ORIGINAL RESEARCH PAPER

Chemical Science

AN ANALYTICAL SURVEY OF USAGE OF BIG DATA AND HADOOP FOR PREDICTION OF DISEASES

KEY WORDS: Big Data, Hadoop, Diseases

Ms Priti D. Sadaria — Department of Computer Science and Information Technology, Rajkot

Dr. Achyut C Patel* — Smt. M.T. Dhamsania Commerce College, Rajkot *Corresponding Author

ABSTRACT

Now a day no field remain untouched with Information Technology. Health care industries are using Information Technology for different purpose. Health data growing quickly because of fast acceptance of Information Technology. Extraction of useful information by analyzing this rapidly growing data for building a useful model which can be applicable in real life is really a challenging task. Knowledge discovery and decision making from such voluminous data is a new trend that is Big Data Computing. Machine learning techniques can be used to make predictive analytics. Cloud computing provides computing services over the internet which includes servers, storage, databases, software and analytics for big data processing. Now a day, analysis of diabetic Big Data is facing lots of problems because of unpredictable growth of data which leads to a big challenge in processing the large and complex datasets manually.

Hypothesis: To propose an algorithm for prediction of diseases like Nephropathy, Retinopathy and Cardio Vascular Diseases and as a result treatment can be given at proper time.

Methodology: This research focuses on the significance of Big Data tools and machine learning algorithms for clustering and classification to develop a predictive algorithm for predicting diabetic related diseases by using Hadoop platform.

INTRODUCTION:

The use of public healthcare data and analysis plays an important role for the healthcare planners. Since last few years due to rapid adoption of Information Technology in healthcare systems, the health data increase exponentially and data is available in different forms. Healthcare providers can use healthcare data and analytics to learn more about patients and can enhance preventive care by utilizing key data. Healthcare data and information are major helpful sources for taking effective decisions. From such rapidly growing huge data, Knowledge discovery and decision-making is a challenge concerning both data organization and timely processing. Big Data computing combines large-scale computing with machine learning techniques is used to build predictive analytics for prediction of some chronic diseases.
To process large amount of Big Data, by human inspection is impossible and because of this reason, it is required to develop a high-speed system to manage such Big Data. A variety of Big Data tools are presently being used to scrutinize data at a faster rate and present users with essential knowledge for decision-making and prediction.

Storing and processing quickly growing data within the specific time using conventional tools is a difficult task. Capable tools are required for data management and analysis for the data which is collected from a lot of applications in scientific and business field [1]. Big Data provides tools to create, manipulate and manage large datasets.

In recent era a service-oriented computing model – Cloud computing is mostly used for processing large volumes of rapidly increasing data at a faster scale and it is actually a demand for Big Data computing. So to fulfill this requirement, Big Data framework Hadoop and Spark are used to complete Big Data tasks with machine learning techniques. This proposal focuses on predictive analytics by using different machine learning strategies to analyze Big Data and as a result of this one can make decisions about future complications of diabetic patients. I will be focus in my research work on a framework for Big Data computing along with Hadoop Map Reduce.

The clinicians can prevent patients from undergoing ineffective treatments or can allow better treatment by utilizing the existing medical data out of Big Data analytics and can come up with potential outcome. At present, Cloud-based Big Data analytic environment is necessary to analyze semi structured, structured and unstructured Big Data from healthcare sector.

Depending on the present life style attributes, data sets can be collected and can be process at a faster rate by developing a new algorithm, it is possible to predict diabetic related diseases. MapReduce is commonly used to process large amount of data at the same time as hiding the complexity of parallel execution across hundreds of servers in a Cloud environment in distributed cloud computing.

Now a day diabetes Mellitus has become a worldwide threat. It is one kind of clinical disorder due to the deficiency insulin. The present research is focus on to predict patients with diabetic, who may have the menace of having Cardio Vascular Disease, Nephropathy, and Retinopathy well in advance using new algorithm on Hadoop plat form in Cloud. By using this new algorithm timely treatment can be given to the patient before dangerous stage.

Objectives of the Research:

- To analyze diabetic Big Datasets in standalone mode using Hadoop platform
- To analyze diabetic Big Data using MapReduce and Spark
- The clustering and classifying of diabetic Big Datasets are to be carried out by using machine learning methods like K-means clustering and SVM classification
- To begin parallel processing of Hadoop in AWS Cloud

Research Methodology:

The literature review give the clarity to understand the standard methodologies like Apache Hadoop MapReduce, Cloud computing, K-means, Support Vector Machine, Hybrid algorithm, RStudio for diabetic Big Data analysis. Big data architecture includes mechanisms for protecting, processing, and transforming data into file systems or database structures.

Review of Literature in the area research:

Wei Fan and Albert Bifet [2] had focused on Big Data mining

from which useful information can be obtained. In the past, data mining operations on large amount of data was not possible. But currently, with the help of software like Apache Hadoop, it is possible. Even in addition to Apache Hadoop, there are Big Data tools like R, MOA, Strom, Vow pal Wabbit which are a few open source software to deal with Big Data. In order to develop smart cities for better services and better customer experience now a day Big Data mining applies to healthcare, technology and business.

In the research, Sachidanand and Nirmala(2013)[3] have given detailed information about the analysis of Big Data and its significance in various fields like manufacturing, healthcare, public sectors, retail, etc. Sadhana and Savitha Shetty(2014) proposed a prediction model to analyze the facts regarding diabetic dataset[4].Due to the enormous size of the data and complexity in the current era, Vikram Phaneendra & Madhusudhan Reddy(2013) recommended HDFS as a appropriate tool for data processing [5]. At present, enormous information is created through many sources using web servers to become huge data set which forms a challenging task to extract useful information out of it.

Every day huge volume of patient data is produced from doctor's notes, doctor's prescription, clinical reports, and even from body sensors. Still the analysis of healthcare parameters and the prediction of the consequent future health conditions are in the toddler stage. In his research as Islam et al.(2015) has discussed that in smart cities, the patients are using variety of electronic devices like mobile networks. Through the use of internet the smart devices help to monitor the health status of the patients in hospitals and outdoor environment. Due to this, huge volume of data can be gathered [6]. Big Data analytic platform is the best way to analyze the structured and unstructured data generated from healthcare management systems through the cloud.

Dhavapriya et al have focused on MapReduce framework and HDFS in the file indexing with mapping and reducing in implementing Big Data analysis [7]. A new technique was developed by Chen He Ying Lu David Swanson [8] to improve map task's data locality and has integrated this technique into Hadoop default FIFO scheduler and Hadoop fair scheduler. Augustine [9] described regarding the contribution of Big Data Analytics and Hadoop to provide the services of healthcare to everyone with best possible cost.

In his research, Weng et al.(2016) have indicated that in the recent year,various disease prediction models have been proposed. For the treatments of patients, the prediction of future disease is very crucial and important [10]. Many partitioning techniques of clustering algorithms and their advantages and applications initiated by Gopi Gandhi and Rohit Srivastava [11]. In his research work Park et al.[12] have developed a modern PDA based personal diabetes management system. The diabetes patients require both self-test and emergency tests kits for urgent use.

There is a strong need for research in diagnosis of diabetes which will be helpful for medical practitioners and patients. Expert systems for both diagnosing and treatment of diabetes can be analyzed. Syeda Farha Shazmeen et al.[13] have worked with different data mining applications and various classification algorithms and came to the conclusion that the competence of the algorithm can be enhanced by applying data pre-processing techniques.

Hypothesis:

To propose an algorithm for prediction of diseases like Nephropathy, Retinopathy and Cardio Vascular Diseases and as a result treatment can be given at proper time.

CONCLUSION:

Now a day, analysis of diabetic Big Data is facing plenty of problems due to unpredictable growth of data leading to big challenges in processing the large and complex datasets manually which is converted in problem known as Big Data trouble. Extracting useful information from this enormous amount of data is highly complex, costly, and time consuming and therefore the problem can be solved by processing huge amount of data by applying machine learning techniques on Hadoop platform in Cloud environment.

The analysis of diabetic Big Data for the classification of diabetic patients and identification of high risk patients is most important to predict chronic diseases, control the diseases at an early stage and improve the medical care. The urgent need is the high speed processing of diabetic Big Data. The outcome of the study will reduce the risk of diabetic related diseases, improves patient well-being and reduce the death rate. Further, detecting at an early stage of diabetic disease will help the public to get a precise medical care.

REFERENCES:

1. Chen C.L.P. and. Zhang, C.Y Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Inform.Sci (2014), http://dx.doi.org/10.1016/j.ins.2014.01.015
2. Wei Fan and Albert Bifet. "Mining Big Data: Current Status, and Forecast to the Future", SIGKDD Explorations. 14(2).
3. SachchidanandSinghandNirmala 2013 International Conferenceon Computing Technology (ICCICT). IEEE,2011. Singh. Big dataAnalytics", Communication, Information&
4. Sadhana and Savitha Shetty, "Analysis of Diabetic Data Set Using Hiveand R", International Journal of Emerging Technology and Advanced Engineering,4(7),2014.
5. Vikram Phaneendra,S.& E.Madhusudhan Reddy, "Big Data- solutions for RDBMS problems-A survey". In12th IEEE/IFIP Network Operations & Management Symposium. 2013.
6. Islam,S. M. R D. Kwak, M. H. Kabir, M. Hossain, and K.-S. Kwak, The internet of Things for health care: A comprehensive survey," IEEE Access, vol. 3,678-708, 2015.
7. Dhavapriya, M. and Yasodha, "Big Data Analytics: Challenges and Solutions Using Hadoop, Map Reduce and Big Table", International Journal of Computer Science Trends and Technology (IJCST), 4(1):5-14.2016.
8. Chen He Ying Lu David Swanson "Matchmaking: A New MapReduce Scheduling" in 10th IEEE International Conference on Computer and Information Technology (CIT'10),2736–2743.2010.
9. Peter Augustine D., Leveraging Big Data Analytics and Hadoop in DevelopingIndia's Healthcare Services, International Journal of Computer Applications, 89(16),2014.
10. Weng C.H, T. C.-K.Huang, and R.-P. Han, ``Disease prediction with different types of neural network classifiers," Telematics Inform., vol. 33,no. 2, pp. 277_292,2016
11. Gopi Gandhi and Rohit Srivastava,"AComparativeStudy on Partitioning Techniques of Clustering Algorithms", International Journal of Computer Application, 2014.
12. Kyung-Soon Park, Nam-Jin Kim, Ju-Hyun Hong, Mi-Sook Park, Eun-Jong Cha, Tae-soo Lee,PDA based Point-of-care Personal Diabetes Management System, Proceedings of the IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China. 1-4, 2005.
13. Syeda Farha Shazmeen, Mirza Mustafa Ali Baig, M. Reena Pawar, Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis, IOSR Journal of Computer Engineering (IOSR-JCE), 10(6):16-2013