A SURVEY ON BIG DATA MINING IN HEALTH CARE SYSTEM

[1]M.Dhanamalar, [2]Dr. Kavitha
[1]Research Scholar, [2]Assistant Professor,
[1] Mother Teresa University, Kodaikanal
[2] Mother Teresa University. Kodaikanal

Abstract: Though the role of big data analytical The domain of healthcare acquired its influence by the impact of big data since the data sources involved in the healthcare organizations are well-known for their volume, heterogeneous complexity and high dynamism techniques, platforms, tools are realized among various domains, their impact on healthcare organization for implementing and delivering novel use-cases for potential healthcare applications shows promising research directions. In the context of big data, the success of healthcare applications solely depends on the underlying architecture and utilization of appropriate tools as evid enced in pioneering research attempts.

Keywords**:** Big Data, Hadoop, MapReduce, Health Care System**.**

1. INTRODUCTION

The idea of `big data' is not new, however, the way it is characterized is constantly changing. Various attempts at characterizing big data essentially characterize it as a collection of data elements whose size, speed, type as well as complexity expect on e to seek, adjust and invent new hardware and software mechanisms keeping in mind the end goal to successfully store, analyze and visualize the data. Healthcare is a prime case of how the three V's of data, velocity (speed of age of data), variety and volume, are a natural aspect of the data it produces. This data is spread among multiple healthcare systems, health insurers, researchers, government entities, and so on. Moreover, every one of these data repositories is siloed and inherently unequipped for providing a platform for global data transparency. To add to the three V's, the veracity of healthcare data is also basic for its significant use towards developing translational research.

Despite the innate complexities of healthcare data, there is potential and advantage in developing and actualizing big data solutions inside this domain. A report by McKinsey Global Institute suggests that on the off chance that US healthcare were to use big data creatively and effectively, the sector could make more than $300 billion in value every year. Two-thirds of the value would be through decreasing US healthcare expenditure. Historical approaches to therapeutic research have by and large focused on the investigation of disease states based on the changes in physiology as a limited view of the specific singular modality of data. In spite of the fact that this approach to understanding diseases is essential, research at this level mutes the variation and interconnectedness that characterize the genuine hidden medicinal mechanisms. Following quite a while of the mechanical laggard, the field of solution has started to acclimatize to the present advanced data age. New technologies make it possible to catch vast amounts of data about every individual patient over an extensive timescale. However, despite the advent of therapeutic electronics, the data captured and gathered from these patients has remained vastly underutilized and thus wasted.

Essential physiological and pathophysiological phenomena simultaneously manifest as changes across multiple clinical streams. This results from strong coupling among various systems inside the body (e.g., interactions between heart rate, respiration, blood pressure, and so on.) along these lines delivering potential markers for clinical assessment. Thus, understanding and predicting diseases require an aggregated approach where structured and unstructured data stemming from a myriad of clinical and non-clinical modalities are used for a more comprehensive perspective of the disease states. An aspect of healthcare research that has as of late picked up footing is in addressing some of the developing pains in presenting concepts of big data analytics to prescription. Researchers are studying the perplexing idea of healthcare data both in terms of characteristics of the data itself as well a s in the taxonomy of analytics that can be definitively performed on them.

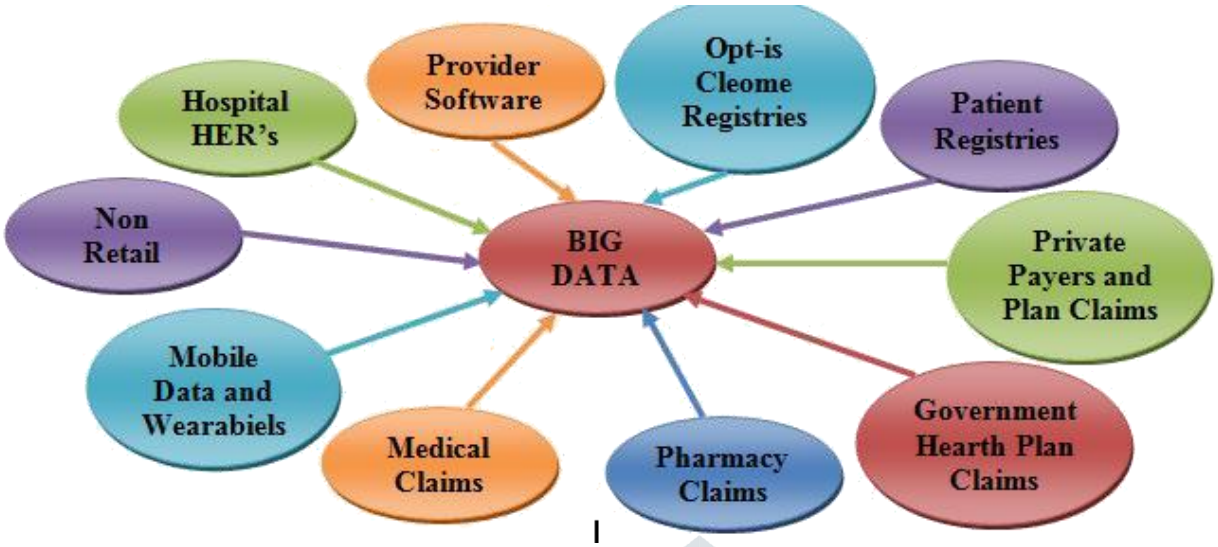

Figure : Big Data in Health Industry

Presently, we can see health monitoring has been a standout amongst the most famous research topics. The customary health monitoring, for the most part, includes the accompanying categories.

Health Cyber-Physical System: Health-arranged mobile Cyber-Physical System (CPS) plays a vital part in existing medicinal monitoring applications, such as diagnosis, disease treatment and emergency rescue, and so on. Some electronic therapeutic astute network systems suitable for countless have been designed. The End-to-End defer of medicinal data delivery is the fundamental concern, especially in the event of a mischance, or in the period when there is pestilence disease episode.

Mobile Health Monitoring: Several years back, mobile health monitoring system based on portable restorative equipment and smart phones was proposed. Smart phones are used to gather physiological signals of a human body from a variety of health monitoring devices by the virtue of committed smart phone application software. At that point those physiological signals are transmitted to therapeutic centers. In the event that necessary, it can also advise caregivers and therapeutic emergency institutions using short message service of mobile telephone.

Wearable Computing for Health Monitoring: Over a long stretch, wearable devices and wearable computing are the key research topics to empower health monitoring. As another sort of body sensor nodes, smart phone and smart watch are adjusted to measure SpO2 and heart rate in, however, such measurement data has low accuracy, few signal types and restricted medicinal uses.

Health Internet of Things: Health IoT is another approach to provide health monitoring service. The mobile sensing, restriction and network analysis based on IoT technologies can be used for healthcare.

Ambient Assisted Living: Ambient Assisted Living (AAL) aims at improving the life nature of patients, and it can tell relevant relatives, caregivers and healthcare experts. AAL-related technologies incorporate sensing innovation, physiological signal monitoring, home environment monitoring, video-based sensing, smart home innovation, pattern analysis and machine learning. Nowadays, AAL focuses on incorporating existing IoT technologies to provide the patients with more life convenience. However, it is not worried about the mobility, flexibility and accuracy of physiological signal acquisition.

Healthcare for Special Population Group: Some researchers focus on health issue of special population groups, such as the elderly individuals, discharge nester, and patients with interminable diseases. Researchers use wearable computing to help elderly individuals to live independently and safely. On basis of the coordination of mobile health software applications and wearable devices, system engineering is designed to diminish the risk of cardiovascular disease, however, there are absence of specific execution, organization and develop application cases.

Body Area Network: Existing work on body area network (BAN) focus on sensor hub's vitality saving, intra-BAN network design, implantable miniaturized scale sensors, physiological signal acquisition, and so on. Portable smart wearable health monitoring system based on BAN has been developed. However, stability, interference, security and reliability of the system should be improved.

II. LITERATURE SURVEY

Health facilities greatly depend on big data in order toprovide care to patients. Big data in the healthcareenvironment constitute a variety of data accumulatedfrom various sources and at different speeds. Tobe more specific, healthcare big data consist of largeand complex patients' data sets, which cannot behandled with traditional systems. According tobig data in healthcare holda lot of potential benefits in its premedical abilityto improve clinical decisions. Other highlighted benefits that can be achieved through big data includeearly disease detection and overall management of health.

S.No	Author & Year	Proposed Method	Advantages
1	Subramaniyaswamy, Vijayakumar, Logesh and Indragandhi - 2015	Map Reduce tasks, Collaborative, Filtering, User's Prediction, Emotion Score.	1. Map Reduce is ashuffling strategy to perform filtering and aggregation of dataanalysis tasks. 2. Collaborative Filtering Technique is used to generaterecommendations basedon user data. 3. Sentiment Analysis is a technique whichuses natural language processing and Text analysis techniquesfor predicting the user sentiments based on polarity.
2	Weiming Lu, Yaoguang Wang, Jingyuan Jiang , Jian Liu , Yapeng Shen , Baogang Wei - 2017	Group based workload balanced partitioning strategy; Group based executiontime balanced partitioning strategy.	1. Proposed the Group based Workload Balanced Partitioning Strategy, which considers the characteristics of different data stores. 2. ProposedGroup based Execution Time Balanced Partitioning Strategy by considering both the volume of data and the execution time.
3	Mohit Dayaland Nanhay Singh – 2016	Health care dataset against different research queries using Pig Latin Script	Analyzed the health care dataset against different research queries using Pig Latin Script, over the last few decades; quality of health care servicesin India has been improved tremendously because of the improved health care services, increased number of private and government hospitals and increased number of doctors with recognized medical qualification.
4	Min Chen, Yujun Ma, Jeungeun Song, Chin-Feng Lai, Bin Hu – 2016	Design Details, Key Technologies	This paper introduces design details, keytechnologies and practical implementation methods of smartclothing system.
5	Deborah A. Marshall, Lina Burgos-Liz, Kalyan S. Pasupathy, William V. Padula, Maarten J. IJzerman, Peter K. Wong, Mitchell K. Higashi, Jordan Engbers, Samuel Wiebe, William Crown, Nathaniel D. Osgood - 2015	Dynamic simulation modeling	DSM can serve as a natural bridge between the wealth of evidence offered by big data and informed decision making as a means of faster, deeper, more consistent learning from that evidence.
6	Matthew Herland, Taghi M	Big Data tools, Approaches	Thispaper will present recent

	Khoshgoftaar and Randall Wald - 2014		research using Big Data tools and approaches for the analysisof Health Informatics data gathered at multiple levels, including the molecular, tissue,patient, and population levels.
7	Venketesh Palanisamy, Ramkumar Thirunavukarasu- 2017	Various analytical avenues, Stakeholders	We have presented various analytical avenues that exist inthe patient-centric healthcare system from the perspective of various stakeholders. We havealso reviewed various big data frameworks with respect to underlying data sources, analyticalcapability and application areas.
8	YichuanWang, Nick Hajli 2016	Resource based theory, Capability building view	Our findings providenew insights to healthcare practitioners on how to constitute big data analytics capabilities for businesstransformation and offer an empirical basis that can stimulate amore detailed investigation of big data analyticsimplementation.
9	YichuanWang,LeeAnn Kung, Terry Anthony Byrd – 2016	Big data analytics, Big data analytics architecture, Big data analytics capabilities, Business value of information technology	We alsomapped the benefits driven by big data analyticsin terms of information technology (IT) infrastructure, operational, organizational, managerial andstrategic areas. In addition, we recommend five strategies for healthcare organizations that are considering toadopt big data analytics technologies.

III. ADVANTAGES OF BIG DATA MINING IN HEALTH CARE SYSTEM

By digitizing, combining and effectively using big data, healthcare organizations running from single-physician offices and multi-provider groups to vast hospital networks and accountable care organizations stand to acknowledge significant benefits. Potential benefits incorporate detecting diseases at prior stages when they can be dealt with all the more easily and effectively; overseeing specific individual and population health and detecting health mind fraud all the more rapidly and proficiently. Numerous questions can be addressed with big data analytics. Certain developments or outcomes might be predicted as well as estimated based on vast amounts of historical data, such as length of stay (LOS); patients who will choose elective surgery; patients who likely won't profit by surgery; complications; patients at risk for medicinal complications; patients at risk for sepsis, MRSA, C. difficile, or other hospital-procured illness; illness/disease progression; patients at risk for advancement in disease states; causal factors of illness/disease progression; and possible comorbid conditions. McKinsey estimates that big data analytics can empower more than $300 billion in savings for every year in U.S. healthcare, two thirds of that through reductions of around 8% in national healthcare expenditures. Clinical operations and R and D are two of the largest areas for potential savings with $165 billion and $108 billion in waste respectively.

Big data could help decrease waste and wastefulness in the accompanying three areas:

1. Clinical operations:

• Comparative effectiveness research to decide all the more clinically relevant and cost-effective ways to diagnose and treat patients.

2. Research and development:

• Predictive displaying to bring down steady loss and deliver a more slender, faster, more focused on R and D pipeline in drugs and devices;

• Statistical tools and algorithms to improve clinical trial design and patient enrollment to better match treatments to individual patients, thus decreasing trial failures and speeding new treatments to advertise;

• Analyzing clinical trials and patient records to distinguish take after on indications and discover adverse effects previously products achieve the market.

3. Public health

• Analyzing disease patterns and following disease outbreaks and transmission to improve public health surveillance and speed response;

• Faster development of all the more accurately focused on vaccines, e.g., choosing the yearly flu strains; and,

• turning a lot of data into significant data that can be used to distinguish needs, provide services, and predict and prevent crises, especially for the advantage of populations.

4. Evidence-based medicine:

Consolidate and analyze a variety of structured and unstructured data-EMRs, budgetary and operational data, clinical data, and genomic data to coordinate treatments with outcomes, predict patients at risk for disease or readmission and provide more proficient care;

5. Genomic analytics:

Execute quality sequencing all the more proficiently and cost effectively and make genomic analysis a piece of the standard restorative care decision process and the developing patient medicinal record;

6. Pre-adjudication fraud analysis:

Quickly analyze extensive numbers of claim requests to diminish fraud, waste and abuse;

7. Device/remote monitoring:

Catch and analyze continuously extensive volumes of fast-moving data from in-hospital and in-home devices, for safety monitoring and adverse event prediction;

8. Patient profile analytics:

Apply advanced analytics to patient profiles (e.g., segmentation and predictive demonstrating) to recognize individuals who might profit by proactive care or lifestyle changes, for instance, those patients at risk of developing a specific disease (e.g., diabetes) who might profit by preventive care.

IV METHODS USED FOR BIG DATA MINING

With the evolution of computing technology, immense volumes can be overseen without requiring supercomputers and high cost. Numerous tools and techniques are available for data management, including Google Big Table, Simple DB, Not Only SQL (No SQL), Data Stream Management System (DSMS), Memcache DB, and Voldemort. However, companies must develop special tools and technologies that can store, access, and analyze a lot of data in close ongoing because Big Data differs from the customa ry data and can't be stored in a single machine. Besides, Big Data lacks the structure of customary data. For Big Data, some of the most usually used tools and techniques are Hadoop, MapReduce, and Big Table. These innovations have re-imagined data management because they effectively process a lot of data productively, cost effectively, and in an opportune way.

4.1 Hadoop

Hadoop is composed in Java and is a best level Apache venture that started in 2006. It emphasizes discovery from the perspective of scalability and analysis to acknowledge close impossible feats. Doug Cutting developed Hadoop as a collection of open-source projects on which the Google MapReduce programming environment could be connected in a distributed system. Presently, it is used on a lot of data. With Hadoop, enterprises can harness data that was previously hard to oversee and analyze. Hadoop is used by around 63% of organizations to oversee gigantic number of unstructured logs and events. Specifically, Hadoop can process to great degree vast volumes of data with varying structures (or no structure by any stretch of the imagination). The accompanying section details various Hadoop projects and their links as per them. Hadoop is composed of HBase, HCatalog, Pig, Hive, Oozie, Zookeeper, and Kafka; however, the most well-known components and surely understood paradigms are Hadoop Distributed File System (HDFS) and MapReduce for Big Data. The Hadoop ecosystem, as well as the connection of various components to each other.

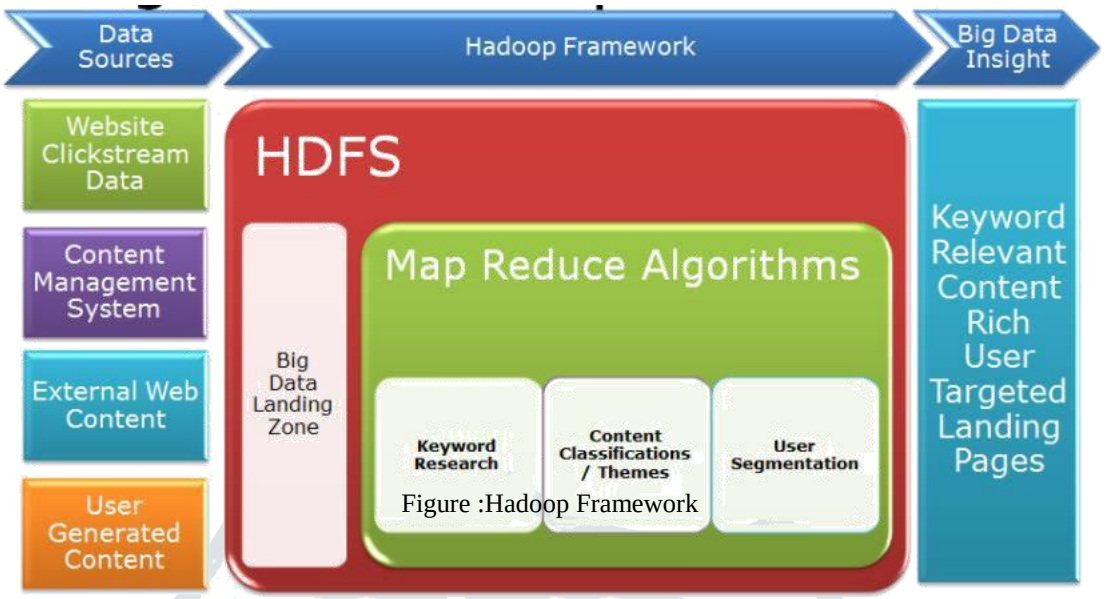

Figure :Hadoop Framework

4.2 MapReduce

MapReduce is a philosophy to process data paralleled by the distribution of data as small chunks across the clusters. The gigantic volume data divided into chunks has to be checked for interdependencies to avoid basic problems while total of these resulting sets to get the required structured data. The data have to be clustered based on their due date scheduled for processing, priorities and data dependencies. In the event that processing of one data requires the yield of other data as its info, at that point it can be joined together to frame a cluster. The clusters can also be framed on the basis of need and processing of the data clusters. MapReduce procedure is essentially used for parallel processing of data sets across various clusters known as separating, performed by the guide work and creating calculation result by total, which is the decrease work. The Map Join Reduce strategy is used for the processing the heterogeneous data items. It does not shuffle the middle results that are to be passed from mapper to reducer and it avoids check pointing of results habitually. In MapReduce, the guide tasks and fragmented lessen tasks will be re executed instead of the whole guide and decrease tasks in case of a single hub disappointment. We can achieve the base execution time. Some frameworks based on MapReduce are proposed which are fit for understanding data semantics, simplifying the written work of analytics applications and potentially improving execution by decreasing MapReduce phases.

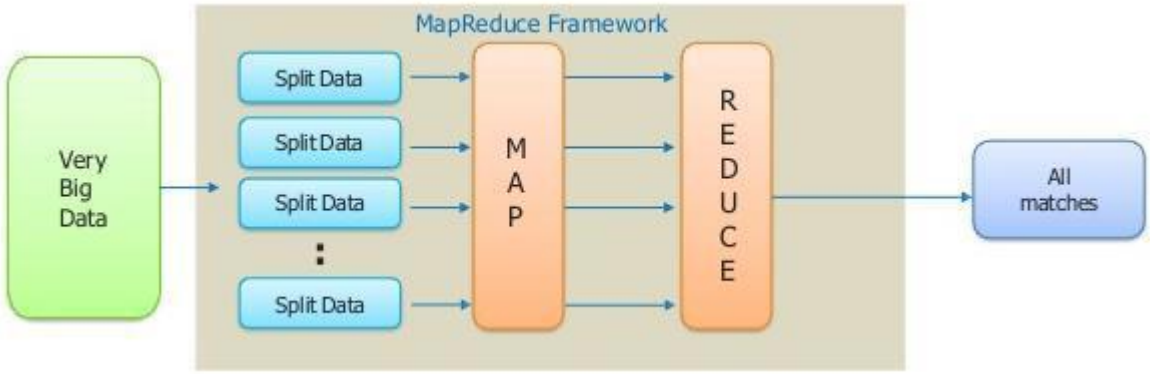

Figure: Map Reduce Framework

V. CONCLUSION

Big data analytics has the potential to transform the way healthcare providers use sophisticated technologies to pick up insight from their clinical and other data repositories and settle on educated decisions. Later on, we'll see the quick, widespread usage and use of big data analytics across the healthcare association and the healthcare industry. With that in mind, the several challenges featured above must be addressed. With the effect of big data, healthcare area was revamped and offer intensive solutions for taking care of diversified big data sources that range from patient health records to medical images. This paper reviews various research attempts at establishing healthcare frameworks and summarizes their significant outcomes. The summary of contributions by various researchers highlights the data source used, adopted analytical techniques and different features.

VI. REFERENCES

1. Subramaniyaswamy V, Vijayakumar V, Logesh R and Indragandhi V, "Intelligent travel recommendation system by mining attributes from community contributed photos" 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)
2. Weiming Lu , Yaoguang Wang, Jingyuan Jiang, Jian Liu, Yapeng Shen, Baogang Wei, "Hybrid storage architecture and efficient MapReduce processing for unstructured data", Elsevier 2017.
3. Mohit Dayal and Nanhay Singh, "An Anatomization of Aadhaar Card data set – A big data challenge", Elsevier-2016.
4. Min Chen, Yujun Ma, Jeungeun Song, Chin-Feng Lai, Bin Hu, "Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring", Springer Science+Business Media New York 2016.
5. Deborah A. Marshall, Lina Burgos-Liz, Kalyan S. Pasupathy, William V. Padula, Maarten J. IJzerman, Peter K. Wong, Mitchell K. Higashi, Jordan Engbers, Samuel Wiebe, William Crown, Nathaniel D. Osgood, "Transforming Healthcare Delivery: Integrating Dynamic Simulation Modelling and Big Data in Health Economics and Outcomes Research", Elsevier Feb 2016.
6. Matthew Herland, Taghi M Khoshgoftaar and Randall Wald, "A review of data mining using big data in health informatics", Springer 2014.
7. Venketesh Palanisamy, Ramkumar Thirunavukarasu, "Implications of Big Data Analytics in developing Healthcare Frameworks – A review", Journal of King Saud University - Computer and Information Sciences, Dec 2017.
8. YichuanWang, Nick Hajli, "Exploring the path to big data analytics success in healthcare", Elsevier, 2016.
9. YichuanWang, LeeAnn Kung, Terry Anthony Byrd, "Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations", Elsevier 2016.
10. Afrati, F.N. & Ullman, J.D. (2011) Optimizing Multiway Joins in a Map-Reduce Environment. IEEE Transactions on Knowledge and Data Engineering, 23(9), 1282-1298.
11. Bakshi, K. (2012) Considerations for Big Data: Architecture and Approach. IEEE Aerospace Conference, (pp.1-7). Big Sky, USA.
12. Gu, R., Yang, X., Yan, J., Sun, Y., Wang, B., Yuan, C. & Huang, Y. (2014) SHadoop: Improving MapReduce Performance by Optimizing Job Execution Mechanism in Hadoop Clusters. Journal of Parallel and Distributed Computing, 74(3), 2166-2179.