Big Data Analytics and its Applications in Healthcare: A Review

Rahul M Bharadwaj, Dr. Surbhi Agrwal

[1]Computer Science Undergraduate Student,[2]Associate Professor
[1]Department of Computer Science and Engineering,
[1]RV Institute of Technology and Management, Bengaluru, India

Abstract - The amount of heterogeneous data generated annually in the modern world has increased dramatically. The generation of data will continue to grow enormously in the years to come thanks to discoveries and developments in numerous fields around the world. Thus, it becomes crucial to effectively store, manage, and use data. The Covid-19 pandemic, which caught everyone off guard, showed our healthcare systems' flaws and got us thinking more about how to combine healthcare with various contemporary technologies. One such tool that has recently gained importance, particularly in the healthcare industry, is big data analytics. The amount of data that will be collected in the near future will expand enormously as a result of numerous countries spending extensively in healthcare infrastructure and technologies in response to the lessons learned from this pandemic. In this review paper, we will attempt to gain useful insights into the necessity of managing and converting unstructured and semi-structured data into structured data, the impact of big data analytics on the healthcare industry, the various tools and handling methods available to deal with big data, applications of big data analytics in the healthcare industry, and finally our understanding of the electronic health record (EHR) system in India.

Index Terms – Hadoop Distributed File System (HDFS), Structured Query Language (SQL), Electronic Health Records (EHR), Relational Database Management System (RDMS), Resource manager (RM), Node Manager (NM), Primary Health Care (PHC), Secondary Health Care (SHC), International Classification of Diseases (ICD 11), Tertiary Health Care (THC), Healthcare Information Technology (HIT), Picture Archiving Communication System (PACS), Ministry of Health and Family Welfare (MoHFW), Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), Logical Observation Identifiers Names and Codes (LOINC), National Drug Code (NDC).

I. Introduction

One of the most vital industries in any community is healthcare. But the majority of nations around the world have long disregarded this important sector. Due to this carelessness, the modern civilization had a very difficult time preparing for an unparalleled catastrophe that brought the entire world to a standstill. In addition to helping us understand the value of maintaining the health of our healthcare systems, this battle against an invisible foe also provided us with the opportunity to learn about some very significant technologies that were already playing significant roles in our lives without our knowledge. A lot of developments in the healthcare industry were also made possible by this sudden visit from an unexpected visitor.

The development of computers and related technologies over the past two decades has fundamentally altered how we gather, examine, and use data. Open information is ushering in a new era in healthcare. The amount of healthcare data being generated is unfathomably large due to the population's rapid growth and considerable advancements in medical treatment techniques. The healthcare industry is being revolutionized by the sheer volume and accessibility of data. Both organized and unstructured data are present in this enormous volume of created heterogeneous data. Since the majority of the data created in the healthcare industry is unstructured, organizing this data to produce relevant outcomes becomes equally difficult. As a result, both structured and unstructured data are extracted, analyzed, processed, and managed using a variety of methods. Now that everyone understands how crucial it is to combine technology with healthcare, big data analytics and all of its tools will guarantee that our global healthcare systems are better prepared to handle any issues that may arise in the future and call into question the continued survival of mankind.

II. WHAT ARE BIG DATA AND BIG DATA ANALYTICS?

Big data refers to vast quantities of heterogeneous data that have been produced in both organized and unstructured formats by several organizations throughout the globe. Effective tools and strategies are required to extract, analyze, process, and manage the vast amounts of data generated. Big data analytics may be summed up as the systematic processing of enormously huge and varied data sets (also known as heterogeneous data sets), which include both structured and unstructured data from various sources all over the world.

In big data analytics, structured data refers to the data that has a pre-defined format. Such data having a particular format and size can be processed easily as compared to unstructured data. Structured data can include a patient's height, weight, and blood pressure in numerical form, as well as blood type information and details about the phases of a disease's diagnosis in categorical form. They are well-organized and simple to comprehend. Unstructured data is a collection of several data kinds in a wide range of forms. They do not have a pre-defined size or structure and hence makes it nearly impossible to organize them using pre-defined structure. It is this type of data that majorly constitutes the large amounts of data generated in the healthcare sector. One of the most basic examples of unstructured data is the prescriptions written by physicians. They do not follow any particular format. Unstructured text is frequently used to record clinical information, such as patient symptoms during a doctor appointment. There is a chance that a doctor's letter describing medical symptoms contains misspellings and acronyms. Such errors cannot be processed without human intervention. Hence, the application of various tools of big data analytics becomes necessary to extract useful results from the huge volumes of varied types of data generated in this sector.

WHY DATA ANALYTICS? WHY IS IT IMPORTANT TO CONVERT UNSTRUCTURED DATA INTO STRUCTURED DATA?

As was previously mentioned, the majority of data produced in the healthcare industry is unstructured. Without organizing the findings in a certain manner that can be processed, it becomes practically difficult to extract any kinds of outcomes due to their complexity and variety. The amount of healthcare data generated by the year 2020 was predicted in one of the surveys done in 2012 to reach 26,000 peta bytes. This demonstrates the requirement for efficient and effective data organization in order to generate useful forecasts. We've attempted to outline a few benefits of turning unstructured data into structured data to help you understand the relevance of this process:

1) The amount of time needed for manual expert assessment will decrease. Medical practitioners can spend less time reading and analyzing free texts and electronic health records.
2) Large amounts of organized data will be available to reviewers in Food and Drug administration committees which enables them to make decisions accordingly and effectively in a short duration.
3) Healthcare enthusiasts and practitioners can keep themselves updated about the latest developments in the medical field due to the availability of large amount of accurate information.
4) Being able to process huge volumes of organized, structured clinical data will play a vital role in implementing algorithms to accurately predict various outcomes like the possibility of a group of population getting a disease due some common ailments or features exclusive to that group. Similarly, based on the analysis of large amount of data over a period of time, hospitals can predict the number of patients it may have to treat over the next few days or weeks. Getting to know about the possibilities will be very crucial in saving many lives if arrangements are made beforehand.

Hence, it becomes extremely necessary to have the data in an organized format to get the best results.

THREE MAJOR AVAILABLE FRAMEWORKS IN DATA ANALYTICS (HEALTHCARE)

The largest hurdle in big data analytics is undoubtedly the processing of different types of data. Given that the majority of the data created in the healthcare industry is unstructured, this becomes the industry's biggest concern. As a result, there is always a need for new and efficient methods and approaches to merge disparate data and provide useful findings. In light of this, we attempt to comprehend the three main and widely-used frameworks in this part, which can be used to analyze healthcare data. [1]:

1) *Predictive analysis in healthcare:*

Predictive analysis is crucial for saving lives, as was covered in the preceding section of this essay. Through the analysis of massive amounts of data created over time, predictive analytics uses statistical techniques like data mining and machine learning to forecast the future. Simply put, data mining is the process of applying a variety of approaches to extract relevant patterns and information from a sizable collection of disorganized data. Doctors can use this prediction data to assist in making crucial decisions about patient treatment.

2) *Machine learning:*

Another crucial framework that is quite similar to data mining is machine learning. Artificial intelligence is often utilized in machine learning, where data is used to increase programme comprehension. Simply said, it enables a system or programme to automatically learn from events and develop without explicit guidance or control. Here, allowing the system to learn autonomously without human input is the major goal. Due to the extremely low likelihood of human error, this increases the effectiveness of the entire predictive and pattern-based decision-making process.

3) *Electronic Health records (Descriptive analytics):*

Electronic health records or EHRs are all about maintaining digitized versions of a patient's health records which includes medical history, symptoms, lab test results, etc. This information is made available to both private and public healthcare providers. These records can be easily updated and modified as and when it's required. This drastically reduces the amount of paper work and duplication of records. We will try to learn more about Electronic health records in the later section of this paper where we have discussed in detail about the proposed Electronic Heath record scheme in India.

III. THE SEVEN V'S OF BIG DATA ANALYTICS

Before knowing the techniques of organizing the data to obtain useful results, we try to understand the seven basic attributes of big data analytics [2][7]. Fig 1 shows the seven V's of big data analytics:

1) Volume:

The term "big data analytics" is self-explanatory. Big data is the term used to describe vast quantities of data, both structured and unstructured. Data organization becomes crucial in big data analytics in order to process the data and produce results. Because of things like cheaper data storage and processing designs, the amount of data collected in the healthcare industry has recently been growing tremendously. Big data analytics implementation and use in the healthcare industry are now required rather than optional.

2) Velocity:

In big data analytics, "velocity" essentially refers to the rate or speed of data collection. As was already said, the amount of data collected in the healthcare industry is increasing exponentially. Given the high rate of data generation, it is crucial to gather data at a similar or faster rate in order to support prompt decision-making based on output. Big data analytics in the healthcare industry are, to put it briefly, a race against the clock.

3) Variety:

In big data analytics, variety refers to the various forms of data produced in the healthcare industry. The majority of the data produced in the healthcare industry is unstructured. Such data needs to be organized and formatted in order to produce useful analysis. Structured data mostly consists of information on things like blood type, disease stages, etc. Emails, pictures, prescriptions, and other unorganized data formats are examples of unstructured or semi-structured data. We shall comprehend the numerous tools utilized to organize unstructured data in following sections of this work.

4) Veracity:

Veracity basically refers to the authenticity of the data generated and processed to get an output. Since different types of data are generated from various sources, it is very important to check the credibility of these sources. In healthcare sector, the decisions taken based on the output of an analysis should be error free since it deals with the lives of thousands of patients. Keeping a check on the veracity of the data is and will continue to be a major challenge for big data analytics in healthcare sector.

5) Visualization:

Massive volumes of structured, unstructured, and semi-structured data are produced and processed as part of big data. It would be convenient, effective and efficient to analyze and represent the results in simple, understandable and visually appealing formats like charts and graphs instead of traditional and conventional formats i.e. spreadsheets and reports. Hence, visualization of huge amounts of data generated and analyzed becomes a very important aspect of big data analytics.

6) Variability:

Variability in big data analytics refers to the data that keeps changing constantly. It is crucial to have the meanings of data constant. The homogeneity of data can be significantly impacted by the frequent changing of data meanings.

7) Value:

The ultimate goal of analyzing huge amounts of data is to obtain meaningful values. This involves efforts, time and usage of resources on the above mentioned V's of big data analytics. If the data is analyzed and processed correctly, the desired results can be obtained.

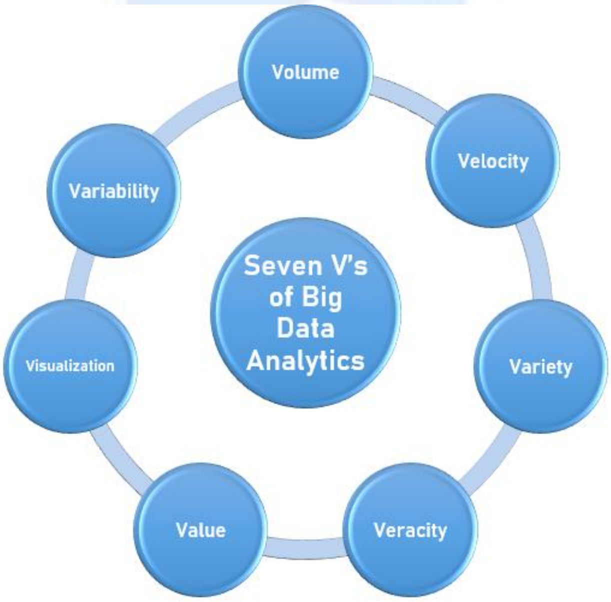

Fig 1. Seven V's of Big Data Analytics

IV. General Phases in the Life Cycle of Big Data Analytics

This section of the paper includes brief description of the eight important phases of the general life cycle of big data analytics [3][9]:

1) Business Evaluation:

This is the commencement phase which includes the definition of the reasons and goals behind conducting an analysis. This is termed as business case evaluation phase.

2) Identification of data and sources of data:

This phase primarily involves identification of various sources of data. The data obtained from these sources can be structured, unstructured or semi-structured.

3) Data Filtering:

The data gathered from numerous sources in the previous step are filtered according to the needs, and the undesirable, corrupt data is removed.

4) Data Extraction:

Though the data is filtered in the previous stage, it might not be entirely compatible with the tools selected for data analysis. In this phase, data that isn't compatible with the selected tools is extracted and then transformed into a compatible format.

5) Data Aggregation:

This is a crucial phase, data having the same fields across different datasets are integrated together. The data is expressed in a summarized form.

6) Data Analysis:

This phase constitutes the core part of performing analysis on the data gathered. Using various statistical and analytical tools selected, analysis is performed on the data to obtain useful information.

7) Visualization of Data:

To understand the obtained results better, they can be represented in a visually appealing and understandable format like histograms, bar charts, heat maps, pie charts, etc. This phase includes usage of various tools like Tableau, Garfana, and Power BI to produce graphic visualization of the results.

8) Final Result Analysis:

This terminal phase of the lifecycle includes publication of the final analysis results to the concerned business stakeholders who can make further decisions based on these results.

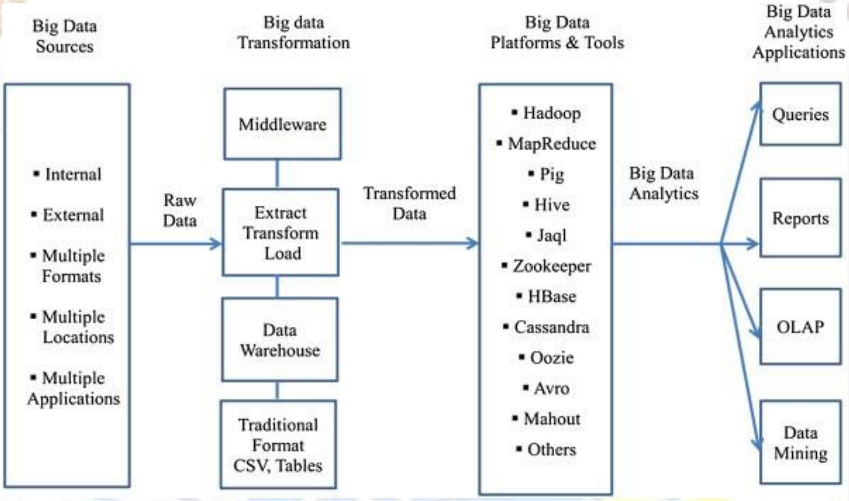

Fig 2. Lifecycle of Big Data Analytics

V. HADOOP TOOLS AND TECHNIQUES

As we have previously noted, the healthcare industry generates a substantial amount of unstructured data. Such data collections are captured, arranged, and analysed using the Hadoop platform.

1) Apache Hadoop:

A collection of open-source software tools popularly called and known as Apache Hadoop is used for the processing, storing, and distributed computation of enormous amounts of data.. The two core components of Hadoop are:

a) Hadoop distributed file system (HDFS)
b) MapReduce

a) *Hadoop distributed file system (HDFS):*

Hadoop distributed file system is a primary data storage system. It can store data both in structured and unstructured form. Within HDFS is the HBase which is a scalable data base exclusively dedicated to structured data storage. HDFS basically works on master-slave architecture. The name node and the data node make up HDFS's architecture. Here, the name node serves as the master and oversees the upkeep of the data nodes, while the data node serves as the slave. The file system's name space is mostly maintained by the name node. It consists of meta data (data about a data) of all the files stored like the location and size of the file in HDFS. Data nodes or the slave nodes are the places where the data is actually stored and handled. Clients here will be the HDFS users. To understand HDFS in a better way, if the index page of a notebook containing the serial numbers, concept titles and the page numbers is considered to be the name node, then the following pages in the notebook containing the actual content of the chapters will become the data nodes. Here, the reader of the book becomes the client.

b) *MapReduce:*

MapReduce is a standard programming framework/ model used for processing and analyzing huge amounts of data. It breaks tasks into sub tasks and efficiently analyze large data sets. MapReduce as the name suggests, majorly performs two tasks:

- Mapping which deals with splitting and mapping of data.
- Reducing which deals with shuffling and reducing the data.

Also, operation of MapReduce architecture is split into three major parts:

1) *Client:* It is in charge of uploading jobs to the job tracker in the form of JAR files, which are Java archives that combine multiple files into one.

2) *Job tracker*: It is responsible for maintaining all the jobs that are executed in MapReduce. Thus, it acts as a master service

3) *Task tracker:* It is in charge of carrying out the tasks given to it by the Job tracker. It serves as a slave service as a result.

In a word, MapReduce is the architecture used for efficiently analyzing enormous amounts of structured and unstructured data, with HDFS serving as the storage location for both types of data.

2) Hadoop tools

In this section, we have tried to understand some of the many Hadoop tools available briefly as depicted in Fig 3 [4]:

Apache Hive: Large data files stored in the HDFS can be managed, read, and written using Apache Hive, an open-source data warehouse programme. Hive resembles typical database code written in SQL (Structured Query Language) in appearance. The vast amount of unstructured data that may be kept in tables with varied rows and columns in the metastore, just like we see in databases of structured query language, makes MapReduce programming with this open-source software simple. However, Apache Hive is less suitable for software that requires quick reaction times.

Apache Pig: Similar to Apache Hive, Apache Pig is open-source software used for big data analysis. Pig Latin is a high-level language used by this software to create data analysis programmes. Pig Latin allows the users to develop their own user defined functions for processing data. This language is very similar to SQL and a programmer well versed in SQL can easily learn this language without having to learn and type Java codes which are relatively complex. The presence of a component called Pig Engine makes analysis of data simpler as it converts the Pig Latin scripts developed by the user into Map and reduce tasks. Hence, it is not mandatory for a programmer to learn complex programming languages to work with Apache Pig

Apache Oozie: Apache Oozie is a complex technique used to run tight system designs or complex systems where there is data dependence between a number of interconnected stations. The biggest advantage of Oozie is that it is integrated with Hadoop jobs like Pig and Hive and system specific jobs like Java. Oozie mainly consists of two components: work flow engine and coordinator engine. Work flow engine is responsible for storing and running work flows containing Hadoop jobs like Pig and Hive. Coordinator engine runs work flow jobs based on schedules and availability of data. Action node and control flow node are two of the nodes that make up Oozie process. Action nodes indicate workflow operations like running map reduce or Pig/Hive jobs, putting files into HDFS, etc. While the control flow node is primarily in charge of managing the execution of the process. It also includes start nodes, finish nodes, and error nodes that, respectively, mark the beginning of a work flow job, the conclusion of a job, and the incidence of an error.

Apache Avro: Apache Avro is an open-source project that provides data exchange and data serialization services. It makes it easier for programmes built in many languages to communicate huge data. One of the most effective Hadoop tools is Apache Pig because it can exchange data from compiled languages like C with scripting languages like Python. Data is serialized and placed in files or messages. Avro is incredibly compact and effective since it stores both the data description and the actual data in a single binary file. Additionally, it has markers that can be used to divide big data sets into smaller ones so that they can be used in the MapReduce process.

Apache Zookeeper: Apache Zookeeper, as the name itself suggests, is an open-source centralized service enables highly reliable distributed coordination. It enables synchronization between many servers in a large Hadoop cluster. With the help of Zookeeper, cross node synchronization is also possible. Numerous Zookeeper servers are used to support large Hadoop clusters, with a master server synchronizing the top-level servers. Also ensuring application dependability is Zookeeper. To resume the duties in the event of a mistake or issue with one of the application masters, a new application master is established.

Apache Yarn: The Hadoop ecosystem uses Apache Yarn (Yet Another Negotiator) as its resource management layer. Resource manager (RM) and Node Manager (NM) are the two main parts of yarn. A Hadoop cluster, which consists of numerous hosts (computers or nodes), is managed by a resource manager. The management of resources present on an independent host falls more specifically under the purview of the node manager. Together, these two parts are in charge of managing memory and scheduling tasks.

Apache Sqoop: It is a powerful open-source tool used for extracting data from the relational database management system (RDMS) and storing them in HDFS for processing and analysis. This is done with the help of MapReduce or other Hadoop tools like Hive and Pig. The data in HDFS can be easily used by Hadoop applications.

Apache Flume: Similar to Sqoop, Apache Flume is an open-source technology that is used to transport data into HDFS. But vast amounts of data from numerous independent computers are gathered in Flume. It is frequently used to store data in HDFS from a variety of data sources, including emails and social media.

Apache HBase: Java-based Apache HBase is a distributed NoSQL database. HDFS, the Hadoop distributed file system, serves as its foundation. Due to the fact that it is column oriented, it is horizontally scalable. It handles extremely huge data sets quite well and enables the cluster to grow by adding more nodes. It is particularly optimized for fast, low-latency data access. Because it duplicates the data across a large number of cluster nodes, it is also fault tolerant. It is difficult to set up and manage, though. Contrary to relational SQL, its language HBase Shell is feature-poor, making it challenging to carry out complicated operations.

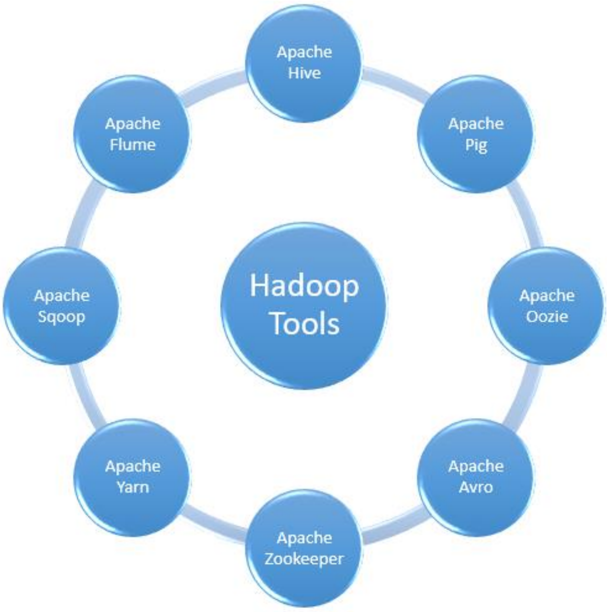

Fig 3. Hadoop Tools

VI. APPLICATIONS OF BIG DATA ANALYTICS IN HEALTHCARE

1) *Treatment of cancer and genomics:*

The human genome consists of 3 million base pairs, which can generate a lot of data which needs to be organized and analyzed. Oncologists have determined that in order to recognize patterns of cancer, the genetic make-up of the patient needs to be ascertained, which is specific to that patient and hence, a specific treatment can be administered to the particular patient. The mapping of the three million base pairs can be done though MapReduce, a subsidiary of Hadoop technology.

2) *Monitoring of patient's vitals:*

Hospitals all over the world use a variety of Hadoop-based file components in the Hadoop Distributed File System (HDFS), such as the HBase, Hive, and Flume frameworks, to convert the enormous amount of unstructured data produced by sensors that take patient vital signs, such as heart rate, blood pressure, blood sugar level, and respiratory rate. This is how hospitals all over the world connect their work output through big data technologies.

3) *Hospital network:*

A number of hospitals use the Hadoop ecosystem, a NoSQL database, to collect and handle their enormous volumes of real-time data from various sources linked to patient care, finances, and payroll, which enables them to identify high-risk patients while also lowering daily costs.

4) *Healthcare intelligence:*

Hospitals and insurance companies process large datasets related to medications, diseases, symptoms, opinions, geographic regions, and other factors using healthcare intelligence applications built using the Hadoop ecosystem, such as Pig, Hive, and MapReduce technologies, in order to extract meaningful information (for example, desired age) for insurance companies.

5) *Detection and Prevention of Frauds:*

Insurance firms have to use a variety of techniques in the early days of big data analytics to spot fraud activities in an effort to stop medical insurance fraud. Businesses employ Hadoop apps based on prediction models to find fraudsters using information from their past health claims, voice recordings, earnings, and demographics.

VII. CASE STUDY: USAGE OF BIG DATA ANALYTICS IN ELECTRONIC HEALTH RECORDS (INDIA)

1) *Study of the existing healthcare information system*

The collection of data in most healthcare systems in India is currently paper based and the medical information is stored in the form of records and reports. Whenever a patient visits a healthcare facility for diagnosis and treatment, large amounts of data must be collected and processed and stored in a manner that doesn't compromise the privacy of the patient. Since these records are physical, it's inconvenient for the patient to carry it with them to every hospital they visit. Due to the delay in receiving medical records, this paper-based documentation method also causes lengthy registration lines and treatment wait times. The data must first be physically captured, which takes time, and it must then be shared in real time, which is challenging. [5].

Indian hospitals come in all shapes and sizes, and both the federal and state governments operate them. This country has a diverse healthcare system. It can roughly be divided into the primary health care (PHC), secondary health care (SHC), and tertiary health care (THC) categories of care. PHC is the patient's first point of contact with the healthcare system. Patients are referred to secondary care levels, and so on, if they need advanced or specialized care. The registration book, the examination book, and the treatment books are all maintained at the PHC level and are primarily paper-based. The patient's five-year records are kept at the medical record department's front desk for quicker access, while the remaining data are kept in a repository for later access. As a result, the employees at the healthcare facility must go through an extremely time-consuming and difficult data collection process, which delays patient care.

The lack of high-speed internet connectivity is one of the key obstacles to establishing IT infrastructure at the PHC level, particularly in rural areas. A combination of paper-based and computerized recording systems are used at the secondary and tertiary levels of healthcare. Demographic information, vital signs, patient health histories, diagnosis specifics, prescriptions, allergy information, lab test results, and medical imaging make up a large portion of the paper-based data. Other non-medical data is kept in digital form, including administrative, billing, and registration data. A small number of healthcare facilities have begun using EMR, however because these Hospital Information Systems employ proprietary software, the data cannot be merged with the EMR of another hospital if the patient is sent to that facility. The former hospital still has the patient's records.

Therefore, it is seen that in the existing system in practice, the patient information and care is able to be documented but the transfer of this information to different levels of health facilities is not supported. Considering the existing challenges, the development of a standardized Electronic Health Record (EHR) has been proposed with the objectives of reducing health care professionals' times for information management and retrieval, allowing for greater productivity and performance using the assistance of rich data provided in standard code. The healthcare application should be developed to be user friendly and should have features for easy documentation and report generation.

2) *Development of standardized Electronic Health Records*

Healthcare information technology (HIT) is reshaping the way the healthcare sector functions and has already started to cut waste while assisting in improving patient outcomes. The Electronic Health Record (EHR) is a key element of HIT. Electronic health records are patient records that are stored digitally and contain information on the patient, such as contact details, their medical history, allergies, test results, and treatment plans. [6].

Delivering lifetime clinical care at all times is the main goal of the healthcare information system. Data's constant syntactic and semantic compatibility must be maintained in order to satisfy this criteria. Making patient data accessible to all levels of healthcare professionals in a country like India, where a huge population is dispersed over a vast geographic area, is challenging since the data is not interoperable and meaningful. The standardized syntax and code should be used to maintain data consistency and interoperability.

EHR is a digital repository of patient health data that is accessible to medical professionals via a computer network. It also includes lab test results and diagnosis images. Healthcare IT solutions are largely being provided through cloud computing. As a result, cloud infrastructure can greatly benefit healthcare organizations and may be a great way to meet the nation's need for better healthcare.

The Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), International Classification of Diseases (ICD 11), Logical Observation Identifiers Names and Codes (LOINC), and National Drug Code (NDC) are the standards recommended by the Ministry of Health and Family Welfare (MoHFW) for use in EHRs.

The suggested EHR paradigm links all users of the health information system to a single location for the purpose of storing and exchanging patient health data across an EHR network. All tiers of healthcare practitioners can contribute to the patient's overall health records in this cloud-based system. The HER universal health service provider database, a global cloud data base, processes and stores all the data collected from every healthcare location. Using the EHR online application, healthcare practitioners can securely access this database from a distance, collaborate with other healthcare facilities, and evaluate the patient's medical history.

The collection of health information from the many modules—administrative, medical, nursing, laboratory, radiology, pharmacy, etc.—is necessary to create an integrated system.

The Administrative System Module is the first part of the EHR system, and it maintains information on patient registration details (patient demographic data), admission, discharge, and record transfers from one hospital to another. The nursing module, which makes up the second part of the system, is where the patient's apparent information, such as height, weight, blood pressure, and BMI values, is mostly recorded. Data from laboratories, including microbiology and biochemistry, are included in the third EHR component. The data from the lab are used to make decisions about 60% to 70% of the quality decisions. The fourth module contains information and images relating to radiology, such as X-ray, CT scan, and MRI. Picture Archiving Communication System (PACS), a medical imaging technology, is used to manage digital images.

The clinical documentation module, which collects patient clinical data such diagnosis, treatment, complication, and prescription, is a crucial and significant part of an EHR system. A clinical record in electronic form aids the healthcare professional in verifying the patient's care. The EHR accomplishes its goals by giving the patient's information in a precise, thorough, and concise manner, assisting the subsequent healthcare practitioner in making an excellent choice. Pharmacy is a key component of the EHR system because it stores detailed information about all of the medications a patient has ever taken, including their name, drug code, dose, amount, and any allergic reactions. This information helps the pharmacist provide the patient with safe and effective medication.

3) *Implementation and Validation:*

The Indian healthcare IT market is predicted to expand quickly, with a 15.8% annual growth rate, to reach a value of USD 390.7 billion by 2024. In this regard, quality assessments for the EHR system are necessary. This test is essential to verify the system's compliance, interoperability, features, and performance since the testing and feedback phases of the software development lifecycle must continue until the final product is prepared for a global rollout. Testing web applications is difficult due to the large number of programme entries and exits, the variety of browsers used by end users, and the individuals who use the product. The following is a list of this EHR system's salient characteristics and advantages:

Expedited registration - In contrast to the previous approach, it offers for a simple method of recording patient records. Patients may only register once. Prior to being officially registered in the hospital, patients had to fill out several, time-consuming papers by hand at the hospital. The administration then had to keep all of these records on paper. This issue can be readily solved by the suggested paradigm, greatly improving the efficiency of patient registration.

Digital signature - Medical forms and prescription information for patients must be digitally signed in the healthcare system to guard against unauthorized access, manipulation, and forgery. The key device used to safely and secretly store potentially sensitive patient data is a digital signature. Digital signatures are the message authentication primitives that use two cryptographic keys that can authenticate each other. A PHC doctor may use a digital signature to sign a paper. Data is fed into the hash function (SHA-512) by the signer, which produces an original representation of the data known as a hash. The signature algorithm known as RSA generates the digital signature on the supplied hash value using the hash value and the signer's private keys as inputs. The document is subsequently delivered to the verifier along with the signature. Using a verification technique and the signer's public key, any other healthcare user at a lower level can validate the signature. As the EHR forbids doctors from signing any papers without a valid digital signature, this feature facilitates and improves the secure capture and sharing of patient data.

Minimal or no configuration - Unlike desktop software, this web application doesn't require any local machine configuration. It is simple to use. In the proposed model, a chat bot is used to introduce new users to the website and direct them to desired pages. This bot can be programmed to respond to English text written in natural language.

Reduced cost - The instantaneous availability of patient data frees up healthcare providers' time to concentrate completely on patient care by reducing the time they must spend recording patient information on paper records. Hospitals can avoid repeating tests and reports when patient records are constantly accessible.

DICOM viewing option - An essential component of EHR systems is the imaging information system called DICOM. The cloud storage uses the DICOM standard to store the diagnostic picture data that has been collected from hospitals, clinics, and imaging facilities globally. The DICOM files can be archived via cloud computing, which also offers data security. Healthcare practitioners can access imaging data thanks to the centralized access from anywhere in the world. When the DICOM Standard is available in the healthcare enterprise network, patients may receive faster and more effective care.

VIII. CHALLENGES AND PATH AHEAD

One of the most significant contemporary technologies that can help the state of health care systems all over the world is big data analytics. After the Covid-19 pandemic swept the globe and exposed the frailty of our healthcare systems, governments and different international agencies have been taking concerns relating to the health care sector more seriously than ever. The need of preparing the healthcare systems to handle upcoming difficulties in the post-pandemic era will also enhance significantly the role of technology like big data analytics. But even as we make great strides in closing the gaps in our healthcare systems, it is crucial to research and address some of the difficulties associated with improvising and implementing new technologies, such as big data, in the healthcare industry. This review paper's final portion aims to explore some of the difficulties encountered while implementing big data analytics in the healthcare sector. [8]:

1) Multiple source information management

The sole aim of big data analytics in healthcare sector is to analyze real world medical data to predict a possible outcome. This process completely depends on how data is stored, prepared and mined. In healthcare sector, most of the data generated is in the heterogeneous form. Even though the number of countries using EHRs is growing, analyzing this vast amount of data to produce results is becoming increasingly difficult due to the lack of interoperability in transferring and storing data of various types. In order for different data providers to cooperate with one another and share the data, an appropriate infrastructure must be developed.

Patients frequently visit numerous clinics and hospitals in search of a medical cure for their disease. The use of this data for additional analysis is made easier and more effective if all the patient data gathered across different clinics can be made available on a single platform. However, there is still a long way to go until hospitals and clinics are able to accomplish this interoperability, without which our healthcare systems' efficiency cannot be increased. In order to acquire data in an organized manner that is simpler to analyze and handle, the healthcare industry should also work towards standardizing specific operations.

2) Security and privacy of Data:

In any industry, providing security for user data obtained from them becomes crucial. Data security cannot ever be an option in the modern world. Mutual trust between the system and the user is essential for any system to succeed. To operate effectively and produce better results, a healthcare system must priorities patient privacy and information security. If there is a lack of trust, the patient may decide not to share some crucial health-related information. The operational effectiveness of the system will be directly hampered by the absence of such data, which will have a significant impact on the analysis's final findings.

A new facet of warfare has recently begun to take shape. It is biological warfare. Unsecured access to vast amounts of a nation's inhabitants' medical information could one day result in the collapse of the entire society. As a result, it is crucial for any healthcare system to guarantee the security of the patient data it collects from millions of people. In order to prevent data misuse, it also becomes crucial to guarantee patient information anonymity. Big data analytics will soon play a crucial role in healthcare systems all around the world. Real-time data availability should be ensured in light of this in order to get the greatest results.

*3) **Need to adopt advanced analyzing techniques:***

The amount of data produced in the healthcare industry is growing tremendously in recent years. Unfortunately, new methods for processing and analyzing such vast amounts of data are not being adopted at a rate that is keeping up. The amount of data being generated at this time does not allow for or need the use of typical machine learning techniques. Adopting cutting-edge techniques like predictive analytics, deep machine learning, and graph analytics becomes crucial for the healthcare industry. Without the use of a specific model, novel methods should be created to infer the links between the data. These methods should make it possible for machines to recognize various patterns in vast amounts of unstructured data.

*4) **Data quality**:*

The amount of data being produced in the healthcare industry is varied. The majority of it is unstructured data. Due to their incompleteness, consistency, and inaccuracy, such vast amounts of data represent a key source of worry in big data analytics. In various places, different approaches are taken to deal with health-related challenges. Such inconsistencies could result in the loss of important data. Most hospitals and healthcare facilities are unable to acquire information on a patient's state or health in real time. It can be exceedingly challenging to recognize and treat certain diseases in their early stages when there is a lack of real-time data. Another significant problem in the healthcare sector is determining the veracity of the data. The entire analytical process can be quickly derailed by incorrect or undesirable data, which will have a direct impact on the analysis's final result. The enormous amounts of unstructured data are still the largest problem facing the healthcare industry. In the analytical process, the quality of the data obtained is crucial. Therefore, in order to achieve the best outcomes, it is important to create new tools and methods for gathering and processing high-quality data.

IX. Conclusion

Humanity still has a long way to go before adopting and spreading these specialized technology to the lowest rungs of society. However, the knowledge gained from the Covid-19 pandemic and our unwavering commitment to improving the world will undoubtedly provide us the boost we need to integrate new technologies into our current healthcare systems. Big data analytics will soon play a crucial role in our healthcare systems, as was previously discussed. Therefore, it becomes crucial that we create new methods for removing barriers that prevent the integration of cutting-edge technology like big data analytics into our healthcare systems.

X. Acknowledgement

I would like to thank my guide Dr. Surbhi Agrwal whose expertise was invaluable in formulating this review paper. I would also like to extend my heartfelt gratitude to Dr. Jayapal R, principal, RVITM and other faculty members for their guidance and support.

XI. References

[1] S. Kumar and M. Singh, "Big data analytics for healthcare industry: impact, applications, and tools," in Big Data Mining and Analytics, vol. 2, no. 1, pp. 48-57, March 2019.

[2] Batko, K., Ślęzak, A. The use of Big Data Analytics in healthcare. J Big Data **9**, 3 (2022).

[3] Rakesh Raja, Indrajit Mukherjee, Bikash Kanti Sarkar, "A Systematic Review of Healthcare Big Data", *Scientific Programming*, vol. 2020, Article ID 5471849, 15 pages, 2020.

[4] Siva Sankara Reddy Donthi Reddy., Udaya Kumara Ramanadham Advances in Science Technology and Engineering Systems Journal 2(4):189-196

[5] Pai, M.M.M., Ganiga, R., Pai, R.M. et al. Standard electronic health record (EHR) framework for Indian healthcare system. Health Serv Outcomes Res Method (2021).

[6] Kruse, C.S., Stein, A., Thomas, H., Kaur, H.: The use of electronic health records to support population health: a systematic review of the literature. J. Med. Syst. 42(11), 214 (2018)

[7] Rehman, A., Naz, S. & Razzak, I. Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. Multimedia Systems (2021).

[8] Sayantan Khanra, Amandeep Dhir, A. K. M. Najmul Islam & Matti Mäntymäki (2020) Big data analytics in healthcare: a systematic literature review, Enterprise Information Systems, 14:7, 878-912

[9] Raghupathi, W., Raghupathi, V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst* **2**, 3 (2014).