International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.6, No.9, September 2017 **DOI:10.15693/ijaist/2017.v6i11.45-50**

Big Data, Big Knowledge: Big Data for Personalized Healthcare

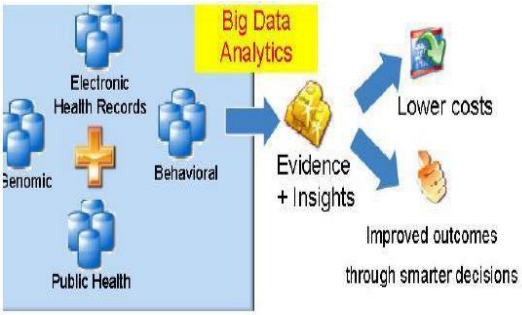
Charles Sherly Supriya[1], G. Radha Devi[2]
PG Student[1], Assistant Professor[2]
Department of Computer Science and Engineering,
Samskruti College of Engineering and Technology,
Kondapur, Ghatkesar, Hyderabad

Abstract Today the healthcare industry is undergoing one of the most important and challenging transitions to date, the move from paper to electronic healthcare records. While the healthcare industry has generally been an incrementally advancing field, this change has the potential to be revolutionarily. Using the data collected from these electronic records exciting tools such as disease recommendation systems have been created to deliver personalized models of an individual's health profile. However despite their early success, tools such as these will soon encounter a significant problem. The amount of healthcare encounter data collected is increasing drastically, and The computational time for these applications will soon reach a point at which these systems can no longer function in a practical timeframe for clinical use. This paper will begin by analyzing the performance limitations of the personalized disease prediction engine CARE (Collaborative Assessment and Recommendation Engine). Next it will detail the creation and performance of a new single patient implementation of the algorithm. Finally this work will demonstrate a novel parallel implementation of the CARE algorithm, and demonstrate the performance benefits on big patient data.

I. INTRODUCTION

A collection of large and complex data sets which are difficult to process using common database management tools or traditional data processing applications. Big data is not just about size. Finds insights from complex, noisy, heterogeneous, longitudinal, and voluminous data. It aims to answer questions that were previously unanswered. Big Data Constantly facing significant challenges like outsized, heterogeneity, noisy labels, and non-stationary distribution.

Capturing, storing, searching, sharing & analyzing. The four dimensions (V's) of Big Data It is important to recognize the full potential of Big Data by addressing these technical challenges with new ways of thinking and transformative solutions. If these challenges are resolved on time, there will be a plenteous opportunities to provide major advancement in science, medicine and business. While there is clearly an important research space examining the fundamental methods and technologies for big data analytics, it is vital to acknowledge that it is also necessary to fund domain-targeted research that allows specialized solutions to be developed for specific applications. Healthcare, in general, deserves to be a natural candidate for this kind of evaluation.



Above diagrammatic representation explains the advantage of the massive amounts of data which provide right intervention to the right patient at the right time. Personalized care to the patient that potent

ially benefit all the components of a healthcare system i.e., provider, payer, patient, and management.

II. LITERATURE REVIEW

M. Viceconti *et al*., [1] described five major problems in healthcare data management systems. These are as follows; 1. Working with sensitive Data. 2. Analytics of complex and heterogeneous data spaces, including nontextual information. 3. Distributed data management under security and performance constraints. 4. Specialized analytics to integrate bioinformatics and systems biology information with clinical observations at tissue, organ and organisms scales 5. Specialized analytics to define the "physiological envelope" during the daily life of each patient. J. Andreu-Perez *et al*., [2] provided an overview of recent developments in big data in the context of biomedical and health informatics. Yunchuan *et al*., [3] promoted the concept of ``smart and connected communities (SCC)", which is evolving from the concept of smart cities. SCC are envisioned to address synergistically the needs of remembering the past (preservation and revitalization), the needs of living in the present (livability), and the needs of planning for the future (sustainability). X. W. Chen and X. Lin [4] has given a brief overview of deep learning, and highlighted current research efforts and the challenges to big data, as well as the future trends. A. Fahad *et al*., [5] performed a survey on a comprehensive study of the clustering algorithms proposed in the literature. In order to reveal future directions for developing new algorithms and to guide the selection of algorithms for big data, they proposed a categorizing framework to classify a number of clustering algorithms. The categorizing framework is developed from a theoretical viewpoint that would automatically recommend the most suitable algorithm(s) to network experts while hiding all technical details irrelevant to an application. L. Xu *et al*., [6] reviewed the privacy issues related to data mining by using a user-role based methodology. They differentiated four different user roles that are commonly involved in data mining applications, i.e. data provider, data collector, data miner and decision maker. A. Belle *et al*., [7] reviewed that the Big Data focused on three areas of interest: medical image analysis, physiological signal processing, and genomic data processing. V. Sujatha *et al*., [8] analyzed that the data sets from statistical models or complex pattern recognition models may be fused into predictive models that combines data set of patients' treatment information and prognostic outcome results. S. Vennila and J. Priyadarshini., [9] promoted that the security in Big data is a challenging research issue. If Integration of MapReduce, a machine for privacy preserving, is designed for the analyzing of data would provide better privacy. Kovalchuk *et al*., [10] represented an early stage of the work aimed to the development of a general-purpose concept of the P4 CDSS rising from a treatment-level scope to a hospital-level scope. J. Cunha, C. Silvaa and M. Antunes [11] proposed a generic functional architecture with Apache Hadoop framework and Mahout for handling, storing and analyzing big data that can be used in different scenarios. Z. Liu *et al*., [12] presented an agent-based model of emergency department that was implemented in Netlogo simulation environment. Case studies have been carried out for proving two of the possible uses of the simulator, one to meet the increasing patient arrival overcrowding problem, and the second a quantitative analysis of the influence of ambulance response time (for departure) over the ED behavior.

M. Srivathsan and Y. Arjun [13] proposed that Prognotive Computing recognize patterns and formulates its own structure to provide a solution or gives a predicted alert so as to find a solution by ourselves .The System provides a handle of Health care and life span of numerous life forms. A. Abbas *et al*., [14] stated that they propose a cloud based framework that effectively manages the health related Big-data and benefits from the ubiquity of the Internet and social media. The framework facilitates the mobile and desktop users by offering: (a) disease risk assessment service and (b) consultation service with the health experts on Twitter. F. Zhang *et al*., [15] proposed a task-level adaptive MapReduce framework. This framework extends the generic MapReduce architecture by designing each Map and Reduce task as a consistent running loop daemon. The beauty of this new framework is the scaling capability being designed at the Map and Task level, rather than being scaled from the compute-node level. Y. Wang, L. Kung and T. A. Byrd [16] examined that health care industry has

not fully grasped the potential benefits to be gained from big data analytics. K. Kambatla *et al*., [17] provided an overview of the state-of-the-art and focus on emerging trends to highlight the hardware, software, and application landscape of big-data analytics. J. Wang, M. Qiu and B. Guo [18] developed a telehealth system that covers both clinical and nonclinical uses, which not only provides store-and-forward data services to be offline studied by relevant specialists, but also monitors the real-time physiological data through ubiquitous sensors to support remote telemedicine. S. M. DeJong [19] proposed that technology is likely to become increasingly important in healthcare. Any professionalism concerns must be weighed against the potential benefits of technology to patients. P. Nadkarni [20] explained that the Institute of Medicine's idea of a learning health system, in which the boundaries between research and clinical practice are blurred.

The historical roots of this idea are identified by exploring initiatives in the business world such as knowledge management, business process reengineering, and enterprise resource planning. M. Legg [21] stated that the standardization required to achieve interoperability for pathology test requesting and reporting. Interoperability is the ability of two parties, either human or machine, to exchange data or information in a manner that preserves shared meaning. A. T. Janke *et al*., [22] explained that clinical research often focuses on resource-intensive causal inference, whereas the potential of predictive analytics with constantly increasing big data sources remains largely unexplored. Basic prediction, divorced from causal inference, is much easier with big data. L.A. Winters-Miner *et al*., [23] predicted the development of a healthcare-centered democracy and seen an explosion in the volume and velocity of patient-generated data. This development has become a driving force in the connection of digital health records to each other and to diagnosis and treatment practitioners.

III. ARCHITECTURE OF HEALTHCARE SYSTEM

Fig.1.Architecture of Healthcare system using Hadoop platform

The architecture of the system includes different sources of data that are as follows

1. Legacy Electronic Medical Records (EMRs)
2. Transcriptions
3. PACS
4. Medication Administration
5. Financial

6. Laboratory (e.g. SunQuest, Cerner)
7. RTLS (for locating medical equipment & patient throughput)
8. Bio Repository
9. Device Integration (e.g. iSirona)
10. Home Devices (e.g. scales and heart monitors)
11. Clinical Trials
12. Genomics (e.g. 23andMe, Cancer Genomics Hub)
13. Radiology (e.g. RadNet)
14. Quantified Self Sensors (e.g. Fitbit, SmartSleep)
15. Social Media Streams (e.g. FourSquare, Twitter)

The data system performs the following processing steps before to transfer data to the YARN data operating system.

1) Loading Healthcare Data

Apache Sqoop is included in Hortonworks Data Platform, as a tool to transfer data between external structured data stores (such as Teradata, Netezza, MySQL, or Oracle) into HDFS or related systems like Hive and HBase. Other tools or standards for loading healthcare data into Hadoop are: Health Level 7 (HL7) International Standards Apache UIMA, JAVA ETL rules

2) Processing Healthcare Data

Depending on the use case, healthcare organizations process data in batch (using Apache Hadoop Map Reduce and Apache Pig); interactively (with Apache Hive); online (with Apache HBase) or streaming (with Apache Storm).

3) Analyzing Healthcare Data

Once data is stored and processed in Hadoop it can either be analyzed in the cluster or exported to relational data stores for analysis there. These data stores might include:

1. Enterprise data warehouse

2. Quality data mart
3. Surgical data mart
4. Clinical info data mart
5. Diagnosis data mart
6. Neo4j graph database

4) Data analysis and visualization

Many data analysis and visualization applications can also work with the data directly in Hadoop. Hortonworks healthcare customers typically use the following business intelligence and visualization tools to inform their decisions: Microsoft Excel, Tableau, RESTful Web Services, EMR Real-time analytics, Metric Insights, Patient Scorecards, Research Portals, Operational Dashboard, and Quality Dashboards.

The next process is to transfer the data to the YARN operating system,YARN is a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users' applications. The last step include the applications business analytics, custom applications, packaged applications.

III. REPRESENTATION OF DATA

A medical data is collection of patient information that contains disease diagnoses, patient-doctor correspondences, laboratory test results, and drug charts. The information can be recorded as a relational database [2]. The most important table, contains the following column attributes: <patient ID>, <code>, <date>, <value1>, <value2>, <text>. <Patient ID> refers to an internal reference of a patient ID, that is, a foreign key in database terminology. A separate Patient table in which <patient ID> is a primary key contains relevant information about the patient such as year of birth, gender, economic deprivation index, etc.

From the machine learning, a medical data can be described along three dimensions, namely the patient dimension, the time dimension, and the concept dimension. The 3D data structure of cells hold the contents of <value1>, <value2>, <text>). We describe how these three dimensions of data can be dealt with below.

1) Patient dimension: Machine learning methods that operate along the patient (or patient record) dimension include but are not limited to the following: mixture of experts, multi-level models, multi-task learning and domain adaptation.

2) Concept dimension: Machine learning methods operating along the concept dimension are closely related to formal naming and definition of the types, properties, and interrelationships of the entities. If patient records are documents, clinical codes are words, then we can use popular text-retrieval models and popular information retrieval techniques which can analyze the relationship between documents (patient records) and the terms (clinical codes).

3) Time dimension: This dimension is related to time. There are a number of algorithms commonly used for temporal analysis. Examples include autocorrelation, cross-correlation, transfer entropy, randomization testing which are used to solve regularized time series problems [3].

These methods work well when the observations are sampled at equal time intervals, such as speech, music, and EEG signals. However, Medical data are often not recorded at regular intervals. For example, blood sugar level samples are only collected as a patient visits his/her clinic, as and when necessary, or else during regular appointments. Irregularities, gaps or missing samples are inevitable because a patient can be absent for the appointment or a clinician may cancel or reschedule the appointment. From the previous studies, an irregular time series can be divided into two types: (1) time series with missing values at random intervals and (2) time-series sampled at non-uniform time intervals. The missing value problem can be regularized using (a) interpolation techniques and (b) regression analysis. The regularization of non uniform time intervals can be addressed using spectral analysis. The idea of spectral analysis is to regularize the time series by generalizing it with Fourier transforms or wavelet transforms.

4) Problems are the in each dimension: Although algorithms that deal specifically with each of the three dimensions already exist and are even well established, there are few algorithms that can operate in all three dimensions simultaneously. For instance, existing solutions in information retrieval [4] may scale well with the number of medical data and concepts but do not consider the evolution of concepts over time.

Dynamic models such as Hidden Markov Models may not scale well with a large number of concepts.
Because of the irregularities in the data sets some of the Data mining algorithms are not suitable for healthcare data Therefore, data representation is very problematic and much research work is required in this field of data modeling specifically healthcare data as it is increasingly important topic in near future.

IV. PRIVACY OF INFORMATION

Cybercriminals are indeed targeting healthcare organizations for their valuable data: cyber attacks and physical criminal activity now have officially surpassed insider negligence as the main cause of a data breach in healthcare organizations. The Ponemon Institute's new Fifth Annual Benchmark Study on Privacy and Security of Healthcare Data, published today, found that close to 45% of all data breaches in healthcare are due to criminal activity such as cybercriminal and nation-state hacks, malicious insiders, and physical theft, a 125% increase in such activity over the past five years. That's a first, since employee or insider negligence, user errors, lost laptops and thumb drives, etc. accounted for the majority of breaches last year and in years past. About 45% of those breaches came via criminal attacks; 43% by lost or stolen computing devices; 40% via employee mistakes; and 12% via a malicious insider. Some $6 billion per year, with an average cost of $2.1 million per healthcare organization, according to the report, which was commissioned by ID Experts."For the first time, criminal attacks constitute the number one root cause [of data breaches], versus user negligence/incompetence or system glitches," says Larry Ponemon, chairman and founder of Ponemon Institute.
Healthcare organizations also are regularly battling security incidents [5][6], such as malware infections. Some 65% say they were hit with cyber attacks in the past two years, and half suffered incidents involving paper-based security incidents. They're not confident in their incident response capabilities, either, with more than half saying their IR isn't adequately funded or manned. And one-third don't have an IR plan at all.
Lost and stolen devices were a problem at 96% of healthcare organizations in the study, as was spear phishing (88%).
Attackers are after insurance information for insurance fraud, as well as employee data from the healthcare providers.
Methods such as K-Anonymization model [7] can be used, which is a sample model based on a set on Techniques such as Generalization, Suppression and others. It convert private data to public data including the data benefits can used it at different processing and also preserving privacy of personal data.
Some of the Preventive measures for the privacy protection:
1) The employees in the private hospitals should be careful about the private data of the patients
2) Proper security measures should be provided for storing the documents of the patients
3) Use of antivirus and license software for prevention against the hackers
4) The systems and devices which stores the sensitive information should be kept secure
5) Insurance companies should also verify the patient identity to avoid fraud.

V. DATA SECURITY

Security requirements a typical healthcare platform should have four principles of data security [8], namely,
1) Diversity: This means that a dataset is unique for a single data extraction and modeling task. If the data of a patient are represented in two data sets, then the two records should be different, e.g., the pseudonimized identity reference or the data should be different.
2) Revocability: This means that if a data set becomes compromised (stolen), a new copy of the data set can be reissued or regenerated.
3) Security: This means that if a data set is stolen, it is computationally difficult to derive the original data set.
4) Utility: This means that if patient data are processed or transformed, they should not reduce their usefulness for analytics.

VI. CONCLUSION

The field of healthcare is very vast and also it has immense impact on the society. The paper discuss the

Apache Hadoop platform which is used worldwide for the processing of bigdata. The above work proposes a three dimensional data representation which consist of patient, concept and time dimensions.

The security and the privacy of the patient data is very important as it may create problems for the individual like financial or physical loses. Privacy preservation method like K-Anonmization is proposed which can be used to get data benefits from the private data and also preserving privacy of patient information. In this paper we tried to identify some of the security problems with the healthcare data present at different levels of processing and also suggested some measures to solve them. We thus conclude that health care data is crucial to the individuals and proper measure suggested above should be taken according to the type of information.

REFERENCES

1) Marco Viceconti, Peter Hunter, and Rod Hose," Big data, big knowledge: big data for personalized healthcare", IEEE Journal of Biomedical and Health Informatics,pp.1,September 2014.
2) Norman Poh, Santosh Tirunagari and David Windridge," Challenges in Designing an Online Healthcare Platform for Personalised Patient Analytics",pp 1-6, Computational Intelligence in BigData, IEEE 2014
3) . M. T. Bahadori and Y. Liu. Granger ",Causality analysis in irregular time series", In SDM, pages 660–671, 2012.
4) Priyanka K, Prof Nagarathna Kulennavar," A Survey On Big Data Analytics In Health Care", International Journal of Computer Science and Information Technologies, Vol. 5 (4) , pp 5865-5868,2014
5) Xindong Wu, Fellow, Xingquan Zhu, Gong-Qing Wu, and Wei Ding," Data Mining with Big Data",IEEE transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp 97-107,january 2014 [6] Lei xu, Chunxiao Jiang, Jian Wang, Jian Yuan, and Yong Ren, "Information Security in Big Data:Privacy and Data Mining", IEEE Access vol 2, pp 1149-1176, Oct 2014.
6) Asmaa Hatem Rashid and Norizan Binti Mohd Yasin,"Generalization Technique for Privacy Preservation of Medical Information", IACSIT International Journal of Engineering and Technology, Vol. 6, No. 4, pp 262-264,August 2014
7) Venkata Narasimha Inukollu , Sailaja Arsi and Srinivasa Rao Ravuri," Security Issues Associated with Bigdata in Cloud Computing", International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, pp 45-56, May 2014.