Improved Healthcare Access using Big Data Analytics through Hadoop and Spark

[1]MUNI KUMAR N, [2]MANJULA R

[1]Research Scholar, [2]Associate Professor
[1]SCOPE, VIT Institute of Technology,
[1]Vellore, Tamil Nadu, India.

Abstract : One of the greatest concerns in both developed and developing countries like India, is access to the primary health care information and maintenance of EHR (Electronic Health Records) and EMR (Electronic Medical Records). Though, the people living in cities have better access to high end health services, the millions of people residing in rural parts of the country are facing serious problems in accessing healthcare. Also the healthcare organizations have become data rich but information poor. Now-a-days, large amounts of heterogeneous medical data have become very common in most of the healthcare organizations. Further, it has become mandatory for the healthcare organizations to use Big Data analytics such as non-linear multivariate predictive analytic models to understand the better diagnoses, origin of diseases, health care fraud detection, provide personal health care and deliver high quality care with low cost to the patients by combing the emerging technologies like Hadoop Framework, Map Reduce and Spark for better data analytics. This paper addresses the difficulty in accessing primary health care, EMR & EHR, big data and its use cases in healthcare sector, also discusses how spark overcomes Hadoop for better analytics.

IndexTerms **- Big Data, Primary Health care, EHR, EMR, Hadoop, Spark, Big Data Healthcare use cases.**

I. INTRODUCTION

The Increase in population in developing countries like India, over-burdens the health care structure. Health care in our country is government financed and government run. But, for many people living in the remote parts of the country, accessing primary health care is still a challenge. The exponential growth [1] of data over the last decade has introduced a new domain in the field of information technology and data science called Big Data. The term Big data is often used to describe a massive volume of data (both structured and unstructured) which is too large to store and difficult to process using traditional database management techniques. As the health care industry is flooded with huge volumes of data which needs validation and analysis, Big Data Analytics can be applied. Big data has the potential to perform critical computing and analytical ability towards the processing of the huge volumes of health care data.

Large amounts of heterogeneous medical data have become available in various healthcare organizations (payers, providers, pharmaceuticals). Those data could be an enabling resource for deriving insights for improving care delivery and reducing waste. The enormity and complexity of these datasets present great challenges in analysis and subsequent applications to a practical clinical environment. In this paper, we focus on the characteristics and related analytic challenges on dealing with clinical data from electronic health records.

As per the statistics available in the Global Health Observatory Data Repository [2], a World Health Organization (WHO) repository, the per capita government expenditure on health care in India during 2011 is at an average of $44, compared to that of $4047 in the USA. The outcome in USA is long lives (increase in the life expectancy), full of sophisticated facilities for health care system, efficient clinical staff, round the clock emergency services and world- class doctors. To provide better e-health care services to the massive Indian people with right care for the right disease at the right time, the big data analytics can be applied. The term, Telemedicine uses the electronic communication technology to exchange patient information among doctors / hospitals and provide health care services at remote locations equivalent to the services rendered by the city hospitals. This innovative technology is gaining increasing attention as a way to improve the performance of health care systems by linking various systems via a data and communications platform to reduce redundant medical tests, improve and expedite clinical decision making, and enable access to all levels of healthcare for a wide range of conditions. With telemedicine, hospitals hope to lower the cost of patient care and increase the effectiveness of chronic disease management. It collects all possible patient information to create thorough electronic health records (EHR's) for each patient.

Healthcare has entered in to the new phase called 'post EMR' deployment [3]. Now, the organizations are keen on gaining insights by using analytics from the vast amounts of data being collected from their EMR systems. Also the participants / stakeholders are much keen about reducing their cost and improving the quality of care by applying advanced analytics.

The rest of the paper is organized as follows: section II discusses what is big data? Section III discusses integrating EMR and EHRs. Section IV Hadoop framework and Apache Spark, Section V deals with Hadoop for Health care, Section VI discusses about various health care specific use cases that can be implemented using big data analytics, Section VII concludes the work.

II. WHAT IS BIG DATA?

The term 'big data' has become the most popular buzzword now-a-days. It seems every industry and organization is focusing on the implementation of big data strategies to analyze the unstructured data. Big Data is a collection of large and complex data sets which are difficult to process using common database management tools or traditional data processing applications. According to zdnet.com, "Big data refers to the tools, processes and procedures allow an organization to create, manipulate and manage very large data sets and storage facilities".

Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it. Big data is arriving from multiple sources at an alarming velocity, volume and variety. To extract meaningful value from big data, you need optimal processing power, analytics capabilities and skills. Big data is changing the way people within organizations work together. It is creating a culture in which business and IT leaders must join forces to realize value from all data. Insights from big data can enable all employees to make better decisions—deepening customer engagement, optimizing operations, preventing threats and fraud, and capitalizing on new sources of revenue. But escalating demand for insights requires a fundamentally new approach to architecture, tools and practices.

The big data are generated from online transactions, emails, videos, audios, images, click streams, logs, posts, search queries, health records, social networking interactions, science data, sensors, mobile phones and their applications

In the year 2012, Gartner has defined the big data using 4 V's as Volume, Velocity, Variety and Veracity. The data which is flowing with huge volume, under high velocity of Veracity (uncertainty) of data is termed as Big Data. In Healthcare, the big data flows from various sources such as [4]

- Human generated data - Electronic mail records, physician notes, paper documents etc.
- Electronic Health Records (EHR) - Health care data of a person in digital form, health claims, hospital revisits, prescriptions, medical expenses, health claims and other records that may be either structured or unstructured.
- Biometrics data - Genetics, fingerprints, handwriting, retinal scan data, x-ray data, medical images, blood pressure and pulse rate data etc.
- Sensor data - Data generated thru various sensors, meters and other health care devices.
- Web and social media data - The social media data of patients, discussions on the twitter and blogs

III. INTEGRATING EMR'S AND EHR'S

The Electronic Health Record (EHR) [5] is a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting. This record includes the patient demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports. The EHR automates and streamlines the clinician's workflow. The EHR has the ability to generate a complete record of a clinical patient encounter - as well as supporting other care-related activities directly or indirectly via interface - including evidence-based decision support, quality management, and outcomes reporting.

An EMR [6] contains the standard medical and clinical data gathered at healthcare provider's office. Unlike EHR, An electronic medical record (EMR) is a digital version of a paper that contains all of a patient's medical history. An EMR is mostly used by providers for better diagnosis and treatment

The maintenance of EMR's has more benefits than paper records because it allows providers to:

- Track data over time
- Identify patients who are due for preventive visits and screenings
- Monitor how patients measure up to certain parameters, such as vaccinations and blood pressure readings
- Improve overall quality of care in a practice

The information stored in EMRs is not easily shared with providers outside of a practice. A patient's record might even have to be printed out and delivered by mail to specialists and other members of the care team.

Now, the study is being carried out to integrate the EMR and EHR so that, the patient data can be transferred thru online and can be accessed by any provider by accessing the records available in cloud storage. If this becomes reality, the patient need not carry any reports or prescriptions when moving from one provider to another. As the patients have EHR, the doctor can access

the relevant data and can reduce the un-necessary delays in the treatment. Further, the better analytics can be performed and gain good insights about the patient's health care status and also the origin for the diseases.

IV. HADOOP FRAMEWORK AND APACHE SPARK

Hadoop is an open source framework which employs a simple programming standard that allows distributed processing of massive data sets on clusters of computers. The entire technology incorporates shared utilities, a distributed file system (DFS), analytics and information storage platforms, plus an application layer which manages the activities like workflow, distributed processing, parallel computation and configuration management.

HDFS

The basic idea of Hadoop is to make use of the Distributed file system for storing and processing the data. This HDFS splits the file into blocks and these blocks are allocated in the Hadoop cluster nodes. The input data in HDFS is given once and it is processed by MapReduce and the outcomes are sent to HDFS. The HDFS data is safeguarded by duplication mechanism among the nodes which gives reliability and availability regardless of node failures.

In Hadoop, there are two types of HDFS nodes:

(1) Data Node (2) Name Node

Data Node stores the data blocks of the files, whereas the Name Node contains the metadata, with record blocks and a list of DataNodes in the cluster.

MapReduce

MapReduce is the programming paradigm that allows for massive scalability across hundreds or thousands of servers in the Hadoop cluster. MapReduce is the heart of Hadoop where the processing is carried out by assigning the tasks to various clusters.

Apache Spark

Apache Spark [10] is an open source big data processing framework built around speed, ease of use, and sophisticated analytics. It was originally developed in 2009 in UC Berkeley's AMPLab, and open sourced in 2010 as an Apache project. Spark is written in Scala Programming Language and runs on Java Virtual Machine (JVM) environment. Spark has several advantages compared to other big data and MapReduce technologies like Hadoop and Storm. First of all, Spark gives us a comprehensive, unified framework to manage big data processing requirements with a variety of data sets that are diverse in nature (text data, graph data etc.). Spark enables applications in Hadoop clusters to run up to 100 times faster in memory and 10 times faster even when running on disk. Spark lets you quickly write applications in Java, Scala, or Python. It comes with a built-in set of over 80 high-level operators. And you can use it interactively to query data within the shell. In addition to Map and Reduce operations, it supports SQL queries, streaming data, machine learning and graph data processing. Developers can use these capabilities stand-alone or combine them to run in a single data pipeline use case.

Hadoop and Spark

Hadoop as a big data processing technology has been around for 10 years and has proven to be the solution of choice for processing large data sets. MapReduce is a great solution for one-pass computations, but not very efficient for use cases that require multi-pass computations and algorithms. Each step in the data processing workflow has one Map phase and one Reduce phase and you'll need to convert any use case into MapReduce pattern to leverage this solution. The Job output data between each step has to be stored in the distributed file system before the next step can begin. Hence, this approach tends to be slow due to replication & disk storage. Also, Hadoop solutions typically include clusters that are hard to set up and manage. It also requires the integration of several tools for different big data use cases (like Mahout for Machine Learning and Storm for streaming data processing). Each of those jobs was high-latency, and none could start until the previous job had finished completely.

Spark allows programmers to develop complex, multi-step data pipelines using directed acyclic graph pattern. It also supports in-memory data sharing across DAGs, so that different jobs can work with the same data.

Spark runs on top of existing Hadoop Distributed File System (HDFS) infrastructure to provide enhanced and additional functionality. It provides support for deploying Spark applications in an existing Hadoop v1 cluster (with SIMR – Spark-Inside-MapReduce) or Hadoop v2 YARN cluster or even Apache Mesos. We should look at Spark as an alternative to Hadoop MapReduce rather than a replacement to Hadoop. It's not intended to replace Hadoop but to provide a comprehensive and unified solution to manage different big data use cases and requirements.

Features of Spark

Spark takes MapReduce to the next level with less expensive shuffles in the data processing. With capabilities like in-memory data storage and near real-time processing, the performance can be several times faster than other big data technologies. Spark also supports lazy evaluation of big data queries, which helps with optimization of the steps in data processing workflows. It provides a higher level API to improve developer productivity and a consistent architect model for big data solutions. Spark holds intermediate results in memory rather than writing them to disk which is very useful especially when you need to work on the same dataset multiple times. It's designed to be an execution engine that works both in-memory and on-disk. Spark operators perform external operations when data does not fit in memory. Spark can be used for processing datasets that larger than the aggregate memory in a cluster.

Spark will attempt to store as much as data in memory and then will spill to disk. It can store part of a data set in memory and the remaining data on the disk. You have to look at your data and use cases to assess the memory requirements. With this in-memory data storage, Spark comes with performance advantage.

Other Spark features include:

- Supports more than just Map and Reduce functions.
- Optimizes arbitrary operator graphs.
- Lazy evaluation of big data queries which helps with the optimization of the overall data processing workflow.
- Provides concise and consistent APIs in Scala, Java and Python.
- Offers interactive shell for Scala and Python. This is not available in Java yet.

V. HADOOP FOR HEALTH CARE

The following are some sources of medical data that Hadoop gathers to make it less expensive and more available, so that patients have more choices, doctors have more insight, pharmacy and health device manufacturers can deliver more effective, reliable products:

- **Access Genomic Data for Medical Trials**
  If we read that a given drug is "40% effective in treating cancer," another interpretation could be that the drug is 100% effective for patients with a certain genetic profile. Matching a particular drug to a specific genomic profile is a big data problem. Each individual's genome is about 1.5 gigabytes of data. Massive data storage and processing power is required to analyze data on a drug's interactions with different genetic combinations. For example, just focusing on 20 genes is a 20,000-choose-20 calculation, with 4.3 x 10^67 possible combinations. Researchers are turning to Apache Hadoop as a cost-effective, reliable platform for storing genomic data and combining that with other data sets (e.g. demographics, trial outcomes) to find out which drugs and treatments work best for groups of patients across the genetic spectrum.
- **Monitor Patient Vitals in Real-Time**
  In a typical hospital setting, nurses do rounds and manually monitor patient vital signs. They may visit each bed every few hours to measure and record vital signs but the patient's condition may decline between the time of scheduled visits. This means that caregivers often respond to problems reactively, in situations where arriving earlier may have made a huge difference in the patient's wellbeing. New wireless sensors can capture and transmit patient vitals at much higher frequencies, and these measurements can stream into a Hadoop cluster. Caregivers can use these signals for real-time alerts to respond more promptly to unexpected changes. Over time, this data can go into algorithms that proactively predict the likelihood of an emergency even before that could be detected with a bedside visit.
- **Reduce Cardiac Re-Admittance Rates**
  Patients with heart disease can be closely monitored while they are in a hospital, but when those patients go home, they may skip their medications or ignore dietary and self-care instructions given by their doctor when they left the hospital. Congestive heart failure causes fluid retention, which leads to weight gain. In one innovative program at UC Irvine Health, patients can return home with a wireless scale and weigh themselves at regular intervals. Algorithms running in Hadoop determine unsafe weight gain thresholds and notify a physician to see the patient proactively, before an emergency re-admittance.
- **Improve Prescription Adherence**
  The Centers for Disease Control and Prevention (CDC) found in 2010 that 48% of Americans take at least one prescription drug. Many people do not take the drugs as prescribed and a separate study by the New England Health Care Institute found that this prescription non-adherence costs the health care system $290 billion annually. Innovative healthcare providers are testing and measuring various communication programs to improve adherence. A successful outcome is a renewal of a prescription in the expected time frame. Hadoop can store renewal information and tie it to social media content and online reminders. Natural language recognition can analyze doctors' hand-written notes. And geo-location data can help direct patients to the nearest pharmacy for a refill.

VI. HEALTH CARE SPECIFIC USE-CASES

Healthcare evolves with the advent of big data technologies and Hadoop framework. Healthcare organizations are taking advantage of the volume, velocity, variety and veracity of data being generated from internal and external sources of the clinics. With the use of big data analytics, the organizations got the opportunity to improve medical outcomes, lower costs and inform strategic planning. The organizations are trying to unlock the key insights from big data in the following health care specific use cases.

(1) Personalized Health Care

Based on the EHR of every individual patient, it is possible to do deep diagnosis about his past health history and treatments already done, diagnosis reports etc. This deeper analysis must be carried by hadoop to predict the better drug and provide high quality treatment during the early stages of the disease so as to avoid the issues that might arise with the advancement of disease. Real time analytics can be done with Map Reduce and Hadoop, based on the analytical results, the patient will be provided personalized care.

(2) Improve at-risk Prioritization [8]

Healthcare organizations can use the analytics to predict which patients are at risk of non-compliance with clinical guidelines for treatable conditions like diabetes which improves medical outcomes and reduces hospitalization. This can be done

in hadoop by building predictive variables from clinical, financial and behavioral data and applying a segmentation and regression models.

(3) Claim Assistance

By studying the various patterns in the medical expenses data, it is possible to estimate the charges and medical expenses which help the healthcare payers to avoid overpayment and detect questionable charges. This helps the patient as well to know the approximate medical expenses during the initial days of treatment.

(4) Staff optimization

Using big data analytic models, it is possible to forecast the patient visits during certain months or seasons. This helps the providers to avoid overstaffing, improve staffing flexibility and lower the overall staffing cost without sacrificing patient care.

(5) Fraud detection

The healthcare providers are facing lot of problems with the frauds and looking for the methods to detect the frauds. Big data analytics can be applied to estimate the charges that are to be paid by the insurance companies thereby reducing the overburden for the insurance companies.

VII. CONCLUSION

Big data & Spark have very good potential to transform health care and significantly improve health outcomes. This paper discussed the importance of EHR, EMR and their integration in providing the better health care at lower costs. The health care data analytics through Map reduce, Hadoop and Apache Spark can help the health care stake holders in assessing the various tasks related to the staffing, budgeting, frauds etc. The health care data analytics through Hadoop dramatically will improve the delivery of primary health care to the people living in rural areas.

VIII. ACKNOWLEDGMENT

I would like to take this opportunity to express my profound gratitude and deep regard to my guide, *Prof. R. Manjula, SCOPE, VIT Institute of Technology*, for her exemplary guidance, valuable feedback and constant encouragement in completing this paper. Her valuable suggestions were of immense help in getting this work done. Her perceptive criticism kept me working to make this paper in a much better way. Working under her, was an extremely knowledgeable experience. Also, I would like to extend my sincere gratitude to my parents, my wife and my lovable daughter for their constant support and encouragement in completing this paper.

REFERENCES

[1] Muni Kumar N and Manjula R, "Survey on Map Reduce Based Apriori Algorithms in Medical Field for the Prediction of Diabetes Mellitus", Research Journal of Fisheries and Hydrobiology, 11(4), 2016. pp. 13-18.

[2] Muni Kumar N and Manjula R, "Role of Big Data Analytics in Rural Health Care - A Step Towards Svasth Bharath", International Journal of Computer Science and Information Technologies, Vol. 5(6), 2014, pp. 7172-7178.

[3] Haritha Chennamsetty et.al. , "Predictive Analytics on Electronic Health Records (EHRs) using Hadoop and Hive", IEEE, 2015.

[4] Padmapriya S, Jaya Kumar P, "Summarization Techniques in Association Rule Data Mining For Risk Assessment of Diabetes Mellitus", International Journal for Trends in Engineering & Technology, Vol.3, Issue.1, pp.52-57, January 2015.

[5] Shruthi M Kulkarni and B Sathish Babu, "Cloud-Based Patient Profile Analytics System for Monitoring Diabetes Mellitus", International Journal of Innovative Technology and Research, pp. 228-231, April 2015.

[6] Thulasi et.al. "Predicting Relative Risk for Diabetes Mellitus using Association Rule Summarization Technique in EMR, "International Journal of Innovative Research in Science, Engineering and Technology, Vol.4, Issue3, pp. 970-975, March 2015.

[7] D. Peter Augustine, "Leveraging Big Data Analytics and Hadoop in Developing India's Healthcare Services", *International Journal of Computer Applications*, vol. 89-No 16, pp. 44-50, Match 2014.

[8] Raghupathi and Raghupathi, "Big data analytics in Healthcare: Promise and Potential", Health Information Science and Systems, Hissjournal, 2014, http://www.hissjournal.com/content/2/1/3

[9] Thirumal and Nagarajan, "Applying Average K Nearest Neighbour Algorithm to Detect Type-2 Diabetes", Australian Journal of Basic and Applied Sciences, 8(7) May 2014, pp.128-134

[10] http://spark.apache.org/