International Journal of Advance Research in Science and Engineering
Vol. No.9, Issue No. 12, November 2020
www.ijarse.com
IJARSE
ISSN 2319 - 8354

BIG DATA ANALYTICS: A BOON TO HEALTHCARE SECTOR. A DETAILED STUDY THROUGH HADOOP AND SPARK

Miss Khushbu A Patel[1], Dr. Mahammad Idrish I. Sandhi[2]

[1]*Assistant Professor, Shree Uttar Gujarat BCA College, Surat, India*
[2]*Associate Professor & HOD ,Sankalchand Patel College of Engineering Department of Computer Application, Sankalchand Patel University, Visnagar, India*

ABSTRACT:*Entry to primary health care information and management of EHR (Electronic Health Records) and EMRR is one of the major challenges in both developed and emerging countries such as India (Electronic Medical Records). While urban residents have greater access to high-end health facilities, millions of people live in rural areas of the country are facing significant healthcare access problems. Healthcare organisations have now become rich in records, but low in information.Huge volumes of heterogeneous patient data have become very common in most healthcare institutions today.In addition, Big Data analytics such as non-linear multivariate predictive computational models have become mandatory for healthcare organisations to understand better diagnosis, disease sources, health care fraud identification, offer personal health care and provide low-cost high-quality care to patients by integrating new technology such as Hadoop Platform, Map Reduce and Map Reduce.This paper addresses the problem of accessing primary health care, EMR & EHR, big data and its use cases in the healthcare industry, as well as how spark overcomes Hadoop for improved analytics.*

Keywords*:Big Data, Big Data Healthcare use cases, EHR, EMR, Hadoop, Primary Health care, Spark.*

I. INTRODUCTION

The population increase in developed nations such as India is overburdening the system of health care. Health insurance in our country is funded by the taxpayer and managed by the government. But accessing primary health care is still a problem for many people living in rural areas of the country. Over the past decade, accelerated data growth [1] has introduced a new domain called Big Data in the area of computer technology and data science.Big data is also used to characterise a vast quantity of information (both organised and unstructured) that is too huge to contain and impossible to handle using conventional methods of database management. Big Data Analytics should be implemented as the health care sector is overwhelmed with large quantities of data that needs evaluation and review. Big technology provides the

potential to conduct basic computing and analytical capability to process the large amounts of data from health care.

In multiple health institutions, vast numbers of heterogeneous patient data have become available (payers, providers, pharmaceuticals). Such data may be an enabling resource for gaining information to enhance patient quality and waste reduction. The enormity and sophistication of these databases provide a realistic healthcare setting with great difficulties in study and subsequent implementations.In this article, we concentrate on the characteristics of working with clinical evidence from electronic health reports and associated methodological problems.

According to the figures available in the World Health Organisation (WHO) archive of the Global Health Observatory Data Repository [2], the per capita government spending on health care in India was on average $44 in 2011, compared to $4047 in the USA. The consequence in the United States is long life (increase in life expectancy), full of modern health care system equipment, efficient nursing personnel, emergency coverage round the clock and world-class physicians.Big data analytics can be implemented to offer improved e-health care services to the large Indian population with accurate care for the right illness at the right time. Telemedicine uses electronic networking systems to communicate patient records between doctors / hospitals to deliver remote health care facilities similar to the services offered by city hospitals.

As a way to increase the efficiency of health care services, this revolutionary technology is gaining growing attention by connecting multiple systems across a data and communications network to eliminate unnecessary diagnostic testing, enhance and expedite clinical decision-making, and enable access to all levels of healthcare across a wide variety of conditions. With telemedicine, hospitals are aiming to reduce health care costs and improve the feasibility of treating chronic diseases.In order to establish accurate electronic health records (EHR's) for each patient, it gathers all available patient information.Healthcare has entered a new process called the implementation of 'post EMR' [3]. Today, by leveraging analytics from the large volumes of data obtained by their EMR programmes, companies are keen on obtaining insights. Participants/stakeholders are also keen to reduce their costs and boost the quality of treatment through the use of applied analytics.

II. WHAT IS BIG DATA?

Now-a-days the phrase 'big data' has become the most popular buzzword. In order to process unstructured data, every industry and organisation appears to be focused on applying big data strategies. Big Data is a series of vast and complicated data sets that using popular database management techniques or conventional data processing software, are difficult to process"Big data refers to the tools, processes and procedures allow an organisation to create, manipulate and manage very large data sets and storage facilities".

Big data is generated at all times from everything around us. It is created by any digital process and social media exchange. It is passed on by structures, sensors and smart devices. Big data comes at an alarming speed, volume and range from various outlets. You need optimal computing resources, analytics capabilities and expertise to derive real value from big data. Big technology is transforming the way entities operate together within organisations.It establishes a community in which business and IT leaders need to join forces to derive value from all knowledge. Big data analytics will help all workers to make smarter choices, such as deepening customer loyalty, improving processes, mitigating breaches and fraud, and capitalising on new revenue streams. Yet a radically new approach to architecture, tools and practises involves an escalating demand for perspectives.

Online transactions, emails, videos, audio, photos, click feeds, logs, messages, search requests, health reports, connections with social networks, scientific data, sensors, cell phones and their apps produce big data

In 2012, using 4 V's as Length, Velocity, Variety and Veracity, Gartner identified big data.Big Data is the data that flows with immense volume, under strong data velocity of veracity (uncertainty). Big data streams from multiple outlets of healthcare, such as [4]

- Electronic Health Information (EHR) - A person's health care documents in digital form, health claims, revisits of hospitals, medications, medical costs, health statements and other records that may be organised or unstructured.
- Electronic mail archives, doctor's notes, paper journals, etc. Human data produced
- Genetics, fingerprints, signatures, retinal scan data, x-ray data, diagnostic images, data on blood pressure and heart rate, etc. Biometrics details
- Online and social media data - Patient social media data, Twitter conversations, and blogs
- Sensor data - Data generated by different sensors, metres and other instruments for health care.

III. INTEGRATING EMR'S AND EHR'S

The Electronic Health Record (EHR) [5] is a longitudinal electronic record of patient health information created in any care delivery system from one or more experiences. Patient demographics, progress notes, complaints, medicines, vital signs, prior medical history, immunizations, laboratory data and information on radiology are contained in this record.The EHR automates and streamlines the process for the clinician. The EHR is able to produce a full record of a clinical patient experience - as well as to actively or implicitly assist other care-related practises through the interface - like evidence-based policy support, quality control and monitoring of results.

The regular diagnostic and therapeutic data obtained at the healthcare provider's office is used in the EMR [6]. An electronic medical record (EMR) is a digital representation of a document that includes all of the medical records of a patient. An EMR is often used for improved diagnosis and care by professionals.

EMR maintenance provides more advantages than paper documents because it helps providers to:

- Over time, track data
- Increase the average level of care in a practise
- Identify patients who are due for scans and preventive appointments
- Track if patients, such as vaccines and blood pressure readings, live up against those criteria

Information contained in EMRs is not readily exchanged with outside-of-practice suppliers. The record of a patient may also have to be written out and sent to physicians and other care team members by mail.

The research is now being undertaken to merge the EMR and EHR so that medical data can be transferred electronically and accessible by any provider by using the cloud storage records available. When this becomes reality, by switching from one service to another the patient need not bring any reports or medications. The doctor can access the correct details when the patients have EHR and can reduce the needless gaps in therapy.

IV. HADOOP FRAMEWORK AND APACHE SPARK

Hadoop is an open source platform that uses a basic programming standard that enables large data sets on computer clusters to be stored distributed. Shared utilities, a distributed file system (DFS), analytics and information retrieval systems are included in the whole infrastructure, plus an application layer that handles operations such as workflow, distributed processing, parallel computing and configuration management.

4.1 HDFS

Hadoop's fundamental concept is to use the Distributed File System for data collection and processing. This HDFS divides the file into blocks, and the Hadoop cluster nodes distribute these blocks. The HDFS input data is generated once and MapReduce processes it and the results are sent to HDFS. HDFS data is secured by a replication mechanism between nodes that provides durability and usability irrespective of node failures.

There are two HDFS node types in Hadoop:

(1) Data Node (2) Name Node

The Data Node stores the data blocks of the files, while the Name Node stores the metadata, the record blocks and the Data Node list of the files in the cluster.

4.2 MapReduce

The programming paradigm that allows for huge scalability in the Hadoop cluster over hundreds or thousands of servers is MapReduce. The heart of Hadoop is MapReduce, where the sorting is carried out by assigning tasks to separate clusters.

4.3 Apache Spark

Apache Spark [10] is an open source platform for the analysis of big data based on speed, ease of use and advanced analytics. It was initially developed at UC Berkeley's AMPLab in 2009 and opened as an Apache

project in 2010. Spark is written and runs on a Java Virtual Machine (JVM) environment in the Scala Programming Language. In contrast to other big data and MapReduce innovations like Hadoop and Wind, Spark has some advantages.

First of all, Spark provides us with a robust, unified system to handle specifications for big data analysis for a number of data sets that are complex in nature (text data, graph data etc.). In Hadoop clusters, Spark helps programmes to run up to 100 times faster in memory and 10 times faster even while operating on a disc. Spark lets you write apps in Java, Scala, or Python easily. It comes with an interconnected collection of more than 80 operators at the highest level.And to query data inside the shell, you can use it interactively. It facilitates SQL queries, streaming data, deep learning and manipulation of graph data in addition to Map and Reduce operations. These features can be used stand-alone by developers or merged to run in a single use case for the data pipeline.

4.4 Hadoop and Spark

Hadoop has been around for 10 years as a big data analysis platform and has proved to be the tool of choice for the processing of large data sets. MapReduce is a fantastic solution for one-pass computations that require multi-pass computations and algorithms, but not very effective for use cases. There is one Map and one Reduce phase for each stage in the data processing workflow and you would need to turn each use case into a MapReduce pattern to utilise this approach.Job output data needs to be saved in the distributed file system between each step before the next step can begin. Hence, due to replication & disc storage, this method appears to be sluggish. Also, generally, Hadoop strategies require clusters that are difficult to set up and maintain.It also includes many tools to be implemented with various cases of big data use (like Mahout for Machine Learning and Storm for streaming data processing). Any of those jobs was high-latency, and once the previous job was fully done, none could begin.

Using a guided acyclic graph pattern, Spark enables programmers to create dynamic, multi-step data pipelines. In-memory data sharing through DAGs is also enabled, so that multiple jobs will operate on the same data.

To have improved and external features, Spark runs on top of the current Hadoop Distributed File System (HDFS) architecture. Help for deploying Spark applications in an existing Hadoop v1 cluster (with SIMR-Spark Inside-MapReduce) or Hadoop v2 YARN or even Apache Mesos clusters is given. Instead of a Hadoop substitute, we can look at Spark as an alternative to Hadoop MapReduce. It is not meant to replace Hadoop, but to provide a robust and unified approach for managing multiple cases and specifications of big data use.

4.5 Features of Spark

With less costly shuffles in the data processing, Spark brings MapReduce to the next stage. The output will be many times quicker than other big data technologies with features such as in-memory data storage and close real-time computing. Spark also supports lazy big data query evaluation, which allows to simplify the

steps in workflows for data processing. It offers a higher-level API to increase the efficiency of developers and a clear architecture model for big data solutions.Instead of writing them to disc, Spark keeps intermediate memory outcomes, which is very handy, particularly when you need to focus on the same dataset many times. It is designed to be an in-memory and on-disk operating execution engine. When data does not fit into the memory, Spark operators execute external operations. You may use Spark to process datasets greater than the combined memory of a cluster.

Spark tries to retain in memory as well as data and then bursts onto the disc. It will store in memory part of a data set and the remaining data on the disc. To analyse the memory criteria, you have to look at the data and usage cases. Spark has a performance edge with this storing of inmemory files.

Additional Spark capabilities include:

- Provides in Scala, Java and Python succinct and reliable APIs.
- Offers for Scala and Python interactive shells. This is not yet usable in Java.
- Lazy assessment of big data queries which helps to improve the overall workflow for data processing.
- Optimizes random graphs for operators.
- Supports more than just elements like Map and Minimize.

V. HADOOP FOR HEALTH CARE

Any medical data points that Hadoop collects to make it less costly and more usable are the following, such that patients have more options, physicians have more knowledge, suppliers of pharmacy and health equipment will produce more effective, safe products:

5.1. Improve commitment to prescriptions

In 2010, the Centers for Disease Control and Prevention (CDC) discovered that 48% of Americans took at least one prescribed medicine. Many patients may not take the medications as recommended and a separate report from the New England Health Care Institute showed that the health care system pays $290 billion annually for this opioid non-adherence. To maximise adherence, creative healthcare providers are evaluating and assessing multiple contact programmes.A good result is a renewal in the anticipated time period of a prescription. Hadoop will store information on renewals and connect it to material from social media and online reminders. Recognition in natural language may analyse hand-written observations from physicians. And data on geo-location can help guide patients to a prescription at the closest pharmacy [7].

5.2. Reducing cardiac re-admission rates

It is possible to closely track patients with heart disease while in a hospital, but when those patients go home, they can miss their medications or ignore their doctor's nutritional and self-care guidelines when they leave the

hospital. Congestive heart disease, which leads to weight gain, causes fluid accumulation. Patients will return home on a wireless scale and measure themselves at daily intervals in one ground breaking service at UC Irvine Fitness.Hadoop algorithms assess unhealthy rates for weight gain and alert a practitioner before an emergency readmission to proactively see the patient.

5.3. Access to Surgical Trial Genomic Details

If we read that a given medication is 40% effective in cancer therapy, another explanation may be that for people with a specific genetic profile, the drug is 100% effective. A big data challenge is matching a given drug to a particular genomic profile. The genome of each person is approximately 1.5 gigabytes of data. To examine data on the interactions of a drug with numerous genetic combinations, large data storage and processing capacity is required.For e.g., only concentrating on 20 genes is a computation of 20,000-choose-20, with potential variations of 4.3 x 10^67. Researchers are looking to Apache Hadoop to figure out which medications and therapies perform well for classes of patients across the genetic continuum as a cost-effective, accurate medium for storing genomic data and integrating them with other data sets (e.g. populations, study results).

5.4. Real-time Hospital Vitals Control

In a normal hospital environment, nurses do rounds and monitor vital patient signs manually. Every few hours, they may visit each bed to assess and report vital signs, although the health of the patient can decline between planned visits. This suggests that caregivers frequently impuls reactively to difficulties, in cases where arriving faster may have made a significant difference in the well-being of the patient.

At far higher frequencies, modern wireless instruments can collect and relay patient vitals, and these readings can feed into a Hadoop cluster. To respond more promptly to sudden shifts, caregivers may use these triggers for real-time warnings. Over time, this data can be entered into algorithms that proactively forecast the possibility of an incident long before a bedside visit can be observed.

VI. HEALTH CARE SPECIFIC USE-CASES

With the introduction of big data technology and the Hadoop paradigm, healthcare is changing. Healthcare organisations take advantage of the number, pace, variety and veracity of data generated by clinics' internal and external sources [9]. The companies were able to enhance patient performance, reduce costs and guide strategic planning through the use of big data analytics.

In the following health care related use cases, the companies attempt to unlock the key lessons from big data.

6.1. Health Treatment Tailored

It is possible to make a detailed diagnosis on the basis of and particular patient's EHR regarding his previous clinical records and procedures already completed, diagnosis notes, etc. Hadoop must carry out this deeper research to predict the right medication to provide high-quality care during the early stages of the disease in order to prevent the complications that can occur with the progression of the disease.Based on the analytical

findings, customised treatment will be given to the patient with real-time analytics with Map Reduce and Hadoop.

6.2. Assistance in Argument

It is possible to quantify the costs and medical expenditures that help healthcare payers prevent overpayment and identify questionable charges by analysing the different trends in the medical cost results. This also allows the patient during the early days of recovery to consider the estimated medication costs.

6.3. Improving prioritisation at risk [8]

Analytics can be used by healthcare organisations to determine which patients are at risk of non-compliance with treatment criteria for treatable diseases such as diabetes, which increases diagnostic performance and lowers hospitalisation. Through constructing predictive variables from health, financial and behavioural data and applying a segmentation and regression model, this can be achieved in Hadoop.

6.4. Detection of Fraud

Healthcare providers face a lot of fraud challenges and are searching for tools for detecting fraud. Big data analytics can be used to predict the premiums to be charged by the insurance providers, thus reducing the insurance firms' overload.

6.5. Optimization of Workers

It is possible to predict patient visits during certain months or seasons using Big Data predictive models. This allows providers to prevent overstaffing, increase the versatility of staffing and reduce the total expense of staffing without reducing patient safety.

VII. CONCLUSION

Big data & Spark have very high prospects for changing patient care and improving health quality dramatically. This paper addressed the value of delivering quality health coverage at reduced rates for EHR, EMR, and their convergence. Via Map Minimize, Hadoop and Apache Spark, health care data analytics will assist health stakeholders in determining the multiple activities relevant to hiring, budgeting, theft, etc.Via Hadoop, health care data analytics would significantly increase the provision of primary health care to people living in rural areas.

VIII. REFERENCES

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.

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

[2] HarithaChennamsetty et.al. , ―Predictive Analytics on Electronic Health Records (EHRs) using Hadoop and Hive‖ , IEEE, 2015.

[3] 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.

[4] Shruthi M Kulkarni and B SathishBabu, ―Cloud-Based Patient Profile Analytics System for Monitoring Diabetes Mellitus‖ , International Journal of Innovative Technology and Research, pp. 228-231, April 2015.

[5] 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.

[6] 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.

[7] 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

[8] 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

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