International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 09, September-2018

Big Data in Health Care System

Shreya Desai
Department of Computer Science
K.J Somaiya Institute of Engineering
and Information Technology, Sion,
Mumbai- 400022

Prof. Mamta Borle
Department of Computer Science
K.J Somaiya Institute of Engineering
and Information Technology, Sion,
Mumbai- 400022

***Abstract*---The healthcare sectors have grown rapidly in the last few decades. The healthcare industry provides facilities to the patients with medicinal, precautionary, reassuring care, etc. Thus, a huge volume of data is generated. A long time ago when the computers were not in use, this data was stored in the hard copy form. But now the current trend is rapid digitalization. Big data consists of a large amount of data which cannot be stored and processed using the traditional approach. This paper basically studies the implementation techniques of the big data in the healthcare sector. This paper also implies the challenges and tools to implement big data in the healthcare sector. It also discusses different machine learning algorithms that are useful for maintaining the trade-off between accuracy and efficiency. [1]**

***Keywords: big data, analytics, healthcare, machine learning*.**

I. INTRODUCTION

Daily a huge amount of data is produced via social media, etc. Now the question arrives that how huge this data to be in order to be classified as big data. The small amount of data can also be considered as a big data depending on the contents it is used. The name 'Big Data' itself is related to a size. Big data does not need to be only in the term of gigabytes, megabytes or anything larger than this in size. Considering an example of attachments in mail, for an instance if we try to attach a document that is of 200 megabytes in size, we won't be able to do so because an email system does not support an attachment of this size. Thus, this size of the document with respect to email can be referred to as big data. Big data in healthcare refers to electronic healthcare records (EHR) which are a digital version of a patient's paper chart where the information available instantly and securely to authorized users.

Health records can contain a patient's medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, test results and also allow access to evidence-based tools that provider can use to make decisions about a patient's health care. [2]

II. 3 V'S OF BIG DATA

Volume – Volume consists of the amount or the junk of data that is generated and stored. In total, we produce approx. 2.5 quintillion bytes of documents of data a day [15]. The volume of data is measured in terms of terabytes, records: that are generated to be processed, transaction: the online transaction that is being done in every second, file: the files which are being saved in form of documents, etc.

Velocity - Velocity is calculating the speed of the flow of the data. Batch in velocity means the data is preceded in batch wise like a set of 20 data so in chunks of data are processed. In near real time, the data is not processed immediately.

Variety – In variety, we consider different types of data generated i.e. both structured and unstructured data that has the possibility of getting generated either by humans or by machines. The data can come from any source like web and social media data, biometric data, human-generated data, etc. This variety is all about classifying the incoming data into different categories.[3]

III. OVERALL ANALYSIS OF DATA IN THE HEALTHCARE SECTOR.

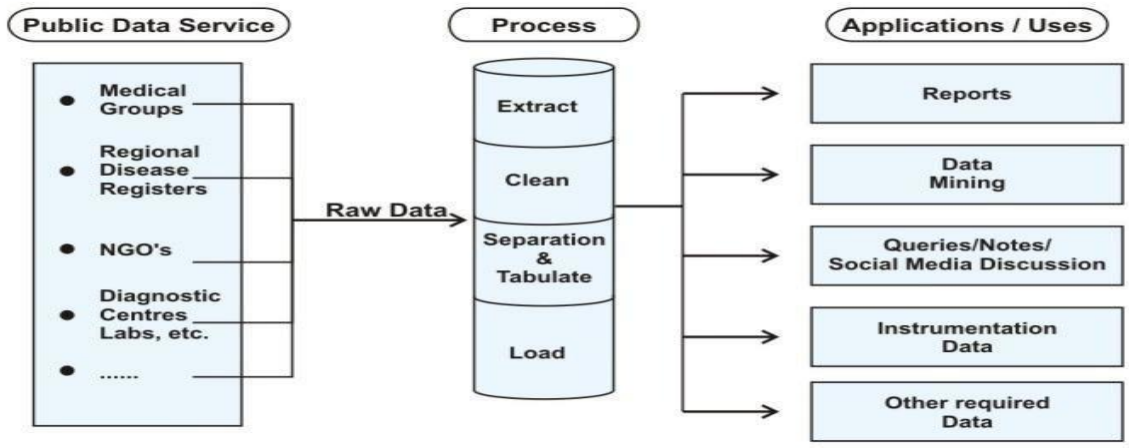

Fig 3.1 Big Data Analytics in Health Care

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 09, September-2018

Healthcare analytics is a term used to describe the healthcare analysis activities of the patients that can be undertaken as a result of data collected from the different areas like medical groups, regional disease registers, NGOs, etc. This data needs to be consolidated and analyzed by using some methods like Regression analysis, which is used for estimating the relationships among variables. Classification tree analysis is a machine learning algorithm used for classifying the data in a tree like structure where branches represent attributes and leaves represent the decision. Sentiment analysis, which is used to systematically identify, extract, quantify and study affective states and subjective information. With the evolution of technology and the increased multitudes of data flowing in and out of organizations daily, there has become a need for faster and more efficient ways of analyzing such data.

The life cycle of analysis of the big data is:

1. Acquire
2. Store
3. Process
4. Utilize

In the above figure, the massive amount of data is collected from the different public sources like medical groups including institute and education complexes, regional disease registers from government/semi-government authorities, various social camps held by NGOs, diagnosis centers or the labs/clinics and or other data providers.

All this raw data is extracted i.e. the data is analyzed and crawled through to retrieve relevant information from data sources (like a database) in a specific pattern like in a tabular form. This extracted raw data is cleaned and arranged. Then the data transformation and separation take place, it is the process of converting data from one format or structure into another format or structure. And now this streamlined data is available for the further use.

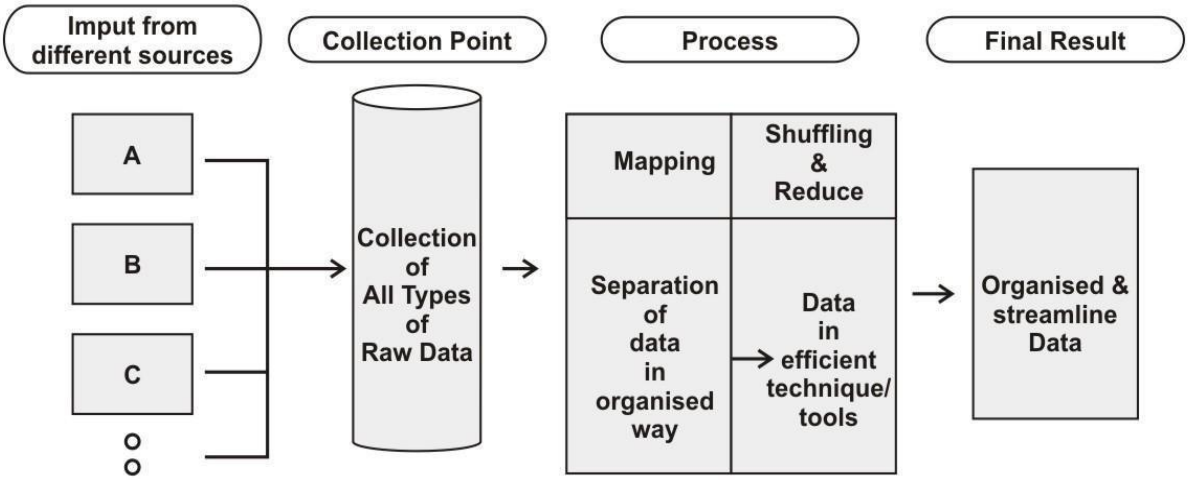

Fig 4.1 MapReduce Method.

After the aggregation process, this data is used as various visualization reports, instrumentation data, data mining which is primarily done to extract and retrieve desired information or pattern from a humongous amount of data etc. All this information is analyzed from sources is sent to the interested users for monitoring patient's health and to prevent the human life. These applications and uses of streamlined data is also a noble cause of social activities for human life. Big data in health informatics can be used to predict the outcome of disease and epidemics, improve treatment and quality of life, and prevent premature deaths and disease development. It also provides information about disease and warning signs for treatment to be administered. This will assist government or in a personal capacity to save the cost of medical treatment.[4]

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 09, September-2018

IV. IMPLEMENTATION USING THE MAPREDUCE METHOD.

Step 1 – Input and the Collection of data
There should be a distributed system for this because the data which is being collected from the different places need to be load without any delay. This is done in the input from the different sources mentioned in the figure above. This is the place where the input data is stored and passed to the collection point. The data from the collection points is sent for processing.

Step 2 – Process and the final result
Mapping – In this stage, the mapper's job is to process the input data and store it in the Hadoop file system. Line by line the input is passed to the mapper function and creates many small chunks of data.

Shuffling and reduce – This stage is the combination of the shuffle stage and reduces stage. Shuffling is the physical movement of data which is done over the network. The output of the mapper is shuffled on reducer nodes. And thus the intermediate output is merged and sorted. This output is provided as input to reduce phase and then runs a reducer function on each of this to generate the output. This output is the final output. Thus, streamlined data is generated. [7]

V. PERFORMANCE TUNING USING MACHINE LEARNING

To deal with a large number of features, dimensionality approaches are applied to obtain relevant features. Dimensionality eliminates unnecessary features to speed up the computation and also predict the accurate results.

The massive amount of data is collected from different public sources. All this raw data is converted into pure data using normalization. Normalization is the process of efficiently organizing data in a database. The goals of this process are that it eliminates redundant data and stores only related data in the table. Thus, using normalization pure database is obtained.

This pure data is preprocessed to form a featured database. Preprocessing can be done using the PCA algorithm and Random Forest algorithm to reduce the noisy features. PCA algorithm is a method of dimensionality reduction without much sacrificing the accuracy. It aims to summarize data with many independent variables to a smaller set of derived variables in such a way, that first component has maximum variance, followed by the second, followed by the third and so on. PCA is used because in many data analysis scenario independent variables are highly correlated which affects the model's accuracy and reliability. Random forest algorithm can be used for both classification and regression. It creates the forest with a number of decision trees. Higher the number of trees gives the high accuracy results.
In this multiple trees are grown as opposed to a single tree in court model to classify a new object based on the attributes each tree gives a classification and we save tree votes for that class. The forest chooses the classification having the highest number of votes and in the case of regression, it takes the average of the outputs of the different trees. This algorithm handles the missing values and maintains the accuracy for the missing data. Hence algorithm is helpful for identifying the disease by analyzing the patient's medical records. Thus, by using the preprocessing algorithms featured data set is obtained. [17]

The data from the featured data set is used to form the classification model. The classification model classifies input data and it uses target class for training and testing. Target class data is given by the classifier for performing correct decision that involves features of input data based on the classifier model for the target class. After training of the classifier, the next phase is testing were the input data is given to perform the prediction about the target class. For a huge amount of data, a decision tree like structure is used for classification. A decision tree is used for gaining accurate and fast results. Support Vector Machine (SVM) is a statistical method used for classification. SVM is capable of making decisions on a large dataset. This method performs prediction very fast after training. Prediction is done based on the optimal hyperplane which maximizes the margin between different classes of the training data. Idea is to choose the hyperplane which is at maximizing margin from both the classes of training data. Maximizing the distance between the nearest points of each class and the hyperplane would result in an optimal separating hyperplane. In this way, SVM will provide accuracy measures and numerical statistics which helps to decide whether the sample has accuracy above a threshold value or not. And thus the instance can be successfully classified in the appropriate labels. [16]

VI. OBSTACLES IN BIG DATA HEALTHCARE

One of the biggest hurdles in the way to use big data in medicine is how medical data is spread across many sources and are governed by different states, hospitals, and administrative departments. Capturing the data that is clean, complete, accurate and formatted correctly for use in multiple systems is an ongoing battle for the organization.

- Security

Over the past few years, data has become one of the most important assets for companies in every field. The tendency towards increasing the volume and detail of the data that is collected by companies will not change in future but will keep on increasing as the social networks, multimedia, internet of things is producing an overwhelming flow of data. Achieving the security in data has become one of the most important barriers that could slow down the spread of new modern technology because we all know that hackers, cyber theft, cheaters and phishing, where the stolen data can be sold for a huge sum. Hence, it is necessary to ensure that the administration, privacy, security of the big data is well protected and secured. Protection health information via transmission security, multi-layer authentication, using anti-virus software, firewalls, and data encryption are indeed vital. In order for individuals to feel comfortable sharing their private data, the healthcare ecosystem must constantly

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 09, September-2018

remain vigilant about protecting data and keeping private information private. The accessibility of the healthcare data needs to be consistently reviewed and monitored. [12] [14].

- Storage

Earlier there were limited space to store the huge amount of data which was in the form of charts, graphs, reports, videos, etc. but in latest year cloud computing has become more and more popular. Cloud storage offers access to data storage, processing, upload data or having the whole system designed in the cloud. It is one of the safe methods where data can be stored online in a flexible, secure and cost-effective basis. Apart from involving word documentation, notes, queries, prescription, etc. cloud storage is also used to store graphics type such as x-ray, videos, MRI, images per patient, etc. The system should also be able to generate graphics presentations from the available data so that clinicians are able to visualize and understand quickly and can take prompt decisions. Cloud computing makes the whole process of managing the enormous amount of data process easier and accessible to all enterprises. Frameworks such as Hadoop, Spark, etc. are used for analyzing and processing Big Data in cloud computing. The cloud has access to a large pool of resources and various forms of infrastructures that can accommodate this integration in the best suitable way possible; with minimum effort, the environment can be set up and managed to allow an excellent workspace for all the big data needs i.e. data analytics. [13]

VII. CONCLUSION

The Big Data technology stands for the future of healthcare solutions. It provides a solid platform for solving and improving health care issues with the huge and ever-growing amount of data available. Data-rich sources often result in correspondingly higher noise and irrelevant features. Thus, preprocessing techniques and machine learning models such as Random Forest Algorithm, Support Vector Machine and Principal Component Analysis are proposed in this paper to increase accuracy along with efficiency and accept the technique having a satisfactory trade-off. With the evolving algorithms and tools, the further focus should be on improvising the challenges in Big Data like data security, computational scalability, and increase in efficiency and improvement in algorithms. Being one of the most critical emerging technologies, the application of Big Data in the healthcare will prove to be a boon.

VIII. 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/V6N2/6216ijist16.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 We Live, Work and Think by Viktor Mayer-Schonberger, Kenneth Cukier
(9) http://healthcare-communications.imedpub.com/the-usefulness-and-challenges-of-big-data-in-healthcare.pdf
(10) Big Data Now Current perspectives from O'Reilly Media
(11) https://en.wikipedia.org/wiki/Big_data
(12) http://www.iosrjournals.org/iosr-ce/papers/Vol18-issue3/Version-5/R18030 5120123.pdf
(13) https://www.sciencedirect.com/science/article/pii/S1877705811065192
(14) http://iopscience.iop.org/article/10.1088/1755 -1315/100/1/012026/pdf
(15) https://www.youtube.com/watch?v=TzxmjbL -i4Y
(16) https://pdfs.semanticscholar.org/26f2/55c768911e5b4f859e7e4bbe1f734a834d45.pdf
(17) http://thesai.org/Downloads/Volume8No6/Paper_46-A_Survey_of_Big_Data_Analytics.pdf