Available online at: http://www.ijmtst.com/vol4issue12.html

International Journal for Modern Trends in Science and Technology

ISSN: 2455-3778 :: Volume: 04, Issue No: 12, December 2018

Disease Analytics in Healthcare System Using Hadoop

Patel Vishruti[1] | Prof. M.D. Ingle[2]

[1,2]Computer Science, JSPM's Jayawantrao Sawant College of Engineering,

To Cite this Article

Patel Vishruti and Prof. M.D. Ingle, "Disease Analytics in Healthcare System Using Hadoop", *International Journal for Modern Trends in Science and Technology*, Vol. 04, Issue 12, December 2018, pp.-31-35.

Article Info

Received on 26-Oct-2018, Revised on 07-Dec-2018, Accepted on 19-Dec-2018.

ABSTRACT

In today's world, due to technological advancements, the amount of data that is getting generated is growing rapidly. Enterprises worldwide will need to perform data analytics with these huge data datasets to make business decisions and stay competitive. Storage of data sets and performing data analytics was traditionally accomplished using RDBMS (Relational Database Management System). However, RDBMS would be inefficient and time consuming when performing data analytics on huge data sets. Many dimensions of big data still present issues in its use and adoption, such as managing the volume, variety, velocity, veracity, and value, the accuracy, integrity, and semantic interpretation are of greater concern in clinical application. Hadoop came into existence recently and overcomes the limitations of existing RDBMS by providing simplified tools for efficient data storage and faster processing times for data analytics. The purpose of this work is to study different Hadoop functionalities in detail and perform data analytics on a health care data set using Hadoop. A health care data set comprising of 1.5 million patient records is considered for the data analysis. Different use cases as to be considered and will perform analytics using MapReduce, Hive and Pig functionalities of Hadoop.

KEYWORDS:*big data analytics, healthcare, clustering, analysis, hadoop*

I. INTRODUCTION

Data in health care industry is originated from different sources like clinical data, diagnosis data, doctor's prescription, electronic health records(HER), medical images, etc. collectively all this data becomes big data in health care industry. Big data is nothing but the data big in size. Health Insights Study, states that the worldwide health care data grew up to 500 petabytes in 2012 and has been estimated to grow more than 25,000 petabytes by 2020. Big data applies to information which cannot be processed or analyzed using traditional tool and technique. To calculate the *value* of the big healthcare data, to deliver the best evidence based, patient-centric, result oriented and accountable care. [1] The main characteristics of big data are volume, velocity, veracity, verity, Value. Volume indicates the amount of data that may in Gbs or Tbs or more than that. Velocity is considered as how fast the data are created and collected. Verity means how many types of data are present that may be structured, semi-structured or unstructured. Veracity means the truthfulness of data. Value means, how much the data is valuable. The Big data processing technologies includes

machine learning percentages determine the algorithms, natural language processing algorithms, predictive modelling and other artificial based techniques. Analysing Big Data will bring out better outputs for the organizations and will improve their performance.

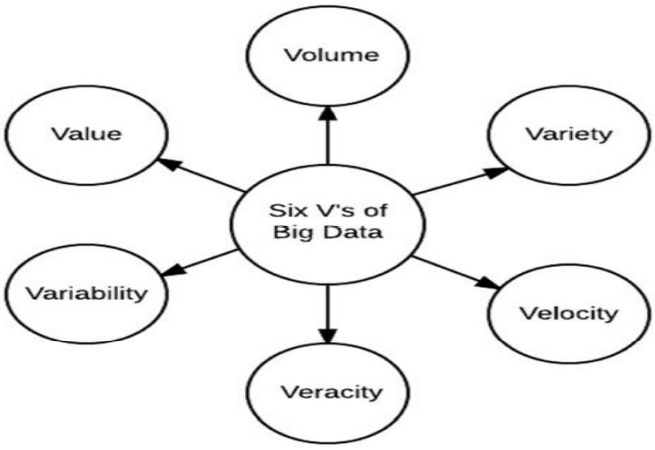

Figure 1.1: 6V's of Big Data.

There are various Big Data technologies which could be put into work. These technologies are reliable, fast, cost effective and robust. The Big Data technologies are mostly open sourced. To name a few; Hadoop, Spark, Cassandra, MongoDB (Documented oriented database) are amongst the various technologies which are on the go. Another important class of application of Big Data in the healthcare domain includes the Medical Body Area Networks (MBANs). MBANs enable a continuous monitoring of patient's condition by sensing and transmitting recorded measurements such as heart rate, electrocardiogram (ECG), body temperature, respiratory rate, chest sounds, and blood pressure etc. MBANs will allow:

1. Real-time and historical monitoring of patient's health;
2. Control of Infection;
3. Tracking and identification of patients; and
4. Geo-fencing and vertical alarming.

To manage and analyse such massive MBAN data from millions of patients in the real-time, healthcare providing organization will need upgrade to a highly intelligent and highly secure ICT infrastructure. [2] These issues can be managed and analysed by deploying Big Data technologies in the healthcare systems. Deploying Big Data analytics with various health IT apps like HER, EMR (Electronic Medical Records), CDSS (Clinical Decision Support Systems) and PHR (Personalized Health Records) etc., will improve the health systems. Predictive analysis provides the ability to make financial and clinical decisions based on predictions made by the system.

II. Literature Survey

It is found that developing an HMS that can alert the pharmacist of the expiry date of drugs at a given time and handle all departments in the hospital. The proposed system addresses the problem of data quality in electronic patient records using a computerized patient records report system as an example [1Physicians extracted five parameters from a traditional free text report and encoded these parameters thus producing a computer process able report. The proposed system is divided into Receptionist's module, Doctor's module and Pharmacist's module.

Proposes a Map Reduce Algorithm method, which is suitable Semi-Structural Data which enhances the Map Reduce works by breaking the processing into two phases: the map phase and the reduce phase[2] . Each phase has key-value pairs as input and output, the types of which may be chosen by the programmer. The programmer also specifies two functions: the map function and the reduce function.Min Chensuggested that key benefit of the method came from: 1) The global interdependence model of EL decisions; 2) The purely collective nature of the MAP REDUCE algorithm, in which evidence for related EL decisions could be reinforced into high-probability decisions. Experimental results showed that the method could achieve significant performance improvement over the traditional EL methods. [9]

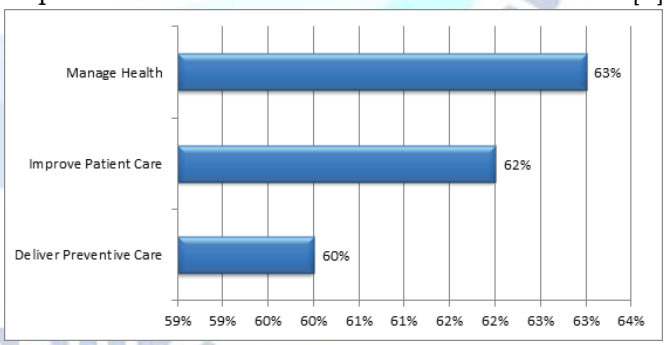

Figure 2.1 Graph depicting benefits of Big Data in Healthcare

Solved problem of data quality in electronic patient records using a compute-rized patient records report system. The limitation was only leading for Semi-structured data. And supports for MapReduce Does not support for Apache Hadoop. Many emergence of massive datasets in a clinical setting presents both challenges and opportunities

in data storage and analysis [4]. This so called "big data" challenges traditional analytic tools and will increasingly require novel solutions adapted from other fields. Advances in information and communication technology present the most viable solutions to big data analysis in terms of efficiency and scalability. It is vital those big data solutions are multithreaded and that data access approaches be precisely tailored to large volumes of semi-structured/unstructured data. The MapReduce programming framework uses two tasks common in functional programming: Map and Reduce[4]. MapReduce is a new parallel processing framework and Hadoop is its open-source implementation on a single computing node or on clusters. Compared with existing parallel processing paradigms (e.g. grid computing and graphical processing unit (GPU)), MapReduce and Hadoop have two advantages: 1) fault-tolerant storage resulting in reliable data processing by replicating the computing tasks, and cloning the data chunks on different computing nodes across the computing cluster; 2) high-throughput data processing via a batch processing Framework and the Hadoop distributed file system (HDFS). Data are stored in the HDFS and made available to the slave nodes for computation.The drawback for existing system is that Does not provide an index mechanism. And does not support for multi structure data.

III. TECHNOLOGY USED

Hadoop is the technology considered as an opensource platform build up by Apache. Hadoop is written in Java . Doug Cutting, named Hadoop as a technology by his son's yellow toy elephant . Hadoop provides storage and distributed process of cluster computations. MATLAB has now used in Big Data for diagnosis, dialysis and analysis of chronic kidney disease datasets. In order to predict the disease there are many modules of MATLAB which is used within data mining classification algorithms. Prediction begins with the identification of symptoms and dataset classified for that in a way to detect the disease. The Hadoop framework provides four core modules:

1. The inbuilt functions are used by the Hadoop modules.
2. HDFS : Hadoop Distributed File System, It is a java based file system, the datas can be expanded and
can store datas across different machines in an organization
3. It provides the software model for processing large data sets in parallel
4. It provides YARN (Yet another Resource Negotiator) framework, for managing, scheduling and handling the resources from middleware applications. [12]

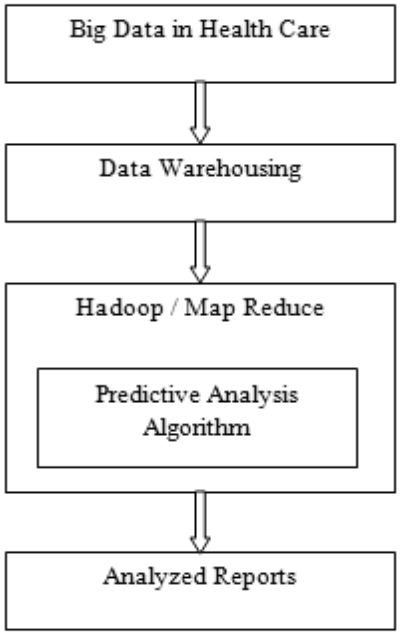

Figure 3.1 Architecture used to manage healthcare data

IV. PROPOSED SYSTEM FOR PROCESS OF BIG DATA ANALYSIS IN HEALTHCARE INDUSTRY

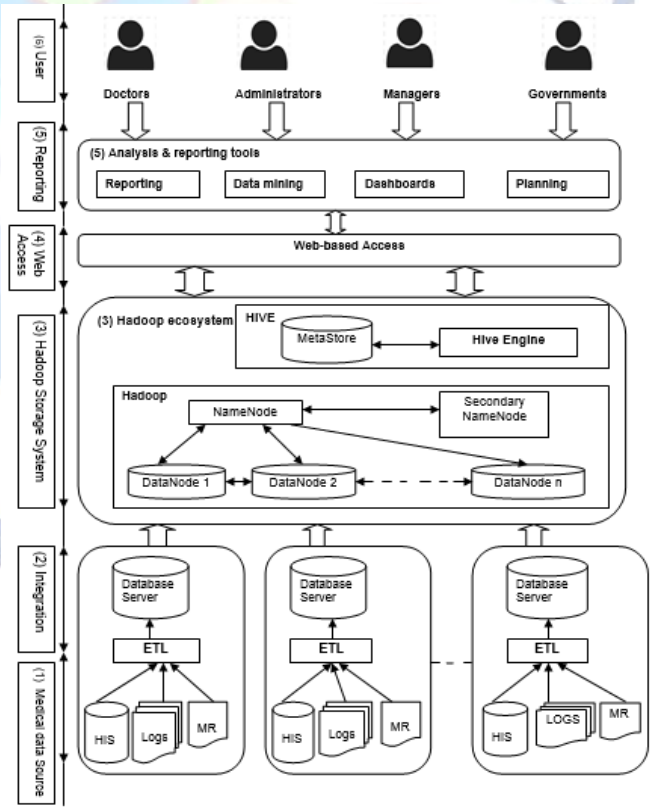

Figure 4.1 Hadoop-based system architecture of medical big data

Big data analysis has the prospective to change the method of healthcare suppliers practice cultured equipment's to increase awareness from their clinical and other data repositories and make a declared conclusion. Big data Healthcare analytics has five processes: Data Acquisition, Data Storage, Data Management, Data Analytics, and Data Visualization & Report. The Figure 5.1 presents the proposed system for data analysis in healthcare management. Hive Partitioning Algorithm we can use. Apache Hive is an ETL and Data warehousing tool built on top of Hadoop which enables ad-hoc analysis over structured and semi-structured data. It makes job easy for performing operations like of:

1. huge datasets
2. Ad-hoc queries
3. Data encapsulation

Apache Hive organizes tables into partitions.Partitioning is a way of dividing a table into related parts based on the values of particular columns like date, disease occurred, and department. Partitioning in Hive will allow distributes execution loadhorizontally.

V. Data management

Data Management in healthcare includes organizing, cleaning, retrieval, data mining, and data governance. It also includes the method of validating whether there is some scrap data or any missing values. Such data needs to be removed [7]. It helps in risk assessment of patients, personalized discharge plan. Major data management tools are Apache Ambari and HCatalog. Data retrieval is a process of extracting file or valuable information from large healthcare databases. Big data analysis in healthcare frequently contains information recovery and data mining [10]. Wang et al. [10] mention that "information retrieval is the process of searching within large document collections, and in healthcare it mainly covers medical text retrieval and medical image retrieval. Data governance refers to overall management of security, integrity, usability and availability of the data employed in an enterprise. Maintaining confidentiality of individual patient records is very important in Healthcare management.

VI. Conclusion and future work

Health care analysis is important to find and there's no better way to collect direct feedback regarding the patients and improve the product or service. However, the way of handling a patient's health using the healthcare analysis is simpler task to provide a finer health quality to the patients as earlier loading large amount of data is very difficult. By using Big data complexity of loading large amount of data can be reduced. The proposed tool enables agencies too easily and economically clean, characterize and analyze the data to identify actionable patterns and trends. There are various Big Data handling tools which can provide us information from unstructured data like medical data but we choose Hadoop, since it is an open source and it can handle huge amount of unstructured data. The information helps us in improving the different areas within a particular region. The information can also be helpful in designing and implementation of healthcare systems in a particular country. The relative time is also reduced and measured as 1x for 1 TB and 10 TB of data.Future research can include First, data security is very important in all data management systems. On large data processing, operations are often running remotely in the server or cloud side, new encryption method is needed to ensure the data privacy.

References

[1] Automated Health Care Management System Using Big Data TechnologyM. Ashish Reddy, Journal of Network Communications and Emerging Technologies, April (2016).

[2] Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring. Min Chen. Springer Science and Business Media New York 2016.

[3] Big Data for Health Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, JULY 2015

[4] Medicine and physical health system Using IoT. P.Elanthiraiyan International journal march 2015.

[5] Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends. Emad A Mohammed , Behrouz H Far, Biodata Mining 2014

[6] Telemedicine-Based WBAN Framework for Patient Monitoring. Chinmay Chakraborty

[7] MARY ANN LIEBERT, INC. AUGUST 2013 TELEMEDICINE and e-HEALTH.

[8] Framework for Dynamic Integration of Multimedia Medical Data Into Distributed m-Health Systems. Liviu Constantinescu, IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 16, NO. 1, JANUARY 2012.

[9] Implementation of a Medical Image File Accessing System on Cloud Computing. Chao-Tung Yang Lung-Teng Chen, 2010 13th IEEE International Conference on Computational Science and Engineering.

[10] Big Data Solutions in Healthcare: Problems and Perspectives. Prabha Susy Mathew,, Dr. Anitha S. Pillai, IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems.2015

[11] L. Wang, R. Ranjan, J. Kołodziej, A. Zomaya, L. Alem, Software Tools and Techniques for Big Data Computing in Healthcare Clouds, Futur. Gener. Comput. Syst. 43 (2015) 38– 39. doi:10.1016/j. future.2014.11.001.

[12] U. S Government, Department of Health and Human Services, Fedral Register, Rules and Regulations, 74(2009) 56123- 56131, Available from: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/administrative/enforcementrule/enfir.

[13] R. Nambiar, R. Bhardwaj, A. Sethi, R. Vargheese, A look at challenges and opportunities of Big Data analytics in healthcare, Proc. - 2013 IEEE Int. Conf. Big Data, (2013) 17–22. doi:10.1109/BigData.2013.6691753.

[14] O. S. Lupse, M. Crisan-Vida, L. Stoicu-Tivadar, E. Bernard, Supporting diagnosis and treatment in medical care based on big data processing, Studies in Health Technology and Informatics. 197 (2014) 65–69.

[15] M. L. Berger and V. Doban, Big data, advanced analytics and the future of comparative effectiveness research, Journal of Comparative Effectiveness Research. 3 (2014) 167–176.

[16] B. Feldman, E. M. Martin, T. Skotnes, Big Data in Healthcare- Hype and Hope, Dr. Bonnie 360 degree (Business Development for Digital Health), 2012. http://www.riss.kr/link?id=A99883549.

[17] W. Raghupathi, V. Raghupathi, Big data analytics in healthcare: promise and potential, Heal. Inf. Sci. Syst. 2 (2014) 1–10. doi:10.1186/2047-2501-2-3.

[18] Jing Bi, Zhiliang Zhu, Ruixiong Tian, and Qingbo Wang. 2010 IEEE 3rd International Conference on Cloud Computing (CLOUD), Miami, Florida, PP 370-377,2010.

[19] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li. 2012 IEEE International Symposium on Pervasive Systems, Algorithms and Networks, PP 17-23, 2012.