NOVATEUR PUBLICATIONS
INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY [IJIERT]
ISSN: 2394-3696
VOLUME 4, ISSUE 2, Feb.-2017

FRAUD CLAIM DETECTION USING SPARK

I. KARTHIKA
Assistant Professor, Department of CSE , M.Kumarasamy college of Engineering,
karthikai.cse@mkce.ac.in

K. P. PORKODI
Research scholar, Department of CSE ,Al-Ameen engineering college ,
porkodiprabhakaran@gmail.com

ABSTRACT
Objective: To reduce the fraud claims in health insurances companies and to improve outcomes in health care industry Analysis: In the existing system, Apache hadoop and Apache hive is used for processing data, it is a batch processing system. In proposed system, Apache spark is used for processing streaming data. Findings: EHR record is used as data source, it contains unique id for patients across world, so it is very easy to detect fraud claim with help of patientid. Apache spark processing streaming data on regular basis. But in existing system Apache hadoop and Apache hive takes hours of time to process the stored data. Improvement: Rule based model machine learning algorithm is used for detecting and automating the fraud claim and Apache spark is used for fast processing data, so it is more accurate and fast.

RELATED WORK
[1]Machine learning algorithm is used for detecting fraud claim. Support vector machine algorithm is trained to detect the boundary between authorized claim and fraud claim, clustering is done with Evolution Clustering method for clustering. it is highly accurate but highly complex for dynamic data
[2] Bayesian co-clustering technique is used for detecting fraud, fraud occurrence is detected based on unusual behavior. This helps the auditors to know fraudulent unusual behavior. But in case of statistical analysis it become difficult, because it combines medical detection, prevention and response.
[3]outlier detection methods is used for detecting fraud. It checks for deviations from the clusters, trend deviation, single deviation and peek deviations on regression models. Accuracy level is low.
[4]Supervised learning decision tree, neural networks and genetic algorithms in same cases. Outlier detection is used for unsupervised learning in cases for detecting fraud calim. This system is slow and less accurate.
[5]Both supervised and unsupervised technique is used for detecting fraud occurrence. This hybrid model uses efficient association rule mining. It produces efficient result but processing speed is low.

INTRODUCTION
Health care insurance companies uses big data analytics as its game changer for finding fraud occurrence in health care . Health care data set are mostly an unstructured data .These data are very tough to process. Since it is health care there is a need for processing real time data. Apache spark is used for processing real-time streaming data. Traditional approach of finding fraud in insurance claims are storing and analyzing historical data about the patient. Traditional approach of processing data take long time and inaccuracy in finding the fault because of more data. Here we are processing with streaming data so fraud occurrence are identified in real-time.

EXISTING SYSTEM
EDI claim data collected from source and put into Apache hadoop for processing, it is responsible for only batch processing. Datas gathered together for day and put for processing. Apache hive open source tool uses structure data and also apply structure to unstructured data. HiveQL queries is used for processing data[7]. But it also batch processing, so instant fraud detection can't be made. Since it is healthcare industry, on a single day huge volume of data is generated and possible of inaccuracy when handling with more volume of data.

PROPOSED SYSTEM

Electronic health and medical data is used for detecting fault occurrence in health care insurance companies. Each patient is assigned with unique patient id across the world. Apache spark is used for processing instantly on regular updates in medical records and finds the fraud occurrence in same city by using map transformation and Reduce transformation is used to find across the world. Rule based model , Machine learning algorithm is used for automating the process Result displays patientid , city, time, hospital of patient who trying to fraud claim[9][6]. So fraud claim is reduced and it is most accurate when compared the existing system.

MODULES

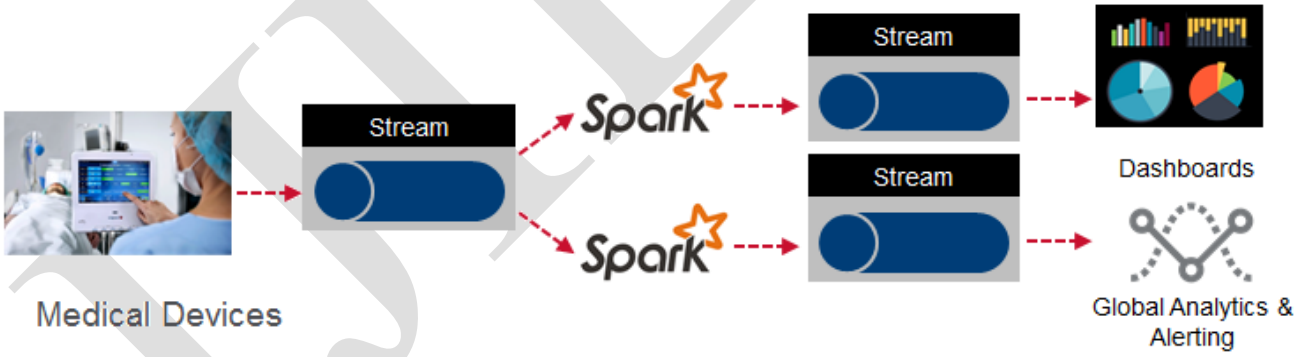
DATA SOURCE

EMR is a main source of collecting health care data's. It will produce real time data. Apart from EMR other sources is also available, digitizing the patient records is Electric health and medical data.EHR technology gets data from the hospitals, labs about personal and medical information of the patient.EHR technology introduces each patients with unique patient id. Every hospital and lab collects personal information about patients are unique id ,Admission Date ,Name, Gender, city ,discharge date and medical information as diagnosis id and diagnosis_ICD10_code.where this diagnosis_ICD10_code. is unique for hospitals in term of diagnosis.

DATA PROCESSING

Information received from hospitals and labs put in to processing. In this decade number of hospitals is high. Since it is digitized update of database is more on seconds basis. If a person admitted in the hospital automatically updated EHR electronic database. information in the electronic database is put into spark for stream process. Patient database is loaded into spark and count the number of patient getting treatment right now.

patientdatabase = sc.textFile('file:///spark/medicaldatasets/patients.csv')
patientdatabase.count()

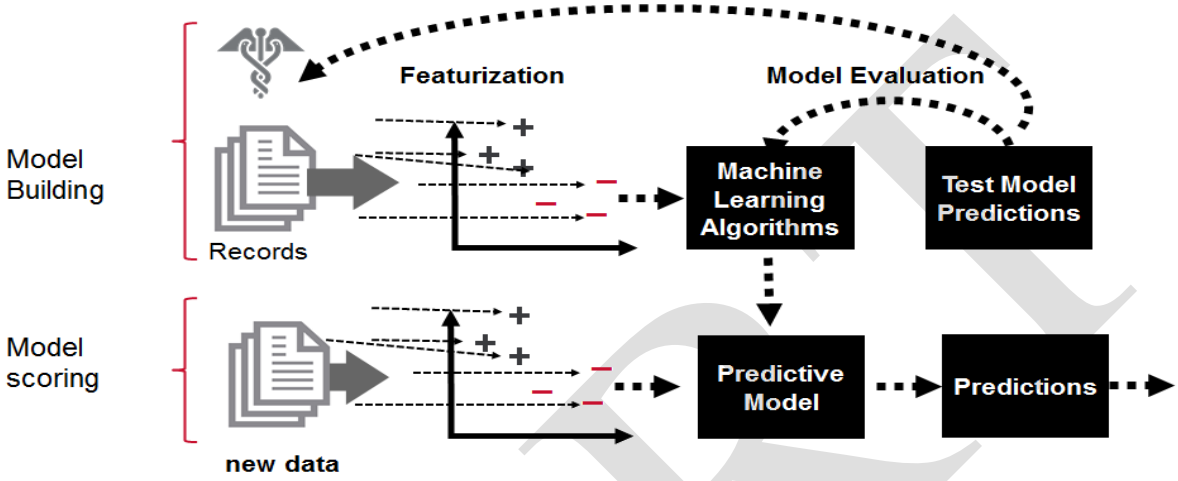


Gives the total count of the user gets treatment right now patient database contains both personal and medical information. To detect the fraud claim RDD transformation map function is used for mapping the patient id and reduce by function is used for knowing the status of the patients treatment according to the city. There is no need processing in head office of insurance companies. Map process can be done in respective branches in cities. So if there is any conflict in the patient id in the same city or branch are identified. once the map process is over it is sent to the head office for reducing, there may be conflict in patient id in various cities so in reduce by key operation it is identified.

patient_city_wise = patient_id.map(lambda line : (line[5],1)).reduceByKey(lambda a,b:a+b).map(lambda line:(line[1],line[0])).sortByKey(False).take(5).

Once data is processed it sent to dashboard for viewing in graphical manner. So it displays the patient id , name, city of the person which shows an conflict or reoccurrence. so we can easily identify the fraud such as claim which is submitted by the same person with same data in different cities. And also with the help of their medical records we can identify the medical claim submitted by the patient is true according their medical record.

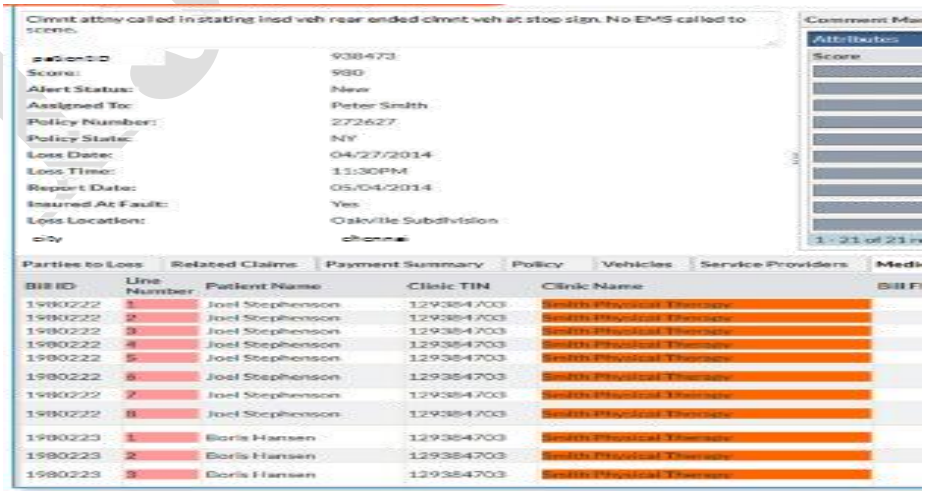
Claims can also be reduced by introducing the machine learning approaches.EHR is used for early diagnosis.EHR consists of 30 million patients data that will help the doctors to know about disease and the treatment given to patients. This will help doctors to make predictions in the treatment.



Model data is put into the machine learning algorithms and again it is put in to test model predictions for predicting. Result given back to the machine learning algorithms for checking features. Based on the result ,score is calculated, model scoring is very helpful for making predictions for new data. If the new data contain patient database with certain symptoms same as old database , predictive model helps the doctors to predict the treatment[8]. So this will reduce the re-admissions in the hospitals. Machine learning approaches is also identifies patient condition and matching treatments, need for re-admission.

RESULT

Processed data is put in to Spring Boot .it is a software for visualizing the result. it will show the Patient Id and city where conflict occurs. Conflict occurrence may be in the same city. Final result show the entire status of the patient. So fraud can be detected easily , when the patient applying for the claim.EHR technology not only helps for detection the fraud occurrence ,but also, giving the efficient treatment to the patients and reducing the readmissions.



CONCLUSION

Health care insurance company consumes 20 millions records from various hospitals, labs. In existing system it tooks 22 hours to process the data which is received on a single day. But in proposed system Apache spark takes only 20 minutes to process the real time data. And also it detects fraud claim not only in the same city across the world. Fraud claim can be identified based on patientid details displays in dashboard. EHR technology also helps in reducing the treatments and readmissions. Fraud claim is highly accurate and fast.

REFERENCES

1) Vipula Rawte, G Anuradha, "*Fraud Detection in Health Insurance using DataMining Techniques", International Conference on Communication, Information & Computing Technology (ICCICT), Jan. 16-17,2015*

2) Tahir Ekina, Francesca Levab, FabrizioRuggeri c, Refik Soyer d, "*Application of Bayesian Methods in Detection of Healthcare Fraud" Chemical Engineering Transactions Vol. 33, 2013*

3) Guido Cornelis van Capelleveen, "*Outlier based Predictors for Health Insurance Fraud Detection within U.S. Medicaid", at the University of Twente December 2013.*

4) Hossein Joudaki, Arash Rashidian, Behrouz Minaei-Bidgoli, Mahmood Mahmoodi,Bijan Geraili, MahdiNasiri "*Using Data Mining to Detect Health Care Fraud and Abuse",Global Journal of Health Science Vol. 7, No. 1; 2015*

5) Shunzhi Zhu, Yan Wang, Yun Wu, *"Health Care Fraud Detection Using Nonnegative Matrix Factorization", International conference on August3-5 2011, Singapore.*

6) Thilagamani, S. and N. Shanthi, 2010. *Literature survey on enhancing cluster quality. Int. J. Comput. Sci. Eng.., 2: 1999-2002. http://www.enggjournals.com/ijcse/doc/IJCSE10- 02-06-26.pdf*

7) S. Chitra, B. Madhusudhanan, G. Sakthidharan, P. Saravanan, *Local Minima Jump PSO for Workflow Scheduling in Cloud Computing Environments, Springer, ISBN 364241673X, 1225–1234, 2014.*

8) E.T. Venkatesh , P. Thangaraj , and S. Chitra , " *An Improved Neural Approach for Malignant and Normal Colon Tissue Classification from Oligonucleotide Arrays ," European J. Scientific Research , vol. 54 , pp. 159 – 164 , 2011*

9) S Saravanan, V Venkatachalam,"*Advance Map Reduce Task Scheduling algorithm using mobile cloud multimedia services architecture" IEEE Digital Explore,pp21-25,2014.*

10) S Saravanan, V Venkatachalam,"*Enhanced bosa for implementing map reduce task scheduling algorithm" International Journal of Applied Engineering Research,Vol 10(85),pp60-65,2015.*