqa Raschid. Proc. Fifth Biennial Conf. Innovative Data Systems Research, Jan. 2011.

[16]. Understanding individual human mobility patterns. Marta C. González, César A. Hidalgo, and Albert-László Barabási. Nature 453, 779-782 (5 June 2008)

[17]. Bamford, James (15 March 2012). "The NSA Is Building the Country's Biggest Spy Center (Watch What You Say)". Wired Magazine. Retrieved 2013-03-18.

[18]. UN GLobal Pulse (2012). Big Data for Development: Opportunities and Challenges (White p. by Letouzé, E.). New York: United Nations.

[19]. Layton, Julia. "Amazon Technology". Money.howstuffworks.com. Retrieved2013-03-05.

[20]. Wingfield, Nick (2013-03-12). "Predicting Commutes More Accurately for Would-Be Home Buyers - NYTimes.com". Bits.blogs.nytimes.com. Retrieved 2013-07-21.

[21]. AMPLab at the University of California, Berkeley". Amplab.cs.berkeley.edu. Retrieved 2013-03-05.