Augmenting The Outcomes Of Health-Care Services With Big Data Analytics: A Hadoop Based Approach

N Pranav, Devavarapu Sreenivasarao, Shaik Khasim Saheb

Abstract: In today's data-driven world, one of the many emergent topics of great significance turns out to be Big Data Analytics, because in the recent years, it has led to a paradigm shift in various sectors, in terms of the strategies they implement to improve their operational effectiveness. However, Big Data Analytics poses great challenges too, because the analytics of huge mounds of structured and unstructured data is a highly arduous task, and hence, to make this task simpler and flexible, a nimble framework has to be equipped. Hadoop is one of many such frameworks, which has proven to be coherent, simple and highly efficient towards processing Big Data. One of the many sectors turning towards Hadoop based analytics is the medical industry, because of its ergonomic design and the accurate predictions made, which have succeeded in counteracting the results of laborious and fallible age-old diagnostic practices still being employed in most healthcare centres. Hence, this paper gives a brief overview of the applications of Hadoop in the healthcare industry, and its contribution towards strengthening the practices adopted in medical diagnostics.

Index Terms: Big Data, Hadoop, HDFS, MapReduce, YARN, Cloudera, HBase.

1. INTRODUCTION

THE healthcare industry is expanding at an exponential rate, owing to an increase in the requirement of healthcare services in every nook and corner of the world. The market value of the Indian healthcare industry is set to hit $372 billion by 2022 [9], which explain the magnitude of growth of medical services, and the industry as well. Added to the financial expansion is the technological advancement the industry has seen in terms of the practices adopted, as the healthcare industry has now evolved, and has seen a complete revolution in the kinds of curative practices used, be it sensor-based wireless information retrieval devices, smart patient-monitoring devices or EHRs (Electronic Health Records). Also, an important point to be noted is that all of these emerging techniques rely completely on Big Data Analytics, because it is the processing of data obtained which is the most vital facet of predicting medical results, also the reason why the healthcare industry has been the first one to embrace Big Data and associated techniques[8], as seen below, in Fig. 1.

- *N Pranav is currently pursuing Bachelors Degree program in Computer Science and Engineering in Sreenidhi Institute of Science and Technology, JNTUH University, INDIA, PH-8985633707. E-mail: npranav.2018@gmail.com*
- *Devavarapu Sreenivasarao is currently working an Assistant Professor in Computer Science & Engineering Department in Sreenidhi Institute of Science and Technology and Pursuing Ph. D (CSE) from Annamalai University, Chidambaram. Chennai. PH-9866014581. E-mail: sreenivasaraodevavarapu@gmail.com*
- *Shaik Khasim Saheb is currently working as Assistant Professor in Computer Science & Engineering Department in Sreenidhi Institute of Science and Technology, and and Pursuing Ph. D (CSE) from Annamalai University, Chidambaram. Chennai. PH-9642097865. E-mail: shaikkhasims@sreenidhi.edu.in*

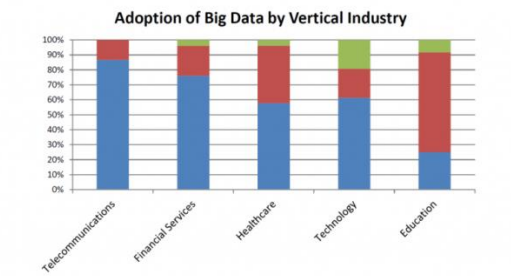

***Fig.1**. Adoption of Big Data by various industries in the year 2017*

Nevertheless, a lot of healthcare centres and hospitals have still clung to archaic medical practices and methods of diagnostics, despite the technological revolution, and this has led to an inhibition of accuracy in diagnosis. In this present-day scenario, where a disease-centric healthcare industry is shifting to a patient-centric one, these age-old techniques do no good to the improvement of medical diagnosis. Hence, this paper puts forward a proposal of Hadoop-based techniques which carry out Big Data Analytics to strengthen healthcare services, improving their speed and quality. The techniques proposed in this paper are primarily aimed at, one – predicting the condition of a patient, and two – predicting the disease a patient has, and the same have been elucidated in the following segments of the paper.

1.1. An Introduction to Big Data

Big Data is often confused with data which is big in terms of size, which is spurious. Big Data, though ambiguous in terms of the kinds of data it includes[16], essentially comprises of humongous volumes of data which are varied in terms of various parameters, i.e, datasets which may be structured or unstructured, pile up to large sizes, and come from an enormous number of sources together comprise Big Data. In order to characterize Big Data, the 'Vs' of Big Data are used, which are substantially increasing in number with the expansion of Big Data[1]. However, as shown in Fig. 2, Big Data is characterized by five basic Vs – Volume, Velocity, Variety, Veracity and Value [22].

Volume
The size of Big Data is described by a parameter called 'Volume', which is usually not less that few millions of Gigabytes, considering the generation of big data on the internet on a daily scale. This is an important parameter used to classify data into – conventional data, and big data.

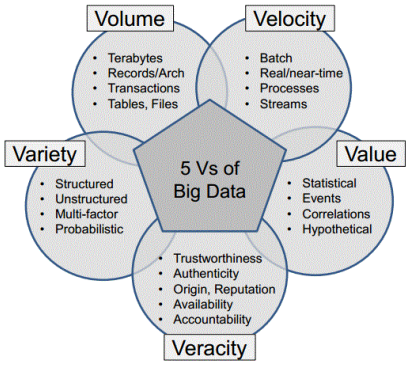

***Fig. 2**. The Five Vs of Big Data*

Velocity
Velocity determines the capacity of big data by measuring how 'fast' it is generated. In simpler terms, it is a term used to characterize the speed of data flowing, originating from a multitude of sources such as mobile phones, computers, machines, networks, etc. into platforms such as social media.

Variety
The homogeneity or heterogeneity of Big Data is described by the term 'Variety'. The variety of Big Data defines the methods of analytics to be used, as it describes if the data obtained fits into a formal format of data, such as a spreadsheet, or doesn't fit, such as audio, video, or text files. Variety describes the 'structured' nature of Big Data.

Veracity
Veracity describes the accuracy of data in the dataset, as it denotes with the abnormalities, inconsistencies and volatility of big data, such as data on social media, where trending topics keep changing often.

Value
The value of big data describes how strategically advantageous it is to organizations which would make use of it later, for analytics. In simpler terms, it delineates the value of that data in the market, if it would be analyzed.

1.2. An Introduction to Hadoop

Structuring huge amounts of data is a herculean task, owing to factors determining the characteristics of Big Data such as scale, speed and heterogeneity [11], as dealing with each of these in the process of filtering Big Data makes the process complex. In order to cope with all of the aforementioned challenges, an advanced computational framework is required that can handle huge amounts of Big Data, and can optimize the time taken for processing such large amounts of data, and this could be made possible by improving data storage and retrieval, and the processing speed. One such a framework is Hadoop, which enables parallel processing and distributed storage together [13], and hence, helps analysts focus on the analytics to be done, not on the resources to be deployed to make Big Data Analytics happen.
Hadoop is a Big Data processing software framework developed by the Apache Software Foundation and is open source. It enables distribution of data into clusters, and processing data stored on various clusters together, That is, Parallel Processing. Hence, in simpler terms, Hadoop is an example of a framework which processes enormous sets of data in a clustered computing environment, and is an implementation based on Google's MapReduce model. In order to facilitate parallel processing and distributed storage, the Hadoop ecosystem comprises of a variety of other frameworks such as HBase, Pig, Hive, Drill and Spark [15], which together help convert amassed Big Data into necessary outputs and results.

Hadoop Components and Ecosystem

The Hadoop framework comprises of two key components.
Hadoop Distributed File Storage System (HDFS)
HDFS is a basic component of the Hadoop framework, and is a distributed file system which stores a huge dataset as multiple files, across various machines or clusters, which makes it highly fault-tolerant [12]. HDFS is designed using low cost hardware, and is hence an effective solution towards storing Big Data effectively. HDFS is based on the master-slave architecture, where data storage is handled by NameNodes and DataNodes[14], as shown in Fig. 3.

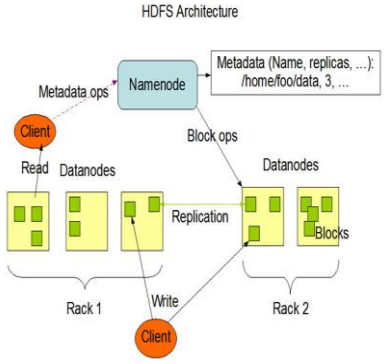

***Fig. 3**. HDFS Architecture*

The implementation of HDFS is completely based on Java, and hence, can be extensively run on ordinary hardware. The key features of HDFS include – high availability of data by means of replication, easy extraction, retrieval and storage of data, and simple distribution of data.

MapReduce
MapReduce is a programming model, or in simpler terms, a means of processing which efficiently carries out parallel processing across multiple computing nodes, thereby cutting down the time required for processing extremely large datasets. Unlike conventional computing in which data is sent to the processing units, MapReduce works on the following paradigm : it sends the computing unit to the point where a chunk of data resides. MapReduce works in two fragments: the Map part and the Reduce part [21], as shown in Fig. 4.

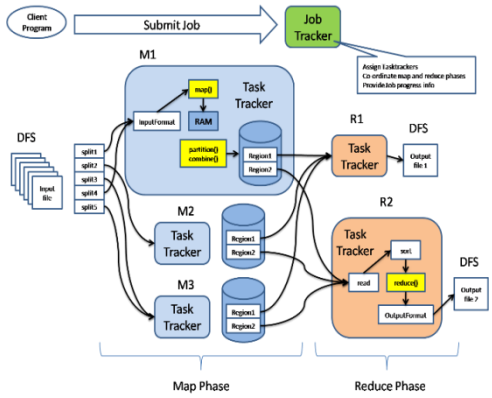

Fig. 4*. MapReduce Architecture*

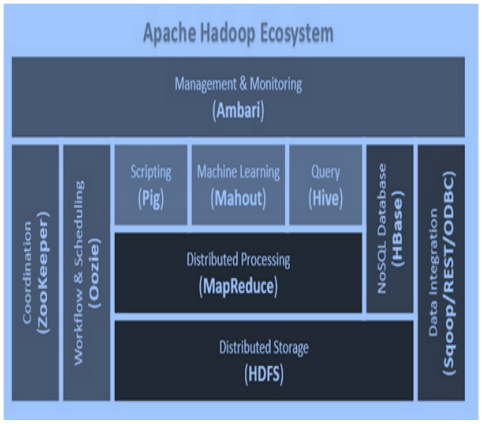

Fig.5*. The Ecosystem of Hadoop*

Mapping

The job of the Mapper is to "map" the data. In other words, the Mapper takes inputs from data stored in the HDFS, and is parsed line by line, after which data is split into chunks, precisely into "key-value" pairs. The outputs obtained from the Mapper are sent as inputs to the Reducer.

Reducing

The Reducer layer takes inputs from the Mapper, and performs the following operations on them. Shuffling, Sorting and Reducing. In this way, the mapped data is further processed by the Reducer. The final outcome of processing led by the Reducer is stored as a new set in the Hadoop Distributed File Storage System. The Map and Reduce tasks are driven by a MasterNode [23], which receives job requests from clients, to map and reduce data. The MasterNode splits the task among various SlaveNodes, where MapReduce tasks are carried out. Also, another element of Hadoop which is sometimes considered a component too is YARN (Yet Another Resource Negotiator), which is responsible for handling, scheduling and monitoring tasks being carried out in the framework. YARN is responsible for scheduling tasks to be executed on various cluster nodes, and also for allocating system resources to various applications and jobs running on different clusters, and hence, contributes to smooth and effective task management in Hadoop. The ecosystem of Hadoop [17] as shown in Fig. 5 comprises of a wide range of frameworks that allow it to carry out operations and work on various kinds of computation and simplification simultaneously. The prime components of Hadoop's Ecosystem are MapReduce, HDFS, Mahout, Flume, Ambari, ZooKeeper, Sqoop, Oozie, HBase, and a lot more, which construct a suite providing a variety of services to facilitate all kinds of data processing[21]. The ecosystem of Hadoop is what makes it more modular than advanced Big Data processing frameworks such as Spark, as Spark primarily optimizes techniques used in Big Data Analytics, and its ecosystem is less modular in comparison with that of the ecosystem of Hadoop.

2. LITERATURE REVIEW

The efficiency of healthcare services can be augmented in numerous ways by Hadoop and Big Data Analytics, and hence, we see no dearth of fields of healthcare Hadoop based techniques can contribute to. Deepthi Yaramala et al.,[6] propose various Big Data Analytics based techniques which can be implemented in hospitals to segregate data from Electronic Health Records associated with patients needing blood in each state per year, companies supplying medications per year, diseases affecting the most number of patients, etc., implemented using frameworks belonging to the Hadoop ecosystem, such as Hadoop, Hive and HBase. B. Durga Sri et al.[7] propose a Hadoop, Hive based model, working on the Cloudera platform, used for the analysis of trends of patients getting admitted to hospitals, and the prevalence of a disease in a given area, on a given day. MS Minu Sanjudharan et al.[5] put forward the idea of using HDFS to store EHRs in a repository, and using MapReduce to carry out an analysis of the drugs prescribed to a patient suffering with a particular disease, and the average expenditure of a household on drugs. Mukesh Borana et al.[4] propose the idea of enhancement of analytics by Hadoop by clustering the given data by pairing MapReduce with the Fuzzy C-Means Clustering Algorithm and the Iterative Dichotomiser 3 (ID3) Classifier, contributing to the simplification of massive volumes of EHRs processed by Hadoop. Vishruti Patel et al.[10], in their paper explain the potential of technology reliant on Hadoop such as Medical Body Area Networks (MBANs), and their contribution to advanced healthcare services.

3. SOLUTIONS TO COMMON HEALTH-CARE PROBLEMS

As described in the introduction to the paper, using Hadoop based techniques to simplify Big Data in medical diagnosis can be used to directly prescribe medication on the basis of evidences obtained from various patients, instead of administering several lab tests to patients. However, these techniques are not being sought after or implemented, owing to the complexity of implementation. Hence, two problems addressed by this paper are:

3.1. Patient Status Prediction

Data obtained from a patient's lab tests stored in EHRs, is further simplified using algorithms and is mapped and reduced using MapReduce. The same is demonstrated in this paper

using ECGs (Electrocardiograms) taken from patients' reports, converted into reducible data, and is then mapped and reduced, to predict if the heart condition of a patient is critical.

3.2. Patient Disease Prediction

Every disease is defined by a set of rules or symptoms measured on the scale of results of lab tests, such as platelet count, blood pressure, and in order to assess the disease a patient is suffering with, data obtained from the results of a patient's lab tests is mapped and reduced in comparison with the prototype of every disease, in order to predict the disease the patient is suffering with.

4. PATIENT MEDICAL STATUS PREDICTION USING HADOOP – WORKING

This paper describes a working proposal of predicting the medical condition of a patient's heart, on the basis of the assessment of his/her Electrocardiogram. All the readings made in the ECG are first converted into data which can be processed by Hadoop, and then, MapReduce is used to simplify this data to predict heart condition. The following flowchart, Fig. 6 describes the working of this proposed MapReduce model.

As shown in the flowchart below, the steps defining the flow of operations in the proposed system are as follows.

1. Preprocessing (PAA)
2. Preprocessing(SAX)
3. MapReduce
4. Generating Outputs

After the ECG is converted into readable values by PAA (Piecewise Aggregate Approximation) and SAX(Symbolic Aggregate Approximation), it is stored on a virtual HDFS tool, from which it is sent into a MapReduce job [3]. On the basis of the number of values obtained corresponding to a key, depending on a threshold set, the heart condition of a patient is predicted to be normal or critical.

4.1 Hardware Requirements

The proposed model has been tested on a computer with the following hardware specifications:

- Processor: Intel CoreTM i3 (dual-core)
- RAM: 8GB
- Hard Drive Space: 500GB
- Platform: Cloudera run on any virtual machine, for instance, VMWare Workstation or Oracle VirtualBox.

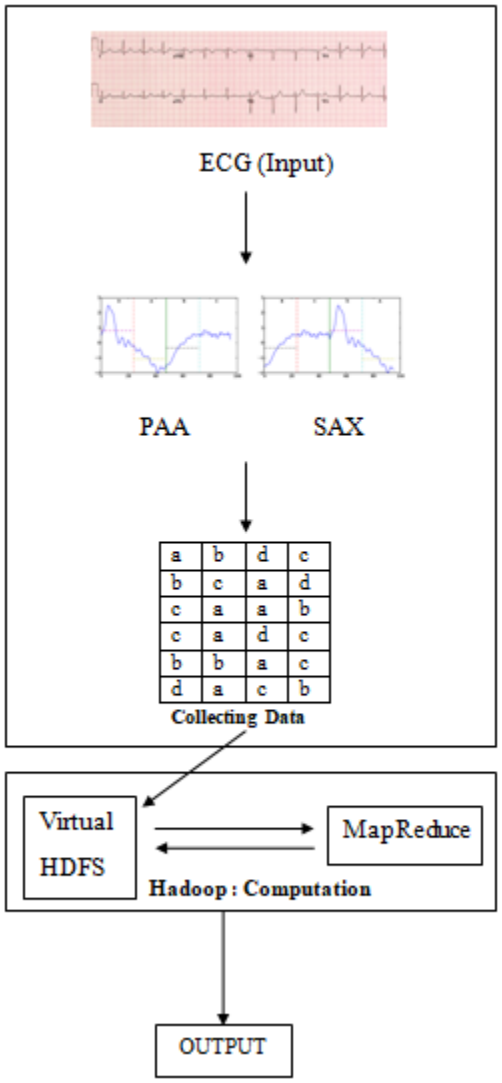

***Fig. 6**. Patient Status Detection Working Model, Flow of Operations*

In this proposed framework, we use Cloudera to carry out Hadoop based operations. Cloudera is a platform built on the vanilla (basic) version of Hadoop[18], used to simplify configuration, installation and execution of commands to carry out large-scale data analytics on Hadoop. Fig. 7 shows the basic user interface of the Cloudera Manager.

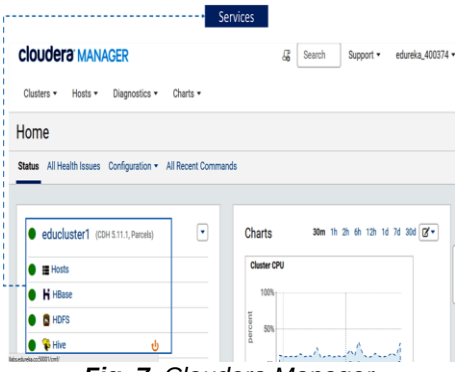

***Fig. 7**. Cloudera Manager*

Cloudera has pre-conFig.d components of the Hadoop ecosystem, such as MapReduce, HDFS (conFig.d according to the system storage/in the cloud), YARN, Hive, HBase, and a lot more, and is hence used in the implementation.

4.2. Implementation

4.2.1 Preprocessing – PAA and SAX

An Electrocardiogram (ECG) is a test which measures the rhythmic activity of a patient's heart, which is recorded on a screen or a strip of paper. Since such graphical data cannot directly be sent to be processed by Hadoop, it has to be converted into readable data. The first step towards doing so is carrying out PAA(Piecewise Aggregate Approximation) of the ECG [19]. Consider the following segment of an ECG, shown in Fig. 8.

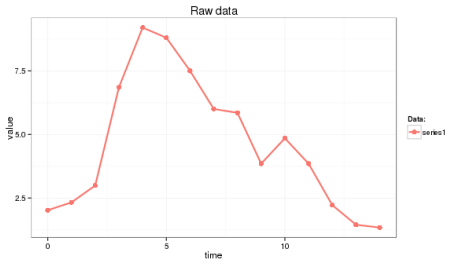

***Fig. 8**. Unprocessed ECG*

Assuming that the ECG is of length 'n', it is split into 'M' segments such that an x_i¯corresponding to each divided segment is given by

$$\bar{x}_i = \frac{M}{n} \sum_{j=n/M(i-1)+1}^{(n/M)i} x_j$$

i.e, when the graph is divided into M equally spaced frames, the mean values of each frame are calculated [19], and are then combined to result in a sequence shown in Fig. 9.

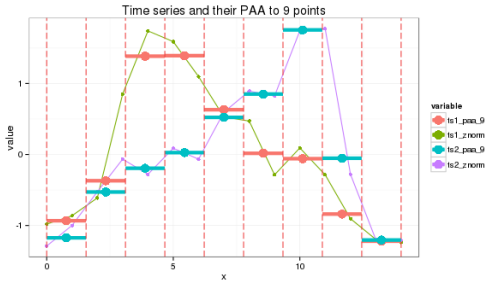

***Fig. 9**. ECG after PAA*

Now, this ECG is subjected to Symbolic Aggregate Approximation, in which the time series is converted into a string. Consider a vector B=$\beta_1,\beta_2,\ldots,\beta_{a-1}$ such that $\beta_{i-1} < \beta_i$ and $\beta_0 = -\infty$, representing the co-ordinates of the lines which divide the graph into segments. Let C¯ represent the vector of PAA coefficients obtained from the previous graph. If this is to be converted into a string C^,

$$\hat{c} * i = alpha * j, \text{ iif, } \bar{c} * i \in [\beta_{j-1}, \beta_j)$$

Where 'alpha' represents an alphabet assigned to values in the specified interval. After applying the above operations, the following split, seen in Fig. 10 is obtained[19].

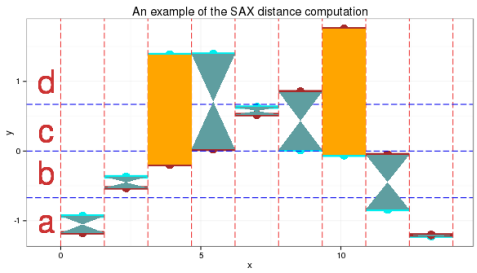

***Fig.10.** ECG after PAA and SAX*

These letters are obtained by calculating the Euclidean and PAA distances between each of these segments, and assigning a letter to each of them. Next, the letters are tabulated, as shown in Table 1.

TABLE 1

TABULATED LETTERS TO BE SENT INTO HDFS, THEN MAPREDUCE

a	b	d	c
b	c	a	d
c	a	a	b
c	a	d	c

4.2.2 MapReduce Operations on the Data Obtained

Tabulated values obtained of thousands of patients are stored in the cloud or in a virtual HDFS environment, facilitated by Cloudera. This data is sent into a MapReduce job. As described in the earlier sections of this paper, MapReduce has two functional components: Mapping and Reducing, both of which deal with data split into key, value pairs. The same happens with the dataset seen above. One of the many advantages of using MapReduce is the feature of giving it a 'Custom Input Format', and hence, in this case, the input format of a patient includes the Patient ID, the SAX String ID, and the tabulated values which may range from 10 to nearly 150-200 columns. In order to demonstrate a simple MapReduce job, the dataset has been simplified to three columns and three entries. As seen in the Fig. below, data sent into a MapReduce job is first mapped into key, value pairs, and is then reduced to count various phases observed in the ECG. Fig. 11 and 12 describe the functioning of a MapReduce job on the tabulated data.

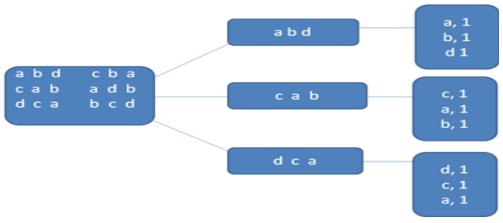

***Fig. 11.** The Mapper*

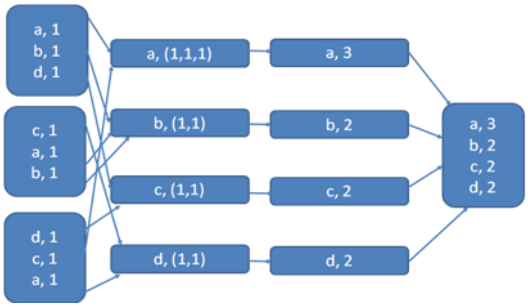

***Fig. 12**. The Reducer*

As per the clinical criterion set and defined, from the obtained output, the model carries out an analysis of the state of every patient, and prints the status of all the patients, and if a patient's heart is in a critical state, the model prints the same too.

5. PATIENT DISEASE PREDICTION USING HADOOP – WORKING

Diagnostic reports have to be reviewed by a doctor, and hence, on the basis of the results seen in the reports, the doctor predicts the ailment the patient has been affected by, but at times, this turns out to be fallible. For instance, neurological syndromes can be mistaken for psychiatric illness [20], owing to the fact that both of these conditions occur with nearly the same symptoms. This proves that manual diagnosis is not reliable and accurate all the time, therefore, there is always the need of computerising disease prediction, and this can be made possible through Hadoop and HBase. Given that a set of rules, symptoms or values from reports are defined corresponding to every disease and are stored in a mega-database, which in this case is HBase, Hadoop can be used to simplify data from the EHRs of patients which can then be classified on the basis of diseases, using ML classifiers. MapReduce is also used in this case to calculate the number of patients suffering with a given disease in a hospital, taken over a period of time.

5.1 Disease Rule Generation and Risk Factor Selection

In order to predict a disease, rules corresponding to each disease have to be generated. In simpler terms, the identification of a disease is done on the basis of rules resembling conditional IF-ELSE statements set. For instance,

IF ((age>45) and (intake_fat>42.4mg/day) and (married=yes)) THEN cardiovascular_disease := yes.

All of these rules are generated only after thorough examination of scores of patient records. In this paper, we have chosen the datasets from the Korean National Healthcare Centre (KNHC), comprising of records of close to 7,00,000 patients with various fields of information on a patient's disease, disease history, lab tests and records, etc, and MIMIC – III, which comprises of information about diseases patients are affected with when they are admitted in the ICU of a hospital, the number of days they stay in the ICU, death date, number of deaths caused by a particular disease, etc.

Once these rules are generated, they are stored in HBase. Also, the risk factor of a disease is taken into consideration in this model, and a risk factor of a disease is one which when possessed by a person increases the probability of him/her getting affected by the disease, and estimating risk factors corresponding to various diseases is a vital facet of disease prediction.

5.2 Working

The flowchart in Fig. 13 explains the sequence of operations used to implement the proposed model. Here, MapReduce is used to simplify the input data (records of a patient) to make it simple enough for classifiers to identify the disease. Using classifiers such as K-means, Random Forest or Decision Trees along with algorithms such as Apriori, probable diseases are filtered on the basis of the risk factor, and are then further classifies to predict the disease. It has been observed that Random Forests give the most accurate outputs as classifiers, and hence, they are used to predict the disease a patient is suffering with. Further, every such disease predicted is stored in HBase, and a simpler version of MapReduce is applied to calculate the number of diseases people have been affected within a given area. Hence, MapReduce is used for two purposed in this model – to simplify data in EHRs and to count the number of diseases people have been affected with in an area.

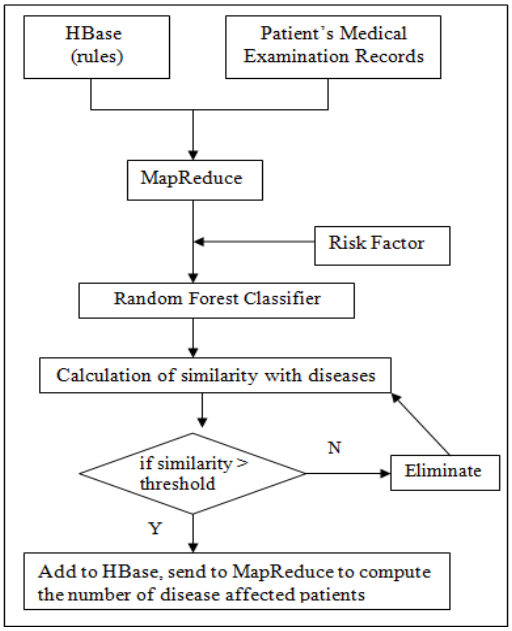

***Fig. 13**.Disease-Prediction Model*

As depicted in the flowchart above, the flow of operations in the proposed model may be defined by the following steps –

1) MapReduce on simplified EHRs
2) Classification using Random Forests
3) Disease Similarity Identification
4) Computation of number of disease affected patients after disease identification

It may also be noted that MapReduce may be used with functional models such as the Multivariate Adaptive Regression Splines model in order to enhance the predictions made, also, to eliminate the need of using a classifier [2].

Towards the end of the model, in order to count the number of disease affected patients in area in one year, a list of all the predicted diseases of all patients are sent into a MapReduce model. The algorithm leading to the simplification of the data is

shown below, in Fig. 14.

```
class Mapper
  method map (HBase table)
       for each instance row in table
          write((diseasei, patientID), 1)

class Reducer
  method reduce ((diseasei, patientID), ones[1,1,1,…n])
     sum=0
     for each one in ones do
         sum+=1 11.
     return ((diseasei, patientID),sum)
```

Fig. 14*. Map and Reduce algorithms used to count the number of patients*

6. RESULTS

To start with, an implementation of the patient-status prediction model gives the results, as seen in Fig. 15. It is seen that the model gives accurate predictions, and prints the status of all the patients, specifically those with a critical heart condition. The model has been implemented on a computer with specifications mentioned in Section 5.1, and is executed by loading data into a virtual HDFS environment hosted by Cloudera, and then, applying MapReduce on the data as explained in the preceding sections.

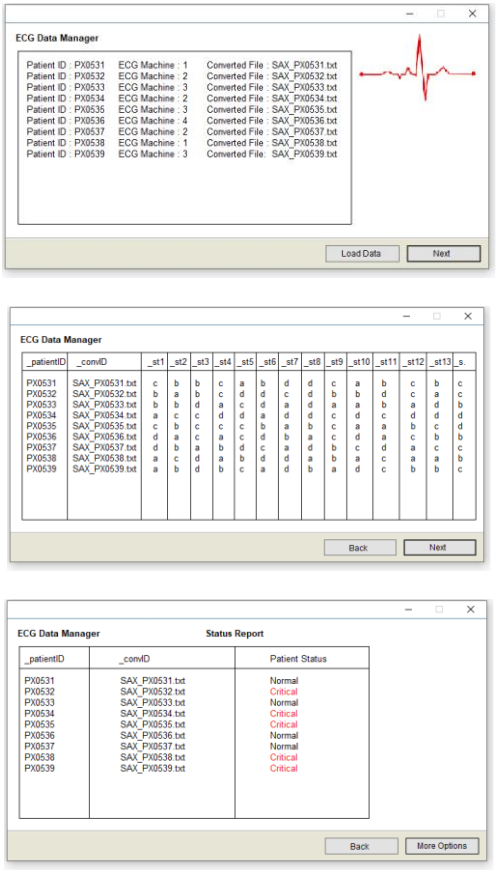

Fig. 15*. Results of implementing the Patient-Status Prediction Model*

Secondly, the disease-prediction model too makes nearly accurate predictions of diseases. The number of diseases per year in an area computed by MapReduce, limited to select diseases, computed from EHRs of patients gives the following results, as seen in Table 2.

TABLE 2

DISEASES COUNT PER YEAR

Disease	Count
Thalassemia	949
Dysplasia	332
Blastomycosis	678
Pneumonia	742
Gastroenteritis	1,534

7. CONCLUSION

This paper briefly outlines the contribution of Big Data Analytics to the healthcare industry, and proposes two techniques which can be brought in retaliation to the age-old techniques still being used in diagnostics, in this age of revolution. This paper puts forward the idea of using Hadoop–enabled Big Data Analytics to diagnose patients in the fastest possible means, and predict the heart condition of an individual, and also, the disease an individual is suffering with, made possible by empowering diagnostics with parallel processing and distributed computing, MapReduce and HDFS respectively. The results of implementing the same have been discussed in the paper, too.

REFERENCES

[1] D. Peter Augustine “Leveraging Big Data Analytics and Hadoop in Developing India’s Healthcare Services”, International Journal of Computer Applications, Vol 89, No.16, March 2014.

[2] Dingkun Li, Hyun Woo Park, Erdenebileg Batbaatar, Yongjun Piao, Keun Ho Ryu “Design of Health Care System for Disease Detection and Prediction on Hadoop Using DM Techniques”, International Conference of Health Informatics and Medical Systems, 2016.

[3] Dr.V.P.Gladis Pushparathi, S.Divya Barathi, S.Kavitha, S.Shakti Nivetha, “Clincial Decision Support System using Hadoop”, International Journal of Innovative Research in Science, Engineering and Technology, Volume 7, Speical Issue 2, March 2018.

[4] Mukesh Borana, Manish Giri, Sarang kamble, Kiran Deshpande , Shubhangi Edake “Healthcare Data Analysis using Hadoop”, International Research Journal of Engineering and Technology , Volume 02, Issue 07, October 2015.

[5] MS.Minu, Ishan Meena, Pratyush, R. Aravind, Vijayditya Sarker, “Healthcare Analysis Using Hadoop Framework”, International Journal for Science and Advance Research In Technology, Volume 4, Issue 10, October 2018.

[6] Deepthi Yaramala “Healthcare Data Analytics using Hadoop” , Thesis, San Diego University.

[7] B. Durga Sri, K.Nirosha, M. Padmaja, “Healthcare Analysis Using Hadoop”, International Journal Of Current Engineering And Scientific Research (IJCESR), Volume-4, Issue-6, 2017

[8] Adoption of Big Data by various industries https://deevita.com/how-is-big-data-used-by-different-industries/.

[9] Expansion of Big Data in the healthcare industry:https://www.investindia.gov.in/sector/healthcare.

[10] Patel Vishruti, Prof. M.D. Ingle, “Disease Analytics in Healthcare System using Hadoop”,International Journal for Modern Trends in Science and Technology, Volume 4, Issue 12, December 2018.

[11] Rahul Beakta, “Big Data and Hadoop : A Review Paper” , ResearchGate, Volume 2, Special Issue 2, 2015.
[12] Iqbaldeep Kaur, Navneet Kaur, AmandeepUmmat, Jaspreet Kaur, Navjot Kaur, “Research Paper on Big Data and Hadoop”, Internal Journal of Computer Science and Technology, Volume 7, Issue 4, October-Dec 2016.
[13] IvaniltonPolato, ReginaldoRé, Alfredo Goldman , Fabio Kon “A comprehensive view of Hadoop research—A systematic literature review”, Journal of Network and Computer Applications, Elsevier, 2014.
[14] Harshawardhan S. Bhosale ,Prof.Devendra P. Gadekar, ”A Review Paper on Big Data and Hadoop”, International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014.
[15] Apache Hadoop Project – https://en.wikipedia.org/wiki/Apache_Hadoop.
[16] MIT Technology Review–“The Big Data Conundrum:How to define it?”, https://www.technologyreview.com/s/519851/the-big-data-conundrum-how-to-define-it/
[17] The Ecosystem of Hadoop - https://www.janbasktraining.com/blog/introduction-architecture-components-hadoop-ecosystem/
[18] Cloudera and Hadoop - https://www.edureka.co/blog/cloudera-hadoop-tutorial/
[19] Principal Aggregate Approximation and Symbolic Aggregate Approximation,https://jmotif.github.io/sax-vsm_site/morea/algorithm/PAA.html, https://jmotif.github.io/sax-vsm_site/morea/algorithm/SAX.html
[20] Reports on diseases often mistaken to be one another https://jnnp.bmj.com/content/76/suppl1/i31
[21] MapReduce–Apache Hadoop, https://hadoop.apache.org/docs/r2.8.0/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html
[22] The Five Vs of Big Data, http://iihtofficialblog.blogspot.com/2014/07/5-vs-of-hadoop-big-data.html
[23] Architecture of MapReduce in Hadoop, http://a4academics.com/tutorials/83-hadoop/840-map-reduce-architecture