Chapter 4
Healthcare Big Data:
A Comprehensive Overview

Pijush Kanti Dutta Pramanik
National Institute of Technology Durgapur, India

Saurabh Pal
Bengal Institute of Technology, India

Moutan Mukhopadhyay
Bengal Institute of Technology, India

ABSTRACT

Big data has unlocked a new opening in healthcare. Thanks to the considerable benefits and opportunities, it has attracted the momentous attention of all the stakeholders in the healthcare industry. This chapter aims to provide an overall but thorough understanding of healthcare big data. The chapter covers the 10 'V's of healthcare big data as well as different healthcare data analytics including predictive and prescriptive analytics. The obvious advantages of implementing big data technologies in healthcare are meticulously described. The application areas and a good number of practical use cases are also discussed. Handling big data always remains a big challenge. The chapter identifies all the possible challenges in realizing the benefits of healthcare big data. The chapter also presents a brief survey of the tools and platforms, architectures, and commercial infrastructures for healthcare big data.

1. INTRODUCTION

Healthcare industry is going through a trailblazing makeover. Due to the significant advancement in digitised and open and pervasive healthcare systems, it is generating a massive amount of data. This healthcare data is truly comparable to the Big Data both in size and nature, hence termed as healthcare Big Data. Though compared to other industries, the healthcare industry has been lagging behind in the adoption of Big Data technologies, the changed medical and clinical landscape has forced the stakeholders to delve into the development quickly. Suddenly, Big Data has become crucial for almost every operational, clinical, and management task (Bresnick, 2017). The healthcare people are now convinced

DOI: 10.4018/978-1-5225-7071-4.ch004

of the benefits of Big Data and persuading themselves to analyse the data for extracting new insights that have given them the access to promising new threads of knowledge which are being transformed into innovative and purposeful actions (Groves et al., 2013).

Several healthcare use cases are well-suited for incorporating Big Data technologies. The healthcare Big Data analytics has opened up many exciting avenues in different healthcare operations including diagnosis and medical care, clinical decision support, population health management etc. (Bresnick, 2017). The success of healthcare Big Data is mostly dependent on the efficient collection and storage of massive quantities of disparate data acquired from diverse sources and also running it through an in-depth analysis (McDonald, 2017). The effective utilization of healthcare Big Data and obtained knowledge through analytical processes has the potential to save a significant amount of money and most importantly, people's lives (Lebied, 2017).

This chapter provides a preliminary and overall understanding of healthcare Big Data. The rest of the chapter is organised as follows. Section 2 discusses the 'big' aspect of the healthcare data including the ten 'V's of healthcare Big Data. It also covers the basics of healthcare Big Data analytics as well as the present healthcare Big Data market. Section 3 identifies several advantages of Big Data in healthcare. Section 4 mentions some application areas while Section 5 lists some specific examples of healthcare Big Data. The associated challenges are discussed in Section 6. Section 7 provides a brief survey of the different platforms and tools, architectures, and practical infrastructure for healthcare Big Data. Section 8 concludes the chapter.

2. BIG HEALTH DATA

2.1. The Data Overload

Back in 2012, a study estimated that healthcare data boast the largest share (30%) in occupying the overall electronic data storage in the world (Brown, 2015). To make the things graver, healthcare data is growing at a rapid pace, in fact seriously rapid. Among the growing digital universe healthcare is one of the fastest growing sectors. A report from the EMC Corporation along with the research firm IDC suggests that the digital healthcare data is growing at 48% per year whereas the growth rate of the overall digital universe 40% per year for the (IDC, 2014). The report estimates that the size of healthcare data will swell to 2,314 Exabytes by 2020 from a figure of 153 Exabytes in 2013 with an annual growth rate of 48%. The report elaborates that if all the digital healthcare data are stored on a stack of tablet computers, the height of the tower, by the year 2020, would cross 82,000 miles scaling from 5,500 miles in 2013 (Leventhal, 2014). The above statistics are sufficient to get a picture of the growth rate of the healthcare data and if this rate is continued, the healthcare data volume will soon reach the zettabyte and yottabyte scale.

2.2. The 'V's of Healthcare Big Data

2.2.1. Volume

The term 'Big' in healthcare Big Data itself certifying that volume of created and accumulated healthcare data is huge. Over the years, it is expected to grow even further. Clinical notes and trial data, lab results, personal medical records, claims data, FDA submissions, medical device data, human genetics

and population data, genomics, radiology images, 3D imaging, and biometric sensor readings are the primary information productive contributors in the exponential growth of healthcare data (Bresnick, 2017) (Raghupathi & Raghupathi, 2014).

2.2.2. Velocity

The healthcare data can be generated either by humans or sensors (Crapo, 2017). Traditional healthcare management systems are typically populated by the human-generated (entered by people) data. In that case, the rate at which the data flows into the system can be manageable. But widespread use of IoT and smart medical devices has led to the generation of real-time health data. Newly accumulated data come in constant and rapid flow which makes collecting and storing really challenging. Also, because of real-time health applications, the unified healthcare system's response time is very crucial. Though trying to react to every health data stream may cause the inefficient and uneconomical use of resources. So, it is important to decide which data require immediate action and which can be deferred.

2.2.3. Variety

Healthcare data come from various sources and, hence, in diverse shapes (e.g. structured, unstructured and semi-structured) formats (e.g. plain text, blobs of text, pictures, video, medical image, etc.) and sizes. Data captured at different estranged locations are typically incompatible because of different data structure and semantics. This requires the development of specialized data structures, communication protocols, and storage systems (Crapo, 2017).

2.2.4. Veracity

Veracity represents the quality of the data. Here, data quality includes integrated, trustworthiness, complete, and bias and noise free. Maintaining these qualities in health data is a serious challenge because of the diversity of sources and channels. Obviously, veracity is, probably, the most sensitive issue among the all 'V's in healthcare Big Data.

2.2.5. Validity

Validity represents the accuracy and correctness of data. The validity of data is also measured by how up to date it is and whether it is generated through standard scientific protocols and methods (Bresnick, 2017).

2.2.6. Viability

From the ocean of healthcare data, it is important to identify the relevant data for each use case (Bresnick, 2017). Relevancy of data is required to maintain for achieving desired and accurate outcome through analytical and predictive measures.

2.2.7. Volatility

As mentioned above, the healthcare data are generated and being changed at a rapid rate. Hence, they tend to live short. But the question is how short? How quickly does the data change? It is important to determine how long the data is relevant, how long to store it and the data of which time period should be considered for analysis (Bresnick, 2017).

2.2.8. Vulnerability

It need not be mentioned explicitly that the privacy and security are of utmost importance for healthcare data especially as data are stored in the cloud and travels to different data junctions (for details, refer Section 6).

2.2.9. Visualization

Healthcare data not only need to be correct and accurate, they have to be presented unambiguously, and attractively to the user. Large and complex clinical reports need to be presented in a way that is meaningful and less time-consuming to understand. Proper visualisation helps in finding valuable insights as it reflects the details in an expressive and usable way.

2.2.10. Value

The ultimate purpose of healthcare Big Data analysis is to gain value in the form of better health services. Better governance, better analytics, and smarter decision making are the core factors for creating maximum value out of healthcare data (Bresnick, 2017). For healthcare organisations, Big Data brings value by shifting profit pools towards the right direction by reducing overall cost.

2.3. Healthcare Big Data vs. Healthcare Smart Data

The healthcare Big Data is basically the data which contains the patient's electronic health records (EHRs) and other related medical information. The objective of this is to provide healthcare services based on the past records. Application of analytics and data science on Big Data gives an opportunity to find the rightful insight. Data analytics instead of trying to combine all the available data, target the right data and thus gives the physician the rightful insights. But still, the gap exists between what is demanded and what is stored in Big Data. But it is not necessary that a voluminous data may always reveal the information which is exactly asked for. Similarly, merely Big Data may not provide valuable and precise insights to the physician.

On the other hand, smart data are data which gives rightful information as per the context. It allows physicians to make smart choices like coordinating the treatment of a patient suffering from multiple illnesses, able to diagnose the diseases better and getting a complete picture of the patient based on the family history. For a particular patient's ailment, based on symptoms smart data accurately advices the right medication and predict any future complications. Smart data is the meaningful data that is extracted from the healthcare Big Data based on the type of data, its volume, and validity. (Bresnick, 2016).

Inferring smart data from Big Data need proper planning and implementation. The focus is what the organization need after Big Data is applied to the system. Based on the system requirement it is wise to enrich the Big Data with the appropriate type of valid data in quality and volume. For example, scanning endless PDF, X-rays of fractures, and blood test data would not necessarily help a physician in diagnosing a patient with stomach ailment or why someone is reacting negatively to a certain medication. To have smart data, the Big Data should be enriched with appropriate and correlated data. It is important to understand that what is being collected is what is being delivered (Bresnick, 2016) (Leventhal, 2014).

2.4. Healthcare Big Data Analytics

Healthcare Big Data analytics is expected to change the face of healthcare. The healthcare sector has been a bit hesitant to embrace the Big Data scenario quickly (TF7 Healthcare subgroup, 2016). With most of the healthcare services now slowly coming to Big Data, a whole new paradigm of services that can be provided to patients can be opened.

Till now the whole treatment and caring of the patient are done by the hospital, but with the advent of Big Data, the treatment has become much easier and more focused (Raghupathi & Raghupathi, 2014). Furthermore, the Big Data analytics will significantly bring down the cost of clinical treatment, administration, medication, etc. The healthcare Big Data will enable the physicians, by providing proper insights, to understand where the attention is needed at any given point of time and take care of the patient accordingly.

The healthcare Big Data, if used properly, has the potential to become the guiding force for the physicians to improve the existing healthcare facilities. It will also help to gain more insight into the health of local population so that the suitable health services can be rendered to them (Crapo, 2017).

2.4.1. Descriptive Analytics

In the descriptive analysis, the various events that have occurred in the past are described. Data mining is performed on the aggregated data to get the details about the event that has happened in the past. The descriptive analysis, in a nutshell, can be summarised by a simple question as “What has happened?” (TF7 Healthcare subgroup, 2016).

Descriptive analytics in healthcare provides insight into the origin and the reason for spreading diseases and how long the quarantine periods might need. It can also shed light on which medicines failed and which combination of drugs succeeded? It can also identify the unhealthy lifestyle that leads to the disease and also the measures taken for its cure.

2.4.2. Diagnostic Analytics

Diagnostic analysis is an analytic technique that uses the descriptive analytics details to identify the reason for an event i.e. it tries to identify the various causes due to which a certain event happened in the past. In a nutshell, diagnostic analytics can be said to ask a simple question “Why did it happen?” (TF7 Healthcare subgroup, 2016).

Diagnostic analytics in healthcare can provide insights into the causes of various disease outbreaks, administrative and patient treatment malpractices, etc. It is also helpful in identifying the reasons - why certain diseases remain dormant and when do they become active i.e. conditions required for the patho-

gens to germinate. Diagnostic analysis deals with finding out the reason behind an event in the past using techniques of data mining, correlation, and data discovery.

2.4.3. Predictive Analytics

Predictive analytics is an advanced analytics technique which uses data mining, machine learning, artificial intelligence, and statistical modelling to predict future events. Predictive analytics in healthcare empower us to follow the saying "prevention is better than cure" in the true sense.

The predictive analytics in healthcare has a great scope as we can predict diseases that will afflict a person based on his/her habits, genetic makeup or history, or based upon the work environment or home environment (Lebied, 2017). This type of analysis can play its role in a larger scale by helping us in judging the health of the population in the future thus allowing us to take proper precautionary measures. Predictive analytics, if used properly, can prevent many future patients from being in a hospital.

2.4.4. Prescriptive Analytics

Prescriptive analytics deals with finding the best course of action for the predicted scenario derived from the predictive analysis. The prescriptive analysis extends the predictive analysis and descriptive analysis to find out what to do with a predicted event. Based on the available predictions, the predictive analysis provides the answer to the question "What should we do?" (TF7 Healthcare subgroup, 2016). It can suggest possible decision-making options that can lead to mitigating risks.

Prescriptive analytics in healthcare can completely change the face of modern healthcare where not only the diagnosis will be done very quickly but also the treatment will be based on the patients' medical history. This will lead to a better treatment in terms of medicine incompatibilities and side-effects.

2.5. The Healthcare Big Data Market

The healthcare industry has gained a massive business value during the last couple of years. Fuelled by the large-scale adoption of EHR by governments, hospitals, and physicians, it is expected to grow to US$18.7 Billion by 2020 ("Revolutionizing the Healthcare Industry with Big Data, Analytics and Visualization", 2015). Whereas the healthcare and medical analytics are expected to grow at a CAGR of 27.1% to reach US$24.55 Billion by 2021 from US$7.39 Billion in 2016 (marketsandmarkets.com, 2016). The whole healthcare market will be changed due to the transformation of the healthcare data from the normal data to the Big Data. Big Data allows for change of so many things in the health industry, for example, it provides better data storage facilities at a low cost and also allows one to work on the data at a low cost. The Health Big Data market is open to various possibilities such as using the analytics power of the Big Data by the pharmaceutical companies to give better drugs at cheaper costs, reducing the administrative loads on the physicians. The healthcare Big Data will allow organisations to meet the consumer demand for affordable medical care. One of the major contributions of the goodies of using Big Data technologies in healthcare is applied to the unprivileged countries/states where there are inadequate doctors and medical facilities.

3. THE BIG DATA ADVANTAGE IN HEALTHCARE

Digitization of data accumulated from patient records, prescriptions, medical imaging, laboratory, pharmacy and insurance results in an enormous amount of data that is known as healthcare Big Data. This enormous amount of data can lead to knowledge traces that would significantly improve the quality of medical and healthcare services. The Big Data analysis would infer meaningful insights that help in making informed decisions, disease surveillance, and other healthcare and medical services. Patients, physicians, healthcare organization, pharmaceutical companies, policymakers and other stakeholders will be suitably benefited from the knowledge inferred. The applications would be individual and people in mass health surveillance; predicting health issues of individuals like calculating the medical complication and risk associated with a patient, the disease advancement/progression of disease; analysis of the particular kind of treatment suitable for the person; analysis of current treatment strategies to detect whether the diagnosis and applied treatment are correct or not, and thus adjusting the treatment plan accordingly. It may inform the patient about their current and future health state which allows them to make better-informed decision. The projection of Big Data application into the current medical and health service is advantageous in scaling the quality and accountability of health services, some of the advantages are discussed below (Raghupathi & Raghupathi, 2014).

3.1. Reducing Healthcare Cost

Today's healthcare system is a disease centred model where based on the symptoms of disease and lab reports, physicians reccomend the treatment procedure. The treatment procedure is quite centred on clinical expertise and medical evidence, further based on treatment results, other alternative treatment options are suggested. The treatment procedures are recursive, time-consuming and costly. Healthcare Big Data with the enormous amount of data on medical cases, treatment patterns, medicines and their effectiveness, genetic data allows for a patient-centred model of treatment. This information allows analysing and diagnosing the patient correctly and further predicting the correct treatment pattern. In another application, Big Data in genetics ensures personalized profiling and hence personalized medicine for patient allowing accurate treatment. Healthcare Big Data infer knowledge do actually reconcile the redundant costly treatment procedure (Srinivasan & Arunasalam, 2013).

3.2. Reducing Hospital Readmissions

The probability of patient returning back within a month after treatment at the hospital are getting high. This incurs cost and resource to hospital. Big Data analytics applied to the patient medical record, history, chart information, and the patient lifestyle record can identify the patients who are at risk of having medical complications and again readmission. Thus, identifying the patients who need additional care would able to reduce the hospital readmissions ("Seven Big Data Examples That Have Improved Healthcare Operations", 2016).

3.3. Optimized Workforce

Hospitals and other health organizations critically face the classical problem of recruiting staffs and scheduling them. For many staff recruited, there will be a risk of under usage of resources while minimizing the numbers can lead to poor customer services. This problem can be overcome by the application of Big Data. Big Data obtain data from various sources can well predict the number of patients that would come on the daily and hourly basis. Further, hospital admission record analysis over a time series of years can get the future trend on the type and number of patient that would be admitted to the hospital. The application of data science for crunching the raw data delivers information that would allow the organization to optimize its workforce based on the future requirements (Lebied, 2017).

3.4. Real-Time Alerting

One of the biggest functional advantages of Big Data is real-time alerting. Big Data incorporated into services in hospital and clinics would analyse the medical records of patients and thereby in real-time provides advice to a physician for prescriptive decision. Further patient's medical and health data collected by wearable devices (Pramanik et al., 2019) are being analysed in the cloud in association to Big Data to feed real-time information of patient current medical issues, future medical ailment, and preventive measures. The patients' data being collected over time would allow the doctors to see the health status of people in society and thereby taking corrective strategy for any deviation observed. The real-time data is also shared by the health organization to react immediately to any disturbing results found in patient monitoring (Lebied, 2017).

3.5. Analysing Electronic Health Records (EHRs)

Analysing the enormous electronic health record would make a physician comprehend the treatment trends, their effects, and side effects. Moreover, the Big Data allows the updated patient data to be shared among doctors for their treatments and, also, medical cases are exchanged among doctors thus reducing duplicate test and redundant treatment procedure ("7 Big Data Use Cases for Healthcare", 2016).

3.6. Analysing Hospital Networks

Analysing hospital care and management data would be useful in administrating critical medical condition case, strategies for reducing postoperative infection and analysing the medicine (antibiotics) and treatment procedure carried out by doctors which found ineffective in a cure. ("7 Big Data Use Cases for Healthcare", 2016)

3.7. Control Data for Public Health Research

Medical and health organizations, hospitals and small clinics produce huge medical data. This is overwhelming data for medical professionals. As most of the data are raw, unstructured and cluttered, probably these data in their direct form are difficult for medical professional and policymakers to suitably

use in public health regulations and decision making. There is a gap between what health professionals want and to what data exists. These data without the application of Big Data is of no use. Application of data analytics can regularize and standardize the public health data, filling the existing gap and thus providing information that can be applied in various regulation and research work for providing better care and health services ("7 Big Data Use Cases for Healthcare", 2016).

3.8. More Efficient Medical Practice

Big Data helps in providing effective treatment for the patient. Big Data analytics gives knowledge about the diseases, their symptoms, treatment pattern, and medicine effectiveness in different stages of illness to the physician that would help them for better-informed healthcare practice while treating a patient.

3.9. Using Health Data for Informed Strategic Planning

On the outbreak of diseases or growth in chronic disease rate among the population, Big Data on public health records like immunizations, distribution, and availability of medical services enables to design strategies for public health. Big Data map the required services over a region in comparison to the available services and thus inferring what necessary steps need to be taken for controlling the public health issues (Eastwood, 2013).

3.10 Optimizing Workflows in Healthcare

Hospitals have multiple departments, limited patient admission capacity, limited doctors and staff with varying schedule and varying flow of incoming patients with different need and treatment make healthcare services in the hospital very dynamic. The patient's health issues span multiple factors like involving treatment from different departments and different physicians at different stages of treatment, laboratory tests, and medication. Further, there are cases of medical complication for inappropriate diagnosis and treatment and hospital readmissions. This volatile state demands completely planned and managed workflow ready for an emergency and changing situations. Healthcare data enriched by multiple sources of data such as real-time patient data (patient health status, location etc.), medical records of patients, nursing information, laboratory data and machine status (available, working etc.) helps to identify the current operational state of hospital and thus allow to take informed decision for better healthcare services by best utilization of resources ("TF7 Healthcare subgroup", 2016).

3.11. Better Safety Practices

The patients who underwent surgery are often suspected of postoperative infections. Big Data analytics of electronic medical record can well inform in previous the susceptibility of patients for getting an infection. This ensures preventive measures like intensive postoperative care, pre- and post-operation preventive medication etc. This predictive analysis promotes better patient safety ("Seven Big Data Examples That Have Improved Healthcare Operations", 2016).

3.12. Patient Engagement

Today's one of the healthcare challenges is patient's engagement leading to patient's awareness, retention and proactive care. A huge spending is done on the healthcare segment though even after that, patients are never properly taken care of resulting in bigger health problems. Patients engagement are found in getting insights, communication, personal wellness tracking, medical management, new disease diagnosis, personal history, immunization, health cost planning, social services research and clinical trial participation etc. Patients want to be engaged so that they are aware, engaged and feel confident (Shah, 2016). This further improves patients' satisfaction and experience leading to the better outcome. Patient interactions with physicians, hospital and clinic staff, pharmacy and pharmaceutical company, laboratory and other counselling staff increases one's engagement. Controlled and updated data about the patient helps the service provider to attune the services to patient's need, helping patients to be more engaged in managing their health issues. To realize the situation Big Data has contributed to a bigger extent. Using Big Data insight, healthcare providers can identify the individuals with critical illness or need support in terms of insurance, medical counselling, care and managing their medical cases. These people are informed regularly. Patients are delivered with a personalized message or called. The personalized messages make people aware of their individual journey as a patient. Utilising Big Data technologies, when the patient makes contact with the clinic or hospital for healthcare service, the physician or the hospital staff may already be aware of patient's history, communication preference and their healthcare journey (Disch, 2016).

3.13. Big Data Is Helping to Prevent Unnecessary ER Visits

Many often patients frequently visit multiple hospitals and clinics for the same or different ailment or disease which causes repetitive treatment procedure, discontinuity in treatment and in the worst case, a wrong treatment that leads to critical health jeopardy and wastage of money and time. In this direction, the application of Big Data would help the patient in reducing unnecessary visits to medical and health agencies, further make sure the treatment is on the right course. Big Data allows the updated medical record of patient to be shared in emergency department so that when patient shifts between hospitals/clinics or take a new treatment for ailment, the staff will have the knowledge of patient's medical history, treatment procedures carried out, drug administered and the reports of various lab test that are already done. This definitely eases the patient diagnosis, reducing patient overhead and increasing satisfaction (Lebied, 2017).

4. APPLICATIONS OF HEALTHCARE BIG DATA

4.1. Battling the Flu

Flu is an epidemic which may outbreak at any moment of time with potential to spread across in a short period of time. Every year around the world Flu takes millions of lives. Centres for Disease Control and Prevention (CDC), an operating component of the Department of Health and Human Services in the USA is fighting against disease and their control and prevention. The CDC receives 700,000 reports of

patients with Flu-like symptoms every week. Doctors, hospitals, and laboratories send a large amount of data to CDC. The reports include the where and what kind of treatment was given by the physician. Big Data analytics has been tremendously useful to get scientific insight from the data stored over time. CDC has made the information public through FluView, an application developed and deployed at CDC (Nambiar et al., 2013). FluView reports in real time how the influenza is spreading across, along to that how vaccines and antivirals can aid patients etc. It gives doctors the answers they need to effectively battle the outbreak, such as knowing which vaccine is effective to which virus strains and whether or not antiviral drugs will be effective in recovery.

Other similar applications where Big Data have been used to capture and analyse the data on the pandemic are FluNearYou[1] and GermTracker[2]. FluNearYou asks the person to input data on their symptoms; based on the symptoms the application generates a map that can allow users to prepare themselves and take preventive measures against infection. GermTracker takes the data from social media posts and analyses them for the disease outbreak. The huge data obtained thus gives a view of a pandemic which doctors might have missed.

4.2. Diabetes and Big Data

Diabetes has become a major health problem in the world. Most of the people are diagnosed with type2 diabetes, a condition where the body cannot use insulin effectively. In these cases, Insulin injections are administered to patients to control sugar level in blood. Diabetes can be controlled by monitoring the patient's sugar levels regularly and further administrating insulin based on the situation.

In managing diabetes and the betterment of patient health, Common sensing has pioneered smart insulin pen cap technology called GoCap[3]. GoCap is a data-driven smart cap that fits over the injection needle of a standard insulin pen and measures the volume of insulin in the pen. GoCap records the time, amount and kind of insulin in the logbook and further sends the data through the Internet to mobile devices of doctors and family members. The information is also displayed and recorded in the mobile device application. The GoCap mobile application helps the patients to enter information about their food intake and glucose levels at the different time of day. These accumulated data on patients helps the doctors to provide personalized treatment to patients.

4.3. Fight Cancer

Big Data on cancer treatment is an accumulation of data collected from cancer patients at different stages of treatment like pre-detection, pre-treatment and end stage. These data can effectively be used in cancer prediction for new cases. Using the historical data of patients, the predictive machine learning algorithm can well identify cancer. Flatironhealth has developed an Oncology cloud service called OncoCloud (Flatiron, 2017) which aims to gather data during diagnosis and treatment and further make it available to the researcher for their advanced study. Another example of Big Data in cancer treatment programme is the Cancer Moonshot Programme (Orcutt, 2016). Cancer is triggered by genetic changes and often genome information is required for therapeutic treatment. In this direction, Big Data as a huge data repository provides enough information about human genomics, patients case reports and their background details. The valuable insight which can be obtained from these data would help to identify the genetic changes which had trigger cancer in a patient and the treatment procedure. Big Data which

consist of millions of readable and usable samples would allow scientist all over the world to analyse and work on the data. For widespread and new trends of treatment, sharing of data is essential so that the medical researcher can use a large amount of data on treatment plans and recovery of cancer patients.

4.4. Improved Diagnosis and Treatment

In the healthcare industry, Big Data plays a pivotal role. The application of Big Data includes analysis of the disease pattern and predicts outbreaks. Furthermore, it allows public health monitoring and tracking epidemic and helps doctors and medical health policymakers to prevent future outbreaks. Big Data allows the doctor to keep an efficient record of patient's medical history and thereby enable doctors to provide proper healthcare service for their patients. Further, the data insight allows doctors to carry diagnosis and treatment of patient located remotely connected to the Internet. The patient can contact doctors and get the advice within a minute. The enablement of IoT helps to monitor patient and raise real-time alerts. Big Data helps both sick and healthy patients to connect doctors for improved health and treatment.

4.5. Tackling Opioid Abuse

Opioid abuse is often described as a new "epidemic" in the world. More than 15000 patients die every year on painkiller drug medication and several thousand are killed for the drug overdose or wrong usage. Pain medication is a tricky area for physicians; depending on individual person's capacity doctors prohibit prescribing the opioid use above a certain threshold. Ignoring the prescribed medication and their dosages, patients misuse the drug intakes and often uses them wrongly to what is called as drug abuse. The responsibility of reducing and preventing the drug abuse is split between the patient himself and the healthcare system. In America, to stop this, monitoring programs have started. This includes maintaining a centralized database with data analytics and decision-making system which detect prescription abnormalities, track drug dispensing patterns and preventing the patient from taking dangerous amounts of opioid. Electronic Prescription for Controlled Substances (EPCS) (HealtIT.gov, 2016) - a standard practice which has helped to reduce the opioid epidemic by enabling the practitioners to prescribe medicine electronically which can be shared and monitored everywhere.

4.6. Precision Medicine

The EHR is a major source of Big Data, the EHR contains sociodemographic, medical, and genetics treatment details which allows scientists and clinicians to predict more accurately for a patient's ailment and further reasons for precision medicine. The precision medicine model allows appropriate therapeutic and preventive approaches based on patients' genetic makeup, lifestyle, and environmental factors that work effectively for an individual.

Doctors are often misled in case of instantaneous diagnosis (for emergency) and symptomatic treatments. Applying data science and analytics over Big Data, precise information based on the patient context can be obtained. This information would serve in making precise prescriptive decisions.

4.7. Evidence-Based Medicine

Around the world, evidence-based medicine (EBM) is a standard way of medication. Clinical trial on diseases or ailment treatment methods includes rigorous experiments on patients and involves risk. It is often difficult to generalize the clinical trial methods testing and often the positive test result on a small group of people may not work on the outside world.

Big Data helps in EBM to sort out and identify the effective clinical methods applied to real patients. By data mining over practice-based clinical data of actual patients, it is easy to find out which patient has gone under what treatment, how well the treatment has done or has it any side effects. The data can then be analysed at an individual level to create a patient data model, which can be aggregated across the population to infer wider insights into disease prevalence, treatment pattern etc. Clinicians match the symptoms to a larger patient database and identify the accurate disease in a faster and efficient way (Marr, 2016).

4.8. Genomic Analytics

Researchers are getting a fascinating new perspective on human genome due to advancements made in Big Data analytics. Since genes are all about information, Big Data is a perfect fit for genetics. Researchers are able to look more closely at human genes and they are applying Big Data analytics in this issue. A typical human genome contains more than 20,000 genes, simply mapping a genome requires 100 gigabytes of data. Sequencing multiple genomes and tracking gene interactions, multiples the number by many times. Much of the work done with the human genome and Big Data analytics deals closely with health and medicine. One of the pioneering application of genetics and Big Data is creating personalized medicine and new drugs. One's gene analysis could help in finding the inheritable traits that can be passed to the next generation. This is specifically important for people who are susceptible to passing disease like diabetes, anaemia, heart disease, cancer, obesity etc to their child. Big Data analytics helps the researcher in understanding the human genome and they can get a view of future diseases.

4.9. Clinical Operations

Big Data accumulating a large amount of clinical data from various sources can assist staff in clinical operations. It helps in enhancing the operation of healthcare facilities, the quality of clinical trial and research. Clinicians can do the comparative studies on different patient cases in the different area of treatments. Big Data has changed the mindset of researchers from hypothesis-based treatment to the research-based analysis on a large amount of data resulting in accurate treatment method.

4.10. Patient Monitoring

The wide adoption of EHR in hospitals and clinics has enriched Big Data. Big Data provides a deep Clinical knowledge and understanding of disease and illness. For patient monitoring, various sensors are used to screen for patient's vital statistic like blood pressure, heartbeat, respiratory, oxygen in the blood, and sugar level etc. Any change in the data pattern is analysed along with Big Data and further alerts are raised for preventative measures. Big Data having patient's past medical knowledge, analyses the current statistics to predict the future medical complication and ailments.

4.11. Medical Device and Pharma Supply-Chain Management

Proper functioning of hospitals and clinics availability medical devices and pharmaceutical supplies are proper in quantity is of utmost importance. Many hospitals are not implementing the appropriate supply chain strategies, this cost to hospitals in terms of poor quality services. Nonstandard ordering methods and unnecessary product may cause stockpiling similarly also may cause in supply deficiency. The supply chain is not only about the product but also about the people who buy, move and use them.

Implementing Big Data analytics and automation tool makes supply chain management much easier. Automation allows clinicians to spend more time delivering high-quality care and less time to find for products availability. Technology like RFID (radio frequency identification) has been used for tracking products. The stocking products are fitted with a scannable tag which can provide information about items, manufacturing date, expiry date and shipping date, which allows providers to track the object throughout its life cycle. Predicting the number and type of patient and the supplies required allows maintaining the stock at its optimum level. For the stock, level decreases automatic stock ordering is done (Castle & Szymanski, 2008) (Uzsoy, 2005).

4.12. Drug Discovery and Development Analysis

Drug discovery requires and processing and analysing of unstructured and structured biomedical data obtained from various surveys and experiments. This includes data regarding gene sequencing, protein interaction data, drug data, electronic patient records and clinical trials, self-reporting data etc. These data collected in a huge amount from various sources over time has enriched the medical information base. The pharmaceutical company, while designing a new drug or applying clinical trial, applies analytics to these data to infer knowledge and to build a predictive model for new drug design. This usage of past data has made the drug discovery and development faster in comparison to the traditional method (Jawadekar, 2016).

4.13. Big Data for Personalised Healthcare

There is a rapid advancement in genomic technologies. With large sources of data, it is easier for healthcare professionals to understand the disease mechanisms which leads to better treatment of patients. Many diseases have preventable risk factors. A clear understanding of disease characteristics helps in personalized healthcare and thus reduces the chances of getting the disease. Healthcare professional takes patient's history, do the physical examination and laboratory testing to identify the risk of disease in future. The application of Big Data in this direction makes the process fast and easy. Analytic on patient's health and medical profile gives a personalized medical view which is specific to a patient. The personalized profile for treatment is derived not only from the particular patient's electronic medical report but also from other similar types of patients' case records. Personalized healthcare enables disease risk profiling, disease management plans and wellness plan for an individual patient.

4.14. Infection Prevention, Prediction, and Control

According to the European Centre for disease prevention & control, every year estimated 100,000 patients are infected with hospital-acquired infection (HAI). It is estimated that around 37,000 people died with direct consequences of infections per year. The World Health Organization (WHO) has strict guidelines which can minimize the risk of spreading infections. Some guidelines are easy to follow and some are hard to implement due to limited technology. In this direction, Big Data technology integrates genomics with epidemiology data in order to not just control but also to prevent as well as envisage the possibility of the spread of infections within hospitals and clinics ("TF7 Healthcare subgroup", 2016).

4.15. Health Insurance Fraud Detection

Healthcare fraud is a big problem all over the world. Over the last decade, the healthcare industry had spent billions of dollars on improper claims. More than 1.5 million people have been victimized. These statistics represent avoidable healthcare cost. Healthcare industry is applying analytical controls throughout the treatment process and also incorporating the claims review process. Review process incorporates rule-based data analytics and predictive modelling. The treatment procedure carried and the medicine administrated over the patient is analysed for similar symptom cases to find out whether the treatment and drug used are legitimate as per the context or vague one ("Seven Big Data Examples That Have Improved Healthcare Operations", 2016).

5. USE CASE EXAMPLES

5.1. Asthmapolis

Asthmapolis has launched a GPS enabled tracker called propeller that records the usage of asthma medicine by asthmatics. It helps the medical practitioner to treat asthma more effectively. Propeller[4] uses Bluetooth sensors-based inhaler, mobile application, and advanced analytics. The sensor records information that when and where the patient suffered from an asthma attack and used the inhaler. The information is transferred to a centralized database to be combined with the information at the Center for Disease Control. The combined information identifies the trends and catalyst of asthma attacks in individuals and aggregating the pattern with mass population. This leading information would help the physician to identify when the risk of an asthma attack is higher, and further help them in personalized treatment for asthma patients.

5.2. 23andMe

23andMe[5] is a privately-held personal genomics and biotechnology company. The company is named for 23 pairs of the chromosome in a normal human cell. It is saliva-based direct to consumer genetic testing business. 23andMe extract the genome information from the person saliva for delivering back to the customer. The company has built a huge genetic information bank. The aggregated customer data

are available for research team employed by 23andMe and other scientific groups for inherited disorders and other diseases. 23andMe provides raw genetic data which in its original form are not meaningful in different genetic studies. Promethease a personalized online tool for health genetic information processes the raw data from 23andMe and provides personal DNA report (Ramsey, 2015).

5.3. USC Medical Monitor

Parkinson's disease is affecting millions of people today. By 2030 it will be doubled. Since this disease affects a patient's movement, it is important to monitor the patient's mobility condition. In the University of Southern California, computer scientists are teamed up with neurologists, kinesiologist & public health experts to fight against Parkinson's disease (Nambiar et al., 2013). They use different types of devices to monitor patient's movement throughout the day and gather a large amount of data about patient's abnormal movements through 3D sensors, mobile devices and from wearable sensors. The acquired patient data is fitted into algorithms which analyses the kind of changes in movements. The analysed data and reports are sent to researchers and caregivers. Depending on the motion abnormalities alert are sent to the caregiver, depending on which they prescribe appropriate medication and exercises (Nambiar et al., 2013).

5.4. GNS Healthcare and Aetna

Applying Big Data analytics, GNS Healthcare[6] in collaboration with Aetna[7], an insurance-based company aims to treat and prevent metabolic syndrome. The metabolic syndrome can increase the risk of heart disease, strokes and diabetes. GNS uses the claims and health information from Aetna, as a platform called - Reverse Engineering, Forward Simulation (REFS) to create data-driven models. GNS healthcare analyses the risk of getting metabolic syndrome depending on the five conditions i.e. increased waist size, high blood pressure, high triglyceride level, HDL cholesterol and high blood sugar. GNS analysed the data of 37000 Aetna's customers who participated voluntarily in the screening program of metabolic syndrome (Nambiar et al., 2013). The analysed data includes claims, records, pharmacy claims, demographics, laboratory test, biometric screening. GNS & Aetna team uses two analytical models:

1. Claim based model
2. Both claim and biometric data based model

Both analytical models predicted future risk of metabolic syndrome in both population and individual (Nambiar et al., 2013).

5.5. WestMed Medical Group

In the WestMed medical group, as the medical practice growing day by day, the number of physicians grows from 16 to 250. The physicians are seeing 250,000 patients with annual revenue of $255 million ("7 Big Data Use Cases for Healthcare", 2016). Using Big Data, the medical practitioner is now capable of analysing more than 2,200 procedures. The Big Data has helped in streamlining the workflow, shifted clinical work from doctors to nurses, reduced unnecessary testing, and improved patient satisfaction.

5.6. HealthCore (WellPoint)

Healthcore[8] is a clinical outcome research subsidiary of Wellpoint Inc. Healthcore has a team of highly experienced researchers including physicians, pharmacists, epidemiologists, health economists and scientists. They measure safety, efficacy, and effectiveness, compliance of drug, medical devices, and care management interventions in the real-world settings ("Epidomiology and Genomics Research Program", 2018). The Healthcore uses data from different sources like health plan providers, patient's reported information, clinical information from physicians that allows the researchers to answer complex clinical, economic and health policy questions. HealthCore interprets health data and estimates the influence of disease, their treatment and care on the outcome.

5.7. Evolent Health

Evolent Health[9] (previously Valence Health) provides value-based care solutions for health sector like hospitals, health systems and physicians to help them to effectively manage patient populations. Based on each client requirements, customized value-based care model for clinically integrated networks, population health management, health plan administration & TPA, risk adjustment, pharmacy benefit management is designed, built and managed. To handle customer's enormous and heterogeneous data, Evolent Health has adopted Big Data technology. Evolent health has been using MapR for building data repository for storing huge data. MapR offers a distributed data platform to store and analyze Big Data in a distributed fashion which is linearly scalable which can be extended just by adding more machines and/or CPUs, without changing the application code. The MapR platform delivers enterprise-grade security, reliability & real-time performance which lowers both hardware and operational costs of most applications and data. Evolent consumes 3000 inbound data feeds with 45 different types of data. This data includes lab test result, patients' health records, prescriptions, pharmacy records and claims and payment records which are used to make decisions about improving healthcare outcomes & reimbursement. Prior to the MapR, the company takes 20 hours to 22 hours to process millions of data. But MapR cuts the time to 20 minutes, which also requires less hardware.

5.8. United Healthcare

United Healthcare[10] provides health benefits and services to millions of people. Their payment team has the tough task of ensuring the claims are paid correctly and on time and not paying for the fraudulent services. They shifted to a predictive modelling environment based on Hadoop which provides a flexible (seamless integration of new tools and technologies) and cost-effective platform with enterprise-grade features (e.g. high availability and disaster recovery). Bringing the information about claims, prescriptions, plan participants, contracted care providers and associated claim review outcomes to a single framework thus people at United Healthcare are able to identify inaccurate claims in a systematic and repeatable way ("UnitedHealthcare uses Hadoop to Detect Health Care Fraud, Waste and Abuse", 2018). With the help of the data integration from different data silo, it has become easier to find out the fraudulent. United Healthcare also started using NLP to better understand customer satisfaction by converting records of customer's voice (who calls to its call enter) into text and search the indications of customer's dissatisfaction.

5.9. OptumInsight

Optum Labs[11], a collaborative research and development initiative of the Optum Inc of United Health Group and the Mayo Clinic, is pursuing a variety of Big Data analytics projects aiming to improve patient care and costs reduction. It is combining electronic medical record from Mayo & other healthcare organization with claims data from United Healthcare to understand and provide better and more effective health care and also to analyze the total cost of care for specific procedures or diseases (Cambridge & Rochester, 2013).

5.10. Liaison Technologies

Liaison Technology[12] offers a cloud-based solution which provides organizations with the ability to integrate, manage and secure data across the enterprise. They are particularly active in the healthcare industry where a patient's record may be stored in many systems in different formats. The storage is updated in real-time as per the streaming data so that the user always gets the up-to-date view of the data in a single location and in a most suitable format (McDonald, 2017).

5.11. Novartis Genomics

Next Generation Sequencing[13] (NGS), a research initiative carried out at Genomics Institute of the Novartis Research Foundation, is a Big Data application that deals with a huge volume of heterogeneous data. The data includes genome information and other medical data. NGS requires heavy interaction with diverse data from external organizations for experimental and other associated data. The Novartis Genomics suffers from two problems, one integrating heterogeneous datasets and the other in integrating public datasets. To address the problem Big Data technologies like MapReduce and Apache Spark are being used by Novartis. As a result, the combined Spark & MapReduce based workflow and integration layer allow the company's life science researchers to meaningfully take advantage of thousands of experiments that different public organization have conducted (McDonald, 2017) and thus helped in accelerating drug research.

6. CHALLENGES

The major challenges of using the Big Data in healthcare is that health data is completely distributed i.e. one cannot find a complete repository of data related to a patient at one location and in a homogeneous form (Lebied, 2017). The above statement basically points out that we'll need new infrastructure so that data can collaborate together. We have to add other new technologies to the existing older ones like predictive analysis, machine learning and graph analytics (Lebied, 2017). Some important challenges in healthcare Big Data are discussed below.

6.1. Expertise

Whenever a new technology is introduced, expertise on that technology is not achieved overnight. We must understand that most of the hospital IT staffs are not trained to used Big Data. They are familiar with SQL programming. To resolve this knowledge gap quickly and cost-effectively Big Data must allow itself to be extremely user-friendly.

The whole scenario here is that if the expertise of the user is not in the current technology he/she will lose interest in it very quickly and the investment in the technology will be dead. To avoid such mishaps, we the organisations which convert to the Big Data architecture must use various tools which help in bridging the gap between the new technology and the old. The tools must allow for the usage of the languages know previously i.e. the architecture must have backward compatibility. Creating a correct balance of backward compatibility and the new architecture must be balanced for the results.

6.2. Costly Processing and Analysis

The most common challenge in any data processing architecture is that the cost of processing the data. The medical data which comes from various pathological tests itself comes at a huge cost due to the requirement of sophisticated equipment's required to conduct the test. Add to this cost the processing charges for the additional analysis of the data to diagnose a specific disease. To understand the cost and analysis of the medical data, we must come to terms that present-day medical diagnosis and the further prognosis are completely based on the symptoms shown by the patient. These symptoms can be the same for multiple diseases thus leading to a further increase in cost for the diagnosis of the disease. We can very well say that proper handling of the data and its analysis is a very big challenge.

6.3. Security & Privacy

The collection of health data and creating EHR is the way of digitalizing the health data. The major problem with digitalization is the security of the data as well as maintaining the privacy of the patient to whom the data belongs to. The analysis of healthcare data must be done with keeping in mind that under no circumstances the privacy and the personal information of the patient are known anyone including the system i.e. a level of privacy encryption must be imposed on the data that is to be analysed (Patil & Seshadri, 2014). The security architecture of the cloud (or where ever the data is stored) must prevent data breaches and even if breaches cannot be completely avoided try to make sure that data is not traceable to any patient to whom they belong.

6.4. Ransomware Threat

Ransomware is classified as programs that infect the victim's system and prevents the victim from using the system until the ransom is paid. Ransomware basically encrypts the user's files and data and withholds the key to decryption as ransom. Hospitals are particularly vulnerable to this type of attack as the security systems employed by them are not up to the mark thus making the hospitals and medical data repositories as soft targets for such attacks. One can easily understand this as hospitals need their data on a day-to-day basis as the patients' lives depend on the data generated. To process healthcare data, we

must up the security from various sources from where the data is collected so that the main systems are not compromised by these attacks.

Ransomware threats can be further reduced by utilising proper antivirus software's as well as having strict monitoring of the data. Also, regular backups of data to places which are completely secure can help from prevention as well as recovery of the valuable data.

6.5. E-Mail Vulnerability in Healthcare

E-mail vulnerability emanates from the various ways in which an email can be used to install a malicious code in the system which in turn can lead to theft of data. As more and more health data are being digitalized, health institutions are falling prey to more and more email-based attacks like phishing, where the people working in those institutions are conned into activating the attack themselves. To remove or to fight against such vulnerabilities, we must follow a three-step approach which includes proactively monitoring the threats to the continuous management of security ("Email Vulnerability in Healthcare", 2017). Also, we must bank for the worst possibility that the data is siphoned off and go for rapid incident response and recovery.

6.6. Data Governance

Data governance is the practice of managing the data assets throughout the lifecycle of the asset so that it meets the organizational needs throughout its lifetime as well as maintains its integrity (Bresnick, 2016). Data governance is required for the health data because as we are digitalizing the data if it is not properly organized, updated and maintained then that data cannot be used for anything other record keeping. We must understand that data governance is important for analytics perspective. To find more from the data we need the data to be organized rather than just being electronic. Some possible solutions to this can be that healthcare institutions hire dedicated data entry operators in their company.

6.7. Data Management

As the digitalization of the healthcare data is done, the need to manage such data also grows. We can't implement traditional methods as our analysis model is based on Big Data architecture. We must manage the data in such a way so that the maximum amount of data is always available for processing. Below are the ways via which data can be managed ("TF7 Healthcare subgroup", 2016):

- **Data Quality:** The quality of data used for processing is important as if we feed low-quality data to the system the analysis of the data will not be worth the effort. We need to focus on data quality as it is the only raw material in the system which will be used to generate valuable results. For every data that is used we must be aware of how the data was collected, conditions under which the observations were made, and how the contents were processed and transformed. This enables reproducibility of experiments as well as the reliability of the data. The quality of the data can significantly influence the conclusion of the whole analysis.

- **Data Quantity:** The quantity of data used for analytics plays a very large role in the healthcare sector. We must understand that this sector is knowledge-intensive and requires data and analytics to improve its practices. The quantity of data used in the analysis, allows us to draw greater insights into the processes. Also, if the quantity of data is less, then the whole analysis can sometimes become questionable as lack of data may not provide proper authenticity to the analysis.
- **Multi-Modal Data:** The healthcare data as mentioned earlier comes from various sources. All these data have various formats used by their respective system. This data can be further classified into structured and unstructured data. The system should accept both the data and try to find synergy among it. This type of integration of data from all sources are helpful in finding new cures for diseases or finding the genomic point of view of disease i.e. can a disease be genetically transferred to the next generation or even if transferred will it remain dormant or not. These types of analysis require data from all the kinds of sources like ancestry of a person, family history of diseases, his/her habits as well as of their family's etc. thus we can say that multi-modal data integration and analysis plays a pivotal role in healthcare analysis.
- **Data Access:** All the analysis of the data and its inference is only possible when and only when access to data is provided. The biggest challenge of today is getting the data i.e. getting access to the data after bypassing the various privacy laws, security barriers etc. and to add hurt to the injury these laws vary from country to country. The health data is highly fragmented i.e. data is distributed among hospitals, pathologies or even specialists whom the patients contact. One of the possible solutions could be that we ask patients to willingly give their data to the analytic system so that all the red tape can be cut easily.
- **Patient-Generated Data:** The patient-generated data (PGD) refers to the data which is generated by the patient's devices and gives us an insight about the patient's habits and can also let us know about the diseases he/she is prone to or maybe suffering from. PGD also provides ways to monitor recently released patients from hospitals and continuously monitor their progress. The major challenge in PGD is that there are various types of devices, a present which is all not compatible with each other as there is no standardization about them. This causes problems and conflicts with the data that is generated.
- **Data Integration:** As it is already known that health data is heterogeneous, so before performing analytics on it we need to integrate the data from all the sources. The data integration in itself is a very big challenge as the data that is provided may be or may not be structurally compatible to each other i.e. they all may exhibit a level of structural heterogeneity if this remains the analysis can be incomplete or inconclusive. To solve this problem, we can use information extraction mechanisms in conjunction with machine learning and semantic web technologies which will help us to get homogeneously integrated data for analysis.

6.8. Fragmented Analytics

As the cost of centralized analytics systems is very high, the industries are now moving towards fragmented analytics model i.e. we analyse the heterogeneous data in fragments and then try to combine the analytics results (Bresnick, 2015). For the healthcare industry, it is assumed that it will not be successful to a very large extent as it is considered to fail at data integration and reporting of states. The Fragmented

analytics model is a cheap way of implementing any data analytics system, thus a possible solution can be to maintain tables which show a correlation which previously exists among data and then use that table for data integration and reporting.

6.9. Choosing the Right Healthcare Big Data Analytics Tools

Till now we have seen the various challenges that plague the Big Data applications in healthcare. Adding further to that, we would like to elaborate on how choosing the right Big Data tool also a challenge. We must understand here that the tools must be chosen on the basis of use otherwise we'll end up increasing the cost of the whole architecture, and not even get the proper result. For example, suppose we want to use the Big Data architecture for clinical analysis and not for qualitative benchmarking (Bresnick, 2015). We must understand that the organisations must self-evaluate them and then order better tools which provide better optimisations and digestible reports (Bresnick, 2016).

7. HEALTHCARE BIG DATA IMPLEMENTATION ENVIRONMENTS

7.1. Platforms and Tools

Big Data is implemented in the healthcare domain using various tools. Below some tools and their uses are mentioned:

7.1.1. The Hadoop Distributed File System (HDFS)

The HDFS (Shvachko et al., 2010) is a cluster file system which is used to store structured and unstructured Big Data across multiple systems in a large-scale cluster (Sarkar, 2017). It allows Hadoop to process the massively-scaled volume of data in a fault-tolerant manner. HDFS is used in the healthcare system to collect and aggregate related data while maintaining the data security and privacy in an effectively better way compared to traditional storage systems (Raghupathi & Raghupathi, 2014) (Sarkar, 2017).

7.1.2. MapReduce

MapReduce[14] is a programming model for implementing, analysing and generating Big Data sets. The programming model is composed of the map and reduce methods. The map methods do filtering and data sorting whereas the reduce method summarizes the data. The MapReduce follows the divide-and-conquer approach i.e. it solves by reducing the problem into smaller sub-problems or the input data into independent chunks which are processed completely in parallel, and then the result is combined (Raghupathi & Raghupathi, 2014) (Sarkar, 2017). This algorithm provides a fault-tolerant and flexible way to analyse large data sets over distributed architecture. For large data sets like health data, the data is analysed over a distributed architecture and then the results are collected from different points of processing and combined.

7.1.3. Pig and PigLatin

Pig[15] is a data processing architecture built on top of Hadoop. Pig is designed for making Hadoop simpler to approach and use. The Pig tool can be used for easy and faster cleaning and analyses of big data sets. The pig execution has two modes the Local and Hadoop mode. The PigLatin is a script-based language to express data flow, data input and operation on data to produce the desired output (Hurwitz et al., 2013). The simple and easy to write PigLatin language is used to write programs that bring the structured, unstructured and semi-structured data to the fold which is used to convert and accommodate heterogeneous health data.

7.1.4. Hive

Apache Hive[16] is data warehouse software built over the top of Apache Hadoop. It facilitates data summarization, query writing and processing, and data analysis. Hive helps us easily write queries for getting the results. Hive is a support architecture which supports the structured query language (SQL). It allows the programmers to write queries similar to SQL thus allowing, people to use the Big Data architecture who are unfamiliar with it (Raghupathi & Raghupathi, 2014). Hive helps in maintaining the backward compatibility with the older systems, which worked on SQL, thus allowing for people trained SQL to directly work in the Big Data architecture.

7.1.5. Zookeeper

Apache Zookeeper[17] is a centralized service for distributed systems, it provides maintaining configuration information, registry naming, distributed synchronization and group services. Zookeeper allows synchronization across the resource cluster i.e. it maintains some control regarding which process should use which resource. It helps in parallel computing, by providing centralized resource management services (Raghupathi & Raghupathi, 2014).

7.1.6. HBase

HBase[18] is a Hadoop database which works on non-SQL approach and is easily able to hold all types of data. It is architecturally placed above the HDFS (Hadoop Distributed File System) (Raghupathi & Raghupathi, 2014). It helps in assimilating heterogeneous data and allows reading and writing, randomly in Big Data.

7.1.7. Cassandra

Cassandra[19] is a distributed database system which helps in handling large quantities of data spread across various utility servers (Raghupathi & Raghupathi, 2014). This database helps in health data analysis by allowing us to easily access the data stored on various servers.

7.1.8. Oozie

Oozie[20] is basically used to streamline workflow among tasks (Raghupathi & Raghupathi, 2014). It works by creating synchs among the various parties/processes involved in the system so that conflict does not occur and the workflow moves smoothly.

7.1.9. Mahout

Mahout[21] provides machine learning applications that are helpful in the predictive analysis of healthcare Big Data (Raghupathi & Raghupathi, 2014). It helps in creating applications such that they can learn from past data and able to correlate and analyse similar situations which give results faster, thus making the system more efficient.

7.1.10. Avro

Avro ("Apache Avro™ 1.8.2 Documentation", 2017) provides data serialization services (Raghupathi & Raghupathi, 2014). In healthcare industries, serialization of patients' medical history helps in diagnosing the disease he/she is suffering from or may suffer (observing some pre-symptoms).

7.1.11. Lucene

Apache Lucene[22] is one of the most widely used frameworks for information retrieval. It offers efficient text searching and analysis. It can be used to explore the EMR to satisfy patient-related queries. It can effectively be used in the diagnosis and analysing patient's reports (Raghupathi & Raghupathi, 2014). Also, Lucene can help in analysing the patient's handwriting which can determine his/her psychological state.

7.2. Architectures

7.2.1. Cloud Computing

Cloud architecture provides the most promising architecture for the health industry to switch over to the Big Data platform. Cloud allows the trouble-free and uncomplicated way to collect data in one place and then distribute it among its specialized facilities for further processing thus allowing for a secure way of processing and analysing the data (Pramanik et al., 2018). The cloud architecture provides a centralized data repository from which data is readily accessible for the Big Data analytics. The security of these structures is dependable; also, they allow the usage of the redundancy for reliability. The redundancy in the cloud architecture allows the system to have a protection against possible crashes.

7.2.2. In-Memory Computing

This is a fairly new architecture which provides real-time analysis (Mian et al., 2014). This architecture involves in-memory SQL databases. The basic advantage of in-memory computing is that it is much faster than the traditional systems as disk access time is removed from the equation. This type of architecture

is suitable where real-time analysis of data is required (Pramanik & Choudhury, 2018). This architecture can be used even with non-SQL databases. The only drawback of it is that if the system crashes for any reason recovery is near impossible.

7.3. Infrastructure for Big Data in Healthcare

7.3.1. Cisco

Cisco major focus of research is IoT (Internet of Things). For such purpose, Cisco is also working on the processing architecture of IoT data called Fog (Cisco, 2015) (Pramanik et al. 2018). The major application of Fog and IoT is in the healthcare sector. The use of IoT has enabled the health industry, to create a right strategy to deliver the proper healthcare to the patients. Physicians are now able to use the clouds and all the computing powers to securely know about their patient's wellbeing as well as are able to diagnose the disease at an early stage even if the patient is at some distance away from them.

The mobility is the key here which allows the physicians, patients and the administration to be on the same page and provide the best medical care available to the patient.

7.3.2. Watson Health

Watson Health[23] is a complete package developed by IBM, which helps in all aspects of health. It has artificial intelligence as well as machine learning capabilities that help in providing effective diagnosis and medication of the diseases and reduces the job of the hospital staff and people in patient care. Watson can learn about the patient's medical history and is able to provide recommendations to the physician of all the possible new drugs or techniques available in the market, thus saving the time of the physicians to go through all the literature.

7.3.3. Philips HealthSuite

HealthSuite ("About HealthSuite", 2018) is an open cloud-based digital platform designed for the continuous health and personalized care of the user. The suite contains the power of analysing, sharing and orchestrating healthcare services. The analysis part utilizes the machine learning algorithms and various predictive analysis techniques. The sharing features are basically the interoperability of the platform form multiple devices. The orchestrating basically implement the workflow synch, communication like task etc. ("A Cloud-based Platform: Purpose-built for Healthcare", 2018).

8. CONCLUSION

Big Data has influenced almost all the industries in recent years. Healthcare is also no exception. In fact, the healthcare industry is the largest producer of the digital data. Big Data technologies have opened up new opportunities in healthcare. It not only has brought the advantages to the patients but also to the healthcare units and hospitals. To maximise the benefits, careful consideration should be given in adopting the right Big Data tools and the underlying architecture. Though Big Data technology has a significant impact on modern healthcare it needs to progress further to realise its fullest potential. The

traditional healthcare system is still lacking in becoming accustomed to the 'big' change. Nevertheless, Big Data has set healthcare industry on the righteous trajectory of rapid transformation and that will surely bring startling benefits to the mankind.

REFERENCES

A Cloud-based Platform: Purpose-built for Healthcare. (2018). Retrieved July 24, 2017, from http://www.usa.philips.com/healthcare/innovation/about-health-suite

About HealthSuite. (2018). Retrieved April 29, 2018, from https://www.philips.co.in/healthcare/innovation/about-health-suite

Apache Avro™ 1.8.2 Documentation. (2017, August 2). Retrieved April 29, 2018, from https://avro.apache.org/docs/current/

7 . Big Data Use Cases for Healthcare. (2016, October 8). Retrieved July 24, 2017, from http://www.ingrammicroadvisor.com/data-center/7-big-data-use-cases-for-healthcare

Bresnick, J. (2015, September 18). *Healthcare Big Data Analytics Suffers from "Fragmented" Approach.* Retrieved May 26, 2017, from http://healthitanalytics.com/news/healthcare-big-data-analytics-suffers-from-fragmented-approach

Bresnick, J. (2015, January 13). *How to Select a Big Data Analytics, Business Intelligence Vendor.* Retrieved July 14, 2017, from https://healthitanalytics.com/news/how-to-select-a-big-data-analytics-business-intelligence-vendor

Bresnick, J. (2016a). *How to Choose the Right Healthcare Big Data Analytics Tools.* Retrieved July 14, 2017, from https://healthitanalytics.com/features/how-to-choose-the-right-healthcare-big-data-analytics-tools

Bresnick, J. (2016b). *The Difference Between Big Data and Smart Data in Healthcare.* Retrieved July 14, 2017, from https://healthitanalytics.com/features/the-difference-between-big-data-and-smart-data-in-healthcare

Bresnick, J. (2016c). *The Role of Healthcare Data Governance in Big Data Analytics.* Retrieved July 14, 2017, from https://healthitanalytics.com/features/the-role-of-healthcare-data-governance-in-big-data-analytics

Bresnick, J. (2017, June 5). *Understanding the Many V's of Healthcare Big Data Analytics.* Retrieved July 20, 2018, from Health IT Analytics: https://healthitanalytics.com/news/understanding-the-many-vs-of-healthcare-big-data-analytics

Brown, N. (2015, September 11). *Healthcare Data Growth: An Exponential Problem.* Retrieved August 8, 2017, from http://www.nextech.com/blog/healthcare-data-growth-an-exponential-problem

Cambridge, Mass & Rochester, Minn. (2013, January 15). *Optum, Mayo Clinic Partner to Launch Optum Labs: An Open, Collaborative Research and Innovation Facility Focused on Better Care for Patients.* Retrieved April 20, 2018, from https://www.optum.com/about/news/optum-labs.html

Castle, B. L., & Szymanski, G. (2008). Supply Chain Management on Clinical Units. In eBusiness in Healthcare (pp. 197-217). London: Springe.

Cisco. (2015). *Fog Computing and the Internet of Things: Extend.* Cisco. Retrieved from https://www.cisco.com/c/dam/en_us/solutions/trends/iot/docs/computing-overview.pdf

Crapo, J. (2017). *Big Data in Healthcare: Separating The Hype From The Reality.* Retrieved July 27, 2017, from https://www.healthcatalyst.com/healthcare-big-data-realities

Disch, W. (2016, August 11). *How to Use Big Data to Improve Patient Engagement*. Retrieved 8 2017, from http://data-informed.com/how-to-use-big-data-to-improve-patient-engagement/

Eastwood, B. (2013, April 23). *6 Big Data Analytics Use Cases for Healthcare IT.* Retrieved July 24, 2017, from http://www.cio.com/article/2386531/healthcare/healthcare-6-big-data-analytics-use-cases-for-healthcare-it.html

Epidomiology and Genomics Research Program. (2018, April 5). Retrieved April 29, 2018, from https://epi.grants.cancer.gov/pharm/pharmacoepi_db/healthcore.html

Evolve, I. P. (2017). *Email Vulnerability in Healthcare.* Retrieved from http://www.evolveip.net/lp/email-vulnerability-healthcare

Flatiron. (2017). *Community Oncology*. Retrieved 8 2017, from https://flatiron.com/community-oncology/

Groves, P., Kayyali, B., Knott, D., & Kuiken, S. V. (2013). *The 'big data' revolution in healthcare*. McKinsey & Company.

HealtIT.gov. (2016, November 4). *Electronic Prescribing of Controlled Substances (EPCS)*. Retrieved 8 2017, from https://www.healthit.gov/opioids/epcs

Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2013). *Hadoop Pig and Pig Latin for Big Data*. Wiley.

IDC. (2014). *The Digital Universe Driving Data Growth in Healthcare.* IDC. Retrieved August 7, 2017, from http://www.emc.com/analyst-report/digital-universe-healthcare-vertical-report-ar.pdf

Jawadekar, D. M. (2016, August 4). *Big Data and Pharmaceutical Drug Discovery*. Retrieved August 2017, from https://pharma.elsevier.com/pharma-rd/big-data-and-pharmaceutical-drug-discovery/

Lebied, M. (2017, May 24). *9 Examples of Big Data Analytics in Healthcare That Can Save People.* Retrieved July 26, 2017, from http://www.datapine.com/blog/big-data-examples-in-healthcare/

Leventhal, R. (2014, January 10). *Industry expert: "Big data is closer than it appears"*. Retrieved August 2017, from https://www.healthcare-informatics.com/article/how-healthcare-organizations-can-turn-big-data-smart-data

Leventhal, R. (2014, December 4). *Report: Healthcare Data is Growing Exponentially, Needs Protection.* Retrieved August 7, 2017, from https://www.healthcare-informatics.com/news-item/report-healthcare-data-growing-exponentially-needs-protection

marketsandmarkets.com. (2016). *Healthcare Analytics/Medical Analytics Market.* Retrieved August 7, 2017, from http://www.marketsandmarkets.com/Market-Reports/healthcare-data-analytics-market-905.html

Marr, B. (2016, February 16). *How Big Data Is Transforming Medicine*. Retrieved August 2017, from https://www.forbes.com/sites/bernardmarr/2016/02/16/how-big-data-is-transforming-medicine/#244b70277ddc

McDonald, C. (2017, February 27). *5 Big Data Production Examples in Healthcare.* Retrieved July 24, 2017, from https://mapr.com/blog/5-big-data-production-examples-healthcare/

Mian, M., Teredesai, A., Hazel, D., Pokuri, S., & Uppala, K. (2014). In-Memory Analysis for Healthcare Big Data. *IEEE International Congress on Big Data*.

Nambiar, R., Sethi, A., Bhardwaj, R., & Vargheese, R. (2013). A Look at Challenges and Opportunities of Big Data Analytics in Healthcare. *IEEE International Conference on Big Data.* 10.1109/BigData.2013.6691753

Orcutt, M. (2016, June 29). *The Rocket Fuel for Biden's "Cancer Moonshot"? Big Data.* Retrieved April 29, 2018, from https://www.technologyreview.com/s/601784/white-house-cancer-moonshot-data/

Patil, H. K., & Seshadri, R. (2014). Big data security and privacy issues in healthcare. *IEEE International Congress on Big Data.*

Pramanik, P. K. D., & Choudhury, P. (2018). IoT Data Processing: The Different Archetypes and their Security & Privacy Assessments. In *Internet of Things (IoT) Security: Fundamentals, Techniques and Applications*. River Publishers. doi:10.4018/978-1-5225-4044-1.ch007

Pramanik, P. K. D., Pal, S., Brahmachari, A., & Choudhury, P. (2018). Processing IoT Data: From Cloud to Fog. It's Time to be Down-to-Earth. In Applications of Security, Mobile, Analytic and Cloud (SMAC) Technologies for Effective Information Processing and Management. pp. 124-148. IGI Global. doi:10.4018/978-1-5225-4044-1.ch007

Pramanik, P. K. D., Upadhyay, B., Pal, S., & Pal, T. (2018). Internet of Things, Smart Sensors, and Pervasive Systems: Enabling the Connected and Pervasive Health Care. In *Healthcare Data Analytics and Management*. Elsevier.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems, 2*(3).

Ramsey, L. (2015, December 23). *I tried 23andMe's new genetics test - and now I know why the company caused such a stir*. Retrieved August 2017, from http://www.businessinsider.in/I-tried-23andMes-new-genetics-test-and-now-I-know-why-the-company-caused-such-a-stir/articleshow/50302664.cms

Revolutionizing the Healthcare Industry with Big Data, Analytics and Visualization. (2015). Retrieved August 2, 2018, from https://www.einfochips.com/whitepaper/Revolutionizing-the-Healthcare-Industry-with-Big-Data-Analytics-and-Visualization.pdf

Sarkar, B. K. (2017). Big data for secure healthcare system: A conceptual design. *Complex Intelligent Systems*, *3*(2), 133–151. doi:10.100740747-017-0040-1

Seven Big Data Examples That Have Improved Healthcare Operations. (2016, April 19). Retrieved July 24, 2017, from Ingram Micro Advisor: http://www.ingrammicroadvisor.com/data-center/seven-big-data-examples-that-have-improved-healthcare-operations

Shah, S. (2016, February 18). *Why patient engagement is so challenging to achieve*. Retrieved August 2, 2018, from http://www.ibmbigdatahub.com/blog/why-patient-engagement-so-challenging-achieve

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop Distributed File System. In *IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST).* IEEE.

Srinivasan, U., & Arunasalam, B. (2013). *Leveraging Big Data Analytics to Reduce Healthcare Costs*. IEEE Computer Society.

TF7 Healthcare Subgroup. (2016, December 21). *Big Data Technologies in Healthcare: Needs, opportunities and challenges.* Retrieved July 24, 2017, from http://www.bdva.eu/sites/default/files/Big%20Data%20Technologies%20in%20Healthcare.pdf

United Healthcare uses Hadoop to Detect Health Care Fraud, Waste and Abuse. (2018). Retrieved April 20, 2018, from https://mapr.com/customers/unitedhealthcare/

Uzsoy, R. (2005). *Supply-Chain Management and Health Care Delivery: Pursuing a System-Level Understanding*. Retrieved 2017, from https://www.ncbi.nlm.nih.gov/books/NBK22867/

ENDNOTES

1 https://flunearyou.org/#/
2 https://twitter.com/search?q=%23Germtracker
3 http://www.gocap.me/
4 https://www.propellerhealth.com/
5 https://www.23andme.com/
6 http://www.gnshealthcare.com/
7 https://www.aetna.com/
8 https://www.healthcore.com/
9 https://www.evolenthealth.com/valence-health
10 https://www.uhc.com/
11 https://www.optumlabs.com/
12 https://www.liaison.com/
13 https://www.novartis.com/tags/next-generation-sequencing
14 http://hadoop.apache.org/
15 http://pig.apache.org/
16 https://hive.apache.org/
17 https://zookeeper.apache.org/
18 https://hbase.apache.org/
19 http://cassandra.apache.org/
20 http://oozie.apache.org/
21 https://mahout.apache.org/
22 https://lucene.apache.org/
23 https://www.ibm.com/watson/health/