JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PERFORMANCE OF BIG DATA ANALYTICS FOR INTERACTIVE HEALTHCARE PREDICTION AND RECOMMENDATION SYSTEM

C K Indira
Assistant Professor of CSE
G Pullaiah College of Engineeiring and Technology, Kurnool,AP,India

K Jashwanth Kiran
IV B.Tech II Sem
Student, G Pullaiah College of Engineering and Technology,Kurnool

P Mahesh Reddy
IV B.Tech II Sem
Student,G Pullaiah College of Engineering and Technology,Kurnool,

B Madhan
IV B.Tech II Sem
Student, G Pullaiah College of Engineering and Technology,kuenool AP,India

T Sai Siddharth
IV B.Tech II Sem
Student,G Pullaiah College of Engineering and Technology,Kurnool AP,India

ABSTRACT: BDA is a rapidly expanding discipline that has begun to become crucial to the development of medical procedures and academic analysis. In the realistic world, where significant amounts of data are occasionally gathered, the healthcare business is a necessary component. The healthcare business generates a wealth of informative data. For data mining and to extract business insights, an in-depth big data approach is required. Big data analytics has many advantages in the healthcare industry, including the ability to identify serious illnesses early on and provide better medical care to the appropriate patient at correct time to enhance quality of life. The effectiveness of BDA is presented in this analysis by an interactive healthcare prediction and recommendation system. The software for big data analytics in interactive healthcare is evaluated for a variety of simulated patient inquiries, employing data ingested into the Hadoop file system and several Apache Spark and web-based Zeppelin apps. The findings indicated that Hadoop required about two hours to process one billion records. Compared to Spark and Zeppelin, Hadoop Distributed File System (HDFS) performed better. When making predictions, MAE (Mean Absolute Error) value is taken into account for patient and physician suggestions.

KEYWORDS: Big Data Analytics (BDA), Healthcare, recommender system, web-based interfaces.

I. INTRODUCTION

Rising prevalence of chronic illnesses, the expanding population, and the difficulty to analyze and extract useful details from a variety of health-related data sets are a few primary drivers of innovation adoption in the healthcare industry nowadays. [1].

Big data is a new innovation that has the potential to upgrade the existing healthcare system and advance it in various ways [2]. The medical sector is experiencing several issues in the current world. First of all, it is quite challenging to acquire and analyze the enormous amount of data, both organized and unstructured. It implies that information is gathered from various resources. While unstructured data must be treated using various methods, structured data may be processed using standard tools. The need for skilled employees to evaluate this information is another difficulty. The vast amount of clinical data that is currently being produced comes from a variety of sources, including body area sensors, cell phones, patients, medical facilities, investigators, practitioners, and associations.

Electronic Health Records (EHRs) [3], Genomic Sequencing (GS), Medical Imaging (MI), Pharmaceutical Research (PR), Clinical Records (CR), wearable technology as well as Medical Devices (MD) are just a few of the possible formats in which the data gathered from diverse sources will be presented. For advanced analysis, these massive clinical information is kept specifically in Medical Server (MS), Clinical DataBase (CDB) and various Clinical Data Repositories (CDR). To make it simpler for people, the storage technologies are largely employed to store, mechanism, evaluate, handle and extract the enormous volumes of information. As a result, it serves to not only teach about symptoms, illnesses, and medications but also to warn, forecast results early, and help people to make the best choices.

BDA is a software featuring an analytical architecture that includes interfaces for users to enter requirements on a web-client application to get information from Big Data sets [4]. Big Data mining is a technique used to draw out important and worthwhile information from huge datasets. From the large data, valuable information is uncovered, including obscure patterns, unknown correlations and the likes. BDA is a combination of two unique ideas. Together, they constitute a novel approach to information management that aims to extract previously undiscovered knowledge and insights from data in order to answer a number of novel and important issues.

In order to use a huge amount of data for healthcare, Health Big Data may be massive, complicated, dispersed, and extremely diversified [5]. Health Big Data enables healthcare suppliers and specialists to utilize analytical techniques on the BDA to evaluate health services, in-patient patient transportation, in-hospital obtained diseases, evidence-based medicine, mechanisms and diagnoses connected to actual data with clinical results, such as scientific and clinical findings. Children's Mercy Hospital in Kansas City, US, implemented a BDA platform called Constellation for the treatment of feasibly incurable disorders, and it successfully correlated children's patient data with complete genome sequencing. In an emergency, the variable diagnosis for a genetic condition in infants might be determined in 50 hours, according to their BDA study [6]. With additional Hadoop improvements, the whole genome sequencing analysis time was reduced from 50 to 26 hours. However, despite these examples of fully functional BDA systems that have been successful in bioinformatics, there are minimal analyses and reports that describe how BDA platforms are utilized by healthcare providers to analyze hospital and patient records [7].

The remaining parts of the analysis are arranged as follows. Literature reviews are included in Section 2, the BDA for interactive healthcare methodology is described in Section III, the result analysis is explained in Section IV, and the conclusion is explained in Section V.

II. LITERATURE SURVEY

Fuad Rahman, Marvin Slepian, Ari Mitra, et. al. [8] demonstrates a revolutionary big-data framework for medical applications. The variety, accuracy, and amount of clinical information make it a good candidate for bigdata computation and analytics. They outline a new method for creating a big data architecture that may be readily used for a variety of healthcare applications, thereby creating "Big-Data-Healthcare-in-a-Box." Xiao Li, Reza Sharifi Sedeh, Liao Wang, Yang Yang, et. al. [9] to connect the two datasets de-identified hospitals, they first did hospital matching. Next, they execute patient matching, but only on two databases of medical records that come from the same institutions. The suggested method is very simple to deploy on BDA platforms like Hadoop.

Cavoukian, Ann, Michelle Chibba, Graham Williamson, Andrew Ferguson et. al. [10] emphasized the significance of privacy concerns in the big data world and covered how Attribute-Based Access Control (ABAC) might shield sensitive data from exploitation. They stated industries like medical, insurance as well as aviation, among others, may profit from ABAC.

TK, Ashwin Kumar, Hong Liu, Johnson P. Thomas, and Goutam Mylavarapu, et. al. [11] Sensitive Data Elements in Hadoop Identification. In the existing Hadoop deployment, only file-level access control is practical, and sensitive material may only be found by humans or using the data owner's details. Privacy can be affected whenever sensitive data is utilized by an unauthorized user or utilized improperly by an authorized individual. The process of manually recognizing sensitive data items is automated in this analysis, which is the initial component of the anticipated access control architecture for Hadoop. The suggested approach makes use of data provenance, usage trends, and context to detect these data objects. The suggested framework also has the ability to follow the history of the data.

Hasani, Zirije, Margita Kon-Popovska, and Goran Velinov, et. al. [12] applied the lambda framework to stream data processing. They created a real-time service to use the data in real-time and give the user with the query view flawlessly. The batch layer of the system as it is now executed repeatedly uses the MapReduce function. Between the batch layer powered by hadoop and the speed layer powered by storm, there exists a merge step. The query view will be provided employing all of the data from each level.

U. Srinivasan and B. Arunasalam, et. al. [13] outline two cutting-edge systems that utilize big data to identify trash, fraud, and other problems in claims for health insurance. This reduces recurring losses and enables better patient care. The findings show that claim anomalies found by these applications assist private medical insurance funds in recovering hidden cost increases that cannot be found by transaction processing devices. This work is a component of a special edition on utilizing business analytics as well as big data.

Zeng, Wenrong, Yuhao Yang, Bo Luo et. al. [14] Content-Based Access Control (CBAC) paradigm was discussed and put into practice. The access control decisions are based on the dynamic similarity between the available material and the requester's personal information. It is a content-centric approach to information sharing. The paradigm allows for appropriate access control choices to be made with little overhead because all users are members of a special role which has access to specific data categories.

Jai-Andaloussi. S, Elabdouli. A, Chaffai. A, Madrane. N, & Sekkaki. A, et. al. [15] implementing MapReduce distributed computing architecture and HDFS storage framework will help to solve the problem of content-based picture retrieval systems. BEMD-GGD approach (Bidimensional Empirical Mode Decomposition with Generalized Gaussian Density functions) as well as the BEMD-HHT approach (BEMD with Huang-Hilbert Transform (HHT)) are two techniques employed to describe the image content. To calculate the distance among image signatures and determine how similar two images are, they compare the BEMD-GGD signatures using Kullback-Leibler Divergence (KLD) and HHT signatures using Euclidean distance. They have confirmed that the findings from the studies upon that Digital Database for Screening *Mammography* (*DDSM*) image database are encouraging, and such study has permitted users to confirm viability as well as effectiveness of using Content-based image retrieval (CBIR) in huge databases of clinical image.

III. BIG DATA ANALYTICS FOR INTERACTIVE HEALTHCARE

The block diagram of Performance of BDA for Interactive Healthcare Prediction and Recommendation System is represented in below Fig. 1.

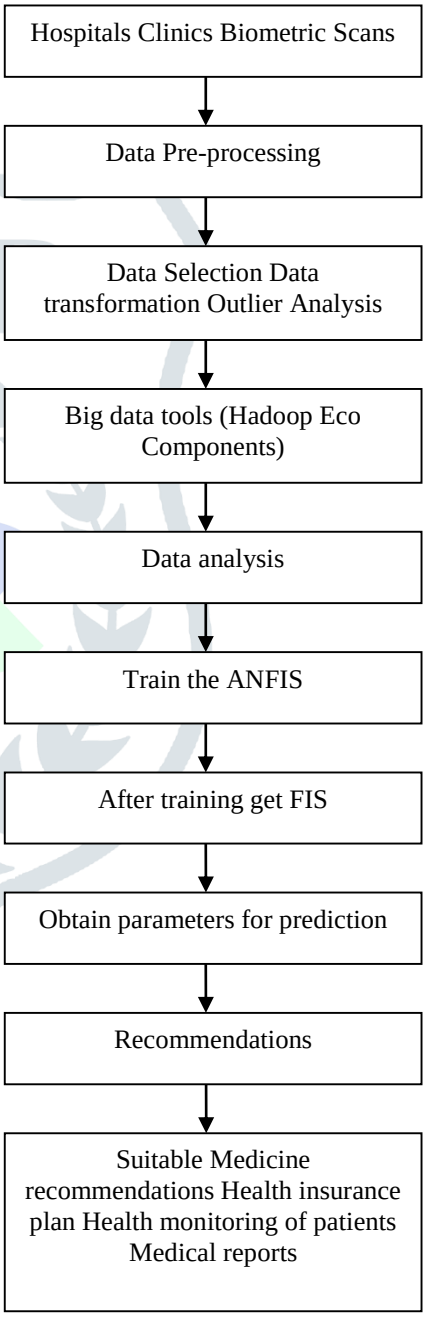

Fig. 1: OVERVIEW OF THE RECOMMENDER SYSTEM USING BIG DATA ANALYTICS

The construction of recommender systems uses machine learning. BDA is the term used whenever big data is taken into account. Recommender systems go through several phases. Information

gathering, learning, and prediction are among the stages. The procedure uses information from data sources like clinics, hospitals, and biometric scanning. For reporting results and analytics, the capability to query information is a crucial duty. First, it needs to fix the interoperability issues that restrict query tools from obtaining the whole information repository used by the healthcare sector. The pre-processing of the gathered data includes data selection, data transformation, and outlier assessment.

The act of finding and erasing erroneous health-related records is known as data cleaning. Since the data must exist in a structured manner in order to execute the appropriate analysis, the data gathered from sensors, doctor's prescriptions, medical picture data, and social media data is frequently not incorrectly formatted. Removing as well as adding the missing values is a constant difficulty during this stage. For example, the data retrieved might comprise medical images (such as MRI (Magnetic Resonance Imaging), CT (Computerized Tomography), PET (PolyEthylene Terephthalate), and ultrasound) and in this condition, data retrieval is frequently application-dependent and extremely challenging to filter relying on its structure. For the purpose of conducting a meaningful analysis, these information must be divided into structured, semi-structured, and unstructured.

The next important step is to deploy big data platform for implementation and evaluation of big data. Now the question arises, which platform for big data needs to utilize for analytical study of data. This process uses Hadoop Eco components big data tools. HDFS is the main data storage file system that manages data processing and storage for massive health data applications running in clustered healthcare systems. The enormous healthcare information collection is divided up into smaller pieces and dispersed among many medical servers. It fulfills the dual roles of data controller as well as processing tool, holding enormous promise for helping businesses manage data that has previously been difficult to manage. The data is either structured or unstructured. Extremely large volumes of data are processed using Hadoop. Support for the distributed Hadoop platform comes from the surrounding ecosystem of additional platforms and technologies.

Data analysis performed for various reasons may produce leads for recommender engines, which produce suggestions. The purpose of data analysis is to extract meaningful information from CR sets employing a variety of analytical techniques and technological tools. For diagnosis, the analysis makes use of medical image recommendations for detection of patients of higher risk patterns of drug use. Data analysis employs cooperative filtering strategies. There are a lot of benefits. They consist of rapid access to medical information, cloud-based data search that makes use of accessibility and scalability, effective service models, a patient-centered method, and more.

The Adaptive Neuro Fuzzy Inference System (ANFIS) architecture has various levels. A neural network is a series of algorithms that intends to intimidate the human brain into discovering relationships in such a data set. The layers deal with membership grades, rule strength, firing strength normalization, function computations and output aggregation. The estimation outcomes are produced in the format of recommendations in the health industry, as shown in the ANFIS framework. The recommender system takes into account suggestions for suitable medications and health insurance plan health monitoring with medical reports. This approach will help caregivers or medical professionals administer the required care to stop any more issues. In the last round, suggestions for enhancing the standard of patient care would be gathered from both patients and physicians.

IV. RESULT ANALYSIS

With many stakeholders participating, the Interactive Healthcare Prediction and Recommendation System is reviewed and found to be effective. The collection of 500 doctors' assessments includes data on around 10,000 patients. Additionally, physicians were divided into five groups. The performance indicator used is the MAE. Its accuracy is higher when the MAE value is smaller. For patient and physician suggestions, the MAE value is taken into account. Errors among matched observations reflecting the identical phenomenon are measured by MAE. Table 1 presents the findings.

Table 1: PERFORMANCE OF THE INTERACTIVE HEALTHCARE PREDICTION AND RECOMMENDATION SYSTEM

Number of Parties	MAE Value	
	Patients	Doctors
1	0.68	0.84
2	0.65	0.80
3	0.69	0.85

4	0.71	0.82
5	0.77	0.81

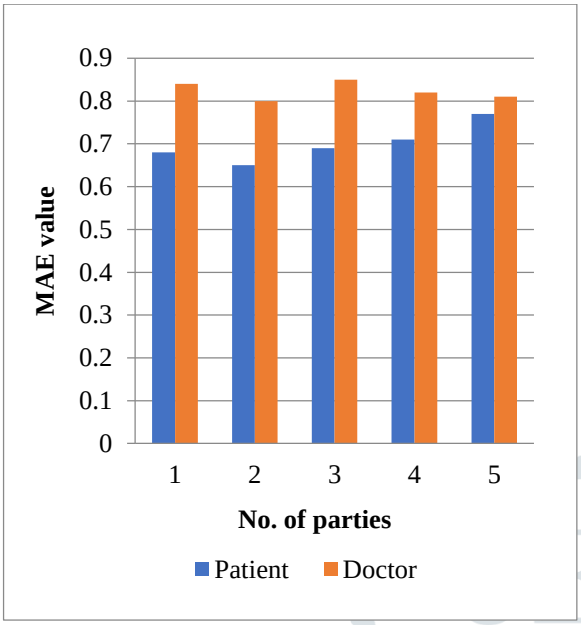

Fig. 2: PERFORMANCE ANALYSIS

The described recommender system performed better, according to the findings. Two CSV (Comma Separated Values) files were imported into HDFS: one with one billion records and one with 50 million records, this is used for the actual benchmarking. Repeated executions of identical query using Zeppelin and Apache Spark showed that the performance times got faster over time. The core of the Hybrid BDA platform's processing activities and filters is Apache Spark, with its action filters, transformations, and Resilient Distributed Dataset (RDD). Zeppelin is another user interface for interacting with Spark. A web-based notebook which was independent of Spark was called Zeppelin. Through its interpreter mechanism, it accommodates a wide range of back-end procedures. The outcomes for ingestion times are shown in Table 2. If they set T_i(N) as time to receive N records, consequently, the data Ingestion Effectiveness (IE) was :

$$Ingestion\ Efficiency = \frac{1B \times T_i(50M)}{50M \times T_i(1B)} \dots (1)$$

As with Spark and Zeppelin, consuming and utilizing CSV files on Hadoop had its benefits (e.g., simplicity, frequent CSV import and exports in medical applications, quick ingestion). Spark, however, was costly to operate.

Table 2: INGESTION TIME ANALYSIS

Tool	Ingestion time	
	50 Million records (23 GB)	1 Billion records (451 GB)
HDFS	6 min	2h 4 min
Apache Spark	8 min	2h 30min
Zeppelin	9 min	2h 40min

The outcomes showed that there was an MAE disparity between estimations for patients and doctors. In addition, accuracy decreases as MAE increases. Another finding is that efficiency improves with more collaboration between parties. Fast ingestion time improves described Interactive Healthcare Prediction and Recommendation System connecting BDA.

V. CONCLUSION

The effectiveness of BDA for an interactive healthcare prediction and recommendation system is discussed in this analysis. Making smarter decisions is a huge benefit of BDA for the healthcare sector. It therefore focuses on the potential distinctive traits, various stages, and analytical techniques of big health data. The outcomes showed that MAE determines patient and physician recommender systems fperformance. The prediction model's accuracy decreases with more MAE data. Hadoop also has benefits from ingesting and utilizing CSV files (For example, CSV imports and exports are frequently used in healthcare applications, and ingestion is quicker than with Spark and Zeppelin). Fast ingestion time improves described Interactive Healthcare Prediction and Recommendation System connecting big data analytics.

VI. REFERENCES

[1] Swapnil Shrivastava, T K Srikanth, “A Dynamic Access Control Policy for Healthcare Service Delivery in Healthcare Ecosystem using Electronic Health Records”, 2021 International Conference on COMmunication Systems & NETworkS (COMSNETS), Year: 2021

[2] Jeong Hyeon Han, Joo Yeoun Lee, “Digital Healthcare Industry and Technology Trends”, 2021 IEEE International Conference on Big Data and Smart Computing (BigComp), Year: 2021

[3] Oluwaseyi Ajayi, Meryem Abouali, Tarek Saadawi, "Secure Architecture for Inter-Healthcare Electronic health records Exchange", 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Year: 2020
[4] Keh Kok Yong, Mohamad Syazwan Shafei, Pek Yin Sian, Meng Wei Chua, "Review of Big data analytics (BDA) Architecture: Trends and Analysis", 2019 IEEE Conference on Open Systems (ICOS), Year: 2019
[5] P. Saranya, P. Asha, "Survey on Big data Analytics in Health Care", 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), Year: 2019
[6] M. Ambigavathi, D. Sridharan, "Big data analytics in Healthcare", 2018 Tenth International Conference on Advanced Computing (ICoAC), Year: 2018
[7] Mark Sterling, "Situated Big data and Big data analytics for healthcare", 2017 IEEE Global Humanitarian Technology Conference (GHTC), Year: 2017
[8] Fuad Rahman, Marvin Slepian, Ari Mitra, "A novel big-data processing framwork for healthcare applications: Big-data-healthcare-in-a-box", 2016 IEEE International Conference on Big Data (Big Data), Year: 2016
[9] Xiao Li, Reza Sharifi Sedeh, Liao Wang, Yang Yang, "Patient-record level integration of de-identified healthcare big databases", 2016 IEEE International Conference on Big Data (Big Data), Year: 2016
[10] Cavoukian, Ann, Michelle Chibba, Graham Williamson, Andrew Ferguson "The importance of ABAC: attribute-based access control to big data: privacy and context." Privacy and Big Data Institute, Ryerson University, Toronto, Canada (2015)
[11] TK, Ashwin Kumar, Hong Liu, Johnson P. Thomas, and Goutam Mylavarapu, "Identifying Sensitive Data Items within Hadoop" 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems, pp. 1308-1313. IEEE, 2015
[12] Hasani, Zirije, Margita Kon-Popovska, and Goran Velinov, "Lambda Architecture for Real Time Big Data Analytic", ICT Innovations, Vol. 15, pp. 133-143, 2015
[13] U. Srinivasan and B. Arunasalam, "Leveraging Big Data Analytics to Reduce Healthcare Costs", IEEE Computer Society, IT Professional, 2013, 15(6), pp. 21 – 28, 2013
[14] Zeng, Wenrong, Yuhao Yang, and Bo Luo, "Access control for big data using data content." In 2013 IEEE International Conference on Big Data, pp. 45-47, IEEE, 2013
[15] Jai-Andaloussi, S., Elabdouli, A., Chaffai, A., Madrane, N., & Sekkaki, A. (2013, May). "Medical content based image retrieval by using the Hadoop framework", In Telecommunications (ICT), 2013 20th International Conference on (pp. 1-5), 2013.