Diseases Inference from Health-Related Information Via Map Reducing

S.Sandhya[1] P.Ranjitha[2] S.Thaiyal Nayagi[3] A.Kalai selvi[4]
[1,2,3]Student [4]Assistant Professor
[1,2,3,4]Department of Information Technology Engineering
[1,2,3,4]Dhanalalakshmi College of Engineering

Abstract— The intention of the work is to ensure the correct diagnosis of any illness with the help of decision support system. Patient's Health records (PHR's) are maintained in the public cloud where each and every patient is provided with an ID. All the data stored should be encrypted. Different encryption keys should be used on different machines and the key information should be stored centrally behind strong firewalls. This way even if a hacker is able to get the data, he cannot extract meaningful information from it and misuse it. Patient data will be stored securely in an encrypted manner. The main goals are ease of retrieval/collection of the specific information, less time consumption, cost effective, scalable, Fault tolerant and increase in security.

Key words: BigData, Map Reduce, Diseases prediction, HDFS, Encrypted details

I. Introduction

The upcoming challenge in healthcare is "working with big data in hospital systems is hugely challenging but at the same time holds tremendous promise in providing more meaningful information to help clinicians treat patients across the continuum of care". The challenge is how to ensure data confidentiality and integrity when storing such data but still make it highly available, process it to extract actionable information for decision makers,including medical professionals, and share it with collaborators, while preserving the privacy of individual patients and giving them the full control of their data at all times.The doctor predicts the disease based upon the patient's unique ID and symptoms by updating the patient record. Previously the trained data is used to predict the disease based upon the symptoms. In this dataset, the symptoms are collected and categorized under some predictable diseases. It uses the classificationand clustering techniques for predicting the disease.

II. Existing System

The patient's records are maintained manually and it consumes more time for the doctors to diagnose the patient. The medical records of the patient are stored in a separate room. Even though it contains the electronic medical records they are just maintained in a database with less security. The records are stored in a local database and it contains only the structured data and uses the traditional database (RDBMS) and hence the details are not shared. Diseases prediction consumes some time. It is necessary to find some diseases as early as possible to save human life.

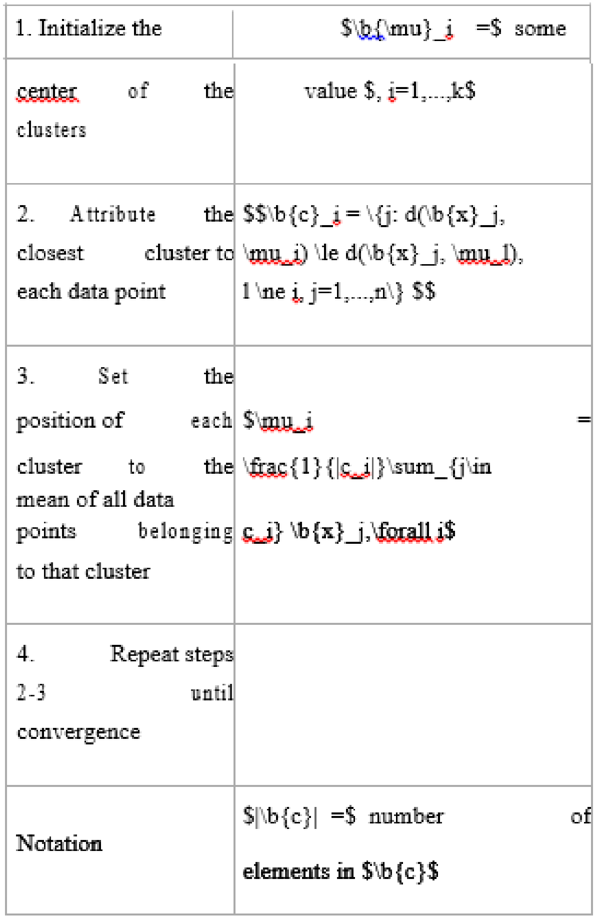
III. Proposed System

Hadoop is used to predict the disease based upon the symptoms. The patients are provided with the unique ID. The Patient's Health Record (PHR's) of the patient are stored in the public cloud. Since the PHR contains the sensitive information each and every patient records are encrypted using the Homomorphic based encryption. When the PHR is needed, they are retrieved from the cloud by decrypting it with the key. So, this results in providing the confidentiality to the data.

IV. Methodology and Materials:

A. K-Mean Clustering:

Clustering is the process of partitioning a group of data points into a small number of clusters.This algorithm is used to predict the diseases from the default symptoms dataset.Here dataset is divided into clusters according to symptoms of the patient.



1. Initialize the	$\b{\mu}_i =$ some
center of the clusters	value $, i=1,...,k$
2. Attribute the closest cluster to each data point	$$\b{c}_i = \{j: d(\b{x}_j, \mu_i) \le d(\b{x}_j, \mu_l), 1 \ne i, j=1,...,n\} $$
3. Set the position of each cluster to the mean of all data points belonging to that cluster	$\mu_i = \frac{1}{\|c_i\|}\sum_{j\in c_i} \b{x}_j,\forall i$
4. Repeat steps 2-3 until convergence	
Notation	$\|\b{c}\| =$ number of elements in $\b{c}$

B. Map Reduce Technique:

In the Map Reduce programming paradigm, the input patient details is splitted into many pieces and these pieces are given to map processes running on different machines. Then the outputs of these map processes are given to many reduce processes which are the final stages of the execution. This is used for fast retrieval of patient details during emergency.

Steps of Algorithm

1) Input

- Data are loaded into HDFS in blocks and distributed to data nodes
- Blocks are replicated in case of failures
- The name node tracks the blocks and data nodes

2) Job Submits the job and its details to the Job Tracker
3) Job initialization

- The Job Tracker interacts with the Task Tracker on each data node
- All tasks are scheduled

4) Mapping

- The Mapper processes the data blocks
- Key value pairs are listed

5) Sorting the Mapper sorts the list of key value pairs
6) Shuffling

- The mapped output is transferred to the Reducers
- Values are rearranged in a sorted format

7) Reduction Reducers merge the list of key value pairs to generate the final result
8) Result

- Values are stored in HDFS
- Results are replicated according to the configuration
- Clients read the results from the HDFS

C. Homomorphic Based Encryption:

This algorithm is used to store patient details in Hive database in encrypted format and hence hackers only get the encrypted data not the original details. This detail is decrypted by the Authorized doctors by private key.

Steps of Algorithm

- The dimension n, which is a power of 2.
- The cyclotomic polynomial $f(x) = x n + 1$.
- The modulus q, which is a prime such that $q \equiv 1$ (mod 2n), Together, n, q and f(x) define the rings R , Z[x]/ hf(x)i and Rq , R/qR = Zq[x]/ hf(x)i.
- The error parameter σ, which defines a discrete Gaussian error distribution χ = DZn,σ with standard deviation σ.
- A prime $t < q$, which defines the message space of the scheme as Rt = Zt[x]/ hf(x)i, the ring of integer polynomials modulo f(x) and t.
- A number $D > 0$, which defines a bound on the maximum number of multiplications that can be performed correctly using the scheme. These parameters will be chosen (depending on the security parameter κ) in such a way as to guarantee correctness and security of the scheme.

V. EXPERIMENTAL WORK

A. Administration:

The role of admin is to maintain all PHR's of the patient on the regular basis. Hospitals admin's are responsible for the day-to-day operation of a hospital, clinic, manages care organization or public health agency. Normally patients contact the hospital admin in order to register their details. The admin collects the details from the concerned patients and maintain it in a cloud database. Related to the project, the admin will be having the electronic registration form for the new patient's. The registration form contains the personal details of the patient like name, address, male/female, DOB, Age, mobile, height, weight, Drugs, Symptoms- High Fever/normal, Considerable weight loss, cough, yellow correlation of eyes, head ache, sweats, anaemia, vomiting, etc., The patient provide his/her personal details and the symptoms suffering from. And the PHR of the patient gets stored in the cloud database. There may be even millions of hospitals in the world where the admin's in each and every hospital performs the same job.

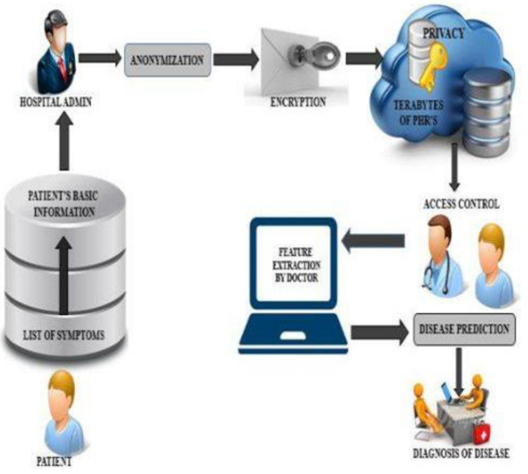
B. Doctor Consultation:

After the login, the patient follow up page will be displayed to the doctor. In that particular page the doctor can able to view the new and the existing patient's information. The new patient's record contains the personal information and the symptoms that they are suffering from. The existing patient record contains the test report values that are taken before and the disease that they have diagnosed by other doctor.The doctor has the authority to make the changes in the patient's record .When the patient visits the regular doctor where he/she knows well about the patient condition and when the doctor is not available the patient who visits the other doctor can provide her ID so that the temporary doctor can gain more knowledge regarding the patient's details within a few seconds. This helps the doctor to better perform with more patients in a lesser time. When the consultation time gets over, the doctor logs out the screen when he /she leaves the hospital.

C. Diseases Prediction:

The doctor diagnoses the existing patients where they are informed to take the prescribed test. When the patient visits the doctor again he/she just provide the unique ID by which the doctor gets the information about the patient and the input values are provided from the test reports. From this, the doctor predicts the disease based upon the patient's unique ID and symptoms by updating the patient record. Previously the trained data is used to predict the disease based upon the symptoms. In this dataset, the symptoms are collected and categorized under some predictable diseases. It uses the classification and clustering techniques for predicting the disease. Now the patients undergo the medical test given by doctor and provide the test result to trained dataset file. Here the doctor analyses the symptoms and compare the trained dataset with it and finally predict the kind of disease. By this the doctor diagnoses the patient and updates the PHR of the patient.

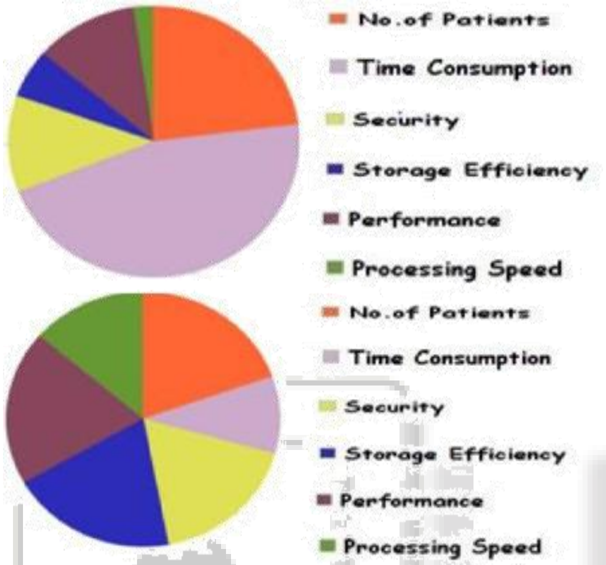
VI. ARCHITECTURAL DIAGRAM



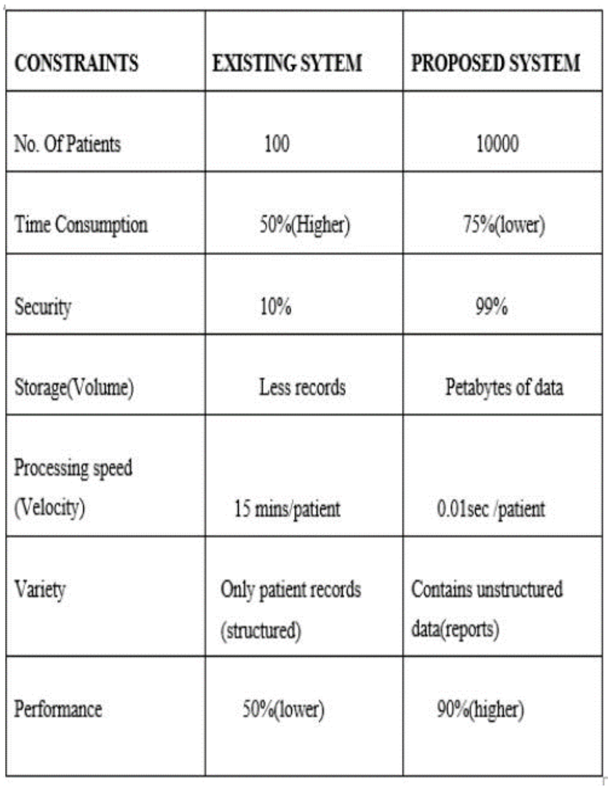
VII. RESULT AND DISCUSSION

To evaluate the performance of proposed methodology, patient's dataset is used. The dataset consists of the patient's personal information and also the health regarding information. Thus, with the help of Big Data the doctor can easily predict the disease and diagnose it, which results in less time consumption. In addition, the large number of patient's can be consulted by the doctor and gets diagnosed soon. It also results in large storage of PHR and contains increase in processing speed. In comparison with the existing system, the performance is high in the proposed system.

VIII. EXISTING SYSTEM:



The comparison diagram involves the difference between the existing system and the proposed system. For example, Here we consider a table which involves the No. of patients treated per day, Based on that it produces the time consumption to treat the patient by the doctor. How the PHR of the patients are stored securely and efficiently. The performance also gets increased in case of the proposed system. The processing speed of retrieving the records makes the difference between the two. The content of the table involves the following details.



CONSTRAINTS	EXISTING SYTEM	PROPOSED SYSTEM
No. Of Patients	100	10000
Time Consumption	50%(Higher)	75%(lower)
Security	10%	99%
Storage(Volume)	Less records	Petabytes of data
Processing speed (Velocity)	15 mins/patient	0.01sec /patient
Variety	Only patient records (structured)	Contains unstructured data(reports)
Performance	50%(lower)	90%(higher)

From the above table it is illustrated that the existing system consists of only 100 patients where the proposed system consists of 10x times greater number of patients. This specifies the time taken by the 100 patients will be 50% higher than that of the proposed system. Regarding the security the patient's records are less secured as they are placed in a separate rooms or in a traditional database. Whereas in the proposed the patient's are given a unique ID which can't be known by any other persons. The processing speed for diagnosing the patient will be high in proposed system by 0.01sec/patient than existing system which take 15min/patient approximately. By this the performance of the proposed system will be by 90% than the existing system which contains lower performance by 50% approximately.

IX. CONCLUSION AND FUTURE ENHANCEMENT

An approach is proposed for predicting the disease based upon the symptoms with the use of Big Data.The terabytes of patient health records are maintained in hive database which help clinicians to predict the correct diagnosis of any illness of the patient by the process of decision support system. Hadoop helps in retrieving the information of the patient with the high processing speed. Here it contains the high volume of PHR's in the database. It also contains the structured, unstructured and semi-structured data in the patient's record.

In future enhancement, add n number of symptoms in dataset to predict even more new diseases.So new diseases can also be predicted in future.

REFERENCES

[1] Ahsananun Nessa, Moshaddique Al Ameen, Sana Ullah, Kyung Kwak (2010),"Applicability of Telemedicine in Bangladesh: Current Status and FutureProspects" The International Arab Journal of Information Technology Vol. 7,pp.138-145,

[2] Pantelopoulos and N. Bourbakis, (2010), “A survey on wearable sensor-based for health monitoring and prognosis” IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 40, no. 1, pp. 1–12.

[3] Chen and Hui Yang*, Member (2014),”Heterogeneous Postsurgical Data Analytics for Predictive Modelling of Mortality Risks in Intensive Care Units”.

[4] P. Groves, B. Kayyali, D. Knott, and S. Van

[5] Kuiken, (2013),“The big data Revolution in Healthcare,” McKinsey & Company.

[6] Marco Viceconti, Peter Hunter and Rod Hose (2015), “Big data, big knowledge: BigData for Personalized Healthcare”.

[7] Xiaoyong Xu, Fen Yu, Jiannong Shi, Institute of Psychology (2010),”Personality and Abusive Supervision: A Study on Leadership in the HealthcareIndustry in China” -2010 3rd International Conference onBiomedical Engineering and Informatics (BMEI 2010)

[8] Projections Team, (2008),"Health and spending projectionsthrough 2017: the baby-boom generation is coming to Medicare, “Health Affairs, vol.27, pp.w