A Study on Big Data in Health Care Industry

Shivani Jain[1], Alankrita Aggarwal[2, #]

[1] *Research Scholar, Department of Computer Science, IGDTUW, New Delhi.*
[2] *Assistant Professor, Department of Computer Science and Engineering*
Panipat Institute of Engineering and Technology, Samalkha (Haryana)
[#]Corresponding Author, Email: alankrita.agg@gmail.com

***Abstract*— With the development in health care industry, more and more healthcare data are being collected from different sources. Extracting, maintain and used further for analysis is a complex task in the health care industry. Health care data is generated enormously every bit of a second analyze that data is difficult by any conventional methods so data big data gain interest among the research communities. Big data is a viable solution to solve these problems. In this paper, authors give an overview of the big data possibility in healthcare domain. It outlines the 5V's of big data in health care and a brief description about how data is generated from different sources and processed though the big data software and how these processes are performed. We have also summarized few of the software tools used for health care domain. At last, discussed the challenges faced to process health care data efficiently.**

***Keywords*— Big Data, Map-Reduce, Hadoop, Healthcare, Clinical data, Software Tools**

1. INTRODUCTION

Health is the major concern in all the human beings. A good heath is essential for a human life. But now a days majorly all person around the globe suffers from various health issues such as hypertension, flu, blood pressure, lung infection, heart diseases, mental disorder etc. A large volume of data is generated from the various methods to record the human health conditions. Such a huge volume of data is generated through these records are so complex and difficult to manage and process through conventional data management tool and techniques. Figure 1 shows the volume of health care data generated around the globe. So, there is a need of high process devices and software tools which can process large data to extract useful and meaningful information. Big data [1] has the potential to manage the large datasets. Every day a zettabyte (1021 gigabytes) of data being generated by healthcare industry in India. Various sources are there, through which health care data is generated such as medical record of a patient, images of infected body part, IOT data, Smartphone health data. Many applications are developed that record the health data of every second of a user. Such a huge data cannot be processed and managed by the conventional methods of software that only record and used for query system. To analyse such records of patient'sBig data software tools are required. It helps health care organization to improve care, save lives and reduce cost and provide quality of health care services [2]. Big data assures to mitigate the change to authenticate the data-driven healthcare, to take decisions for individual patients based on customized analysis and to detect the risk at an early state. It is vitally important for the healthcare organization to be aware of sources of data generated, how to process data to leverage big data effectively so as to improve efficiency and quality of health care delivery. Different software tools, machine learning techniques are used in big data for extraction of meaningful information from the health care domains. In this research paper, authors try to summarize some of techniques used in health care domain for better understating of health records of human.

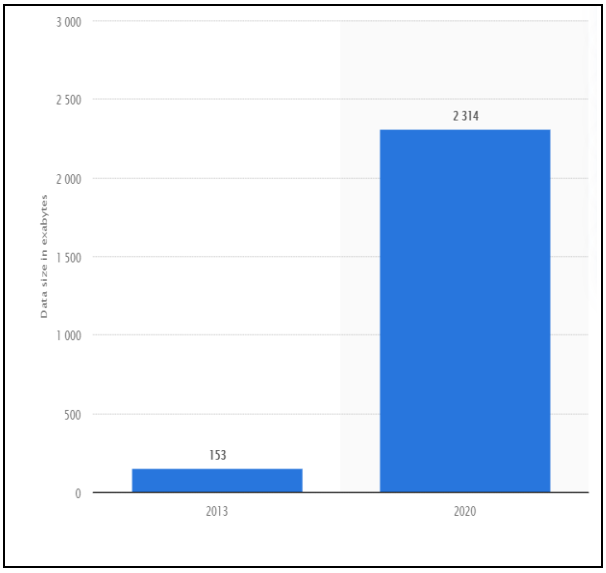

Fig. 1: Health care data generated around the Globe.

This paper is structured as follows: Ssection IIdescribe the big data in healthcare and various source of data generated in the health care domain; Ssection III explains the techniques used for processing big data and the major software used in health care sector; Ssection IV discusses the various challenges faced in healthcare domain, Section V has the conclusion and the future work of the research paper.

2. BIG DATA IN HEALTHCARE

Healthcare produces structured and unstructured data from the different source at a rapid speed such as electronic patient's records contains a different type of information as blood pressure, heart rate, pulse rate, body temperature, past diseases information. U.S healthcare only generated 1150 Exabyte of data in the year 2017. At this rapid pace, health care industry data in U.S. will reach to zeta-byte (1021 gigabytes) in numbers or even yotta-bytes (1024 gigabytes) in numbers [3].A patient hospitalisation generates large amount of data records that includes diagnosis, medical records, digital image processing, laboratory rreports, and hospital bills, based on these data available certain outcomes can be predicated.In U.S healthcare, analysis of big data helps to save roughly $300 billion every year, 66% of that throughdiminishments of proximately 8% in expenditures value of national healthcare[4] as estimated by McKinsey. Healthcare problems manifest all of the V's and it is inevitable that we will use big data techniques to solve them. Features of bigdata are expressed in 3V's [5]. Many other characteristics of big data have been proposed features like 'Value' and 'Veracity'. Volume deals with a enormous amount of data being generated; Variety of data coming from variety of data sources; Velocity is the speed at which data is generated, captured and gathered; Veracity deals with trustworthiness of data and variety of data results in data inconsistency and ambiguity and a value is to identify which data is valuable than transformed and analyzed. Fig.1 illustrates the characteristic of big data.

2.1 5 V's of Big Data in Healthcare-

Volume: Volume comes from the large number of datasets that include Personal Medical Records, Electronic Medical Records (EMR), sensor reading, clinical reports, X-rays reports and laboratory results.

Velocity: Healthcare related data is accumulated in at high speed while monitoring patients' current condition through sensor or tracking web-posts (such as from twitter) for clinical decision support.

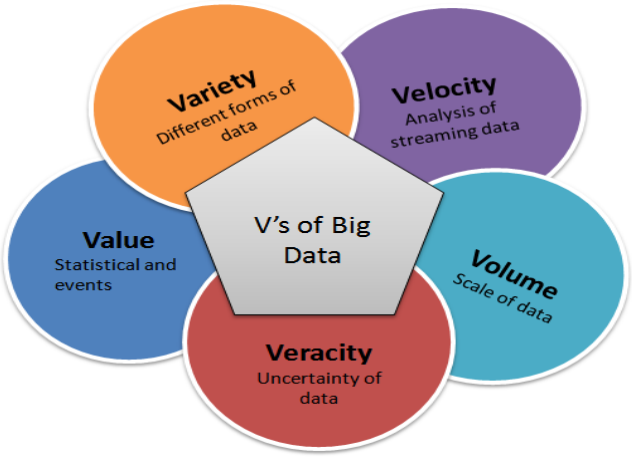

Fig. 2: Characteristics of Big Data

Variety: Large heterogeneous data generated every day from the different source which includes sensors data, clinical data report, prescriptions for detection and diagnosis of disease.

Veracity: Patient may have uncertain and dubious information. Let's take an example-Patients Health Records may subsist of typing error, acronyms, and

incorrect spellings. Data from social networking sites can prompt to inaccurate decision as they are unaware of the context of data and data coming from uncontrolled active environments. Data qualities are major cause of concern for decision making in healthcare.

Value: Large set of valuable and accurate data are generated from sensor data and electronic monitoring of data for the detection of disease.

HealthCare data sources: Health care data can be generated through various sources, which are different in structure and have different complexity to extract the useful pattern from these records. In this section we have shown the various sources from which health care data can be recorded. Fig. 3 shows the sources of big data in health domain.

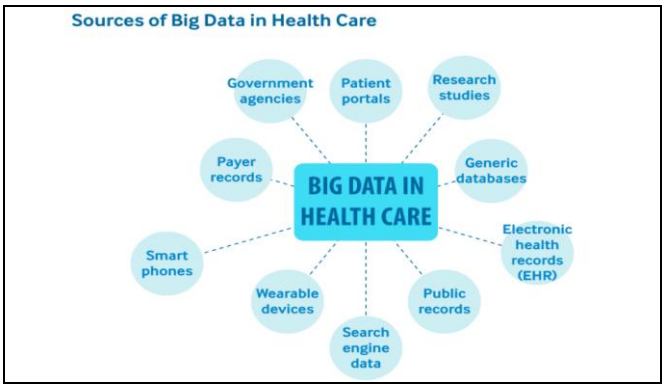

Fig. 3: Various Data sources in Health care

2.2 Electronic Health Records: Traditionally, the health records are written by the doctors and stored by the patients only. No copy is available with the doctors and hospital. Next time if the patient visits the doctor, no previous data is available with the doctor. To overcome this problem now the doctors and hospital are maintaining patient's data using Electronic Health Record of the patients that can used for analysis and accurately assessing the health disorder.Doctors composed notes and prescriptions, Medical Imaging like X-rays, CT scan, MRI, Laboratory report, Pharmacy, and insurance claims are used in electronic form to determine the cost-effective way to diagnosis and treat patients accurately. Electronic Health Records (EHR) [6] also comes under clinical data records. The California based healthcare network called Kaiser Permanente, which has memberships around 9 million, archived around 26.5 and 44 petabytes of data from sources like medical images, EHRs and comments [7]. EHR measures outcomes and encourages patients to take more care of their health. These EHR records further used for predicative analysis now a days. As such data is huge so we need Big data tools to process them and to identify useful pattern from the health records.

Wearable devices and Sensor Data: Now a day's sensors and wearable devices are used to measure the health records of the user. Many health applications are developed to measure the temperature, bp, pulse record, heart beat record of the patients. Through, these application health records are stored in mobile memory. Companies like apple launched apple watch (http://www.appple.com/watch/) which measures heart rate for diabetic patients. Wearable and implantable sensors were developed, which providepersistent monitoring and glucose level which is outstanding to be eating regimen subordinate [8]. Sensors senses, monitors and capture critical data and provide continuous health-related information to find useful pattern for early detection of health issues of patients.

Social media or web data: Social media helps in increase communication between patients, physician, and communities. Twitter tweets, blogs, status posted on platforms, web pages like Facebook, all these include health related pages, websites, pages, smart phone application etc. For analysis of spread of disease or transmission many online social networks data are generally used [9, 10]. Johns Hopkins School of Medicine Researchers analyzes that "google flu trends" data could be utilized to anticipate surges in flu-related crisis. Additionally, Twitter updates were as precise as official reports, Haiti to track the effect of cholera Haiti in the month of January after the earthquake they were moreover two weeks earlier [11]. Twitter data can efficiently track an epidemic across a given population (in real time). In the recent time Covid -19 flu outbreak is analysed using the google trends. These data is used to predict the Ppatients with perpetual diseases, for

example- diabetics, cancer and cardiac problems patients sharing experience with other patients with the similar condition on social media resulting in big data and status updates, text messages and posts on Twitter, Facebook provides healthcare related information.

2.3 Patients Records: Medical images such as X-rays, Brain images, fingerprints, retrieve scan, computed tomography are important sources of data used for early diagnosis of diseases using image segmentation and machine learning techniques[12-16].Healthcare systems in recent scenario utilizes various divergent and interminable tracking devices that use particular discretized or physiological waveform information to give alert direction in case of obvious events.These records are stored in the form of Images, so analysis are done on these records with the help of image process software tools. Different machine learning and deep learning algorithm are applied in these areas to early identification of health patterns and disease.

2.4 Payer Records: Billing records and health care claims are available in semi-structured and unstructured formats. Healthcare insurers are using big data to develop customizing insurance product, managing risk and reducing health care costs. These health records are used for insurance companies for their strategy building.

2.5 Publicly Available data: Health data of the human being are used by the government, private sector to analyse the country growth and to measure the overall health of people of their countries. These data can be used for research, to find the vaccine and cure for the particular diseases. Government agencies record the heath condition of their country people to prepare the health policy. A huge amount of revenues is spent to increase the health care of people.

3. PROCESSING BIG DATA

Big data needs to be processed and store large datasets. MapReduce [17-18] and its open source implementation Hadoop [19], leading computing platforms for big data. Hadoop is built to gather and analyse data in petabytes and Exabyte scale. Hadoop [20] uses Hadoop distributed file systemis extremelyfault tolerantt. HDFS is manufactured to handle large extent data and deployed on hardware with minimal-cost and provide high processing speed. In 2011, an adverse drug event (ADE) detection has been developed based on MapReduce algorithm [21]; for mining useful data for clinical decision-making purpose. MapReduce [22] is the combination of map and reduce function. Job Tracker and Task Tracker deals with the map and reduce tasks [23]. Job Tracker is the master of the system, handles and stores the blocks which are in Task Tracker whereas task tracker is the slave of the system, which provides storage for data and takes the responsibility of the implementation of tasks. The figure below describes the Hadoop Map-Reduce process, whichincludes input data and five phases i.e., split, map, shuffle, reduce and output.

For complex task, combine together an array of Map-Reduce jobs and process them in sequentally. Using directed acyclic graph (DAG) pattern developers were allowed to design miscellaneous, multistep data pipelines in Spark [24- 27], in-memory data sharing among DAGs is also supported by the spark to perform diverse jobs with same data and results in better performance than other big data technologies.

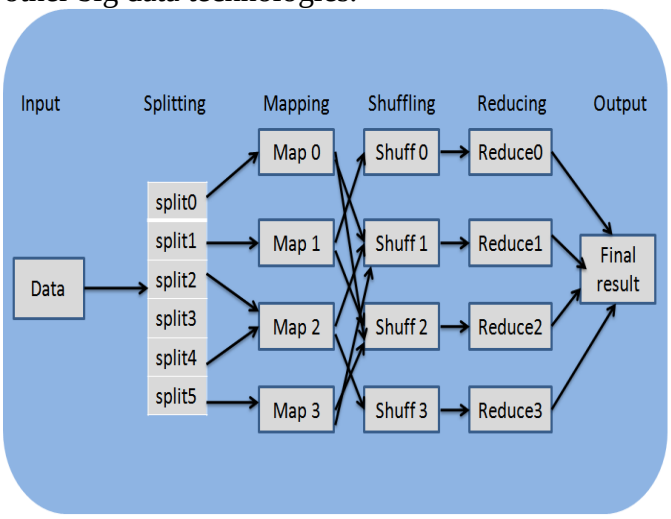

Figure 4: Hadoop Map-Reduce process.

To serve upgraded and extra functionality, Spark runs over existing Hadoop Distributed File System (HDFS) platform. Operating drivers in Spark utilizes a pair of operation i.e. (1) Action and (2) Transformation. Action is same as the reduce and

transformation is same as the map operation. Here we have shown the working of the Hadoop map-reducer used in big data technology, but using the big data technology our main purpose is to handle the data that is generated by the health care sector. The data generated by the health sector is having lot of complexity as its volume is high and highly diverse in nature. EMR health records are maximally written in hard form, first change into electronic readable device. For this scanner are used but scanned documents are difficult to process and read by the machine as every handwritten document is different. Different AI based algorithms are used to fetch the information from the scanned documents. All data is un-structural format and different machine learning techniques are used to read, analyze and predict the information from these records. Maximum data used in health care sector in the Image form. So, to process the thousands of image file is a big issue in health care system. Here we have summarized a few software used in medical domain for storing, analyzing and predicting health care data.

4. CHALLENGES

There are various challenges are faced to mine the information in the field of health care sector

Clinical data are in unstructured and semi-structured format, needs to be processed and mined by the different platform, having different context. Every practitioner has its own opinion to deal with the problem so standardization is difficult for health care data. Healthcare data is rarely standardized, often fragmented with incompatible format. Privacy and Security is a major problem as data is heterogenous and are generated from the diverse environment. Health records having patients' personal records so security and privacy of the data is the main concern for any health care company. Leverage healthcare data deliver quality healthcare services to patients' while

1. preserving it private and secure.
2. Genomic data analysis is a computationally hard job and integration with standard clinical data adds more difficulty and add an extra layer of complexity.

Data structure of health care data is highly complex and difficult to handle by any machine learning algorithm, DNA one sequence conations millions of possible combinations, to identify useful pattern is a complex task.

Table 1. Software used to store and analyze medical health records of the patients.

Name of Software	Company Name	Software Used	Tools for Analysis
ClearHealth	Clear health Inc	PHP, Java Script	Used EMR, Scheduling and billing
Open Dental	Open Dental Software	PHP, MySQL	Used for Dental Records
OpenHospital	Informatici Senza Frontiere	Java	Used for billing, scheduling and maintaining records
CamBA	GNU GPL	Java, Python	Collecting neuroimaging
GIMIAS	CISTIB University of Sheffield	C++	Biomedical Imaging processing & simulation.
MITK Medical Imaging Interaction Toolkit	NVIDIA	AI, Transfer learning, CNN	Used to create interactive medical image for better understanding.
Caisis	GNU GPL	C#, XML, HTML, Java Script	Web based information system for collecting, storing and analyzing the cancer patient data
cTAKES "clinical Text Analysis Knowledge Extraction Software"	Mayo Clinic	Python	It uses NLP processing system to identify and extract different entities from un-structural text data.
LabKey Server	LabKey	Java	Used for analyzing, sharing, integrating and storing medical research data, web-based application for querying and reporting across different platform
Glucosio	GNUGPL	IOS, IOT	This is mobile application used to measure the glucose level of diabetic patients.

1.
2.

Smart phone applications generate billions of data every second, storing and processing all the data without human intervention is a difficult task.

.

5. CONCLUSION

Big data has great potential to change the way healthcare provider uses data sources and technologies to get vision from the clinical and other health related repositories. Big data mining software's improved efficiency of health services and providing quality services at low cost. Machine learning and deep learning methods have gained much interest among the research communities to solve health care data. Health care data is complex in nature and are generated from the heterogenous sources so standardization is not possible. In this paper, we have summarized the sources of health-related data and the processes involved to analyse the data for information extraction. Authors have also discussed the various software used in health care domain. Although various studies are available is the field however this research paper focus on health sector, big data and shows different challenges in this area that leads a better understanding of healthcare services at an early stage. We also surveyed major software used in healthcare sector.

In future, authors will work on various methods of machine learning and deep learning applied in Health domain for delivering quality health services at minimal cost.

REFERENCES

[1] Zikopoulos PC, Eaton C, DeRoos D, Deutsch T and Lapis G. "Understanding Bigdata," New York et al: McGraw-Hill, 2012.

[2] Knowledgent: Big Data and Healthcare Payers; 2013. http:// knowledgent.com / media page /Insights/whitepaper/482.

[3] Cerrato, P. "Is Population Health Management the Latest Health IT Fad?", 2012, Information Week: http://www.info4mationweek.com/ health care/clinical-systems/is-population-health-management-latest-h/240004578.

[4] Manyika J, Chui M, Brown B, Buhin J, Dobbs R, Roxburgh C, Byers AH: Big Data: The Next Frontier for Innovation, Competition, and Productivity. USA: McKinsey Global Institute; 2011.

[5] D. Laney, "3D Data Management: Controlling data volum e, velocity and variety," ed: Meta Group, 2001

[6] G. N. Forrest, T. C. Van Schooneveld, R. Kullar, L. T. Schulz, P. Duong, and M. Postelnick, "Use of electronic health records and clinical decision support systems for antimicrobial stewardship," Clin. Infectious Dis., 59,122–133, 2014

[7] IHTT: Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry, 2013. http://ihealthtran.com/wordpress/2013/03/iht%C2%B2-releases-big-data-research-reportdownload-today/

[8] M. Breton, A. Farret, D. Bruttomesso, S. Anderson, L. Magni, S. Patek,C. D. Man, J. Place, S. Demartini, S. Del Favero, C. Toffanin, C. HughesKarvetski, E. Dassau, H. Zisser, F. J. Doyle, G. De Nicolao, A. Avogaro,C. Cobelli, E. Renard, and B. Kovatchev, "Fully integrated artificial pancreas in type 1 diabetes: Modular closed-loop glucose control maintains near normoglycemia," Diabetes, 61, 2230–2237, 2012

[9] A. Sadilek, H. Kautz, V. Silenzio, "Predicting disease transmission from geotagged micro-blog data", Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.

[10] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters",6th USENIX Symp. Oper. Syst. Des. Implementation, 137-150, 2004.

[11] A. Sadilek, H. Kautz, V. Silenzio, Modelling spread of disease from social interactions, in: Sixth AAAI International Conference on Weblogs and Social Media (ICWSM), 2012, http://www. cs.rochester.edu /~kautz/ papers /Sadilek-Kautz-Silenzio_Modeling-Spread-of-Disease-from-Social Interactions_ICWSM-12.pdf.

[12] TechAmerica Foundation, "Demystifying Big Data: A Practical Guide to Transforming the Business of Government". Washington, D.C.: TechAmerica Foundation,2012

[13] Mittal, A., Kumar, D., Mittal, M., Saba, T., Abunadi, I., Rehman, A., & Roy, S. "Detecting Pneumonia Using Convolutions and Dynamic Capsule Routing for Chest X-ray Images". Sensors, 20(4), 1068.

[14] Mittal, M., Kaur, I., Pandey, S. C., Verma, A., & Goyal, L. M. "Opinion Mining for the Tweets in Healthcare Sector using Fuzzy Association Rule". EAI Endorsed Transactions on Pervasive Health and Technology, 4(16).

[15] Kaur, B., Sharma, M., Mittal, M., Verma, A., Goyal, L. M., & Hemanth, D. J. "An improved salient object detection algorithm combining background and foreground connectivity for brain image analysis". Computers & Electrical Engineering, 71, 692-703.

[16] Chhetri, B., Goyal, L. M., Mittal, M., Gurung S. "Consumption of Licit and Illicit Substances leading to Mental Illness: A Prevalence Study". EAI Endorsed Transactions on Pervasive Health and Technology, 6(21).

[17] Mittal, M., Arora, M., Pandey, T., & Goyal, L. M. "Image Segmentation Using Deep Learning Techniques in Medical Images. In Advancement of Machine Intelligence in Interactive Medical Image Analysis" 41-63, Springer, Singapore.

[18] Mittal, M., Singh, H., Paliwal, K. K., & Goyal, L. M., "Efficient random data accessing in MapReduce", 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), IEEE, 552-556.

[19] Hadoop. (2014) [Online]. Available: http://hadoop.apache.org/.

[20] XieJiong, Yin Shu, RuanXiaojun, Ding Zhiyang, TianYun, "Improving Mapreduce performance through data placements in heterogeneous Hadoop cluster", 2010.

[21] Wang W, Haerian K, Salmasian H, Harpaz R, Chase H, Friedman C: A drug-adverse event extraction algorithm to support pharmacovigilance knowledge mining from PubMed citations. In AMIA Annual Symposium Proceedings: 2011.Bethesda, Maryland – USA: American Medical Informatics Association; 2011:1464.

[22] E. A. Mohammed, B. H. Far, and C. Naugler, "Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends," Bigdata Mining,7(22), 1–23, 2014.

[23] Huang Lu, Hu Ting-ting, Chem Hai-shan," Research on Hadoop cloud computing model and its applications", 2012 Third International Conference on Networking and Distributed Computing,59-63.

[24] Singh, A., Mittal, M., & Kapoor, N, Data processing framework using Apache and Spark technologies in big data. In Big Data Processing Using Spark in Cloud, 107-122, Springer, Singapore.

[25] Mittal, M., Balas, V. E., Goyal, L. M., & Kumar, R. (Eds.) "Big data processing using spark in cloud", Springer.

[26] Mittal, M., Balas, V. E., & Hemanth, D. J. (Eds.), "Data Intensive Computing Applications for Big Data" 29, IOS Press.

[27] "Big Data Processing with Apache Spark – Part 1: Introduction", 2015.
https://ww.infoq.com/articles/apache-spark-introduction.