International Journal of Research in Information Technology (IJRIT)

www.ijrit.com ISSN 2001-5569

BIG DATA IN HEALTHCARE

Pranav Patil [1], Rohit Raul [2], Radhika Shroff [3] Mahesh Maurya[4]

Abstract

The big-data revolution is in its early days, and most of the potential for value creation is still not explored. But it has set the industry on a path of rapid change which lead to extensive research in this field and contribute to betterment of human life. As a result of rapid progress in digitizing medical records in recent times, healthcare organization and pharmaceutical companies have started collecting and storing more and more healthcare data in order to analyze it and obtain insights on how to solve problems related to variability in healthcare quality ,skyrocketing healthcare cost, monitor , safety of healthcare systems etc .This paper provides information about all the significant developments that have carried out so far in the field of Big data analysis in healthcare sector .It first explains how we can incorporate and effectively use big data analytics in the field of healthcare by listed all the significant sources of data that can be used for providing effective solution for the problems faced in the healthcare industries and then it describes the new value pathways that could be adopted to increase overall profit pools , reduce overall healthcare cost in the near future and provide new ways which could be adopted to improve the healthcare quality .Our paper also covers key big data implementation challenges and Big Data solutions which attempt to cost-effectively solve the challenges of large and fast-growing data volumes and realize its potential analytical value. We also have discussed different efforts which have been taken by various companies to effectively provide solutions using big data analytics to different areas in healthcare sector.

1. Introduction

Due to the fast expansion and growth of the IT industry, new technologies have been developed in healthcare domain as a result of it health care stakeholders now have access to new sources of knowledge which can create a revolution in the healthcare sector . This information is a form of “big data,” so called not only for its sheer volume but for its complexity, diversity, and timeliness. Now days Advanced medical information management system are been used to store medical data as a result the healthcare providers are working efficiently by to storing and maintaining the patient’s records with use of various techniques like EMR (Electronic Medical Record) and PHR (Personal Healthcare Record) record systems. By access this data researchers can mine the data to see what treatments are most effective for particular conditions, identify patterns related to drug side effects as well as find drugs which can cure certain diseases based on the healthcare data and gain other important information that can help patients as well as reduce the overall healthcare costs. As a result of all the recent technologic advances in the industry have improved their ability to work with such data, even though the files are enormous and often have different database structures and technical characteristics. There is geometric growth increase in the data i.e structural and non-structural data with the expansion of users. Therefore, it has become an important issue of storage, distribution and management of data. In healthcare industry, as huge amount of data is involved , data security and interoperability becomes a huge challenge

2. Different Dimensions of Big Data in Healthcare

There are four main “dimensions” to Big Data, commonly referred to as the Four V’s

1) **Volume**: The big data that is stored in these healthcare systems in terabytes and petabytes(1,000 terabytes).It is a big challenge to store such a complex data. It requires scalable storage and also support for these distributed queries across the data sources. Health care systems should be able to analyze , identify the data and locate it in this huge data structure

2) **Variety** : The data that is stored in the health care systems is of various types like the data stored regarding the patient's health record, data regarding various diseases, data about the medicines used for particular diseases. The data that is stored is both structured and un-structured. Various technologies exists to deal with this highly variable data.

3) **Velocity** : Data that is stored in the healthcare systems is updated on daily, weekly and even monthly basis so it is vital that the data that is stored is correct and without any errors. Therefore, the big data needs to be processed and analysed. It is also important to leverage data to control and reduce the healthcare costs. The data needs to be analyzed in such a way that the root cause of the disease is also analyzed –in other words predictive analysis needs to be done.

4) **Veracity**: Data of varying quality, relevance and meaning .In order to provide efficient and accurate solution the data on which analysis is to be done must be accurate in all sense .Improving coordination of care, avoiding errors and reducing costs depend on high-quality data if the data is of good quality then the results will be better and can be used to draw some effective conclusion to present healthcare problems like eg Which is the most effective ways of treating cancer , preventive measures which can be undertaking to avoid heart strokes

3. Effective new pathways that can be adapted to Improve Health care Industry

There are five effective new pathways that can be adopted to improve the health care industry:

1. Right Living- Living in the right way would enable the patients to live a healthy lifestyle. It would also enable the patients to play a vital role in improving their own health by preventing themselves from being affected by any disease. A patient can lead a healthy lifestyle by eating proper diet food, by exercising regularly and drinking plenty of water. Thus making informed lifestyle choices that lead to overall well being by taking own care

2. Right Care- Patients should get right treatment at the right time. This can achieved if all the healthcare providers work towards this in a the co-ordinated manner. so that all poses accurate information and work towards the goal of provided best care to the patients .

3. Right Provider- The treatment that is provided to the patients has to be appropriate and must be provided by right people. The doctor giving treatment to the patients need to have complete in depth knowledge about problem as well the patient health history so that he can take appropriate decision about which treatment would be right for him/her.

4. Right Value- It is very important to keep on enhancing the healthcare value by constantly improving its quality to provide best care to patients by using the most latest technology . While improving the quality of the healthcare system we need to ensure that the patient gets most cost effective outcome from the healthcare system .We must also take preventive measure to eliminate fraud ,abuse of the patients and wastage of medical resources and in the system.

5. Right Innovation . We need to constantly find new and effective ways of providing right treatment to the patients by giving impetus to find most advance medicines by promoting R&D in the healthcare system . This goal can also be achieved by developing most innovative machine's which can help in early diagnosis of diseases

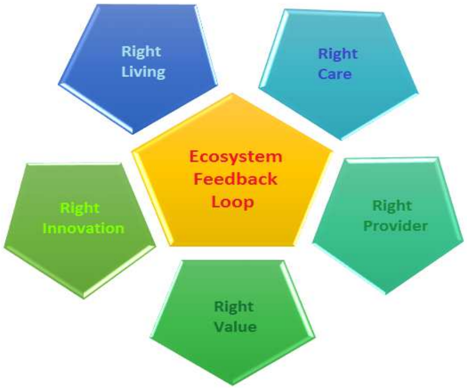

Fig-1: Five pathways

4. Primary sources of Healthcare data which can be used for analysis purpose

Healthcare data has shown phenomenal growth in the past years .Clinical data has grown exponentially as a result of the IT revolution. For example in 2005 only 30 percent of the physicians and doctors used electronic medical records but by the end of 2011 more than 50 percent of physicians and nearly 75 percent hospitals adopted the new technology. Since majority of the hospitals have till now adopted the new technology as a result of which they are now either participating in local or regional health-information exchanges (HIEs) . These developments allow stakeholders access to a broader range of information. For instance, customers who use tools developed by Epic, can access the benchmark and reference information from the clinical records of all other Epic customers. As another example, the HIE in the state of Indiana now connects over 80 hospitals and has information on more than ten million patients. Over 18,000 physicians can take advantage of the data.

In addition to clinical data, several other sources are contributing to the big-data revolution, including:

- Medical insurance claims data which provides us information about the services were provided and how they were reimbursed.
- The pharmaceutical R&D departments describe about the side effects and other harmful actions caused by specific drug treatments.

Various mechanisms are be used to determine the patients information such patient's medical history, patient's finances. Mechanisms such as Acxiom and Accurint are used widely in the today's market.

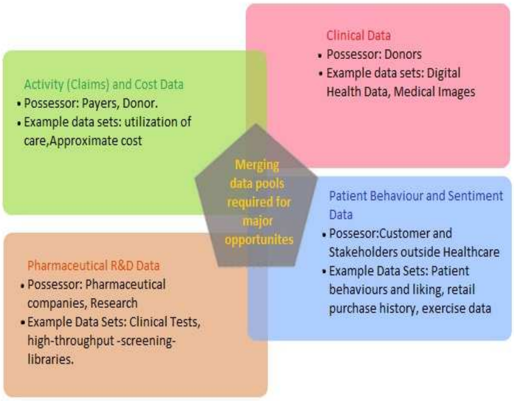
Clinical Data
• Possessor: Donors
• Example data sets: Digital Health Data, Medical Images
Activity (Claims) and Cost Data
• Possessor: Payers, Donor.
• Example data sets: utilization of care,Approximate cost
Merging data pools required for major opportunites
Patient Behaviour and Sentiment Data
• Possesor:Customer and Stakeholders outside Healthcare
• Example Data Sets: Patient behaviours and liking, retail purchase history, exercise data
Pharmaceutical R&D Data
• Possessor: Pharmaceutical companies, Research
• Example Data Sets: Clinical Tests, high-throughput -screening-libraries.

Fig-2: Major Sources of Healthcare data

5. Present Implementations of Big Data in Healthcare industry

5.1 CMC-HIBIS & CMC I+PLUS for health insurance claim analytics: CMC Insurance Solutions (Pty) Limited, Australia has developed two distinct application in the area of health insurance claim analytics that leverages big data to detect and flag fraud, abuse, waste and errors (FAWE) in the health insurance claims sector, thus providing an important step in reducing recurrent losses and facilitating enhanced patient care.

I. The first application is CMC-HIBIS which applies a combination of leading-edge business rules and business intelligence technologies to hospital, medical and ancillary claims data to identify FAWE, and generate alerts along with relevant explanations that can be understood and actioned by claims processing and risk compliance staff. The system also generates a variety of reports related to different types of metrics on alerts and potential costs of leakage in specific areas that are of concern to a claims processing and loss reduction departments

II. The second application is CMC I+PLUS which provides advanced performance analytics using claims-scoring and predictive modelling techniques applied to hospital and medical claims data. This application is targeted at decision makers, and specialists who have the knowledge and ability to perceive and derive value using the sophisticated statistics-based visualisations tools. In addition to automatically detecting FWAEs, the system provides a number of indicators to measure and compare provider performance both from quantitative aspects such as costs, as well as qualitative aspects such as infections and unplanned readmission rates that indicate quality of care.

5.2 Using Map Reduce for Large–scale Medical Image Analysis : In this work, Map Reduce was used to speed up and make possible three large–scale medical image processing use–cases:

I. **Parameter optimization for lung texture classification using support vector machines (SVM)** : It was carried out on the Hadoop cluster. A map task was defined for each coupled value of (C, σ). A clear link between the runtime of a map task and the resulting classification accuracy was observed and most of the tasks with long runtimes resulted in poor classification accuracies. The interruption of such map tasks allowed a reduction of the total runtime from 50h to 9h15m, while keeping all coupled values (C, σ) leading to best classification performance.

II. **Content–based medical image indexing**: Two approaches for content–based image indexing were compared and implemented in the MapReduce framework: component–based versus monolithic indexing. The former is convenient to separately r3optimize feature extraction and the indexer because it does not require to run the whole pipeline for each optimization. However, it requires to write the features to a very large CSV (Comma–Separated Values) file of approximately 100 Gb for 100,000 images. This resulted in an unexpectedly long runtime for the feature extractor with the MapReduce framework in the component–based approach. The result is consistent with previous work that showed that MapReduce was not performing well with input-output (IO)–intensive tasks . The monolithic strategy showed to be well–suited for MapReduce, which allowed indexing 100,000 images in about one hour using 24 concurrent tasks.

III. **Three–dimensional directional wavelet analysis for solid texture classification** : The parallelization of solid texture processing based on non–separable three–dimensional wavelets allowed to reduce a total runtime from more than 130h to less than 6h, while keeping the code in the original Matlab /Octave programming language with Hadoop streaming.

5.3 Intel’s Solution for Healthcare Industry :

The healthcare industry is at an inflection point. To control costs and provide the best care possible, organizations must begin to take advantage of big data to drive better informed, faster decisions, without sacrificing security and privacy. Using Intel Distribution optimized for Intel Xeon processor-based servers, the healthcare industry has access to a powerful platform that can provide both highly scalable and low-cost data storage integrated with affordable, scalable processing. Offering performance, density, reliability, scalability, and operational savings across the entire infrastructure consisting of compute servers, storage, and network, Intel Distribution provides a comprehensive solution architecture. Intel Manager offers robust tools to streamline setup, management, security, and troubleshooting for Hadoop clusters. Healthcare providers and researchers are now able to begin to unlock the power of their data to solve increasingly complex problems. The capability to understand and act upon that data will help lead to a more efficient and effective healthcare industry

5.4 UFIDA Medical Big Data Application Case for a Healthy City Strategy in China :

Intel and UFIDA Medical leveraged an Intel Xeon® E5 Processor platform and Intel Distribution of Hadoop through repeated single business load tests, big data tests, optimizations and other technical means to successfully build a complete regional medical big data computing architecture at the Jinzhou Regional Health Data Centre. The architecture can meet performance requirements of high concurrency retrieval and real-time data analysis of mass data (with more than 100 million records). The architecture based on Intel Xeon® E5 Processor platform and Intel Distribution of Hadoop provides a smart health cloud service platform for data processing, retrieval, analysis and other data services to meet Jinzhou’s healthy city goals.

6. Various Technology options available to implement Big data

In today’s arena, there are many technology solutions available to implement Big Data. Various batch analytics facilities are used for storage and processing of data. Some of the options include Teradata, Vertica (HP) and Netezza (IBM). Maintenance and value of these solutions is relatively low. Ownership cost of these solutions is very high thus making it expensive for purchase.

The barriers in Big Data are reduced or removed using cloud hosted software as a service (SAAS). SaaS-based solutions allow healthcare entities that control subsets of data to expose access through services that eliminate some of the aggregation and integration challenges. Additional services that facilitate analytics, both basic and advanced, can be made part of the overall offering. Google , Amazon implements Big Data by using big computers that can store terabytes of data on it. When there a big problems to be dealt in Map Reduce , the data is divided into segments which is done by using Map Reduce framework. These tasks are distributed over these large computers for execution using Map Reduce algorithms . Other cloud-based solutions include Tableau, which supports visualization.

Hadoop which is open source framework is used by many companies as a robust, scalable, high performance and low cost option when dealing with big data. Vendors such as Greenplum (a division of EMC), Microsoft, IBM and Oracle have commercialized Hadoop and aligned and integrated it with the rest of their database and analytic offerings.

7. Challenges faced while implementation of big data in Healthcare industry

There are three main challenges in big data:

1) **Awareness**: We must fully realize the necessity, complexity and imperativeness of data services.

2) **Talent**: Powerful companies and enterprises are expected to actively promote the application of big data technology in the medical industry and drive big data development and talent cultivation.

3) **Exploration of business application model**: Many enterprises are not familiar with the medical business and thus are unable to discover key business intelligence with big data application in medical practices, constraining big data development in the sector

8. Conclusions

Big Data in Healthcare systems allows to Leverage tremendous amounts of data and processing resources related to healthcare. Healthcare systems need to allocate more time and resources to planning and forecasting.

Our recommendations for healthcare organizations looking to leverage big data include:

- Establish centre for business intelligence that will put the entire focus on big data.

- Decide on an appropriate big data strategy based on the organization’s current and target business and technological enhancements and objectives.

- It is required to assess the various big data initiatives that can be deployed to meet overall corporate objectives, focusing initially on short-time.

- It is also necessary to work with a partner that understands the full range of big data technologies and implementations, including trends, security, internal and external system integration, hosting and development platforms, and application and solution development.

.

9. Acknowledgement

I take this opportunity to express my profound gratitude and deep regards to my guide **Professor Mahesh Maurya** for his exemplary guidance, monitoring and constant encouragement throughout the course of this project . The blessing, help and guidance given by him time to time shall carry me a long way in the journey of life on which I am about to embark.

10. References

1) Research and Implementation of Massive Health Care Data Management and Analysis Based on Hadoop.

2) Leveraging Big Data Analytics to reduce Healthcare Costs- Uma Srinivasan Bavani Arunasalam.

3) The 'Big Data' revolution in healthcare –MCKinsey & Company – White paper.

4) Using Map Reduce for large scale Medical Image analysis – DOI –IEEE Paper.

11. BIOGRAPHIES

Name : Pranav Patil

Student of BTech 4th year Computer Engineering

College : Mukesh Patel School of Technology Management & Engineering

Name : Rohit Raul

Student of BTech 4th year Computer Engineering

College : Mukesh Patel School of Technology Management & Engineering

Name : Radhika Shroff

Student of BTech 4th year Computer Engineering

College : Mukesh Patel School of Technology Management & Engineering

Name : Mahesh Maurya

Assistant Professor Computer Engineering at Mukesh Patel School of Technology Management & Engineering