A Survey on Big Data Frameworks and Approaches in Health Care Sector

Jay L. Borade[1], Joel Dsouza[2], Gunjan Munde[3], Divya Varghese[4]

[1]Assistant Professor Department of Information Technology, Fr. Conceicao Rodrigues College of Engineering, Mumbai, Maharashtra, INDIA
[2,3,4]Student, Department of Information Technology, Fr. Conceicao Rodrigues College of Engineering, Mumbai, Maharashtra, INDIA

Abstract - *Big data is becoming a booming technology in many industries. There are many aspects of big data that have been recognized by researchers and technology consultants around the world. Big data analytics shows high potential in the healthcare industry and can be used to reveal causal relationships. Analysis of big data in healthcare can reveal valuable relationship between IT transformation, practices, benefits and business value. In this paper we mainly look at many approaches proposed by different researchers about how with the collaboration of big data and electronic healthcare records we can bring in many benefits in the field of healthcare and well-being. In this paper we breakdown the discussion into four sections starting with introduction, methodology, results and conclusion.*

Key Words: Big Data, Electronic Health Records, Healthcare, Hadoop, Apache Framework

1. INTRODUCTION

Data processing is one of the biggest motivations in the areas of research with the main focus of analyzing huge amounts of data. In the present era the combination of big data, cloud computing and data warehousing are becoming popular in the field of healthcare. Healthcare information in the form of Electronic Health Records (EHR) can be created, used, stored and retrieve important patient data. The volume of EHR is too much to be applicable using any scale. Any data can be related to healthcare such as physician's notes, biological data, lab reports, patient metadata, case history diet regime and list of doctors in a particular hospital [1]. We can consider many sources of data in healthcare such as PubMed, SEER and NCI database, HCUP (Health Cost and Utilization Project) and CMS. We need to understand what type of data that we are dealing with and what significant role does it play in acquiring knowledge. Knowledge can be briefly be thought of as tacit or explicit [2]. Tacit knowledge can be said to be knowledge developed due to individual experience. This type of knowledge can be very subjective and could be very complex. On the other hand explicit knowledge is when it easy to collect format and distribute data of various persons. When it comes to healthcare we can consider different things like medical reports and records that help in acquiring explicit knowledge. It could also include medical procedures and diagnosis of certain diseases. Different data sources can be considered when it comes to healthcare analysis like the following:-

1. Electronic Health Record (EHR)
2. Genetic Data
3. Medical Imaging Data
4. Unstructured Clinical Notes
5. Documentation and Test Reports

Beneficiaries of Big Data analytics in healthcare:-

1. Doctors: A doctor can check the health records of a patient at any place. Consider a patient that has moved to a new place. The treatment of that patient can continue without any hassle as the data is available in the form of EHR present online

2. Government:- The usage of this data can help the government to provide subsidies to private hospital where government health centres are not available

3. Insurance companies: Insurance companies can use this information in order to prevent false claims. It can also help in improving health products

4. Pharmaceutical Companies:-Pharmaceutical companies can use these records for research and development work. Customizable drug can created for diseases based on gained information

5. Marketing and Advertisement Agencies:-They can use big data information about different diseases prevalent in society. They can then make appropriate training programs for educating the rural people of India

In the following section we will have a look at how different big data tools and architecture can be used to in providing the in depth knowledge of healthcare analytics.

2. METHODOLOGY

This section briefly gives an overview of different tools that can be used in big data and healthcare analysis.

2.1 Tools and Technology in Big Data

Hadoop Architecture: Apache Hadoop uses the master slave architecture where in you have two nodes namely the datanode and namenode. The namenode performs as a master and datanode as a slave [3]. The namenode manages the access to datanodes. The datanodes on the other hand administrate and store data across multiple nodes

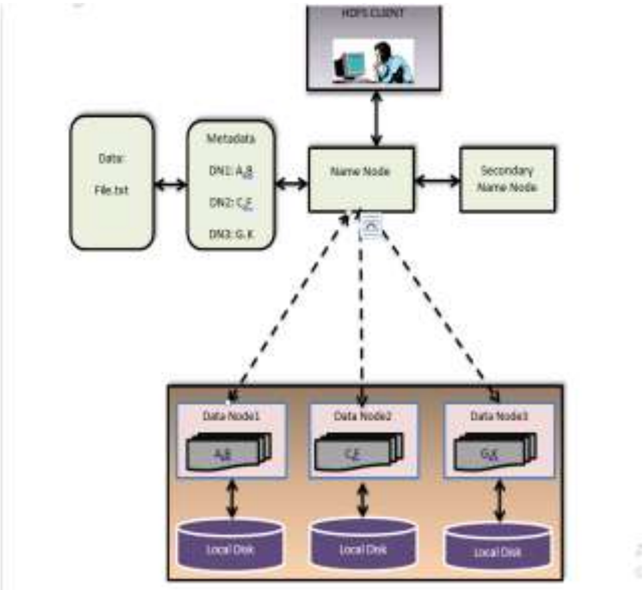

Fig -1: Hadoop File System Architecture

Hadoop MapReduce:

The Hadoop MapReduce is a programming model used to process huge size of data sets across a Hadoop Cluster [3]. Hadoop framework also provides the scheduling, distribution, and parallelization services to process the big data [3].

Apache Sqoop:

Apache Sqoop can extract data from Hadoop Distributed File System and export it to some external data storage such as relational database.

2.2 Data Warehouse Based EHR platforms

OpenEHR and EHR4CR are the most common platforms used as data warehouse for EHR analytics. The EHR4CR use something known as a Common Information Model (CIM). A CIM is a higher order query applied on EHR data warehouse. EHR data is unstructured data according to CIM. We need to apply the ETL operations in order to get unstructured data into structured data and the EHR4CR takes care of local information models and the EHR4CR CIM. OpenEHR is designed based on requirement captured through many years. The minimum requirements need to build openEHR system are, data warehouse includes EHR, archetype repositories, terminology, and demographic or identity information [3]. The demographic repository is used as a front end to store patient master index (PMI). The EHR can be configured to include either no demographic or some identifying data [3]. It is clear that the above platforms are a good option for the analysis of historical healthcare data. However in the present era the driving needs tend to implement new technology other than data warehousing in order to handle the new requirements.

2.3 EHR, Big Data Systems and Retrieval of EHR records

A cumulative approach of structured and unstructured data stemming from clinical and nonclinical modes of existence will help us to understand and predict diseases. Fig 2 shows a basic framework of more complicated platform which can process big data of EHR and give us more desirable performance in complicated analytics rather than data warehouse platform. In fact, this method allows healthcare institutions to productively document a total clinical confrontation rapidly and recover necessary data from EHR big data cluster in high execution processing. Therefore, this approach gives most demands that are needed in big EHR era. To examine tasks and to produce insight that enable decision makers and to take steps and enhance health performance and functional impact, Healthcare institutions created data warehouses. An increment in data should encourage the Healthcare organizations to move to big data technology to provide new features that are mentioned in data warehouse approach. Parallel and distributed computing are some of the basic architectures that are used in the handling of big data, because of which it can execute processes concurrently on number of machines. An open-source package called Hadoop was released by Apache for distributed data handling recently. Hadoop Distributed File System also known as HDFS can access, handle, and retrieve all data files simultaneously among the Hadoop group. In [4] it is discussed that how big data systems can be used for designing platforms for conducting elastic search. In order to conduct elastic search we can crawl a set of websites. Extract data from these web structures and index them by elastic search methods. After the entire procedure is conducted a client can search by setting a search key and search option. The steps of conducting elastic search is also briefly elaborated in [4] as below:

Step 1: Collection and storage

Step 2: Classification

Step 3: Create Index

Step 4: Search and Analyze

Some applications of big data and elastic search can therefore be used in order to develop sustainable applications for ease of search useful patterns that can be used in long term for healthcare.

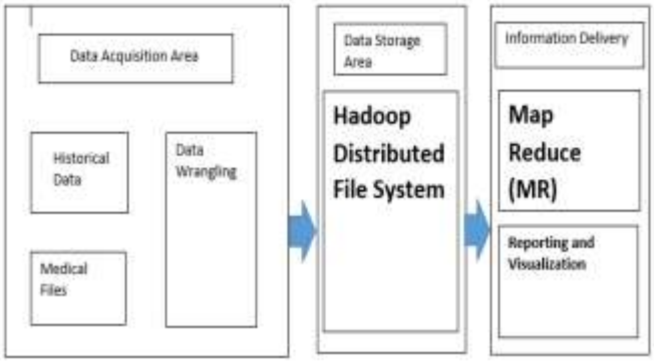

Fig 2:- Big Data Analytics in Healthcare

3. Some Applications in Discussed Papers

On observation we see that there exists many technology and tools that can be used in order to develop systems that can provide easy access to healthcare data. On combining the Big Data like Hadoop and openEHR or EHR4CR we can develop systems where easy access to EHR is possible. Following are some of the results that can be obtained on the combination of the above tools

1. Healthcare analysis: On analysis a patients past reports and current health checks a doctor could possibly get better discernment on providing prescription
2. Prognosis: Prognosis is nothing but the early diagnosis of chronic age related diseases.
3. Report Management: Quick access to ones reports and health info any doctor can consider what treatment needs to be given to a patient
4. Healthcare Service Recommendations: Better services could be provided by analysis of Healthcare Service Providers on the basis of user ratings and feedback
5. Follow Up Systems: Follow Up Systems could be used in order to keep track of a patients current health status with regards to his/her last visit.

Thus the above mentioned applications can possibly be brought to reality by the integration of two common available technology stacks

4. Conclusion

Nowadays, space (from hospital to home and carry) and time (from discrete sampling to continuous tracking and monitoring) are no longer a stumbling stone for modern healthcare by using more powerful analysis technologies. Medical diagnosis is evolving to patient-centric prevention, prediction, and treatment. The big data technologies have been developed over the years and will be implemented everywhere. Consequently, healthcare will also enter the big data era. More precisely, the big data analysis technologies can be used as guide in lifestyle, as a tool to support in the decision-making, and as a source of innovation in the evolving healthcare ecosystem. This paper has presented a smart health system assisted by cloud and big data, which includes 1) a unified data collection layer for the integration of public medical resources and personal health devices, 2) a big data enabled and data-driven platform for multisource heterogeneous healthcare data storage and analysis, and 3) a unified API for developers and a unified interface for users. Supported by Health-CPS, various personalized applications and services are developed to address the challenges in the traditional healthcare, including centralized resources, Information Island, and patient passive participation. In the future, we will focus on developing various applications based on the Health-CPS to provide a better environment to humans.

REFERENCES

[1] Gunasekaran Manogaran, Chandu Thota, Daphne Lopez, V. Vijayakumar, Kaja M. Abbas and Revathi Sundarsekar, “Big Data Knowledge System in Healthcare” Springer International Publishing AG 2017 C. Bhatt et al. (eds.), Internet of Things and Big Data Technologies for Next Generation Healthcare, Studies in Big Data 23, DOI 10.1007/978-3-319-49736-5_7

[2] Komal Sindhi, Dilay Parmar and Pankaj Gandhi, “A Study on Benefits of Big Data for Healthcare Sector of India” Springer Nature Singapore Pte Ltd. 2019 D. K. Mishra et al. (eds.), Data Science and Big Data Analytics, Lecture Notes on Data Engineering and Communications Technologies16

[3] Youssef M.Essa, Gamal ATTIYA2, Ayman El-Sayed and Ahmed ElMahalawy, “Data processing platforms for

electronic health records" Received: 15 February 2017 /Accepted: 3 January 2018 IUPESM and Springer-Verlag GmbH Germany, part of Springer Nature 2018

[4] Aiqin Yang, Shiwei Zhu, Xianyi Li, Junfeng Yu, Moji Wi and Chen Li, "The research of Policy Big Data Retrieval and Analysis based on Elastic Search" 978-1-5386-6987-7/18/$31.00 ©2018 IEEE