A Survey on Big Data Analytics in Healthcare

Vijendra Singh
University of Petroleum & Energy Studies

A Survey on Big Data Analytics in Healthcare

Vijendra Singh
Department of Computer Science and Engineering
School of Engineering and TechnologyThe NorthCap University
Gurugram, INDIA
Email ID: vsingh.fet@gmail.com

Madhu Kumari
Department of Computer Science and Engineering
School of Engineering and Technology The NorthCap University
Gurugram, INDIA
Email ID: madhunain@gmail.com

Abstract— **Big data has evolved as a most challenging field of study and research area. It has drawn much attention during the last few years and influence our modern society including business, government, healthcare, and in research in almost every discipline. Due to amplified usage of data intensive technique, it's become extremely crucial to handle colossal amount of data generated with on hand data processing techniques and make optimum decision on time. Big data has a huge potential to improve healthcare. An appropriate analysis of healthcare data will fabricate most suitable predictive results and also greatly decrease the cost and time of treatment. This paper attempts to consolidate the description of big data by integrating various concepts and definitions from the diverse practioners and academics. We locate various factors which make big data different from traditional datasets. We discuss the challenges associated with big data analytics. In addition we highlight the potential of big data in healthcare sector and summarize various tools and technologies developed to handle healthcare data.**

Keywords—Big data; potential ; Challenge; healthcare; analytical tools and technologies.

I. INTRODUCTION

In this era of digitization, gigantic amount of information is generated and captured into verity of different format from various different sources. Data has become a currency of information economy [1]. It has become the subject of interest from corporate to government sector and from learner to the researchers. It has been observed that that big data has penetrated into every area of today's industry and business functions and has become an important factor in production .Big Data has gained much attention during last few years. According to IDC (International Data Corporation) "Big data absolutely has the potential to change the way governments, organizations, and academic institutions conduct business and make discoveries, and its likely to change how everyone lives their day-to-day lives," said Hauser .The amount of data generated due to the rapid digitization of society is very difficult to handle. By 2020, world's data will reach 35 trillion gigabytes [2]. Big data analytics enable organizations to analyze structured, semi-structured and unstructured data which scales to terabytes, petabytes and Exabyte to scale of valuable information and insights[3]. Accumulating, preprocessing, formatting and analysis of such oversized data is a big challenge. Unstructured data is much more complex and difficult to handle. The rate of growth of unstructured data is increasing very rapidly. According to Computer World, unstructured information may account for more than 70% to 80%of all data in organizations. These data, which mostly originate from social media, constitute 80% of the data worldwide and account for 90% of Big Data [4].

According to Industrial Development Corporation (IDC) and EMC Corporation, the amount of data generated in 2020 will be 44 times greater [40 zettabytes (ZB)] than in 2009. This rate of increase is expected to persist at 50% to 60% annually [5].

Right data, at right time and in right hand is assets in any organization. Once scrap can be others assets, only an attempt is required to identify it. Almost every sector realizes the potential of big data analysis and tries to use it to improve their outcomes. One wrong decision in any organization, business, and industry can spoil everything and leads to sever consequences.

This paper is organized into various sections which provide better understanding about the topic. We first try to define the big data following with its various different characteristics. We uncover the fact that big data is not only about its volume but there are a variety of other factors also which contribute to it. This paper provides an overview of big data's big potential in healthcare sector along with some of the work done. We also summarize the various tools and technologies developed to harness the healthcare data.

II. BIG DATA'S BASIC CONCEPT

Big data can be defined in various different ways depending upon its context. It represents great challenge in front of researchers and academics. Due to the rapid advancement of technology, the size of the data set that qualifies for big data will also increase. Some data which seems to be small for some organization may be big for some other.

According to Wikipedia, big data is defined as "an all-inclusive term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications" [6].Big data is different from traditional data sets in many ways like depending upon data source, volume, data type etc. Table 2 shows how big data is different from traditional small data sets.

TABLE I. DIFFERENCE BETWEEN TRADITIONAL DATASETS AND BIG DATA.

S.no.	Property	Traditional Small Data Sets	Big Data
1.	Data Source	Any kind of transactional data.	User generated content. Social media. Spatial and GPS data. Business data. HD video, Audio, images. SMS, MMS etc
2.	Volume	Small, generally in Terabytes.	Very large Sometimes in Petabytes.
3.	Velocity	Batch velocity	Real time velocity.
4.	Data Type	Structured	Structured, semi structured and unstructured.
5.	Availability	Frequency of availability varies according to working hours.	24 ×7
6.	Architecture	Centralized	Distributed
7.	Tools	OLTP, Relational DBMS	NOSQL, Hadoop, stream computing, In- memory etc.

A widely recognized definition is given by IDC: "big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/ or analysis" [7].In general big data is any property that possess a challenge to store and process it with the help of traditional data processing applications. Initially big data is considered to have three characteristics i.e. 3V's of big data. According to Gartner ''Big data is high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight, and decision making insight discovery and process optimization.[8]''

Similarly, Laney (2001) suggested that Volume, Variety, and Velocity (or the Three V's) are the three dimensions of challenges in data management [9].

Tech America Foundation defines big data as follows: "Big data is a term that describes huge volumes of high velocity, complex and variable data that call for sophisticated techniques and technologies to facilitate the capture, storage, distribution, management, and analysis of the information."[10]. As the study grows it is found that three V's are inadequate to illustrate the big data, as a result a new V i.e. veracity is also integrated to big data. Now it has turn out to be 4 V's. A further ''V'': value (sometimes changed to validity or verification), can be added to the 4 V's of IBM. Big data now be characterize by 5V's i.e. namely, volume, variety, veracity, velocity, and Value. The basis of challenges of big data are its ever-increasing volume (amount of data), the velocity (speed of data), and the variety (range of data types and sources). Big data is data which is too big(volume), frequency of arrival is too fast(Velocity), also arrives along with too much noise(veracity) and it is too diverse in nature(verity) that it is difficult to be processed using on hand data management tools or traditional data processing applications.

III. BIG DATA'S BIG CHALLENGES

Major challenges in big data analysis include data inconsistency and incompleteness, scalability, timeliness, and security [6] [11]. What really matters about big data is, what it does? Aside from how we define big data as a technological phenomenon, the wide variety of potential uses for big data analytics raises crucial questions about whether our legal, ethical, and social norms are sufficient to protect privacy and other values in a big data world[12].Some of challenges are:

1. Data preparation: Before analyzing the data it must be prepared suitable for the analyzing tools. Captured data consist of about 80% of unstructured and semistructred data which present a immense challenge in the preparation of data. Captured data also contain noise, inconsistency and incomplete information which require great attention. Various preprocessing techniques are required to apply in order to make it suitable for processing.
2. Efficient algorithms for analysis: Once the data is ready for processing an efficient algorithm must be applied in order to get value from it. Algorithms should be capable of handling large datasets in less time. Big Data analysis algorithm must run in linear or even in sub linear time (O (n)). It must use up to poly algorithmic space.
3. Privacy: Privacy becomes major concern when comes to big data. Some of the data floating over the network are personal to the individual. No rules or standard are set to ensure the privacy of this personal data. It has been found that some security agencies are using this personal data for their benefits. Therefore, some rules standards or policy should be developed to ensure the privacy of personal data.
4. Talent gap: More and more experts are required to handle the big data. Application developer needs to upgrade their knowledge in order to deal with big data tools. But when many expert comes to practical aspect of data modeling, data integration, data analysis are found naive. Thus a gap is created between the experts and their capabilities According to analyst firm McKinsey & Company, "By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as

1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.

IV. BIG DATA IN HEALTHCARE INDUSTRY

Effective analysis of big data provides insights into many complex issues and improves the accuracy on time. Decision making capabilities has improved remarkable, which results ever improving results. A considerable amount of data is generated by healthcare sector in recent years. Due to the digitization of world, healthcare industry also likes to have computerized data. The health care data includes Electronic Health Reports (EHR) of patients data, clinical reports, doctor's prescription, diagnostic reports, medical images, pharmacy information, health insurance related data, data from social medias and medicinal journals [13]. All these data contribute to big data in healthcare sector. An appropriate analysis healthcare data will produce most suitable predictive results and also greatly decrease the cost and time of treatment. According to McKinsey and Company, by the use of big data analytics in healthcare, approximately $300 billion to $450 billion can be saved [14]. Big data in healthcare makes it much smarter than earlier and helps in moving one step forward. It basically focuses on

- Improving efficiency
- Lowering cost
- Reducing death rate
- Improving Personalized medication
- Improving Personalized care
- Preventing diseases, early diagnosis and better treatment

Many researches have been carried out to utilize the correct potential of big data in healthcare. Several recommendation system and prediction system has been developed to benefit the healthcare. Recommendation system is the system that suggests or recommends the action to be taken in order to improve the decision making. For example, a nursing care recommendation system has been proposed by Duan [15] which provide nursing education, clinical decision support and quality control. Hoens [16] proposed a recommendation system, which allows choosing a best physician of their need. Prediction systems are designed, which analyze the healthcare data and predict the future health conditions. "Prevention is better than cure" is a well known quote and big data analysis makes it possible. Predicting the disease gives the opportunity to preventing it. Early diagnosis of illness helps in curing it in an efficient manner and also avoiding the related health issues. IBM predictive analysis is promising example, which aims to detect the patients at risk of heart failure. Virginia health system Carilion clinic in collaboration with IBM, achieve a great success in detecting the patient which are at high risk of heart failure. Natural language processing technology of IBM analyzed the clinical data such as electronic medical record (EMR), discharge notes, doctor's prescription etc. In the pilot phase 85% accuracy was achieved. Early detection allows deciding plan of action, applying preventive measures on time, results in improved outcomes. Analysis of big biological data helps in identifying various undesirable conditions and future implications. Big data opens new opportunities to handle some life threatening disease like cancer. One of the successful examples is the treatment of lung cancer which is particularly deadly. On scanning hundreds of thousands of gene expressionof various different cell type and tissues, Atul Butte and Julien Sage used a algorithm to identify a possible new treatment for small cell lung cancer[17].

Moon short program is also initialized to increase the survival rate of cancer patients. It is a joint effort of various industries and cancer researchers. Another notable example is locating the outbreaks of EBOLA virus in 2014. Monitoring the movement of mobile phones help in predicting the movements of typical population in particular region and thereby deciding the locations to set the treatment centers.

V. BIG DATA TOOL AND TECHNOLOGIES IN HEALTHCARE

As the big data is a exiting and promising phenomenon, every sector now try to adopt it for their benefits. Many companies have geared up to develop and provide big data solutions for the world. Various tools and technologies are developed to handle big data and help other to effectively use it. For example, Hadoop is like a revolution in the era of big data. Hadoop is the most widely used tool to store and mange big data. Mapreduce is another technology to deal with big data. Mapreduce is a big data analytical platform. The collaboration of IT industry and healthcare sector greatly changed the idea of handling healthcare. The place of traditional healthcare system is now taken by the fact based healthcare system. Facts are the meaningful and interesting patterns extracted from the healthcare big data.Health decisions are taken on the basis of these facts which greatly improve the efficiency of the healthcare delivery system. People are now days more aware about their health and keen to know more about the various health illnesses, their causes, and cure and prevention methods. They want to collect, store and manage all their clinical data at one secure place which can be easily accessed when required. One solution is to digitize the data and store it on online repository. These online repositories allow people to store, manage and share their healthcare data. Availability of healthcare data online over online repository will speed up the analysis task and lowers the time factor. Many companies are working to provide such solutions. The main objective of these companies is to provide a solution where people can store their healthcare data in their online repository and efficiently manage retrieve and share their healthcare data.Table 2 lists some of the tool developed by different companies to handle big data.

TABLE II. TOOLS DEVELOPED BY DIFFERENT COMPANY ALONG WITH THEIR FUNCTIONALITY.

S.no.	Provider	Product	Functionality
1.	Google	Hadoop	Store and mange big data.

2.	IBM	DB2	Database capable of handling lager volume of data. Capable of optimizing performance.
		SPSS	Predictive analytical software. Predict the future conditions and helps in making smarter decisions.
3.	Teradata	Aster	Store and mange large volume of data sets. Analysis of datasets. Built-in functions for big data application.
4.	SAP	HANA	In memory database for big data. Support approximately 80 terabytes of data. Support predictive analysis.
5.	Amazon Web Service	DynamoDB	Big Data database.

Understanding human genome is a tricky task but once understand, different health related issues can be solved. Many studies and research projects are initiated to understand the human genome sequencing. Heredity diseases and by birth disorder can be prevented and even can be cure if genomic data is analyzed carefully and properly.

Many solutions have been developed by various different companies. Integrating data mining and healthcare data effectively and applying suitable big data analysis techniques will no doubt impact healthcare delivery system positively in terms of cost and also improves healthcare results via well informed decision making [14]. Liquidity of healthcare data is necessary and also crucial as collecting and sharing the patient health data involve many consideration such as technical, ethical and public policy. Effective integration of data mining and medical information and its subsequent analysis using big data will no doubt impact healthcare delivery costing and improved healthcare results via well informed decision making[18].Table 3 list the various company's which are currently working on big data and using it for improving healthcare.

TABLE III. COMPANY'S WORKING ON BIG DATA

S.no.	Company name	Nature of service
1.	Explorys	It captures the clinical, financial and operational data. Provide them who need it.
2.	Propller health	Respiratory health management that uses remote monitoring.
3.	NextBio	Provide platform for pharmaceutical companies and researchers to search, discover and share valuable information across public and proprietary data.
4.	GNS Healthcare and Aetna	Developed Reverse engineering forward simulation technology that identifies people at risk of certain condition depending upon five information about them.
5.	Common sensing	Diabetes management. Record insulin dosage and times of specific dosage. This data is made available to healthcare professionals to take timely and correct decision.
6.	DNAnexus	Analysis of biological data.
7.	Appistry	It deals with genomic data. It simplifies the analysis of next generation sequencing data with an ease to use solution that is complaint, scalable and flexible to meet the need of organization.

VI. CONCLUSION

The objective of this paper is to enlighten the facts about the big data. Various descriptions of big data are presented by integrating various concept and definitions from different practitioner and experts. We discussed 5V's i.e. volume (size of data), velocity (speed of data generation), variety (data in different format from different sources), veracity (inconsistency and inaccuracy in data) and value (valuable information hidden). Big data is big and growing bigger and bigger by time and presents various challenges in front of researches. Data preparation, privacy, efficient algorithm for analysis, data liquidity and talent gap are some of the challenges which we have discussed in this paper. Healthcare big data touches new horizon of dealing with various diseases, its identification, prevention and cure. Many initiatives are taken to promote the analysis of healthcare big data and many more yet to be started. Convergence of advanced computing and big data technologies developed by IT industry helps in achieving high performance, scalability at low cost. Big data has a great potential to improve healthcare. This is great field which have lot of potential and worth studying it.

REFERENCES

[1] Chen, M., Mao, S. and Liu, Y., 2014. Big data: a survey. Mobile Networks and Applications, 19(2), pp.171-209.

[2] Amant, K.S. and Ulijn, J.M., 2009. Examining the information economy: exploring the overlap between professional communication activities and information-management practices. IEEE Transactions on Professional Communication, 52(3), pp.225-228.

[3] Rao, S., Suma, S.N. and Sunitha, M., 2015, May. Security Solutions for Big Data Analytics in Healthcare. In Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference on (pp. 510-514). IEEE.

[4] Khan, N., Yaqoob, I., Hashem, I.A.T., Inayat, Z., Mahmoud Ali, W.K., Alam, M., Shiraz, M. and Gani, A., 2014. Big data: survey, technologies, opportunities, and challenges. The Scientific World Journal, 2014.

[5] IDC, “Analyze the futere,” 2014, http://www.idc.com/.

[6] Matturdi, B., Zhou, X., Li, S. and Lin, F., 2014. Big Data security and privacy: A review. China Communications, 11(14), pp.135-145.

[7] Gantz, J. and Reinsel, D., 2011. Extracting value from chaos. IDC iview,1142, pp.1-12.

[8] Gartner IT Glossory(n.d). retrived from http://www.gartner.com/it-glossary/big_data/

[9] Laney, D., 2001. 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6, p.70.

[10] TechAmerica Foundation’s Federal Big Data Commission.(2012).Demystifying big ta: A practical guide to transforming the business of Government. Retrieved from http://www.techamerica.org/Docs/fileManager.cfm?f=techamerica-gdatareport-final.pdf

[11] Kouzes, R.T., Anderson, G.A., Elbert, S.T., Gorton, I. and Gracio, D.K., 2009. The Changing Paradigm of Data-Intensive Computing. IEEE Computer, 42(1), pp.26-34.

[12] United States. Executive Office of the President and Podesta, J., 2014. Big data: seizing opportunities, preserving values.

[13] Muni Kumar, N. and Manjula, R., 2014. Role of Big data analytics in rural health care-A step towards svasth bharath.

[14] Nambiar, R., Bhardwaj, R., Sethi, A. and Vargheese, R., 2013, October. A look at challenges and opportunities of big data analytics in healthcare. InBig Data, 2013 IEEE International Conference on (pp. 17-22). IEEE.

[15] Duan, L., Street, W.N. and Xu, E., 2011. Healthcare information systems: data mining methods in the creation of a clinical recommender system.Enterprise Information Systems, 5(2), pp.169-181.

[16] Hoens, T.R., Blanton, M., Steele, A. and Chawla, N.V., 2013. Reliable medical recommendation systems with patient privacy. ACM Transactions on Intelligent Systems and Technology (TIST), 4(4), p.67.

[17] http://scopeblog.standford.edu/2013/09/27/big-data-big-finds-clinical-trail-for-deadly-lung-cancer-launched-by-standford-study/

[18] Sun, J. and Reddy, C.K., 2013. Big Data Analytics for Healthcare Tutorial presentation at the SIAM International Conference on Data Mining. Austin, TX.

[19] G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (references)

[20] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[21] I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[22] K. Elissa, “Title of paper if known,” unpublished.

[23] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[24] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[25] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.