IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Research on the Secure Medical Big Data Ecosystem Based on Hadoop

Sudhir Allam, *Sr. Data Scientist, Department of Information Technology, USA*

Abstract*—The application of big data strategies to increase the efficiency of healthcare delivery has become unavoidable owing to the overwhelming quantity of data in the healthcare system. Big Data Analytics like Hadoop plays an important role in doing meaningful real-time analyses on massive amounts of data and predicting emergencies before they occur [1]. The Electronic Health Record (EHR) is a computer database that stores important details from patient records. Because of government directives and technical advancements, the volume of data in the EHR is growing. Sensors and diagnostic records are used to document patient details. Given the massive quantities of data sets in the EHR, appropriate methods for storing and analyzing this data for practical interpretations are needed. This paper aims to address the uses of big data in the healthcare sector. The usage of medical big data for secondary uses is becoming more widespread in healthcare and medical studies. Understanding the rationale underlying medical big data reveals trends in healthcare information management which has significant implications for hospital information systems that are developing and improving medical systems [1]. This paper investigates a secure medical big data ecosystem built on Hadoop to expand the sophistication of the medical system. The framework in this paper incorporates healthcare information based on the Hadoop big data network, with the original centralized on data being processed and evaluated internally through a modern networking system and an autonomous data acquisition system. The customized health information framework for diseases has been developed to offer personalized health care systems for patients and to promote the treatment of patients by medical professionals by using the capabilities of the Hadoop big data system.*

Keywords: Medical data, Hadoop, Big Data, Healthcare, Map-Reduce, Hadoop Distributed File System (HDFS)

I. RESEARCH PROBLEM

The problem that this research will look at solving is how to improve the medical systems utilizing big data analytical techniques like Hadoop. Many issues face the healthcare sector in terms of utilizing data to diagnose diseases and predict a disease outcome. These challenges can be solved by understanding the significance of implementing a medical system based on

Hadoop Big data analytics. Cased of errors and lack of medical know-how in disease prediction will be a thing of the past. The problem that this study would aim to resolve is how to address different medical problems in medical care by using technology development methods such as Hadoop to build a stable infrastructure. Despite the incorporation of big data analytics techniques and frameworks into traditional data processing structures for healthcare systems, such architectures struggle to mitigate emergencies. This paper explains how we can get more benefit out of the data produced by public health care. The healthcare institutions produce a significant volume of heterogeneous data. Such data, however, became worthless due to a lack of proper data analytics techniques [2]. This paper would examine how the deployment of a stable medical big data ecosystem built on the Hadoop big data model would enhance medical care and services to enhancing the sophistication of the medical system. Additionally, the paper suggests how a comprehensive clinical management system that helps people to consider their care and recovery condition whenever and wherever, and all primary healthcare data is spread in separate independent medical databases. Medical practitioners will more effectively create their MapReduce software to operate on a Hadoop platform that scales from a single node to multiple machines. The main aim of the paper is to look at a clear overview of a Hadoop-based medical big data analysis framework.

II. INTRODUCTION

Big Data in healthcare is derived from vast electronic health datasets, which are extremely challenging to manage using traditional hardware and software [3]. The usage of legacy data collection systems and software still prevents the usefulness of all this data. Big Data analytics is a daunting phenomenon, not just about the sheer amount of

data, but also because of the various forms of data and the intensity at which healthcare data processing must be handled. The "Big Data" challenge in the healthcare industry is the overall total of data relating to the client and their health. Data Analytics is now becoming an increasingly obvious concern in health systems. By addressing novel problems as a result of data processing, database architecture, data compiling, and clinical decision making, healthcare informatics leads to the advancement of Big Data analytic technologies [4].

In today's digital environment, digitization of these records is needed. To increase the efficiency of healthcare while reducing costs, a vast amount of data must be accurately analyzed to respond to emerging problems. Every day, the government still produces petabytes of data. It necessitates technologies that allow for real-time review of the vast data collection. This would assist the government in providing residents with value-added facilities. With the assistance of deep learning techniques, big data analytics aids in the discovery of useful choices through understanding data structures and their relationships [4]. The clustered algorithms display potential in comparison with single nodes to facilitate effective data analysis using medical big data. Furthermore, with the usage of medical big data analytics, patient information systems may be configured to be even more knowledgeable and user-friendly by providing customized suggestions. This paper aims to provide an analysis of big data analytics in medical systems. It addresses how these structures produce big data, data features, reliability concerns in processing big data, and also how big data analytics assists in obtaining a holistic assessment of these data sets.

III. LITERATURE REVIEW

A. Healthcare as a big-data repository

Healthcare is a multi-faceted infrastructure created for the primary purpose of preventing, diagnosing, and treating health-related conditions or impairments. Health practitioners have a range of records, including personal health records (including diagnoses and prescription medicines), research and clinical details (including data from imaging and lab tests), and other private or confidential medical details. In the past, it was standard practice to hold a patient's medical records in the context of handwritten documents or printed files [6]. Also, medical test reports were contained in a paper filing system. In reality, the earliest case records can be found in an Egyptian papyrus text dating back to 1600 BC. The digital technology of all therapeutic examinations and medical reports in health systems is becoming a common and generally accepted procedure nowadays, due to the invention of electronic systems and their ability. Electronic health records (EHR) have many benefits when it comes to processing current healthcare results [6].

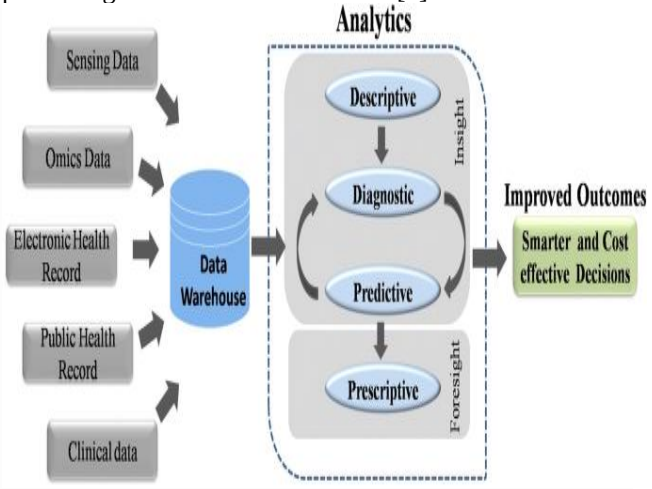

Fig i: Data inflow in medical systems

Health diagnosis, medications, details about documented allergies, demographics, health histories, and the outcomes of different laboratory tests are among the records [6]. As a consequence of the shorter lag time between previous test findings, recognizing and treating medical problems is more effective. We've seen a substantial reduction in unnecessary and extra tests, as well as missed instructions and ambiguities created by incorrect data, and better patient management between different healthcare professionals, over time [6,7]. By minimizing prescription dosage and frequency mistakes, such logistical errors have culminated in a decline in the number of drug allergies. Healthcare workers have since learned that through digital alerts and reminders for vaccines, irregular test findings, cancer tests, and other routine checkups, they will dramatically enhance their medical practices.

B. Hadoop

Hadoop is a major data analysis framework. Working with big data involves loading vast volumes of (big) data into the secondary storage even for the most efficient computer clusters. As a result, the only logical way to analyze massive amounts of large big data is to share and process it in parallel through many nodes. Even so, data is normally so massive that it requires lots of processing devices to transmit and process it in a reasonable timeframe [7]. While dealing with hundreds or even thousands of nodes, challenges such as concurrent processing, data transmission, and malfunction management need to be solved. Hadoop is one of the most commonly deployed open-source distributed software for this task. Hadoop uses MapReduce to process and generate massive datasets. The map and reduce algorithms are used in MapReduce to convert each theoretical entry in the inputs into a series of intermediate digital certificates, with the minimized process combining all the variables that use the same key. It parallels calculations effectively, addresses errors, and manages inter-system coordination through vast machine clusters [7].

C. Components of Hadoop

1. Hadoop Distributed File System (HDFS)

HDFS is a storage framework that helps for scaling, efficiency, and replica-based storage systems across multiple nodes in a cluster. HDFS is a file system built for holding massive files with streamed shared data trends and operating on commodity hardware clusters [9]. It is a framework with a master/slave file design, as well as a single name node or the master server that handles the log files registry and monitors user permissions. Furthermore, the cluster's storing partitions the data into "blocks," which are then stored repetitively around the virtual server. The HDFS structure includes 3 versions within each file, which are referred to as the name node or master node, the data node, and the HDFS users or edge nodes [9]. Name Node is the centralized node that displays details regarding the Hadoop file system. The HDFS name node serves as a master node by storing details regarding the current systems, like metadata and attributes, as well as the precise position of files. This always contains details regarding nodes that have been inserted, updated, or deleted. Data nodes operate as slave nodes for other nodes. As name nodes order blocks, data nodes usually store and extract them. The collections of blocks that were processed by the data nodes are regularly sent out to the name node. Hadoop promotes fault tolerance by operating a secondary name node and utilizing data backup (prolonged status of the command line metadata) [10].

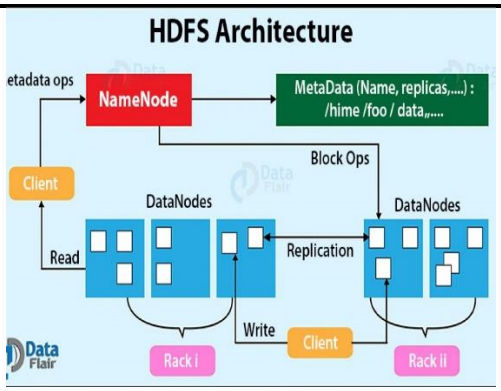

Fig ii: Hadoop Distributed File System (HDFS) architecture

2. *MapReduce Architecture*

MapReduce is a Hadoop computing system that serves two distinct tasks: map and reduce [10]. For batch data analysis, MapReduce is a commonly used solution. The data is broken down into tiny chunks and scattered through several nodes to retrieve intermediate results, which are then combined to provide final results. The first stage of MapReduce is the Map Task, which scans and executes a set of data to generate digital certificates as intermediate outputs [11]. The performance from the mapping (key-value pairs) is fed into the Reducer in a reduced task. The reducer, on the other side, gets the key-value pair from different charts. Finally, the reducer incorporates all of the intermediate public keys into a single final result.

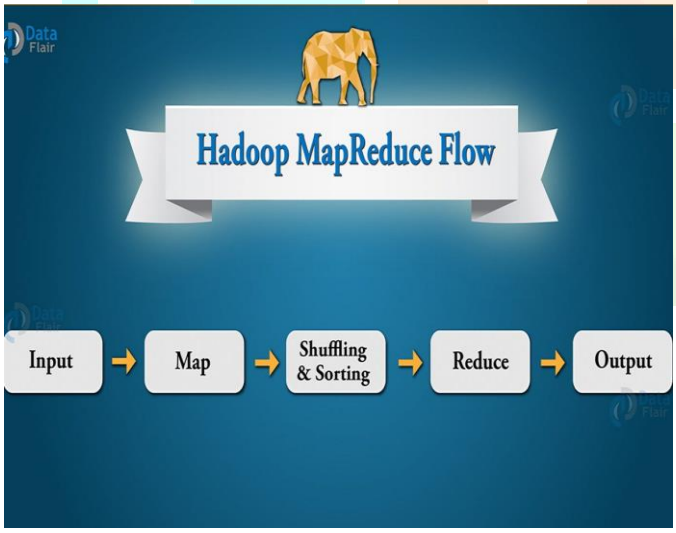

Fig iii: MapReduce data flow

Hadoop has other technologies that boost the storage and computing elements, which have been easily embraced by several big corporations such as Yahoo, Facebook, and many others [11]. Hadoop has allowed researchers to deal with data sets that would have been difficult to manage otherwise. Hadoop is being used in various applications, including determining a connection between pollution levels information and asthma diagnoses, drug discovery utilizing genomic and proteomic data, and other areas of medicine. As a result of the Hadoop system's deployment, healthcare analytics could not be set back.

D. *How Hadoop Supports Secure Medical Big Data Ecosystem*

Hadoop deployment in the healthcare data management infrastructure allows data repositories to store and interpret organized and unstructured data for effective patient treatment. Hadoop is a distributed data management and analytics technology that is open source [12]. It is a computing platform that handles both organized and unstructured data. It is not a data center in the traditional sense. Hadoop distributes vast volumes of data through many computing nodes before combining the output. Since the device is dealing with smaller batches of regional data rather than the whole warehouse's contents, this method enables data to be handled more quickly [12,13]. Most healthcare organizations are also searching for the most effective means of large data analytics to optimize patient services and allow them to take part in predictive analytics and public health management.

The term "structured data" refers to information that is contained in a certain format, such as a spreadsheet [1]3. Since it has appropriate limits and is generated and processed consistently, structured data is simpler to interpret and process. Clinical information details, diagnostic and intervention information, prescription codes, and other data from the electronic medical record are usually created in a consistent, coordinated manner. Structured data is normally managed by standard data centers.

Healthcare organizations can experience further problems as a consequence of unstructured data. Unstructured data will take several types, from texts, voice recordings, images, text files, and social media network updates, to name a few [14]. Unstructured data is amorphous and cannot be processed in the same manner as structured data can. Since several EHRs support unlimited text feedback for medical documentation as well as other descriptive data collection areas, they present a unique challenge to medical institutions.

To be evaluated, unstructured data must be extracted, processed, and normalized. The extraction requires time which is an unnecessary cost for organizations that could be on a limited budget. Hadoop's distributed data method could be able to assist. Hadoop uses the Hadoop Distributed File System (HDFS) and MapReduce to store and process data [14]. Hadoop implementations use HDFS as their main distributed storage. HDFS isn't a physical database; alternatively, it captures knowledge and saves it in clusters before an entity can utilize it. Hadoop divides unstructured data into nodes, which are independent components of a broader data system. The nodes are connected and will integrate the data stored inside to generate results based on criteria defined by an entity.

The data is analyzed using MapReduce. MapReduce is a collection of Java applications that pull data from Hadoop clusters when required [15]. Using Hadoop as part of the complete center helps organizations control and analyze data that was historically inaccessible. It's possible that fully integrating Hadoop into a data warehouse would necessitate server upgrades. Scalability and availability problems may be avoided by including more on-premise servers or looking for a hybrid storage system. Hadoop is a big project, but companies can think of what type of data they'll be analyzing and if their existing database will support it [15].

When incorporating a modern approach into their infrastructure, healthcare institutions must still weigh cost-effectiveness. To enjoy the advantages of a solution like Hadoop, companies must be completely invested and trained. Standard systems and data centers have not been utilized and can yet be applied easily in Hadoop hybrid technologies [16].

E. *Hadoop's disease prediction*

Diagnostic reports must be checked by a psychiatrist, who also predicts the ailment that the patient is suffering from depending on the observations of the reports, but this may be unreliable at times. Because all of these disorders

present with almost identical signs, neurological syndromes may be confused for mental disease [16]. This demonstrates that manual diagnosis is not often effective and precise, necessitating the use of Hadoop and HBase to computerize disease prediction. Since a series of report guidelines, symptoms or statistics are described for each illness and are kept in a mega database termed HBase, Hadoop may also be used for simplification utilizing ML classification of data from EHRs of patients who can then be categorized by disease [16]. MapReduce is often used in this context to measure the number of people in a hospital that are recovering from a particular illness over time.

IV. ITS FUTURE IN THE U.S

The future of the U.S medical system will depend on the improvements made in how data is stored and analyzed to improve medical outcomes. The implementation of Hadoop in the U.S medical system will be beneficial in boosting how diseases are predicted and therefore the future will involve increase implementation of the framework. Healthcare practitioners in the United States would be required to combine conventional expertise in care, teamwork, and leadership with advanced skills in technology and analytics in the future. Value-based treatment would become a more feasible alternative as technology advances, enabling medical providers to properly monitor the continuity of care, special packages, and migrate to Electronic Health Records (EHRs) [17].

Kaiser Permanente takes the lead in the United States and may give the EU a blueprint to pursue. They've completed the implementation of a framework named HealthConnect, which shares data from all of their locations and simplifies the utilization of health records. A McKinsey Big Data Healthcare report suggests, the automated approach has increased cardiovascular disease performance and gained an unprecedented $1 billion in cuts from fewer office visits and laboratory testing. Medical imaging is important and about 600 million imaging operations are carried out in the United States per year [17]. The manual processing and preservation of these pictures is both money and time-costly because radiologists have to review each picture separately, while hospitals have to preserve it for many years.

Continued developments in healthcare are therefore going to influence the management of a myriad of patient needs. Telehealth — the opportunity to access treatment conveniently and easily through technologies such as smartphones, tablets, and online media — represents an enormous change in the current healthcare landscape. Ease of access to healthcare services is enhanced, since people who do not reside in the proximity of a healthcare facility may utilize telehealth to communicate daily [17]. Additional treatments may be provided from home via video calling to people who need support for certain illnesses, like communicative or mental health conditions. Precision medicine often extends to the preventive field, where patients will choose a lifestyle to remain healthy as well as monitor their physical health. As the future of clinical technology becomes increasingly patient-centered, health providers will be using existing tools such as wearables to capture and interpret individual health data so that clients can directly understand and use them.

V. ECONOMY BENEFITS

Medical costs in the United States are high, almost twice as high and continue to rise exponentially in much of the other developing countries. The unsustainable planned trend of the costs of US health insurance has contributed to demands for better health care coverage. The Affordable Care Act, though, has been blamed for not doing more to control rates, despite being the most important regulatory change of US health care in decades [17]. The opportunity to minimize healthcare expenses has become a huge motivation for hospitals to engage in Big Data since the United States pays much more on healthcare than the rest of the world on comparable or in certain instances worse outcomes. Its healthcare costs accounted for about 18 percent of the gross domestic product in 2009 [18]. The application of big data analytics such as Hadoop in healthcare needs saves time, resources, and money. Whenever the supply chain of a medical facility is compromised or disrupted, everything from patient services and medication, to long-term funding and beyond is likely to suffer [18]. In other words, the next major data in the medical field emphasizes the significance of analytics to make the supply chain smooth and effective from start to finish. The use of analytical instruments to follow the measurements for the supply chain and to take reliable, data-driven decisions on procedures and expenditure will save up to $10 million in hospitals each year.

Because of increasing healthcare prices in the United States, there is a massive demand for big data analytics. According to a McKinsey report: after over 20 years of continuous rises, health care expenditures already make up 17.6 percent of GDP — about $600 billion higher than anticipated for a country of scale and prosperity like the United States [19]. In other words, the costs are significantly bigger and have risen for the last 20 years. This sector will need some intelligent, data-driven thinking. And existing encouragement is also changing: more businesses move from fee-for-service programs to plans that give priority to medical results (they are rewarding using costly and often needless procedures and easy handling of vast numbers of patients).

VI. CONCLUSION

In this paper, the Big Data Analytics application in the healthcare system is briefly outlined and different processes are proposed to capture the effect of old analytical procedures in this revolutionary era. This paper suggests the concept of utilizing the Hadoop– driven Big Data Analytics to identify and forecast patients' illnesses as quickly as possible, and even the illness a person has by simultaneous treatment and distributed machines, MapReduce and HDFS, etc. The paper also addressed the outcomes of introducing the same. Consideration of a Hadoop-scale database solution is an essential first move for the stable development of an institution's health IT systems. Healthcare providers are continuing to look for better ways of treating patients by gathering and analyzing as much information as possible. Organizations that gather data for both patients and staff will more readily recognize situations where changes are required and counterproductive initiatives can be minimized.

References

[1] A. Ergüzen and E. Erdal, "An Efficient Middle Layer Platform for Medical Imaging Archives", *Journal of Healthcare Engineering*, vol. 2018, pp. 1-12, 2018.

[2] C. Anastasopoulos, M. Reisert and E. Kellner, "Nora Imaging": A Web-Based Platform for Medical Imaging", *Neuropediatrics*, vol. 48, no. 01, pp. S1-S45, 2017.

[3] S. Sakr and A. Zomaya, "Editorial for Special Issue of Journal of Big Data Research on "Big Medical/Healthcare Data Analytics", *Big Data Research*, vol. 13, pp. 1-2, 2018.

[4] D. Chrimes, "Interactive Big Data Analytics Platform for Healthcare and Clinical Services", *Global Journal of Engineering Sciences*, vol. 1, no. 1, 2018.

[5] M. Fan and S. Xu, "Massive medical image retrieval system based on Hadoop", *Journal of Computer Applications*, vol. 33, no. 12, pp. 3345-3349, 2013.

[6] M. Supriya and A. Deepa, "A Survey on Prediction Using Big Data Analytics", *International Journal of Big Data and Analytics in Healthcare*, vol. 2, no. 1, pp. 1-15, 2017.

[7] D. Chrimes and H. Zamani, "Using Distributed Data over HBase in Big Data Analytics Platform for Clinical Services", *Computational and Mathematical Methods in Medicine*, vol. 2017, pp. 1-16, 2017.

[8] E. Kolker and E. Kolker, "Healthcare Analytics: Creating a Prioritized Improvement System with Performance Benchmarking", *Big Data*, vol. 2, no. 1, pp. 50-54, 2014.

[9] B. Ristevski and M. Chen, "Big Data Analytics in Medicine and Healthcare", *Journal of Integrative Bioinformatics*, vol. 15, no. 3, 2018.

[10] M. Ghazi and D. Gangodkar, "Hadoop, MapReduce, and HDFS: A Developers Perspective", *Procedia Computer Science*, vol. 48, pp. 45-50, 2015.

[11] D. Peter Augustine ―Leveraging Big Data Analytics and Hadoop in Developing India's Healthcare Servicesl, International Journal of Computer Applications, Vol 89, No.16, March 2014.

[12] S. Sakr and A. Elgammal, "Towards a Comprehensive Data Analytics Framework for Smart Healthcare Services", *Big Data Research*, vol. 4, pp. 44-58, 2016.

[13] H. Zhou, T. Liu, F. Lin, Y. Pang, J. Wu and J. Wu, "Towards efficient registration of medical images", *Computerized Medical Imaging and Graphics*, vol. 31, no. 6, pp. 374-382, 2007.

[14] S. Sakr and A. Elgammal, "Towards a Comprehensive Data Analytics Framework for Smart Healthcare Services", *Big Data Research*, vol. 4, pp. 44-58, 2016.

[15] D. E. Laxmi Lydia and M. Srinivasa Rao, "Applying compression algorithms on Hadoop cluster implementing through apache tez and Hadoop MapReduce", *International Journal of Engineering & Technology*, vol. 7, no. 226, p. 80, 2018.

[16] Q. Yao, Y. Tian, P. Li, L. Tian, Y. Qian, and J. Li, "Design and Development of a Medical Big Data Processing System Based on Hadoop", *Journal of Medical Systems*, vol. 39, no. 3, 2015.

[17] M. Zhang, R. Zhang and C. Liu, "Design of Smart Healthcare Data Management System Based on Hadoop", *Advanced Materials Research*, vol. 998-999, pp. 1121-1124, 2014.

[18] Y. Rochd and I. Hafidi, "Performance Improvement of PrePost Algorithm Based on Hadoop for Big Data", *International Journal of Intelligent Engineering and Systems*, vol. 11, no. 5, pp. 226-235, 2018.

[19] J. Griss, Y. Perez-Riverol, S. Lewis, D. Tabb, J. Dianes, N. del-Toro, M. Rurik, M. Walzer, O. Kohlbacher, H. Hermjakob, R. Wang and J. Vizcaíno, "Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets", *Nature Methods*, vol. 13, no. 8, pp. 651-656, 2016.