Implementation of Electronic Health Record System using Hadoop

Mr. C.S. Arage, Mr. M.P. Gaikwad, Mr. Akshay Gandhi,
Mr. Sanket Majale, Mr. Vivek Patil

Department of Computer Science and Engineering, Sanjay Ghodawat Institute, Atigre.
arage.cs@sginstitute.in, gaikwad.mp@sginstitute.in, akshay0045@gmail.com
sanketmajale1008@gmail.com, vivekpatil.patil1@gmail.com

Abstract— Big data and the related technologies have improved health care enormously, from understanding the origins of diseases, better diagnoses, helping patients to monitor their own conditions. By digitizing, combining effectively using big data, healthcare organizations can improve their quality of service by analyzing the effectiveness of a treatment and also the efficiency of the healthcare delivery process and drug abuse more quickly and efficiently. General goals to use analytics are, we can predict readmission risks, increase the efficiency of clinical care, and finding opportunities for cost savings. This paper gives various solutions for how and where big data can be applied in the health care system. Apache Hadoop is open source software used to process huge data sets in the distributed computing environment using clusters and commodity hardware. MapReduce is a programming model for processing such huge data sets. Further we propose a MapReduce Program to efficiently Analyse Electronic Health Records (EHR) database.

Index Terms — **Predictive Analytics, Electronic Health Records, Big Data, Map-Reduce Architecture, JAVA.**

I. Introduction

Big data is a collection of techniques and technologies which needs new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale. Big data can also be defined as large volume of unstructured data which cannot be handled by traditional data management tools like relational database management system. The increasing digitization of healthcare information is opening new possibilities for providers and payers to enhance the excellence of care, improve healthcare outcomes, and reduce costs. Due to advance technologies the paper works are converted into digital format (digital health records or Electronic Health Records (EHR). Since information is in digital form, healthcare providers can use some available tools and technologies to analyze that information and generate valuable insights. As health care data is generated in variety of devices, with high velocity and huge volume the big data solutions are required to solve the problems of storage and processing. There are many big data technologies available to solve these issues. But as health care data need to be handled in a different way we many need to customize according to the specific purpose. Big data analysis in health care data can reduce the costs and improve the quality of health care by providing a personalized health care. Big Data in healthcare industry promises to support a diverse range of healthcare data management functions such as population health management, clinical decision support and disease surveillance. The Healthcare industry is still in the early stages of getting its feet wet in the large scale integration and analysis of big data.

With 80% of the healthcare data being unstructured, it is a challenge for the healthcare industry to make sense of all this data and leverage it effectively for Clinical operations, Medical research, and Treatment courses.

LITERATURE REVIEW AND PROBLEM DEFINITION

i. Problem Definition

Human health information from healthcare system can provide important diagnosis data and reference to doctors. However, continuous monitoring and security storage of human health data are challenging personal privacy and big data storage. To build secure and efficient healthcare application, Hadoop-based healthcare security communication system is proposed

ii. Drawback

Limitations of existing work

- Time-consuming process
- Manual Process
- Inefficient
- Require more human interaction

II. METHODOLOGY

The major challenge in big data analytics is to locate the required information from tables and extract the information contained in database .Defining appropriate tools and techniques for the same in highly desired in order to enrich the healthcare research. Obviously this involves a number of preprocessing and classification steps. For example, a big data analytics can be employed to solve the problem of searching relevant information about a particular individual from the huge database. The big data analytics research focuses on analyzing the data automatically, update the database, extract the most informative data. In the EHR system, information gathered from patients, doctor's pharmacy having representation are considered for data analytics. In this proposed research work our approach is to study the existing data and to develop and implement a new system for data analysis which have a better degree of performance. To develop this system we are making the use of Hadoop cluster. For the purpose of data analysis and extraction of resultant data "K-means Clustering" algorithm has been used.

The proposed System has 3 modules:
A. Data Collection
C. Big-Data Analysis
D. Prediction of disease based on algorithm

MapReduce:- MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

K-Means Algorithm:- K-means is one of the simplest unsupervised learning algorithm that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centres, one for each cluster. These centers should be place in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point, we need to re-calculate k new centroids as barycentre of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centre's change their location step by step until no more changes are done or in other words centres do not move any more. Finally, this algorithm focus minimizing a objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (||x_i - v_j||)^2$$

Where,

'||*xi - vj*||' is the Euclidean distance between *xi* and *vj*. *'ci'* is the number of data points in i^{th} cluster. *'c'* is the number of cluster centers.

Algorithmic steps for k-means clustering

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Big-Data Analysis

This module plays important role in decision making for the disease prediction based on the current symptoms identified for particular patient. Big data analysis is is required as we will have data of extremely high number of patients. The K-means algorithm is implemented to perform this data analysis on Hadoop cluster.

As this database collects information over the years for patient symptoms and detailed disease information based on symptoms. As K-means algorithm for Hadoop cluster only works on numerical values, we have mapped a numbers starting from 1 to each symptoms. This number later mapped to particular disease and finally after the analysis result is displayed to user.

- **System Architecture:**

The architecture includes patients, big data analysts and specialized doctors. In the first step the individual inputs of a set of data sets. The individual's diseases from various health records like Electronic Health Records (EHR), Age, Gender and Symptoms are the7n compared to all other patients available in the existing database and an early filtering is done. The data analyst's outcome is a list of diseases which the patients have. The patient's data are kept in a big data storage area as a dataset.

Which is then processed by the doctors as well as those who are handling the patient's when they are in critical stage of sick.

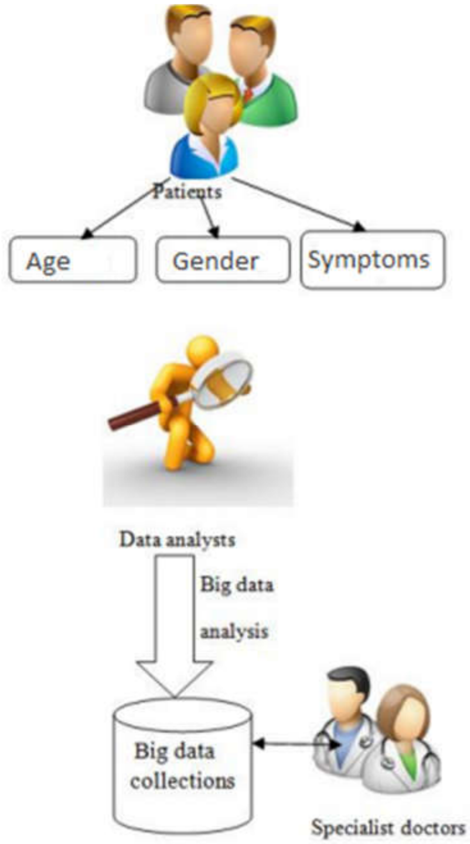
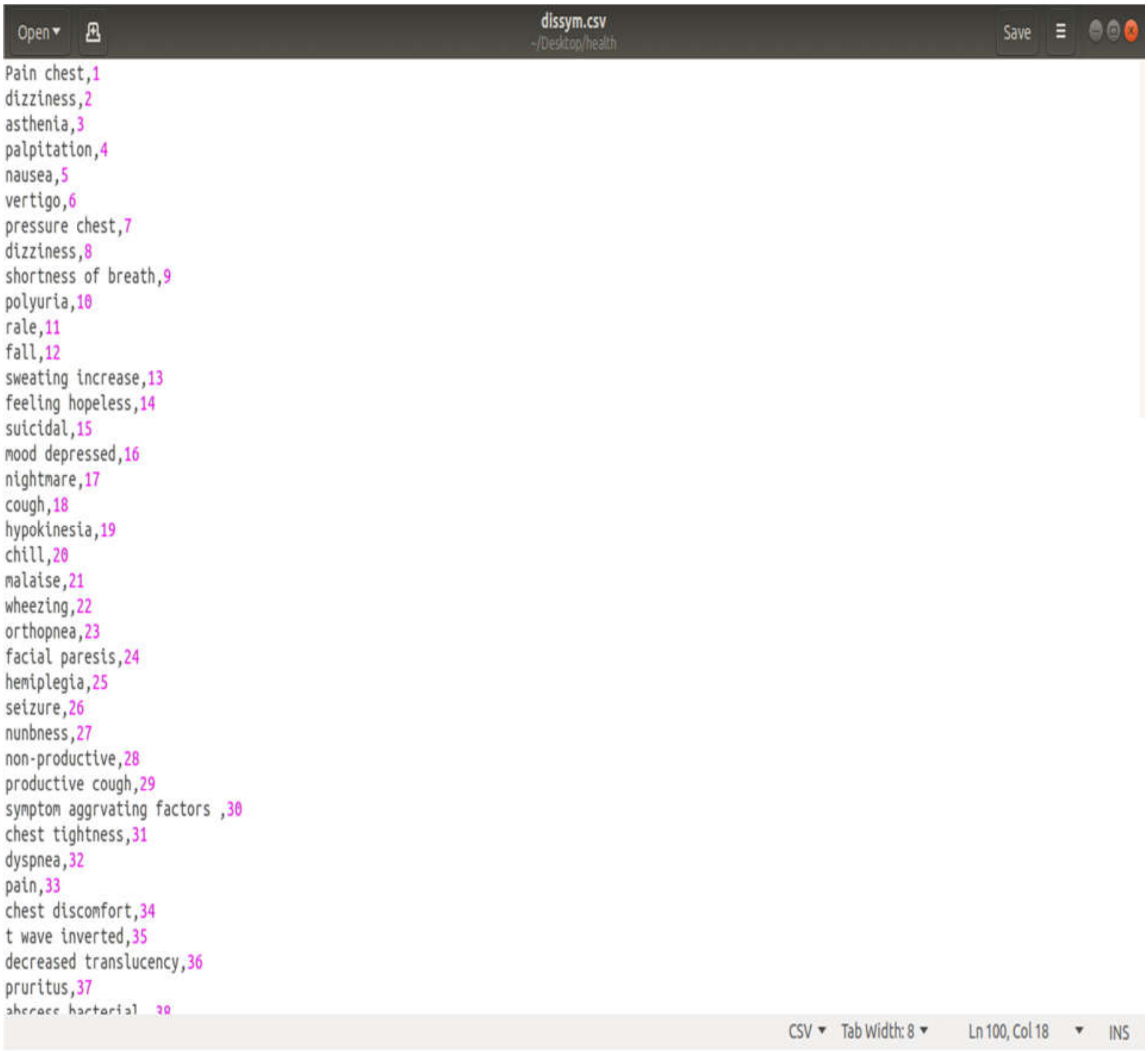


Figure 1: System architecture for proposed system

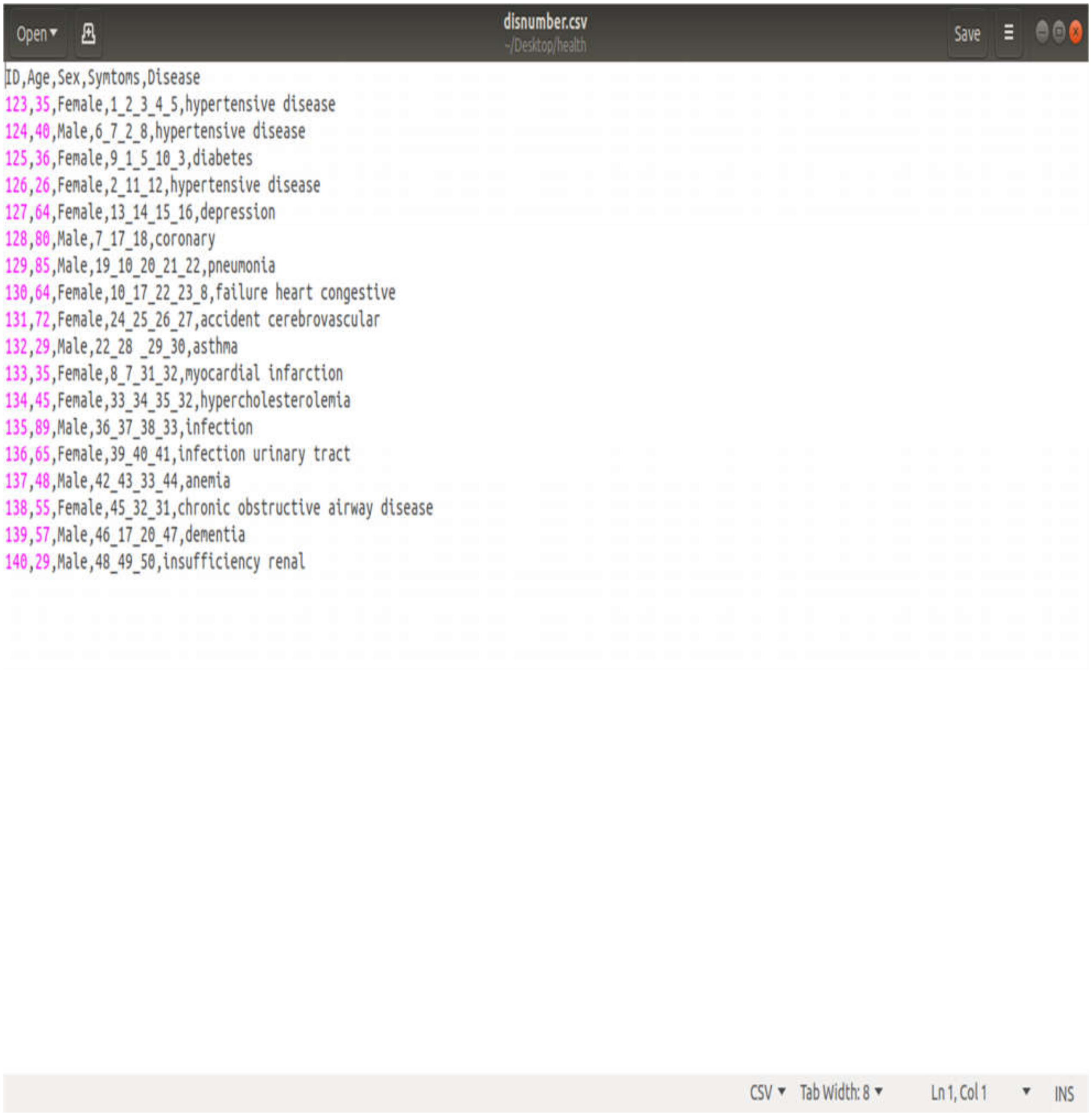
The data in the health care sector is growing rapidly and is coming from various domestic as well as exterior sources like portable devices, wearable sensor devices, medical notes, social media etc. The effective analysis of the present health data can help in offering newer answers to the present diseases

III. RESULT ANALYSIS



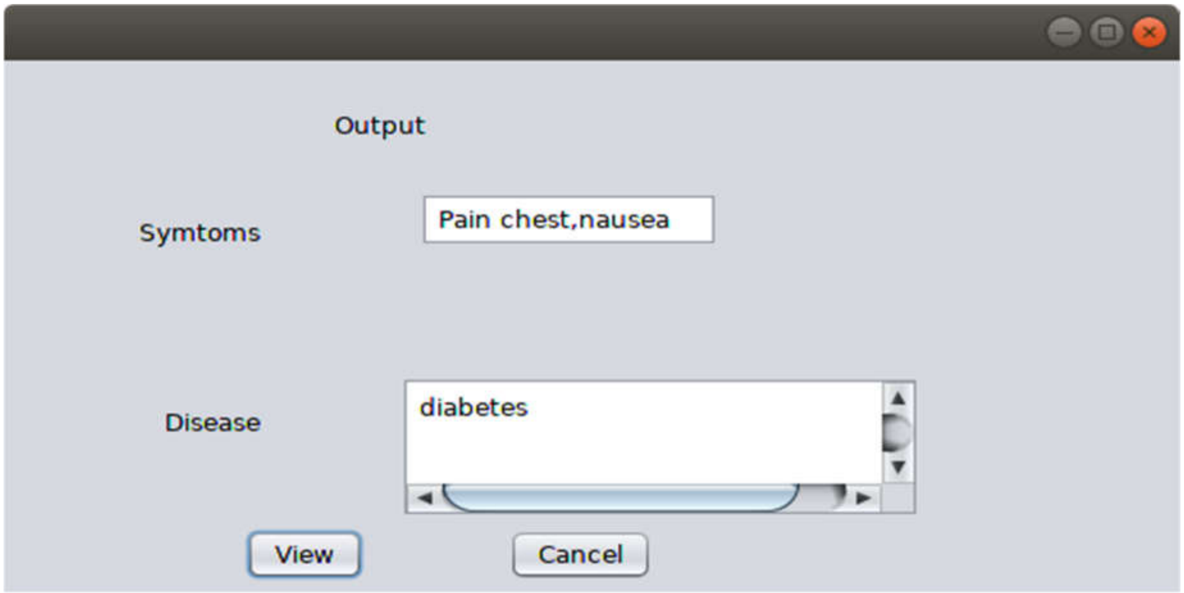
```
Pain chest,1
dizziness,2
asthenia,3
palpitation,4
nausea,5
vertigo,6
pressure chest,7
dizziness,8
shortness of breath,9
polyuria,10
rale,11
fall,12
sweating increase,13
feeling hopeless,14
suicidal,15
mood depressed,16
nightmare,17
cough,18
hypokinesia,19
chill,20
malaise,21
wheezing,22
orthopnea,23
facial paresis,24
hemiplegia,25
seizure,26
nunbness,27
non-productive,28
productive cough,29
symptom aggrvating factors ,30
chest tightness,31
dyspnea,32
pain,33
chest discomfort,34
t wave inverted,35
decreased translucency,36
pruritus,37
abscess bacterial ,38
```

There are various symptoms where each symptoms has been assigned with unique identity number. The number have been assigned in order to run the K-means clustering algorithm successfully.



```
ID,Age,Sex,Symtoms,Disease
123,35,Female,1_2_3_4_5,hypertensive disease
124,40,Male,6_7_2_8,hypertensive disease
125,36,Female,9_1_5_10_3,diabetes
126,26,Female,2_11_12,hypertensive disease
127,64,Female,13_14_15_16,depression
128,80,Male,7_17_18,coronary
129,85,Male,19_10_20_21_22,pneumonia
130,64,Female,10_17_22_23_8,failure heart congestive
131,72,Female,24_25_26_27,accident cerebrovascular
132,29,Male,22_28 _29_30,asthma
133,35,Female,8_7_31_32,myocardial infarction
134,45,Female,33_34_35_32,hypercholesterolemia
135,89,Male,36_37_38_33,infection
136,65,Female,39_40_41,infection urinary tract
137,48,Male,42_43_33_44,anemia
138,55,Female,45_32_31,chronic obstructive airway disease
139,57,Male,46_17_20_47,dementia
140,29,Male,48_49_50,insufficiency renal
```

Here the description of the dataset is given where we get the exact details whether the user is Male or Female. We also get the details about the age of the user, symptoms entered by the user. The above dataset has been prepared by mapping with the data shown in the previous figure. The dataset at last has been concluded with the disease related with entered symptoms.



Output

Symtoms: Pain chest,nausea

Disease: diabetes

In above shown window accepts the symptoms as the input on which K-means algorithm runs and result for that symptoms appears in next text view.

IV. Conclusion

The big data is a growing technology which maintains all kind of data's in the real world. Though there are several challenges like combining heterogeneous data, infrastructure issues, insufficient real time processing, data quality that must be addressed, Big Data has the potential to transform and revolutionize the way healthcare systems use technologies to gain valuable insight from the data repositories. In the future we are sure to see widespread use of big data analytics across the different areas of healthcare industry. This paper provides the mechanism to improve the quality of big data analytics by improving the performance quality of map reduce, whose proper selection can give promising results. Big data is helps to maintain the health care analysis and gives the wonderful results to the user. In this proposed research work, the objective is to develop the system in which the result should be depends on data we input if the data is correct then it should give consistent result.

V. FUTURE SCOPE

We can extend to this system in the way that , for only particular disease we could collect the symptoms form the patient and then predict the percentage or chances of that disease could be occur. More refined techniques for data pre-processing, in order to extract required information efficiently comprises future work on this study. In order to generate substitute methods and various clustering techniques which can be investigated for further enhanced analysis, other algorithms for pattern detection shall also be incorporated in the system.

VI. REFERENCES

1) Wimalasiri, J. S., Ray, P. and Wilson, C. S., "Maintaining Security in an Ontology Driven Multi-Agent System for Electronic Health Records", Proceedings of the IEEE Healthcom 2004, Odawara, Vol 3, June 2004. Page no:47-52

2) Sreekanth et al. "MapReduce Program to Efficiently Analyse Big Data Electronic Health Records Database using Hadoop Cluster on Amazon Elastic Compute Cloud", International Journal of Advanced Research in Computer Science and Software Engineering. Volume 5, Issue 8, August 2015.Page no: 774-777.

3) HAN HU, YONGGANG WEN, TATSENG CHUA, AND XUELONG LI, Toward Scalable Systems for Big Data Analytics, Vol: 2, April2014. Page No: 652658

4) Sreekanth Rallapalli "Improving Healthcare-Big Data Analytics for Electronic Health Records on Cloud" IEEE Jounal of Advances in Information Technology Vol. 7, No. 1, February 2016.Page no:65-68

5) Haritha Chennamsetty"Predictive Analytics on Electronic Health Records (EHRs) using Hadoop and Hive". IEEE Journals of Advances in Computer Engineering. Vol 9, Sept 2009, Page no: 978-982.

6) Dr. Mahendra P. Dhore, Madhura A. Chinchmalatpure" Review of Big data Challenges in Healthcare Application" IOSR Journal of Computer Engineering (IOSR-JCE)., NCRTCSIT-2016, e-ISSN: 2278-0661,p-ISSN: 2278-8727, PP 06-09.

7) V. Kavitha, S. Kannudura" Health Care Analytics with Hadoop Big Data Processing" International Journal of Advanced Research in Computer and Communication Engineering. Vol. 5, Issue 5, May 2016.