# International Journal of Engineering Sciences & Research Technology

(A Peer Reviewed Online Journal)
Impact Factor: 5.164

IJESRT

RESEARCHERID
THOMSON REUTERS

[Bhattacharjee *et al.*, 9(4): April, 2020]
IC™ Value: 3.00

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

# A BIG DATA ARCHITECTURE FOR HEALTHCARE

**Nidhi Bhattacherjee[1], D. Mustafi [2] & V. Bhattacharjee*[3]**
[1]Department of Computer Science and Engineering, KIIT, Bhubaneswar
[2&*3]Department of Computer Science and Engineering, BIT Mesra, Ranchi

**DOI**: 10.5281/zenodo.3778449

## ABSTRACT

These are testing times for organizations in the healthcare industry as it is transforming at an incredible velocity. The industry has been focusing on delivering better quality care and outcome at better prices in terms of affordability i.e. the industry is focusing more on value. The shift of healthcare towards digitization, ever increasing due to implementation and adoption of digital medical records, is now being utilized and combined with new age technologies like mobile communication, social networking, cloud computing and analytics. This paper presents an overview of applying Big data in health care, a proposed framework for the same and the challenges associated.

**KEYWORDS**: Healthcare, Big data, Analytics, Machine learning.

## 1. INTRODUCTION

Handling Big data is becoming a very imminent need of the era, and several researchers are attracted towards investigating the possibilities of developing new models or improving upon the existing ones. With the increasing amount of data being collected by sensors, the huge number of transactions, social networking sites, it is obvious that new software tools are required to capture, process and manage this data so as to generate action within tolerable time limits. In most cases, processing this data and extracting useful information is a big challenge. Handling big data generates the requirement for a new set of technologies and paradigms. Moreover, the explosive growth of data from non-traditional sources such as IoT sensors, security logs and web applications has made streaming data an important component of enterprise data architecture. It would be very beneficial to act on the data as soon as it arrives. Usually, the term streaming data refers to data that is continuously generated, often in large volumes and at high velocity.

This era of Big data is bringing about a revolutionary change in our perception and the working of health care industry. The foundation for the future is being laid by this digital re-invention. Even individuals and organizations are being empowered by it to improve overall health of everyone. The organization of the rest of this report is as follows. In Section 2 we shall present the related work, in Section 3 we present the proposed Big data architecture for health care, in Section 4 we discuss the challenges. In Section 5 we present a summary and the future scope.

## 2. RELATED WORK

Healthcare and life sciences organizations require precise and improved information on their actual costs, in order to successfully transform from volume to value. It is thus natural that several researchers have been interested in the field of big data analytics in healthcare. Big data and IoT need to be fused with computational intelligence before being converted into nay impactful action. Health care big data has different features since it is not easy to access, is somewhat structured and has legal complications and privacy issues [1-2]. For effective processing of medical big data, machine learning algorithms have been combined with Hadoop [3]. In another paper [4], authors propose a new automated system for the efficient analysis of big data in healthcare technology. They have developed a 5-node distributed Hadoop cluster to extract the features of the distributed system. In [5] authors have applied Random Forest classification technique, with feature selection to EEG dataset. In[6], the authors study the architecture, analytics, developments and functionalities of big data for its tactical enactment in health care industry. According to the content and results of numerous big data analytics

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

cases many capabilities have been pinpointed. In [7], authors study the relationship of big data and IoT. The data collected through sensors, bank transactions, mouse clicks, tweets, social networking sites is continuous in nature and has redundancy as well as noise. Hence mechanisms are required to process this data to make it useful, these are very well presented in [8]. To derive meaningful insights from data, it is essential to know the steps to follow for efficient data analytics. Big Data analysis has been split into four steps: Acquisition, Assembly, Analyze and Action, which are collectively called the 4A's of Big Data analytics. These 4A's have been explained in detail in [9]. The healthcare industry has various sources of big data as hospital records, medical records of patients, etc., giving birth to the requirement of managing this big data. [10] explains this very aspect- big data in healthcare, its management, analysis and future prospects.Further, the management of big data in healthcare systems and its related areas like its architecture and implementation have been explained in detail in [11].

## 3. PROPOSED ARCHITECTURE FOR BIG DATA HEALTHCARE

The characteristics of data collected in the healthcare domain can be listed out as in Fig. 1. Our proposed architecture is developed so as to be able to handle these characteristics.

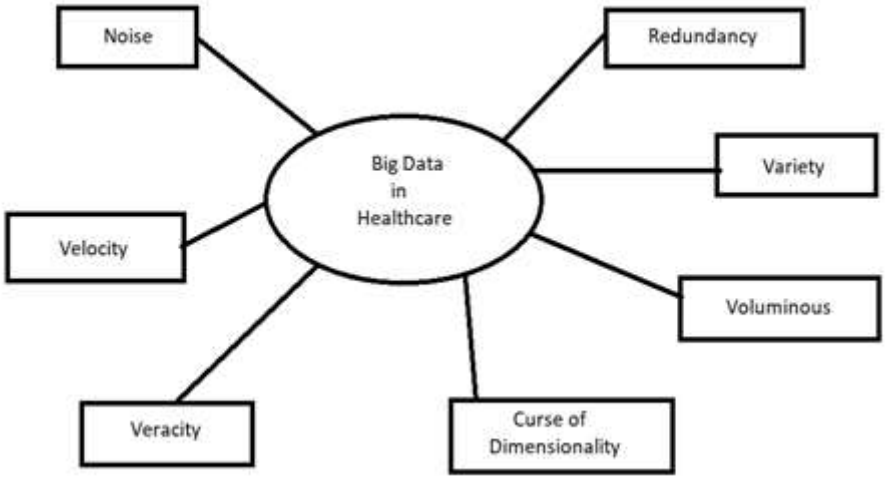


***Fig. 1 Characteristics of Big Data Healthcare***

The proposed system that we are developing is generic and deals with diverse situations like disease diagnosis ahead of time and emergency detection of ailment. The proposed architecture aims to handle medical advancements in disease diagnosis through big data streaming technology by accumulating data from heterogeneous sources having varying formats, and further using machine learning techniques to make meaningful predictions.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

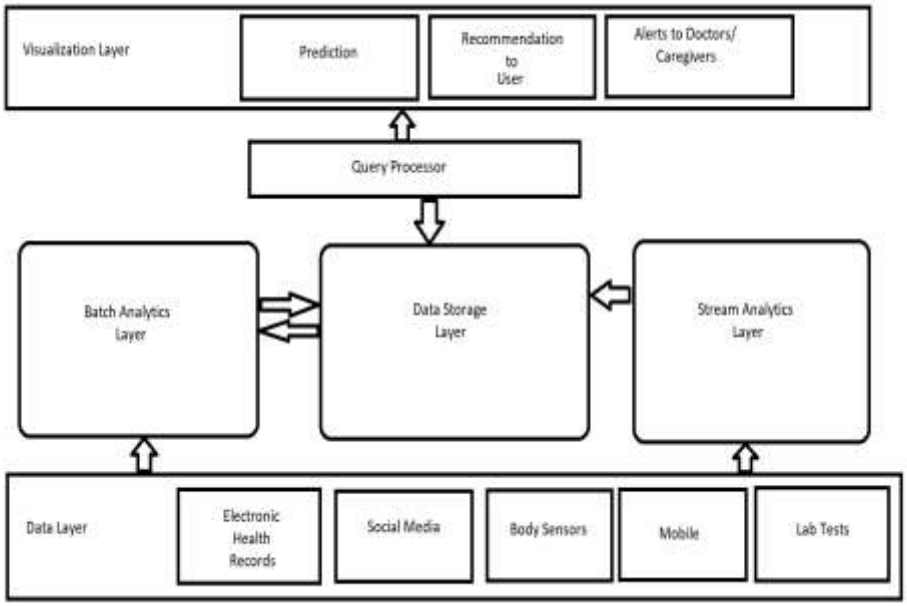


***Fig. 2 Architecture for Big Data Healthcare***

The proposed architecture aims to handle medical advancements in disease diagnosis through big data streaming technology by accumulating data from heterogeneous sources having varying formats, and further using machine learning techniques to make meaningful predictions.

The data is forwarded concurrently to the batch as well as streaming layer present in the analytics layer. After preprocessing, the data is then analysed while passing through various phases including feature selection and extraction. The prepared data in the final stage is further used in design models which are significant for predicting health conditions of a patient in future.

In case of streaming data, data is taken from a number of different sources such as medical sensors attached to the patient's body which provides information regarding the present real time data about the condition of the patient like blood pressure, heart rate etc. The adaptive preprocessor divides the data into blocks and then extracts the required relevant information which is further transmitted to the predictive processor for building predictive models.

Our aim is to provide accurate and precise predictive models which are also cost effective. For this purpose we need to remove the need for extra computational power and storage area requirement which can be achieved by extracting the data in earlier stages so as to send only the relevant data for analysis. This can be done through cleaning and filtering of data.

- Data cleaning: Various data cleaning techniques invlude normalisation, handling missing values, reducing noise etc.
- Data filtering: It separates useful and data relevant to our model from other non-useful data present in abundance. Extraction of useful information is helpful in further processing of data.

The procreation of devices used in the medical field for collecting data is immensely dependent on the features and instances in the medical field which are increasing exponentially. Moreover, it becomes imperative to select the essential features which will be the deciding factors in the designing of predictive models.

Different methodologies are used but the most significant ones are feature extraction and feature selection.

- Feature Extraction: It is focused on reducing the number of attributes that are to be used among thousands of features present in the data acquired.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

- Feature selection: Requires use of statistical data to determine the inevitable features that must be present while designing the predictive model.

Predictive model: This encompasses the objective of building a predictive model which is capable of producing predictions for current observations based on previously stored data accurately. It incorporates the usage of data analysis, statistical data and machine learning techniques.

The data present is vast but the storage space available is significantly too less in comparison which turns out to be one of greatest challenges in big data architecture. In accordance with the proposed model of disease diagnosis and monitoring the data produced by the sensors itself is too large and the results are required on a real time basis else it could result in hazardous or life threatening situations.

Furthermore, since new data is continuously flowing in, we need an adaptive system to maintain the precision and accuracy depending on the incoming newer values. The adaptive system to be formed is associated with adaptive preprocessor and adaptive predictor. In Adaptive preprocessor, sliding window techniques are used to split the incoming data in time windows. If the value of average exceeds a particular threshold immediate alert is generated, else directly sent to the storage layer. The features with highest relevance are transmitted to the predictor. There are constant interactions through message sending and sharing feedback among adaptive preprocessors and predictors regarding the change in values and required updates to be done.

The query processor is responsible for determining the patient's status by mapping the queries sent to batch and stream processing. The user interface and interactions based on the user's view is done through the visualization layer including the display of reports in an uniform manner and show the real time values on the dashboard.

Consider a framework for detecting emergency cases. Such a framework directly gets data from medical sources and detects abnormal situations based on user's thresholds. The streaming system would then send alerts to Data base which would be used to notify doctors about emergencies. Analytics techniques like classification, clustering are adopted for real-time monitoring and online learning to predict whether the current state of patients is danger or not. Machine learning models such as the KMeans clustering model or K-Nearest neighbour classification can be trained with some static data and then used to classify events in a data stream. Having classified a disease related data, specific recommendations can be given to the user or medical assistance can be sought by means of alerts to the doctor or the caregiver.

## 4. CHALLENGES AND ISSUES

This massive change requires the industry to be transparent, in case of both cost and efficiency of care, as well as use the gigantic amount of information related to health today and in the future to derive effective insights. There are several challenges which arise in the field of medical big data. These are as follows:

- Creation of standard terminology and vocabulary for sharing of information.
- Creation of standard representation of information.
- Integrating the different types of information.
- Removing or correcting erroneous data
- Storing the information for re-usability.
- Need to resolve data privacy and security issues.

It is important that the above issues be addressed since the healthcare organizations need precision and better information on the quality and relevancy of the services they deliver.

## 5. FUTURE SCOPE

The massive quantities of medical, genomic and life sciences data, as well as new cognitive systems to derive value from it all, are essentially the reason why the healthcare industry is on the verge of new learnings and discoveries. And the practitioners and researchers are discovering and applying the answers in innovative ways, all due to the help from the new tools in analysis and gleaning of relevant information from this whirlpool of data. This paper proposes a big data framework for healthcare, which comes with several challenges as

discussed above. The standardization of terminology for data collection and representation can be done by joint collaboration of the healthcare organization and the data scientists. The machine learning algorithms which we propose to implement in our analytics layer would handle the missing or erroneous data. Finally, as software engineers, it is our duty to apply our knowledge of big data and computational intelligence, ethically and with social concerns when dealing with any kind of data especially healthcare data, thus taking care of the privacy and security issues. If this becomes true, patients can become partners in monitoring their own wellness. This would lead us to a future of good health for one and all.

## REFERENCES

[1] Raghupati, W. and Raghupati, V. 2014. Big Data analytics in healthcare: Promise and Potential. Health Information Science and Systems, Springer, 2(3), pp. 2-10.
[2] 2. Lee, H. C. and Yoon, H.-J. 2017. Medical Big Data: Promises and Challenges. Kidney Research and Clinical Practice, 36(1), pp. 3-11.
[3] Lakshmi, J. V. N. and Sheshasaayee, A. Machine Learning approaches on Map Reduce for big data analytics. 2015 International Conference on Green Computing and Internet of Things, Noida, 2015, pp. 480-484
[4] Yao. Q., Tian Y., Li P. F. et al. Design and development of a medical big data processing system based on Hadoop. The Journal pof Medical Systems, vol.39, no. 23, 2015.
[5] Jha R., Bhattacharjee V., Mustafi A. (2020) IoT in Healthcare: A Big Data Perspective. In: Pattnaik P., Mohanty S., Mohanty S. (eds) Smart Healthcare Analytics in IoT Enabled Environment. Intelligent Systems Reference Library, vol 178. Springer, Cham.
[6] Pal, C, Motwani, D. Patel, A. Survey of health care industries by study the architecture and functionality of big data analytics. International Journal Of Engineering Sciences & Research Technology, 7(8), August 2018, 618-626.
[7] Tulasi, B. and Vemulkar, G. J. Blending IoT and Big Data Analytics. International Journal Of Engineering Sciences & Research Technology, 5(4), April 2016, 192-196.
[8] Gubbi, J., Buyya, R. Arusic, S., Palaniswami, M. Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions. Future Generation Computer Systems 29 (2013) 1645–1660,Elsevier
[9] Nimankar, S. and Dagare, S. Big Data Analytics:4 A's. International Journal Of Engineering Sciences & Research Technology, 7(2), February 2018, 328-332.
[10] Dash, S., Shakyawar, S., Sharma, M. and Kaushik, S. Big data in healthcare: management, analysis and future prospects. Journal of Big Data, 6:54, 2019, 1-25
[11] El aboudi, N. and Benhlima, L. Big Data Management for Healthcare Systems: Architecture, Requirements, and Implementation. Advances in Bioinformatics, Volume 2018, Article ID 4059018, 10 pages.