



18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

Big Data Analytics in Healthcare – A Comprehensive Literature Review

Fatima Zahra¹, Banu M²

Sr. Inventory Specialist, Target Corporation, Karnataka, India¹,

Assistant Professor, St. Joseph's College, India²

ABSTRACT: Big data is gigantic measures of data that can do some incredible things. It has gotten a subject specifically compelling for as long as two decades in view of a high potential that is covered up in it. Different open and private part ventures create, store, and break down huge information to improve the administrations they give. In the social insurance industry, various hotspots for huge information incorporate emergency clinic records, clinical records of patients, aftereffects of clinical assessments, and gadgets that are a piece of the web of things. Biomedical examination additionally creates a critical bit of enormous information pertinent to open medicinal services [4]. This information requires legitimate administration and examination to determine important data. Something else, looking for an answer by breaking down large information rapidly gets tantamount to finding a needle in the pile. There are different difficulties related with each progression of dealing with huge information which must be outperformed by utilizing very good quality registering answers for huge information investigation. That is the reason, to give significant answers for improving general wellbeing, social insurance suppliers are required to be completely outfitted with proper framework to produce and examine huge information methodically. Effective administration, examination, and understanding of large information can change the game by opening new roads for present day human services. That is exactly why different ventures, including the human services industry, are finding a way to change over this potential into better administrations and budgetary focal points. With a protected mix of biomedical and social insurance information, present day human services associations can upset the clinical treatments and customized medication.

KEYWORDS: Healthcare, EHR, Omics, Hadoop, Data integration, Digital Imaging

I. INTRODUCTION

Healthcare is a highly data-concentrated industry. The enormous growth of the health care volume can be observed by increasing trend. It was foreseen that the data of the U.S.A healthcare sector would reach the Yottabyte. The exponential growth in healthcare data has been forecasted to continue expanding in various forms, such as electronic health records (EHR) [2], patient-reported outcomes, medical imaging, genomics, biomarker and biometric data information [4]. Some of the primary sources of healthcare data are medical service providers, pharmaceutical industries, public healthcare organizations, and medical insurance etc. The Integration of data from Electronic Medical Record (EMR) and some biochemical and hormonal analyses, leads to the disclosure of new symptomatic and helpful devices. Both structured (EHR data) and unstructured data (MRI images) are present in the healthcare sector which substantiates that technologies used in healthcare can handle a variety of data range. [7] Healthcare data consists of four primary attributes: volume, velocity, variety, and uncertainty or veracity, which together constitute the notion of Big Healthcare Data. These properties cause a variety of challenges for the users of healthcare data and for technologies used. The few significant challenges involved in Big Healthcare Data are data integration, data processing and analysis. Out of all, Integration of unstructured data, in particular, is a considerable obstacle, mainly due to the heterogeneity of data.

We will study the developments in advanced solutions for big healthcare data integration and present the design and construction of integration strategies that are commonly adopted by the research communities. We will discuss current challenges concerning Big Healthcare data integration. Furthermore, we will discuss some possible future research directions in this area.

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

II. LITERATURE REVIEW

1. "In Search of Big Medical Data Integration Solutions - A Comprehensive Survey"

Authors Houssein Dhayne, Rafiqul Haque. Rima Kilany, Yehia Taher

Publisher: *IEEE Access*, vol. 7, pp. 91265-91290, 2019

The paper focuses on finding strengths and weaknesses of existing Big Data technologies, tools and applications in various operations of Healthcare sector such as data collection, data integration, processing and analysis and their efficiency over the recent decades. The limitations were found in existing Hadoop based model technologies which work on NoSQL and parallel programming.

2. "Big Data Analytics for Healthcare Industry: Impact, Applications, and Tools"

Authors: Sunil Kumar and Maninder Singh

Publisher: TUP (Tsinghua University Press), Volume: 2, [Issue: 1](#), March 2019

To study the tools available in Hadoop ecosystem for handling the huge data and understanding the conceptual architecture of big data analytics involving electronic health records, text and image records and others. Authors also proposed the use of a conceptual architecture for solving healthcare problems on basis of Hadoop technologies, which involves the utilization of the big data, generated by different levels of medical data and developing effective methods for data analysis to obtain solutions for healthcare queries. However these proposed solutions could not solve the problems associated with big data as big data contains big systems and is in early stages of development.

3. "Big Data Analytics in Healthcare"

Authors: Ashwin Belle, Raghuram Thiagarajan, S.M.Reza Soroushmehr, Fatemeh Navidi, Daniel A. Beard and Kayvan Najarian

Publisher: *Journal of Biomedicine and Biotechnology*, January 2015

To discuss on upcoming areas of medical research, medical image analysis, physiological signal and integration of physiological data with genomic data and their impact on healthcare. These three areas face challenges in dealing with structured and unstructured big data sources. Author suggests further research in the area of big data, aggregating and coordinating uninterrupted and distinct medical data formats, developing signal processing techniques specialized towards physiological signals are essential to overcome the challenge.

III. EXISTING SYSTEM

In recent advances in big data for health informatics and their role to tackle disease management are presented, for instance, diagnosis prevention and treatment of several illnesses. Clinical context produces unstructured data or a semi-structured from data such as handwritten doctor notes. These data may have differences in meanings and interpretations. The process of big data in the healthcare industry is broken into five stages: Data Acquisition, Data storage, Data management, Data Analytics, and Data visualization and report. The further content explains each one of the mentioned stages:

Big data in healthcare organizations consist of both internal (Patient's health history) and external data (Third party data/data from public providers) [3]. Both of these can be housed on cloud computing [2]. The data management process includes, storing, organizing, maintaining, retrieving, data mining, data monitoring and data validating. Data Analytics is broadly a process of converting raw data into information. Big data analytics in healthcare is segregated into Descriptive, Diagnostic, Predictive, and Prescriptive Analytics. Data Visualization is defined as graphical representation of analytics result obtained from healthcare data analysis used for better understanding of the correlation of data. Big-data analytics in current framework system is processed by clustering and scanning multiple nodes of clusters in the network.

In Hospital Network, NoSQL database is used to collect and manage the enormous amount of real-time data from diverse sources, which assists the management in administering high-risk patients and reducing everyday expenditures. In terms of Monitoring of Patient's vital information, use of Hadoop-based components in the Hadoop Distributed File System (HDFS), including the Impala, HBase, Hive, Spark, and Flume frameworks are common, to convert the unstructured data generated using sensors which take patient's vital signs. Using Hadoop, healthcare staff can analyze these unstructured data. Hadoop technology supports the healthcare intelligence applications. Hadoop ecosystem's Pig, Hive, and MapReduce technologies process large datasets related to medicines and other factors to extract meaningful

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

information for medical institutes. Companies, with help of Hadoop application, have begun to use a prediction model to determine the scams and fraud committers before the action is taken place.

IV. PROBLEM STATEMENT

Data integration: This is caused by two major reasons of data heterogeneity where in the data is sourced from different source describe the same domain using various schemas and data disparity which refers to incorrect or incomplete and often outdated data. [2]

Data Quality: Collation of heterogeneous data from varied sources causes low quality, meaningless medical information and sometimes loss of data as well.

Real time responses: Currently used large and complex algorithm leads to blockage in data processing

V. PROPOSED SOLUTIONS

Solution for Integrating Medical Data Streams and Data quality: A unified web based system can be developed having the features of not just in taking Unicode but also image, audio and voice format.

Level 1 Solution: Inclusion of “Integration Bucket” functionality in landing zone. This concentrates on Data filtering which maintains high quality by removing all redundant data. Data collected from various sources are directed to landing zone causing repetition of data in cloud. To avoid this, the adding a functionality of “Integration” to remove repeated and redundant data can save time and complexity. Such a solution is tremendously useful for performing a critical analysis on the fly. (Figure 1)

Solution for integration with real-time responses: Most of the existing solutions are built on the Hadoop-MapReduce framework, which mostly solved the data volume challenge. However, due to the extravagant sorting algorithm that Hadoop relies heavily on for performing causes the performance to bottleneck. Based on our study we concluded that this problem can be resolved by incorporating it with cloud computing

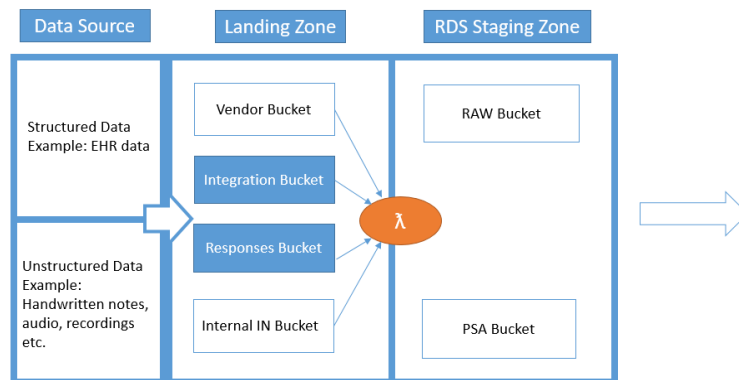


Figure 1. Components in Cloud Computing

Level 2 Solution: Inclusion of “Responses Bucket” function in landing zone. Real time responses occur only when there are real time interactions. Interactions are not unique; repeated interactions can be captured in landing stage with the use of a pointer to call particular function of real time response (Figure 1). This will help save tremendous amount of processing time. Responses are processed and to give out a relevant solution, relevant interactions can be captured by this function to avoid duplication.

All the filtered solutions are then provided to LAMBDA bucket. Since LAMBDA is python based programming, it reduces the number of operations. The data is retrieved using SQL query.

VI. CONCLUSION

In this paper, we have studied description and overview of big data in healthcare system thoroughly, which plays a significant influence on healthcare system. Data was collected from the work published during the year 2018 to 2020. This paper sketches the significant features of the healthcare big data, techniques, the applications, effectual

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

management of the healthcare big data. The advent of Big Data promoted the growth of Big Data technologies for various operations, including collection, integration, storage, analysis, management and visualization. Our study focused on finding the areas of limitations and challenges and working toward them consequently. A company working largely around big data, can implement a centralized location to store the distributed data through Cloud Computing. Advancement in cloud computing can lead to solving major problems in big data.

The most challenging part of big data in healthcare are data privacy, data leakage, efficient handling of large voluminous data. We proposed the use of a conceptual architecture for solving some of the complications in big data using Hadoop-based terminologies. It utilizes the big data to generate different levels of medical data and develop methods for analyzing its analysis obtaining solutions to medical problems. This combination can lead to producing treatments that are patient-centric. As the study shows, big data analytics are in their premature state of development and current tools and methods cannot cope up with existing problems of the healthcare data. Therefore, a great deal of research and invention is required in the field of technology to strengthen and advance the big data in healthcare sector.

VII. FUTURE WORK

This paper reviewed the published works of literature on big data in healthcare. This paper reviewed some of the most challenging areas of big data in healthcare. Few key points of the paper are understanding unstructured clinical data in the right context, extracting potentially useful information to utilize them to create tools and their management for the advancement of the healthcare system and improve the overall medical experience. Some limitations have been listed below:

- *Heterogeneous data and unstructured data:* Health care produces a varied type of data. These data may be diverse in terms of Data types and data model [5]. Clinical context produces unstructured data or a semi-structured form such as handwritten doctor notes, images that may have differences in meanings and interpretations
- *Patient privacy and data security:* Essential factors related to patient privacy and consent and other legal issues are taken into consideration during the data integration process. These regulatory aspects could create a possible roadblock to the proper implementation of the data integration
- *Non-expert users:* Given the scale and heterogeneity of medical data, automatic integration is not entirely accurate. End-users of integrated data are generally physicians, nurses, and health professionals with limited training in the field.

This reading addressed the major experiments on data governance in healthcare. Big data has a great perspective to progress healthcare management and transform the healthcare industry to the next level. The review article will be benefiting the healthcare academicians, practitioners, researchers who are engaged in the areas of healthcare Management.

REFERENCES

- [1] S. Nazir et al., "A Comprehensive Analysis of Healthcare Big Data Management, Analytics and Scientific Programming," in IEEE Access, vol. 8, pp. 95714-95733, 2020, doi: 10.1109/ACCESS.2020.2995572.
- [2] H. Dhayne, R. Haque, R. Kilany and Y. Taher, "In Search of Big Medical Data Integration Solutions - A Comprehensive Survey," in IEEE Access, vol. 7, pp. 91265-91290, 2019, doi: 10.1109/ACCESS.2019.2927491.
- [3] Senthilkumar SA, Bharatendara K Rai, Amruta A Meshram, Angappa Gunasekaran, Chandrakumarmangalam S. Big Data in Healthcare Management: A Review of Literature. American Journal of Theoretical and Applied Business. Vol. 4, No. 2, 2018, pp. 57-69. doi:10.11648/j.ajtab.2018.0402.14
- [4] Sai, B & Jyothi, Singaraju. (2015). A STUDY ON BIG DATA MODELING TECHNIQUES. B. SAI JYOTHI , S. JYOTHI.
- [5] Ezhilraman, S.V. & Srinivasan, Sujatha. (2018). State of the art in image processing & big data analytics: Issues and challenges. International Journal of Engineering and Technology(UAE). 7. 195-199. 10.14419/ijet.v7i2.33.13885.
- [6] Dash, S., Shakyawar, S.K., Sharma, M. et al. Big data in healthcare: management, analysis and future prospects. *J Big Data* 6, 54 (2019). 10.1186/s40537-019-0217-0
- [7] Belle, Ashwin & Thiagarajan, Raghuram & Soroushmehr, S.M.Reza & Navidi, Fatemeh & Beard, Daniel & Najarian, Kayvan. (2015). Big Data Analytics in Healthcare. BioMed Research International.
- [8] Sunil Kumar and Maninder Singh , "Big Data Analytics for Healthcare Industry: Impact, Applications, and Tools" ISSN 2096-0654 05/ 06 pp48-57 Volume 2, Number 1, March 2019 DOI: 10.26599/BDMA.2018.9020031



18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

- [9] How Data Visualization Is Transforming the Health Care Industry by Erin McCoy Using Visual Analytics, Big Data Dashboards for Healthcare Insights by Jennifer Bresnick
[10] Explore the application of Big Data technologies in Healthcare by Aishwarya Anantharam