CLINICAL DATA ANALYTICS IN BIG DATA USING HADOOP

Kavitha. G, IV B.Tech IT, Dr. D.Prabha, Associate Professor, CSE Department,
Sri Krishna College of Engineering and Technology, Coimbatore.

Abstract — The increasing digitization of healthcare information is opening new possibilities for providers and payers to enhance the quality of care, improve healthcare outcomes, and reduce costs. The technology advancement has accelerated the move from paper to digital health records. With information in digital form, healthcare organizations can use available tools and technologies to analyze that information and generate valuable insights. The main idea of this project is to integrate the massive clinical data, perform a wide range of medical and healthcare functions to infer knowledge and to address numerous analytical questions using Hadoop, an open source tool that enables distributed parallel processing of large datasets across various nodes.

Index Terms — Healthcare, Big data, Hadoop, Data analytics, Clinical data, MapReduce, Big data analytics

1 INTRODUCTION

The latest technology advancements have been accelerated the move from paper to digital health records. With information in digital form, healthcare organizations can use available tools and technologies to analyze that information and generate valuable insights. Processing the huge volume of data produced by the healthcare organizations requires a fast, flexible and scalable platform to support ongoing analytics and to fulfill the complex analysis needs. The main idea is to integrate the massive clinical data, perform a wide range of medical and healthcare functions to infer knowledge and to address numerous analytical questions using Hadoop, an open source tool that enables distributed parallel processing of large datasets across various nodes. Although the traditional database management systems can handle data effectively, when it comes to analytical processing of high volume, velocity and veracity of unstructured data, it becomes very challenging to perform complex reporting within a reasonable amount of time as the size of the data grows exponentially as well as C growing demands of customers. Hence the usage of Hadoop framework will be the scalable solution to analyze healthcare datasets. In healthcare, data mining is becoming increasingly popular. Data mining applications can greatly benefit all parties involved in the healthcare industry.

Healthcare organizations are leveraging big data technology to capture all of the information about a patient to get a more complete view for insight into care coordination and outcomes-based reimbursement models, population health management, and patient engagement and outreach. Big data is already changing the way business decisions are made and it's still early in the game. However, because big data exceeds the capacity and capabilities of conventional storage, reporting and analytics systems, it demands new problem-solving approaches. Big data solutions attempt to cost-effectively solve the challenges of large and fast-growing data volumes and realize its potential analytical value.

Big Data is the hottest trend in the business and IT world right now. We are living in the age of big data where due to the rapid development in the computational power and the WWW, we are producing an overwhelming amount of data, which has led to the need of a change in the existing architectures and mechanisms of the data processing systems. Big data- as these large chunks of data is generally called has redefined the current data processing scenario. From consumers to companies, people have an unquenchable appetite for data and all that can be done with it. Not only are we relying on data for movie suggestions and gift recommendations but are depending on data for multidisciplinary climate and energy research, building adaptable roads and buildings, better foresighted healthcare, new ways to identify fraud, and keeping a check on consumer behavior and sentiment. It's a data feast and is not going to end any time soon.

Map Reduce & Hadoop are the most widely used models used today for Big Data processing. Hadoop is an opensource large-scale data processing framework that supports distributed processing of large chunks of data using simple programming models. The Apache Hadoop project consists of the HDFS and Hadoop Map Reduce in addition to other modules.

ROLE OF BIG DATA AND BIG DATA ANALYTICS IN HEALTHCARE

Health care monitoring systems are generating loosely structured data from different sensors that are connected to the patient over a period of time. And these are large complex system requiring efficient algorithms to process these raw data's and require huge computational power. Big data refers to the data generated from different sensors this includes medical, traffic and social data. Some of the characteristics of big data are volume, velocity and value.

International Journal of Scientific & Engineering Research Volume 6, Issue 5, May-2015
ISSN 2229-5518

2 LITERATURE REVIEW

BDA processes unstructured data to find patterns, whereas DW systems process structured data. Outcomes-based research is used to determine which treatments will work best for specific patients. Pre-diagnosis that automatically mines medical literature to create a medical expertise database capable of suggesting treatment options to clinicians based on patients' health records. Clinical decision support (CDS) solutions - Enables self learning, question and answer sessions. Helps to understands categorizes the data and predict outcomes. It also recommend alternative treatments to clinicians and patients [1].

Improving Quality with External Data - Integrating data from outside of the healthcare system has additional challenges, such as privacy, security and legal concerns, as well as questions about authenticity, accuracy and consistency. Implication of Regionalization and Globalization - Aggregating data regionally and globally also provides healthcare researchers with larger populations for clinical studies, trending and disease monitoring for epidemics, as well as early detection and the potential for improved results. There are many potential use cases for BDA in health care. BDA can be used to: help researchers find causes of, and treatments for diseases; actively monitor patients so clinicians are alerted to the potential for an adverse event before it occurs; and personalize care so precious resources associated with a treatment are not administered to a patient who cannot benefit from the intervention[2].

Security and privacy concerns Health Insurance Portability and Accountability Act (HIPAA) Providers, patients and other interested parties such as researchers need secure access, data access should be controlled by group, role and function. Big data plays an important role in medical and clinical research and has been leveraged in clinically relevant studies. Major research institute centers and funding agencies have made large investments in the arena. For example, the National Institutes of Health recently committed US $100 million for the big data to Knowledge (BD2K) initiative. The BD2K defines biomedical big data as large datasets generated by research groups or individual investigators and as large datasets generated by aggregation of smaller datasets [3].

This paper addresses how to analyze massive data in a reliable manner. To take in real world medical data from all levels of human existence to help advance our understanding of medicine and medical practice. Big Data tools and approaches for the analysis of Health Informatics data gathered at multiple levels, including the molecular, tissue, patient, and population levels. The huge body of medical research that has been performed using large datasets demonstrates the broad spectrum of data resources used and shows that the structure of the medical dataset depends on the research question. Data from different subareas of medical research have broad diversity in terms of numbers of entries, types of data stored (or levels), dimensionality, and sample size [4].

Identifying causality of patient symptoms, in predicting hazards of disease and in improving primary-care quality. Describe employed computational algorithms, statistical methods, and software toolkits for data manipulation and analysis. Discussed issues related to storage and analysis of type of data. Technologies for Data Storage and Handling, due to the massiveness and complexity of big data, non-relational and distributed databases such as Apache Hadoop, Google BigTable, NoSQL, and massively parallel-processing databases are used rather than traditional relational databases to store data. A large number of biostatistics software packages have been used to handle large clinical datasets, some of which enabled the features of cloud-based or distributed computing. Popular software packages include, but are not limited to, SAS, Mplus, SPSS, PP-VLAM, Stata, and R. These technologies and tools greatly facilitate the handling of big data.

Methodologies for Data Preprocessing, clinical raw big data can be highly diverse and uninformative without preprocessing. Extraction of a diagnosis from raw computer tomography data is an example of one of the predominant manners in which clinical big data are preprocessed. This type of processes relies on a specialist's personal expertise and can be a source of bias. Most early analyses of big data, including that collected by the Framingham Heart Study adopted some form of preprocessing; therefore, challenges exist in curation. As an alternative to expert preprocessing, computational algorithms or statistical approaches, including compression methods, significance testing, or normalization can be implemented to preprocess raw big data. This methodology may also introduce bias and can cause uncertainty problems during data integration[5].

Purpose – Healthcare is among the fastest-growing sectors in both developed and emerging economies. E-healthcare is contributing to the explosive growth within this industry by utilizing the internet and all its capabilities to support its stakeholders with information searches and communication processes. The purpose of this paper is to present the state-of-the-art and to identify key themes in research one-healthcare. Design / methodology / approach – A review of the literature in the marketing and management of e-healthcare was conducted to determine the major themes pertinent to e-healthcare research as well as the commonalities and differences within these themes [6].

E-healthcare is the use of web-based systems to share and deliver information across the internet. With this ability, privacy and security must be maintained according to the Health Insurance Portability and Accountability Act (HIPAA) standards [7].

International Journal of Scientific & Engineering Research Volume 6, Issue 5, May-2015
ISSN 2229-5518

Large-scale distributed systems, such as e-healthcare systems, are difficult to develop due to their complex and decentralized nature. The service oriented architecture facilitates the development of such systems by supporting modular design, application integration and interoperation, and software reuse [8].

Advantages of using MapReduce in clinical data analytics: MapReduce handles failures by re-executing the failed job on some other machine in a network. The master process, JobTracker, periodically pings the worker nodes, TaskTrackers. JobTracker and TaskTracker are the main processes in Hadoop, but the original MapReduce has similar processes. If the master receives no response from a worker, that worker is marked as failed, and its job is assigned to another node. Even completed Map tasks have to be re-executed on failure, since the intermediate results of the Map phase are on the local disk of the failed machine, and are therefore inaccessible. Completed Reduce tasks, on the other hand, do not need to be re-executed, as their output is stored in a global, distributed file system [9].

The disadvantages of Existing are Traditional DBMS systems cannot handle tera and peta bytes of data, they are not capable of handling unstructured data, they uses static schema for data storage, takes more time to process massive amount of data, relatively not as sound as the proposed system

3 PROPOSED SYSTEM:

The proposed system integrates all the internal and external and multi format data, the data that is being collected from various applications. Those raw data can then be transformed into useful facts by using a data processing tool (for example weka). During data processing the raw data is fed into ETL cycle. Then the transformed data can be analyzed using big data analytics framework and tools. In this project we focus on Hadoop framework, an open source distributed system which processes tera and peta bytes of data efficiently. Once the hadoop environment is set, the transformed data will be stored on the storage space. Now all kind of analytical queries can be addressed effectively. We can infer knowledge from complex heterogeneous medical sources.

Advantages of Proposed System is the proposed system processes real time medical data and analyzes them in a robust manner. Based on the outcome of the analysis, new insights can be derived. Again the same can be fed into the analytics system to enhance the strength of the solution provided. Scalable, Hadoop is a highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel. Cost effective, Hadoop also offers a cost effective storage solution for businesses' exploding data sets. Flexible, Hadoop enables businesses to easily access new data sources and tap into different types of data (both structured and unstructured) to generate value from that data. Fast, Hadoop's unique storage method is based on a distributed file system that basically 'maps' data wherever it is located on a cluster. Resilient to failure, A key advantage of using Hadoop is its fault tolerance. When data is sent to an individual node, that data is also replicated to other nodes in the cluster, which means that in the event of failure, there is another copy available for use. Fault tolerant - storage for large quantities of data. Flexible data model - Near real-time lookups. Atomic and strongly consistent row-level operations. Automatic sharding and load balancing of tables. Metrics exports via File and Ganglia plugins. High availability through automatic failover. In-memory caching via block cache and bloom filters. Server side processing via filters and co-processors. Replication across the data center.

The identified methodologies are Medical Data Collection, Data Transformation, Setting up the working platform environment, Conceptual Model development Analytic techniques Implementation and producing Results and Insights

4 CONCLUSION:

The novelty and significance of this work is that by introducing the idea of efficient information utilization to develop a distributed clinical data analytics system, this work advances the state of art in the healthcare arena. Real-time big data analytics is a key requirement in healthcare. Efficiently utilizing the colossal healthcare data repositories yield some immediate returns in terms of patient outcomes and lowering healthcare care costs, infers knowledge from complex heterogeneous health records enhances the personalized care given to the patients and addresses numerous analytical questions. Data with more complexities keep evolving in healthcare thus leading to more opportunities for big data analytics.

REFERENCES

[1] Emerging Technology Series Big Data Analytics in Health (White Paper) Canada Health Info December, 2013
[2] Bill Hamilton: Big data is the future of healthcare, Cognizant Business Consulting, March 2014, vol 30.
[3] Clinical Analytics System Vol 2 Findings Center for US Health system reform business Technology Office April, 2014
[4] Matthew Herland, Taghi M Khoshgoftaar and Randall Wald : A review of data mining using big data in health informatics, April , 2014 (SIGMOD Record vol 12 No 3)
[5] Science Weiqi Wang, PhD; Eswar Krishnan : Big Data and Clinicians: A Review on the State of the, MD, MPH School of Medicine, Stanford University, United States May,2014
[6] Avinandan Mukherjee, John McGinnis : E-healthcare: an analysis of key themes in research–International Journal of Pharmaceutical and Healthcare Marketing 2007
[7] Avinandan Mukherjee, John McGinnis : E-healthcare: an analysis of key themes in research–International Journal of Pharmaceutical and Healthcare Marketing 2007
[8] F. Gengxin Miao Moser, L.E. Melliar-Smith : A Distributed e-Healthcare System Based on the Service Oriented Architecture Kart,P.M. Dept of Electr. & Comput. Eng., Univ. of California, Santa Barbara, CA;

International Journal of Scientific & Engineering Research Volume 6, Issue 5, May-2015
ISSN 2229-5518

[9] Tomi Aarnio: Parallel big data processing with MapReduce, Helsinki University of Technology, tomi.aarnio@hut.fi