International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 03 | Mar-2018 www.irjet.net p-ISSN: 2395-0072

"Analyse Big Data Electronic Health Records Database using Hadoop Cluster"

Mr. C.S. Arage[1], Mr. M.P. Gaikwad[2], Mr. Rohit Tadasare[3], Mr. Ronak Bhutra[4]

[1,2,3,4] *Sou. Sushila Danchand Ghodawat Charitable Trust's Sanjay Ghodawat Group of Institutions. Computer Science and Engineering, Shivaji University Kolhapur, Maharashtra, India*

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 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.

Keywords: Big Data, HealthCare Solutions, Hadoop Use Cases, Clinical Decision Support.

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 analyse 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.

II. LITERATURE REVIEW

The increasing digitization of healthcare information is opening new possibilities for providers and payers to improve the quality of care, health care results, and minimize the costs. The latest tools and technologies are used on digital information of health care organizations can generate valuable insights. Organizations must also analyse internal and external patient information to more accurately measure risk and outcomes. At the same time, many providers and payers are working to increase data transparency to produce new insight knowledge. Existing analytical techniques can be applied to the vast amount of existing (but currently unanalysed) patient related health and medical data to reach a deeper understanding of results, which can be applied at the point of care. Ideally, these data would inform each physician and their patients during the decision-making process and used to identify the appropriate treatment option for that particular patient.

A. Tools and Application in Health Care System

The health care system has a large volume of unstructured data, so it is impossible to do research and diagnoses without an appropriate tool or technique. Hadoop is a tool that is designed to process huge volumes of data, which is integrated with map-reduce concept. Map reduce can divide the data set into multiple chunks, each will be processed in parallel among multiple nodes. MapR can overcome the limitation of Hadoop, as it has dynamic read-write data layer that provides unparalleled dependability.

B. Application of Big Data in Health Care:

1. **Personalized Treatment Planning**: Based on the medical histories of every individual patient, diagnoses can be done, which can be used to decide the appropriate treatment and medicine for that patient. Real time analysis will be done using MapR and Hadoop, based on the analytics results, the patient can have personalized care for them.

2. Assisted Diagnosis: Physicians can isolate and treat the patient based on some factors like symptoms, medical history, and side effects. Using prediction modeling and Hadoop can provide information which will be helpful to the doctors.

3. Utilization Review: To assist evidence-based treatment, which is considered to be the best form of treatment, the big data analytics of health information are required. The analysis can be further improved by getting information from non-traditional sources like social and other electronic media for more insightful information using big data analytics tools and techniques like Hadoop and MapReduce.

C. K-Means Clustering:

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:

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

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

III.CONCLUSION

Today healthcare providers, payers, physicians generate huge data which requires analysis to provide better health care to patients. Patients can have personalized health care and in order to reduce the cost analytics play a major role. In this paper we have seen the importance of Big Data analytics in health care. MapReduce model is an efficient programming paradigm for processing such huge data. . Efficient in MapReduce model can be increased by improving the efficiency of algorithm. In this paper, we propose an algorithm where the EHR database can be handled efficiently with minimum time taken for reading the database. The time taken to reduce the records and write the results back to the file is also minimized. The MapReduce algorithm run on Hadoop clusters.

IV. 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, TAT-SENG CHUA, AND XUELONG LI, Toward Scalable Systems for Big Data Analytics, Vol: 2, April2014. Page No: 652-658

4. Sreekanth Rallapalli "Improving Health care - 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: 978-98.