IJETMR
International Journal of Engineering Technologies and Management Research
A Knowledge Repository

# BIG DATA ANALYSIS IN HEALTH CARE DOMAIN: A SYSTEMATIC REVIEW

**Abhishek Bajpai [*1], Dr. Sanjiv Sharma [*2]**
[*1] Student of Master of Technology, Department of Computer Science Engineering & Information Technology, Madhav Institute of Technology and Science, Gwalior, India
[*2] Assistant Professor, Department of Computer Science Engineering & Information Technology, Madhav Institute of Technology and Science, Gwalior, India

**Abstract:**
***As the Volume of the data produced is increasing day by day in our society, the exploration of big data in healthcare is increasing at an unprecedented rate. Now days, Big data is very popular buzzword concept in the various areas. This paper provide an effort is made to established that even the healthcare industries are stepping into big data pool to take all advantages from its various advanced tools and technologies. This paper provides the review of various research disciplines made in health care realm using big data approaches and methodologies. Big data methodologies can be used for the healthcare data analytics (which consist 4 V's) which provide the better decision to accelerate the business profit and customer affection, acquire a better understanding of market behaviours and trends and to provide E-Health services using Digital imaging and communication in Medicine (DICOM).Big data Techniques like Map Reduce, Machine learning can be applied to develop system for early diagnosis of disease, i.e. analysis of the chronic disease like- heart disease, diabetes and stroke. The analysis on the data is performed using big data analytics framework Hadoop. Hadoop framework is used to process large data sets Further the paper present the various Big data tools , challenges and opportunities and various hurdles followed by the conclusion.***

**Keywords:** *Big Data Analysis; Data Mining; Machine Learning; Map Reduce.*

**Cite This Article:** Abhishek Bajpai, and Dr. Sanjiv Sharma. (2018). "BIG DATA ANALYSIS IN HEALTH CARE DOMAIN: A SYSTEMATIC REVIEW." *International Journal of Engineering Technologies and Management Research,* 5(2:SE), 1-8. DOI: 10.5281/zenodo.1195065.

## 1. Introduction

To understand what actually big data is, it is the advanced technology to store and analyse the huge amount of the data in the form of the terabytes, petabytes, and Exabyte **[1]** which are not used in the old or manual methods. The big data is one of the buzz words in the information Technology. Storing and analyses of this high volume information or data is to provide the business profit and better decision making process. Big data is characterized by three properties i.e. volume, velocity and variety **[2]**. It represents huge volume of data, many variety of information and velocity at which speed the collected information must be processed. In this

present era, there are 6V's (Volume, Velocity, Variety, Veracity, Validity, and Volatility) of big data, evolving into value of data **[3]**. The data can be structured, unstructured and semi structured.80% of the data is unstructured. Structured data has pre design arrangement i.e. banking data etc. Unstructured data has no predesign arrangement i.e. audio and video files, social websites etc. Semi structured data is the combination of the structured data and unstructured data **[4]**. Traditional searching, sorting and processing algorithms would not able to handle the data in this range, and that too most of them are unstructured. The Big data processing technologies includes machine learning algorithms, natural language processing algorithms, predictive modelling and other artificial based techniques.

Recently, the healthcare industry generated large amounts of data in the form of various health care data like imaging data (CT scan, MRI, angiography, Ultrasonic data etc.), clinical notes derived by record of patient's databases, compliance & regulatory requirements, and patient care using the Wearable sensors, Mobile devices, Genomic Sequences and Social Media etc. In the study of the EMC digital universe study, shows that the health care data is growing at the rate of the 48% of the year **[5]**. The advantages of the Big Data term in the 'Health Care Industry' are that to improve the quality of the health care delivery as well as reducing the cost. The purpose of the Health care data analysis in health care domain is to retrieve and study heterogeneous healthcare data which helps to healthcare providers to deliver right intervention to the acute patient at the right time with minimum possible cost. A simple Definition of the Big Data in medicine is "the totality of data related to patient healthcare and well-being" (Raghupathi 2014). Big data in health care refers to electronic health data set i.e. large and complex medical data that they are difficult (or impossible) to manage with traditional Machine Learning Algorithms or more specifically traditional machine learning infrastructure works on centralised databases; nor can they be easily managed with traditional or common data management tools and methods. But in the case of the Big data may be in the huge volume (in the form of the petabytes, Exabyte) which is not suitable to process on a single machine so we need to improve the traditional algorithm or to develop the new methods or algorithms which can accept these challenges of managing the large amount of the data in the different systems **[6]**.

Applications of big data analytics in health care sector take advantage to extract useful knowledge for making better informed decisions in order to provide the better care to the active patient at the minimum possible cost. When analytics is applied in the context of big data is the process of explore large amounts of data, from a variety of data sources such as- warehouses , databases etc. and in different formats, to deliver knowledge or useful insights that can able to take decisions in real or near real time. Big data analysis in Health care sector takes some various challenges such as- Security, Visualization, Number of data integrity concerns etc.

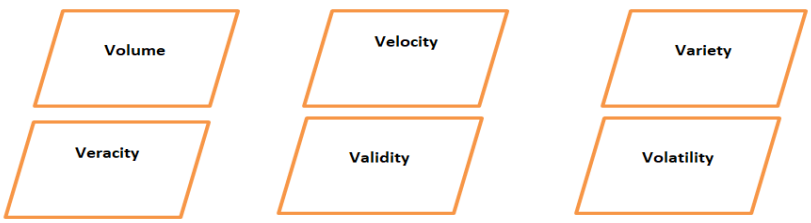


Figure 1: 6v's of the Big Data

## 2. Overview of the Health System

A health system or health care system is the group of people, clinical society, and resources that deliver health care services to the patients to make better decision making system. We described the health care system via five attributes which are given below:

### 2.1. Patient

In the first level of hierarchy, patient is a central entity of the healthcare system because patient's care defines the overall healthcare system. A patient is the person who receives the medical treatment from the group of the organization i.e. Hospital, doctor. Generally, patients play the active role rather than the passive role in the healthcare system because an active patient involves in the all phases of the big data analysis i.e. analysis, design, implementation and maintenance (coordination) of his/her care. Most of the big data technologies set up the patient to share their structured and unstructured data in order to make the good decision making system with possible minimum care cost. It is the central entity because every stages of the healthcare system used this entity as a unique entity.

### 2.2. Health Care Providers

In the second level of the hierarchy, health care providers are everyone who provides the care to the patient i.e. as physicians, doctors, nurses, pharmacies and even family member of the patients to deliver the better care to the patient but this entity should be authorized to practice by the medical rules or state law. Health care providers responsible for stratify the patients depends on his/her disease to deliver the better care. Healthcare providers are also responsible for offering actual treatment and other services to the patients.

### 2.3. Organization

In the third level of hierarchy, the organization that offers the physical existence of the care in the terms of the infrastructure and the other required resources like clinical notes, medicines etc. Organization is the combination of the hospitals, clinic, nursing homes etc. as a result to offer the coordinated care, improve the quality of the patient's care. According to Ferlie and Shortell **[12] [13]**, "organization is a critical level that manages the culture for the care to the patient via better decision-making systems and meaningful human resources.

### 2.4. Health Insurer

The purpose of health insurance is to improve the care of the patients at the real time of care in the form of the finance policy or rules. It protects you and your family financially in the event of an unexpected serious illness or injury that could be very expensive. Health insurers are the third party entities whose provides the insurance schemes to the patient. Buyers are responsible for setting up health plans, selling health plans to patients (individual or group), manage patient care, and manage patient claims. These are the economical environments which assist the patients in the chronic condition.

### 2.5. Pharmacy and the Diagnosis Centre

Pharmacy and diagnostic centres are responsible for disbursement of drugs and diagnostic tests respectively. There are quite some pharmacies and diagnostic centres, which are an integral, part of healthcare providers physically, though technically they might be independent entities.

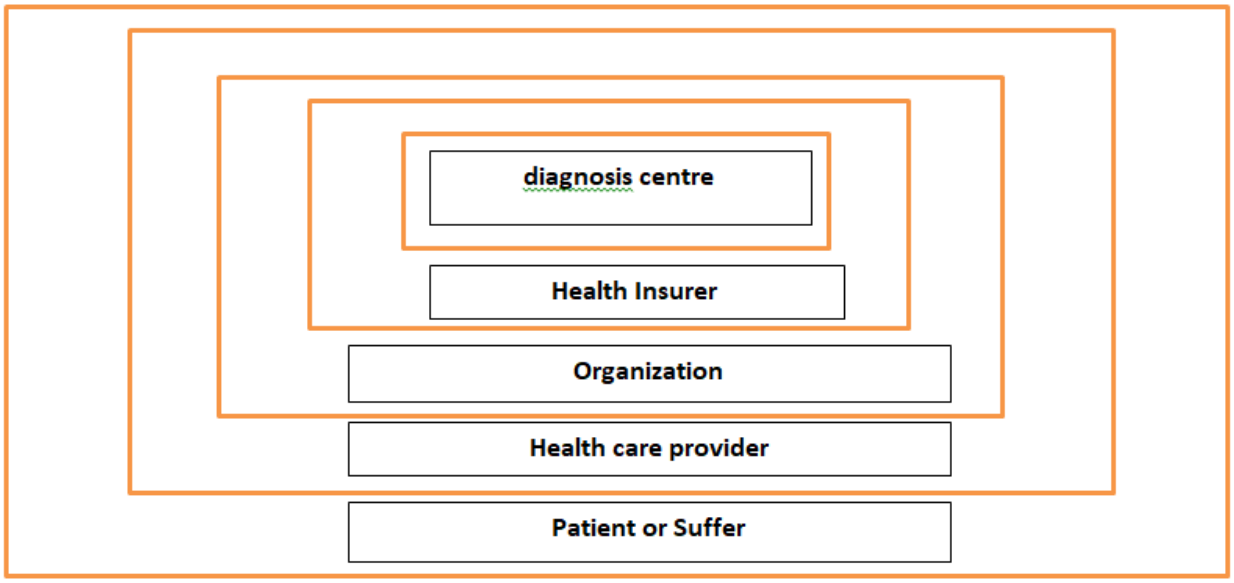


Figure 2: Health Care System

## 3. Review of the Related Work

This section present brief review about the various healthcare streams where Big Data technologies are applied. Big data is stepping onto predictive analysis of epidemiology which is Important to control and prevent chronic disease and morality rates. Hence in this review an effort is made to provide brief information about the ways how the health care domain is benefitting from Big Data analytics technology nowadays.

Jigna Ashish Patel describes **[7]** beautifully how we can think of using Big Data on health care industry by providing the consequences of few surveys done on the usage of Big Data in organization.Liu and Dr E.K Park says that Big Data tools can be used to provide Digital Healthcare services.

If the patent's data and information is exchanged over the network then it becomes most essential to provide the security and safety for the patient's data. Blobel **[8]** says that achieving privacy and security for the patients for their personalized treatment is a challenge for Big and analytics. They say that by using de identification, proper ID management and authentication like single sign on it is possible to establish the trust in digital services. Xindong Wu et al **[1]** proposed HACE theorem that evaluate the features of the big data evolution and proposed a big data processing model using the data mining prospective.

Kiyana Zolfaghar et al **[9]** gives the model which predict the 30 day risk of the Congestive heart failure disease. They developed the scalable data mining models to predict risk of readmission using the integrated data set scalable data mining models to predict risk of readmission using the integrated data set based on the random forest algorithm.

IBM Watson offers cognitive computing powers to be applied to the stored data. Watson's cognitive powers are similar to the powers that humans possess to inform their decisions: Observe; Interpret; Evaluate; and Decide **[10].**

*Stanford Health Care (previously Stanford Hospital* and Clinics) is an academic health system and part of *Stanford* Medicine, which includes the *Stanford* University School of Medicine and Lucile Packard Children's *Hospital Stanford.The drug making process takes 12 years of the research and get the maximum cost. Google research in collaboration with Stanford's Scientist to discover the new drugs with in minimum time and low cost using the machine learning algorithms, neural network, data mining algorithms or advanced technologies like Map reduce on the Hadoop framework [**11]**.*

## 4. Tools & Techniques

| Tools | Features |
|---|---|
| **Hadoop** | Framework for parallel processing in distributed environment on commodity hardware, comprise a set of primitives to perform batch processing. |
| **Pentaho Business Analytics** | Branching into big data by making it easier to absorb information from the new sources. |
| **Karmasphere Studio and Analyst** | Designed to simplify the process of ploughing through all of the data in a Hadoop cluster. |
| **Map Reduce** | It is computational paradigm which works on mapper and reducer functions which can be executed and re-executed on any node in the cluster. |
| **Splunk** | It creates an index of your data as if your data were a book or a block of text. It is just like text search process |
| **Apache Hive** | The tool used for data warehouse infrastructure placed upon Hadoop which help for data analysis as well as querying. |
| **Apache Sqoop** | It is a tool to exchange the data between Hadoop and relational database. |
| **Talend Open Studio** | It is written in Latin language used as high level platform with map reduce program used to analyse larger data set. |

### 4.1. Hadoop

Apache Hadoop is an open source software framework which allows the pre-processing of the large amount of the datasets based on the commodity clusters. Apache Hadoop written in java technology which provides the distributed processing of the big data based upon group of the clusters of the computer systems using simple programs. Apache Hadoop is a software framework for reliable, scalable, parallel and distributed computing [14]. Hadoop is designed to

run on a large number of machines that don't share any memory or disks. Hadoop is the open source platform that enables the processing of the large amount of data. But if we will implement the work using Hadoop. It introduces security threats such as Breach of privacy by unauthorized release of data, manipulation of data in the database and denial of information **[18]**.

### 4.2. Map Reduce

Map reducing is the processing technique or the programming model that allows being performed multiple tasks on multiple systems in a parallel way. Hadoop MapReduce is provided for writing applications which process and analyse large data sets in parallel on large multi node clusters of commodity hardware in a scalable, reliable and fault tolerant manner. Map reduce operates in the two phases i.e. Map phases (Mapper) and Reduce phases (Reducer).It operates on the (key, value) pair of the framework. Mappers are used to process the input data and create the new several chunks of data. And reducer are used to process the data in the form of the (key, value) pair as a result to produce the new set of the output in the list format stored in the Hadoop distributed file format **[15]**. Saravana N et al **[16]** proposed the algorithm in Hadoop/map reduce framework or model to stratify the diabetes types and the type of treatment that to be provided.
Map reduce techniques are used to process the huge volume of the data with high fault tolerance in the parallel way on the various distributed clusters of computer system in the low computational time **[17].**

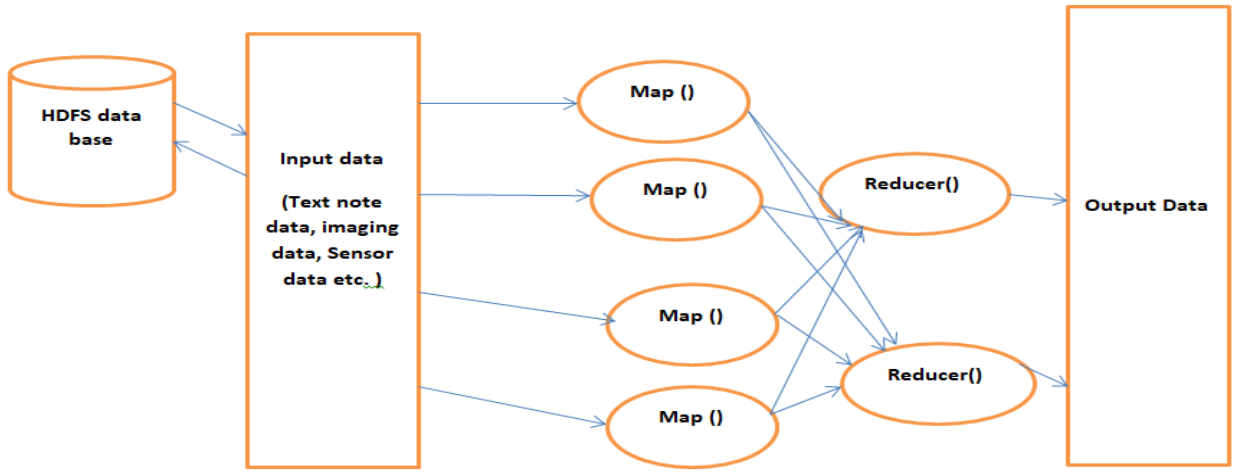

Figure 3: Map Reduce Techniques

## 5. Health Care Big Data Challenges

Big data provides significance with improved performance in the healthcare domain but still several challenges also exist:

1) Clinical notes are difficult to understand in the right manner as data of the clinical reports are in the different format and the medical test result also varies on the basis on the different symptoms of diseases.
2) Handling large volumes of medical imaging data efficiently and extracting potentially useful information. So there is no proper standardization in the health care domain

because massive amount of health related data is generating from different sources such as physicians, social websites and wearable sensor and many other health Intensive care devices.

3) In the health care data, complexity is too high because everyone cannot analyse or process the medical report and genomic data i.e. imaging data, or unstructured data.
4) In the health care domain **[19]**, data are generated from the various heterogeneous sources so it is difficult to interaction between the old system and modern system is become difficult. Like PHR, EHR, Account and labs.

Existing Electronic Health Records are limited to data acquisition than analytics: Data acquisition of EHR is possible but don't have the ability to aggregate, transform, or create actionable analytics from it.

Now days advanced technologies are used in health domain but traditional clinical equipment and physical resource are not compatible to work with them like- hospitals, medical reports, and imaging sensor devices **[20]**. To reduce the incompatibility, middle ware systems are used to match the type of the medical data (structured or unstructured) to the desired compatible type required by the advanced techniques. Big data analysis provide the opportunities to the health care industry in order to aggregate the huge amount of patients care data to understand, classify and make some learning techniques that are used as an alternative treatments to care provider and patient at the time of patient's care to support the clinical decision support system In addition to Big data analysis provides personalized care i.e. insulin injections, DNA sequences for HIV AIDS etc. to the patients using processing data mining or stratify the analytical solutions that are helps to the patients at the real time of care which enables the early detection and diagnosis before a patient develops disease symptoms.

## 6. Conclusion

The role of big data is beyond the description. Healthcare stakeholders have begun experiencing the immense power that data possess. The researches, inventions, innovations and discoveries in the field are incomplete without the medical practitioners realizing their necessities. The promising advantages of big data such as evidence-based diagnosis and drugs; personalized care and treatment; decreased costs; faster and effective decisions can bring value into the lives of not only patients but also caregivers. The healthcare's future is clearly in real-time intelligent decision making from the data. Finally, after looking at the challenges in processing the data by healthcare analysts and researchers, it is proposed that there is a need of a common platform which can be leveraged by all the researchers to pursue common tasks of feature engineering and data preparation. This way more time will be spent on invention rather than on time-consuming task that can be automated.

This paper provide an insight for the upcoming researchers to know the impact of Big data on healthcare and to get an awareness of researches efforts made in healthcare using Big Data so far . This is booming area, there is a lot of scope for the researches to come up with innovative ideas and conclusion to efficiently use the Big data tools in healthcare and also to overcome the challenge of Big data in healthcare and thus help to increase the life expectancy of the people.

## References

[1] David Becker, Bill mcmullen, Trish Dunn King, "Big Data, Big Data Quality Problem", IEEE International Conference on the Big data (Big data ), 2015

[2] Aslam M.A and Abdullah.A "A Methodology and a tool to prepare Agro-meterological maps as source of Big Data" Multimedia Big Data (Big MM) ,2015 IEEE International Conference, Beijing, 2015, pp. 208-211

[3] Khan, M.A., Uddin, M.F., Gupta, N.: Seven V's of big data. In: ASEE Zone 1, Proceedings of 2014 Zone 1 Conference of the American Society for Engineering Education (ASEE Zone 1) (2014).

[4] Xindong Wu, Fellow, Xingquan Zhu,Gong-Qing Wu, and Wei Ding (2014) "Data Mining with Big Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No 1, pp.97-107.

[5] EMC Digital Universe & IDC: The digital universe: driving data growth in healthcare, challenges and opportunities (2004).

[6] Suthaharan, Shan. "Big data classification: Problems and challenges in network intrusion prediction with machine learning." ACM SIGMETRICS Performance Evaluation Review 41.4 (2014): 70-73.

[7] Prof Jigna Ashish Patel, Dr. Priyanka Sharma "Big data for better health planning",IEEE International conference on advances in Engineering & Technology research, August 2014

[8] B.Blobel , D M Lopez & C Gonzalez, "Patient privacy and security concerns on Big Data for personalized medicine", Springer – Verlag Berlin Heidelberg

[9] Kiyana Zolfaghar, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy, Si-Chi Chin and Brian Muckian (2013) "Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure", IEEE International Conference on Big Data Vol.3, No 6, pp. 64-71.

[10] What is IBM Watson http://www.ibm.com/smarterplanet/us/en/ibmwatson/whatiswatsonhtml.

[11] Large-scale machine learning for drug discovery. http://googleresearch.blogspot.in/2015/03/large-scale-machine-learning-for-drug.html

[12] Ferlie, E.B., Shortell, and S.M.: Improving the quality of healthcare in the United Kingdom and the United States: a framework for change. 79, 281–315 (2001)

[13] Sun, J., Reddy, C.K.: Big data analytics for healthcare. In: SIAM International Conference on Data Mining (2013)

[14] Mohd Rehan Ghazia, Durgaprasad Gangodkara: Hadoop, MapReduce and HDFS: A Developers Perspective: International Conference on Intelligent Computing, Communication & Convergence (ICCC-2014).

[15] G.Vaishali, V.Kalaivani : BIG DATA ANALYSIS FOR HEART DISEASE DETECTION SYSTEM USING MAP REDUCE TECHNIQUE.

[16] Saravana N, M Ramachandran and S Lavanya Kumar (2015)," Predictive Methodology for Diabetic Data Analysis in Big Data", ScienceDirect - Procedia Computer Science Vol 50, pp. 203 – 208.

[17] Rama Satish, K. V., and N. P. Kavya. "Big data processing with harnessing hadoop-MapReduce for optimizing analytical workloads." Contemporary Computing and Informatics (IC3I), 2014 International Conference on. IEEE, 2014.

[18] Victor L. Voydock and Stephen T. Kent (1983): Security mechanisms in high-level network protocols. ACM Comput. Surv 15(2):135–171,

[19] Puneet Singh Duggal, Sanchita Paul, ― Big Data Analysis: Challenges and Solutions‖, International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV.

[20] Raghupathi, Wullianallur, and Viju Raghupathi. "Big data analytics in healthcare: promise and potential." Health Information Science and Systems 2.1 (2014): 3