ISSN No. 0976-5697

Volume 8, No. 1, Jan-Feb 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Big Data: Meaning, Challenges, Opportunities, Tools

Vinti Parmar
Dept. of Computer Science
Indira Gandhi University
Meerpur, Rewari, India

Jyoti Yadav
Dept. of Computer Science
Indira Gandhi University
Meerpur, Rewari, India

Abstract: Big data is certainly one of the biggest buzz phrases in Information Technology today. The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the Next Wave of InfraStress". Big data is a technological capability that will force data centers to significantly transform and evolve within the next five years. Big data infrastructure is distinctive and can create an architectural upheaval in the way systems, storage, and software infrastructure are connected and managed. Big data is an amalgam of large and varieties of data sets including structured data, semi structured data and unstructured data so it's beyond the capability of traditional tools to capture, store, process and analysis of big data. It is true that big data have potential of unlocking new sources of development in many fields but at the same time researchers are being confronted challenges with big data. This paper divulge the various challenges faced with big data and opportunities realized with big data.

Keywords: Big data, Challenges, Opportunities, Security Issues.

I. INTRODUCTION

Big data refers to the collection and subsequent analysis of any significantly large collection of data that may contain hidden insights or intelligence (user data, sensor data, machine data). When examine properly, big data can bring new business insights, open new markets, and create competitive advantages. The growth of mobile users has increased enterprise aggregation of user statistics—geographic, sensor, capability, data—that can, if properly combined and analyzed, provide extremely powerful business intelligence. In addition, the better use of sensors for everything from traffic patterns, purchasing behaviors, and real-time inventory management is a primary example of the vast increase in data. Much of this data is gathered in real time and provides an opportunity if it can be analyzed and acted upon quickly. Machine-to-machine interchange is another often unidentified source of big data. The rise of security information management (SIM) and the security Information and event Management industry is at the heart of gathering, analyzing, and proactively responding to event data from active machine log files. Although it may be clear that new technologies and new forms of personal communication are driving the big data trend, consider that the global Internet population grew by 6.5% from 2010 to 2011 and now represents over two billion people[1]. This may seem large, but it suggests that the vast majority of the world's population has yet to connect, while it may be that we never reach 100% of the world's population online (due to resource constraints, cost of goods, and limits to material flexibility), increasingly those that are online are more connected than ever. Just a few years ago, it was realistic to think that many had a desktop (perhaps at work) and maybe a laptop at their disposal. However, today we also may have a connected smartphone and even a tablet computing device. So, of today's two billion connected people, many are connected for the vast majority of their waking hours, every second generating data:

- In 2011 alone, human beings created over 1.2 trillion gb of data.
- Data volumes are estimated to grow 50 times by 2020.
- 72 hours of video are added to Youtube every minute.
- There are 217 new mobile Internet users every minute.
- Twitter users send over 100,000 tweets every minute (that's over 140 million per day).
- Companies, brands, and organizations receive 34,000 "likes" on social networks every minute.

International data Corporation (IdC) predicts that the market for big data technology and services will reach $16.9 billion by 2015 with 40% growth over the prediction horizon[2]. Not only will this technology and services influence big data technology providers for related SQL database technologies, Hadoop or Mapreduce file systems, and related software and analytics software solutions, but it also will impact new server, storage, and networking infrastructure that is purposely designed to leverage and optimize the new analytical solutions. Major attributes of Big Data are:

- **Volume**: It means size or amount of big data that is between terabyte and petabyte. So which data is big or not can to be said accurately. It depend on size of organisation.
- **Variety**: Big data does not mean structured data only. It's a combination or mixture of varieties of data that is structured, unstructured and semi structured data. Today's generation of unstructured data is more than structured data and analysis of such data reveals valuable information which is not possible to get from structured data only on which business world relied in past years so there is need of new technologies for managing such diverse types of data. So it's a challenging task to deal with different varieties of data captured from various sources that arriving.
- **Velocity**: speed at which data is produced and processed continuously at high pace. For cope up with that challenging task there is need of adoption of

new technology that have potential of extracting meaningful information from large and diversified data that is arriving continuously.

- **Veracity**: refers to the degree in which a leader trusts the used information in order to take decision. So getting the right correlations in Big Data is very important for the business future. Combining dissimilar data from different sources provide valuable insight rather than isolated data.
- **Variability**: Variability refers to data whose meaning is constantly changing. This is particularly the case when gathering data relies on language processing.
- **Value**: The potential value of Big Data is huge. Speaking about new Big Data initiatives in the US healthcare system last year, McKinsey estimated if these initiatives were rolled out system-wide, they "could account for $300 billion to $450 billion in reduced health-care spending, or 12 to 17 percent of the $2.6 trillion baseline in US health-care costs"[3]. However, the cost of poor data is also huge- it's estimated to cost US businesses $3.1 trillion a year. In essence, data on its own is virtually worthless. The value lies in rigorous analysis of accurate data, and the information and insights this provides.

So what does all of this tell us about the characteristics of Big Data? Well, it's vast and rapidly-escalating, but it's also noisy, messy, constantly-changing, in hundreds of formats and virtually worthless without analysis and visualization. In the world of Big Data, data and analysis are totally interdependent- one without the other is futile, but the power of them combined is virtually inexhaustible. Big data encompasses high velocity, huge volume and variety of data like structured data, semi structured and unstructured data that have strength of unlocking new sources of development, providing valuable insights and decision making.

- **Unstructured data** – Unstructured data is not in a particular format so can not fit properly in database. It is in alike form in which it is collected and heterogeneous in nature. For example: pdf, audio, video, images, emails. It is estimated that creation of unstructured data is more than structured data in an organization because analysis of unstructured data have potential of providing better accurate insights. Unstructured data can also be computer or human generated Satellite images, video, audio are machine generated data while mobile data, social media data is human generated data.
- **Semi Structured data** – It is in between structured and unstructured data that is to some extent processing is done on them. That data is not in proper organised form. For example: html, xml Since big data is collection of structured, unstructured and semi structured data so big data have enough potential to do tasks that were impossible earlier like disease management, crime management, providing new direction in business enterprises. Many corners or fields of science currently facing exponential growth in volume of data as compared to past years. It is true that big data revolutionized the research field but at same time challenges are faced in dealing with big data so there is need of emergence of new technology to utilize full potential of big data and addressing confronted challenges. As discussed by The Economist "Managed well, the data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments to accounts" [4].

II. RELATED WORK

Provide an approach for research efforts towards developing highly scalable and autonomic data management systems associated with programming models for processing Big Data. Aspects of such systems should address challenges related to data analysis algorithms, real-time processing and visualization, context awareness, data management and performance and scalability, correlation and causality and to some extent, distributed storage[1]. Summarize opportunities and challenges with big data. Recent technological advances and novel applications, such as sensors, cyber-physical systems, smart mobile devices, cloud systems, data analytics and social networks are making possible to capture, process, and share huge amounts of data referred to as big data and to extract useful knowledge, such as patterns, from this data and predict trends and events. Big data is making possible tasks that before were impossible, like preventing disease spreading and crime, personalizing healthcare, quickly identifying business opportunities, managing emergencies, protecting the homeland, and so on[2]. Provide an approach for sources of structured and unstructured big data. Unstructured data is everywhere. In fact, most individuals and organizations conduct their lives around unstructured data. Successful decision-making will increasingly be driven by analytics-generated insights. From the lowest-level network enablers to high-level business support systems, there will be an opportunity to utilize insights gathered from careful analysis of network data. Big data is at the core of this opportunity[3]. But the traditional view of big data is not enough. Rather than focusing exclusively on what technology big data brings, we have to look at what value it can create[4]. This paper explores big data analytics, the opportunities it gives rise to, and how big data should be expanded to support analytics. The enormous growth in the amount of data that the global economy now generates has been well documented, but the magnitude of its potential impact to drive competitive advantage has not. It is my hope that this briefing urges all stakeholders— executives who must fund analytics initiatives, IT teams that support them and data scientists, who uncover and communicate meaningful insight—to go boldly in the direction of "Big Analytics." This opportunity is enormous and without precedent.

III. CHALLENGES WITH BIG DATA

Already some success is achieved from big data in some fields like Sloan digital sky survey so it's mean there is some potential in big data and benefits are also real but still some challenges like scalability, heterogeneity, integration, privacy, security etc. need to be addressed for realizing full potential of big data. One of the major challenge is transformation of unstructured data to structured form for accurate and timely processing[5]. Challenges with big data starts with very first phase of big data analysis pipeline that is data acquisition phase. It's a challenging task to determine what data to keep,

what to discard and how to efficiently store the data. Other challenges are faced in data cleaning, integration and data analysis phase of big data analysis pipeline.

Few major challenges of big data are as below:

- Short of efficient tools and techniques for safely organizing large-scale data and distributed data sets. Both companies and law enforcement agencies increasingly rely on video data for surveillance and criminal investigation. Closed-circuit television (CCTV) is ubiquitous in many commercial buildings and public spaces. Police cars have cameras to record pursuits and traffic stops, as well as dashcams for complaint handling. Many agencies are now experimenting with body-worn video cameras to record incidents and gather direct evidence from a crime scene for use in court, obviating the need for eyewitness versions of events. Taser guns also now come equipped with tiny cameras. Because all of these devices can quickly generate a large amount of data, which can be expensive to store and time-consuming to process, operators must decide whether it is more cost effective to let them run continuously or only capture selective images or scenes. In order to get actionable knowledge, big data must be filtered that is a major challenge. Also research must be done on how to intelligently filter raw data acquired from different sources without missing useful information.
- Production of right metadata (data about data) is also a difficult task because it is the metadata that describes data and its recording procedure. It also essential for proper interpretation of result.
- Security and privacy issues while sharing data and susceptible ever increasing public databases. Security and privacy affairs are growing as big data becomes more and more approachable.

- The accumulation and compilation of massive quantities of heterogeneous data are now possible. Data sharing on large scale is becoming routine among scientists, clinicians, businesses, governmental agencies, and citizens. However, the tools and technologies that are being developed to manage these massive data sets are often not designed to incorporate adequate security or privacy measures, in part because we lack sufficient training and a fundamental understanding of how to provide large-scale data security and privacy.
- Planned or malicious leakage of data leakage represents a big problem. Data hackers have become more damaging in the era of big data due to the availability of massive volumes of publically available data, the ability to store huge amounts of data on portable devices such as USB drives and laptops, and the accessibility of simple tools to acquire and integrate disparate data sources. According to the Open Security Foundation's DataLossDB project hacking accounts for% of all data breach incidents, with theft accounting for an additional 24%, fraud accounting for 12%, and web-related loss accounting for 9% of all data loss incidents. Greater than half (57%) of all data loss incidents involve external parties, but 10% involve malicious actions on the part of internal parties, and an additional 20% involve accidental actions by internal parties today's era of big data.
- Big data is constitution of varieties of data that is structured, unstructured and semi structured. Different sources generate diverse types of data. For example data obtained from sensors is in structured form while data obtained from mobile, satellite is unstructured data. But for analysis of big data for getting valuable insight all data must be converted into structured form that is a challenging task.

All these challenges requires evolution and implementation of new powerful technology for dealing with big data and to get full potential of big data. Current analysis techniques can be enhanced with some more advancement to deal with some of the challenges. Regardless of various challenges there are myriads of Opportunities in Big Data. Some major Big Data opportunities are listed below:

- Growing customer demands for smarter products, higher individualization, and mass customization.
- Enhanced use of freely available data.
- Building new products and services supplemented with Big Data analytics and privacy by design, developing products adapted to European privacy standards . The power and opportunity of big data applications used well, big data analysis can improve economic productivity, drive improved consumer and government services, prevent terrorists, and save lives. Examples include:
- Big data and the rising "Internet of Things" have made it possible to merge the manufacturing and information economies. Jet engines and delivery trucks can now be outfitted with sensors that monitor hundreds of data points and send automatic alerts when maintenance is needed. This makes repairs easy, reducing maintenance costs and increasing safety.
- The Centers for Medicare and Medicaid Services have begun using predictive analytics software to flag likely instances of reimbursement fraud before claims are paid. The Fraud Prevention System helps identify the highest risk health care providers for fraud, waste and abuse in real time, and has already stopped, prevented or identified $115 million in fraudulent payments—saving $3 for every $1 spent in the program's first year.
- During the most violent years of the war in Afghanistan, the Defense Advanced Research Projects Agency (DARPA) deployed teams of data scientists and visualizers to the battlefield. In a program called Nexus 7, these teams embedded directly with military units and used their tools to help commanders solve specific operational challenges. In one area, Nexus 7 engineers fused satellite and surveillance data to visualize how traffic flowed through road networks, making it easier to locate and destroy improvised explosive devices[5].

IV. TOOLS: OPEN SOURCE REVOLUTION

The Big Data is basically related to the open source software revolution. Large companies as Facebook, Yahoo!, Twitter, LinkedIn promote and benefit working on open source

projects. Big Data infrastructure deals with Hadoop, and other related software as:

• **Apache Hadoop**: software for data-intensive distributed applications, based in the MapReduce programming model and a distributed file system called Hadoop Distributed Filesystem (HDFS). Hadoop allows writing applications that rapidly process large amounts of data in parallel on large clusters of compute nodes. A MapReduce job divides the input dataset into independent subsets that are processed by map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to obtain the final result of the job.

• **Apache Hadoop related projects**: Apache Pig, Apache Hive, Apache HBase, Apache ZooKeeper, Apache Cassandra, Cascading, Scribe and many others.

• **Apache S4**: platform for processing continuous data streams. S4 is designed specifically for managing data streams. S4 apps are designed combining streams and processing elements in real time.

• **Storm**: software for streaming data-intensive distributed applications, similar to S4, and developed by Nathan Marz at Twitter.

In Big Data Mining, there are many open source initiatives. The most popular are the following:

• **Apache Mahout**: Scalable machine learning and data mining open source software based mainly in Hadoop. It has implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent pattern mining.

• **R**: open source programming language and software environment designed for statistical computing and visualization. R was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand beginning in 1993 and is used for statistical analysis of very large data sets.

• **MOA**: Stream data mining open source software to perform data mining in real time. It has implementations of classification, regression, clustering and frequent item set mining and frequent graph mining. It started as a project of the Machine Learning group of University of Waikato, New Zealand, famous for the WEKA software. The streams framework [6] provides an environment for creating and running stream processes using simple XML based definitions and is able to use MOA, Android and Storm. SAMOA is a new upcoming software project for distributed stream mining that will combine S4 and Storm with MOA.

• **Vowpal Wabbit**: open source project started at Yahoo! Research and continuing at Microsoft Research to design a fast, scalable, useful learning algorithm. VW is able to learn from terafeature datasets. It can exceed the throughput of any single machine network interface when doing linear learning, via parallel learning.

More specific to Big Graph mining we found the following open source tools:

• **Pegasus**: big graph mining system built on top of MapReduce. It allows finding patterns and anomalies in massive real-world graphs. See the paper by U. Kang and Christos Faloutsos in this issue.

• **GraphLab**: high-level graph-parallel system built without using MapReduce. GraphLab computes over dependent records which are stored as vertices in a large distributed data-graph. Algorithms in GraphLab are expressed as vertex-programs which are executed in parallel on each vertex and can interact with neighboring vertices[7].

V. CONCLUSION

We conclude that proper analysis of big data reveals valuable actionable knowledge which proves to be very useful in decision making in various areas like medical, scientific research, agricultural, organisation etc. Big data analysis give rise opportunities in designing of competitive offer packages for customers, configuring network to provide reliable services but analysis must be accurate and timely for successful decision making[13]. Review of various challenges faced with big data has been outlined in this paper and these challenges must be addresses in order to realize full potential of big data. Also all the challenges outlined are not domain specific, they are common across varieties of domain. In future research must be done to address outlined challenges[8].

VI. REFERENCES

[1] Big Data A New World of Opportunities , NESSI White Paper, December 2012.

[2] Cardinaels, K., Meire, M., & Duval, E. (2005). Automatic metadata generation: the simple indexing interface. Proceedings of the 14th international conference on World Wide Web, May 10-14, 2005, Chiba, Japan (pp. 548 – 556). New York: ACM Press.

[3] F.C.P, Muhtaroglu, Demir S, Obali M, and Girgin C. "Business model canvas perspective on big data applications." Big Data, 2013 IEEE International Conference, Silicon Valley, CA, Oct 6-9, 2013, pp. 32 - 37.

[4] Zhao, Yaxiong , and Jie Wu. "Dache: A data aware caching for big-data applications using the MapReduce framework." INFOCOM, 2013 Proceedings IEEE, Turin, Apr 14-19, 2013, pp. 35 - 39.

[5] Xindong Wu , Xingquan Zhu, Gong-Qing Wu, Wei Ding, “Data Mining with Big Data”, IEEE Computer Society, Volume:26, Issue: 1, P 97 – 107, 2014.

[6] Ericsson White paper 284 23-3211 Uen | August 2013 Big Data Analytics.

[7] Bo Li, Prof. Raj Jain “Survey of Recent Research Progress and Issues in Big Data”, 2013.

[8] Vinti Parmar, Jyoti, Chanderkant, “ Innards of Big data” , International Journal of Engineering Research, Volume:4, Issue 2, P 216-222, 2016.