IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10729

BIG DATA ANALYTICS FOR HEALTHCARE INDUSTRY: IMPACT AND APPLICATIONS

Ms. Sujata Ghonmode[1], Prof. Hirendra Hajare[2]

MTech Student[1], Assistant Professor[2]

Department of CSE, Ballarpur Institute of Technology (BIT), Ballarpur

Abstract: In recent years, huge amounts of structured, unstructured, and semi-structured data have been generated by various institutions around the world and, collectively, this heterogeneous data is referred to as big data. The health industry sector has been confronted by the need to manage the big data being produced by various sources, which are well known for producing high volumes of heterogeneous data. Various big-data analytics tools and techniques have been developed for handling these massive amounts of data, in the healthcare sector. In this paper, we discuss the impact of big data in healthcare, and various tools available in the Hadoop ecosystem for handling it. We also explore the conceptual architecture of big data analytics for healthcare which involves the data gathering history of different branches, the genome database, electronic health records, text/imagery, and clinical decisions support system.

Key words: big data; healthcare; Hadoop; MapReduce

1. INTRODUCTION

Every day, data is generated by a range of different applications, devices, and geographical research activities for the purposes of weather forecasting, weather prediction, disaster evaluation, crime detection, and the heath industry, to name a few. In current scenarios, big data is associated with core technologies and various enterprises including Google, Facebook, and IBM, which extract valuable information from the huge volumes of data collected [1–3]. An era of open information in healthcare is now under way. Big data is being generated rapidly in every field including healthcare, with respect to patient care, compliance, and various regulatory requirements. As the global population continues to increase along with the human lifespan, treatment delivery models are evolving quickly, and some of the decisions underlying these fast changes must be based on x data [4].

BIG DATA ANALYTICS IN HEALTH INFORMATICS

The main difference between traditional health analysis and big-data health analytics is the execution of computer programming. In the traditional system, the healthcare industry depended on other industries for big data analysis. Many healthcare shareholder's trust information technology because of its meaningful outcomes—their operating systems are functional and they can process the data into standardized forms.

Today, the healthcare industry is faced with the
challenge of handling rapidly developing big healthcare data. The field of big data analytics is growing and has the potential to provide useful insights for the healthcare system. As noted above, most of the massive amounts of data generated by this system is saved in hard copies, which must then be digitized [7].

A number of conceptual approaches can be employed to recognize irregularities in vast amounts of data from different datasets. The frameworks available for the analysis of healthcare data are as follows:

Predictive Analytics in Healthcare: For the past
two years, predictive analysis has been recognized
as one of the major business intelligence approaches, but its real-world applications extend far beyond the business context. Big data analytics includes various methods, including text analytics and multimedia analytics [14]. However, one of the most crucial categories is predictive analytics which includes statistical methods like data mining and machine learning that examine current and historical facts to predict the future. Predictive methods which are being used today in

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10729

the hospital context to determine if patient may be at risk for readmission [15]. This data can help doctors to make important patient care decisions.

Machine Learning in Healthcare: The concept of machine learning is very similar to that of data mining [4], both of which scan data to identify patterns. Rather than extracting data based on human understanding, as in data mining applications, machine learning uses that data to improve the program's understanding. Machine learning identifies data patterns and then alters the program function accordingly [16].

Electronic Health Records: EHR represents the most widespread health application of big data in healthcare. Each patient has his/her own medical records, with details that include their medical history, allergies diagnosis, symptoms, and lab test
results. Patient records are shared in both public and
private Sectors with healthcare providers via a secure information system. These files are modifiable, in that doctors can make changes over time and add3. new medical test results, without the need for paper work or duplication of data.

FOUR Vs OF BIG DATA IN HEALTHCARE

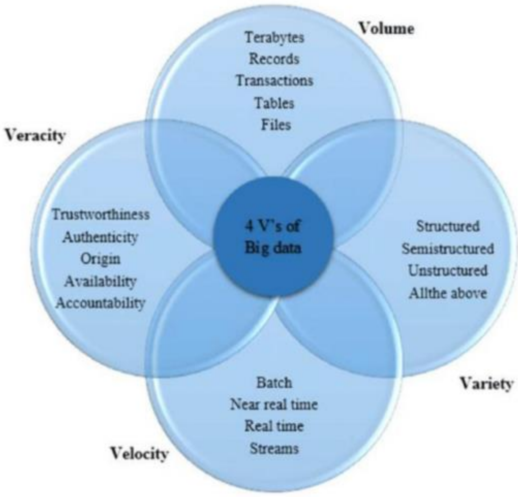

Fig. 1 Big data's four Vs in healthcare

Four Vs of Big Data in Healthcare Four primary attributes (shown in Fig. 1) that are associated with big data: volume, velocity, variety, and veracity.

Volume: Big data is a term to referring to huge volumes of collected data. There is no fixed threshold for the volume of this data. Typically, the term is used with respect to massive-scale data which must be managed, stored, and analyzed using traditional databases and data processing architecture [14]. The volume of data generated by modern IT and the healthcare system has been growing and is driven by the reduced costs of data storage and processing architectures and the need to extract valuable insights from data to improve business processes, efficiencies, and services to consumers [4].
Velocity: Velocity, which represents primary reason for the exponential growth of data, refers to how fast data is collected [14]. Healthcare systems are generating data at increasingly higher speeds. In the volume and variety of the structured or unstructured data collected, the velocity of the generation of this data after processing requires a decision based on its output.

Variety: Variety refers to the form of the data, i.e., unstructured or structured, text, medical imagery, audio, video, and sensor data. Structured data information includes clinical data (patient record data), which must simply be collected, stored, and processed by a particular device. Structured data comprises just 5% to 10% of healthcare data. Unstructured

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10729

or semi-structured data includes e-mails, photos, videos, audios, and other health related data such as hospital medical reports, physician's notes, paper prescriptions, and radiograph films [13].

Veracity: The veracity of data is the degree of assurance that the meaning of data is consistent. Different data sources vary in their levels of data credibility and reliability [9]. The outcomes of bigdata analytics must be credible and error-free, but in healthcare, unsupervised machine learning algorithms make decisions that are used by automated machines based on data that may be worthless or misleading [4]. Healthcare analytics are tasked with extracting useful insights from this data to treat patients and make the best possible decisions

IMPACT OF BIG DATA ON THE HEALTHCARE SYSTEM

The potential of big data is that it could revolutionize outcomes regarding the most suitable or accurate patient diagnosis and the accuracy information used in the health informatics system[15]. As such, the investigation of huge amounts of information will have a powerful effect on medicinal services framework in five respects, or "pathways" (shown in Fig. 2). Improving outcomes for patients with respect to these pathways, as described below, will be the focus of the healthcare system and will directly impact the patient.

Right Living: Right living refers to the patient living a better and healthier life [15]. By right living, patients could manage themselves by making the best decisions for themselves, based on the utilization of information mining better choices and enhancing their wellbeing. By choosing the right path for their daily health, regarding their diet, preventive care, exercise, and other activities of daily living, patients can play an active role in realizing a healthy life [16].

Right Care: This pathway ensures that patients receive the most appropriate treatment available and that all providers obtain the same data and has the same objectives to avoid redundancy of planning and effort[17]. This aspect has become more viable in the era of big data.

Right Provider: Healthcare providers in this pathway can obtain an overall view of their patients by combining data from various sources such as medical equipment, public health statistics, and socioeconomic data [15]. The accessibility of this information enables human service providers to conduct targeted investigations and develop the skills and abilities to identify and provide better treatment options to patients [18].

Right Innovation: This pathway recognizes that new disease conditions, new treatments, and new medical will continue to evolve [15]. Likewise, advancements in the provision of patient services, for research and development efforts, will enable new ways to promote wellbeing and patient health via national social insurance system [17]. The availability of early trial data is important for stakeholders. This data can be used to explore high-potential targets and identify techniques for improving traditional clinical treatment methods.

Right Value: To improve the quality and value of health-related services, providers must pay careful and ongoing attention to their patients. Patients must obtain the most beneficial results identified by their social insurance system [18]. Measures that could be taken to ensure the intelligent use of data includes, for example, identifying and destroying data misrepresentation, manipulations, and waste, and improving resources [19].

HADOOP-BASED APPLICATIONS FOR HEALTH INDUSTRY

In light of the fact that healthcare data exists primarily in printed form, there is a need for the active digitization of print form data. The majority of this data is also unstructured, so it is a major challenge for this industry to extract meaningful information regarding patient care, clinical operations, and research. The collection of software utilities known as the Hadoop ecosystem can help the healthcare sector to manage this vast amount of data. The various applications of the Hadoop ecosystem in the healthcare sector are as follows:

Treatment of Cancer and Genomics: We know that human DNA contains three billion base pairs. To fight cancer, it is vital that large amounts of data are efficiently organized. The patterns of cancer mutations and their reactions vary based on individual genetics, which explains the non-curability of some cancer.

Oncologists have determined that in recognizing the patterns of cancer, it is important to provide specific treatment for specific cancers, based on the patient's genetic makeup. The Hapdoop technology MapReduce facilitates the mapping of three billion DNA base pairs to determine the appropriate cancer treatment for each particular patient. Arizona State University is working on project to develop a healthcare model that takes individual genomic data and selects a treatment based on identification of the patient's cancer gene. This model provides basis for treatment through big data analysis to improve the chances of saving patients' lives.

Monitoring of Patient Vitals: Hospital staff

throughout the world connect their work output using big-data technology. Various hospitals around the globe use Hadoop-based components in the Hadoop Distributed File System (HDFS), including the Impala, HBase, Hive, Spark, and Flume frameworks, to convert the huge amount of unstructured data generated by sensors that take patient vital signs, heartbeats per minute, blood pressure, blood sugar level, and respiratory rate. Without Hadoop, these healthcare staff could not analyse this unstructured data being generated by patient healthcare systems. In Atlanta, Georgia, there are 6200 Intensive Care Units (ICUs) for pediatric healthcare, where children can stay for more than one month depending on their problem. These ICUs are equipped with a sensor technology that tracks the child's health status with respect to heartbeat, blood pressure, and other vital signs. If any problem occurs, an alert is automatically generated to medical staff to ensure the child's safety.

Hospital Network: Several hospitals use the Hadoop ecosystem's NoSQL database to collect and manage their huge amounts of real-time data from diverse sources related to patient care, finances, and a payroll, which helps them identify high-risk patients while also reducing day-to-day expenditures.

Healthcare Intelligence: Hadoop technology also supports the healthcare intelligence applications used by hospitals and insurance companies. Hadoop ecosystem's Pig, Hive, and MapReduce technologies process large datasets related to medicines, diseases, symptoms, opinions, geographic regions, and other factors to extract meaningful information (e.g., desired age) for insurance companies.

Prevention and Detection of Frauds: In the early faces of big data analytics, health-based insurance groups utilize multiple paths to identify

fraud activity and establish methods to prevent medical fraud. With Hadoop, companies use applications based on a prediction model to identify those committing fraud via data regarding their previous health claims, voice recordings, wages, and demographics. Hadoop's

NoSQL database is also helpful in preventing fraud related to medical claims at an early stage by the use of real-time Hadoop based health applications, authentic medical claim bills, weather forecasting data, voice data recordings, and other data sources.

BIG DATA ANALYTICS ARCHITECTURE FOR HEALTH INFORMATICS

Currently, the main focus in big-data analytics is to gain an in-depth insight and understanding of big data rather than to collect it[20]. Data analytics involves the development and application of algorithms for analyzing various complex data sets to extract meaningful knowledge, patterns, and information. In recent years, researchers have begun to consider the appropriate architectural framework for healthcare systems that utilize big-data analytics, one of which uses a four-layer architecture that comprises a transformation layer, data-source layer, big data platform layer, and analytical layer [14]. In this layered system, data originates from different sources and has various formats and storage systems. Each layer has a specific data-processing functionality for performing specific tasks on the HDFS, using the MapReduce processing model. The other layers perform other tasks, i.e., report generation, query passing, data mining processing, and online analytical processing.

The main requirement in big-data analytical processing is to bundle the data at high speed to minimize the bundling time. The next priority in big-data analytical processing is to efficiently update and transform queries at a constant time [21]. The third requirement in the big-data analytical processing is to utilize and efficiently manage the storage area space. The last specification of big-data analytics is to efficiently become familiar with the rapidly progressing workload notations. Big-data analytics frameworks differ from traditional healthcare processing systems with respect to how they process big data [22]. In the current health care system, data is processed using traditional tools installed in a single stand-alone system like a desktop computer. In contrast, big data is processed by clustering and scans multiple nodes of clusters in the network [23]. This processing is based on the concept of parallelism to handle large medical data sets [24]. Freely available frameworks, such as Hadoop, MapReduce, Pig, Sqoop, Hive, and HBase Avro, all have ability to process the health-related data sets for healthcare systems.

Big-data technologies broadly refer to scientific innovations that mimic those used for large datasets [25]. In the first component is the requirement for big data sources for processing. In the second component clusters with a centralized big-data processing infrastructure are at the peak of high performance [24]. It has been observed that the tools mainly available for big-data analytics processing provide data security, scalability, and manageability with the help of the MapReduce paradigm. In the third component, big data analytics applications have a storage domain to integrate accessed

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10729

databases that use different applications [26]. In the fourth component, are the most popular big-data analytics applications in healthcare systems, which include reports, Online
Analytical Processing (OLAP), queries, and data
mining.
As shown in Fig. 3, healthcare data come from a
range of sources including EHRs, genome databases, genome data files, text and imagery (unstructured data sources), clinical decision support systems, government related sources, medical test labs and pharmacies, and health insurance companies. These data are frequently available in different scheme tables, and are in ASCII/text and stored at various locations.

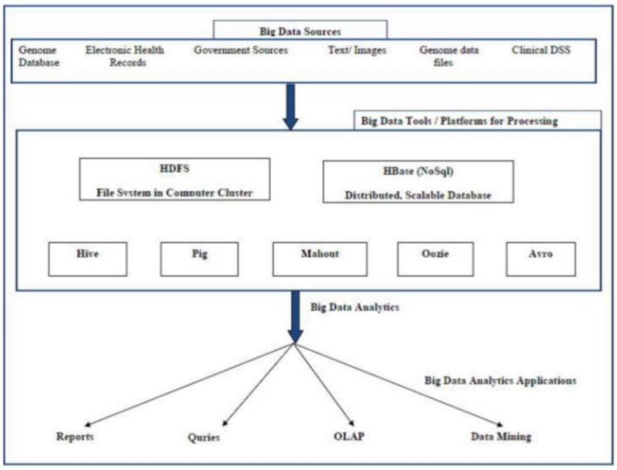

Fig. 2 Conceptual architecture of big data analytics for health informatics.

CONCLUSION

In this paper, we have provided an in-depth description and a brief overview of big data in general and in healthcare system, which plays a significant role in healthcare informatics and greatly influences the healthcare system and the big data four Vs in healthcare. We also proposed the use of a conceptual architecture for solving healthcare problems in big data using Hadoop-based terminologies, which involves the utilization of the big data, generated by different levels of medical data and the development of methods for
analysing this data and to obtain answers to medical questions. The combination of big data and healthcare analytics can lead to treatments that are effective for specific patients by providing the ability to prescribe appropriate medications for each individual, rather than those that work for most people. Big data may be viewed as big systems, which present huge challenges. Therefore, a great deal of research in this field will be required to solve the issues faced by the healthcare system.

2. REFERENCES

[1] A. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods and analytics, International Journal of Information Management, vol. 35, no. 2, pp. 137–144, 2015.

[2] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "Big Data", Hadoop and cloud computing in genomics, Journal of Biomedical Informatics, vol. 46, no. 5, pp. 774–781, 2013.

[3] C. L. P. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences, vol. 275, pp. 314–347, 2014.

[4] M. Herland, T. M. Khoshgoftaar, and R.Wald, A review of data mining using big data in health informatics, Journal of Big Data, vol. 1, no. 1, p. 2, 2014.

[5] D. H. Shin and M. J. Choi, Ecological views of big data: Perspective and issues, Telematics and Informatics, vol. 32, no. 2, pp. 311–320, 2015.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10729

[6] B. Saraladevi, N. Pazhaniraja, P. V. Paul, M. S. Basha, and P. Dhavachelvan, Big data and Hadoop-A study in security perspective, Procedia Computer Science, vol. 50, pp. 596– 601, 2015.

[7] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, Data mining with big data, IEEE transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97–107, 2014.

[8] S. Sharma and V. Mangat, Technology and trends to handle big data: Survey, in Proc. 5th International Conference on Advanced Computing & Communication Technologies, 2015, pp. 266–271.

[9] R. Mehmood and G. Graham, Big data logistics: A healthcare transport capacity sharing model, Procedia Computer Science, vol. 64, pp. 1107–1114, 2015.

[10] D. P. Augustine, Leveraging big data analytics and Hadoop in developing India healthcare services, International Journal of Computer Applications, vol. 89, no. 16, pp. 44– 50, 2014.

[11] J. A. Patel and P. Sharma, Big data for better health planning, in Proc. International Conference on Advances in Engineering and Technology Research, 2014, pp. 1–5.

[12] A. E. Youssef, A framework for secure healthcare systems based on big data analytics in mobile cloud computing environments, International Journal of Ambient Systems and Applications, vol. 2, no. 2, pp. 1–11, 2014.

[13] MAPR, Healthcare and life science use cases, https:// mapr.com/solutions/industry/healthcare-and-lifescienceuse- cases/, 2018.

[14] W. Raghupathi and V. Raghupathi, Big data analytics in healthcare: Promise and potential, Health Information Science and Systems, vol. 2, no. 1, p. 3, 2014.

[15] J. Sun and C. K. Reddy, Big data analytics for healthcare, in Proc. 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 1525–1525.

[16] C. Mike, W. Hoover, T. Strome, and S. Kanwal. Transforming health care through big data strategies for leveraging big data in the health care industry, http://ihealthtran.com/iHT2 BigData 2013.pdf, 2013.

[17] J. Anuradha, A brief introduction on big data 5Vs characteristics and Hadoop technology, Procedia Computer Science, vol. 48, pp. 319–324, 2015.

[18] M. Viceconti, P. J. Hunter, and R. D. Hose, Big data, big knowledge: Big data for personalized healthcare, IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 4, pp. 1209–1215, 2015.

[19] Y. Sun, H. Song, A. J. Jara, and R. Bie, Internet of things and big data analytics for smart and nnected communities, IEEE Access, vol. 4, pp. 766–773, 2016.

[20] A. Jain and V. Bhatnagar, Crime data analysis using Pig with Hadoop, Procedia Computer Science, vol. 78, pp. 571–578, 2016.

[21] T. Jach, E. Magiera, and W. Froelich, Application of Hadoop to store and process big data gathered from an urban water distribution system, Procedia Engineering, vol. 119, pp. 1375–1380, 2015.

[22] C. Uzunkaya, T. Ensari, and Y. Kavurucu, Hadoop ecosystem and its analysis on tweets, Procedia-Social and Behavioral Sciences, vol. 195, pp. 1890–1897, 2015.

[23] S. G. Manikandan and S. Ravi, Big data analysis using Apache Hadoop, in Proc. International Conference on IT Convergence and Security, 2014, pp. 1–4.

[24] V. Ubarhande, A. M. Popescu, and H. Gonzalez- Velez, Novel data-distribution technique for Hadoop in heterogeneous cloud environment, in Proc. 9th International Conference on Complex, Intelligent, and Software Intensive Systems, 2015, pp. 217–224.

[25] S. Maitrey and C. K. Jha, handling big data efficiently by using map reduce technique, in Proc. International Conference on Computational Intelligence & Communication Technology, 2015, pp. 703–708.

[26] Cloudera, Whole genome research drives healthcare to Hadoop, https://www.cloudera.com/content/dam/www/ marketing/resources/solution-briefs/whole-genomeresearch-inhealthcare.pdf.landing.html., 2018.

[27] Apache Hadoop, http://hadoop.apache.org/, 2018.

