Utilizing Extensive Data in the Health Care Management System

Dr. Nischith. S[1], Dr. Rakesh D[2]
[1]*BIMS, University of Mysore, Mysuru, India.*
[2]*JSS – Centre for Management Studies, JSS Science and Technology University, Mysuru, India.*

***Abstract*—Over the past few decades, the healthcare industry has experienced remarkable growth. The healthcare sector offers medical, preventative, consoling, and other services to patients. As a result, enormous amounts of data are produced. This data was kept on paper in the distant past, when the computers were not in use. But fast digitalization is the present trend. Big data is made up of a lot of data that can't be handled and stored in the conventional way. This article primarily examines big data deployment strategies in the healthcare industry. The difficulties and resources for implementing big data in the healthcare industry are also hinted to in this research. Additionally, it covers various machine learning algorithms that are helpful in upholding the trade-off between accuracy and efficiency.**

***Index Terms*—big data, analytics, healthcare, machine learning.**

I. INTRODUCTION

Massive amounts of data are generated every day via social media, etc. The question now arises as to how large the data must be in order for it to qualify as big data. Depending on the contents it is used for, a tiny amount of data may also be seen as big data. The term 'Big Data' itself has a numerical connotation. Big data does not have to be limited to sizes measured in gigabytes, megabytes, or any other unit of measurement. For example, if we were to try to attach a 200-megabyte document, the email system would not allow it. This is because attachments of this size are not supported by email systems. Thus, big data can be used to describe the size of the document in relation to email. In the healthcare industry, "big data" refers to

A. Overall Analysis of Data in the Health care Sector

electronic health records, or EHRs, which are digital copies of patient charts that are quickly and securely accessible to authorized users.

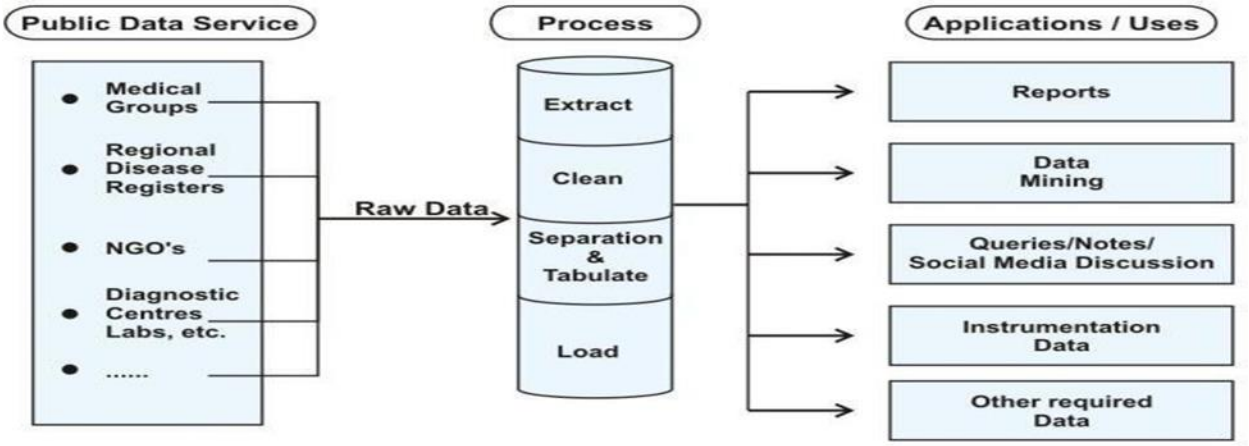
A patient's medical history, diagnosis, prescriptions, treatment plans, dates of vaccinations, allergies, radiological images, test results, and other information can all be found in their health records. Access to evidence-based tools can also be obtained, which clinicians can use to make decisions regarding a patient's care.[2]

II. 3 V'S OF BIG DATA

Volume: Volume is the total amount of generated and stored data, or the trash data. We generate over 2.5 quintillion bytes of data documents per day [15]. Terabytes are used to describe the amount of data that is generated, processed, and stored in various forms such as documents, files, transactions, and records. Every second, an online transaction is completed.

Velocity: Calculating velocity involves determining how quickly data flows. When data is handled in batches of velocity, it indicates that it is processed in batches of 20, for example. The data is not processed instantly in almost real time.

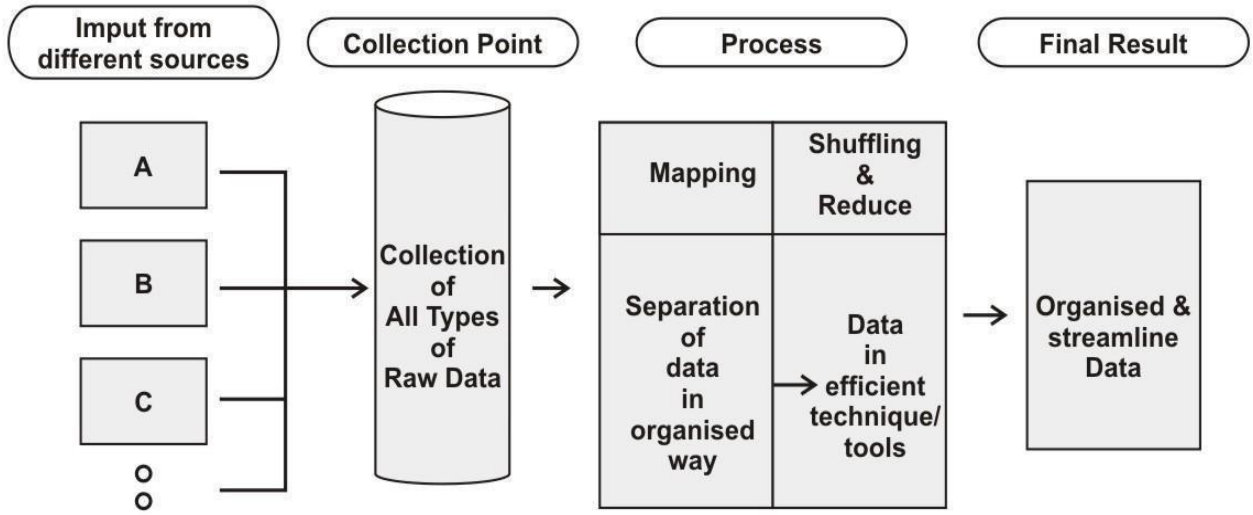
Variety: In diversity, we take into account many forms of data generation, such as organized and unstructured data, which may be produced by computers or by people. Any source of data is acceptable, including biometric data, data created by humans, data from the internet and social media, etc. The key to this type is the categorization of the incoming data.[3]



The phrase "healthcare analytics" refers to patient analysis tasks that can be carried out using information gathered from many sources, such as regional illness registers, medical groups, non-profit organizations, etc. Regression analysis, a tool for calculating the associations between variables, must be utilized to consolidate and analyze this data. A machine learning approach called classification tree analysis is used to classify data into a structure like a tree, with leaves denoting decisions and branches representing attributes. Sentiment analysis is a methodical approach to recognizing, obtaining, measuring, and analyzing affective states and subjective data. As technology has advanced and the volume of data entering and leaving businesses on a daily basis has expanded, there has analysis of the big data is:

- Acquire
- Store
- Process
- Utilize

The vast amount of data in the above figure is gathered from a variety of public sources, such as diagnostic centres, labs/clinics, medical groups (including institutes and education complexes), regional disease registers from government and semi-government authorities, various social camps hosted by non-governmental organizations, and other data providers. All this unprocessed data is extracted, meaning it is examined, analyzed, and crawled through to extract pertinent information in a predetermined format, such as a tabular form, from data sources (like databases). The raw data that was extracted is cleaned and organized. The process of transforming data from one format or structure into another is known as data transformation and separation, and it happens next. And it is now possible to utilize this optimized data going forward.



Map Reduce Method

Following the aggregation process, the data is utilized for a variety of purposes such as instrumentation data, visualization reports, and data mining, which is essentially used to extract and recover required information or patterns from enormous amounts of data. All of this data is examined from sources and distributed to relevant parties in order to track patient health and save lives. These optimized data applications and uses are also a commendable cause of social activities for human life. Big data in health informatics can be used to improve treatment and quality of life, forecast the course of diseases and epidemics, and stop the onset of new diseases and early deaths. It also offers details on illnesses and symptoms that should be taken seriously so that treatment can be started. This will support the government or in a personal capacity to save the cost of medical treatment.[4]

III. IMPLEMENTATION USING THE MAPREDUCE METHOD

A. Step 1: Input and the Collection of data

This calls for the usage of a distributed system since the data being gathered from various locations must load instantly. This is accomplished in the input from the many sources indicated in the preceding figure. The input data is kept here before being sent to the collection point. The collection points provide their data to be processed.

B. Step 2: Process and the final result

Data processing and storage in the Hadoop file system are the responsibilities of the mapper at this level. The mapper function receives the input line by line and generates several little data chunks.

Reducing and shuffling: This step combines the reduction and shuffling stages. The actual physical transfer of data over a network is called shuffling. On reducer nodes, the mapper's output is shuffled. The intermediate output is then combined and sorted as a result. This output is fed into the reduce phase, which applies a reducer function to each of these to produce the desired output. The finished product is this one. Streamlined data is produced as a result. [7].

C. Performance Tuning Using Machine Learning

Dimensionality techniques are used to extract meaningful features when dealing with a high number of features. To expedite computation and provide accurate results prediction, dimensionality removes superfluous information.

The vast quantity of information is gathered from several open sources. Normalization is used to transform all of this raw data into pure data. The technique of effectively arranging data in a database is called normalization. The technique is to remove unnecessary information and keep only relevant information in the table. Therefore, a pure database is obtained through normalization.

Random Forest technique is applicable to both regression and classification tasks. It uses several decision trees to form the forest. Results with higher precision are obtained from a larger number of trees.

To categorize a new item based on the attributes that each tree provides a classification for, we grow several trees in this model instead of a single tree in the court model, and we save tree votes for that class. When it comes to regression, the forest takes the average of the outputs from each tree, and it selects the classification with the most votes. This approach maintains the accuracy of the missing data while handling missing values. As a result, using an algorithm to examine the patient's medical records can help detect the illness. Thus, the highlighted data set is obtained by the use of preprocessing methods. [17]

The classification model is formed using the data from the highlighted data set. The target class is used for training and testing while the classification model classifies the input data. The classifier provides the target class data in order to make an accurate choice involving input data attributes based on the target class classifier model. Following the classifier's training, the next step is testing, during which input data is used to make predictions about the target class. A structure similar to a decision tree is used for classification when dealing with large amounts of data. A decision tree is utilized to provide precise and quick outcomes. A statistical technique for classification is called Support Vector Machine (SVM). SVM has the ability to make judgments on huge datasets. After training, this approach makes predictions fairly quickly. The best hyperplane that optimizes the margin between the various training data classes is used to make predictions. The goal is to select the hyperplane with the maximum margin from both training data classes. An ideal separating hyperplane would be produced by maximizing the distance between each class's nearest

points and the hyperplane. SVM will thus offer numerical data and accuracy measurements that aid in determining whether or not the sample's accuracy is over a threshold value. Consequently, the instance can be successfully assigned to the relevant labels for classification. [16]

D. Obstacles in Big Data Health Care

The fact that medical data is dispersed across numerous sources and is overseen by various states, hospitals, and administrative bodies is one of the largest obstacles to the use of big data in medicine. For the organization, gathering data that is accurate, full, clean, and properly organized for usage in a variety of systems is a never-ending struggle.

E. Security:

Data has emerged as one of the most valuable resources for businesses across all industries in recent years. As social networks, multimedia, and the internet of things provide an overwhelming flow of data, businesses will continue to collect more and more data, both in volume and detail. This trend will not change in the future. Attaining data security has emerged as a critical obstacle that could impede the proliferation of contemporary technology. This is because, as we all know, hackers, cybercriminals, con artists, and phishers can profit greatly from the sale of stolen data. Therefore, it is essential to guarantee that large data management, privacy, and security are well-protected and secured. It is crucial to protect health information with data encryption, firewalls, multi-layer authentication, anti-virus software, and transmission security. The healthcare ecosystem must continuously be on guard to secure data and keep personal information private if people are to feel comfortable sharing their private information. The availability of medical data must be regularly examined and tracked. [12] [14].

F. Storage:

Prior to this year, there wasn't enough room to keep the enormous amount of data that came in the form of reports, graphs, charts, movies, and so on. However, cloud computing has grown in popularity. Access to data processing, uploading, storing, and even designing the entire system in the cloud are all made possible by cloud storage. It is one of the safest ways to store data online in a way that is adaptable, safe, and economical. In addition to word documents, notes, inquiries, prescriptions, and the like, cloud storage is also utilized for the storing of graphic files like MRIs, movies, x-rays, and patient photographs. In order for physicians to rapidly see, comprehend, and make judgments, the system should also be able to create graphic presentations from the available data. All businesses can now more easily and conveniently manage the massive amounts of data they generate thanks to cloud computing. In cloud computing, Big Data is processed and analyzed using frameworks like Spark, Hadoop, and others. The cloud offers access to a vast array of resources and diverse infrastructures that can best support this integration; with little effort, the environment can be configured and maintained to provide an outstanding workspace for all big data requirements, such as data analytics. [13]

IV. CONCLUSION:

Big Data technology represents the solutions for healthcare in the future. With the vast and constantly expanding amount of data available, it offers a strong platform for resolving and enhancing health care challenges. Sources with a lot of data frequently provide a lot more noise and extraneous characteristics. Preprocessing methods and machine learning models, including Principal Component Analysis, Support Vector Machine, and Random Forest Algorithm, are therefore suggested in this work in order to accept the technique's acceptable trade-off and boost accuracy and efficiency simultaneously. As algorithms and tools continue to advance, more attention should be paid to addressing Big Data's inherent issues, such as data security, computational scalability, efficiency gains, and algorithmic enhancements. The use of big data, one of the most important new technologies, in healthcare will be very beneficial.

REFERENCES

[1] https://link.springer.com/article/10.1007/s12553-016-0152-4
[2] https://cloudxlab.com/blog/big-data-introduction/
[3] https://www.whishworks.com/blog/big-data/understanding-the-3-vs-of-big-data volume-velocity-and-variety

[4] http://www.aircconline.com/ijist/V6N2/6216ijist 16.pdf
[5] https://files.eric.ed.gov/fulltext/EJ1136190.pdf
[6] https://healthitanalytics.com/news/top-10-challenges-of-big-data-analytics-in healthcare
[7] http://www.tmrfindia.org/ijcsa/v13i12.pdf
[8] Big Data: A Revolution That Will Transform How WeLive, Work and Think by Viktor Mayer-Schonberger,
[9] Kenneth Cukier
[10] http://healthcare-communications.imedpub.com/the-usefulness-and-challenges-of-big-data-in-healthcare.pdf
[11] Big Data Now Current perspectives from O'Reilly Media
[12] https://en.wikipedia.org/wiki/Big_data
[13] http://www.iosrjournals.org/iosr-ce/papers/Vol18-issue3/Version-5/R18030 5120123.pdf
[14] https://www.sciencedirect.com/ science/article/pii/S1877705811065192
[15] http://iopscience.iop.org/article/10.1088/1755 - 1315/100/1/012026/pdf
[16] https://www.youtube.com/watch?v=TzxmjbL - i4Y
[17] https://pdfs.semanticscholar.org/26f2/55c768911e 5b4f85 9e7e4bbe1f734a834d45.pdf
[18] http://thesai.org/Downloads/Volume8No6/Paper_ 46A_Survey_of_Big_Data_Analytics.pdf