International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Emerging Strategies to Big Data Analytics in Healthcare

Tanmayee Tushar Parbat[1], Rohan Benhal[2], Honey Jain[1], Dr. Vinayak Musale[3]
[1]B.E IT, Dr. Vishwanath Karad MIT World Peace University, Pune, Maharashtra, India
[2]BBA IT, Dr. Vishwanath Karad MIT World Peace University, Pune, Maharashtra, India
[3]Dr. Vishwanath Karad MIT World Peace University, Pune, Maharashtra, India

Article Info
Volume 7, Issue 6
Page Number: 254-260

Publication Issue :
November-December-2021

Article History
Accepted : 05 Dec 2021
Published : 15 Dec 2021

ABSTRACT

Big data is gigantic measures of data that can do some incredible things. It has gotten a subject specifically compelling for as long as two decades in view of a high potential that is covered up in it. Different open and private part ventures create, store, and break down huge information to improve the administrations they give. In the social insurance industry, various hotspots for huge information incorporate emergency clinic records, clinical records of patients, aftereffects of clinical assessments, and gadgets that are a piece of the web of things. Biomedical examination additionally creates a critical bit of enormous information pertinent to open medicinal services. This information requires legitimate administration and examination to determine important data. Something else, looking for an answer by breaking down large information rapidly gets tantamount to finding a needle in the pile. There are different difficulties related with each progression of dealing with huge information which must be outperformed by utilizing very good quality registering answers for huge information investigation. That is the reason, to give significant answers for improving general wellbeing, social insurance suppliers are required to be completely outfitted with proper framework to produce and examine huge information methodically. Effective administration, examination, and understanding of large information can change the game by opening new roads for present day human services. That is exactly why different ventures, including the human services industry, are finding a way to change over this potential into better administrations and budgetary focal points. With a protected mix of biomedical and social insurance information, present day human services associations can upset the clinical treatments and customized medication.

Keywords :- Healthcare, EHR, Omics, Hadoop, Data integration, Digital Imaging

I. INTRODUCTION

To obtain the best services and care for the patients, healthcare organizations in many countries have proposed various models of healthcare information systems. These models for personalized, predictive, participatory and preventive medicine are based on using of electronic health records (EHRs) and huge amounts of complex biomedical data and high-quality – omics data [1].

Contemporarily genomics and postgenomics technologies produce huge amounts of raw data about complex biochemical and regulatory processes in the living organisms [2]. These -omics data are heterogeneous, and very often they are stored in different data formats. Similar to these - omics data, the EHRs data are also in heterogeneous formats. The EHRs data can be structured, semi-structured or unstructured; discrete or continuous.

Big data in healthcare and medicine refers to these various large and complex data, which they are difficult to analyse and manage with traditional software or hardware [3], [4]. Big data analytics covers integration of heterogeneous data, data quality control, analysis, modeling, interpretation and validation [5]. Application of big data analytics provides comprehensive knowledge discovering from the available huge amount of data.

Particularly, big data analytics in medicine and healthcare enables analysis of the large datasets from thousands of patients, identifying clusters and correlation between datasets, as well as developing predictive models using data mining techniques [2]. Big data analytics in medicine and healthcare integrates analysis of several scientific areas such as bioinformatics, medical imaging, sensor informatics, medical informatics and health informatics. A survey of big data cases in medical and healthcare institutions/organizations is given in [6].

The new knowledge discovered by big data analytics techniques should provide comprehensive benefits to the patients, clinicians and health policy makers [7].

The remainder of the paper is organized as follows. Related work is described in the second section. Section 3 describes characteristics of big data, while big data analytics is depicted in the subsequent section. The next section explains some challenging issues about big data analytics techniques, while big data privacy and security are described in Section 6. Last section concludes this paper with discussion and further works.

II. RELATED WORK

The rapid development of the emerging information technologies, experimental technologies and methods, cloud computing, the Internet of Things, social networks supplies the amounts of generated data that is growing tremendously in numerous research fields [8].

On this point, contemporarily genomics and postgenomics technologies produce huge amounts of raw data about complex biochemical and regulatory processes in the living organisms [2]. These high throughput – omics data provide comprehensive insight towards different kinds of molecular profiles, changes and interactions, such as knowledge allied to the genome, epigenome, transcriptome, proteome, metabolome, interactome, pharmacogenome, diseasome, etc. [9]. These – omics data are heterogeneous and very often stored in different data formats. Similar to these – omics data, the EHRs data are also stored in heterogeneous formats. The EHRs data, which can be structured, semi-structured or unstructured; discrete or continuous, contain personal patients' data, clinical notes, diagnoses, administrative data, charts, tables, prescriptions, procedures, lab tests, medical images, magnetic resonance imaging (MRI), ultrasound, computer tomography (CT) data. Some of these data are acquired from wearable sensors or capture from medical monitoring devices,

with different collection frequency [5] that makes these data to have complex features and high dimensions [10]. Dealing with noisiness and incompleteness of EHRs are still challenging task and these shortcomings should be consider while applying data mining techniques [11].

These growing amounts of various – omics data need to be collect, clean, store, transform, transfer, visualize and deliver in a suitable manner to be represented to the clinicians [12]. The processing of these big data in medicine and healthcare can be accelerating by using cloud computing and powerful multicore central processing units (CPUs), graphics processing units (GPU) and field-programmable gate arrays (FPGAs) with parallel processing methods.

III. EXISTING SYSTEM

In recent advances in big data for health informatics and their role to tackle disease management are presented, for instance, diagnosis prevention and treatment of several illnesses. Clinical context produces unstructured data or a semistructured from data such as handwritten doctor notes. These data may have differences in meanings and interpretations. The process of big data in the healthcare industry is broken into five stages: Data Acquisition, Data storage, Data management, Data Analytics, and Data visualization and report. The further content explains each one of the mentioned stages: Big data in healthcare organizations consist of both internal (Patient's health history) and external data (Third party data/data from public providers) [3]. Both of these can be housed on cloud computing [2]. The data management process includes, storing, organizing, maintaining, retrieving, data mining, data monitoring and data validating. Data Analytics is broadly a process of converting raw data into information. Big data analytics in healthcare is segregated into Descriptive, Diagnostic, Predictive, and Prescriptive Analytics. Data Visualization is defined as graphical representation of analytics result obtained from healthcare data analysis used for better understanding of the correlation of data. Big-data analytics in current framework system is processed by clustering and scanning multiple nodes of clusters in the network. In Hospital Network, NoSQL database is used to collect and manage the enormous amount of real-time data from diverse sources, which assists the management in administering high-risk patients and reducing everyday expenditures. In terms of Monitoring of Patient's vital information, use of Hadoop-based components in the Hadoop Distributed File System (HDFS), including the Impala, HBase, Hive, Spark, and Flume frameworks are common, to convert the unstructured data generated using sensors which take patient's vital signs. Using Hadoop, healthcare staff can analyze these unstructured data. Hadoop technology supports the healthcare intelligence applications. Hadoop ecosystem's Pig, Hive, and MapReduce technologies process large datasets related to medicines and other factors to extract meaningful information for medical institutes. Companies, with help of Hadoop application, have begun to use a prediction model to determine the scams and fraud committers before the action is taken place.

IV. BIG DATA CHARACTERISTIC

The term big data is described by the following characteristic value, volume, velocity, variety, veracity and variability, denoted as 6 "Vs" [13], [14], shown in Figure Figure1.1. Besides these 6 "Vs", some authors have defined more than these 6 properties to describe big data characteristics [15]. The volume of health and medical data is expected to rise intensely in the years ahead, usually measured in terabytes, petabytes even yottabytes [14], [16]. Volume refers to the amount of data, while velocity refers to data in motion as well as and to the speed and frequency of data creation, processing and analysis. Complexity and heterogeneity of multiple datasets, which can be structured, semi-structured and unstructured, refer to

the variety. Veracity referrers to the data quality, relevance, uncertainty, reliability and predictive value [14], while variability regards about consistency of the data over time. The value of the big data refers to their coherent analysis, which should be valuable to the patients and clinicians.

Considering the big data characteristics, data searching, storage and analysis, a very appropriate and promising software platform for development of applications that can handle big data in medicine and healthcare is the open-source distributed data processing platform Apache Hadoop MapReduce [1], [17] that is based on data-intensive computing and NoSQL data modeling techniques [18].

4.1 Big Data Analytics

Applications of big data analytics can improve the patient-based service, to detect spreading diseases earlier, generate new insights into disease mechanisms, monitor the quality of the medical and healthcare institutions as well as provide better treatment methods [7].

Data mining techniques employed on EHRs, web and social media data enable identifying the optimal practical guidelines in the hospitals, identifying the association rules in the EHRs [8] and revealing the disease monitoring and health-based trends. Moreover, integration and analysis of the data with different nature, such as social and scientific, can lead to new knowledge and intelligence, exploring new hypothesis, identifying hidden patterns [14].

Figure 1: The 6 V's of big data

Nowadays, smart phones are excellent platforms to deliver personal messages to patients to involve them in behavioral changes to improve their wellbeing and health conditions. The mobile phone messages can substitute delivering of medical and motivational advices to the patients [14].

4.2 Challenges in Big Data Analytics

Regarding collection of large amount data, some challenging issues should be considered. Obtaining high-throughput – omics data is tied to the cost of experimental measurements. Concerning heterogeneity of the data sources, the noise of the experimental – omics data and the variety of the experimental techniques, environmental conditions, biological nature should be considered, before integration of these heterogeneous data and before employing of the data mining methods. Different data mining techniques can be applied on these heterogeneous biomedical data sets, such as: anomaly detection, clustering, classification, association rules as well as summarization and visualization of those big data sets.

These shortcomings might lead to the unreliability of some of the data points, such as missing values or outliers. Despite of these drawbacks of the – omics data, EHRs data are very influenced by the staff who entered the patient's data, which can lead to entering missing values, incorrect data as a result of mistakes, misunderstanding or wrong interpretation of the original data [5]. Integration of data from various databases and standardization for laboratory protocols and values still remain challenging issues [10].

High dimensionality of the – omics data means, that there have many more dimensions or features than the number of samples, and on the other side the EHRs data which regard to the individuals/patients, makes data mining techniques to be more challenging task.

The subsequent stage is the pre-processing of the data, which usually envelop handling noisy data, outliers, missing values, data transformation and normalization. This data pre-processing enables to be applied statistical techniques and data mining methods and thus the big data analytics quality and outcomes can improve and can result with discovering of novel knowledge. This novel knowledge obtained by integration of the – omics and EHRs data should results with improving of the implemented healthcare to the patients as well to advanced decision making by the healthcare decision policy makers.

4.3 Big data Privacy and Security

Two important issues towards big data in healthcare and medicine are security and privacy of the individuals/patients [4], [3]. All medical data are very sensitive and different countries consider these data as legally possessed by the patients [2]. To address these security and privacy challenges, the big data analytics software solutions should use advanced encryption algorithms and pseudo-anonymization of the personal data. These software solutions should provide security on the network level and authentication for all involved users, guarantee privacy and security, as well as set up good governance standards and practices.

V. METHODOLOGY

Solution for Integrating Medical Data Streams and Data quality: A unified web based system can be developed having the features of not just in taking Unicode but also image, audio and voice format. Level 1 Solution: Inclusion of "Integration Bucket" functionality in landing zone. This concentrates on Data filtering which maintains high quality by removing all redundant data. Data collected from various sources are directed to landing zone causing repetition of data in cloud.

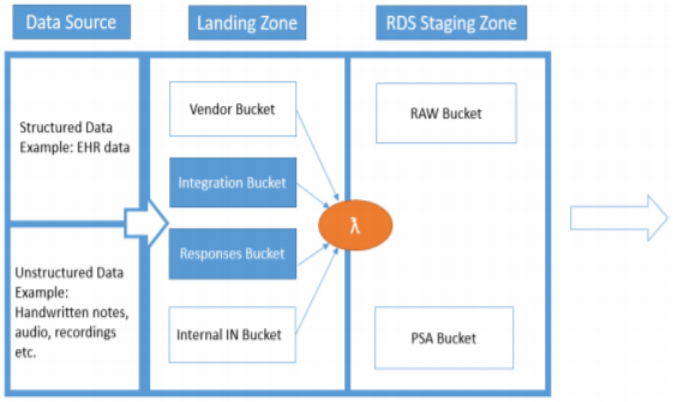

Figure 2: Components in Cloud Computing

To avoid this, the adding a functionality of "Integration" to remove repeated and redundant data can save time and complexity. Such a solution is tremendously useful for performing a critical analysis on the fly. (Figure 1) Solution for integration with real-time responses: Most of the existing solutions are built on the Hadoop-MapReduce framework, which mostly solved the data volume challenge. However, due to the extravagant sorting algorithm that Hadoop relies heavily on for performing causes the performance to bottleneck. Based on our study we concluded that this problem can be resolved by incorporating it with cloud computing.

Level 2 Solution: Inclusion of "Responses Bucket" function in landing zone. Real time responses occur only when there are real time interactions. Interactions are not unique; repeated interactions can be captured in landing stage with the use of a pointer to call particular function of real time response (Figure 2). This will help save tremendous amount of processing time. Responses are processed and to give out a relevant solution, relevant interactions can be captured by this function to avoid duplication. All the filtered solutions are then provided to LAMBDA bucket. Since LAMBDA is python based programming, it reduces the number of operations. The data is retrieved using SQL query.

VI. CONCLUSION

Big data analytics in medicine and healthcare is very promising process of integrating, exploring and analysing of large amount complex heterogeneous data with different nature: biomedical data, experimental data, electronic health records data and social media data. As a further work, the big data characteristics provide very appropriate basis to use promising software platforms for development of applications that can handle big data in medicine and healthcare. One such platform is the open-source distributed data processing platform Apache Hadoop MapReduce that use massive parallel processing (MPP). These applications should enable applying data mining techniques to these heterogeneous and complex data to reveal hidden patterns and novel knowledge from the data.

Recent hardware innovations in processor technology, newer kinds of memories/network architecture will minimize the time spent in moving the data from storage to the processor in a distributed setting.

VII. REFERENCES

[1] Yang C, Li C, Wang Q, Chung D, Zhao H. Implications of pleiotropy: challenges and opportunities for mining big data in biomedicine. *Front Genet.* 2015;6:229.

[2] Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. *IEEE J Biomed Health Inform.* 2015;19:1209–15.

[3] Kankanhalli A, Hahn J, Tan S, Gao G. Big data and analytics in healthcare: introduction to the special section. *Inform Syst Front.* 2016;18:233–5.

[4] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inform Sci Syst.* 2014;2:3.

[5] Wu PY, Cheng CW, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. –Omic and Electronic Health Record Big Data Analytics for Precision Medicine. *IEEE Trans Biomed Eng.* 2017;64:263–73.

[6] Wang Y, Kung LA, Wang WY, Cegielski CG. An integrated big data analytics-enabled transformation model: application to health care. *Inf Manag.* 2017;55:64–79.

[7] El-Gayar O, Timsina P. Opportunities for business intelligence and big data analytics in evidence based medicine. System Sciences (HICSS); 47th Hawaii international conference on 2014.2014. pp. 749–57.

[8] Gu D, Li J, Li X, Liang C. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. *Int J Med Inform.* 2017;98:22–32.

[9] Gligorijević V, Malod-Dognin N, Pržulj N. Integrative methods for analyzing big data in precision medicine. *Proteomics.* 2016;16:741–58.

[10] Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health

care: a literature review. *Biomed Inform Insights.* 2016;8:1.

[11] Gaitanou P, Garoufallou E, Balatsoukas P. The effectiveness of big data in health care: a systematic review. *In: Metadata and semantics research.* 2014:141–53.

[12] Lillo-Castellano JM, Mora-Jimenez I, Santiago-Mozos R, Chavarria-Asso F, Cano-González A, García-Alberola A. et al. Symmetrical compression distance for arrhythmia discrimination in cloud-based big-data services. *IEEE J Biomed Health Inform.* 2015;19:1253–63.

[13] Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang GZ. Big data for health. *IEEE J Biomed Health Inform.* 2015;19:1193–1208.

[14] Archenaa J, Anita EM. A survey of big data analytics in healthcare and government. *Procedia Comput Sci.* 2015;50:408–13.

[15] Borne K. *Top 10 big data challenges – a serious look at 10 big data V's.* MAPR; 2014. NO4, 80.

[16] Hermon R, Williams PA. Big data in healthcare: what is it used for?; Australian Ehealth Informatics and Security Conference; 2014. pp. 40–9.

[17] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. *Commun ACM.* 2008;51:107–13.

[18] Trifonova OP, Il'in VA, Kolker EV, Lisitsa AV. Big data in biology and medicine. *Acta Naturae.* 2013;5:13–6.

[19] Agarwal M, Adhil M, Talukder AK. *International Conference on Big Data Analytics.* Cham, Switzerland: Springer International Publishing; 2015. Multi-omics multi-scale big data analytics for cancer genomics; pp. 228–43.

Cite this article as :

Tanmayee Tushar Parbat, Rohan Benhal, Honey Jain, Dr. Vinayak Musale, "Emerging Strategies to Big Data Analytics in Healthcare", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 6, pp. 254-260, November-December 2021.

Journal URL : https://ijsrcseit.com/CSEIT217672